RISK MANAGEMENT TRENDS

Edited by **Giancarlo Nota**

INTECHWEB.ORG

Risk Management Trends
Edited by Giancarlo Nota

Published by InTech
Janeza Trdine 9, 51000 Rijeka, Croatia

Publishing Process Manager Ana Pantar
Technical Editor Teodora Smiljanic
Cover Designer Jan Hyrat
Image Copyright 18percentgrey, 2010. Used under license from Shutterstock.com

First published July, 2011
Printed in Croatia

A free online edition of this book is available at www.intechopen.com
Additional hard copies can be obtained from orders@intechweb.org

Risk Management Trends, Edited by Giancarlo Nota,
p. cm.
ISBN 978-953-307-314-9

Contents

Preface

Although the etimology of the word *risk* is not certain, two possible sources are truly revealing: riscus and rizq. The mediaeval Latin word *riscus* signifies a reef or a rock sheer from the sea, evoking a sense of danger for the ships. The Arabic *rizq* can instead be interpreted as: all that comes from God, the bare essentials, from which an advantage can be taken. These two different meanings reflect the essential aspects of risks. They express the danger of suffering a loss as a consequence of adverse events but they could also relate to the acquisition of some kind of gain.

As a matter of fact, a given scientific field adopts its own definition of risk. The standard ISO 31000:2009 applicable to any kind of organization, emphasizes the role of uncertainty: "risk is the effect of uncertainty on objectives". According to ISO 31000 and other standards as well, risk management is necessary to achieve objectives, trying to keep away undesirable events but also trying to catch opportunities often related to risks.

Business, social and natural phenomena evolve rapidly and in unforeseen ways today. Things change all the time and risk management requires new concepts and ideas to cope with the uncertainty that comes with the evolving world. Now, more than ever before, it is essential to understand the challenges posed by the new facets that risks can assume. At the same time, acquiring further knowledge on risk management methods can help us to control potential damage or to gain a competitive advantage in a quickly changing world.

The book *Risk Management Trends* offers the results of researchers and practitioners on two wide areas of risk management: business and social phenomena. Chapters 1 and 2 are rather general and could be exploited in several contexts; the first chapter introduces a model where a traditional risk management process is augmented with information and knowledge management processes to improve model quality and usefulness respectively. The second chapter discusses a soft computing based risk management.

Chapters from 3 to 5 deal with project risks. This is a research area where advances are expected in the future. In chapter 3, the attention is on the strategic planning phase of a project when decision maker have to pick a roadmap among several alternatives. In

chapter 4 the assessment of risks associated with climate change in construction projects is approached through the non parametric leave-one-out-cross validation technique. A framework made up of a conceptual architecture and a risk knowledge package structure for collecting and sharing risk knowledge in software projects is presented in chapter 5.

Chapters from 6 to 8 are devoted to the finance. Chapter 6 presents a methodology for risk management in equity portfolios from a long term and short term point of view. Chapter 7 shows an approach to currency trading using the fractal market hypothesis. Chapter 8 focuses on a risk-taking insurance company managing a balance between financial and insurance risks.

Chapter 9 addresses the reorganization for internal security and safety in Norway. This is an emerging research field that has received impulse from the severe shocks such as 9/11 terror attack and the Japanese nuclear reactor hit by the tsunami that caused the evacuation of more than 180,000 people amid meltdown fears.

The last four chapters aim at reporting advances in medicine and pharmaceutical research. In Chapter 10, the concept of Good Dispensing and Delivery Practice (GDDP) is proposed as a system building for risk management on medication. A method to evaluate mental fatigue induced during a visual display terminal task is introduced in chapter 11. Finally, risk management in the development of new products in the pharmaceutical industry and safety monitoring of similar biological products are discussed in chapters 12 and 13 respectively.

I hope that the reader will enjoy reading this book; new ideas on risk management in several fields and many case studies enrich the theoretical presentations making the discussion concrete and effective.

I would like to thank all the contributors to this book for their research efforts. My appreciation also goes to the InTech team that supported me during the publication process.

Giancarlo Nota
Dipartimento di Informatica
Università di Salerno,
Italy

1

Augmenting the Risk Management Process

Jan Emblemsvåg
STX OSV AS[*]
Norway

1. Introduction

> I have seen something else under the sun:
> The race is not to the swift
> or the battle to the strong,
> nor does food come to the wise
> or wealth to the brilliant
> or favour to the learned;
> but time and chance happens to them all.
>
> King Salomon
> Ecclesiastes 9:11

Time and chance happens to them all… – a statement fitting one corporate scandal after the other, culminating by a financial crisis that has demonstrated that major risks were ignored or not even identified and managed, see for example (The Economist 2002, 2009). Before these scandals, risk management was an increasingly hot topic on a wider scale in corporations. For example, the Turnbull Report made at the request of the London Stock Exchange (LSE) '… is about the adoption of a risk-based approach to establishing a system of internal control and reviewing its effectiveness' (Jones and Sutherland 1999), and it is a mandatory requirement for all companies at the LSE. Yet, its effectiveness might be questioned as the financial crisis shows.

Furthermore, we must acknowledge the paradox that the increasing reliance on risk management have in fact lead decision-makers to take risks they normally would not take, see (Bernstein 1996). This has also been clearly demonstrated by one financial institution after the other in the run-up to the financial crisis. Sophisticated risk management and financial instruments lead people astray, see for example (The Economist 2009). Thus, risk management can be a double-edged sword as we either run the risk of ignoring risks (and risk management), or we fall victim to potential deception by risk management.

Nonetheless, there exists numerous risk management approaches, but all suffer from a major limitation: They cannot produce consistent decision support to the extent desired and subsequently they become less trustworthy. As an example; three independent consulting companies performed a risk analysis of a hydro-electric power plant and reached widely different conclusions, see (Backlund and Hannu 2002).

---

[*]Note that the views presented in this chapter are those solely of the author and do not represent the company or any of its stakeholders in any fashion.

This chapter therefore focuses on reducing these limitations and improve the quality of risk management. However, it is unlikely that any approach can be developed that is 100% consistent, free of deception and without the risk of reaching different conclusions. There will always be an element of art, albeit less than today.

The element of art is inescapable partly due to a psychological phenomenon called framing which is a bias we humans have ingrained in us to various degrees, see (Kahneman and Tversky 1979). Their findings have later been confirmed in industry, see for example (Pieters 2004). Another issue is the fact that often we are in situations where we either lack numerical data, or the situation is too complex to allow the usage of numerical data at all. This forces us to apply subjective reasoning in the process concerning probability- and impact estimates regardless whether the estimates themselves are based on nominal-, ordinal-, interval- or ratio scales. For more on these scales, see (Stevens 1946).

We might be tempted to believe that the usage of numerical data and statistics would greatly reduce the subjective nature of risk management, but research is less conclusive. It seems that it has merely altered it. The subjective nature on the individual level is reduced as each case is based on rational or bounded rational analysis, but on an industry level it has become more systemic for a number of reasons:

1. Something called herding is very real in the financial industries (Hwang and Salmon 2004), which use statistical risk management methods. Herding can be defined as a situation when '…a group of investors following each other into (or out of) the same securities over some period of time [original italics]", see (Sias 2004). More generally, herding can be defined as '…behaviour patterns that are correlated across individuals", see (Devenow and Welch 1996).
2. Investors have a tendency to overreact (De Bondt and Thaler 1985), which is human, but not rational.
3. Lack of critical thinking in economic analyses is a very common problem particularly when statistical analyses are involved – it is a kind of intellectual herding. For example, two economists, Deirdre McCloskey and Stephen Ziliak studied to what degree papers in the highly respected journal American Economic Review failed to separate statistical significance from plausible explanations of economic reality, see (The Economist 2004). Their findings are depressing: first, in the 1980s 70 % of the papers failed to distinguish between economic - and statistical significance, and second, in the 1990s more than 80 % failed. This is particularly a finding that researchers must address because the number among practitioners is probably even worse, and if researchers (and teachers) cannot do it correctly we can hardly expect practitioners to show the way.

Clearly, subjectivity is a problem for risk management in one way or the other as discussed. The purpose of this chapter is therefore to show how augmenting the risk management process will reduce the degree of subjectivity to a minimum and thereby improve the quality of the decision support.

Next, some basic concepts – risk and uncertainty – are introduced. Without useful definitions of risk and uncertainty, an enlightening discussion is impossible. Then, in Section 3, a common – almost 'universal' – risk management approach is presented. Then, in Section 4, an improved approach – the augmented risk management approach – is presented. Critical evaluation of the approach and future ideas are discussed in Section 5. A closure is provided in Section 6. A simple, functional case is provided along for illustrational purposes.

2. Introducing risk and uncertainty

Risk and uncertainty are often used interchangeably. For example, (Friedlob and Schleifer 1999) claim that for auditors 'risk is uncertainty'. It may be that distinguishing between risk and uncertainty makes little sense for auditors, but the fact is that there are many basic differences as explained next. First, risk is discussed from traditional perspectives, and the sources of risks are investigated. Second, the concept of uncertainty is explored. Finally, a more technical discussion about probability and possibility is conducted to try to settle an old score in some of the literature.

2.1 Risk

The word 'risk' derives from the early Italian word *risicare*, which originally means 'to dare'. In this sense risk is a choice rather than a fate (Bernstein 1996). Other definitions also imply a choice aspect. Risk as a general noun is defined as 'exposure to the chance of injury or loss; a hazard or dangerous chance' (Webster 1989). Along the same token, in statistical decision theory risk is defined as 'the expected value of a loss function' (Hines and Montgomery 1990). Thus, various definitions of risk imply that we expose ourselves to risk by choice. Risk is measured, however, in terms of 'consequences and likelihood' (Robbins and Smith 2001; Standards Australia 1999) where likelihood is understood as a 'qualitative description of probability or frequency', but frequency theory is dependent on probability theory (Honderich 1995). Thus, risk is ultimately a probabilistic phenomenon as it is defined in most literature.

It is important to emphasize that 'risk is not just bad things happening, but also good things not happening' (Jones and Sutherland 1999) – a clarification that is particularly crucial in risk analysis of social systems. Many companies do not fail from primarily taking 'wrong actions', but from not capitalizing on their opportunities, i.e., the loss of an opportunity. As (Drucker 1986) observes, 'The effective business focuses on opportunities rather than problems'. Risk management is ultimately about being proactive.

It should also be emphasized that risk is perceived differently in relation to gender, age and culture. On an average, women are more risk averse than men, and more experienced managers are more risk averse than younger ones (MacCrimmon and Wehrung 1986). Furthermore, evidence suggests that successful managers take more risk than unsuccessful managers. Perhaps there are ties between the young managers' 'contemporary competence' and his exposure to risks and success? At any rate, our ability to identify risks is limited by our perceptions of risks. This is important to be aware of when identifying risks – many examples of sources of risks are found in (Government Asset Management Committee 2001) and (Jones and Sutherland 1999).

According to a 1999 Deloitte & Touche survey the potential failure of strategy is one of the greatest risks in the corporate world. Another is the failure to innovate. Unfortunately, such formulations have limited usefulness in managing risks as explained later – is 'failure of strategy' a risk or a consequence of a risk? To provide an answer we must first look into the concept of uncertainty since 'the source of risk is uncertainty' (Peters 1999). This derives from the fact that risk is a choice rather than a fate and occurs whenever there are one-to-many relations between a decision and possible future outcomes, see Figure 1.

Finally, it should be emphasized that it is important to distinguish between the concept of probability, measures of probability and probability theory, see (Emblemsvåg 2003). There is much dispute about the subject matter of probability (see (Honderich 1995)). Here, the idea

that probability is a 'degree of belief' is subscribed to, but that it can be measured in several ways out of which the classical probability calculus of Pascal and others is the best known. For simplicity and generality the definition of risk found in (Webster 1989) is used here – the 'exposure to the chance of injury or loss; a hazard or dangerous chance'. Furthermore, 'degree of impact and degree of belief' is used to measure risk.
One basic tenet of this chapter is that there are situations where classic probability calculus may prove deceptive in risk analyses. This is not to say, however, that probability theory should be discarded altogether – we simply believe that probability theory and other theories can complement each other if we understand when to use what. Concerning risk analysis, it is argued that other theories provide a better point of departure than the classic probability theory, but first the concept of uncertainty is explored, which is done next.

2.2 Uncertainty

Uncertainty as a general noun is defined as 'the state of being uncertain; doubt; hesitancy' (Webster 1989). Thus, there is neither loss nor gain necessarily associated with uncertainty; it is simply the not known with certainty – not the unknown.
Some define uncertainty as 'the inability to assign probability to outcomes', and risk is regarded as the 'ability to assign such probabilities based on differing perceptions of the existence of orderly relationships or patterns' (Gilford, Bobbitt *et al.* 1979). Such definitions are too simplistic for our purpose because in most situations the relationships or patterns are not orderly; they are complex. Also, the concepts of gain and loss, choice and fate and more are missed using such simplistic definitions.
Consequently, uncertainty and complexity are intertwined and as an unpleasant side effect, imprecision emerges. Lotfi A. Zadeh formulated this fact in a theorem called the Law of Incompatibility (McNeill and Freiberger 1993):

As complexity rises, precise statements lose meaning
and meaningful statements lose precision.

Since all organizations experience some degree of complexity, this theorem is crucial to understand and act in accordance with. With complexity we refer to the state in which the cause-and-effect relationships are loose, for example, operating a sailboat. A mechanical clock, however, in which the relationship between the parts is precisely defined, is complicated – not complex. From the Law of Incompatibility we understand that there are limits to how precise decision support both can and should be (to avoid deception), due to the inherent uncertainty caused by complexity. Therefore, by increasing the uncertainty in analyses and other decision support material to better reflect the true and inherent uncertainty we will actually lower the actual risk.
In fact, Nobel laureate Kenneth Arrow warns us that '[O]ur knowledge of the way things work, in society or in Nature, comes trailing clouds of vagueness. Vast ills have followed a belief in certainty' (Arrow 1992). Basically, ignoring complexity and/or uncertainty is risky, and accuracy may be deceptive. The NRC Governing Board on the Assessment of Risk shares a similar view, see (Zimmer 1986). Thus, striking a sound balance between meaningfulness and precision is crucial, and possessing a relatively clear understanding of uncertainty is needed since uncertainty and complexity are so closely related.
Note that there are two main types of uncertainty, see Figure 1, fuzziness and ambiguity. Definitions in the literature differ slightly but are more or less consistent with Figure 1.

Fuzziness occurs whenever definite, sharp, clear or crisp distinctions are not made. Ambiguity results from unclear definitions of the various alternatives (outcomes). These alternatives can either be in conflict with each other or they can be unspecified. The former is ambiguity resulting from discord whereas the latter is ambiguity resulting from nonspecificity. The ambiguity resulting from discord is essentially what (classic) probability theory focuses on, because 'probability theory can model only situations where there are conflicting beliefs about mutually exclusive alternatives' (Klir 1991). In fact, neither fuzziness nor nonspecificity can be conceptualized by probability theories that are based on the idea of 'equipossibility' because such theories are 'digital' in the sense that degrees of occurrence is not allowed – it either occurs or not. Put differently, uncertainty is too wide of a concept for classical probability theory, because it is closely linked to equipossibility theory, see (Honderich 1995).

Kangari and Riggs (1989) have discussed the various methods used in risk analysis and classified them as either 'classical' (probability based) or 'conceptual' (fuzzy set based). Their findings are similar:

> *… probability models suffer from two major limitations. Some models require detailed quantitative information, which is not normally available at the time of planning, and the applicability of such models to real project risk analysis is limited, because agencies participating in the project have a problem with making precise decisions. The problems are ill-defined and vague, and they thus require subjective evaluations, which classical models cannot handle.*

To deal with both fuzziness and nonspecific ambiguity, however, Zadeh invented fuzzy sets – 'the first new method of dealing with uncertainty since the development of probability' (Zadeh 1965) – and the associated possibility theory. Fuzzy sets and possibility theory handle the widest scope of uncertainty and so must risk analyses. Thus, these theories seem to offer a sound point of departure for an augmented risk management process.

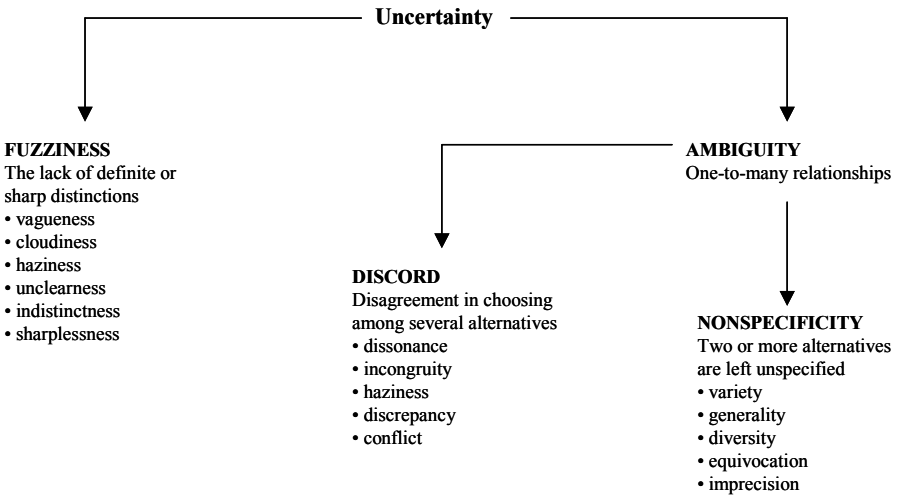


Fig. 1. The basic types of uncertainty (Klir and Yuan 1995)

For the purpose of this chapter, however, the discussion revolves around how probability can be estimated, and not the calculus that follows. In this context possibility theory offers some important ideas explained in Section 2.3. Similar ideas seem also to have been absorbed by a type of probability theory denoted 'subjective probability theory', see e.g. (Roos 1998). In fact, here, we need not distinguish between possibility theory and subjective probability theory because the main difference between those theories lies in the calculus, but the difference in calculus is of no interest to us. This is due to the fact that we only use the probability estimates to rank the risks and do not perform any calculus.
In the remainder of this chapter the term 'classic probability theory' is used to separate it from subjective probability theory.

2.3 Probability theory versus possibility theory

The crux of the difference between classic probability theory and possibility theory lies in the estimation technique. For example, consider the Venn diagram in Figure 2. The two outcomes A and B in outcome space S overlap, i.e., they are not mutually exclusive. The probability of A is in other words dependent on the probability of B, and *vice versa*. This situation is denoted nonspecific ambiguity in Figure 1.

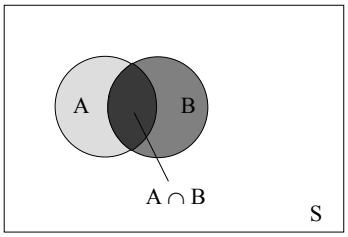


Fig. 2. Two non-mutually exclusive outcomes in outcome space S

In classic probability theory we look at A in relation to S and correct for overlaps so that the sum of all outcomes will be 100% (all exhaustible). In theory this is straightforward, but in practice calculating the probability of A * B is problematic in cases where A and B are interdependent and the underlying cause-and-effect relations are complex. Thus, in such cases we find that the larger the probability of A * B, the larger may the mistake of using classic probability theory become.
In possibility theory, however, we simply look at the outcomes in relation to each other, and consequently S becomes irrelevant and overlaps do not matter. The possibility of A will simply be A to A + B in Figure 2. Clearly, possibility theory is intuitive and easy, but we pay a price - loss of precision (an outcome in comparison to outcome space) both in definition (as discussed here) and in its further calculus operations (not discussed here). This loss of precision is, however, more true to high levels of complexity and that is often crucial because 'firms are mutually dependent' (Porter 1998). Also, it is important that risk management approaches do not appear more reliable than they are because then decision-makers can be lead to accept decisions they normally would reject, as discussed earlier.
This discussion clearly illustrates that '[classic] probabilistic approaches are based on counting whereas possibilistic logic is based on relative comparison' (Dubois, Lang *et al.*). There are also other differences between classic probability theory and possibility theory, which is not discussed here. It should be noted that several places in the literature the word

'probability' is used in cases that are clearly possibilistic. This is probably more due to the fact that 'probability' is a common word – which has double meaning (Bernstein 1996) – than reflecting an actual usage of classic probability theory and calculus.
One additional difference that is pertinent here is the difference between 'event' and 'sensation'. The term 'event' applied in probability theory requires a certain level of distinctiveness in defining what is occurring and what is not. 'The term 'sensation' has therefore been proposed in possibility theory, and it is something weaker than an event' (Kaufmann 1983). The idea behind 'sensation' is important in corporate settings because the degree of distinctness that the definition of 'event' requires is not always obtainable.
Also, the term 'possibility' is preferred here over 'probability' to emphasize that positive risks – opportunities, or possibilities – should be pursued actively. Furthermore, using a possibilistic foundation (based on relative ordering as opposed to the absolute counting in classic probability theory), provides added decision support because 'one needs to present comparison scenarios that are located on the probability scale to evoke people's own feeling of risk' (Kunreuther, Meyer *et al.* 2004).
To summarize so far: the (Webster 1989) definition of risk is used – the 'exposure to the chance of injury or loss; a hazard or dangerous chance' – while risk is measured in terms of 'degree of impact' and 'degree of belief'. Furthermore, the word 'possibility' is used to denote estimate the degree of belief of a specific sensation. Alternatively, probability theoretical terms can be employed under the explicit understanding that the terms are not 100% correct – this may be a suitable approach in many cases when practitioners are involved because fine-tuned terms can be too difficult to understand.
Next, a more or less standard risk management process is reviewed.

3. Brief review of risk management approaches

All risk management approaches known to the author are variations of the framework presented in Figure 3. They may differ in wording, number of steps and content of steps, but the basic principles remain the same, see (Meyers 2006) for more examples and details. The discussion here is therefore related to the risk management process shown in Figure 3. The depicted risk management process can be found in several versions in the literature, see for example public sources such as (CCMD Roundtable on Risk Management 2001; Government Asset Management Committee 2001; Jones and Sutherland 1999) and it is employed by risk management specialists such as the maritime classing society Det Norske Veritas (DNV)[1]. The fact that the adherence to the same standards leads to different implementations is also discussed by (Meyers 2006).
Briefly stated, the process proceeds as follows: In the initial step, all up-front issues are identified and clarified. Proposal refers to anything for which decision support is needed; a project proposal, a proposal for a new strategy and so on. The objectives are important to clarify because risks arise in pursuit of objectives as discussed earlier. The criteria are essentially definitions of what is 'good enough'. The purpose of defining the key elements is to provide relevant categorization to ease the risk analysis. Since all categorization is deceptive to some degree, see (Emblemsvåg and Bras 2000), it is important to avoid unnecessary categories. The categories should therefore be case specific and not generic.

[1] Personal experience as consultant in Det Norske Veritas (DNV).

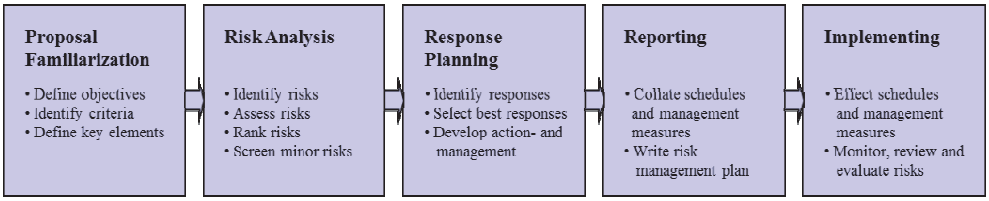


Fig. 3. Traditional risk management process. Based on (Government Asset Management Committee 2001)

The second step is the analysis of risks by identification, assessment, ranking and screening out minor risks. This step is filled with shortcomings and potential pitfalls of the serious kind. This step relies heavily on subjectivism, and that is a challenge in itself because it can produce widely different results as (Backlund and Hannu 2002) point out. The challenge was that there existed no consistent decision support for improving the model other than to revise the input – sadly sometimes done to obtain preconceived results. For example, suppose we identified three risks – A, B and C – and want to assess their probabilities and impacts, see Figure 4. The assessment is usually performed by assigning numbers that describe probability and impact, but the logic behind the assignment is unclear at best, and it is impossible to perform any sort of analysis to further improve this assignment. Typically, the discussion ends up by placing the risks in a matrix like the ones shown in Figure 4, but without any consistency checks it is difficult to argue which one, if any, of the two matrices in Figure 4 fit reality the best. Thus, the recommendations can become quite different, and herein lays one of the most problematic issues of this process. In the augmented risk management process this problem is overcome, as we shall see later.

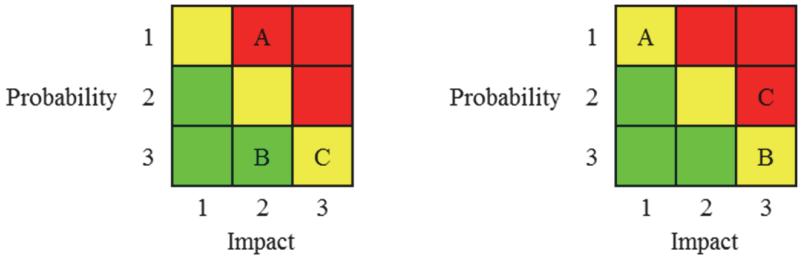


Fig. 4. The arbitrary assignment of probability and impact in a risk ranking matrix

The third step – response planning, or risk management strategies – depends directly on the risk analysis. If the assignment is as arbitrary as the study of (Backlund and Hannu 2002) shows, then the suggested responses will vary greatly. Thus, a more reliable way of analysing risks must be found, which is discussed in Section 4. Nonetheless, there are four generic risk management strategies; 1) risk prevention (reduce probability), 2) impact mitigation (reduce impact), 3) transfer (risk to a third party such as an insurance company) or simply 4) accept (the risk). Depending on the chosen risk management strategy, specific action plans are developed.

The fourth step is often an integral part of step three, but in some projects it may be beneficial to formalize reporting into a separate step, see (Government Asset Management Committee 2001) for more information.

The fifth step – implementation (of the action schedules, management measures and allocation of management resources and responsibilities) is obviously an important step in risk management. It is vital that the effectiveness of these measures must be monitored and checked to secure effective implementation. Possible new risks must also be identified, and so the risk management process starts all over again. Just like the famous PDCA circle, the risk management process never stops.

In addition to the obvious problems with the risk analysis as argued earlier, the entire risk management process lacks three important aspects that aggravate the problems:

1. The capabilities of the organization – the strengths and weaknesses – are either ignored or treated as implicit at best. This is a problem in itself because we cannot rely on responses that cannot be implemented. Understanding that risks are relative to the organization's capabilities is a leap for risk analysis in direction of strategic analysis, which has often incorporated this factor. In other words, risk management should be regarded just as much as management of capabilities as management of risks. If an analysis shall provide recommendations for actions, it is clear that the capabilities, which can be managed, are needed in the risk analysis as well. In this chapter using risk management in strategy is not discussed, so interested readers on how this can be done are referred to (Emblemsvåg and Kjølstad 2002).
2. There is no management of information quality. Management of information quality is crucial in risk management because uncertainty is prevalent. Uncertainty can be defined as a state for which we lack information, see (Emblemsvåg and Kjølstad 2002). Thus, uncertainty analysis should play an integral part in risk management to ensure that the uncertainty in the risk management process is kept at an economically feasible level. The same argument also holds for the usage of sensitivity analyses; both on risk- and uncertainty analyses. This idea is also supported by (Backlund and Hannu 2002).
3. There is no explicit management of either existing knowledge that can be applied to improve the quality of the analyses, or to improve the knowledge acquired in the process at hand which can be used in the follow-up process. The augmented risk management approach therefore incorporates Knowledge Management (KM). KM is believed to be pivotal to ensure an effective risk management process by providing context and learning possibilities. In essence, risk management is not just about managing risks – the entire context surrounding the risks must be understood and managed effectively. Neef (2005) states that 'Risk management is knowledge management', but the point is that the reverse is also important.

This is where the greatest methodological challenge for the augmented risk management process lies – how to manage knowledge. According to (Wickramasinghe 2003), knowledge management in its broadest sense refers to how a firm acquires, stores and applies its own intellectual capital, and according to (Takeuchi 1998), Nonaka insisted that knowledge cannot be 'managed' but 'led'. Worse, we are still not sure what knowledge management really involves (Asllani and Luthans 2003). These aspects, along with the augmented risk management process are elaborated upon in the next section.

4. The augmented risk management process

The augmented risk management process is presented in Figure 5, and it is organized into five steps as indicated by a number, title and colour band (greyish or white). Furthermore, each step consists of three parallel processes; 1) the actual risk management process, 2) the

information management process to improve the model quality and 3) the KM process to improve the usefulness of the model. These steps and processes are explained next, section by section. At the end of each section a running, real-life case is provided for illustrational purposes.

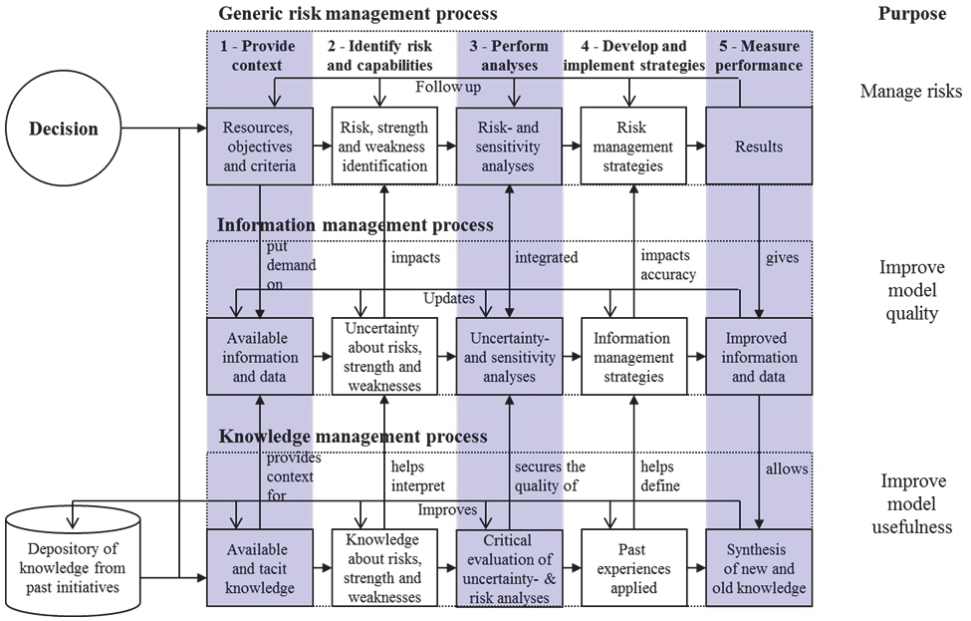


Fig. 5. The augmented risk management process

4.1 Step 1 – provide context

A decision triggers the entire process (note that to not make a conscious decision is also a decision). The context can be derived from the decision itself and the analyses performed prior to the decision, which are omitted in Figure 5. The context includes the objectives, the criteria, measurements for determining the degree of success or failure, and the necessary resources. Identifying relevant knowledge about the situation is also important. The knowledge is either directly available or it is tacit[2], and the various types of knowledge may interplay as suggested by the SECI model[3], see (Nonaka and Takeuchi 1995). Tacit knowledge can be either implicit or really tacit (Li and Gao 2003), and it is often the most valuable because it is a foundation for building sustainable competitive advantage, but it is unfortunately less available, see (Cavusgil, Calantone *et al.* 2003). Residing in the mind of employees, as much tacit knowledge as possible should be transferred to the organization

[2] The dichotomy of tacit- and explicit knowledge is attributable to (Polanyi, M. 1966), who found that tacit knowledge is a kind of knowledge that cannot be readily articulated because it is elusive and subjective. Explicit knowledge, however, is the written word, the articulated and the like.

[3] SECI (Socialization, Externalization, Combination, and Internalization) represents the four phases of the conversions between explicit and tacit knowledge. Often, the starting points of conversion cycles start from the phase of socialization (Li, M. and F. Gao (2003).

and hence become explicit knowledge, as explained later. How this can be done in reality is a major field of research. In fact, (Earl 2001) provides a comprehensive review of KM and proposes seven schools of knowledge management. As noted earlier, even reputed scholars of the field question the management of knowledge…

Therefore, this chapter simply tries to map out some steps in the KM process that is required without claiming that this is *the* solution. The point here is merely that we must have a conscious relationship towards certain basic steps such as identifying what we know, evaluate what takes place, learn from it and then increase the pool of what we know. How this (and possibly more steps) should be done most effectively, is a matter for future work. Currently, we do not have a tested solution for the KM challenge, but a potentially workable idea is presented in Section 5.

From the objectives, resources, criteria and our knowledge we can determine what information is needed and map what information and data is available. Lack of information at this stage, which is common, will introduce uncertainty into the entire process. By identifying lacking information and data we can already early in the process determine if we should pursue better information and data. However, we lack knowledge about what information and data would be most valuable to obtain, which is unknown until Step 3.

Compared to traditional risk management approaches the most noticeable difference in this step is that explicit relations between context and knowledge are established to identify the information and knowledge needs. Typical procedures- and systems of knowledge that can be used include (Neef 2005):

1. Knowledge mapping – a process by which an organization determines 'who knows what' in the company.
2. Communities of practice – naturally-forming networks of employees with similar interests or experience, or with complementary skills, who would normally gather to discuss common issues.
3. Hard-tagging experts – a knowledge management process that combines knowledge mapping with a formal mentoring process.
4. Learning – a post-incident assessment process where lessons learned are digested.
5. Encouraging a knowledge-sharing culture – values and expectations for ethical behaviour are communicated widely and effectively throughout the organization.
6. Performance monitoring and reporting – what you measure is what you get.
7. Community and stakeholder involvement – help company leaders sense and respond to early concerns from these outside parties (government, unions, non-governmental or activist groups, the press, etc.), on policy matters that could later develop into serious conflicts or incidents.
8. Business research and analysis – search for, organize and distribute information from internal and external sources concerning local political, cultural, and legal concerns.

Running case

The decision-maker is a group of investors that wants to find out if it is worth investing more into a new-to-the-world transportation concept in South Korea. They are also concerned about how to attract new investors. A company has been incorporated to bring the new technology to the market. The purpose of the risk management process is to map out potential risks and capabilities and identify how they should be handled. The direct objectives of the investors related to this process are to; 1) identify if the new concept is viable, and if it is to 2) identify how to convince other investors to join.

The investors are experienced people working in mass transit for years, so some knowledge about the market was available. Since the case involves a new-to-the world mass transit solution, there is little technical- and business process knowledge to draw from other than generic business case methods from the literature.

4.2 Step 2 – Identify risks and capabilities

Once a proper context is established, the next step is to identify the risks and the capabilities of the organization. Here, the usage of the SWOT framework is very useful, see (Emblemsvåg and Kjølstad 2002), substituting risks for threats and opportunities, and organizational capabilities for strengths and weaknesses. Identifying the capabilities is to determine what risk management strategies can be successfully deployed.

This step is similar to the equivalent step in traditional approaches, but some differences exist. First, risks are explicitly separated from uncertainties. Risks arise due to decisions made, while uncertainty is due to lacking information, see (Emblemsvåg and Kjølstad 2002). Risks lurk in uncertainty as it were, but uncertainties are not necessarily associated with loss and hence are not interchangeable with risks. Separating uncertainties from risks may seem of academic interest, but uncertainty has to do with information management and hence improvement of model quality, see Figure 5, while risks is the very objective of the model. The findings of (Backlund and Hannu 2002) also support this ascertainment.

Second, the distinction between capabilities and risks is important because capabilities are the means to the end (managing risks in pursuit of objectives). Often, risks, uncertainties and capabilities are mingled which inhibits effective risk management.

Third, for any management tool to be useful it must be anchored in real world experiences and knowledge. Neither the risk management process nor the information management process can provide such anchoring. Consequently, it is proposed to link both the risk management and information management processes to a KM process so that knowledge can be effectively applied in the steps. Otherwise we run the risk of, for example, only identifying obvious risks and falling prey to local 'myths', stereotypes and the like. For more information on how to do this, consult the 'continuous improvement' philosophy and approaches of Deming as described in (Latzco and Saunders 1995) and double loop learning processes as presented by (Argyris 1977, 1978).

Running case

The viability of the concept was related to 5 risk categories; 1) finance, 2) technology, 3) organizational (internal), 4) market and 5) communication. The latter is important in this case because an objective is to attract investors.

We started by reviewing all available documentation about the technology, business plans, marketing plans and whatever we thought were relevant after the objectives had been clarified. We identified more than 200 risks. Then, we spent about a week with top management, in which we also interviewed the director of a relevant governmental research institute and other parties, for a review of the technology and various communication and marketing related risks.

Based on this information we performed a SWOT after which 39 risks remained significant. The vast reduction in the number of risks occurs, as the documentation did not contain all that was relevant. In due course, this fact was established as a specific communication risk. To reduce the number of risks even further we performed a traditional screening of the 39 risks down to 24 and then proceeded to Step 3. This screening totally eliminated the organizational (internal) risks, so we ended up with 4 risk categories.

4.3 Step 3 – perform analyses

As indicated in Figure 5, we propose to have four types of analyses that are integrated in the same model; 1) a risk analysis, 2) a sensitivity analysis of the risk analysis, 3) an uncertainty analysis and 4) a sensitivity analysis of the uncertainty analysis. The purpose of these analyses is not just to analyse risks but to also provide a basis for double-loop learning, that is, learning with feedback both with respect to information and knowledge. Most approaches lack this learning capability and hence lack any systematic way of improving themselves. The critical characteristic missing is consistency.

All these four analyses can be conducted in one single model if the model is built around a structure similar to Analytical Hierarchy Process (AHP). The reason for this is that AHP is built using mathematics, and a great virtue of mathematics is its consistency – a trait no other system of thought can match. Despite the inherent translation uncertainty between qualitative and quantitative measures, the only way to ensure consistent subjective risk analyses is to translate the subjective measures into numbers and then perform some sort of consistency check. The only approach that can handle qualitative issues with controlled consistency is AHP and variations thereof.

Thomas Lorie Saaty developed AHP in the late 1960s to primarily provide decision support for multi-objective selection problems. Since then, (Saaty and Forsman 1992) have utilized AHP in a wide array of situations including resource allocation, scheduling, project evaluation, military strategy, forecasting, conflict resolution, political strategy, safety, financial risk and strategic planning. Others have also used AHP in a variety of situations such as supplier selection (Bhutta and Huq 2002), determining measures of business performance (Cheng and Li 2001), and in quantitative construction risk management of a cross-country petroleum pipeline project in India (Dey 2001).

The greatest advantage of the AHP concept, for our purpose, is that it incorporates a logic consistency check of the answers provided by the various participants in the process. As (Cheng and Li 2001) claim; 'it [AHP] is able to prevent respondents from responding arbitrarily, incorrectly, or non-professionally'. The arbitrariness of Figure 4 will consequently rarely occur. Furthermore, the underlying mathematical structure of AHP makes sensitivity analyses both with respect to the risk- and the uncertainty analysis meaningful, which in turn guides learning efforts. This is impossible in traditional frameworks. How Monte Carlo methods can be employed is shown in (Emblemsvåg and Tonning 2003). The theoretical background for this is explained thoroughly in (Emblemsvåg 2003), to which the interested reader is referred.

The relative rankings generated by the AHP matrix system can be used as so called subjective probabilities or possibilities as well as relative impacts or even relative capabilities. The estimates will be relative, but that is sufficient since the objective of a risk analysis is to effectively direct attention towards the critical risks so that they will be attended to. However, by including a known absolute reference in the AHP matrices we can provide absolute ranking if desired.

The first step in applying the AHP matrix system is to first identify the risks we want to rank, which is done in step 2. Second, due to the hierarchical nature of AHP we must organize the items as a hierarchy. For example, all risks are divided into commercial risks, technological risks, financial risks, operational risks and so on. These risk categories is then broken down into detailed risks. For example, financial risks may consist of cash flow exposure risks, currency risks, interest risks and so forth. It is important that the number of

children below a parent in a hierarchy is not more than 9, because human cognition has great problems handling more than 9 issues at the same time, see (Miller 1956). In our experience, it is wise to limit oneself to 7 or less children per parent simply because being consistent across more than 7 items in a comparison is very difficult. Third, we must perform the actual pair-wise comparison.

To operationalize pair-wise comparisons, we used the ordinal scales and the average Random Index (RI) values provided in Tables 1 and 2 – note that this will per default produce 1 on the diagonals. According to (Peniwati 2000), the RIs are defined to allow a 10% inconsistency in the answers. Note that the values in Table 1 must be interpreted in its specific context. Thus, when we speak of probability of scale 1 it should linguistically be interpreted as 'equally probable'. This may seem unfamiliar to most, but it is easier to see how this work by using the running example. First, however, a quick note on the KM side of this step should be mentioned.

Intensity of Importance (1)	Definition (2)	Explanation (3)
1	Equal importance	Two items contribute equally to the objective
3	Moderate importance	Experience and judgment slightly favor one over another
5	Strong importance	Experience and judgment strongly favor one over another
7	Very strong importance	An activity is strongly favored and its dominance is demonstrated in practice
9	Absolute importance	The importance of one over another affirmed on the highest possible order
2, 4, 6, 8	Intermediate values	Used to represent compromise between the priorities listed above
Reciprocals of above numbers		If item *i* has one of the above non-zero numbers assigned to it when compared to with item *j*, the *j* has the reciprocal value when compared with *i*

Table 1. Scales of measurement in pair-wise comparison. Source: (Saaty, Thomas Lorie 1990)

Matrix Size	Random Index	Recommended CR Values
1	0.00	0.05
2	0.00	0.05
3	0.58	0.05
4	0.90	0.08
5	1.12	0.10
6	1.24	0.10
7	1.32	0.10
8	1.41	0.10
9	1.45	0.10
10	1.49	0.10

Table 2. Average Random Index values. Source: (Saaty, Thomas Lorie 1990)

From a KM perspective the most critical aspect of this step is to critically review the aforementioned analyses. A critical review will in this context revolve around finding answers for a variety of 'why?' questions as well as judging to what extent the analyses provide useful input to the risk management process and what must be done about significant gaps. Basically, we must understand how the analyses work, why they work and to what extent they work as planned. The most critical part of this is ensuring correct and useful definitions of risks and capabilities (Step 2). In any case, this step will reveal the quality of the preceding work – poor definitions will make pair-wise comparison hard.

Running case

From Step 2 we recall that there are 4 risk categories; 1) finance (FR), 2) technology (TR), 3) market (MR) and 4) communication (CR). Since AHP is hierarchical we are tempted to also rank these, but in order to give all the 39 risks underlying these 4 categories the same weight – 25% - we do not rank them (or give them the same rank, i.e. 1). Therefore, for our running example we must go to the bottom of the hierarchy and in the market category, for example, we find the following risks:

1. Customer decides to not buy any project (MR1).
2. Longer lead-times in sales than expected (MR2).
3. Negative reactions from passengers due to the 90 degree turn (MR3).
4. Passengers exposed to accidents/problems on demo plant (MR4).
5. Wrong level of 'finished touch' on Demo plant (MR5).

The pair-wise comparison of these is a three-step process. The first step is to determine possibilities, see Table 3, whereas the second step is to determine impacts. When discussing impacts it is important to use the list of capabilities and think of impact in their context.

From Table 3 we see that MR2 (the second Market Risk) is perceived as the one with the highest possibility (47%) of occurrence. Indeed, it took about 10 years from this analysis first was carried out – using the risk management approach presented in (Emblemsvåg and Kjølstad 2002) – until it was decided to build the first system. We see from Table 2 that the CR value in the matrix of 0.088 is less than the recommended CR value of 0.10. This implies that the matric internally consistent and we are ready to proceed. A similar matrix should have been constructed concerning impacts, but this is omitted here. The impacts would also have been on a 0 to 1 percentage scale, so that when we multiply the possibilities and the impacts we get small numbers that can be normalized back on a 0 to 1 scale in percentages. This is done in Table 4 for the top ten risks.

	R1	R2	R3	R4	R5	Possibility
R1	1	0.14	0.20	3.00	0.33	8 %
R2	7.00	1	3.00	5.00	4.00	47 %
R3	5.00	0.33	1	4.00	3.00	26 %
R4	0.33	0.20	0.25	1	0.33	6 %
R5	3.00	0.25	0.33	3.00	1	14 %
Sum	16.33	1.93	4.78	16.00	8.67	
CR value						0.088

Table 3. Calculation of possibilities (subjective probabilities)

From Table 4 we see that the single largest risk is Financial Risk (FR) number 5, which is 'Payment guarantees not awarded'. It accounts for 27% of the total risk profile. Furthermore, the ten largest risks account for more than 80% of the total risk profile.

The largest methodological challenge in this step is to combine the risks and capabilities. In (Emblemsvåg and Kjølstad 2002), the link was made explicit using a matrix, but the problem of that approach is that it requires an almost inhuman ability of thinking of risks independently of capabilities first and then think of it extremely clearly afterward when linking the risks and capabilities. The idea was good, but too difficult to use. It is therefore much more natural – almost inescapable, less time consuming and overall better to implicitly think of capabilities when we rate impacts and possibilities. A list of the capabilities is handy nonetheless to remind ourselves of what we as a minimum should take into consideration when performing the risk analysis.

At the start of this section, we proposed to have four types of analyses that are integrated in the same model; 1) a risk analysis, 2) a sensitivity analysis of the risk analysis, 3) an uncertainty analysis and 4) a sensitivity analysis of the uncertainty analysis. So far, the latter three remains. The key to their execution is to model the input in the risk analysis matrices in two ways;

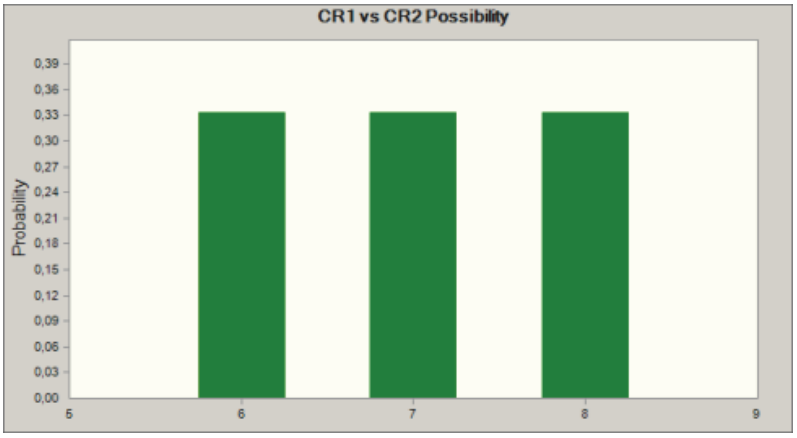
1. Using symmetric distributions, such as symmetric 1 (around the values initially set in the AHP matrices) and uniform distributions shown to the left in Figure 6. It is important that they are symmetric in order to make sure that the mathematical impact on the risk analysis of each input is traced correctly. This will enable us to trace accurately what factors impact the overall risk profile the most – i.e., key risk factors.
2. Modelling uncertainty as we perceive it as shown to the right in Figure 6. This will facilitate both an estimate on the consequences of the uncertainty in the input in the process as well as sensitivity analysis to identify what input needs improvement to most effectively reduce the overall uncertainty in the risk analysis – i.e., key uncertainty factors.

Risks		Possibility	Impact	Risk norm	Risk, acc.
FR5	Payment guarantees not awarded	10 %	12 %	27 %	27 %
FR4	No exit strategy for foreign investors	8 %	8 %	15 %	42 %
TR7	Undesirable mechanical behavior (folding and unfolding)	6 %	6 %	9 %	50 %
TR1	Competitors attack NoWait due to safety issues	9 %	3 %	6 %	56 %
MR3	Negative reactions from passengers due to the 90 degree turn	6 %	4 %	6 %	62 %
MR2	Longer lead-times in sales than expected	12 %	2 %	5 %	67 %
CR1	Business essentials are not presented clearly	9 %	2 %	5 %	72 %
MR4	Passengers exposed to accidents/problems on demo plant	1 %	13 %	4 %	76 %
CR4	Business plan lack focus on benefits	6 %	2 %	3 %	79 %
MR1	Customer decides to not buy any project	2 %	5 %	3 %	82 %

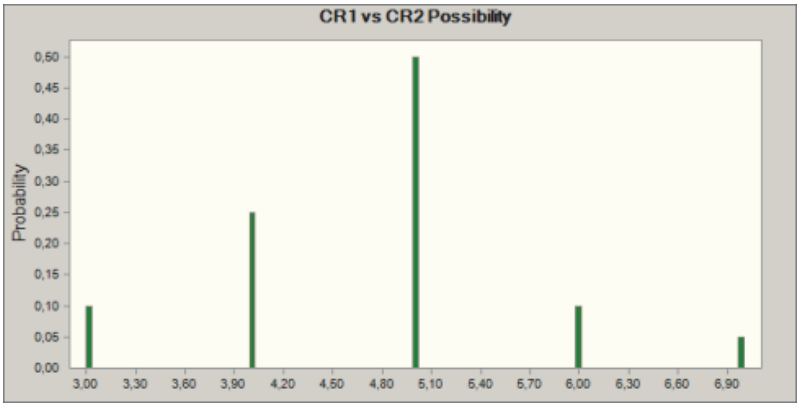
Table 4. The ten largest risks in descending order

Before we can use the risk analysis model, we have to check the quality of the matrices. With 4 risk categories we get 8 pair-wise comparison matrices (5 with possibility estimates and 5 with impact estimates). Therefore, we first run a Monte Carlo simulation of 10,000 trials and record the number of times the matrices become inconsistent. The result is shown in the histogram on top in Figure 7. We see that the initial ranking of possibilities and impacts created only approximately 17% consistent matrices (the column to the left), and this is not good enough. The reason for this is that too many matrices had CR values of more than approximately 0.030. Consequently, we critically evaluated the pair-wise comparison matrices to reduce the CR values of all matrices to below 0.030. This resulted in massive improvements – about 99% of the matrices in all 10,000 trials remained consistent. This is excellent, and we can proceed to using the risk analysis model.

A small sample of the results is shown in Figures 8 and 9. In Figure 8 we see a probability distribution for the 4 largest risks given a ±1 in all pair-wise comparisons. Clearly, there is very little overlay between the two largest risks indicating that the largest risk is a clear number 1. The more overlay, the higher the probability that the results in Table 4 are inconclusively ranked. Individual probability charts that are much more accurate are also available after a Monte Carlo simulation.



for tracing



for uncertainty

Fig. 6. Modelling input in two different ways to support analysis of risk and uncertainty

In Figure 9 we see the sensitivity chart for the overall risk profile, or the sum of all risks, and this provides us with an accurate ranking of all key risk factors. Similar sensitivity charts are available for all individual risks, as well. Note, however, that since Monte Carlo simulations are statistical methods there are random effects. This means that the inputs in Figure 9 that have very small contribution to variance may be random. In plain words; when the contribution of variance is less than an absolute value of roughly 3% - 5% we have to be careful. The more trials we run, the more reliable the sensitivity charts become.
Similar results to Figures 8 and 9 can also be produced for the uncertainty analysis of the risk analysis. Such analysis can answer questions such as what information should be improved to improve the quality of the risk analysis, and what effects can we expect from improving the information (this can be simulated). Due to space limitations this will be omitted here. Interested readers are referred to (Emblemsvåg 2010) for an introduction. For thorough discussions on Monte Carlo simulations, see (Emblemsvåg 2003).

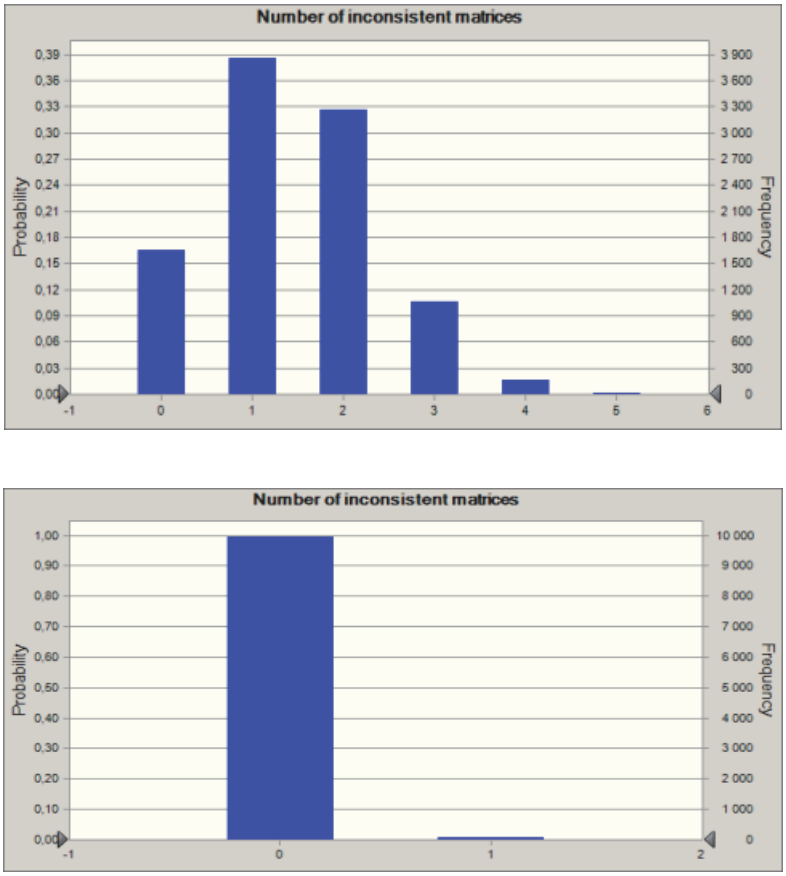


Fig. 7. Improving the quality of the pair-wise comparison matrices

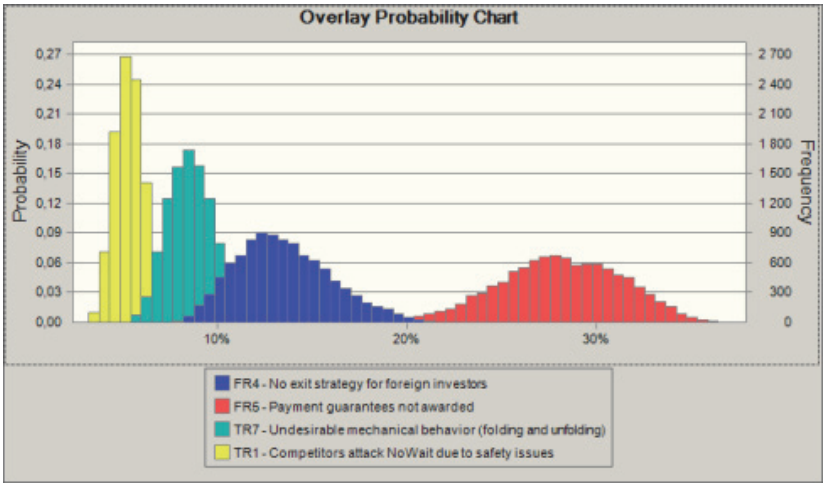


Fig. 8. The 4 largest risks in a subjective probability overlay chart given ±1 variation

The final part of this step is to critically evaluate these analyses. Due to the enormous output of analytical information in this step, the analyses lend themselves to also critically evaluate the results. It should be noted that the AHP structure makes logic errors in the analysis very improbable. Hence, what we are looking for is illogic results, and the most important tool in this context is the sensitivity analyses.

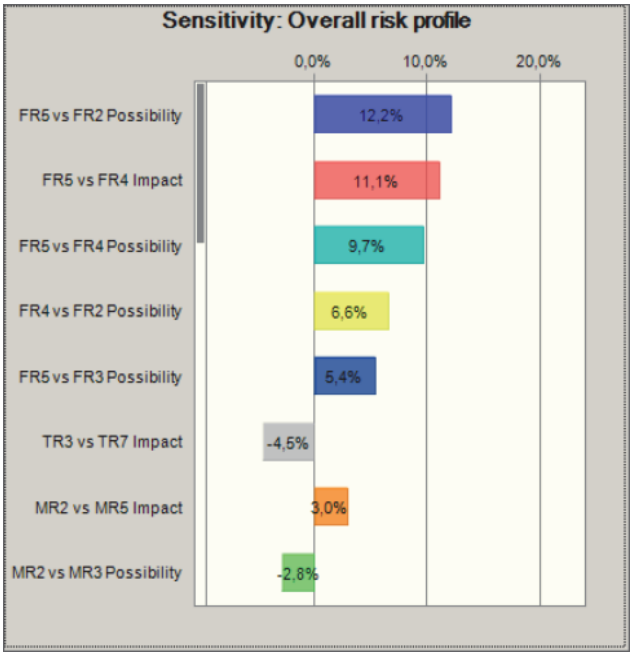


Fig. 9. Sensitivity chart for the overall risk profile given ±1 variation

The next step in the augmented risk management process is Step 4. Running case is omitted from here and onwards since this is not significantly different from traditional approaches except the KM part, which at the time was not conducted in this case due to the fact that this was a start-up company with no KM system or prior experience.

4.4 Step 4 – Develop and implement strategies

After step 3 we have abundant decision-support concerning developing risk management strategies and information management strategies. For example, from Figure 10 we can immediately identify what risk management strategies are most suited (i.e., risk prevention and impact mitigation). Since most inputs are possibilities, risk prevention is the most effective approach. When the issues are impacts, impact mitigation is the most effective approach, and as usual it is a mix between the two that works the best.

These strategies are subsequently translated into both action plans (what to do) and contingency plans (what to do if a certain condition occurs). According to different surveys, less than 25% of projects are completed on time, on budget, and on the satisfaction of the customer, see (Management Center Europe 2002), emphasizing the focus on contingency planning as a vital part of risk management. "Chance favours the prepared mind", in the words of Louis Pasteur. How to make action- and contingency plans is well described by (Government Asset Management Committee 2001) and will not be repeated here.

Information management strategies will concern the cost versus benefit of obtaining better information. This is case specific, and no general guidelines exist except to consider the benefits and costs before making any information gathering decisions. Before making such decisions, it is also wise to review charts similar to Figures 8. If the uncertainty distributions do not overlap, there is no need to improve the information quality; it is good enough. If they do overlap, sensitivity charts can be used to pinpoint what on inputs we need to improve the information quality.

KM in this step will include reviewing what has been done before, what went wrong, what worked well and why (knowledge and meta-knowledge). To the extent this is relevant for a specific case will greatly vary but the more cases are assembled in the KM system the greater the chances are that something useful can be found and therefore aid the process. In this step, however, it is equally important to use the results of the risk- and capability analyses performed in Step 3. These results can help pose critical questions which in turn can be important for effective learning.

4.5 Step 5 – Measure performance

The final step in the augmented risk management process is to measure the actual performance, that is, to identify the actual outcomes in real life. This may sound obvious, but often programs and initiatives are launched without proper measurement of results and follow-up, as (Jackson 2006) notes "Many Fortune 100 companies can *plan* and *do*, but they never *check* or *act*" [original italics][4]. Checking relies on measurement and acting relies on checking, hence, without measuring performance it is impossible to gauge the effectiveness of the strategies and consequently learn from the process. This is commented on later.

Finally, it should be noted that although it may look like the process ends after Step 5 in Figure 5 – the risk management process is to continuously operate until objectives are met.

[4] The PDCA (Plan-Do-Check-Act) circle is fundamental in systematic improvement work.

5. Critical evaluation and future ideas

The research presented here is by no means finished. It is work in progress although some issues seem to have received a more final form than others. What future work that should be undertaken, are the following:

1. Using the AHP matrices for pair-wise comparison is incredibly effective, but it is not workable for practitioners and personally we also believe that many academics will have a hard time making these matrices manually (as done here). Therefore, if the augmented risk management approach is to become commonly used and accepted it will need software that can create the matrices, help people fill in the reciprocal values and to simplify the Monte Carlo simulations.
2. There is significant work to be done on the KM side. Today, we have quite good grasp of handling risks that occur quite frequently as shown by (Neef 2005). The ultimate test of such systems would be how the system could help people deal with risks that are highly infrequent – so called high impact, low probability (HILP) events. This is a common type of events in the natural world, but here even professional bodies treat risks the wrong way, as shown in (Emblemsvåg 2008). Such events are also far more common than what we believe in the corporate world with enormous consequences as (Taleb 2007) shows. Thus, KM systems that could tap from a large variety of sources, to give people support in dealing with such difficult cases, would be useful.
3. Many risks cross the organizational boundaries between business units (known and unknown when the risk management process is initiated) and this raises the issues of interoperability and managing risk in that context, see (Meyers 2006). He defines interoperability as 'the ability of a set of communicating entities to (1) exchange specified information and (2) operate on that information according to a specified, agreed-upon, operational semantics'. The augmented risk management process has not been tested in such a setting, which should be done to prove that it works across organizational boundaries. In fact, due to the more consistent risk analysis, terminology and decision support of the augmented risk management process, it is expected that many of the problems (Meyers 2006) raises are solved, but it must be proven. KM, however, in an interoperable environment is a difficult case.
4. Whether the augmented risk management process would work for statistical risk management processes is also something for future work. Intuitively, the augmented risk management process should work for statistical risk management processes because statistical risk management also has a human touch, as the discussion in Section 1 shows. .

There are probably other, less pressing issues to solve, but this is the focus forward. Item 1 is mostly a software issue and is not commented further here. If this issue is resolved the risk analysis itself and the information management part would be solved. The issue concerning interoperability is a matter of testing, development of specifications and definitions necessary when crossing organizational boundaries. KM, however, is a difficult case – particularly if we include the issues of interoperability.

Since the origin of risks is multi-layered, it is important with a systemic approach towards KM. Also, some risks materialize quite seldom, from an individual perspective, but quite often on a corporate perspective, such as financial crises. This is another argument for a

systemic approach on a wide scale that assembles knowledge from many arenas. Finally, the very rare risks, those that are high impact and low probability, can only be handled in a systematic way because any one person is not likely to experience more than one such risk in decades and hence memory becomes too inefficient. Another issue is that to evoke the right feeling of risk, people must internalize the risks and this can be really difficult. To borrow from (Nonaka and Takeuchi 1995) - explicit knowledge must be internalized and become tacit. Otherwise, the risk profile will not be understood.

From the literature we learn that the SECI model is an effective approach in handling tacit knowledge. Kusunoki, Nonaka *et al.* (1995), for example, demonstrated that the SECI model is good at explaining the successes where system-based capacities are linked with multi-layered knowledge. This is directly relevant for risk management, just mentioned previously. Thus, the SECI model seems to be a promising avenue for improving or complementing the more information-based KM systems that (Neef 2005) discusses. But, according to (Davenport and Prusak 1997), the philosophical position to Nonaka is in striking contrast to scholars subscribing to the information-based view of knowledge, which leads to IT based KM systems. Therefore, we must bridge the gap between these two main avenues of KM: 1) the information-based KM systems which are good at getting hold of large quantities of explicit knowledge, and 2) the SECI process which is good at converting all knowledge into action and *vice versa*. The SECI process is also good at generating new knowledge and making it explicit. How this bridge will work is still unclear and hence needs future work.

6. Closure

This chapter has presented an augmented risk management process. Compared to the traditional process there are many technical improvements, such as the usage of AHP matrices to ensure much more correct and consistent risk assessments, the usage of Monte Carlo simulations to improve the risk analysis and facilitate information management. However, it is still work in progress and the real issue that needs to be resolved in the future, for risk management to really become as important as it should be, is the establishment of a reliable KM process – particularly for HILP events. It is these events that cause havoc and need increased and systematic attention. How this can be achieved is currently unclear, except that it seems that we must listen to Albert Einstein's famous statement that "Imagination is more important than knowledge".

7. Acknowledgment

I greatly acknowledge the cooperation with Lars Endre Kjølstad concerning some parts of this chapter and the project which led to the case presented here. Also, I greatly appreciate the request to contribute to this book, as well as cooperation, from the publisher.

8. References

Argyris, C. (1977). "Double loop learning in organizations." Harvard Business Review 55(Sept./Oct.).

Argyris, C. (1978). Organizational Learning: A Theory of Action Perspective, Addison Wesley Longman Publishing Company.p. 304.

Arrow, K. J. (1992). I Know a Hawk from a Handsaw. Eminent Economists: Their Life and Philosophies. M. Szenberg. Cambridge, Cambridge University Press:pp. 42-50.

Asllani, A. and F. Luthans (2003). "What knowledge managers really do: an empirical and comparative analysis." Journal of Knowledge Management 7(3):pp. 53-66.

Backlund, F. and J. Hannu (2002). "Can we make maintenance decisions on risk analysis results?" Journal of Quality in Maintenance Engineering 8(1):pp. 77-91.

Bernstein, P. L. (1996). Against the Gods: the Remarkable Story of Risk. New York, John Wiley & Sons.p. 383.

Bhutta, K. S. and F. Huq (2002). "Supplier selection process: a comparison of the Total Cost of Ownership and the Analytic Hierarchy Process approaches." Supply Chain Management: An International Journal 7(3):pp. 126-135.

Cavusgil, S. T., R. J. Calantone and Y. Zhao (2003). "Tacit knowledge transfer and firm innovation capability." Journal of Business & Industrial Marketing 18(1):pp. 6-21.

CCMD Roundtable on Risk Management (2001). A foundation for developing risk Management learning strategies in the Public Service. Ottawa, Strategic Research and Planning Group, Canadian Centre for Management Development (CCMD).p. 49.

Cheng, E. W. L. and H. Li (2001). "Analytic Hierarchy Process: An Approach to Determine Measures for Business Performance." Measuring Business Excellence 5(3):pp. 30-36.

Davenport, T. H. and L. Prusak (1997). Working Knowledge: How Organizations Manage What they Know. Boston, MA, Harvard Business School Press.p. 224.

De Bondt, W. F. M. and R. H. Thaler (1985). "Does the Stock Market Overreact?" Journal of Finance 40(3):pp. 793-805.

Devenow, A. and I. Welch (1996). "Rational herding in financial economics." European Economic Review 40(3):pp. 603-615.

Dey, P. K. (2001). "Decision support system for risk management: a case study." Management Decision 39(8):pp. 634-649.

Drucker, P. F. (1986). Managing for Results: Economic Tasks and Risk-Taking Decisions. New York, HarperInformation.p. 256.

Dubois, D., J. Lang and H. Prade Possibilistic logic. Toulouse, Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier.p. 76.

Earl, M. (2001). "Knowledge management strategies: toward a taxonomy." Journal of Management Information Systems 18(1):pp. 215-233.

Emblemsvåg, J. (2003). Life-Cycle Costing: Using Activity-Based Costing and Monte Carlo Methods to Manage Future Costs and Risks. Hoboken, NJ, John Wiley & Sons.p. 320.

Emblemsvåg, J. (2008). "On probability in risk analysis of natural disasters." Disaster Prevention and Management: An International Journal 17(4):pp. 508-518.

Emblemsvåg, J. (2010). "The augmented subjective risk management process." Management Decision 48(2):pp. 248-259.

Emblemsvåg, J. and B. Bras (2000). "Process Thinking - A New Paradigm for Science and Engineering." Futures 32(7):pp. 635 - 654.

Emblemsvåg, J. and L. E. Kjølstad (2002). "Strategic risk analysis - a field version." Management Decision 40(9):pp. 842-852.

Emblemsvåg, J. and L. Tonning (2003). "Decision Support in Selecting Maintenance Organization." Journal of Quality in Maintenance Engineering 9(1):pp. 11-24.

Friedlob, G. T. and L. L. F. Schleifer (1999). "Fuzzy logic: application for audit risk and uncertainty." Managerial Auditing Journal 14(3):pp. 127-135.

Gilford, W. E., H. R. Bobbitt and J. W. Slocum jr. (1979). "Message Characteristics and Perceptions of Uncertainty by Organizational Decision Makers." Academy of Management Journal 22(3):pp. 458-481.

Government Asset Management Committee (2001). Risk Management Guideline. Sydney, New South Wales Government Asset Management Committee.p. 43.

Hines, W. W. and D. C. Montgomery (1990). Probability and Statistics in Engineering and Management Science. New York, John Wiley & Sons, Inc.p. 732.

Honderich, T., Ed. (1995). The Oxford Companion to Philosophy. New York, Oxford University Press.p. 1009.

Hwang, S. and M. Salmon (2004). "Market stress and herding." Journal of Empirical Finance 11(4):pp. 585-616.

Jackson, T. L. (2006). Hoshin Kanri for the Lean Enterprise: Developing Competitive Capabilities and Managing Profit. New York, Productivity Press.p. 206.

Jones, M. E. and G. Sutherland (1999). Implementing Turnbull: A Boardroom Briefing. City of London, The Center for Business Performance, The Institute of Chartered Accountants in England and Wales (ICAEW).p. 34.

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decisions under Risk." Econometrica 47:pp. 263-291.

Kangari, R. and L. S. Riggs (1989). "Construction risk assessment by linguistics." IEEE Transactions on Engineering Management 36(2):pp. 126 - 131.

Kaufmann, A. (1983). Advances in Fuzzy Sets - An Overview. Advances in Fuzzy Sets, Possibility Theory, and Applications. P. P. Wang. New York, Plenum Press.

Klir, G. J. (1991). "A principal of uncertainty and information invariance." International Journal of General Systems 17:pp. 258.

Klir, G. J. and B. Yuan (1995). Fuzzy Sets and Fuzzy Logic: Theory and Applications. New York, Prentice-Hall.p. 268.

Kunreuther, H., R. Meyer and C. van den Bulte (2004). Risk Analysis for Extreme Events: Economic Incentives for Reducing Future Losses. Philadelphia, The National Institute of Standards and Technology.p. 93.

Kusunoki, T., I. Nonaka and A. Nagata (1995). "Nihon Kigyo no Seihin Kaihatsu ni Okeru Soshiki Noryoku (Organizational capabilities in product development of Japanese firms)." Soshiki Kagaku 29(1):pp. 92-108.

Latzco, W. and D. M. Saunders (1995). Four Days With Dr. Deming: A Strategy for Modern Methods of Management, Prentice-Hall.p. 228.

Li, M. and F. Gao (2003). "Why Nonaka highlights tacit knowledge: a critical review." Journal of Knowledge Management 7(4):pp. 6-14.

MacCrimmon, K. R. and D. A. Wehrung (1986). Taking Risks: The Management of Uncertainty. New York, The Free Press.p. 400.

Management Center Europe (2002). "Risk management: More than ever, a top executive responsibility." Trend tracker: An executive guide to emerging management trends(October):pp. 1-2.

McNeill, D. and P. Freiberger (1993). Fuzzy Logic. New York, Simon & Schuster.p. 320.

Meyers, B. C. (2006). Risk Management Considerations for Interoperable Acquisition. Pittsburgh, PA, Software Engineering Institute, Carnegie Mellon University.p. 28.

Miller, G. A. (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Psychological Review 63:pp. 81 - 97.

Neef, D. (2005). "Managing corporate risk through better knowledge management." The Learning Organization 12(2):pp. 112-124.

Nonaka, I. and H. Takeuchi (1995). The Knowledge-creating Company: How Japanese Companies Creat the Dynamics of Innovation. New York, Oxford University Press.p. 298.

Peniwati, K. (2000). The Analytical Hierarchy Process: Its Basics and Advancements. INSAHP 2000, Jakarta.

Peters, E. E. (1999). Complexity, Risk and Financial Markets. New York, John Wiley & Sons.p. 222.

Pieters, D. A. (2004). The Influence of Framing on Oil and Gas Decision Making: An Overlooked Human Bias in Organizational Decision Making. Marietta, GA, Lionheart Publishing.p. 55.

Polanyi, M. (1966). The Tacit Dimension. New York, Anchor Day Books.

Porter, M. E. (1998). Competitive Strategy: Techniques for Analyzing Industries and Competitors. New York, Free Press.p. 407.

Robbins, M. and D. Smith (2001). BS PD 6668:2000 - Managing Risk for Corporate Governance. London, British Standards Institution.p. 33.

Roos, N. (1998). An objective definition of subjective probability. 13th European Conference on Artificial Intelligence, John Wiley & Sons.

Saaty, T. L. (1990). The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. Pittsburgh, RWS Publications.p. 480.

Saaty, T. L. and E. Forsman (1992). The Hierarchon: A Dictionary of Hierarchies, Expert Choice, Inc.

Sias, R. W. (2004). "Institutional Herding." The Review of Financial Studies 17(1):pp. 165-206.

Standards Australia (1999). AS/NZS 4360:1999 - Risk Management. Sydney, Standards Australia.p. 44.

Stevens, S. S. (1946). "On the Theory of Scales of Measurement." Science 103(2684):pp. 677-680.

Takeuchi,H.(1998).Beyond knowledge management: lessons from Japan. www.sveiby.com.au/.

Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable. London, Allen Lane.p. 366.

The Economist (2002). Barnevik's bounty. The Economist. 362:pp. 62.

The Economist (2004). Signifying nothing? The Economist. 370:pp. 63.

The Economist (2009). Greed - and fear: A special report on the future of finance. London, The Economist.p. 24.

Webster (1989). Webster's Encyclopedic Unabridged Dictionary of the English Language. New York, Gramercy Books.p. 1854.

Wickramasinghe, N. (2003). "Do we practise what we preach? Are knowledge management systems in practice truly reflective of knowledge management systems in theory?" Business Process Management Journal 9(3):pp. 295-316.

Zadeh, L. A. (1965). "Fuzzy Sets." Information Control 8:pp. 338 - 353.

Zimmer, A. C. (1986). What Uncertainty Judgements can tell About the Underlying Subjective Probabilities. Uncertainty in Artificial Intelligence. L. N. Kanal and J. F. Lemmer. New York, North-Holland. 4:pp. 249-258.

2

Soft Computing-Based Risk Management - Fuzzy, Hierarchical Structured Decision-Making System

Márta Takács
Óbuda University Budapest
Hungary

1. Introduction

Since its introduction in the mid-sixties (Zadeh, L. A., (1965)), fuzzy set theory has gained recognition in a number of fields in the cases of uncertain, or qualitative or linguistically described system parameters or processes based on approximate reasoning, and has proven suitable and applicable with system describing rules of similar characteristics. It can be successfully applied with numerous reasoning-based systems while these also apply experiences stemming from the fields of engineering and control theory.
Generally, the basis of the decision making in fuzzy based system models is the approximate reasoning, which is a rule-based system. Knowledge representation in a rule-based system is done by means of IF…THEN rules. Furthermore, approximate reasoning systems allow fuzzy inputs, fuzzy antecedents, fuzzy consequents. "Informally, by approximate or, equivalently, fuzzy reasoning, we mean the process or processes by which a possibly imprecise conclusion is deduced from a collection of imprecise premises. Such reasoning is, for the most part, qualitative rather than quantitative in nature and almost all of it falls outside of the domain of applicability of classical logic", (Zadeh, L. A., (1979)).
Fuzzy computing, as one of the components of soft computing methods differs from conventional (hard) computing in its tolerant approach. The model for soft computing is the human mind, and after the earlier influences of successful fuzzy applications, the inclusion of neural computing and genetic computing in soft computing came at a later point. Soft Computing (SC) methods are Fuzzy Logic (FL), Neural Computing (NC), Evolutionary Computation (EC), Machine Learning (ML) and Probabilistic Reasoning (PR), and are more complementary than competitive (Jin, Y. 2010).
The economic crisis situations and the complex environmental and societal processes over the past years indicate the need for new mathematical model constructions to predict their effects (Bárdossy,Gy., Fodor, J., 2004.). The health diagnostic as a multi-parameter and multi-criteria decision making system is, as well, one of the models where, as in the previous examples, a risk model should be managed.
Haimes in (Hames, Y. Y. 2009.) gives an extensive overview of risk modeling, assessment, and management. The presented quantitative methods for risk analysis in (Vose, D. 2008) are based on well-known mathematical models of expert systems, quantitative optimum calculation models, statistical hypothesis and possibility theory. The case studies present applications in

the fields of economics and environmental protection. It is observable that the statistical-based numerical reasoning methods need long-term experiments and that they are time- and computationally demanding. The complexity of the systems increases the runtime factor, and the system parameter representation is usually not user-friend. The numerical methods and operation research models are ready to give acceptable results for some finite dimensional problems, but without management of the uncertainties. The complexity and uncertainties in those systems raise the necessity of soft computing based models.

Nowadays the expert engineer's experiences are suited for modeling operational risks, not only in the engineering sciences, but also for a broad range of applications (Németh-Erdődi, K., 2008.). Wang introduces the term of risk engineering related to the risk of costs and schedules on a project in which there is the potential for doing better as well as worse than expected. The presented case studies in his book are particularly based on long-term engineering experiences, for example on fuzzy applications, which offer the promised alternative measuring of operational risks and risk management globally (Wang, J. X., Roush, M. L., 2000.).

The use of fuzzy sets to describe the risk factors and fuzzy-based decision techniques to help incorporate inherent imprecision, uncertainties and subjectivity of available data, as well as to propagate these attributes throughout the model, yield more realistic results. Fuzzy logic modeling techniques can also be used in risk management systems to assess risk levels in cases where the experts do not have enough reliable data to apply statistical approaches. There are even more applications to deal with risk management and based on fuzzy environments. Fuzzy-based techniques seem to be particularly suited to modeling data which are scarce and where the cause-effect knowledge is imprecise and observations and criteria can be expressed in linguistic terms (Kleiner, Y., at all 2009.).

The structural modeling of risk and disaster management is case-specific, but the hierarchical model is widely applied (Carr, J.H. , Tah, M. ,2001). The system characteristics are as follows: it is a multi-parametrical, multi-criteria decision process, where the input parameters are the measured risk factors, and the multi-criteria rules of the system behaviors are included in the decision process. In the complex, the multilayer, and multi-criteria systems the question arises how to construct the reasoning system, how to incorporate it into the well structured environment. In terms of architectures next to the hierarchical system the cognitive maps (Kosko, 1986.) or ontology (Neumayr, B, Schre, M. 2008.) are also often used. A further possibility is for the system to incorporate the mutual effects of the system parameters with the help of the AHP (Analytic Hierarchy Process) methods (Mikhailov, L., 2003).

Considering the necessary attributes to build a fuzzy-based representation of the risk management system, the following sections will be included in the chapter:

- Fuzzy set theory (fuzzy sets and fuzzy numbers; operators used in fuzzy approximate reasoning models; approximate reasoning models).
- Fuzzy knowledge-base: rule system construction and the approximate reasoning method (Mamadani-type reasoning method).
- Different system architecture representations (hierarchical and multilevel structure of the rule system; weighted subsystems).

Case studies and examples are represented particularly in the Matlab Fuzzy Toolbox environment, particularly in the self-improved software environment representing risk assessment problems.

2. Fuzzy set theory background of risk management

Let X be a finite, countable or overcountable set, the Universe. For the representation of the properties of the elements of X different ways can be used. For example if the universe is the set of real numbers, and the property is "the element is negative", it can be represented in an analytical form, describing it as a subset of the universe : $A=\{x \mid x<0,\ x\in R\}$. The members x of subset A can be defined in a *crisp form* by using characteristic function, where 1 indicates the membership and 0 the non-membership:

$$\chi_A = \begin{cases} 1 & if \quad x \in A \\ 0 & if \quad x \notin A \end{cases} \tag{2.1}$$

Let we assume, that the characteristic function is a mapping $\chi_A : X \to \{0,1\}$.

Fuzzy sets serve as a means of representing and manipulating data that is not precise, but rather *fuzzy*, vague, ambiguous. A fuzzy subset A of set X can be defined as a set of ordered pairs, each with the first element x from X, and the second element from the *interval*. This defines a mapping $\mu_A : X \to [0,1]$. The degree to which the statement "x is in A" is true is determined by finding the ordered pair $(x, \mu_A(x))$.

Definition 2.1

Let be X an non-empty set. A fuzzy subset A on X is represented by its membership function

$$\mu_A : X \to [0,1] \tag{2.2}$$

where the value $\mu_A(x)$ is interpreted as the degree to which the value $x \in X$ is contained in A. The set of all fuzzy subsets on X is called set of fuzzy sets on X, and denoted by $F(X)$[1].
It is clear, that A as a *fuzzy set* or *fuzzy subset* is completely determined by $A = \{(x, \mu_A(x)) \mid x \in X\}$. The terms *membership function* and *fuzzy subset* (get) are used interchangeably and parallel depending on the situation, and it is convenient (to write) for writing simply $A(x)$ instead of $\mu_A(x)$.

Definition 2.2

Let be $A \in F(X)$. Fuzzy subset A is called *normal*, if $(\exists x \in X)(A(x) = 1)$ Otherwise A is *subnormal.*

Definition 2.3

Let be $A \in F(X)$.
The *height* of the fuzzy set A is $\text{height}(A) = \sup(\mu_A(x))$.
The *support* of the fuzzy set A is $\text{supp}(A) = \{x \in X \mid \mu_A(x) > 0\}$.
The *kernel* of the fuzzy set A is $\ker(A) = \in \{x \in X \mid \mu_A(x) = 1\}$.
The *ceiling* of the fuzzy set A is $\text{ceil}(A) = \{x \in X \mid \mu_A(x) = \text{height}(A)\}$.
The α*-cut* (an α *level*) of fuzzy the set A is

[1] The notions and results from this section are based on the reference (Klement, E.P. at all 2000.)

$$[A]^{\alpha} = \begin{cases} \{x \in X | \mu_A(x) \geq \alpha\} & \text{if} \quad \alpha > 0 \\ \text{cl}(\text{supp}(A)) & \text{if} \quad \alpha = 0 \end{cases}$$

where cl(supp(A)) denotes the closure of the support of A.

Definition 2.4

Let be $A \in F(X)$. A fuzzy set A is *convex*, if $[A]^{\alpha}$ is a convex (in the sense of classical set-theory) subset of X for all $x \in X$.

It should be noted, that supp(A), ker(A), ceil(A) and $[A]^{\alpha}$ are ordinary, *crisp sets* on X.

Definition 2.5

Let be $A,B \in F(X)$. A and B are equal (A=B), if $\mu_A(x) = \mu_B(x), (\forall x \in X)$. A is subset of B, (A<B or $A \subset B$), (i.e. B is superset of A), if $\mu_A(x) < \mu_B(x), (\forall x \in X)$.

Definition 2.6

For fuzzy subsets $A_1(x), A_2(x), ..., A_n(x) \in F(X)$ their *convex hull* is the smallest convex fuzzy set $C(x)$ satisfying $A_i(x) \leq C(x)$ for $\forall i \in \{1,2,...n\}$ and for $\forall x \in X$.

Example 2.1.

The Body Mass Index (BMI) is a useful measure of too much weight and obesity. It is calculated from the patients' height and weight. (NHLB, 2011.) The higher their BMI, the higher their risk for certain diseases such as heart disease, high blood pressure, diabetes and others. The BMI score means are presented in the following Table 1.

	BMI
Underweight	BMI<18.5
Normal	18.5≤BMI<24.9 -
Overweight	24.9≤BMI<30
Obesity	BMI≥30

Table 1. The BMI score means

Representing the classification (BMI property) of the patients on the scale (BMI universe) of [0,40] with fuzzy membership functions Underweight ($U(x)$), Normal ($N(x)$), Overweight ($OW(x)$) and Obesity ($Ob(x)$) more acceptable descriptions are attained, where the crisp bounds between classes are fuzzified. Figure 1. shows the BMI universe covered over with four fuzzy subsets, representing the above-mentioned, linguistically described meanings, and constructed in Matlab Fuzzy Toolbox environment .

2.1 Fuzzy sets operations

It is convenient to introduce operations on set of all fuzzy sets like in other ordinary sets. So union and intersection operations are needed for fuzzy sets, to represent respectively in the fuzzy logic environment *or* and *and* operators. To represent fuzzy *and* and *or* t-norm and conorms are commonly used.

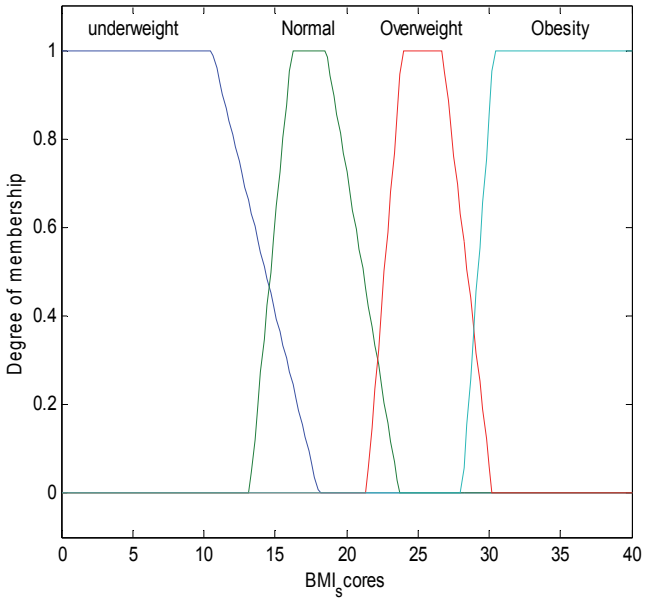


Fig. 1. The BMI universe is covered over with four fuzzy subsets

Definition 2.1.1

A function $T:[0,1]^2 \to [0,1]$ is called *triangular norm (t-norm)* if and only if it fulfils the following properties for all $x,y,z \in [0,1]$

(T1) $T(x,y) = T(y,x)$, i.e., the t-norm is commutative,

(T2) $T(T(x,y),z) = T(x,T(y,z))$, i.e., the t-norm is associative,

(T3) $x \le y \Rightarrow T(x,z) \le T(y,z)$, i.e., the t-norm is monotone,

(T4) $T(x,1) = x$, i.e., a neutral element exists, which is 1.

The basic t-norms are:

$T_M(x,y) = \min(x,y)$, the minimum t-norm,

$T_P(x,y) = x \cdot y$, the product t-norm,

$T_L(x,y) = \max(x+y-1,0)$, the Lukasiewicz t-norm,

$T_D(x,y) = \begin{cases} 0 & if \quad (x,y) \in [0,1[^2 \\ 1 & otherwise \end{cases}$, the drastic product.

Definition 2.1.2

The associativity (T2) allows us to extend each *t*-norm T in a unique way to an *n*-ary operation by induction, defined for each *n*-tuple $(x_1,x_2,...x_n) \in [0,1]^n$, $(n \in N \cup \{0\})$ as

$$\underset{i=1}{\overset{0}{T}} x_i = 1, \quad \underset{i=1}{\overset{n}{T}} x_i = T\left(\underset{i=1}{\overset{n-1}{T}} x_i, x_n\right) = T(x_1,x_2,...x_n) \tag{2.3}$$

Definition 2. 1.3

A function $S:[0,1]^2 \to [0,1]$ is called *triangular conorm (t-conorm)* if and only if it fulfils the following properties for all $x,y,z \in [0,1]$:

(S1) $S(x,y) = S(y,x)$, i.e., the t-conorm is commutative,

(S2) $S(S(x,y),z) = S(x,S(y,z))$, i.e., the t-conorm is associative,

(S3) $x \le y \Rightarrow S(x,z) \le S(y,z)$, i.e., the t-conorm is monotone,

(S4) $S(x,0) = x$, i.e., a neutral element exists, which is 0.

The basic *t*-conorms are:

$S_M(x,y) = \max(x,y)$, the maximum *t*-conorm,

$S_P(x,y) = x + y - x \cdot y$, the probabilistic sum,

$S_L(x,y) = \min(x+y,1)$, the bounded sum,

$S_D(x,y) = \begin{cases} 1 & if \quad (x,y) \in]0,1]^2 \\ \max(x,y) & otherwise \end{cases}$, the drastic sum.

The original definition of *t*-norms and conorms are described in (Schweizer, Sklar (1960)).
At the beginnings of fuzzy theory investigations (and in applications very often today also) *min* and *max* operators are favourites, but new application fields, and mathematical background of them prefers generally t-norms and t-conorms.
Introduce the *fuzzy intersection* $\cap_T$ and *union* $\cup_S$ on $F(X)$, based on t-norm T, t-corm S, and negation N respectively (Klement, Mesiar, Pap (2000a)) in following way

$\mu_{A \cap_T B}(x) = T\left(\mu_{A(x)}, \mu_{B(x)}\right)$ or shortly $\mu_{A \cap_T B}(x) = T\left(A(x), B(x)\right)$,

$\mu_{A \cup_S B}(x) = S\left(\mu_{A(x)}, \mu_{B(x)}\right)$ or shortly $\mu_{A \cup_S B}(x) = S\left(A(x), B(x)\right)$.

The properties of the operations $\cap_T$ and $\cup_S$ on $F(X)$ are directly derived from properties of the t-norm T and t-conorm S. The details about operators you can find in (Klement, E. P. at all, 2000.).

2.2 Fuzzy approximate reasoning

Approximate reasoning introduced by Zadeh (Zadeh, L. A., 1979) plays a very important rule in Fuzzy Logic Control (FLC), and also in other fuzzy decision making applications. The theoretical background of the fuzzy approximate reasoning is the fuzzy logic (Fodor, J., Rubens, M., 1994.),(De Baets, B., Kerre, E.E., 1993.), but the experts try to find simplest user-friend models and applications. One of them is the Mamdani approach (Mamdani , E., H., Assilian, 1975.).
Considering the input parameter x from the universe X, and the output parameter y from the universe Y, the statement of a system can be described with a rule base (RB) system in the following form:

Rule1: IF $x = A_1$ THEN $y = B_1$

Rule2: IF $x = A_2$ THEN $y = B_2$

Rule n: IF $x = A_n$ THEN $y = B_{n1}$

This is denoted as a *single input, single output* (SISO) system.
If there is more than one rule proposition, i.e. the i^{th} rule has the following form
Rule*i*: IF $x_1 = A_{1i}$ AND $x_2 = A_{2i}$ ….. THEN $y = B_i$,
then this is denoted as a *multi input, single output* (MISO) system.
The global structure of an FLC approximate reasoning system is represented in Figure 2.

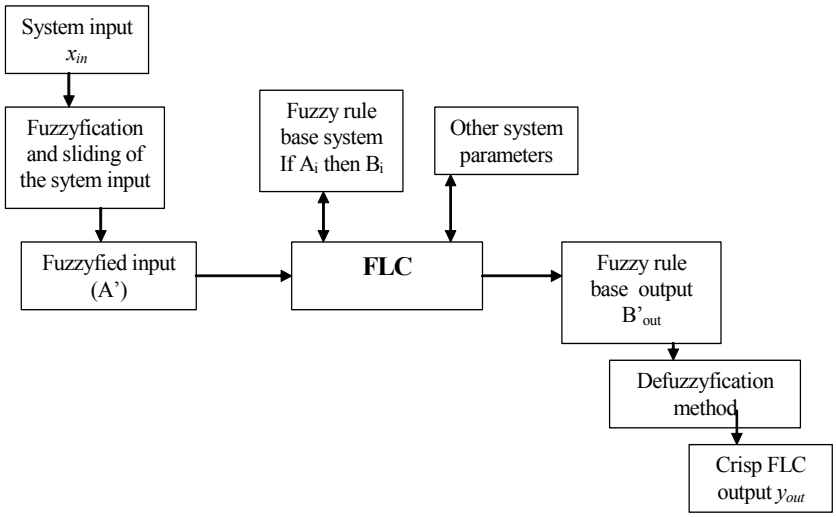


Fig. 2. The global structure of an FLC approximate reasoning system

In the Mamadani-based fuzzy approximate reasoning model (MFAM) the rule output $B_i^{'}(y)$ of the i^{th} rule if x is A_i then y is B_i in the rule system of n rules is represented usually with the expression

$$B_i^{'}(y) = \sup_{x \in X}\left(T\left(A'(x), T\left(A_i(x), B_i(y)\right)\right)\right) \tag{2.4}$$

where $A'(x)$ is the system input, x is from the universe X of the inputs and of the rule premises, and y is from the universe of the output.
For a continuous associative t-norm T, it is possible to represent the rule consequence model by

$$B_i^{'}(y) = T\left(\sup_{x \in X} T\left(A_i(x), A'(x)\right), B_i(y)\right) \tag{2.5}$$

The consequence (rule output) is given with a fuzzy set $B_i'(y)$, which is derived from rule consequence $B_i(y)$, as an upper bounded, cutting membership function. The cut,

$$DOF_i = \sup_{x \in X} T\left(A_i(x), A'(x)\right) \tag{2.6}$$

is the generalized degree of firing level of the rule, considering actual rule base input $A'(x)$, and usually depends on the covering over $A_i(x)$ and $A'(x)$, i.e. on the *sup* of the membership function of $T(A'(x),A_i(x))$.If there is more than one input in a rule, the degree of firing for the i^{th} rule is calculated as the minimum of all firing levels for the mentioned inputs x_i in the i^{th} rule. If the input A' is not fuzzified (i.e. it is a crisp value), the degree of firing is calculated with $DOF_i = \sup_{x \in X} T\left(A_i(x), A'\right)$.

Rule base output, B'_{out} is an aggregation of all rule consequences $B_i'(y)$ from the rule base. As aggregation operator usually *S* conorm fuzzy operator is used.

$$B'_{out}(y) = S(B_n'(y), S(B_{n-1}'(y), S(...., S(B_2'(y), B_1'(y))))). \tag{2.7}$$

If the crisp MFAR output y_{out} is needed, it can be constructed as a value calculated with a defuzzification method., for example with the Central of Gravity (COG) method:

$$y_{out} = \frac{\int_Y B'_{out}(y) \cdot y dy}{\int_Y B'_{out}(y) \cdot dy} \tag{2.8}$$

In FLC applications and other fuzzy approximate reasoning applications based on the experiences from FLC, usually minimum and maximum operators are used as t-.norm and conorm in the reasoning process.

If the basic expectations of this fuzzy decision method are satisfied (Moser, B., Navara., M., 2002.), then the B'_{out} rule subsystem output belongs to the convex hull of disjunction of all rule outputs $B_i(y)$, and can be used as the input to the next decision level in the hierarchical decsison making or reasoning structure without defuzzification. Two important issues arise: the first is, that the B'_{out} is usually not a normalized fuzzy set (should not have a kernel). The solution of the problem can be the use of other operators instead of t-norm or minimum in Mamdani approximate reasoning process to calculate expression(2.6). The second question is, how to manage the weighted output, representing the importance of the handled risk factors group in the observed rule base system. The solution can be the multiplication of the membership values in the expression of B'_{out} with the number from [0,1].

Example 2.2

Continuing the previous example let us consider one more risk factor (risk factor2), and calculate the risk level for the patient taking into account the input risk factors BMI and riskfactor2. Figure 3. shows the membership functions representing the riskfactor2 categories (scaling on the interval [0,1], representing the highest level of risk with 1 and the lower level with 0, i.e. on an unipolar scale). Figure 4. represents the membership functions of the output risk level categories (scaling on the unipolar scale too). Figure 5. shows the system structure, Figure 6. the graphical representation of the Mamdani type reasoning method, and Figure 7. the so called *control surface*, the 3D representation of the risk level calculation, considering both inputs. (Constructions are made in Matlab Fuzzy Toolbox environment).

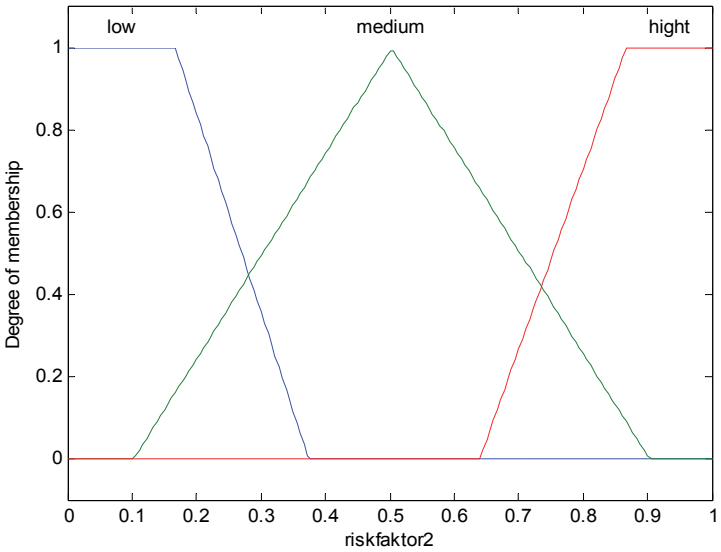


Fig. 3. The membership functions representing the riskfactor2 categories

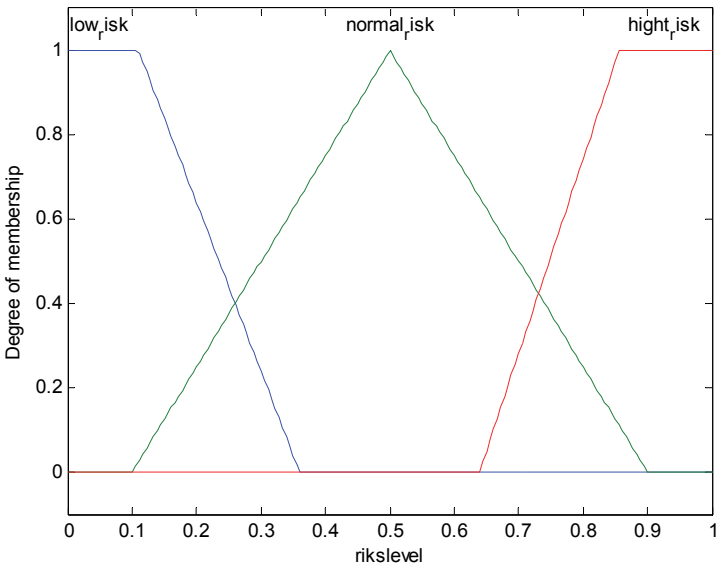


Fig. 4. Represents the membership functions of the output risk level categories

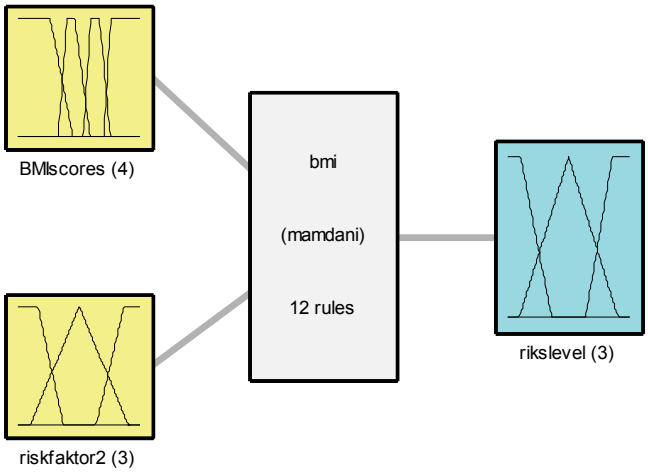


Fig. 5. The system structure

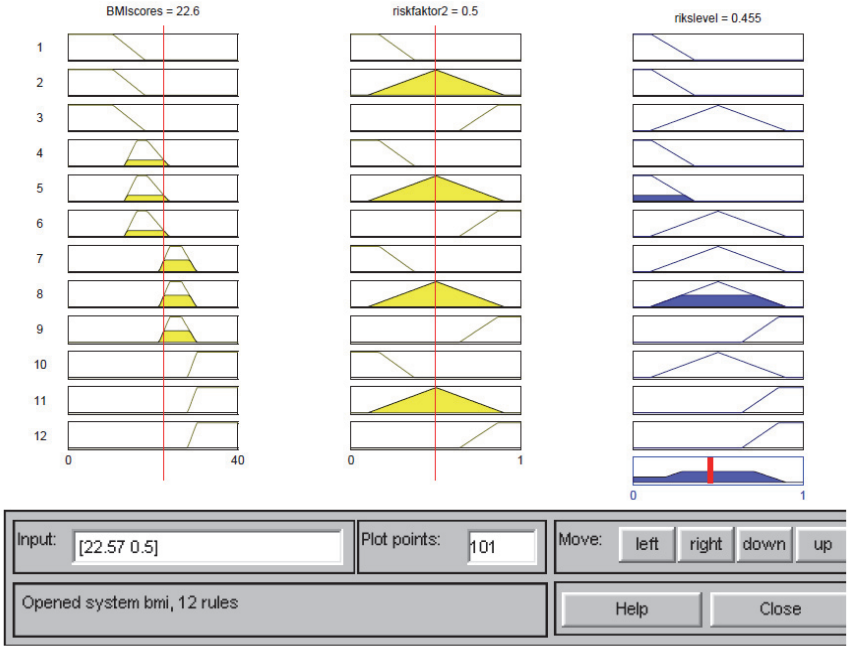


Fig. 6. The graphical representation of the Mamdani type reasoning method

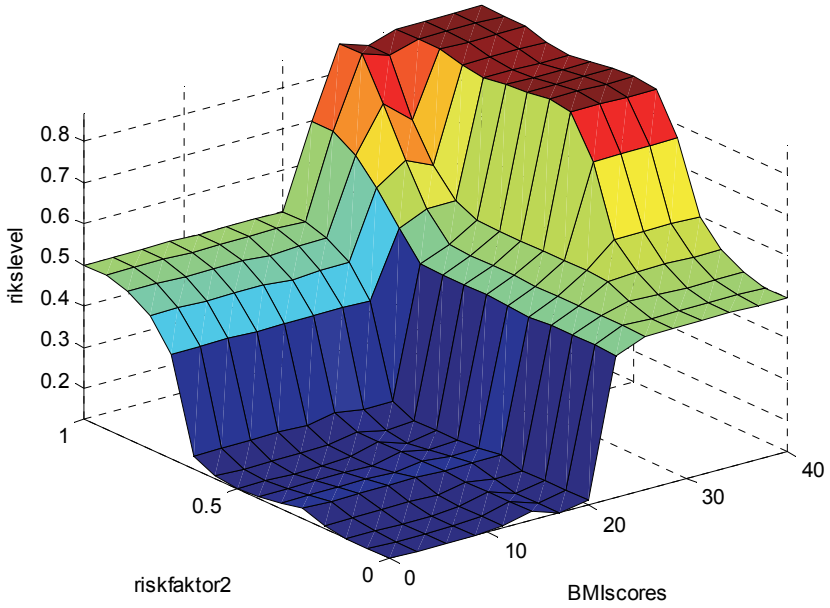


Fig. 7. C*ontrol surface*, the 3D representation of the risk level calculation, considering both inputs

3. Fuzzy logic based risk management

Risk management is the identification, assessment, and prioritization of risks, defined as the effects of uncertainty of objectives, whether positive or negative, followed by the coordinated and economical application of resources to minimize, monitor, and control the probability and/or impact of unfortunate events (Douglas, H. 2009.).

The techniques used in risk management have been taken from other areas of system management. Information technology, the availability of resources, and other facts have helped to develop the new risk management with the methods to identify, measure and manage the risks, or risk levels thereby reducing the potential for unexpected loss or harm (NHSS, 2008.). Generally, a risk management process involves the following main stages.

The first step is the identification of risks and potential risks to the system operation at all levels. Evaluation, the measure and structural systematization of the identified risks, is the next step. Measurement is defined by how serious the risks are in terms of consequences and the likelihood of occurrence. It can be a qualitative or quantitative description of their effects on the environment. Plan and control are the next stages to prepare the risk management system. This can include the development of response actions to these risks, and the applied decision or reasoning method. Monitoring and review, as the next stage, is important if the aim is to have a system with feedback, and the risk management system is open to improvement. This will ensure that the risk management process is dynamic and continuous, with correct verification and validity control. The review process includes the possibility of new additional risks and new forms of risk description. In the future the role of complex risk management will be to try to increase the damaging effects of risk factors.

3.1 Fuzzy risk management

Risk management is a complex, multi-criteria and multi-parametrical system full of uncertainties and vagueness. Generally the risk management system in its preliminary form contains the identification of the risk factors of the investigated process, the representation of the measured risks, and the decision model. The system can be enlarged by monitoring and review in order to improve the risk measure description and decision system. The models for solving are knowledge-based models, where linguistically communicated modelling is needed, and objective and subjective knowledge (definitional, causal, statistical, and heuristic knowledge) is included in the decision process. Considering all these conditions, fuzzy set theory helps manage complexity and uncertainties and gives a user-friendly visualization of the system construction and working model.

Fuzzy-based risk management models assume that the risk factors are fuzzified (because of their uncertainties or linguistic representation); furthermore the risk management and risk level calculation statements are represented in the form of *if premises then conclusion* rule forms, and the risk factor or risk level calculation or output decision (summarized output) is obtained using fuzzy approximate reasoning methods. Considering the fuzzy logic and fuzzy set theory results, there are further possibilities to extend fuzzy-based risk management models modeling risk factors with type-2 fuzzy sets, representing the level of the uncertainties of the membership values, or using special, problem-oriented types of operators in the fuzzy decision making process (Rudas, I., Kaynak, O., 1998.).

The hierarchical or multilevel construction of the decision process, the grouped structural systematization of the factors, with the possibility of gaining some subsystems, depending on their importance or other significant environment characteristics or on laying emphasis on risk management actors, is a possible way to manage the complexity of the system. Carr and Tah describe a common hierarchical-risk breakdown structure for developing knowledge-driven risk management, which is suitable for the fuzzy approach (Carr, J.H. , Tah, M. 2001.).

Starting with a simple definition of the risk as the adverse consequences of an event, such events and consequences are full of uncertainty, and inherent precautionary principles, such as sufficient certainty, prevention, and desired level of protection. All of these can be represented as fuzzy sets. The strategy of the risk management may be viewed as a simplified example of a precautionary decision process based on the principles of fuzzy logic decision making (Cameron, E., Peloso, G. F. 2005.).

Based on the main ideas from (Carr, J.H. , Tah, M. 2001.) a risk management system can be built up as a hierarchical system of risk factors (inputs), risk management actions (decision making system) and direction or directions for the next level of risk situation solving algorithm. Actually, those directions are risk factors for the action on the next level of the risk management process. To sum this up: risk factors in a complex system are grouped to the risk event where they figure. The risk event determinates the necessary actions to calculate and/or increase the negative effects. Actions are described by 'if … then' type rules.

With the output those components frame one unit in the whole risk management system, where the items are attached on the principle of the time-scheduling, significance or other criteria (Fig. 8). Input Risk Factors (RF) grouped and assigned to the current action are described by the Fuzzy Risk Measure Sets (FRMS) such as 'low', 'normal', 'high', and so on. Some of the risk factor groups, risk factors or management actions have a different weighted role in the system operation. The system parameters are represented with fuzzy sets, and the

grouped risk factors values give intermitted results. Considering some system input parameters, which determine the risk factors' role in the decision making system, intermitted results can be weighted and forwarded to the next level of the reasoning process.

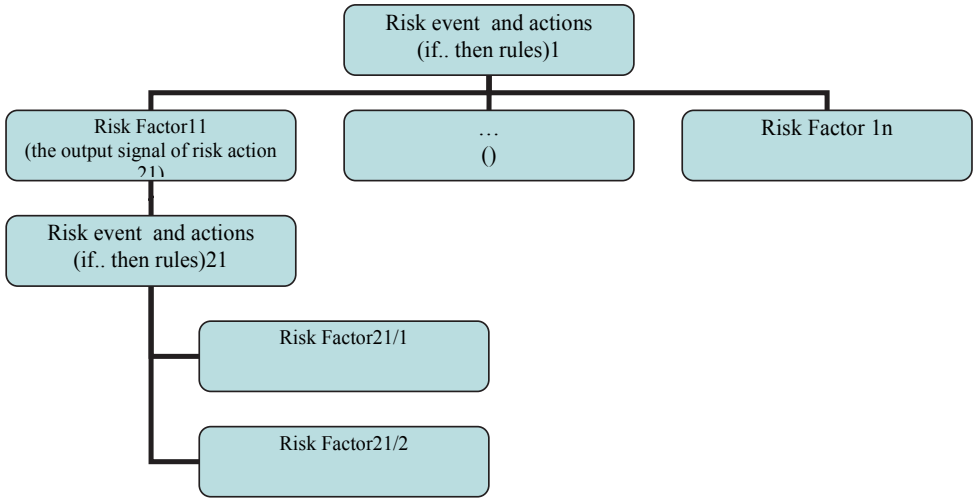


Fig. 8. The hierarchical risk management construction

3.2. Case studies

3.2.1 The brain stroke risk level calculation

Health is commonly recognized as the absence of disease in the body. The fundamental problem with using probability–based statistics for patient diagnosis and treatment is the long time statistical data collection, complex calculation process and the elimination of the real-time human experience (at the actual medical examination) (Helgason, C. M., Jobe, T. H., 2007).The influence of human perception, information collection, experiences involved in diagnosis and therapy realizations support the main fact, namely, that patients are unique. Medical staff has various levels of expertise and the perceptions are often expressed in language. Diagnosis and treatment decisions are determined factors which are either unknown or are not represented within the framework of probability based statistics.

As it is stated in the information brochure published for patients by the University of Pittsburgh Medical Center, the risk factor in health diagnostics is anything that increases chance of illness, accidents, or other negative events. Stroke is one of the most important health issues, because it is not only a frequent cause of death, but also because of the high expenses the treatment of the patients demands. Stroke occurs when the brain's blood flow stops or when blood leaks into brain tissue. The oxygen supply to a part of the brain is interrupted by a stroke, causing brain cells in that area to die. This means that some parts of the body may not be able to function. There are a large number of risk factors that increase the chances of having a stroke. Risk factors may include medical history, genetic make-up, personal habits, life style and aspects of the environment of the patient.

This classification is suitable for grouping the factors, but further different aspects can be applied for grouping.One of them is the classification depending on the possibilities of elimination. Some risk factors cannot be reversed or changed. They are uncontrollable. But some of the risk factors can be eliminated, like smoking, for example. There are other risk factors that the patient cannot get rid of, but can control, like diabetes.

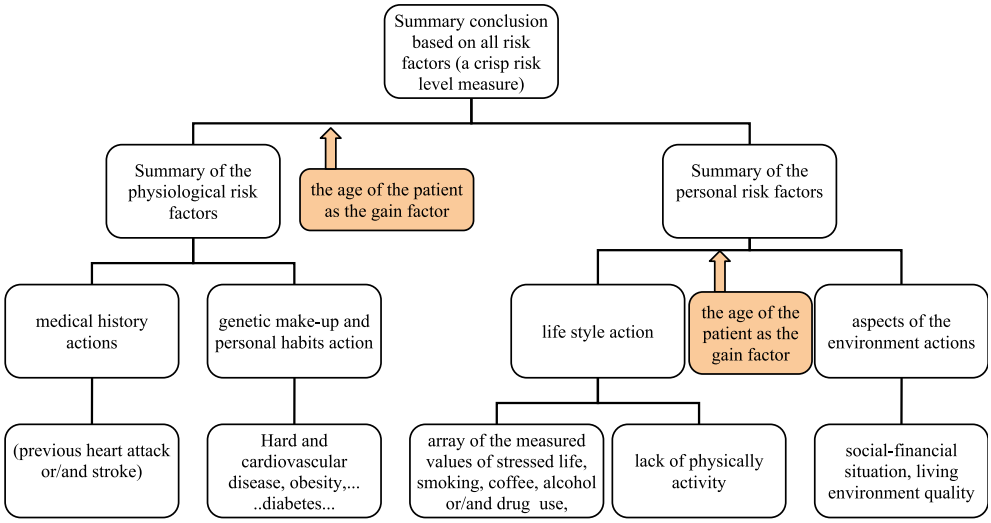


Fig. 9. Brain stroke risk factors - classification, gains and actions

In regard to the theoretical introduction, in the present application a restricted risk factors set is used. The factors are classified in the next *events* – groups (all of risk factors and theirs values are represented with fuzzy membership values)

- medical history (heart attack, previous stroke, …)
- genetic make-up and personal habits (diabetes, obesity, Heart and cardiovascular disease,…)
- life style (stressed life, smoking, coffee, alcohol and drug use, Lack of Physical Activity, …)
- aspects of the environment (social-financial situation, living environment,…).

Grouping physiological events (medical history, genetic make-up and personal habits) and personal controllable events (life style and aspects of the environment) in the separated next level actions, there are two inputs on the final level of actions: summarized physiological factors and summarized personal controllable factors. The final output is the global stroke risk factor based on hierarchically investigated elementary risk factors.

The risk calculation actions are the if then rules regarding to the input variables of the current action level. The outputs at the actions are calculated using the Mamdani type reasoning method, the crisp outputs are achieved with the central of gravity defuzzification. The complex risk calculation system is constructed in a Matlab Fuzzy and Simulink environment.

It ought to be considered, that different events or risk factors have different impact on the stroke occur. Very often the sex or age of the patient will significantly affect the illness. In this experimental system these factors will be the input variables of the system, by which some of the risk factors or events will be gained before the transmission to the next level of action (Figure 9.).
Figure 10. shows the final risk calculation surface.

3.3.2 Disaster management

Disaster event monitoring as one of the steps in risk and crisis management is a very complex system with uncertain input parameters. Fuzzified inputs, the fuzzy rule base, which is constructed using objective and subjective definitional, causal, statistical, and heuristic knowledge, is able to present the problem in a user-friendly form. The complexity of the system can be managed by the hierarchically-structured reasoning model, with a thematically-grouped, and if necessary, gained risk factor structure.
Crisis or disaster event monitoring provides basic information for many decisions in today's social life. The disaster recovery strategies of countries, the financial investments plans of investors, or the level of the tourism activities all depend on different groups of disaster or crisis factors. A disaster can be defined as an unforeseen event that causes great damage, destruction and human suffering, evolved from a natural or man-made event that negatively affects life, property, livelihood or industry. A disaster is the start of a crisis, and often results in permanent changes to human societies, ecosystems and the environment.

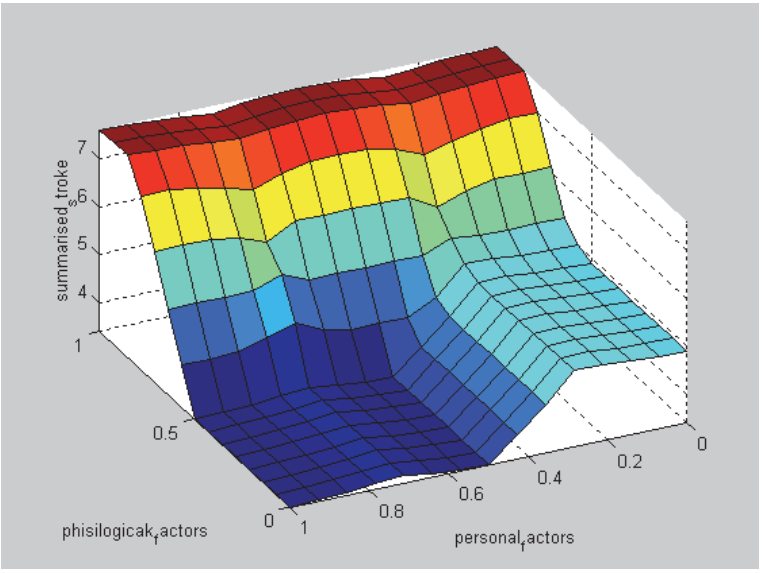


Fig. 10. The summarized risk factor' decision surface

Based on the experts' observations (Yasuyuki S., 2008.), the risk factors which predict a disaster situation can be classified as follows:

- natural disasters;
- man-made disasters (unintended events or wilful events).

Natural disasters arise without direct human involvement, but may often occur, because of human actions prior, during or after the disaster itself (for example, a hurricane may cause flooding by rain or by a storm surge).

The natural disasters can also be grouped primarily based on the root cause:

- hydro-meteorological disasters: floods, storms, and droughts;
- geophysical disasters: earthquakes, tsunamis and volcanic eruptions;
- biological disasters: epidemics and insect infestations;

or they can be structured hierarchically, based on sequential supervention.

The example, presented in this paper, is constructed based on the first principle, with fuzzified inputs and a hierarchically-constructed rule base system (Figure 11.). The risk or disaster factors, as the inputs of one subsystem of the global fuzzy decision making system, give outputs for the next level of decision, where the main natural disaster classes result is the total impact of this risk category.

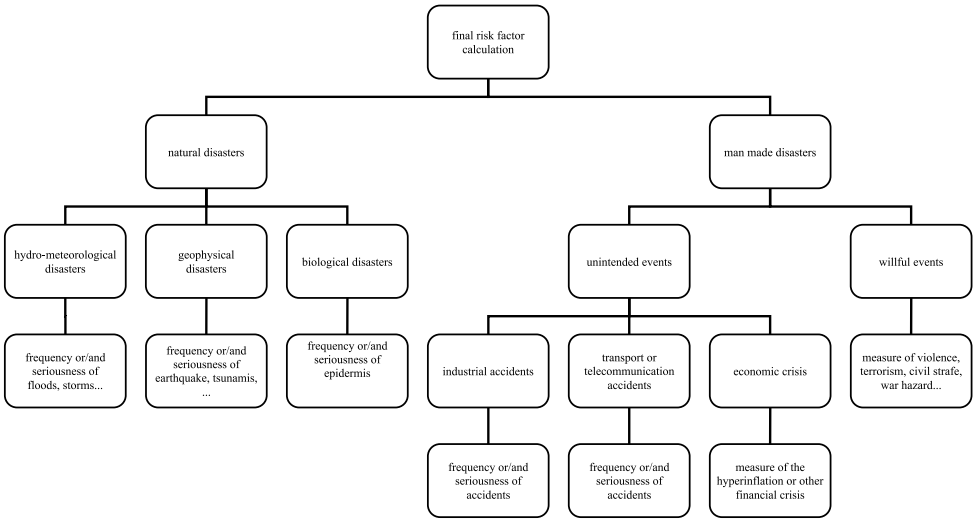


Fig. 11. Hierarchically constructed rule base system representing disaster management

This approach allows additional possibilities to handle the set of risk factors. It is easy to add one factor to a factors-subset; the complexity of the rule base system is changed only in the affected subsystem.

In different seasons, environmental situations etc., some of the risk groups are more important for the global conclusion than others, and this can be achieved with an importance factor (number from the [0,1]). Man-made disasters have an element of human intent or negligence. However, some of those events can also occur as the result of a natural disaster. Man-made factors and disasters can be structured in a manner similar to the natural risks and events.

One of the possible classifications of the basic man-made risk factors or disaster events (applied in our example) is as follows:

- Industrial accidents (chemical spills, collapses of industrial infrastructures);
- Transport or telecommunication accidents (by air, rail, road or water means of transport);
- Economic crises (growth collapse, hyperinflation, and financial crisis);
- wilful events (violence, terrorism, civil strife, riots, and war).

In the investigated example, the effects of man-made disasters as inputs in the decision making process are represented with their relative frequency, and the premises of the related fuzzy rules are very often represented with the membership functions: never, rarely, frequently, etc.[2]

The input parameters are represented on the unit universe [0,1] with triangular or trapezoidal membership functions describing the linguistic variables such as the frequency of the floods, for example: "low", "medium" or "high" (Fig. 12). The system was built in the Matlab Fuzzy Toolbox and Simulink environment.

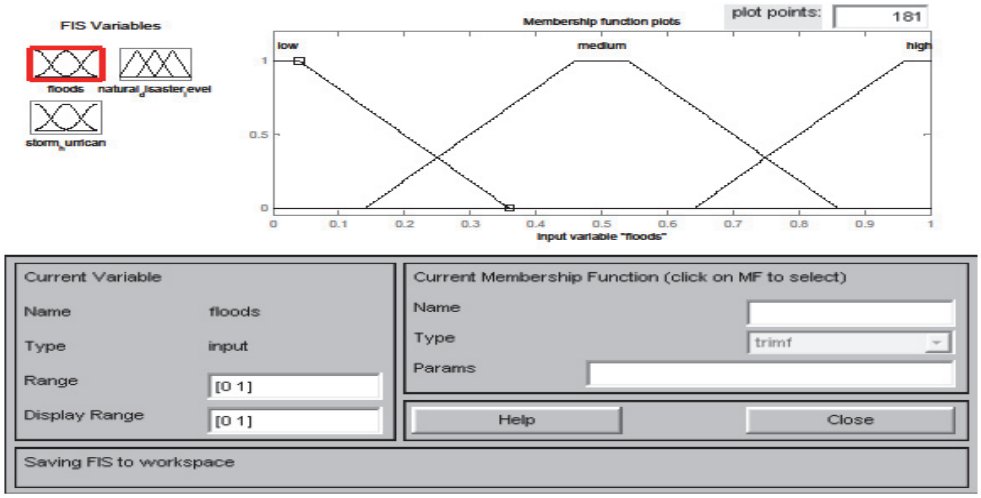


Fig. 12. Membership functions of the flood frequencies

The risk and disaster factors are grouped in two main groups: human- and nature-based group. The inputs are crisp, but the rule base system is hierarchically constructed (Fig. 13), and the decision making is Mamdani type approximate reasoning with basic *min* and *max* operators.

The final conclusion based on both disasters' as risk factors' groups is shown in Figure 14.

[2] The Matlab Fuzzy Toolbox and Simulink elements were in the preliminary, partial form constructed by Attila Karnis, student of the Óbuda University as part of the project in the course "Fuzzy systems for engineers".

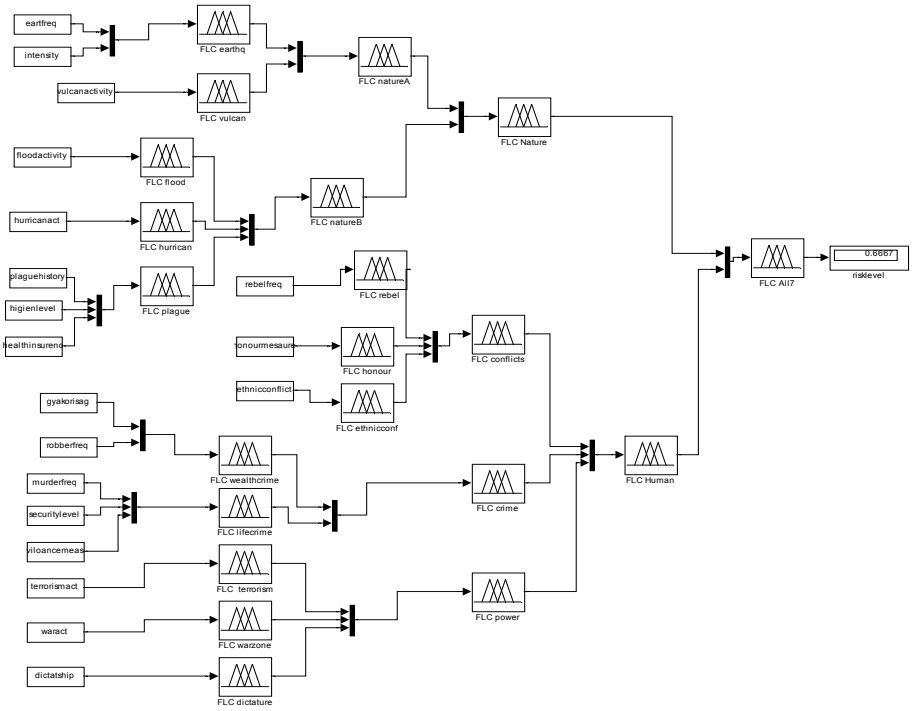


Fig. 13. The Simulink model construction calculating the travel risk level in a country

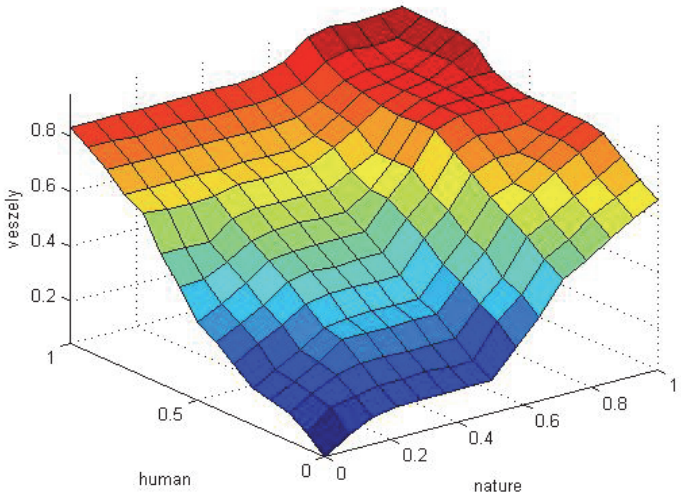


Fig. 14. The final conclusion based on both disasters' as risk factors' groups

4. Conclusion

In this chapter a preliminary system construction of the risk management principle is given based on the structured risk factors' classification and further, based on the fact that some risk factor groups, risk factors or management actions have a weighted role in the system operation. The system parameters are represented with fuzzy sets, and the grouped risk factors' values give intermitted result. Considering some system input parameters, which determine risk factors role in the decision making system, intermitted results can be weighted and forwarded to the next level of the reasoning process. The experimental applications are related to the disaster risk level and stroke risk level calculation.
Considering the fuzzy logic and fuzzy set theory results, there are further possibilities to extend the fuzzy-based risk management models:

- modeling of the risk factors with type 2 fuzzy sets, representing the level of the uncertainties of the membership values;
- use of special, problem oriented types of operators in the fuzzy decision making process;
- the hierarchical or multilevel construction of the decision process, with the possibility of gaining some subsystems, depending on their importance or other significant environment characteristics or on laying emphasis on risk management actors'.

5. Acknowledgment

The research was partially supported by the Research Foundation of Óbuda University, by the project "MI Almanach", (MIEA-TAMOP-412-08_2_A_KMR) and by the project "Model of Intelligent Systems and their Applications" of Vojvodina Provincial Secretariat for Science and Technological Development.

6. References

Bárdossy,Gy., Fodor, J., (2004), Evalution of Uncertainties and Risks in Geology, *Springer*, ISBN 3-540-20622-1.

De Baets, B., Kerre, E.E., (1993), The generalized modus ponens and the triangular fuzzy data model, *Fuzzy Sets and Systems* 59., pp. 305-317.

Cameron, E., Peloso, G. F. (2005). Risk Management and the Precautionary Principle: A Fuzzy Logic Model, *Risk Analysis*, Vol. 25, No. 4, pp. 901-911, August (2005)

Carr, J.H. , Tah, M. (2001). A fuzzy approach to construction project risk assessment and analysis: construction project risk management system, in *Advances in Engineering Software*, Vol. 32, No. 10, 2001. pp. 847-857.

Fodor, J., Rubens, M., (1994), Fuzzy Preference Modeling and Multi-criteria Decision Support. *Kluwer Academic Pub.*,1994

Haimes, Y. Y., (2009). *Risk Modeling, Assessment, and Management*. 3rd Edition. John Wiley & Sons, ISBN: 978-0-470-28237-3, Hoboken, New jersey

Helgason, C. M., Jobe, T. H. (2007). Stroke is a Dynamic Process Best Captured Using a Fuzzy Logic Based Scientific Approach to Information and Causation, In *IC-MED,* Vol. 1, No. 1, Issue 1, Pp.:5-9

Jin, Y., (2010). A Definition of Soft Computing - adapted from L.A. Zadeh, In *Soft Computing Home Page,* 09.04.2011. Available from http://www.soft-computing.de/def.html

Kleiner, Y., Rajani, B., Sadiq, R., (2009). Failure Risk Management of Buried Infrastructure Using Fuzzy-based Techniques, *In Journal of Water Supply Research and Technology: Aqua*, Vol. 55, No. 2, pp. 81-94, March 2006.

Klement, E. P., Mesiar, R., Pap, E.,(2000a), *Triangular Norms*, Kluwer Academic Publishers, 2000a, ISBN 0-7923-6416-3

Kosko, B., (1986). Fuzzy cognitive maps, In *Int. J. of Man-Machine Studies*, 24, 1986.,pp 65-75.

Mamdani , E., H., Assilian ,S. (1975), *An experiment in linguistic syntesis with a fuzzy logic controller,* Intern.,. J. Man-Machine Stud. 7. 1-13

Mikhailov, L. (2003). Deriving priorities from fuzzy pairwise comparison judgements, In *Fuzzy Sets and Systems* Vol. 134, 2003., pp. 365-385.

Moser, B., Navara., M., (2002), *Fuzzy Controllers with Conditionally Firing Rules*, IEEE Transactions on Fuzzy Systems, 10. 340-348

Németh-Erdődi, K., (2008). Risk Management and Loss Optimiyation at Design Process of products*, In* Acta Polytechnica Hungarica*,* Vol. 5, No. 3.

Neumayr, B, Schre, M. (2008). Comparison criteria for ontological multi-level modeling, presented at *Dagstuhl-Seminar on The Evolution of Conceptual Modeling*, Institute für Wirtschaftsinformatik Nr. 08.03, Johannes Kepler Universität Linz.

NHLB, (2011). Assessing Your Weight and Health Risk, *National Heart, Lung, and Blood Institute,* 09.04.2011. Available from http://www.nhlbi.nih.gov/health/public/heart/obesity/lose_wt/risk.htm

NHSScotland model for organisational risk management,(2008) *The Scottish Government,* 09.04.2011. Available from http://www.scotland.gov.uk/Publications/2008/11/24160623/3

Rudas, I., Kaynak, O., (1998), *New Types of Generalized Operations,* Computational Intelligence, Soft Computing and Fuzzy-Neuro Integration with Applications, Springer NATO ASI Series. Series F, Computer and Systems Sciences, Vol. 192. 1998. (O. Kaynak, L. A. Zadeh, B. Turksen,I. .J. Rudas editors), pp. 128-156

Schweizer, B., Sklar, A. (1960), *Statistical metric spaces.* Pacific J. Math. 10. 313-334

Yasuyuki S., (2008). The Imapct of Natural and Manmade Disasters on Household Welfare, 09.04.2011. Available from http://www.fasid.or.jp/kaisai/080515/sawada.pdf

Vose, D., (2008). *Risk Analysis: a Quantitative Guide*, 3rd Edition. John Wiley & Sons, ISBN 978-0-470-51284-5, West Sussex, England

Wang, J. X., Roush, M. L., (2000) What Every Engineer Should Know About Risk Engineering and Management, *Marcel Dekker Inc,* 09.04.2011. Available from http://www.amazon.com/gp

Zadeh, L. A., (1965). Fuzzy sets, *Inform. Control* 8., pp. 338-353

Zadeh, L. A., (1979). A Theory of approximate reasoning*, In Machine Intelligence*, Hayes, J., (Ed.),Vol. 9, Halstead Press, New York, 1979., pp. 149-194.

3

Selection of the Desirable Project Roadmap Scheme, Using the Overall Project Risk (OPR) Concept

Hatefi Mohammad Ali, Vahabi Mohammad Mehdi
and Sobhi Ghorban Ali
Project management department,
Research Institute of Petroleum Industry (RIPI)
Iran

1. Introduction

The findings of researches in the state-of-the art suggest that most errors are due to poor planning of project particularly early in the life of a project. Indeed, project success is positively correlated with the investment in requirements' definition and development of technical specifications (Dvir et. al., 2003). On the other hand, regarding the current business environment of rapid change, one of the main advantages of applying a proactive strategy in planning of projects is the greater flexibility in the competition conditions.
In the strategic planning phase of a project, the below question is outlined to one of the most significant issues in project management:

Which project roadmap scheme (PRS) is the desirable option to execute the project?

The PRSs will be formed by alternative responses to the questions such as:
Which contractor is the desirable option to engineer a given discipline?
Which machinery is the desirable option to produce a given part?
Which technology is the desirable option to montage a given product?
Which supplier is the desirable option to supply a given material?
Evaluating the feasible PRSs is recognized to be a considerable component of a sound project management. An important approach to evaluate the PRSs is the risk efficiency concept, which was originally developed by Markowitz (2002) for managing portfolios of investment opportunities. According to Chapman & Ward (2003), the PRSs can be viewed in a portfolio analysis framework. In fact, each PRS can be considered as an individual project.
The approaches to the solution of the above question "which PRS is the desirable option to execute the project?" can be classified in six groups:

1. Profile and checklist methods,
2. Project scoring methods,
3. Financial measures,
4. Mathematical programming models,
5. Multi Criteria Decision Making (MCDM) models, and
6. Fuzzy approaches.

Project scoring methods do not necessarily ensure the quality of PRS selection, because they do not explicitly take into account PRS level considerations, such as multiple resource constraints and other project interactions. Too often, financial measures are made based solely on criteria such as Net present Value (NPV) and Internal Rate of Return (IRR). Mathematical programming models often solve an integer linear programming to determine the optimal composition of the options subject to resource and other constraints. MCDM models (Keeney & Raiffa, 1999), on the other hand, consider the multi-criteria project values. For data which cannot be precisely assessed, fuzzy sets (Zadeh, 1965) can be used to denote them. The use of fuzzy set theory allows us to incorporate unquantifiable information, incomplete information, non-obtainable information, and partially ignorant facts into the decision model. The first four approaches offer the ability to rate PRSs with a quantitative monetarily unit. Henriksen & Traynor (1999) found that decisions made by managers and those made by a multi-criteria decision making model differ. These differences reflect that such techniques typically do week in simulation of the reality about the projects. It seems the risky world about the projects is usually neglected during the evaluation. In most of the real-world problems, projects are multidimensional in nature and have risky outcomes and decisions and must consider strategy and multidimensional measures (Meade & Presley, 2002).

It is stressed that most significant risks will be subjected to quantitative risk analysis of their impact on project (Project Management Institute [PMI], 2008; United State Department of Energy [US DOE], 2005). Several quantitative models have been introduced to provide valuable predictions for decision-makers. The most common risk valuation technique is expert elicitation. Using this method, the magnitude of consequences may be determined, through the use of expert's opinions. This could be applied using techniques such as interviewing (PMI, 2008). Risks can be represented by probability distribution functions. According to Kahkonen (1999), probability distributions are not widely used, because they are perceived to unlink the assessment from every-day work of project managers. To avoid direct application of probability distributions, the point-estimates (Kahkonen, 1999) are developed such as the Program Evaluation and Review Technique (PERT). Also, Critical Chain Project Management (CCPM) uses the same statistical basis as PERT, but only uses two estimates for the task duration, which are the most likely and the low risk estimates. Many assessment approaches deal with cost and schedule separately in order to simplify the process. Despite this, approaches such as the proposed method by Molenaar (2005) consider both cost and schedule, although schedule modeling tends to be at the aggregate level. Another method to deal with uncertainty is contingency allowance that is an amount of money used to provide for uncertainties associated with a project. The most common method of allowing for uncertainty is to add a percentage figure to the most likely estimate of the final cost of the known works. The amount added is usually called a contingency (Thompson & Perry, 1994).

The present paper introduces a technique to identify the PRS efficient frontier and choose the desirable scheme. According to the introduced model, in responding the question of "which PRS is the desirable option to execute the project?" the decision maker wishes to simultaneously satisfy two objectives, time and cost, with considering positive and negative risks. Most often, these multi-objectives will be in conflict, resulting in a more complicated decision making task. For this purpose, a new modeling approach is proposed to estimate the expected impacts of project risks quantitatively in terms of the project cost and the project time. This framework incorporates Directed A-cyclic Graph (DAG) into the Overall Project Risk (OPR) concept.

2. The proposed modelling approach

Fig. 1 presents process of the proposed model including six phases. The proposed model is structured based on a screening mechanism including three filters as presented in Fig. 2.

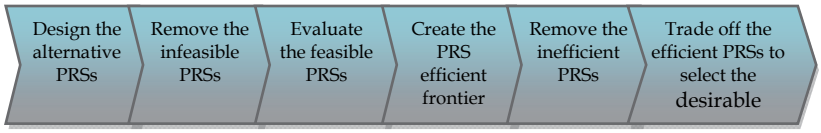


Fig. 1. Process of the proposed technique including six phases

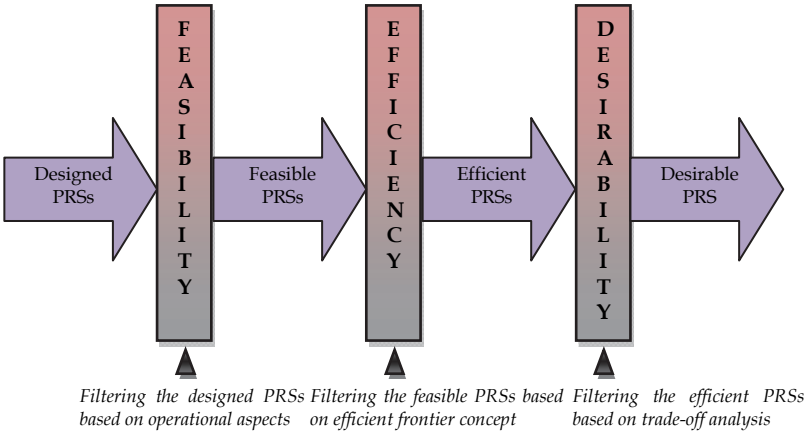


Fig. 2. Screening mechanism of the proposed model

2.1 Designing the alternative PRSs

In the first phase of the process, the project analysts consider different core managerial functions of project and, design the alternative PRSs. Core managerial functions in the field of business and strategic management are (Jaafari, 2007):

- Customers and markets,
- Stakeholders,
- Technology,
- Facility design and operational requirements,
- Supply chain system,
- Learning and innovation,
- Finance,
- Project delivery strategy,
- Risks and due diligence.

Besides, core managerial functions in the field of implementation management are:

- Governance and leadership,
- Engineering, detail design and specifications,
- Procurement, transportation and warehousing,

- Planning and control,
- Team performance,
- Information and communication management,
- Quality management,
- Offsite management,
- Risk management.

2.2 Removing the infeasible PRSs

Some of the designed PRSs may be operationally (technically, conceptually, socially, politically, etc.) inconsistent to implement, so should be removed from the candidate list. The following instances are some inconsistent cases which are experienced in real-world projects:

- An assumed material and a given processing technology may be technically inconsistent.
- Due to some political circumstances, two contractors may keep away to incorporate in a common partnership contract.
- An assumed agent who has not enough experiences should not be assigned for managing a discipline.
- A special mechanical tool may be infeasible to operate in a moist climate.

2.3 Evaluating the Feasible PRSs

In the third phase of the process, including computational core of the model, all of the feasible PRSs are separately evaluated. For a given PRS, it carried out the following stages as shown in Fig. 3:

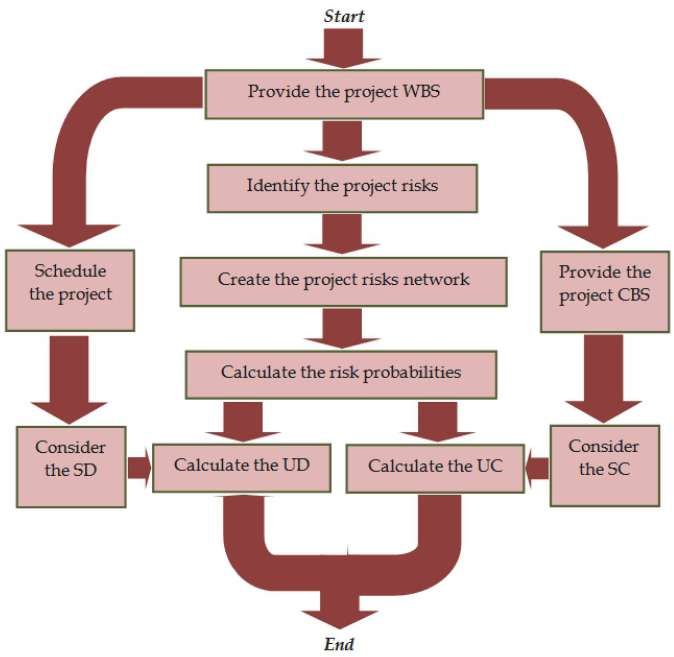


Fig. 3. The evaluation process of an individual PRS

Stage 1: Create the project Work Breakdown Structure (WBS): Complex projects can be overwhelming to the project managers. Instinctively, many project analysts break project down into smaller, more manageable parts. These decompositions are called breakdown structures (US DoE, 2005). WBS is a top-down hierarchical chart of tasks and subtasks required to complete project. WBS can focus on a product, a function, or anything describing what needs to be accomplished (PMI, 2008).

Stage 2: Schedule the project and, calculate Scope Duration (SD): A scheduling methodology defines the rules and approaches for project scheduling. Scheduling is carried out in advance of the project commencing and involves:

- Identifying the activities that need to be carried out;
- Defining activities dependencies which its result is the so called preceding or succeeding activity list.
- Drawing activities network which its result is a graphical portrayed set of activity relationships.
- Estimating how long the activities will take which its result is the so called activity duration.
- Allocating resources to the activities;
- Applying a technique to calculate the earliest/latest start and finish dates of each activity. The present model recommends the better known techniques include Critical Path Method (CPM) or Critical Chain (PMI, 2008).

After scheduling, the project aim on time (SD) will be obtained.

Stage 3: Create the project Cost Breakdown Structure (CBS) and, calculate Scope Cost (SC): The proposed model uses CBS to measure cost elements. Each item in WBS is generally assigned a unique identifier; these identifiers can provide a structure for a hierarchical summation of costs and resources (PMI, 2008). Therefore, CBS represents the hierarchical breakdown of the project costs, so CBS is derived from WBS. After establishing CBS, the target cost of project (SC) will be obtained.

Stage 4: Identify the project risk events: Risk event is an uncertain event or condition that, if it occurs, has a positive or negative effect on at least one project objective: scope, schedule, cost, and quality (PMI, 2008). In the proposed model, risks that have direct or indirect effects on the time and cost of project will be considered. For identifying the risks, the analyzer may benefit from typology of risks mapped in Risk Breakdown Schedule (RBS). For instance, the list below presents a useful typology of common project risks (Mc-Connel, 1996):

- Schedule creation risks such as "excessive schedule pressure reduces productivity".
- Organization and management risks such as "project lacks an effective management sponsor".
- Development environment risks such as "facilities are not available on time".
- End user risks such as "end user ultimately finds product to be unsatisfactory, requiring redesign and rework";
- Customer risks such as "customer has expectations for development speed that developers cannot meet ";
- Contractor risks such as "contractor does not buy into the project and consequently does not provide the level of performance needed";
- Requirement risks such as "vaguely specified areas of the product are more time-consuming than expected";
- Product risks such as "operation in an unfamiliar or unproved software environment causes unforeseen problems";

- External environment risks such as "product depends on government regulations, which change unexpectedly";
- Personnel risks such as "problem team members are not removed from the team, damaging overall team motivation";
- Design and implementation risks such as "necessary functionality cannot be implemented using the selected code or class libraries; developers must switch to new libraries or custom-build the necessary functionality";
- Process risks such as "management-level progress reporting takes more developer time than expected";

Stage 5: Create project risks network and, calculate risks probabilities: Two following criteria are used to characterize risks:

- Risk probability that is the probability of occurring risk event (Kerzner, 2009).
- Risk impact that is the impact of occurring risk event (Kerzner, 2009).

In the proposed model, risk impact reflects the magnitude of effects, either negative or positive, on SC and SD if a risk event occurs. For calculating the risk probability and the risk impacts, the model uses risks network that is a DAG with the following considerations:

i. DAG is a graph $G(N,A)$, where $N=\{E_1,E_2,\ldots,E_m\}$ is a finite set of nodes and $A \subseteq N \times N$ a set of arcs. Each node E_i $(i=1,2,3,\ldots,m)$ refers to a risk event and each arc $(E_i,E_j) \in A$ indicates direct conditional dependencies between two risk events E_i and E_j. If two nodes E_i and E_j within arc (E_i,E_j) are ordered, then the arcs have a direction assigned to them. This is called a directed graph. For a given arc $(E_i,E_j) \in A$, the node E_i is called parent node and the node E_j is called child node.

ii. A conditional probability of P_{ij} which equals $P(E_j \mid E_i)$ is placed for each arc (E_i,E_j). Also, for each node E_i a free probability P_i $(i=1,2,3,\ldots,m)$ is dedicated that is the probability of its occurrence due to risk sources outside risks network. We assume that both P_i and P_{ij} are point estimates or the mean value of a Probability Density Function (PDF) provided by simulation techniques such as the Monte Carlo analysis (PMI, 2008).

iii. Risks network accepts only the acyclic relationships among the risk events. A cycle within a graph is a path that starts and ends at the same node.

Path is a sub-graph of risks network including series of nodes where each node is connected to another node by an arc and all connecting arcs are unidirectional. Each node can occur in the path once only. Each path starts with a source event and ends with a sink event. A path could be depicted as continuum $E_{i_1} \to E_{i_2} \to E_{i_3} \to \ldots \to E_{i_K}$. To simplify this continuum, it could be presented as $\overrightarrow{i_1 i_2 i_3 \ldots i_K}$. We also, denote a specific path as $Path_t$ $(t=1,2,3,\cdots,T)$, which T is the number of the paths within risks network. All paths are placed in the set of R as (1).

$$R=\{Path_t \mid t=1,2,3,\ldots,T\} \tag{1}$$

In a path, the first node is called source and the last node is called sink. As Eq. (2) and Eq. (3), the functions *Source*() and *Sink*() respectively indicates the source event and the sink event of a path.

$$Source(\overrightarrow{i_1 i_2 i_3 \cdots i_K}) = E_{i_1} \tag{2}$$

$$Sink(\overrightarrow{i_1 i_2 i_3 \cdots i_K}) = E_{i_K} \tag{3}$$

As Eq. (4) and Eq. (5) set S_i includes all the paths starting with risk event E_i and set F_i includes all the paths finishing with risk event E_i .

$$S_i = \{Path_t \mid Source(Path_t) = E_i, t = 1,2,3,\ldots,T_i\} \tag{4}$$

$$F_i = \{Path_t \mid Sink(Path_t) = E_i, t = 1,2,3,\ldots,T_i\} \tag{5}$$

As Eq. (6), the plus function $\oplus$ can be used to add a part to the end of a path.

$$(\overrightarrow{i_1 i_2 i_3 \ldots i_k} \oplus \vec{i}_{k+1}) = \overrightarrow{i_1 i_2 i_3 \ldots i_k i_{k+1}} \tag{6}$$

As in term (7) $Path_1$ is subset of $Path_2$, if $Source(Path_1)$ is equal to $Source(Path_2)$, and $Path_1$ contains the complete structure of $Path_2$.

$$\overrightarrow{i_1 i_2 i_3 \ldots i_{v-1} i_v i_{v+1} \ldots i_{K-1} i_K} \subseteq \overrightarrow{i_1 i_2 i_3 \ldots i_{v-1} i_v} \tag{7}$$

According to Eq. (8), each path has a probability, which is defined as the product of free probability of its source event and the conditional probabilities related to its arcs.

$$P(\overrightarrow{i_1 i_2 i_3 \ldots i_k}) = P_{i_1} \times P_{i_1 i_2} \times P_{i_2 i_3} \times \ldots \times P_{i_{K-1} i_K} \tag{8}$$

Probability of the intersection of some paths equals the product of the probabilities of these paths divided by probabilities of common source event or common arcs. Besides, probability of the union of the paths, simply, could be calculated using conventional set union function. As Eq. (9), the occurrence probability of an individual risk event E_i equals the probability of union of all the paths ending with this event. Also, as Eq. (10), the occurrence probability of at least one of the events equals union probability of all paths ending with these events. In addition, as Eq. (11), the occurrence probability of all of events equals intersection probability of all paths ending with these events. It should be noted that for the purpose of identifying the paths within risks network, a labeling algorithm is considered.

$$P(E_i) = P(\bigcup_{\forall Path_t \in F_i} Path_t) \tag{9}$$

$$P(\bigcup_{k=1}^{K} E_{i_k}) = P(\bigcup_{k=1}^{K} \bigcup_{\forall Path_t \in F_{i_k}} Path_t) \tag{10}$$

$$P(\bigcap_{k=1}^{K} E_{i_k}) = P(\bigcap_{k=1}^{K} \bigcup_{\forall Path_t \in F_{i_k}} Path_t) \tag{11}$$

For the purpose of identifying the paths within risks network, a labeling algorithm is considered as Fig. 4, in which F_i is the set of labels for E_i (see Eq. (5)); B_i is a binary index that equals zero until the algorithm completes labeling of risk event E_i. To create the label of a given risk event E_i, if $(E_j, E_i) \in A$, as term (12), the part "i" is added to the end of the labels for risk event E_j. The algorithm does create any labels for a risk event that its free probability is equal to zero.

$$F_i = F_i + \{Path_t \oplus \vec{i} \mid , Path_t \in F_j , \{j \mid (E_j, E_i) \in A\}\} \tag{12}$$

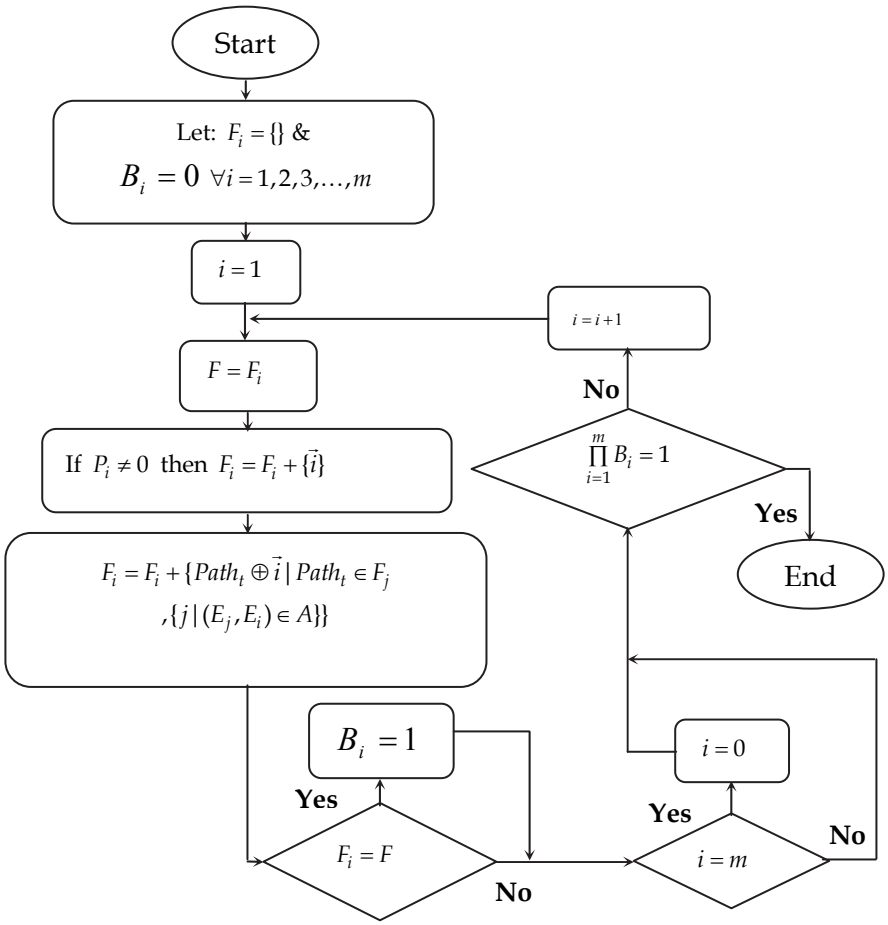
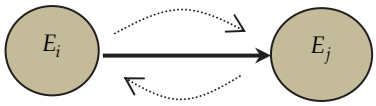


Fig. 4. The labeling algorithm to identify the paths within risks network

Stage 6: Calculate Ultimate Schedule (UD) & Ultimate Cost (UC): UC is the ultimate state of the project cost with considering risk events. UD is the ultimate state of the project duration with considering risk events. The project owners may be interested in knowing the total risk of their project. Indeed, it is often desirable to combine the various risk events into a single quantitative project risk estimate. This estimate is OPR that may be used as input for a decision about whether or not to execute a project, as a rational basis for setting a contingency, and to set priorities for risk response actions (US DOE, 2005). The proposed technique uses the OPR for calculating UC and UD. The main concept here is the relationship between two nodes connected with a direct arc in risks network. According to Fig. 5, the occurrence of a parent node E_i affects the occurrence of a child node E_j (forward circuit), consequently, the impacts of occurrence of the child node E_j, is also transferred to the parent E_i (backward circuit).

Forward circuit: the occurrence of E_i affects the occurrence of E_j



Backward circuit: the impacts of occurrence of E_j, also, are transferred to E_i

Fig. 5. Relationships between child and parent nodes of an arc in risks network

Assume that by use of a suitable level of CBS, the risk impacts on the project cost are as vector (13) that is named as Cost Impact Vector (CIV). It should be noted that each $C_j \in CIV$ is negative value for cost increscent (unwelcome) and is positive value for cost decrement (welcome). The risk analyst can establish the cost matrix (14) in which the rows indicate risk events and the columns stand for the elements of vector (13). The elements of cost matrix (14) are binary parameters c_{ij} as definition (15). Using CIV and cost matrix, UC could be calculated as Eq. (16).

$$CIV^t = \begin{bmatrix} C_1 & C_2 & \cdots & & \cdots & C_c \end{bmatrix} \tag{13}$$

$$C = \left[c_{ij}\right]_{m\times c} = \begin{matrix} E_1 \\ E_2 \\ \vdots \\ E_m \end{matrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1c} \\ c_{21} & & & \\ \vdots & & & \vdots \\ c_{m1} & & \cdots & c_{mc} \end{bmatrix} \tag{14}$$

$$c_{ij} = \begin{cases} 1 & \text{If occurring } E_i \text{ causes cost } C_j \\ 0 & \text{Otherwise} \end{cases} \tag{15}$$

$$UC = SC - \sum_{j=1}^{c} C_j \times P(\bigcup_{\{i|c_{ij}=1\}} E_i) \tag{16}$$

For calculating UD, let $N' \subseteq N$ contain all the risk events that affect the project scheduling. Consider the set β including all non-empty subset of $N' \subseteq N$ as Eq. (17). Now, for all $\beta_w \in \beta$ calculate Eq. (18) in which SD_w is the project duration for subset β_w. For calculating SD_w, we should consider the occurrence of all risk events $E_i \in \beta_w$. In Eq. (18), the second part $\ddot{P}(\bigcap E_i)$ indicates that all risk events $E_i \in \beta_w$ must have occurred. The double-dots sign on the top of this term means that before calculating this probability we are required to apply some conditions related to the third part of Eq. (18). For calculating $\ddot{P}(\bigcap E_i)$, temporarily remove all risk events in which $E_i \in N' \,\&\, E_i \notin \beta_w$. The third part of Eq. (18) indicates that all risk events in which $E_i \in N'$ and $E_i \notin \beta_w$ should not occur. Finally, UD could be calculated as Eq. (19).

$$\beta = \{\beta_w \mid \beta_w \subseteq N', w = 1,2,3,\cdots,W\} \tag{17}$$

$$\lambda_w = (SD - SD_w) \times \ddot{P}(\bigcap_{E_i \in \beta_w} E_i) \times 1 - P(\bigcup_{E_i \in N' - \beta_w} E_i) \quad \forall w = 1,2,3,\ldots,W \tag{18}$$

$$UD = SD - \sum_{w=1}^{W} \lambda_w \tag{19}$$

2.4 Creating the PRS efficient frontier

When evaluating a particular PRS in relation to alternative schemes, we can consider the project cost as the first basic measure of performance and the project time as the second one. The PRS efficient frontier is the set of the feasible PRSs that provides a minimum level of project time for any given project cost, or minimum level of project cost for any given level of project time. This concept is most easily pictured using a graph like Fig. 6. In this figure, B, C, D, E, F and G are the alternative feasible PRSs (schemes A and H are the infeasible PRSs which have been removed in the second phase of the process); the PRS efficient frontier is portrayed by the curve B-C-D-E.

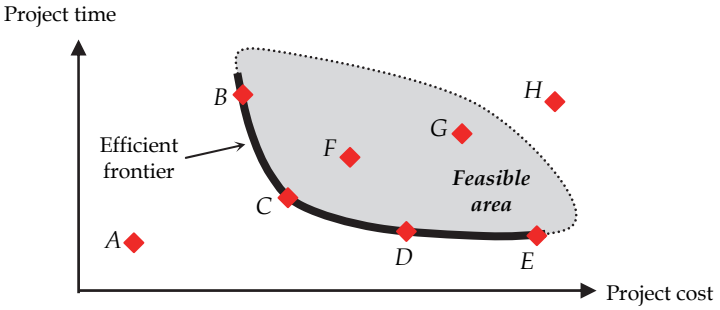


Fig. 6. The PRS efficient frontier concept

2.5 Removing the inefficient PRSs

In the 5th phase of the process, the entire inefficient schemes should be removed from the list of candidate PRSs. Regarding the above discussions in previous section, 2.4, any points inside the frontier, like F and G in Fig. 6, represent the inefficient PRSs. F is more efficient than G, but F can be improved on with respect to both project cost and project time (e.g. moving to C).

2.6 Trading off the efficient prss to select the desirable scheme

In the 6th phase of the process, the efficient PRSs should be pair-wise compared. In each pair-wise comparison, one of the PRSs is removed as Eq. (20). The parameter α is defined as the payment (dollars) that project owners will be admitted for one time-unit (i.e. 1 day) increment in the project duration. More α results in more importance of the project time than the project cost. Regarding Eq. (20), it should be noted that the desirable PRS is the nearest point to the tangent point between the PRS efficient frontier and the line by gradient $-\alpha^{-1}$.

$$\begin{cases} \text{If } \left(\dfrac{UC_i - UC_j}{UD_j - UD_i}\right) \le \alpha \text{ then} & \begin{cases} \text{If } UD_i \ge UD_j \text{ remove PRS \# } i \\ \text{If } UD_i < UD_j \text{ remove PRS \# } j \end{cases} \\ \text{If } \left(\dfrac{UC_i - UC_j}{UD_j - UD_i}\right) > \alpha \text{ then} & \begin{cases} \text{If } UC_i \ge UC_j \text{ remove PRS \# } i \\ \text{If } UC_i < UC_j \text{ remove PRS \# } j \end{cases} \end{cases} \tag{20}$$

3. Analytical results

For analyzing the model, we consider a project includes Engineering, Procurement, and Construction (EPC) of a powerhouse cavern elevator, which has been drawn from a hydro-mechanical power plant. The project includes four sub-products cabin, hoisting machine, suspension guides and control equipments.
The entire outputs of the process phases are at one glance mapped in Table 1 that presents that twelve PRSs were designed.

Phase 1: The project experts considered the following alternatives to design candidate PRSs. They designed twelve PRSs (see Table 1).

- Two alternatives for supplying the elevator cabin:
  - (a1) fabricating the cabin in the firm and then transporting it to the erection site;

  - (a2) fabricating the cabin in the erection site.
- Three alternatives for supplying the elevator hoisting machine:
  - (b1) buying the hoisting machine from the foreign supplier 1;
  - (b2) buying the hoisting machine from the foreign supplier 2;
  - (b3) buying the hoisting machine from the present inside supplier.
- Two alternatives for basic designing the control equipment:
  - (c1) employing a sub-contractor for basic designing the control equipment;
  - (c2) buying a present basic design.

PRS code	PRS contents	Feasibility	UC ($)	UD (days)	Efficiency	Desirability
S1	(a1), (b1), (c1)	Feasible	148,900	540	Inefficient	-
S2	(a1), (b1), (c2)	Infeasible	-	-	-	-
S3	(a1), (b2), (c1)	Feasible	137,000	390	Efficient	Undesirable
S4	(a1), (b2), (c2)	Feasible	165,800	485	Inefficient	-
S5	(a1), (b3), (c1)	Feasible	125,975	525	Efficient	Undesirable
S6	(a1), (b3), (c2)	Feasible	192,900	340	Efficient	Undesirable
S7	(a2), (b1), (c1)	Infeasible	-	-	-	-
S8	(a2), (b1), (c2)	Feasible	158,800	350	Efficient	Desirable
S9	(a2), (b2), (c1)	Infeasible	-	-	-	-
S10	(a2), (b2), (c2)	Feasible	175,698	490	Inefficient	-
S11	(a2), (b3), (c1)	Feasible	138,000	500	Inefficient	-
S12	(a2), (b3), (c2)	Feasible	210,550	335	Efficient	Undesirable

Table 1. The designed PRSs for the typical project

Phase 2: The operational discussions about the feasibility of the schemes resulted in the schemes S2, S7 and S9 are not feasible to execute; consequently, these schemes were removed from the candidate list.

Phase 3: The nine feasible PRSs were evaluated. As a sample, table 2 exhibits the WBS, Fig. 7 shows the CBS and, Fig. 8 shows the risks network for PRS S4. According to Table 2, for PRS S4, SC=137,700 $ & SD=420 days; by considering the occurrence of the risk events, UC=137,700 $ and UD=485 days (see Table 1). Table 1 shows UC and UD for the nine feasible PRSs.

Phase 4: The nine feasible PRSs have been portrayed in Fig. 9.

No.	WBS code	Activity	Duration (days)	Cost ($)
1	**1**	**Powerhouse cavern elevator**	420	137,700
2	**1.1**	**Cabin**	420	56,000
3	1.1.1	Designing	44	9,000
4	1.1.2	Material supply	90	23,000
5	1.1.3	Manufacturing & Assembly	310	4,000
6	1.1.4	Transportation to erection site	10	2,000
7	1.1.5	Erection	40	18,000
8	**1.2**	**Hoisting machine**	**401**	**29,500**
9	1.2.1	Designing	37	6,600
10	1.2.2	Material supply	110	12,800
11	1.2.3	Manufacturing & Assembly	50	2,200
12	1.2.4	Transportation to erection site	17	1,300
13	1.2.5	Erection	20	6,600
14	**1.3**	**Suspension guides**	**381**	**48,000**
15	1.3.1	Designing	60	4,200
16	1.3.2	Material supply	115	2,600
17	1.3.3	Manufacturing & Assembly	155	19,200
18	1.3.4	Transportation to erection site	19	1,400
19	1.3.5	Erection	32	9,600
20	**1.4**	**Control equipment**	**240**	**21,800**
21	1.4.1	Designing	35	3,200
22	1.4.2	Material supply	100	5,100
23	1.4.3	Manufacturing & Assembly	75	4,500
24	1.4.4	Transportation to erection site	15	1,100
25	1.4.5	Erection	15	1,300

Table 2. The project WBS including durations and costs for PRS S4

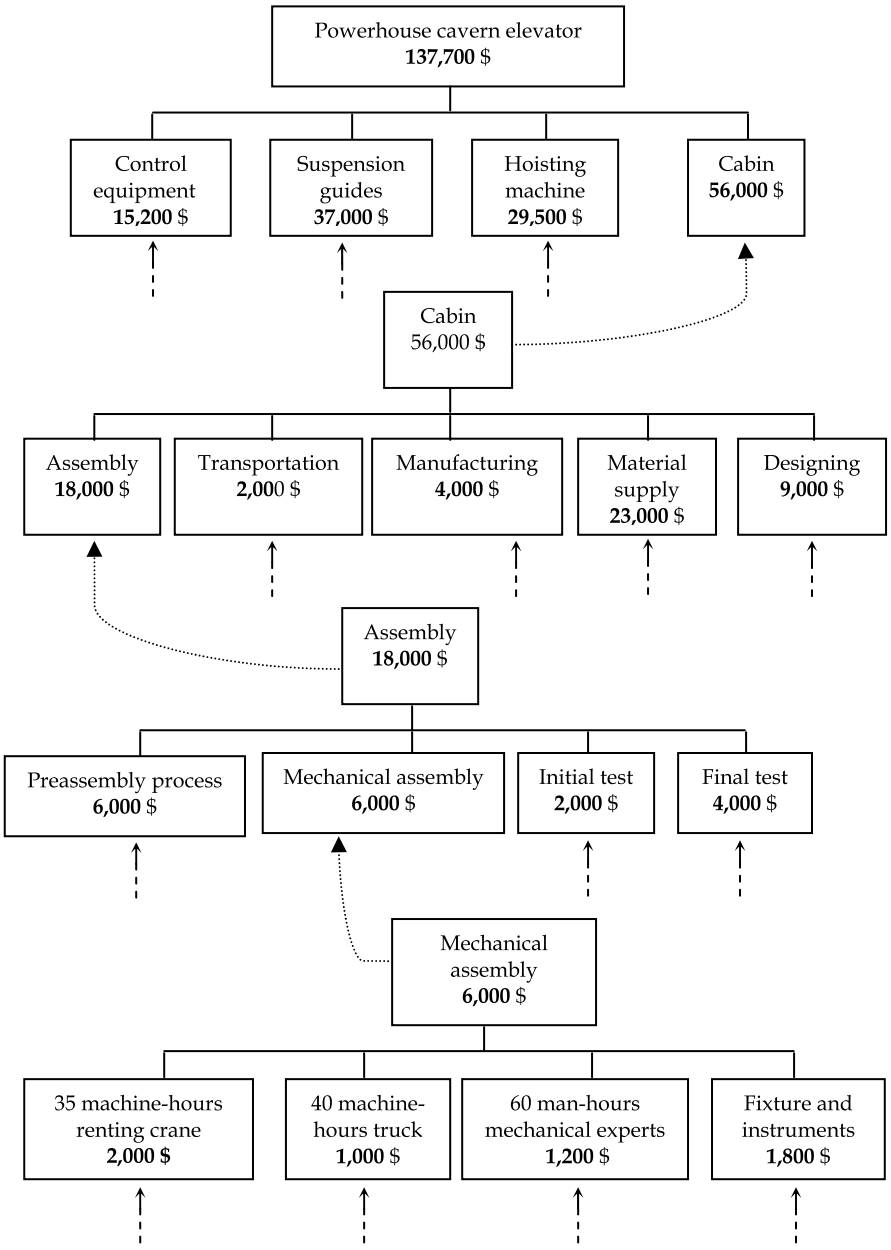


Fig. 7. A part of the CBS for PRS S4

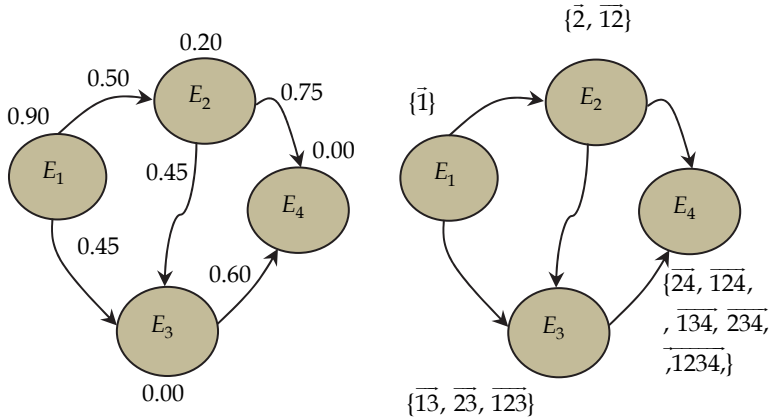


Fig. 8. The risks network for PRS S4 (left) and its labels (right)

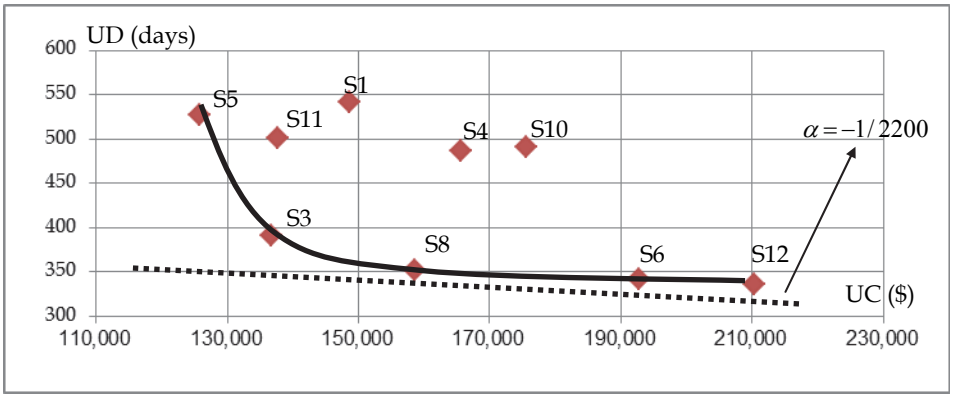


Fig. 9. The PRS efficient frontier for the typical project

Phase 5: The schemes S1, S4, S10 and S11 were considered as the inefficient PRSs and were removed from the candidate list of the PRSs.

Phase 6: For choosing the desirable PRS among the reminded schemes S3, S5, S6, S8 and S12, the experts did the pair-wise comparisons as Eq. (19). By assuming $\alpha = 2200$ \$/day, for instance the term ((158,800-137,000) / (390-350)) = 545 \$/day was calculated for the pair-wise comparison between the schemes S3 & S8. Because $545 < \alpha$ & $350 < 390$ thus PRS S3 was removed; in pair-wise comparison between PRSs S6 & S8, because $3410 > \alpha$ & $158,800 < 192,900$ thus PRS S6 was removed; in comparison between PRSs S5 & S8, because $188 < \alpha$ & $350 < 525$ thus PRS S5 was removed and finally, in comparison between PRSs S8 & S12, because $3450 > \alpha$ & $158,800 < 210,550$ so PRS S12 was removed. Finally PRS S8 was considered as the desirable scheme. The selected scheme, PRS S8, contains fabricating the cabin in the erection site, buying the hoisting machine from the foreign supplier 1, and

buying a present basic design for control equipment. As it has been shown in Fig. 9, the PRS S8 is the nearest point to the tangent point between the efficient frontier and the line with gradient-1/2200.

4. Discussions

Several characters of the proposed model are worthwhile emphasizing:

- The risk researchers believe that project risk analysis should be strongly integrated to the project elements (Chapman & Ward, 2003; Kerzner, 2009; Seyedhoseini et. al., 2008a, 2008b; Ward & Chapman, 2003). In our approach, WBS plays a central role in the quantification of risks. So, the main contribution of the proposed technique is in demonstrating how overall project plan and project risk analysis could be integrated through a united framework. It should be explained that a common technique in estimating the risk probability and risk impact is the use of scales that are usually quantified directly through the expert elicitation. We believe that there is a gap between the scale tablets and the expert's opinion. The proposed model acts a means for bridging the mentioned gap.
- Another key feature of the model is explicitly allowing for dependency relationships among risk events. This is made possible by using DAG.
- The model considers both upside and downside risks within a united perspective. Therefore one can observe that this perspective is a step toward the uncertainty management (Ward & Chapman, 2003).
- Regarding the project environment, since no data record was available about project risk analysis in previous similar projects, probability distribution elicitation for task duration or cost may be difficult for projects, which in turn could limit the applicability of techniques. According to Chapman & Ward (2003), too often this precision is false, because the initial data may be too vague to be fitted into a probability function or the assumptions behind the distributions do not hold true. So, in the proposed technique, all of input data to the model is considered to be one-point estimates. These estimates are easy to understand (Kahkonen, 1999), and do not include a range of values, standard deviation and variance, or confidence intervals, so they do not include the effects of uncertainty and are simply based on the summation of a number of point estimates for items of work.

The technique presented here can be expanded to allow for additional features of the problem.

- Based on the two-polar concept of project risk management (Seyedhoseini et al., 2008a), one such extension is considering the implementation of risk response actions to calculate UC and UD that results in more effective the technique.
- Another extension of the model aims to address the cyclic dependencies among the risk events. Naturally when cyclical feedbacks are considered, it is more difficult but more useful.
- Finally we recall that the proposed model does not guarantee the inclusion of the quality aspects of project. It could be worthwhile to investigate the risk impacts on the project quality. Regarding this area, the reader is encouraged to study work of Seyedhoseini et al. (2008b).

5. Conclusion

Most of the real-world projects are multidimensional in nature and include many risky phenomena, while in the state-of-the art of Project Roadmap Scheme (PRS) selection, risks are usually neglected. In this chapter, we proposed a risk-based modeling approach to support evaluating the alternative PRS and choosing the desirable scheme. In the proposed model, the PRSs are designed then, within a screening mechanism including three criteria of feasibility, efficiency and desirability are filtered. The project cost and the project time play a central role to identify the PRS efficient frontier. The chapter also introduced the development and application of Directed Acyclic Graph (DAG) for estimation of the expected impacts of the project risks. The main contribution of this research was in demonstrating how project plan and Overall Project Risk (OPR) could be integrated through a united framework. We conclude that applying the proposed model helps the project experts to evaluate the feasible PRSs and to choose the desirable scheme in most effective and productive manner dealing with in real world's uncertainties.

6. References

Chapman, C.B., & Ward, S.C. (2003). *Project risk management: processes, techniques and insights (2nd Edition)*, ISBN: 978-0470853559, John Wiley, Chichester (UK).

Dvir, D., Raz, T., & Shenhar, A.J. (2003). An empirical analysis of the relationship between project planning and project success, *International Journal of Project Management*, Vol. 21, No. 2, pp. (89–95).

Henriksen, A.D., & Traynor, A.J. (1999). A practical R&D project selection tool. *IEEE Transactions on Engineering Management*, Vol. 46, No. 2, pp. (158-170).

Hopkinson, M. (2006). Top down techniques for Project Risk management, *PMI Global Congress*, Madrid (Spain), September 1 2006.

Jaafari A. (2007). Project and program diagnostics: a systemic approach, *International Journal of Project Management*, Vol. 25, No. 8, pp. (781-790).

Kahkonen, K. (1999). Integration of qualitative and quantitative risk analysis, *15th Conference of the International Federation of Operational Research Societies*, Beijing (China).

Keeney, R.L., & Raiffa, H. (1999). *Decisions with multiple objectives: preferences and value trade-offs (2nd Edition)*, ISBN: 978-0521438837, Cambridge University Press, UK.

Kerzner, H. (2009). *Project management: a systems approach to planning, scheduling, and controlling (10th Edition)*, ISBN: 978-0470278703, Wiley, TX (USA).

Markowitz, H.M. (2002). *Portfolio selection: efficient diversification of investments (2nd Edition)*, Blackwell Publishers Ltd., ISBN: 978-1557861080, Massachusetts (USA).

Mc-Connel, S. (1996). *Rapid development: taming wild software schedules (1st Edition)*, Microsoft Press, ISBN: 978-1556159008, Washington (USA).

Meade, L.M., & Presley, A. (2002). R&D project selection using the analytic network process. *IEEE Transactions on Engineering Management*, Vol. 49, No. 1, pp. (59-66).

Molenaar, K. (2005). Programmatic cost risk analysis for highway mega-projects, *Construction Engineering and Management*, Vol. 131, No. 3, pp. (343-353).

PMI (Project Management Institute). (2008). *A guide to the project management body of knowledge (4th Edition)* ISBN: 978-1933890517, Newtown Square, PA (USA).

Seyedhoseini, S.M., Noori, S., & Hatefi, M.A. (2008a). Two-polar concept of project risk management, In: *New Frontiers in Enterprise Risk Management*, David L. Olson &

Desheng Wu, pp. (69-92), ISBN: 978-3642097409, Springer Berlin Heidelberg, Berlin (Germany).

Seyedhoseini, S.M., Noori, S., & Hatefi, M.A. (2008). An integrated decision support system for project risk response planning, *Kuwait Journal of Science and Engineering*, Vol. 35, No. 2B, pp. (171-192).

Thompson, A., & Perry, J.G. (1994). *Engineering construction risks: a guide to project risk analysis and risk management (2nd Edition)*, ISBN: 978-0727716651, Thomas Telford, London (UK).

U.S. DoE (Department of Energy). (2005). *The owner's role in project risk management (1st Edition)*, National Academies Press, ISBN: 978-0309095181, NY (USA).

Ward, S.C., & Chapman, C.B. (2003). Transforming project risk management into project uncertainty management, *International Journal of Project Management*, Vol. 21, No. 2, pp. (97-105).

Zadeh L. (1965). Fuzzy sets, *Information Control*, Vol. 8, pp. (338–353).

4

A New Non-Parametric Statistical Approach to Assess Risks Associated with Climate Change in Construction Projects Based on LOOCV Technique

S. Mohammad H. Mojtahedi[1] and S. Meysam Mousavi[2]
[1]*School of Civil Engineering, The University of Sydney, NSW*
[2]*Department of Industrial Engineering, College of Engineering, The University of Tehran, Tehran,*
[1]*Australia,*
[2]*Iran*

1. Introduction

During the last two decades, Iran government has implemented a major program to extend and upgrade construction projects in oil and gas industry. In conjunction with the increasing growth, there are many types of potential risks that affect the construction projects. Risks can be defined as an uncertain event or condition that has a positive or negative effect on project objectives, such as time, cost, scope, and quality (Caltrans, 2007; PMI, 2008). Thus, there is a need for a risk management process to manage all types of risks in projects. Risk management includes the processes of conducting risk management planning, identification, analysis, response planning, monitoring, and control on a construction project. Risk management encourages the project team to take appropriate measures to: (1) minimize adverse impacts to the project scope, cost, and schedule (and quality, as a result); (2) maximize opportunities to improve the project's objectives with lower cost, shorter schedules, enhanced scope and higher quality; and (3) minimize management by crisis (Caltrans, 2007).

In project risk management, one of the major steps is to assess the potential risks (Ebrahimnejad et al., 2009, 2010; Makui et al., 2010; Mojtahedi et al., 2010). The risk assessment process can be complex because of the complexity of the modeling requirement and the often subjective nature of the data available to conduct the analysis in construction projects. However, the complexity of the process is not overwhelming and the benefits of the outcome can be extremely valuable (Mousavi et al., 2011).

Many decisions come with a long-term commitment and can be very climate sensitive. Examples of such decisions include urbanization plans, risk management strategies, infrastructure development for water resource management or transportation, and building design and norms. These decisions have consequences over periods of 50–200 years. Urbanization plans influence city structures over even longer timescales. These kinds of decisions and investments are also vulnerable to changes in climate conditions and sea level

rise. For example, many building are supposed to last up to 100 years and will have to cope in 2100 with climate conditions that, according to most climate models, will be radically different from current ones. So, when designing a building, architects and engineers have to be aware of and account for future changes that can be expected (Hallegatte, 2009)
Not considering of the climate change impacts on projects, especially those are established for a long term use, can cause massive costs for government and public in future. Nicholls et al. (2007) showed that, in 2070, up to 140 million people and more than US$ 35,000 billion of assets could be dependent on flood protection in large port cities around the world because of the combined effect of population growth, urbanization, economic growth, and sea level rise.
Recently, resampling techniques are rapidly entering mainstream data analysis; some statisticians believe that resampling procedures will supplant common nonparametric procedures and may displace most parametric procedures (e.g., Efron and Tibshirani, 1993). These techniques are the use of data or a data gathering mechanism to produce new samples, in which the results can be examined in various fields. In resampling, estimates of probabilities are offered by numerical experiments. Resampling offers the benefits of statistics and probability theory without the shortcomings of common techniques. Because it is free of mathematical formulas and restrictive assumptions. In addition, it is easily understood and computer user friendly (Simon and Bruce, 1995; Tsai and Li, 2008). The purpose of resampling techniques is to find the distribution of a statistic by repeatedly drawing a sample, thus making use of the original sample. The leave-one-out-cross-validation (LOOCV) first originated as generic nonparametric estimators of bias and standard deviation (SD). Moreover, to the best of our knowledge, no LOOCV technique and resampling application was found regarding climate change risk assessment of these projects. On the other hand, a risk data analysis in construction projects often encounters the following situations (Mojtahedi et al., 2009):

- It cannot be answered in a parametric framework.
- It may need to be examined by standard and existing tools.
- It can be assessed only by specially tailored algorithms.

For these reasons, the LOOCV resampling approach is presented to use for assessing risks in construction projects. This approach is flexible, easy to implement, and applicable in non-parametric settings. In this paper, we contribute to this area by providing an effective framework for the application of the LOOCV to climate change risk data obtained from experts' judgments in construction projects.
The chapter is organized as follows: In Section 2, the researchers review related literature and discuss the existing gap in the field. In Section 3, we describe the proposed a new non-parametric LOOCV approach to assess risks associated with climate changes in construction projects. In Section 4, computational results in construction of a gas refinery plant as a case study is presented. The discussion of results is given in Section 5. Finally, conclusion is provided in Section 6.

2. Literature review

Construction projects are subject to many risks due to the unique features of construction tasks, such as long period, complicated processes, undesirable environment, financial intensity and dynamic organization structures (Zou & Zhang, 2009), and such organizational and technological complexity generates enormous risks. The diverse interests

of project stakeholders on a construction project further exacerbate the changeability and complexity of the risks (Zou & Zhang, 2009).

The purpose of project risk management is to identify risky situations and develop strategies to reduce the probability of occurrence and/or the negative impact of risky events on projects. In practice, project risk management includes the process of risk identification, analysis and handling (Gray & Larson, 2005). Risk identification requires recognizing and documenting the associated risk. Risk analysis examines each identified risk issue, refines the description of the risk, and assesses the associated impact. Finally, risk handling/response identifies, evaluates, selects, and implements strategies (e.g., insurance, negotiation, reserve, etc.) in order to reduce the likelihood of occurrence of risk events and/or lower the negative impact of those risks to an acceptable level. The risk-handling process contains the documentation of which actions should be taken, when they should be taken, who is responsible, and the associated handling costs (Fan et al., 2008).

It is widely accepted that construction project' activity is particularly subject to more risks than other business activities because of its complexity, and a wide range of risks associated with construction businesses have been previously identified. A typical classification of risks includes technical risks, management risks, market risks, legal risks, financial risks, and political risks (Shen, 1997).

Identified risks are assessed to determine their likelihood and potential effect on project objectives, allowing risks to be prioritized for further attention. The primary technique for this is the Probability–Impact matrix, where the probability and impacts of each risk are assessed against defined scales, and plotted on a two-dimensional grid. Position on the matrix represents the relative significance of the risk, and high/medium/low zones may be defined, allowing risks to be ranked (Hillson, 2002). While it is not practical to discuss the full implications of all the risks identified in the survey, this section intends to demonstrate the pattern of the risk environment by presenting some practical examples discussed in the five in-depth interviews following the survey. Not all the risks addressed in this section respond to the ''most important risks'' ranked in the risk significance index as interviewees have different experiences, and their perception or judgment may not be fully in harmony with the calculated average index scores (Shen et al., 2001).

Previous studies have been focused on the risk management in mega projects. Grabowski et al. (2000) discussed the challenges of risk modeling in large-scale systems, and suggested a risk modeling approach that was responsive to the requirements of complex, distributed, large-scale systems. Florice & Miller (2001) showed that achieving high project performance requires strategic systems that are both robust with respect to anticipated risks and governable in the face of disruptive events by comparing the features and performance of three common types of project. Miller & Lessard (2001) developed strategies to understand and manage risks in large engineering projects. Wang et al. (2004) tried to identify and evaluate these risks and their effective mitigation measures and to develop a risk management framework which the international investors/ developers/ contractors can adopt when contracting large construction projects' work in developing countries.

Iranmanesh et al. (2007) proposed a new structure called RBM to measure the risks in EPC projects. By combining risk breakdown structure with work breakdown structure (WBS), a new matrix (RBM) is constructed. Hastak & Shaked (2000) presented a risk assessment model for international construction projects. The proposed model (ICRAM-1) assists the user in evaluating the potential risk involved in expanding operations in an international

market by analyzing risk at the macro (or country environment), market, and project levels. Zeng et al. (2007) proposed a risk assessment model based on modified analytical hierarchy process (AHP) and fuzzy reasoning to deal with the uncertainties arising in the construction projects. Mojtahedi et al. (2008) presented a group decision making approach for identifying and analyzing project risks concurrently. They showed that project risk identification and analysis can be evaluated at the same time. Moreover, they applied the proposed approach in one mega project and rewarding results were obtained. Ebrahimnejad et al. (2008) introduced some effective criteria, and attributes was used for risk evaluating in construction projects. They presented a model for risk evaluation in the projects based on fuzzy MADM. Makui et al. (2010) presented a new methodology for identifying and analyzing risks of mega projects (oil and gas industry) concurrently by applying fuzzy multi-attribute group decision making (FMAGDM) approach. Risk identification and classification is the first step of project risk management process, in which potential risks associated with an EPC project are identified. Numerous techniques exist for risk identification, such as brainstorming and workshops, checklists and prompt lists, questionnaires and interviews, Delphi groups or NGT, and various diagramming approaches such as cause-effect diagrams, systems dynamics, influence diagrams (Chapman, 1998; Ebrahimnejad et al., 2008, 2010; Mojtahedi et al., 2009, 2010). There is no a ''best method'' for risk identification, and an appropriate combination of techniques should be used. As a result, it may be helpful to employ additional approaches to risk identification, which were introduced specifically as broader techniques in group decision making field (Hashemi et al., 2011; Makui et al., 2010; Mousavi et al., 2011; Tavakkoli-Moghaddam et al., 2009).

There has been an increasing agreement that many decisions relating to long term investments need to take into account climate change. But doing so is not easy for at least two reasons. First, due to the rate of climate change, new infrastructure will have to be able to cope with a large range of changing climate conditions, which will make design more difficult and construction more expensive. Second, the uncertainty in future climate makes it impossible to directly use the output of a single climate model as an input for infrastructure design, and there are good reasons to think that the required climate information will not be available soon. Therefore, Instead of optimizing based on the climate conditions projected by models, future infrastructure should be made more robust to possible changes in climate conditions. This aim implies that users of climate information must also change their practices and decision making frameworks, for instance by adapting the uncertainty management methods they currently apply to exchange rates or R&D outcomes.

Water resource management is one of the most important fields which has attracted a lot attention. Qin et al. (2008) developed an integrated expert system for assessing climate change impacts on water resources and facilitating adaptation. The presented expert system could be used for both acquiring knowledge of climate change impacts on water resources and supporting formulation of the relevant adaptation policies. It can also be applied to other watersheds to facilitate assessment of climate change impacts on socio-economic and environmental sectors, as well as formulation of relevant adaptation policies. Yin (2001) developed an integrated approach based on the AHP for evaluating adaptation options to reduce climate change effects on water resources facilities.

There are many studies of climate change impacts and the relevant policy responses. For instance, Yin & Cohen (1994) developed a goal programming approach to evaluate climate change impacts and to identify regional policy responses. Huang et al. (1998) proposed a

multi-objective programming method for land-resources adaptation planning under changing climate. Smith (1997) proposed an approach for identifying policy areas where adaptations to climate change should be considered. Lewsey et al. (2004) provided general recommendations and identified challenges for the incorporation of climate change impacts and risk assessment into long-term land-use national development plans and strategies. They addressed trends in land-use planning and, in the context of climate change, their impact on the coastal ecosystems of the Eastern Caribbean small islands. They set out broad policy recommendations that can help minimize the harmful impacts of these trends. Teegavarapu (2010) developed a soft-computing approach and fuzzy set theory for handling the preferences attached by the decision makers to magnitude and direction of climate change in water resources management models. A case study of a multi-purpose reservoir operation is used to address above issues within an optimization framework.

The review of the literature indicates that risk and uncertainty associated with climate changes in construction projects in the developing countries, particularly in Iran, has not been received sufficient attention from the researchers. In addition, climate change risk assessment in construction projects has been focused within a framework of parametric statistics. Among the techniques used in these studies, such as the multi criteria decision making or mathematical modeling, most researchers have assumed that the parameters for assessing risks are known and that sufficient sample data are available. Moreover, parametric statistics, in which the population was assumed to follow a particular and typically normal distribution, was used. However, in risk assessment of construction projects, particularly in developing countries such as Iran, this assumption cannot be made either because of a shortage of professional experts or due to time constraints. Hence, large-sample techniques are not often functional in such projects. Non-parametric cross-validation resampling approach is presented to utilize for assessing risks associated with climate changes in construction projects. This approach is flexible, easy to implement, and applicable in non-parametric settings.

This paper assumes that the risk data distributions in the construction projects are unknown. We cannot find enough professional experts to gather adequate data, and questioning experts about project risk to gather data is a time-consuming and non-economical process. Moreover, few experts are interested in answering or filling out questionnaires. Hence, this paper presents a non-parametric resampling approach based on cross-validation technique to overcome the lack of efficiency of existing techniques and to apply small data sets for risk assessment in the construction projects.

Theoretical studies and discussions about the cross-validation technique under various situations can be found, in (Stone, 1974, 1977; Efron, 1983). The cross-validation predictive density dates at least to (Geisser and Eddy, 1979). Shao (1993) proved with asymptotic results and simulations that the model with the minimum value for the LOOCV estimate of prediction error is often over specified. Sugiyama at al. (2007) proposed a technique called importance weighted cross validation. They proved the almost unbiased even under the covariate shift, which guarantees the quality of the technique as a risk estimator. Hubert & Engelen (2007) constructed fast algorithms to perform cross-validation on high-breakdown estimators for robust covariance estimation and principal components analysis. The basic idea behind the LOOCV estimator lies in systematically recomputing the statistic estimate leaving out one observation at a time from the sample set. From this new set of observations for the statistics, an estimate for the bias and the SD of the statistics can be calculated. A non-

parametric LOOCV technique provides several advantages over the traditional parametric approach as follows: This technique is easy to describe and apply to arbitrarily complicated situations. Furthermore, distribution assumptions, such as normality, are never made (Efron, 1983). The cross-validation has been used to solve many problems that are too complicated for traditional statistical analysis. There are numerous applications of the LOOCV in the various fields (Bjorck et al., 2010; Efron & Tibshirani, 1993).

3. Proposed approach for construction projects

The objectives of this section are as follows: (1) establish a project risk management team, (2) identify and classify potential risks associated with climate changes in construction projects in Iran, (3) present a statistical approach for analyzing the impact of risks using a non-parametric LOOCV technique, and (4) test the validity of the proposed approach.

We implement the proposed approach in the risk assessment of the real-life construction project in Iran. This construction project in oil and gas industry is considered. The project is subject to numerous sources of risks. Designing, constructing, operating, and maintaining of the project is a complex, large-scale activity that both affects and is driven by many elements (e.g., local, regional, political entities, power brokers, and stakeholders). We aim at assessing the climate change risks in order to enable them to be understood clearly and managed effectively. There are many commonly used techniques for the project risk identification and assessment (Chapman & Ward, 2004; Cooper et al., 2005). These techniques generate a list of risks that often do not directly assist top managers in knowing where to focus risk management attention. The analysis can help us to prioritize identified risks by estimating common criteria, exposing the most significant risks. Hence, in this paper a case study which can assess risks of climate changes in a non-parametric statistical environment is introduced.

Data sizes of construction project risks are often small and limited. In addition, there are no parametric distributions on which significance can be estimated for risks data. On the other hand, the LOOCV is the powerful tool for assessing the accuracy of a parameter estimator in situations where traditional techniques are not valid. Moreover, the LOOCV technique is computationally less costly when the sample size is not large (Efron, 1983). A major application of this approach is in the determination of the bias. It answers some questions, such as what is the bias of a mean, a median, or a quantile. This technique requires a minimal set of assumptions.

In the light of the above mentioned issues, in this section one practical approach is proposed to use in assessing risks for construction projects in three phases. Establishing a project risk management team is considered in the first phase which is called phase zero. In this phase, organizational and project environmental in which the risk managing is taking place are investigated. After constructing the project risk management team, we construct the core of the proposed approach in the next two phases. Phase one in turn falls into two steps. In the first step, risk data of construction projects are reviewed in order to identify them. In the second step, the risk breakdown structure (RBS) is developed in order to organize different categories of the project risks. Phase two of the proposed approach falls into four steps. These steps are as follows: (1) determine descriptive scales for transferring linguistic variables of probability and impact criteria to quantitative equivalences, (2) filter the risks at the lowest level of the RBS regarded as initial risks, (3) classify the identified climate change

risks (initial risks) into the significant and insignificant risks, and (4) apply the non-parametric LOOCV technique for final ranking. This phase attempts to understand potential project problems after identifying the mega project risks. Risk assessment is considered in this phase. The proposed mechanism for construction projects is depicted in Fig. 1.

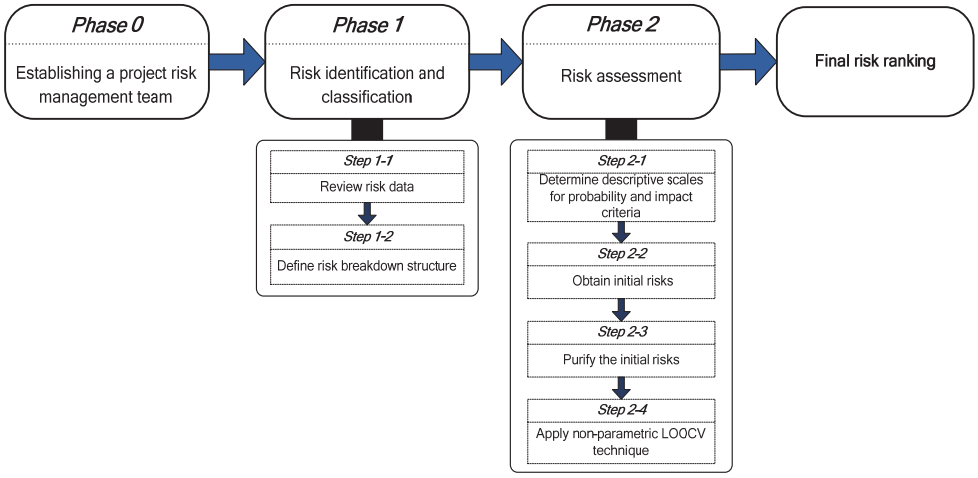


Fig. 1. Proposed non-parametric statistical approach for risk assessment in construction projects.

3.1 Principles of the LOOCV

Step 1. In the first step, principles of non-parametric cross-validation technique are described in order to resample project risks data from original observed risks data.

Step 2. In the second step, the cross-validation principle for estimating the SD of risk factors (RFs) is demonstrated in order to compare cross-validation resampled risk data with original observed project risks data.

Based on the first step of proposed approach, the cross-validation technique is a tool for uncertainty analysis based on resampling of experimentally observed data. Application of the cross-validation is justified by the so-called ''plug-in principle'', which means to take statistical properties of experimental results (=sample) as representative for the parent population. The main advantage of the cross-validation is that it is completely automatic. It is described best by setting two ''Worlds'', a ''Real World'' where the data is obtained and a ''Cross-validation World'' where statistical inference is performed, as shown in Fig. 2. The cross-validation partitions the data into two disjoint sets. The technique is fit with one set (the training set), which is subsequently used to predict the responses for the observations in the second set (assessment set).

Cross-validation techniques an intuitively appealing tool to calculate a predicted response value is to use the parameter estimates from the fit obtained with the entire data set with the exception of the observation to be predicted. This predicted response value of the y_i value is denoted by $\hat{y}_i$ (i=1, 2, ..., n). The LOOCV estimate of average prediction error is then computed using this predicted response value as:

$$\hat{\Delta}_{CV,1} = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2. \quad (1)$$

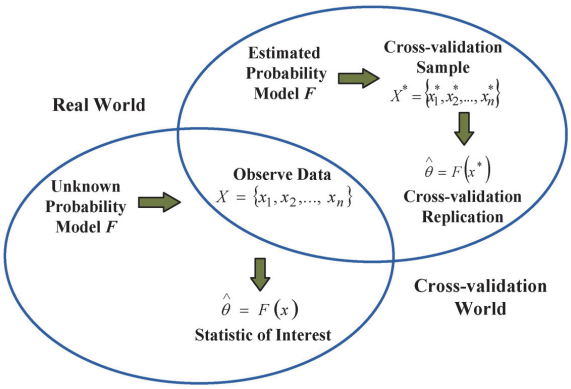


Fig. 2. Schematic diagram of the cross-validation technique.

Generally, in *K*-fold cross-validation, the training set omits approximately n/K observations from the training set. To predict the response values for the *k*th assessment set, $S_{k,a}$, all observations apart from those in $S_{k,a}$ are in the training set, $S_{k,t}$. $S_{k,t}$ is used to estimate the model parameters. The *K*-fold cross-validation average prediction error computed as:

$$\hat{\Delta}_{CV,K} = n^{-1}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(k,t)}\right)^2, \quad (2)$$

where $\hat{y}_{(k,t)}$ is the *i*th predicted response from $S_{k,a}$ (Wisnowski et al., 2003).

K-fold cross-validation: This is the algorithm in detail:

- Split the dataset D_N into *k* roughly equal-sized parts.
- For the *k*th part *k=1,…,K* , fit the model to the other *K-1* parts of the data, and calculate the prediction error of the fitted model when predicting the *k*th part of the data.
- Do the above for *k=1,…,K* and combine the *K* estimates of prediction error.

Let $k(i)$ be the part of D_N containing the *i*th sample. Then the cross-validation estimate of the MSE prediction error is:

$$\mathrm{MSE}_{\mathrm{CV}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i^{-k(i)}\right)^2, \quad (3)$$

where $\hat{y}_i^{-k(i)}$ denotes the fitted value for other *i*th observation returned by the model estimated with the $k(i)$ th part of the data removed.

Leave-one-out cross-validation (LOOCV)*:* The cross-validation technique where *K=N* is also called the leave-one-out algorithm. This means that for each *i*th sample, *i=1,…, N.*

- Carry out the parametric identification, leaving that observation out of the training set.
- Compute the predicted value for the *i*th observation, denoted by $\hat{y}_i^{-i}$

The corresponding estimate of the mean squared error (MSE) is:

$$\mathrm{MSE}_{\mathrm{loo}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i^{-i}\right)^2. \tag{4}$$

The LOOCV often works well for estimating generalization error for continuous error functions such as the mean squared error, but it may perform poorly for discontinuous error functions such as the number of misclassified cases.

3.2 The linear case: mean integrated squared error

Let us compute now the expected prediction error of a linear model trained on D_N when this is used to predict for the same training inputs X a set of outputs y_{ts} distributed according to the same linear law but independent of the training output y. We call this quantity mean integrated squared error (MISE):

$$\begin{aligned}
&\mathrm{MISE} = E_{D_N, y_{ts}}\left[\left(y_{ts} - X\hat{\beta}\right)^T\left(y_{ts} - X\hat{\beta}\right)\right] \\
&= E_{D_N, y_{ts}}\left[\left(y_{ts} - X\beta + X\beta - X\hat{\beta}\right)^T\left(y_{ts} - X\beta + X\beta - X\hat{\beta}\right)\right] \\
&= N\sigma_{\mathrm{w}}^2 + E_{D_N}\left[\left(X\beta - X\hat{\beta}\right)^T\left(X\beta - X\hat{\beta}\right)\right].
\end{aligned} \tag{5}$$

Since

$$\begin{aligned}
&X\beta - X\hat{\beta} = X\beta - X\left(X^TX\right)^{-1}X^Ty \\
&= X\beta - X\left(X^TX\right)^{-1}X^T\left(X\beta + w\right) = -X\left(X^TX\right)^{-1}X^Tw,
\end{aligned} \tag{6}$$

we have

$$\begin{aligned}
&N\sigma_{\mathrm{w}}^2 + E_{D_N}\left[\left(X\beta - X\hat{\beta}\right)^2\right] = N\sigma_{\mathrm{w}}^2 + E_{D_N}\left[\mathrm{w}^TX(X^TX)^{-1}X^TX\left(X^TX\right)^{-1}X\right] \\
&= N\sigma_{\mathrm{w}}^2 + E_{D_N}\left[\mathrm{tr}\left(\mathrm{w}^T\mathrm{w}\right)\right] = \sigma_{\mathrm{w}}^2\left(N + p\right).
\end{aligned} \tag{7}$$

Then, we obtain that the residual sum of squares SSEemp returns a biased estimate of MISE, that is

$$E_{D_N}\left[\hat{\mathrm{SSE}}_{\mathrm{emp}}\right] = E_{D_N}\left[\mathrm{e}^T\mathrm{e}\right] \neq \mathrm{MISE}. \tag{8}$$

Replace the residual sum of squares with

$$\mathrm{e}^T\mathrm{e} + 2\sigma_{\mathrm{w}}^2 p \tag{9}$$

4. Case study (onshore gas refinery plant)

In this section, the proposed approach based on non-parametric cross-validation technique is applied in the construction phase of an onshore gas refinery plant in Iran. The purposes of

this case study are assessing the important risks of climate changes for the onshore gas refinery project.

Onshore gas refinery plants or fractionators are used to purify the raw natural gas extracted from underground gas fields and brought up to the surface by gas wells. The processed natural gas, used as fuel by residential, commercial and industrial consumers, is almost pure methane and is very much different from the raw natural gas.

South Pars gas field in one of the largest independent gas reservoirs in the world situated within the territorial waters between Iran and the state of Qatar in the Persian Gulf. It is one of the country's main energy resources. South Pars gas field development shall meet the growing demands of natural gas for industrial and domestic utilization, injection into oil fields, gas and condensate export and feedstock for refineries and the petrochemical industries (POGC, 2010).

This study has been implemented into 18 phases of south pars gas field development in Iran. The location of the onshore refinery plant is illustrated in main WBS of South Pars Gas Field Development (SPGFD) in Fig. 3. The objectives of developing this refinery plant are as follows:

- Daily production of 50 MMSCFD (Million Metric Standard Cubic Feet per Day) of natural gas
- Daily production of 80,000 bls of gas condensate
- Annual production of 1 million tons of ethane
- Annual production of 1.05 million tons of liquid gas, butane and propane
- Daily production of 400 tons of sulphur

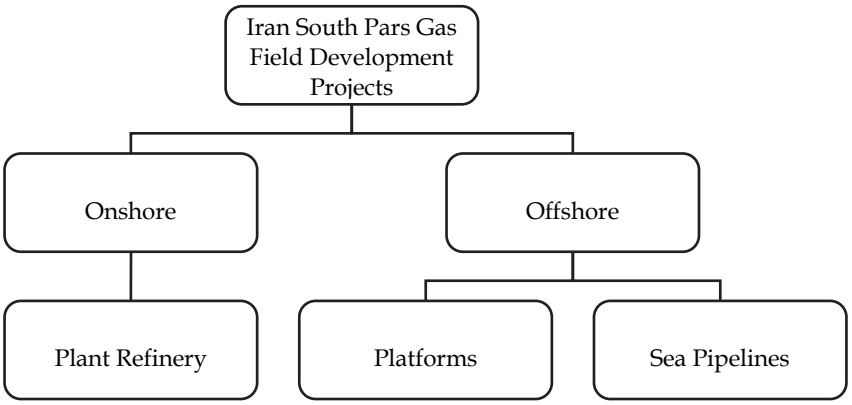


Fig. 3. Location of the onshore refinery plant in South Pars Gas Field Development

The contract type of above mentioned project is MEPCC, which includes management, engineering, procurement, construction and commissioning. In MEPCC contract, the MEPCC contractor agrees to deliver the keys of a commissioned plant to the owner for an agreed period of time. The MEPCC way of executing a project is gaining importance worldwide. But, it is also a way that needs good understanding, by the MEPCC, for a profitable contract execution. The MEPCC contract, especially in global context, needs thorough understanding. The MEPCC must be informed of the various factors that impact on the process of work, the results and success or failure of the contract, in global arena. The MEPCC must have data and expertise in all the required fields.

In this paper, risks of climate changes are considered from general contractor's (GC) perspective. The GC receives work packages from the owner and delivers them to subcontractors by bidding and contracting. This contractor is in charge of monitoring the planning, engineering, designing, and constructing phases. Moreover, the installation, leadership, and the payment of the subcontractors are burdened by the GC. The following risks of climate changes in Table 1 are identified by gathering historical information often performed in construction phase of gas refinery projects in Iran.

Risk	Description
1	Sea level rise
2	Flood
3	Earthquake
4	Bush fire
5	Tsunami
6	Sand storm
7	Increased atmospheric CO2
8	Precipitation patterns & amount
9	Increased temperature
10	Hurricane

Table 1. Climate change risk description

4.1 Apply the proposed approach to assess the risks of climate changes

In this sub-section, we show how the proposed approach can be used in a risk assessment according to the lack of risk sample data and periodic features of the construction projects. Hence, the comparison of the mean and the SD between the original sample distribution and the cross-validation resampled distribution can produce a better result.

In a risk analysis, we consider two indexes, which are probability and impact. The probability of a risk is a number between 0-1; however, the impact of a risk is qualitative. Though, it should be changed to a quantitative number, just like probability, a number between 0-1. The definitions of two indexes are as follows:

- Probability criterion: Risk probability assessment investigates likelihood that each specific risk will occur.
- Impact criterion: Risk impact assessment investigates potential effect on a project objective such as time, cost, scope, or quality.

The RF is computed as follows (Chapman & Ward, 2004; Chapman, 2001):

$$RF_{ij} = P_{ij} + I_{ij} - (P_{ij} \times I_{ij}) \tag{10}$$

The RF, from (0) low to 1 (high), reflects the likelihood of a risk arising and the severity of its impact. The risk factor will be high if the likelihood of P is high, or the consequence I is high, or both. Note that the formula only works if P and I are on scales from 0 to 1. Mathematically it derives from the probability calculation for disjunctive events:

$$\text{Prob (A or B)} = \text{Prob(A)} + \text{Prob(B)} - \text{Prob(A)} * \text{Prob(B)} \tag{11}$$

Two events are said to be independent if the occurrence or nonoccurrence of either one in no way affects the occurrence of other. It follows that if events A and B are independent events, then Prob(A and B)= Prob(A)*Prob(B). Two events are said to be mutually exclusive if the occurrence of either one precludes the occurrence of the other, then Prob(A and B)=0.
As far as probability and impact of project risks are independent; therefore, the formula functions properly in risk analysis and is merely a useful piece of arithmetic for setting risk ranking and priorities. Ten different risks have been identified for which we consider ten probabilities and ten impacts each that form our sample. It means that according to Eq. (10) we have P_{ij} which is the probability of the ith risk and jth observation and I_{ij} which is the impact of the ith risk and the jth observation. It is worthy to mentioning that experts are asked to estimate the probability and impact of each risks in a scale of very low (VL) to very high (VH) based on Table 2 (Chapman, 2001), their estimation are gathered and provided in Table 3. Consequently, gathered data (linguistic variables) are converted to numerical value and results are shown in Table 4.

Scale	Probability	Impact		
		Time	Cost	Performance
Very Low (VL)	< 10%	< 1 week	< 0.1 M USD	Failure to meet specification clause
Low (L)	10-30%	1-5 weeks	0.1-0.5 M USD	Failure to meet specification clauses
Medium (M)	31-50%	5-10 weeks	0.5-5 M USD	Minor shortfall in brief
High (H)	51-70%	10-15 weeks	5-20 M USD	Major shortfall in satisfaction of the brief
Very High (VH)	> 70%	> 15 weeks	> 20 M USD	Project does not satisfy business objectives

Table 2. Measures of probability and impact (M USD: Million US Dollar).

Risk	DMs									
	1		2		3		4		5	
	P	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*
R_1	VH	VH	H	VH	VH	H	H	VH	M	H
R_2	H	M	H	H	M	H	H	VH	M	H
R_3	VH	H	H	M	H	H	M	H	H	M
R_4	L	M	L	H	M	L	L	H	L	M
R_5	M	M	L	H	L	M	M	H	L	L
R_6	H	H	H	M	M	VH	H	H	M	H
R_7	H	VH	H	H	VH	VH	VH	VH	VH	H
R_8	VH	H	VH	VH	M	H	VH	M	H	M
R_9	M	H	L	H	H	H	M	M	H	VH
R_{10}	H	H	VH	H	M	H	H	H	H	H
Risk	DMs									
	6		7		8		9		10	
	P	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*
R_1	H	H	VH	H	H	H	VH	VH	VH	VH
R_2	M	H	VH	M	H	H	M	H	VH	H
R_3	M	M	H	H	M	H	M	H	H	M
R_4	L	H	H	L	L	M	M	M	M	L
R_5	M	L	M	L	M	M	L	L	H	M
R_6	H	M	H	VH	M	H	H	H	H	H
R_7	H	H	H	VH	M	H	VH	H	VH	VH
R_8	VH	M	H	H	VH	H	M	VH	H	H
R_9	M	H	H	M	M	H	H	H	M	H
R_{10}	M	M	M	H	VH	H	M	VH	H	H

Table 3. Risk observed data presented by linguistic variables.

Risk	DMs									
	1		2		3		4		5	
	P	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*
R_1	0.85	0.85	0.60	0.85	0.85	0.60	0.60	0.85	0.40	0.60
R_2	0.60	0.40	0.60	0.60	0.40	0.60	0.60	0.85	0.40	0.60
R_3	0.85	0.60	0.60	0.40	0.60	0.60	0.40	0.60	0.60	0.40
R_4	0.20	0.40	0.20	0.60	0.40	0.20	0.20	0.60	0.20	0.40
R_5	0.40	0.40	0.20	0.60	0.20	0.40	0.40	0.60	0.20	0.20
R_6	0.60	0.60	0.60	0.40	0.40	0.85	0.60	0.60	0.40	0.60
R_7	0.60	0.85	0.60	0.60	0.85	0.85	0.85	0.85	0.85	0.60
R_8	0.85	0.60	0.85	0.85	0.40	0.60	0.85	0.40	0.60	0.40
R_9	0.40	0.60	0.20	0.60	0.60	0.60	0.40	0.40	0.60	0.85
R_{10}	0.60	0.60	0.85	0.60	0.40	0.60	0.60	0.60	0.60	0.60
Risk	DMs									
	6		7		8		9		10	
	P	*I*	*P*	*I*	*P*	*I*	*P*	*I*	*P*	*I*
R_1	0.60	0.60	0.85	0.60	0.60	0.60	0.85	0.85	0.85	0.85
R_2	0.40	0.60	0.85	0.40	0.60	0.60	0.40	0.60	0.85	0.60
R_3	0.40	0.40	0.60	0.60	0.40	0.60	0.40	0.60	0.60	0.40
R_4	0.20	0.60	0.60	0.20	0.20	0.40	0.40	0.40	0.40	0.20
R_5	0.40	0.20	0.40	0.20	0.40	0.40	0.20	0.20	0.60	0.40
R_6	0.60	0.40	0.60	0.85	0.40	0.60	0.60	0.60	0.60	0.60
R_7	0.60	0.60	0.60	0.85	0.40	0.60	0.85	0.60	0.85	0.85
R_8	0.85	0.40	0.60	0.60	0.85	0.60	0.40	0.85	0.60	0.60
R_9	0.40	0.60	0.60	0.40	0.40	0.60	0.60	0.60	0.40	0.60
R_{10}	0.40	0.40	0.40	0.60	0.85	0.60	0.40	0.85	0.60	0.60

Table 4. Converted risk observed data.

A sampling distribution is based on many random samples from the population. In place of many samples from the population, create many resamples by repeatedly sampling with replacement from this one random sample. The sampling distribution of a statistic collects the values of the statistic from many samples. The cross-validation distribution of a statistic collects its values from many resamples. This distribution gives information about the sampling distribution. A set of n values are randomly sampled from the population. The sample estimates RF is based on the 10 values $(P_1, P_2, ..., P_{10})$ and $(I_1, I_2, ..., I_{10})$. Sampling 10 values with replacement from the set $(P_1, P_2, ..., P_{10})$ and $(I_1, I_2, ..., I_{10})$ provides a LOOCV sample $(P_1^*, P_2^*, ..., P_{10}^*)$ and $(I_1^*, I_2^*, ..., I_{10}^*)$. Observe that not all values may appear in the

cross-validation sample. The LOOCV sample estimate RF^* is based on 10 cross-validation values $\left(P_1^*, P_2^*, ..., P_{10}^*\right)$ and $\left(I_1^*, I_2^*, ..., I_{10}^*\right)$. The sampling of $\left(P_1, P_2, ..., P_{10}\right)$ and $\left(I_1, I_2, ..., I_{10}\right)$ with replacement is repeated many times (say n times), each time producing a LOOCV estimate RF^*.

Call the means of these resamples $\overline{RF}^*$ to distinguish them from the mean $\overline{RF}$ of the original sample. Find the mean and SD of the $\overline{RF}^*$ in the usual way. To make clear that these are the mean and SD of the means of the cross-validation resample rather than the mean $\overline{RF}$ and standard deviation of the original sample, we use a distinct notation:

$$mean_{LOOCV} = \frac{1}{n}\sum \overline{RF}^* \tag{12}$$

$$SD_{LOOCV} = \sqrt{\frac{1}{n-1}\sum\left(\overline{RF}^* - mean_{LOOCV}\right)^2} \tag{13}$$

Due to the fact that a sample consists of few observed samples, which is the nature of the construction projects, we use the LOOCV technique to improve the accuracy of the calculation of the mean and SD for the RF of the risks which may occur in a project.

4.2 Results

To do the resampling replications, we used resampling Stat Add-in of Excel software. We compare the original sample and LOOCV resample of the data provided by the Excel Add-in to see what differences it makes. In Table 5, the statistical data of the original sample is presented.

Risk	P (mean)	I (mean)	RF (mean)	P (SD)	I (SD)	RF (SD)
R_1	0.705	0.725	0.913	0.164	0.132	0.074
R_2	0.570	0.585	0.827	0.175	0.125	0.078
R_3	0.545	0.520	0.782	0.146	0.103	0.078
R_4	0.300	0.400	0.596	0.141	0.163	0.081
R_5	0.340	0.360	0.576	0.135	0.158	0.145
R_6	0.540	0.610	0.825	0.097	0.151	0.065
R_7	0.705	0.725	0.913	0.164	0.132	0.074
R_8	0.685	0.590	0.879	0.189	0.165	0.076
R_9	0.460	0.585	0.774	0.135	0.125	0.084
R_{10}	0.570	0.605	0.831	0.175	0.107	0.092

Table 5. Statistical data of the original sample.

After LOOCV resample replications, we obtain the mean for P, I and RF, and the SD for them. The data are reported in Table 6.

Risk	*P* (mean)	*I* (mean)	*RF* (mean)	*P* (SD)	*I* (SD)	*RF* (SD)
R_1	0.728	0.731	0.918	0.048	0.052	0.024
R_2	0.554	0.584	0.819	0.059	0.048	0.027
R_3	0.547	0.529	0.787	0.052	0.029	0.035
R_4	0.293	0.409	0.598	0.047	0.042	0.037
R_5	0.331	0.358	0.571	0.055	0.050	0.034
R_6	0.553	0.574	0.814	0.019	0.054	0.025
R_7	0.693	0.722	0.910	0.077	0.049	0.034
R_8	0.672	0.595	0.872	0.044	0.058	0.016
R_9	0.469	0.564	0.770	0.035	0.029	0.020
R_{10}	0.559	0.598	0.823	0.067	0.030	0.043

Table 6. Statistical data of the LOOCV resample.

Then MSE are calculated for all significant risks for *P*, *I* and RF, based on the LOOCV principles. Results are shown in Table 7.

Risk	*P* MSE	*I* MSE	*RF* MSE_{LOOCV}
R_1	0.042	0.031	0.009
R_2	0.056	0.025	0.010
R_3	0.035	0.017	0.009
R_4	0.033	0.042	0.014
R_5	0.037	0.045	0.036
R_6	0.015	0.049	0.009
R_7	0.058	0.034	0.011
R_8	0.060	0.054	0.010
R_9	0.031	0.023	0.011
R_{10}	0.059	0.023	0.019

Table 7. MSE calculation for risk data.

5. Discussion and test

In this section, according to the computational results in construction of the gas refinery plant, discussion and testing of the proposed approach are presented. Reduction of standard variations, MSE comparison and normality plot are the main topics of discussion.

5.1 Reduction of standard deviations

In Fig. 4, the SDs for higher risks of the project are reduced remarkably and it shows the efficiency of the proposed approach in the project risk assessment. The results show that the proposed approach is practical and logical for estimating the SD particularly in the construction projects.

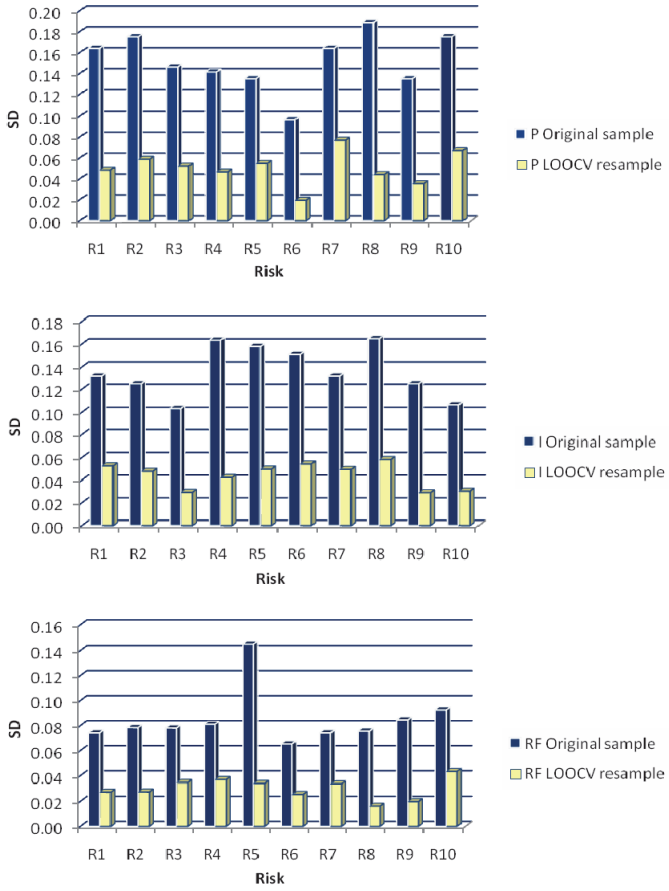


Fig. 4. Standard deviation comparison between original sample and LOOCV resample

Comparison between the SD of the original sample and LOOCV resample in the RF point of view shows that, for instance the SD of risk 1 of the original sample is 0.074 where the SD of the same risk with LOOCV resample is 0.024. In other words, the SD has been reduced about 63% for this risk. These reductions can emphasize that the LOOCV is making a better result in accuracy of the RF for each risk in the construction projects. Then, the SD reduction rate is computed by:

$$SD_{\mathrm{Red}}\% = \frac{SD_O - SD_{LOOCV}}{SD_O} \times 100, \tag{14}$$

where SD_{Red} denotes the rate of SD reduction through the LOOCV, SD_O represents the SD for the original RF data sample and SD_{LOOCV} indicates the SD for the LOOCV. The SD reduction rate is presented in Table 8 for each risk.

Risk	SD Reduction %		
	P	*I*	*RF*
R_1	70.57	60.19	63.36
R_2	66.40	61.84	65.12
R_3	64.24	71.67	55.56
R_4	67.04	74.18	54.37
R_5	59.34	68.54	76.61
R_6	79.86	64.01	61.06
R_7	53.05	62.55	54.67
R_8	76.60	64.67	78.67
R_9	73.74	76.81	76.56
R_{10}	61.72	71.55	53.24

Table 8. Rate of SD reduction for each risk

5.2 MSE interpretation and comparison

$RF_{Traditional}$ and RF_{LOOCV} are shown in Table 8; moreover, the absolute variances are calculated to illustrate that there is no significant difference between two techniques. Therefore, we should take advantage of MSE_{LOOCV} value for ranking project risks. For this purpose we ranked the risks based on MSE_{LOOCV} value in Table 7, smaller MSE_{LOOCV} means higher rank and priority of the project risk, results are shown in Tables 9 and 10. It is obvious that there is significant difference between traditional risk ranking techniques and ranking based on MSE_{LOOCV}. Based on traditional risk ranking technique, risk 1 (sea level rise) stands in the first priority, but based on MSE_{LOOCV} risk 3 (earthquake) stands in the first priority, which this results are much applicable in gas refinery construction projects in Iran.

Risk	$RF_{Traditional}$	RF_{LOOCV}	$Abs(RF_{Traditional} - RF_{LOOCV})$
R_1	0.913	0.918	0.005
R_2	0.827	0.819	0.008
R_3	0.782	0.787	0.005
R_4	0.596	0.598	0.002
R_5	0.576	0.571	0.005
R_6	0.825	0.814	0.011
R_7	0.913	0.910	0.003
R_8	0.879	0.872	0.007
R_9	0.774	0.770	0.004
R_{10}	0.831	0.823	0.008

Table 9. Comparison between traditional risk ranking and LOOCV ranking.

Traditional risk ranking	Risk ranking based on MSELOOCV
R1	R3
R7	R6
R8	R1
R10	R2
R2	R8
R6	R7
R3	R9
R9	R4
R4	R10
R5	R5

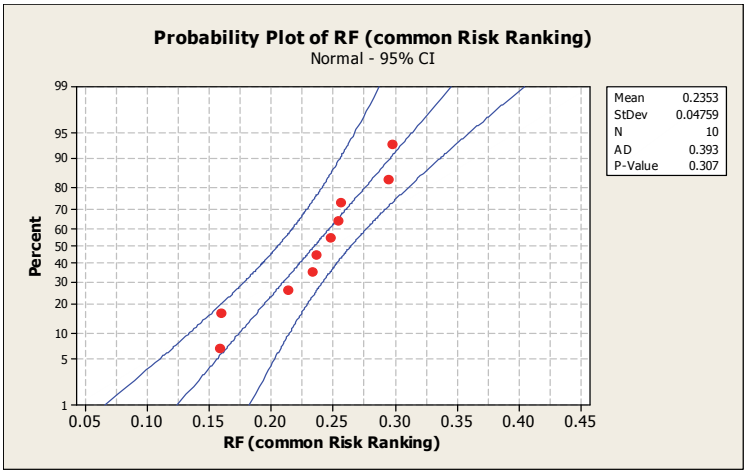
Table 10. Ranking comparison between traditional risk ranking and LOOCV ranking.

In the following, we compare the original sample (i.e., traditional approach) and cross-validation resample (i.e., cross-validation approach) in two respects: normal probability plot (NPP) and matrix plot (MP).

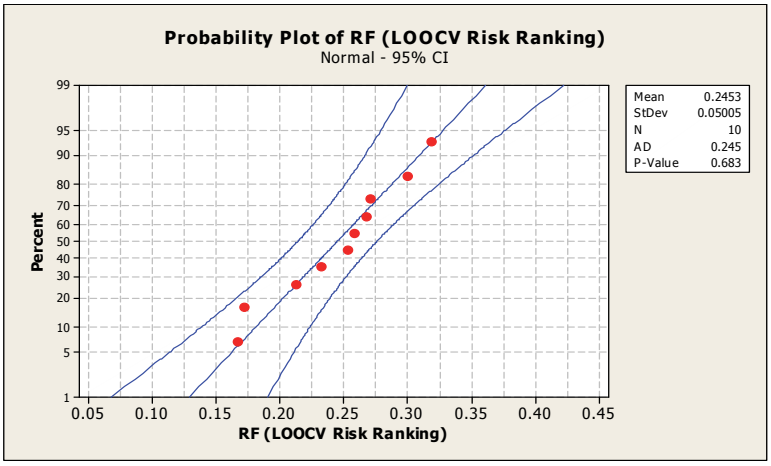
5.3 Normal probability plot and matrix plot

Normal probability plot: The NPP is a graphical presentation for normality testing; assessing whether or not a data set is approximately normally distribution. In other words, the NPP is a standard graphical display that can be used to see deviations from normality. The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. In other words, the NPP is a special case of the probability plot, for the case of a normal distribution. Two NPP are illustrated for RFs in the studied case in Fig. 5. Parts A and B of Fig. 5 are the NPP for the initial and final RFs, respectively. We notice that the NPP is basically straight.. Consequently, when the NPP is straight, we have evidence that the data is sampled from a normal distribution. Before running the LOOCV, the normality plot is drawn in part A, as it is clear, the data are not distributed straightly. The NPP is generally not normal. But, based on part B, the data are distributed closer to mean. Therefore, the normality degree in part B (after running the cross-validation) is higher than normality degree in part A (before running the cross-validation) or data has a distribution that is not far from normal.

To apply the LOOCV idea, we should start with a statistic that estimates the parameter, in which we are interested in. We come up with a suitable statistic by appealing to another principle that we often apply without thinking about it. In this sub-section, the proposed approach clearly shows that the distribution of the original samples do not exactly follows the normal distribution; however, the distribution of the LOOCV follows the normality when the LOOCV is applied for the construction project risks (see Fig. 5.). In comparison between the original samples and LOOCV resamples for different risks, it is evident that the LOOCV resamples are close to the normal distribution in comparison with the original samples of project risks.



Part A: NPP for the RFs before running the LOOCV



Part B: NPP for the RFs after running the LOOCV

Fig. 5. NPP for the RFs in the project.

Matrix plot: *A MP is a kind of scatter plot which enables the user to see the pair wise relationships between variables. Given a set of variables Var1, Var2, Var3, ... the MP contains all the pair wise scatter plots of the variables on a single page in a matrix format. The matrix plot is a square matrix where the names of the variables are on the diagonals and scatter plots everywhere else. That is, if there are k variables, the scatter plot matrix will have k rows and k columns and the ith row and jth column of this matrix is a plot of Vari versus Varj. The axes and the values of the variables appear at the edge of the respective row or the column. One can observe the behavior of variables with one another at a glance. The comparison of the variables under study and their interaction with one another can be studied easily as depicted in Fig. 6 for the construction project. This is why the matrix plots are becoming increasingly traditional in the general purpose statistical software programs.*

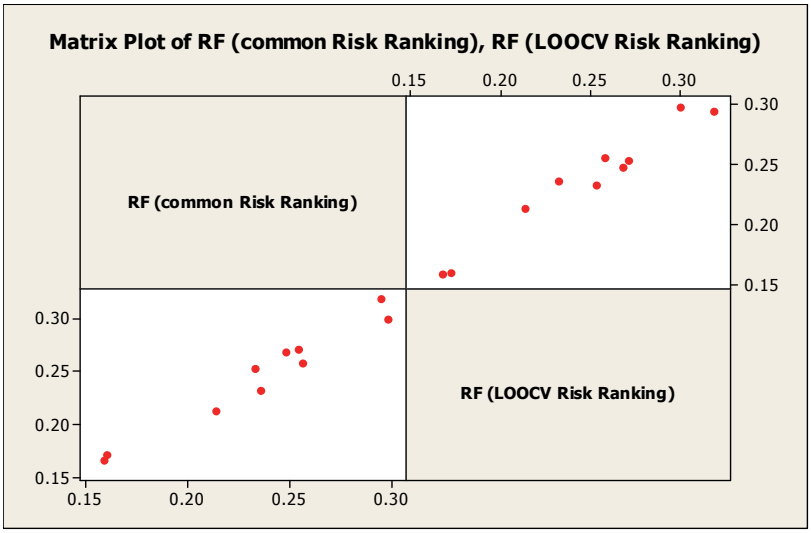


Fig. 6. MP for the RFs befor and after running the LOOCV in the project

The researchers have shown that resampling-based procedure based on the cross-validation) can be easily applied to risk assessment in construction projects. In this paper, routines for implementing the procedures described were calculated in Stat Add-in of Excel. Having considered all different aspects involved in the projects' characteristics, proposed LOOCV approach is very useful for risk assessment in these projects, because of the fact that it provides accurate calculation which was discussed in this section. To ensure the performance of the approach, the potential experts in the construction projects are requested to check the risk approach prepared by using non-parametric statistical technique for applicability, efficiency, and the overall performance of the approach. They confirmed the results of proposed approach in the real world of large-scale construction projects.

6. Conclusion

In this paper, we have attempted to introduce the effective framework based on LOOCV technique to the academia and practitioners in real –life situations. The cross-validation technique has been used subsequently to solve many other engineering and management problems that would be complicated for a traditional statistical analysis. In simple words, the cross-validation does with the computer what we would do in practice, if it was possible, we would repeat the experiment. Moreover, the LOOCV technique is extremely valuable in situations where data sizes are small and limited, which is often the real case in applications of project risk assessment. In the proposed model, the basic principle of the LOOCV technique was explained for analyzing risks where a particular family of probability distributions is not specified and original risk data sizes are small. In particular, we have explained the LOOCV principle for estimating the SD of RFs associated with climate change issues in the construction project. We have found that the LOOCV has greater accuracy for estimating the SD of RFs than estimating the SD from original risks data. SDs for RFs were remarkably reduced when the non-parametric LOOCV was applied.

It has been found that the distribution of the original samples did not exactly follow the normal distribution; however, the distribution of the LOOCV followed the normality when the proposed approach was applied (normality checking). Then, the NPP was provided in order to compare the traditional ranking and the proposed ranking for the climate change risks. The related results demonstrated that the proposed approach could assist top managers to better assess the risks of climate changes in the gas refinery plant construction in Iran. In future research, we may work on comparison of different non-parametric resampling techniques on risk data of climate changes of construction projects.

7. References

Bjorck, J.P.; Braese, R.W., Tadie, J.T., & Gililland, D.D. (2010). The adolescent religious coping scale: development, validation, and cross-validation, *J. Child Fam. Stud.,* Vol. 19, No. 3, pp. 343–359.

Caltrans. (2007). *Project risk management handbook*, California Department of Transportation (Caltrans), Office of Project Management Process Improvement, Sacramento, CA.

Chapman, C. & Ward, S. (2004). *Project risk management: processes, techniques and insights*, Second ed., John Wiley and Sons Ltd.

Chapman, R.J. (1998). The effectiveness of working group risk identification and assessment techniques. *International Journal of Project Management*, Vol.16, pp. 333-343.

Chapman, R.J. (2001). The controlling influences on effective risk identification and assessment for construction design management. *International Journal of Project Management*, Vol. 19, pp. 147–160.

Cooper, D.F.; Grey, S., Raymond, G., & Walker, P. (2005). Project risk management guidelines: managing risk in large projects and complex procurements, John Wiley & Sons.

Ebrahimnejad, S.; Mousavi, S.M. & Mojtahedi, S.M.H. (2008). A model for risk evaluation in construction projects based on fuzzy MADM, *Proceedings of 4th IEEE International Conferences on Management of Innovation & Technology*, Thailand, pp. 305–310,.

Ebrahimnejad, S.; Mousavi, S.M. & Mojtahedi, S.M.H. (2009). A fuzzy decision making model for risk ranking with application to the onshore gas refinery. *International Journal of Business Continuity and Risk Management*, Vol.1, No.1, pp. 38–66.

Ebrahimnejad, S.; Mousavi, S.M. & Seyrafianpour, H. (2010). Risk identification and assessment for build-operate-transfer projects: a fuzzy multi attribute decision making model. *Expert Systems with Applications*, Vol.37, No.1, pp. 575–586.

Efron, B., (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation, *J. Am. Stat. Assoc.*, Vol. 78, No. 382, pp. 316–331.

Efron, B.; and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Fan, M.; Lin, N.-P. & Sheu, C. (2008). Choosing a project risk-handling strategy: An analytical model. No.2, pp. 101–105.

FHWA. (2006). *Risk assessment and allocation for highway construction management.* <http://international. fhwa.dot.gov/riskassess/>. Accessed 14.11.07.

Florice, S.; & Miller, R. (2001). Strategizing for anticipated risks and turbulence in large-scale engineering projects. *International Journal of Project Management,* Vol. 19 pp. 445–455.

Geisser, S., (1975). The predictive sample reuse method with applications, *J. Am. Stat. Assoc.*, Vol. 70, pp. 320–328.

Geisser, S.; & Eddy, W.F. (1979). A predictive approach to model selection. *J. Am. Stat. Assoc.*, Vol. 74, No.365, pp. 153-160.

Grabowski, M.; Merrick, J.R.W.; Harrold, J.R.; Massuchi, T.A.; van Dorp, J.D.; Bus. Dept.; Le Moyne Coll.; & Syracuse, N.Y. (2000). Risk modelling in Distributed, Large-Scale Systems. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions*, Vol. 30, pp. 651-660.

Gray, C.F.; & Larson, E.W. (2005). *Project Management: The Management Process*, 3rd ed., New York: McGraw-Hill.

Hallegatte, S. (2009). Strategies to adapt to an uncertain climate change, *Global Enviromental Change*, Vol.19, pp.240-247.

Hashemi, H.; Mousavi, S.M. & Mojtahedi, S.M.H. (2011). Bootstrap technique for risk analysis with interval numbers in bridge construction projects, *Journal of Construction Engineering and Management*, doi:10.1061/(ASCE)CO.1943-862.0000344.

Hastak, M.; & Shaked, A. (2000). ICRAM-1: Model for international construction risk assessment. *Journal of Management in Engineering*, Vol. 16, No. 1, pp. 59-69.

Hillson, D.; (2002). Extending the risk process to manage opportunities. *International Journal of Project Management*, Vol. 20, pp. 235–240.

Huang, G.H., Cohen, S.J., Yin, Y.Y., and Bass, B. (1998). Land resources adaptation planning under changing climate – A study for the Mackenzie Basin, Resources. *Conservation and Recycling*, Vol. 24, pp. 95–119.

Hubert, M.& Engelen, S. (2007), Fast cross-validation of high-breakdown resampling methods for PCA. Computational Statistica and Data Analysis, Vol.51, No. 10, pp. 5013-5024.

Iranmanesh, H.; Jalili, M., & Pirmoradi, Zh. (2007). Developing a new structure for determining time risk priority using risk breakdown matrix in EPC projects. *2007 IEEE International Conference on Industrial Engineering and Engineering Management*, Singapore.

Lewsey, C.; Cid, G. & Kruse, E. (2004). Assessing climate change impacts on coastal infrastructure in the Eastern Caribbean. *Marine Policy,* Vol. 28, pp. 393–409.

Makui, A.; Mojtahedi, S.M.H. & Mousavi, S.M. (2010). Project risk identification and analysis based on group decision making methodology in a fuzzy environment. *International Journal of Management Science and Engineering Management*, Vol.5, No.2, pp. 108–118.

Miller, R.; & Lessard, D. (2001). Understanding and managing risks in large engineering projects. *International Journal of Project Management,* Vol. 19, pp. 437–443.

Mojtahedi, S.M.H.; Mousavi, S.M., & Aminian, A. (2008). Fuzzy group decision making: A case using FTOPSIS in mega project risk identification and analysis concurrently, *The 5th of IEEE International Conference on Industrial Engineering and Engineering Management*, Singapore.

Mojtahedi, S.M.H.; Mousavi, S.M. & Makui, A. (2010). Project risk identification and assessment simultaneously using multi-attribute group decision making technique. *Safety Science*, Vol.48, No.4, pp. 499–507.

Mojtahedi, S.M.H.; Mousavi, S.M. & Aminian, A. (2009). A non-parametric statistical approach for analyzing risk factor data in risk management process. *Journal of Appled Science*, Vol.9, No.1, pp. 113–120.

Mousavi, S.M.; Tavakkoli-Moghaddam, R.; Azaron, A.; Mojtahedi, S.M.H. & Hashemi H. (2011). Risk assessment for highway projects using jackknife technique. *Expert Systems With Applications*,Vol.38, No.5, pp. 5514-5524.

Nicholls, R.J.; S. Hanson, C. Herweijer, N. Patmore, S. Hallegatte, J. Corfee-Morlot, J. Chateau, R. Muir-Wood, (2007). *Screening Study: Ranking Port Cities with High Exposure and Vulnerability to Climate Extremes*, OECD Working Paper, available on {http://www.oecd.org/document/56/0,3343,en_2649_201185_39718712_1_ 1_1_1,00.html}.

PMI. (2008). *A guide to the project management body of knowledge (PMBOK Guide)*, (4th ed.). USA: Project Management Institute Inc, (Chapter 11).

POGC, (2010), {Http://www.pogc.ir}.

Qin, X.S.; Huang, G.H., Chakma, A., Nie, X.H., Lin, Q.G. (2008). A MCDM-based expert system for climate-change impact assessment and adaptation planning, A case study for the Georgia Basin, Canada. *Expert Systems with Application*s, Vol. 34, pp. 2164–2179.

Shao, J., (1993). Linear model selection by cross-validation. J. *Am. Stat. Assoc*., Vol. 88, pp. 486–494.

Shen, L.Y.; Wu, G.W.C. & Ng, C.S.K. (2001). Risk assessment for construction joint ventures in China, *Journal of Construction Engineering and Management*, Vol.127, No. 1., pp. 76-81

Simon, J. L., & Bruce, P. (1995). The new biostatistics of resampling. MD computing, Vol. 12, pp. 115–121.

Smith, N.J. (1999). Managing risk in construction project. Oxford: Blackwell.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.36, No.2, pp. 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol.39, No.1, pp. 44–47.

Sugiyama, M.; Krauledat, M., & Müller, K. –R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, Vol. 8, pp. 985-1005.

Tavakkoli-Moghaddam, R.; Mojtahedi, S.M.H.; Mousavi, S.M. & Aminian A. (2009). A jackknife technique to estimate the standard deviation in a project risk severity data analysis. *in Proc IEEE Int Conf Comput Ind Eng (CIE39)*, France, pp. 1337–1341.

Teegavarapu, R. (2010). Modeling climate change uncertainties in water resources management models. *Environmental Modeling & Software*, Vol. 25, pp. 1261-1265

Tsai, T.-I., & Li, D.-C. (2008). Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems. Expert Systems with Applications, Vol. 35, No. 3, pp. 1293–1300.

Wang, S.Q.; Dulaimi, M.F., & Aguria, M.Y. (2004). Risk management framework for construction projects in developing countries. Construction Management and Economics, Vol. 22, pp. 237–252.

Wisnowski, J.W.; Simpson, J.R., Montgomery, D.C., & Runger G.C. (2003). Resampling methods for variable selection in robust regression. *Computational Statistics & Data Analysis*, Vol. 43, pp. 341–355.

Yin, Y. (2001). Designing an integrated approach for evaluating adaptation options to reduce climate change vulnerability in the Georgia Basin. *Final Report*, Climate Change Action Fund, Adaptation Liaison office.

Yin, Y.; & Cohen, S. (1994). Identifying regional policy concerns associated with global climate change. *Global Environmental Change*, Vol. 4, No. 3, pp. 245–260.

Zeng, J.; An, M., & Smith, N.J. (2007). Application of a fuzzy based decision making methodology to construction project risk assessment. *International Journal of Project Management*, Vol. 25, pp. 589–600.

Zou, P.X.W.; & Zhang, G., (2009). Managing risks in construction projects: life cycle and stakeholder perspectives. *The International Journal of Construction Management*, Vol. 9, No. 1, pp. 61-77.

5

Towards Knowledge Based Risk Management Approach in Software Projects

Pasquale Ardimento[1], Nicola Boffoli[1],
Danilo Caivano[1] and Marta Cimitile[2]
[1]*University of Bari Aldo Moro, Department of Informatics*
[2]*Faculty of Economy Unitelma Sapienza, Rome*
Italy

1. Introduction

All projects involve risk; a zero risk project is not worth pursuing. Furthermore, due to software project uniqueness, uncertainty about final results will always accompany software development. While risks cannot be removed from software development, software engineers instead, should learn to manage them better (Arshad et al., 2009; Batista Webster et al., 2005; Gilliam, 2004). Risk Management and Planning requires organization experience, as it is strongly centred in both experience and knowledge acquired in former projects. The larger experience of the project manager improves his ability in identifying risks, estimating their occurrence likelihood and impact, and defining appropriate risk response plan. Thus risk knowledge cannot remain in an individual dimension, rather it must be made available for the organization that needs it to learn and enhance its performances in facing risks. If this does not occur, project managers can inadvertently repeat past mistakes simply because they do not know or do not remember the mitigation actions successfully applied in the past or they are unable to foresee the risks caused by certain project restrictions and characteristics. Risk knowledge has to be packaged and stored over time throughout project execution for future reuse.

Risk management methodologies are usually based on the use of questionnaires for risk identification and templates for investigating critical issues. Such artefacts are not often related each other and thus usually there is no documented cause-effect relation between issues, risks and mitigation actions. Furthermore today methodologies do not explicitly take in to account the need to collect experience systematically in order to reuse it in future projects.

To convey these problems, this work proposes a framework based on the Experience Factory Organization (EFO) model (Basili et al., 1994; Basili et al., 2007; Schneider & Hunnius, 2003) and then use of Quality Improvement Paradigm (QIP) (Basili, 1989).

The framework is also specialized within one of the largest firms of current Italian Software Market. For privacy reasons, and from here on, we will refer to it as "FIRM". Finally in order to quantitatively evaluate the proposal, two empirical investigations were carried out: a post-mortem analysis and a case study. Both empirical investigations were carried out in the FIRM context and involve legacy systems transformation projects. The first empirical investigation involved 7 already executed projects while the second one 5 in itinere projects. The research questions we ask are:

Does the proposed knowledge based framework lead to a more effective risk management than the one obtained without using it?
Does the proposed knowledge based framework lead to a more precise risk management than the one obtained without using it?
The rest of the paper is organized as follows: section 2 provides a brief overview of the main research activities presented in literature dealing with the same topics; section 3 presents the proposed framework, while section 4 its specialization in the FIRM context; section 5 describes empirical studies we executed, results and discussions are presented in section 6. Finally, conclusions are drawn in section 7.

2. Related works

Efficient risk management methodologies must be devised and implemented in order to avoid, minimize or transfer the risks to external entities. For this reason risk management should be a mature process integrated with all other enterprise processes (Kànel et al., 2010). Unfortunately, risk analysis is rarely fully integrated with project management in Software Engineering. While Boehm (Boehm, 1989) has lain the foundations and Charette (Charette, 1990) outlined the applications, there have been few widely developed and used formal risk methodologies tailored for software development industry. Today risk methodologies are usually based on the identification, decomposition and analysis of events that can determine negative impacts on the projects (Farias et al., 2003; Chatterjee & Ramesh, 1999; Gemmer, 1997; Costa et al., 2007). Different approaches can be adopted to deal with the key risk factors: in (Arshad et al., 2009; Hefner, 1994; Donaldson & Siegel, 2007) some risk management activities and strategies are described. In (Hefner, 1994) authors propose a methodology based on the use of capabilities and maturity models, combined with risk and value creation factors analysis to reduce risk levels. In (Donaldson & Siegel, 2007), authors propose a five step process for incorporating risk assessment and risk derived resource allocation recommendations into project plan development. Furthermore, in (Kontio, 2001; Hefner, 1994) the Riskit approach is presented. It is a risk management process that provides accurate and timely information on the risks in a project and, at the same time, defines and implements cost efficient action to manage them.
Other assessment methods for risk and hazard analysis (Petroski, 1994; Croll et al., 1997; Stratton et al., 1998) rely on people making judgments based on their experience. For safety systems a detailed knowledge of what can go wrong is an essential prerequisite to any meaningful predictions regarding the cause and effects of systems failures. In (Petroski, 1994), Petroski takes this argument further by stating that teaching history of engineering failures should be a core requirement in any engineering syllabus and take the same importance as the teaching of modern technology. Without an understanding of history or direct experience for a given application then more is unknown and hence risks are higher (Croll et al., 1997). For this reason there is a big interest towards the techniques and tools for storing and share risk knowledge. Nevertheless, the major part of today known risk management methodologies lack in doing this. They do not use any mechanism, except for human memory, to address these needs. In (Dhlamini et al., 2009) SEI SRM methodologies risk management framework for software risk management is presented. This approach is based on the adoption of three groups of practices supporting the experience sharing and communication in enterprise.

In this sense the proposed framework can be considered a complementary infrastructure for collecting and reusing risk related knowledge. Thus it can be used jointly with all the existing methodologies that it contributes to enhance.

3. Proposed framework

The proposed framework is made up of two main components: a conceptual architecture and a risk knowledge package structure for collecting and sharing risk knowledge.

3.1 Conceptual architecture

Conceptual architecture (Figure 1) is based on two well-known approaches: EFO (Schneider, 2003) and the QIP (Basili, 1989; Kànel et al., 2010).

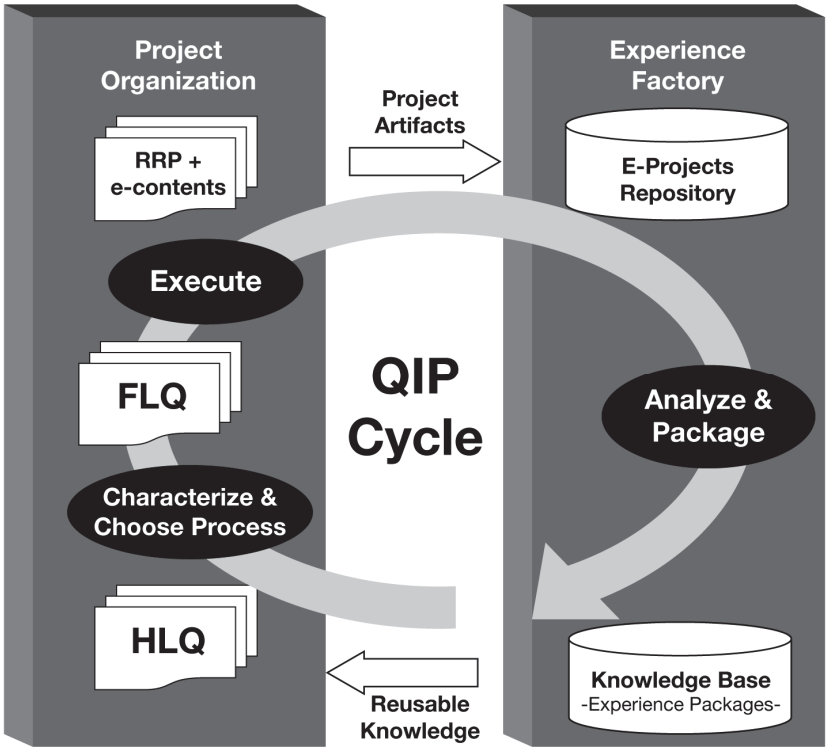


Fig. 1. Conceptual Architecture

EFO is an organizational approach for constructing, representing and organizing enterprise knowledge by allowing stakeholders to convert tacit into explicit knowledge. It distinguishes project responsibilities from those related to collection, analysis, packaging, and experience transfer activities. In doing so, it identifies two different organizational units: Project Organization (PO) and Experience Factory (EF). The first uses experience packages for developing new software solutions and the second provides specific knowledge ready to

be applied. To support these two infrastructures the QIP is used. It is based on the idea that process improvement can be accomplished only if the organisation is able to learn from previous experiences. During project execution measures are collected, and data are analysed and packaged for future use. In this sense QIP can be seen as organized in different cyclic phases (Characterize, Choose Process, Execute, Analyze and Package), that used in the organizations, perform and optimize the process of knowledge collection, packaging and transferring.

- CHARACTERIZE: it deals with the characterization of the project, the description of goals, project strategy to adopt and project planning. Such information are carried out by using focused assessment questionnaires which could have different abstraction levels (i.e. HLQ=High Level Questionnaire, FLQ=Functional Level Questionnaire). The information collected is interpreted by using the Knowledge-Base that suggests the appropriate actions to undertake in order to manage project risks.
- CHOOSE PROCESS: on the basis of the characterization of the project and of the goals that have been set, choose the appropriate processes, using the knowledge packages if present, for improvement, and supporting methods and tools, making sure that they are consistent with the goals that have been set.
- EXECUTE: it deals with the project plan execution and includes all the activities to perform for project execution. In this activities project and risk management knowledge is produces throughout project artefacts produced (e-contents) i.e. project documents, code, diagrams etc., identified risks together with the adopted mitigation actions (RRP - Risk Response Plan). They are stored in the E-Project Repository.
- ANALYZE: this phase continuously collects, analyses and generalises the information related to the executed/closed projects. After the closure of a project, such phase implies the comparison between planned and actual results, the analysis and generalization of strengths and weaknesses, risks occurred, response plans used and their effectiveness.
- PACKAGE: this phase packages experiences in the form of new, or updated and refined, models and other forms of structured knowledge gained from this and prior projects, and stores it in an experience base in order to make it available for future projects.

The proposed architecture supports the synergic integration between PO and EF. Such integration makes knowledge acquisition and reuse process incremental according to the QIP cycle that determines the improvement of the entire organization.

3.2 Structure of a knowledge package on the risk

The EFO model results to be independent from the way knowledge is represented. Nevertheless, its specialization in an operative context requires it to be tailored by using a specific knowledge representation approach.

Knowledge can be collected from several and different sources: document templates, spreadsheets for data collection and analysis, project documents, etc. In this work, an innovative approach for knowledge packaging has been defined. It is based on the use of decision tables (Ho et al., 2005; Vanthienen et al., 1998; Maes & Van Dijk, 1998).

In particular, a set of decision tables have been used to formalize knowledge first and then make it available for consultation. Knowledge means: project attributes exploitation of relations among the attributes, risks identified during project execution and consequent list of mitigation actions. According to the decision tables structure, an example of how they

have been used for representing knowledge on risk is provided (Figure 2). In the CONDITION quadrant there are project attributes (for example cost, time, available personnel, communication etc.); in CONDITIONAL STATES quadrant there are the possible values of project characteristics riskiness (based on scales different by type, at least ordinal, and depending on the attribute value which they relate to); in the ACTION QUADRANT there are risk drivers which must be taken into account (for example schedule, scarceness of personnel, etc.) together with the possible mitigation actions to carry out (for example increase the number of human resources allocated on a project, define a new date for project conclusion, etc). Finally in the RULES quadrant there are relationships between project characteristics, risk drivers and corresponding mitigation action to carrying out.

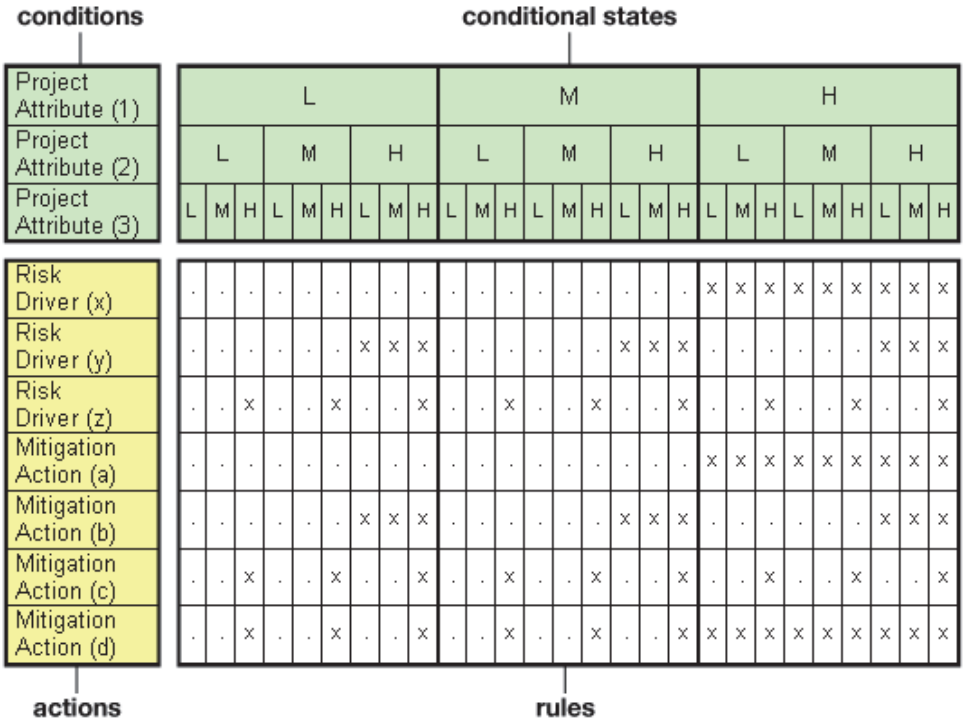


Fig. 2. An example of decision table oriented to risk management

This structure allows formalizing manager experience in risks management and, at the same time, to verify the effectiveness of the mitigation actions (Risk Response Plan). It also allows extending and updating previously acquired knowledge by adding, removing or modifying project attribute, risk driver and mitigation actions. For example, if a mitigation action results to be ineffective it can be deleted from knowledge package; project characterization, in ANALYZE and PACKAGE phases, can be enriched by adding new project attribute (i.e. context parameters).

4. Framework specialization

In order to obtain the information about FIRM context for formalizing the questionnaires and consequently the structure of the decision-tables, we carried out interviews with 50 FIRM project managers (according to the risk questionnaire in (Costa et al., 2007)). They deal with projects executed in a period of seven years. Collected data were analyzed to identify the suitable questions for risk investigation, the related risk drivers and mitigation actions. All this information was formalized as decision tables and was used to populate risk knowledge base. The steps followed were:

- Collected data by interviews were analyzed in order to extract risks from the projects occurred during their execution;
- Common risks were identified and their abstraction led us to define Risk Drivers (RD);
- Each identified risk was related to the effective mitigation actions (MA) executed;
- The most suitable questions to detect risks were identified and then related to risks;
- Questions, risks and mitigation actions were classified in relevant functional areas (Communications, Procurement, Cost, Quality, Resource, Schedule, and Scope).

The products of these activities were:

- two assessment questionnaires used to identify potential risk drivers;
- a knowledge base made of a set of decision tables used for formalizing the relationships between functional areas, risk drivers and mitigation actions

4.1 Assessment questionnaires

To identify project risks, usually the risk management process implies the use of assessment questionnaires during Risk Evaluation activity. Each questionnaire is made up of questions that support the project manager in discovering potential risks.

Typically, risk managers are supported through two different kinds of assessment questionnaires, their aim is to characterize the project by analyzing the different project management functional areas in order to assess, point out and further manage, the risks affecting a project.

In the FIRM context, two different types of questionnaires were used (example in figure 3):

High-Level Questionnaire (HLQ): questionnaire that assesses the general aspects of the projects, its aim is to generally characterize a project.

Functional-Level Questionnaire (FLQ): more specific questionnaire that points out specific issues related to the project (i.e. potential risks to mitigate), there is one specialized section for each project management functional area.

The questions of the questionnaire are answered by using a Low (L), Medium (M), High (H) scale.

The project manager starts with the HLQ for highlighting the general aspects of his project and then he uses one or more of the FLQ sections to discover the critical risk drivers and the mitigation actions related to a particular project management function (i.e. FLQs support the RRP definition).

A generalization of the relationships between HLQ, Project Management Functional Area assessed within FLQ and RD is shown in Figure 5.

It is important to underline that the use of questionnaires for knowledge execution is much diffused in industrial context, but typically, these relations between the different questionnaires and between questionnaire results and the consequent mitigation action

choice are tacit knowledge of the risk manager. Thus even when risks investigation is supported by assessment questionnaires it is usually quite subjective.
This implies the need of a risk knowledge package for collecting individual knowledge/experience previously acquired by managers during the execution of a project. The following section presents the knowledge base (i.e. a set of decision table) structured in the FIRM context.

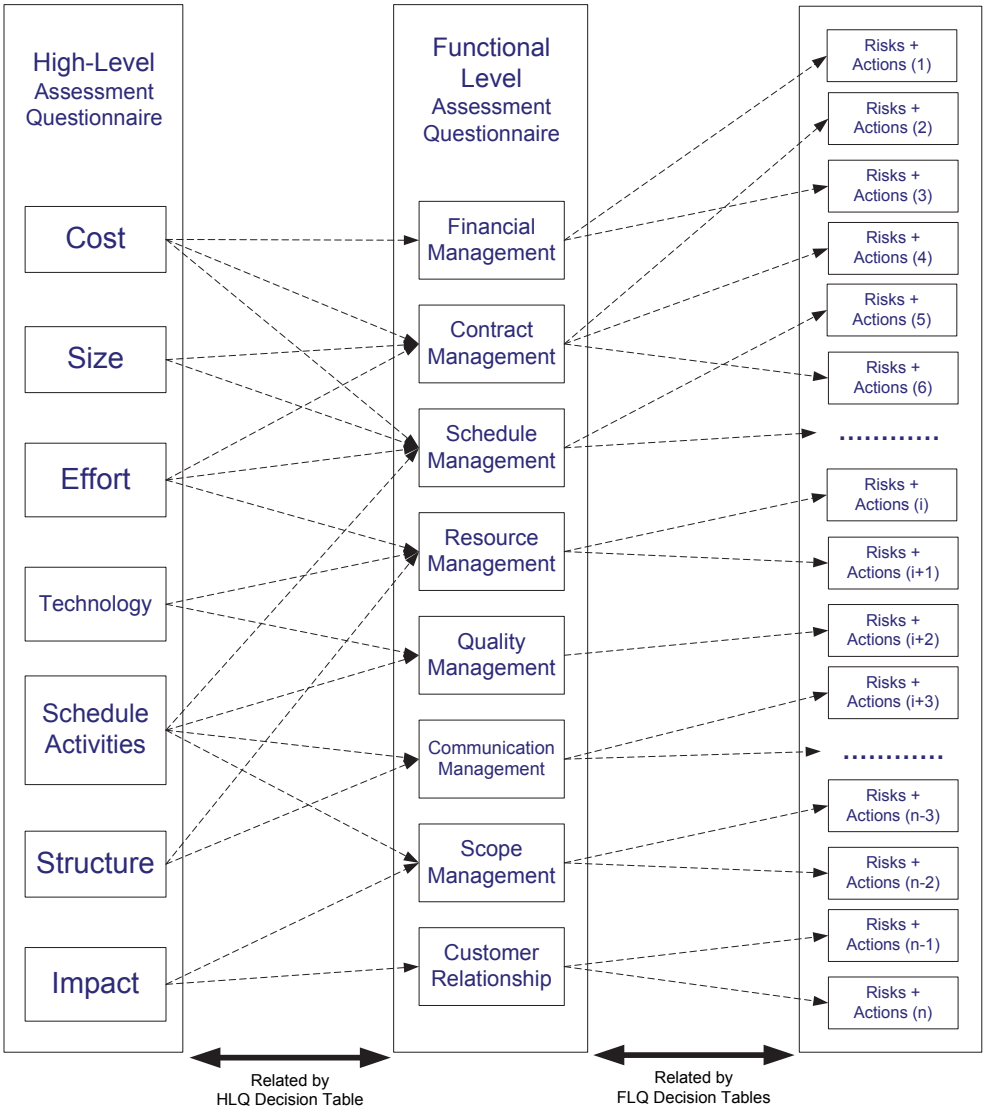


Fig. 3. Relationship schema HLQ-FLQ-Risks

4.2 Knowledge base

A set of decision tables have been used to formalize the relations between HLQ and FLQ; and to guide the project manager during the risk investigation activities. This set can be considered as an experience base. In particular, the following decision tables were introduced:

- 1 decision table for the HLQ: it relates to the general aspects of the project to more specific issues such as the functional areas of the project management that need further investigations (figure 4).
- 1 decision table for each functional area of FLQ: it relates to the specific issue of the functional area to the critical risk driver and consequent mitigation actions (figure 5).

4.2.1 Decision table supporting HLQ

In the CONDITION quadrant there are project attributes referring to general aspects of the project (for example cost, size, effort, technology, etc.); in the CONDITIONAL STATE quadrant there are possible values of project attributes (typically a Low-Medium-High scale); in the ACTION QUADRANT there are the functional areas of project management for further investigations. Finally, in the RULE quadrant there are the relationships between project attributes and the critical functional areas that need more specific investigation. An example is shown in figure 4.

1. Cost	M																H										
2. Size	H																-										
3. Effort	L or M				H												L or M										
4. Technology	H				L or M						H						L										
5. Schedule	H				L or M				H		L or M			H			L or M										H
6. Structure	L		M or H		L or M			H			-			-			-			L			M			H			L
7. Impact	L	M or H	L	M or H	L	M	H	L	M	H	L	M or H	L	M	H	L	M or H	L	M	H	L	M	H	L	M	H	L	M	H	L	M or H
1. ^Financial	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x
2. ^Contract	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
3. ^Schedule	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
4. ^Resource	.	.	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.	.	.	x	x	x	x	x	x	.	.
5. ^Quality	x	x	x	x	.	.	.	.	.	.	x	x	x	x	x	x	x	.	.	.	.	.	.	.	.	.	x	x
6. ^Communication	x	x	x	x	.	.	.	x	x	x	x	x	x	x	x	x	x	.	.	.	.	.	.	x	x	x	x	x
7. ^Scope	x	x	x	x	.	.	x	.	.	x	x	x	.	.	x	x	x	.	.	x	.	.	x	.	.	x	x	x
8. ^Customer	.	x	.	x	.	x	x	.	x	x	.	x	.	x	x	.	x	.	x	x	.	x	x	.	x	x	.	x
	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283

Fig. 4. An example of decision table supporting HLQ

4.2.2 Decision tables supporting the FLQ

The CONDITION quadrant contains specific issues related to the functional area; the CONDITIONAL STATE quadrant contains the possible value of each specific issue (typically a Low-Medium-High scale); in the ACTION QUADRANT there are risk drivers that must be faced (for example schedule, scarceness of personnel, etc.) together with the possible mitigation actions to carry out (for example increasing the number of human resources allocated on a project, defining a new date for project conclusion, etc). Finally the RULE quadrant identifies the relationship between specific issues, risk drivers and corresponding mitigation actions to carry out. An example is shown in figure 5.

1. Type of project organization	L									M									H									
2. Relationship of the organizational units in the project effort	L			M			H			L			M			H			L			M			H				
3. Preparation and commitment to project status reporting	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
1. RD: Project plan requires matrix combination of personnel and production functions	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.
2. MA: Require periodical meetings in the communication plan	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	x	.	.	.	.	.	.	.	.	.	.	.	.
3. RD: Project organization follows a matrix	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	x
4. MA: Require periodical meetings in the communication plan	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	x
5. RD: No management activity assigned to the project	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	x
6. MA: Periodical meetings to be held with the head of personnel	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	x	x	x	x	x	x	x	x	x	x	x	x
7. RD: Cooperation among organization units generates confliction	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.
8. MA: Identify a person for each organization unit, together with tasks for each role	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.
9. MA: Meetings in the communication plan with managers of each organization unit	.	.	.	x	x	x	x	x	x	.	.	.	x	x	x	x	x	x	.	.	.	x	x	x	x	x	x	x	x	x
10. RD: Relation among organization units generates confliction	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.
11. MA: Also involve the higher levels of the client organizations	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.	.	.	.	x	x	x	.	.	.
12. RD: Project team planned a status report, but no agreement about format/frequency	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.
13. MA: Define a standard template for the status report and identify its frequency	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.
14. RD: Project team has not planned to produce a status report	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x
15. MA: Periodical meetings with client management for illustrating project status	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x	.	.	x
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27			

Fig. 5. An example of decision table supporting FLQ

4.3 Scenario of consulting activity

The project manager answers to HLQ, each question included in HLQ corresponds to a condition (project attribute) of the related decision table; then the table interprets these responses and the actual actions are extracted. These actions are related to the functional areas of project management that need further investigation, therefore the actions guide the project manager in answering corresponding sections in the FLQ. Each section in FLQ corresponds to a specific decision table and then each selected question corresponds to a condition (specific issue) of the table which interprets these responses and then extracts the action. These actions are related to risk drivers and correspondent mitigation actions to carrying out. Therefore project managers might use issues, risk drivers and mitigation actions extracted in order to build the final Risk Response Plan (Figure 6).

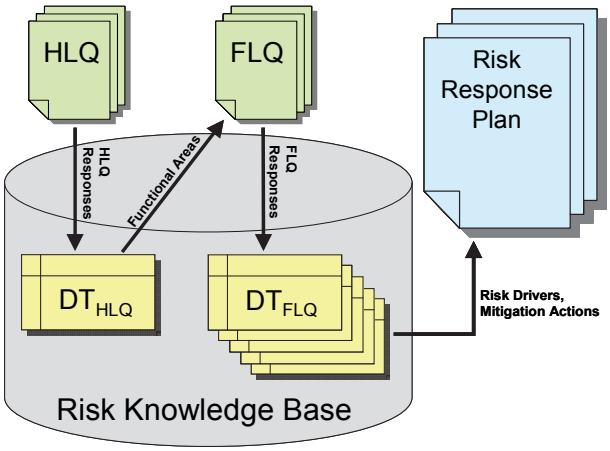


Fig. 6. Scheme of consulting activity

For example, according to Figure 4, one of the tuple corresponding to HLQ answers is (Cost, Size, Effort, Technology, Schedule, Structure, and Impact) = (M, H, H, L, L H, and L) for this tuple "Communication" is one of the critical areas to investigate. In figure 5, Communication area is investigated and one of the tuple obtained by the related FLQ is (Type of project Organization, Relationship of the organizational units in the project effort, Preparation and commitment to project status reporting) = (M, L, M). For this tuple, two selected RD corresponding to row 1 and raw 12 of decision table in Figure 5 are selected and two MA corresponding to row 2 and 13 are suggested.

5. Empirical investigation

The proposed framework has been investigated through two different types of empirical investigations: post-mortem analysis and case study.
Post-mortem analysis can be defined as "a series of steps aimed at examining the lessons to be learnt from products, processes and resources to benefit on-going and future projects. Post-mortems enable individual learning to be converted into team and organizational learning" (Myllyaho et al., 2004).
Case studies (Yin, 2003; Kitchenham et al., 1995), instead, are investigations on real projects being carried out in an industrial setting. Consequently, all variables are defined a priori, but the level of control is low. These are strongly influenced by the context of the enterprise providing the experimental environment. Also, the independent variables of the study may change due to management decisions or as a consequence to a natural evolution of the process variables considered during project execution. Generally, a case study is carried out to investigate a phenomenon within a specific range of time. A case study can be used as a means to evaluate the efficiency of a possible innovation or as a comparative study which evaluates and compares results deriving from the application of an innovative method, technique or tool and the one already in use within the enterprise.
Both post-mortem analysis and case study were executed on industrial project data of a large software firm. The goal of this firm is to embed the risk assessment/treatment in its primary processes in order to support its project execution by the experience acquired in former projects. Therefore FIRM, jointly with the Department of Informatics of Bari, has introduced the approach for highlighting and managing the risks occurred.
To execute post mortem analysis, also called simulation, 54 projects of FIRM have been analyzed, all executed in a period of seven years, and seven of them, considered homogeneous in terms of duration, project size and development team experience, were selected.
Furthermore to execute the case study, 5 projects have been analyzed in-itinere in order to directly evaluate the appropriateness of the proposed framework.
Both investigations aim at evaluating the proposed approach with respect to the same factors, in the same context and with the same viewpoint. For these reasons the experiment definition and the metric model adopted, explained in the following, are the same.

5.1 Experiment definition

The aims of the empirical investigation are to verify whether risk management resulting from the application of Proposed Approach (PA) is more efficient and precise than risk management carried out using traditional Management Support (MS), i.e. the traditional risk management.

Effectiveness means the ability to undertake mitigation actions that, for each expected risk, avoid that a risk degenerates in one or more problems. While Precision is the ability to foresee all the occurred risks.
Research goals are thus formalized as follow:

RG1. *Analyze* the proposed approach *for the purpose of* comparing it to risk management obtained by only using management support with respect to Effectiveness *from viewpoint of* FIRM risk manager *in the context of* industrial FIRM projects	***RG2***. *Analyze* the proposed approach *for the purpose of* comparing it to risk management obtained by only using management support with respect to Precision *from viewpoint of* FIRM risk manager *in the context of* industrial FIRM projects

The consequent research hypotheses to test were:

- H_0Effectiveness: there is no statistically significant difference in effectiveness between PA and MS.
- H_1Effectiveness: there is statistically significant difference in effectiveness between PA and MS.
- H_0Precision: there is no statistically significant difference in precision between PA and MS.
- H_1Precision: there is statistically significant difference in precision between PA and MS.

Independent variables represent the two treatments: risk management using proposed approach (PA) and risk management using only management support (MS).
Dependent variables were quality characteristics of research goals, i.e. effectiveness and precision. Both these variables were operatively quantified by using the metrics presented in the next paragraph.

5.2 Metric model

The following metrics were used to quantitatively asses the research goals:

$$Effectiveness = \left(1 - \frac{NOP}{NMR}\right) * 100$$

$$Effectiveness\,Gain = \left(\frac{Effectiveness\,of\,PA}{Effectiveness\,of\,MS} - 1\right) * 100$$

$$Precision = \left(\frac{NER}{NER + NUR}\right) * 100$$

$$Precision\,Gain = \left(\frac{Precision\,of\,PA}{Precision\,of\,MS} - 1\right) * 100$$

Where:

- Number of Expected Risk (NER): number of Expected Risks during project execution taken into account by project manager.
- Number of Unexpected Risk (NUR): number of occurred risks that are not foreseen (Unexpected Risk).
- Number of Managed Risk (NMR): number of expected risks managed by using a specific strategy.

- Number of Occurred Problems (NOP): number of problems (OP) raised during project execution because of degeneration of an expected risk badly managed. Each OP identifies a single occurred problem or a set of them. For these reasons NMR >= NOP

Figure 7 shows relationships among metrics in terms of sets.

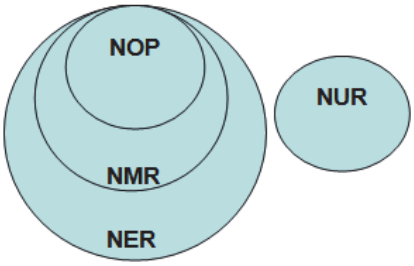


Fig. 7. Relationships between metrics

Note that Effectiveness can be equal to zero or, at maximum, equal to 100%. When Effectiveness is:

- Tends to 100% all the Expected Risks are well managed, in particular when NOP tends to zero;
- Tends to 0% when no one of the Expected Risk is well managed. In particular when NOP tends to NMR.

Therefore Effectiveness means the capability to manage the risks and to put to use the related mitigation actions in the way to avoid they became problems during the project execution. For this reason Effectiveness is as grater as smaller is the NER that became problems.

Precision can tend to zero or, at maximum tend to 100%. When Precision:

- Tends to 100%, when all the possible risks were detected, in particular when UR tends to 0.
- Tends to 0% at the NUR increasing, in particular it means that number of UR is much greater than NER.

In fact Precision means the capability to foresee all the risks that can occur during project execution NUR decreases.

At the beginning of each project and iteratively during project execution, a manager points out a set of Expected Risks. Part of this set, composed by the most critical and relevant risks for the project, will be managed, while the remaining ones, will not. In general terms, a risk is managed when a risk response plan is developed for it.

Action strategy defined by the manager for each MR in some cases is successful and in other cases transforms a risk into an OP. The last case is indicative of ineffectiveness of the strategy action adopted in the project execution. Finally it is also possible that some problems (UP), raised during project execution, are related to UR.

6. Data analysis

Proposed approach was validated using “Post Mortem Analysis” and “Case Study”. Both in Post Mortem Analysis and in Case Study according to the experimental design, statistical analysis were carried out. First of all descriptive statistics were used to interpret data

graphically. Data collected during experimentation have been synthesized through descriptive statistics. Finally, data have been analysed through hypothesis testing, where initial hypothesis were statistically validated with respect to a significance level. The dependent variables were tested in order to investigate the significance of the differences observed in the values collected.
In next paragraphs the results of data analysis are given. The first paragraph (6.1) refers to post mortem analysis and the second one (6.2) to the case study

6.1 Post mortem analysis

This investigation aims at evaluating the PA effectiveness and precision by observing the model behaviour used on legacy data related to projects already executed with the traditional approach.
Data set includes 14 observations, 2 for each project.

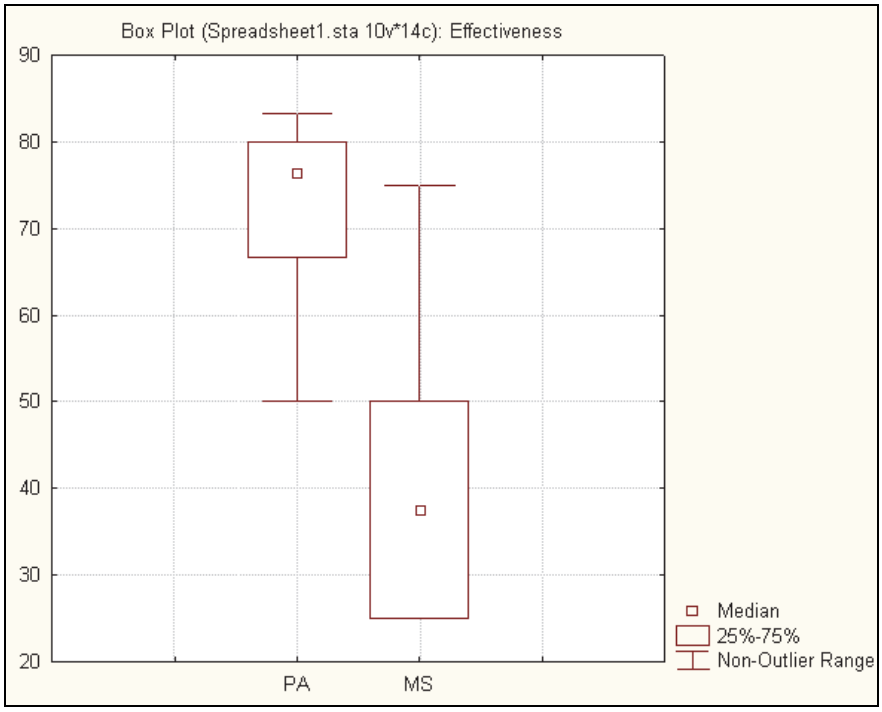


Fig. 8. Box plot for effectiveness (median)

Figure 8 shows the box plot related to the effectiveness of MS and PA. As it can be seen there is a greater effectiveness of PA than MS in terms of median value, 76.39% against 37.50%, and of lower and upper quartiles values , [66.67%, 80.00%] for PA and [25.00%, 50.00%] for MS.
The result of the descriptive analysis is statistically significant according to the Wilcoxon test. Wilcoxon test (Wilcoxon, 1945) is the nonparametric alternative to t-test for dependent samples. Since normality conditions were not always met, non parametric test was chosen. We used Shapiro-Wilk W test to verify if normality conditions were always satisfied or not.

Experimental Group	p-level	Results
Effectiveness	0.0210	reject $H_{0Effectiveness}$ and accept $H_{1Effectiveness}$

Table 1. P-level value of the Wilcoxon test for Effectiveness value

The test points out a significant difference in the Effectiveness between the two approaches. Therefore the null hypothesis can be rejected and we can conclude that the proposed approach is more efficient than traditional risk management.
Figure 9 shows the median values of precision of PA and MS. As it can be seen there is a greater precision of PA than MS in terms median value, 71.43% against 50.00%, and of lower and upper quartiles values, [50.00%, 75.00%] for PA and [33.33%, 66.67%] for MS.

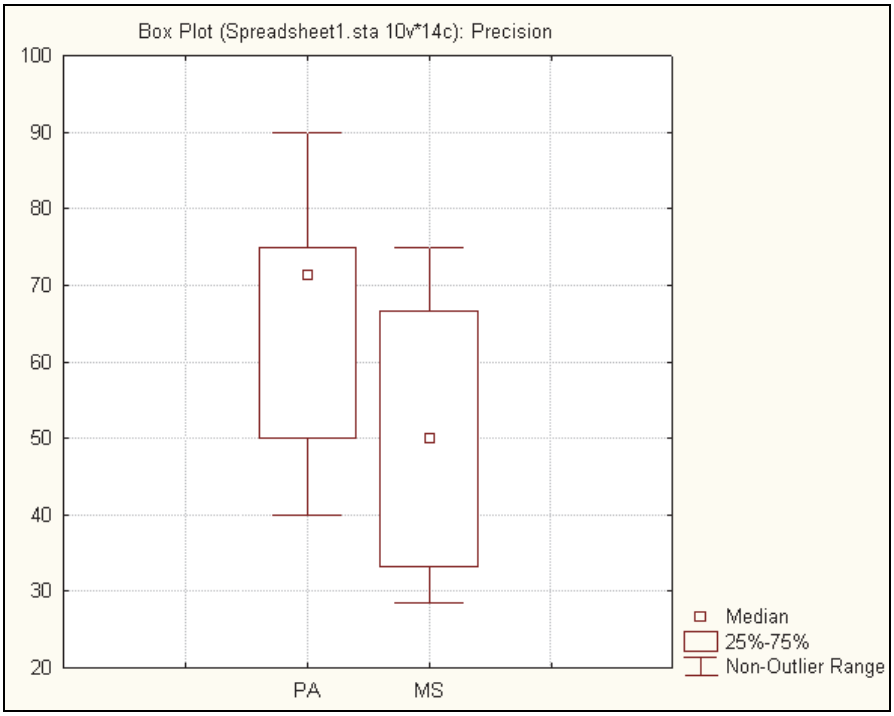


Fig. 9. Box plot for precision (median)

Also in this case Shapiro-Wilk W test was used to test normality. Since observed values were not normally distributed, also in this case, the Wilcoxon test was used.
Table 2 reports the values of the p-level obtained by using Wilcoxon test, applied to Precision of the two approaches. The test points out a significant difference between the two approaches. Therefore the null hypothesis can be rejected and we can conclude that the proposed approach is more precise in risk management.

Experimental Group	p-level	Results
Precision	0.0253	reject $H_{0Precision}$ and accept $H_{1Precision}$

Table 2. P-level value of the Wilcoxon test for Precision value

6.2 Case study data analysis

This kind of investigation evaluates PA effectiveness and precision, compared with MS, measuring it "on the field" during the execution of some processes. For this purpose, 5 projects that conducted with the both approaches were selected. As for the post mortem analysis, also in this case, the collected values appeared as not be normally distributed and thus the Wilcoxon non parametric test was used for the hypotheses testing the α-value was fixed at 5%.

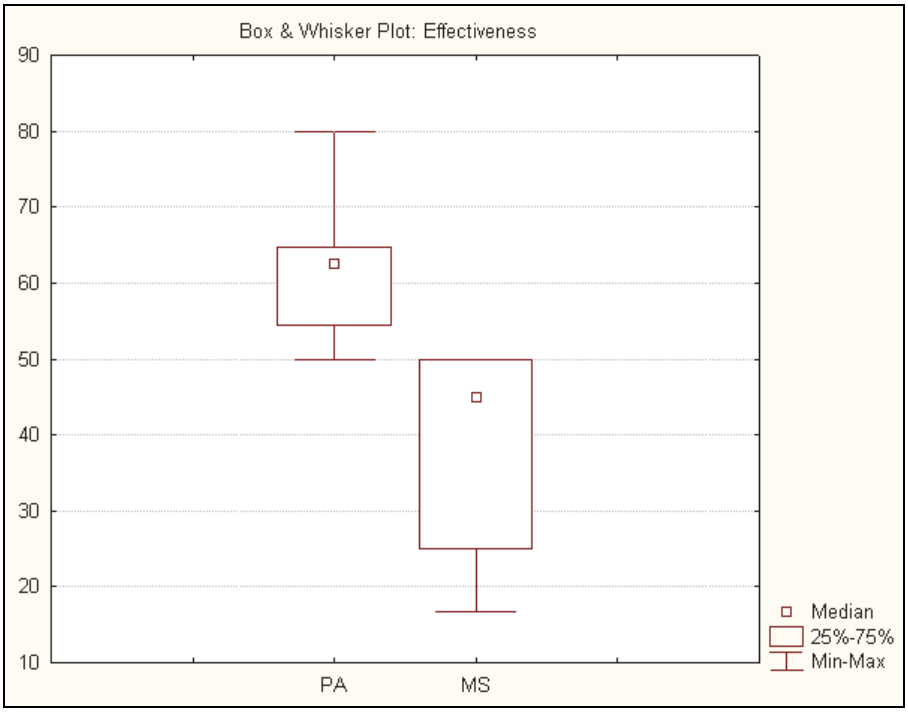


Fig. 10. Box plot for effectiveness (median)

Figure 10 shows the box plot related to the effectiveness of MS and PA. As it can be seen there is a greater effectiveness of PA than MS in terms of median value, 62.50% against 45.00%, and of lower and upper quartiles values, [54.54%, 64.70%] for PA and [25.00%, 50.00%] for MS. Regarding the distribution, data seem to confirm what was seen in post mortem analysis.

Table 3 reports the values of the p-level resulted from the Wilcoxon test applied to the Effectiveness of MS and PA.

Experimental Group	p-level	Results
Effectiveness	0.0009	reject $H_{0Effectiveness}$ and accept $H_{1Effectiveness}$

Table 3. P-level value of the Wilcoxon test for Effectiveness value.

The test points out a significant difference in the Effectiveness between the two approaches. Therefore the null hypothesis can be rejected and we can conclude that the proposed approach is more efficient than the manager approach. The Case Study allows to reject the null hypothesis with the error probability lower than in the case of the post-mortem analysis.
Figure 11 shows the median values of precision of PA and MS. As it can be seen there is a greater precision of PA than MS in terms median value, 71.57% against 50.00%, and of lower and upper quartiles values, [50.00%, 80.00%] for PA and [50.00%, 60.00%] for MS.

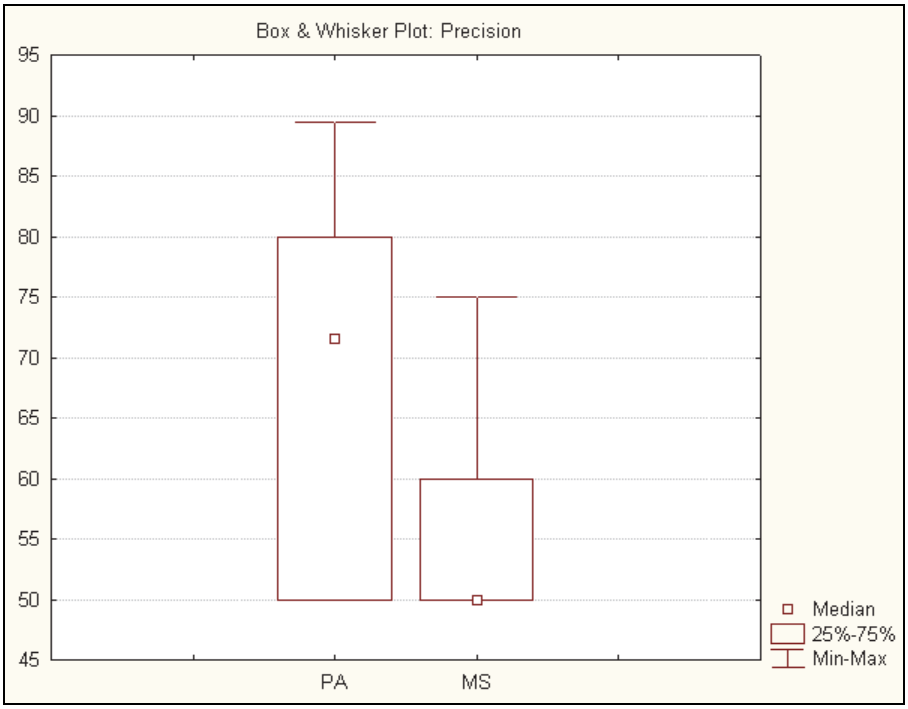


Fig. 11. Box plot for precision (median)

Table 4 reports the values of the p-level obtained by using the Wilcoxon test, applied to Precision of the two approaches. There is, also in this case, a statistically significant difference in the Precision between the two approaches, i.e. the null hypothesis can be rejected concluding that the proposed approach is more precise than the manager approach. Also in this case, the test points out a significant difference between the two approaches. Therefore the null hypothesis can be rejected and we can conclude that the proposed approach is more precise in risk management.

Experimental Group	p-level	Results
Precision	0.005	reject $H_{0Precision}$ and accept $H_{1Precision}$

Table 4. P-level value of the Wilcoxon test for Precision value

6.3 Lessons learnt

An additional experimentation data analysis allowed us to make some general qualitative considerations completing the comparison between the PA and the MS.

To make these analyses we consider the issues areas that were listed in the FLQ (Figure 3):

- Financial Management
- Contract Management
- Schedule Management
- Resource Management
- Quality Management
- Communication Management
- Scope Management
- Customer Relationship

We decided to consider the FLQ issues areas because we valuate this detail level the better one on the base of the number of collected data.

According to the Post Mortem data, the critical areas (the areas that were characterized by the higher number of problems) were: Resource Management, Quality Management, and Scope Management.

Resource Management consists of human resources and infrastructure management. Infrastructure management requires the identification and acquisition of all necessary equipment capable to carry out the project.

Quality Management consists of planning, constructing and evaluating product and service quality. This function requires, in particular, planning and conducting of quality assurance reviews, or reviews aimed at evaluating the quality of the process.

Finally, Scope Management consists of Defining the product or service expected by the consumer (product scope) and the corresponding work necessary to achieve it (project scope); also monitoring changes during the project execution.

For the critical areas we observed that MS finds a lower NER than the PA. Moreover, while in MS only a few part of NER are managed, in PA all the NER are managed. Moreover, in the PA the NUR is lower than in MS, it could be a consequence of the better capacity of PA to find risks and to manage them. These observations could consequently motivate the quantitative Precision Post Mortem results.

According to the Post Mortem data, we found in the Case Study the same critical issues areas but in this case there was a decreasing of the PA criticality. This reduction could confirm that the approach based on the EF tends to improve the capacity to manage the risk in the critical areas towards the past experiences that were acquired in these areas. In fact the higher number of experiences, of data and of practices is usually related to the most critical areas. According to this consideration, we observed that the reduction of occurred problem in PA is consequence of the increase of the number of mitigation action efficacy.

7. Conclusions

This paper proposes a Knowledge based Risk Management Framework able to collect, formalize and reuse the knowledge acquired during past projects execution. As instrument for supporting such methodology, an appropriate set of assessment questionnaires and decision-tables have been proposed. The innovative use of decision tables allowed to capture risk knowledge during the entire project lifecycle and to improve the knowledge collected in the Knowledge Base.
Thanks to knowledge expressed through decision tables, the proposed approach allows to combine the results of each document for evaluating the effects and the possible mitigation actions. In other words it allows express:
The relations between generic and specific issues;
The relations between issues, risks and actions to undertake to mitigate the risks as they occur.
To evaluate the proposed approach the framework has been transferred and investigated in an industrial context through two different types of empirical investigations: post-mortem analysis and case study.
Research goals aimed at assessing whether the proposed approach was more effective and precise for supporting risk management in software processes compared to traditional risk management approaches for Management Support.
Data analysis pointed out a statistically significant difference between the proposed approach and the traditional one in software process risk management with respect to effectiveness and precision. Such empirical results confirm that better structured risk knowledge, customizable according to the context, helps a manager to achieve more accurate risk management. Moreover we observed that the proposed approach allowed, especially in the critical areas such as Resource Management, Quality Management, Scope Management, to obtain better results. Obviously, in order to generalize the validity of the proposed approach further studies extended to other contexts are needed. For this reason, the authors intend replicating the empirical investigations.

8. References

Arshad, N.H., Mohamed, A., & Mansor, R. (2009). Organizational structural strategies in risk management implementation: best practices and benefits, WSEAS Transactions on Information Science and Applications, Vol. 6, No. 6, June 2009

Basili, V.R. (1989). Software Development: A Paradigm for the Future, Proceedings of the Annual Computer Software and Applications Conference, Orlando, September 1989

Basili, V.R., Bomarius, F., & Feldmann. (2007). Get Your Experience Factory Ready for the Next Decade: Ten Years After Experience Factory: How to Build and Run One, Proceedings of IEEE Int. Conf. on Software Maintenance, Minneapolis, May 2007

Basili, V.R., Caldiera, G., & Rombach, H.D. (1994). The Experience Factory. Encyclopedia of Software Engineering, John Wiley & Sons, Inc., 1994

Batista Webster, K.P., de Oliveira, K.M., & Anquetil, N. (2005). A Risk Taxonomy Proposal for Software Maintenance, Proceedings of IEEE Int. Conf. on Software Maintenance, Budapest, September, pp. 2005

Boehm, B.W. (1989). Tutorial: Software Risk Management, IEEE Computer Society Press, New. York, 1989

Charette, R.N. (1990). Application Strategies for Risk Analysis, McGraw-Hill Book Company, ISBN 0-07-010888-9, 1990

Chatterjee, D., & Ramesh, V.C. (1999). Real Options for Risk Management in Information Technology Projects, Proceedings of Hawaii International Conference on System Sciences, Maui, January 1999

Costa, H.R., de O. Barros, M., & Travassos, G. H. (2007). Evaluating Software Project Portfolio Risks, Journal of Systems and Software, Vol. 80, No. 1, pp. 16-31, 2007

Croll, P.R., Chambers, C., Bowell, M., & Chung, P.W.H. (1997). Towards Safer Industrial Computer Control Systems, Proceedings of International Conference On Computer Safety, Reliability and Security, York, September, 1997

Dhlamini, J., Nhamu, I. & Kaihepa. (2009). Intelligent risk management tools for software development, Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association (SACLA '09), ACM, New York

Donaldson S. E. & Siegel, S. G. (2007). Enriching Your Project Planning: Tying Risk Assessment to Resource Estimation, IT Professional, Vol. 9, No 5, pp.20-27, 2007

Farias, L., Travassos, G.H., & Rocha, A.R. (2003). Managing Organizational Risk Knowledge, Journal of Universal Computer Science, Vol. 9, No. 7, 2003

Gemmer, A. (1997). Risk Management: Moving Beyond Process, IEEE Computer, Vol. 30, No. 5, pp. 33-43 , 1997

Gilliam, P.D. (2004). Security Risks: Management and Mitigation in the Software Life Cycle, Proceedings of IEEE Int. Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, Modena, June 2004

Hefner, R. (1994). Experience with Applying SEI's Risk Taxonomy, Proceedings of Conference on Software Risk Management, Pittsburgh, 1994

Ho, T.B., Cheung, D., & Liu, H. (2005). Advances in Knowledge Discovery and Data Mining, LNCS Springer, Heidelberg

Kànel V. J., Cope, E. W., Deleris, L. A., Nayak, N. & Torok, R. G. (2010). Three key enablers to successful enterprise risk management. IBM Journal of Research and Development, Vol. 54, No. 3, May 2010

Kitchenham, B., Pickard, & L., Pfleeger, S.L. (1995), Case Studies for Method and Tool Evaluation, IEEE Software, Vol. 12, No 4, pp. 52-62

Kontio, J. (2001). Software Engineering Risk Management: A Method, Improvement Framework and Empirical Evaluation, R&D-Ware Technical Report, 2001

Loon, H.V. (2007). A Management Methodology to Reduce Risk and Improve Quality, IT Professional, Vol. 9, No 6, pp.30-35, 2007, ISBN 1520-9202

Maes, R.J., & Van Dijk, E.M. (1998). On the Role of Ambiguity and Incompleteness in the Design of Decision Tables and Rule-Based Systems, The Computer Journal, Vol. 31, No. 6, pp. 481-489

Myllyaho, M., Salo, O., Kääriäinen, J., Hyysalo, J., & Koskela, J. (2004) . A Review of Small and Large Post-Mortem Analysis Methods, Proceedings of International Conference on Software and Systems Engineering and Their Applications, Paris, 2004

Petroski H. (1994). Design Paradigms: Case Histories of Error and Judgment in Engineering, Cambridge University Press, ISBN 0-521-46649-0

Schneider K., & Hunnius, J.V. (2003). Effective Experience Repositories for Software Engineering, Proceedings of International Conference on Software Engineering, Portland, May 2003

Stratton, A., Holcombe, M., and Croll, P.R. (1998). Improving the Quality of Software Engineering Courses through University based Industrial Projects, Proceedings of International Workshop on the Projects in the Computing Curriculum, Sheffield, 1998

Vanthienen, J., Mues, C., Wets, G., & Delaere, K. (1998). A Tool Supported Approach to Inter-Tabular Verification, Expert Systems with Applications, Vol. 15, No. 3-4, pp. 277-285

Wilcoxon, F.(1945). Individual Comparisons by Ranking Methods, Biometrics Bulletin, Vol. 1, No 6, pp. 80-83

Yin, R.K.(2003), Case Studies Research Design and Methods, Sage Publications, ISBN 0-7619-2552-X

6

Portfolio Risk Management: Market Neutrality, Catastrophic Risk, and Fundamental Strength

N.C.P. Edirisinghe[1] and X. Zhang[2]
[1]*College of Business, University of Tennessee, Knoxville*
[2]*College of Business, Austin Peay State University, Clarksville*
U.S.A.

1. Introduction

Design of investment portfolios is the most important activity in the management of mutual funds, retirement and pension funds, bank and insurance portfolio management. Such problems involve, first, choosing individual firms, industries, or industry groups that are expected to display strong performance in a competitive market, thus, leading to successful investments in the future; second, it also requires a decision analysis of how best to periodically rebalance such funds to account for evolving general and firm-specific conditions. It is the success of both these functions that allows a portfolio manager to maintain the risk-level of the fund within acceptable limits, as specified by regulatory and other policy and risk considerations. This chapter presents a methodology to deal with the above two issues encountered in the management of investment funds.

While there is an abundance of literature on portfolio risk management, only a few investment managers implement disciplined, professional risk management strategies. During the stock market bubble of the late 90s, limiting risk was an afterthought, but given the increased stock market volatilities of the last decade or so, more managers are resorting to sophisticated quantitative approaches to portfolio risk management. Active risk management requires considering long-term risks due to firm fundamentals as well as short-term risks due to market correlations and dynamic evolution. The literature related to the former aspect often deals with discounted cash flow (DCF) models, while the latter topic is mainly dealt within a more quantitative and rigorous risk optimization framework. In this chapter, we propose new approaches for these long- and short-term problems that are quite different from the traditional methodology.

The short-term portfolio asset allocation (or weight determination) is typically optimized using a static mean-variance framework, following the early work on portfolio optimization by (Markowitz, 1952), where a quadratic programming model for trading off portfolio expected return with portfolio variance was proposed. Variants of this approach that utilize a mean absolute deviation (MAD) functional, rather than portfolio variance, have been proposed, see for instance, (Konno & Yamazaki, 1991). Asset allocation is the practice of dividing resources among different categories such as stocks, bonds, mutual funds, investment partnerships, real estate, cash equivalents and private equity. Such models are expected to lessen risk exposure since each asset class has a different correlation to the

others. Furthermore, with passage of time, such correlations and general market conditions do change, and thus, optimal portfolios so-determined need to be temporally-rebalanced in order to manage portfolio risks consistent with original specifications, or variations thereof due to changes in risk preferences. Consequently, a more dynamic and multistage (rather than a static single stage) treatment of the risk optimization problem must be employed, see (Edirisinghe, 2007). For multi period extensions of the mean-variance risk framework, see Gulpinar et al. (2003), where terminal period mean-variance trade off is sought under proportional transaction costs of trading portfolio management with transaction costs, under a discrete event (scenario) tree of asset returns. Also, see Gulpinar et al. (2004) where tax implications are considered within a multi period mean-variance analysis.

Mean-variance optimal portfolios are shown to be (stochastically) dominated by carefully constructed portfolios. Consequently, general utility functions (rather than quadratic) have been proposed as an alternative to mean-variance trade-off, where the expected utility of wealth is maximized. The first formal axiomatic treatment of utility was given by von Neumann & Morgenstern (1991). Other objective functions are possible, such as the one proposed by Zhao & Zeimba (2001). The relative merits of using Markowitz mean-variance type models and those that trade off mean with downside semi-deviation are examined in Ogryczak & Ruszczynski (1999). The semi-deviation risk trade-off approach yields superior portfolios that are efficient with respect to the standard stochastic dominance rules, see Whitmore & Findlay (1978). When quadratic penalty is applied on the downside deviations, with target defined at the portfolio mean, it is called the downside semi-variance risk metric. Semi-variance fails to satisfy the positive homogeneity property required for a coherent risk measure.

The concept of coherent risk measures was first introduced by Artzner et al. (1999). This landmark paper initiated a wealth of literature to follow on coherent risk measures with several interesting extensions, see Jarrow (2002) for instance. Coherent risk measures scale linearly if the underlying uncertainty is changed, and due to this linearity, coherency alone does not lead to risk measures that are useful in applications. As discussed in Purnanandam et al. (2006), one important limitation of coherent risk measures is its inability to yield sufficient diversification to reduce portfolio risk. Alternatively, they propose a methodology that defines risk on the domain of portfolio holdings and utilize quadratic programming to measure portfolio risk.

Another popular method of risk measurement is to use the conditional value-at-risk (CVaR), see, e.g. Rockafellar and Uryasev (2000) and Ogryczak and Ruszczynski (2002). Risk measures based on mean and CVaR are coherent, see Rockafellar et al. (2002). Such risk measures evaluate portfolio risk according to its value in the worst possible scenario or under the probability measure that produces the largest negative outcome. Consequently, to alleviate the computational burden associated with computing the risk metric, a discrete sample of asset returns vector must be available. This is also the case when computing semi-deviation risk models or convex risk measures, an approach proposed by Follmer & Schied (2002) as a generalization of coherent risk measures. In situations where only partial information on probability space is available, Zhu & Fukushima (2009) proposed a minimization model of the worst-case CVaR under mixture distribution uncertainty, box uncertainty, and ellipsoidal uncertainty. Chen & Wang (2008) presented a new class of two-sided coherent risk measures that is different from existing coherent risk measures, where both positive and negative

deviations from the expected return are considered in the new measure simultaneously. This method allows for possible asymmetries and fat-tail characteristics of the loss distributions.
While the convex and sub-additive risk measure CVaR has received considerable attention, see Pirvu (2007) and Kaut et al. (2007), the availability of a discrete return sample is essential when using (efficient) linear programming-based methods to evaluate the risk metric. In practice, however, estimating mean and covariance parameters of asset return distributions in itself is a daunting task, let alone determining specific return distributions to draw return samples for risk metric evaluation. If one strives to eliminate possible sampling biases in such a case, a sufficiently large sample must be drawn for computing CVaR. Such a practice would lead to enormous computational difficulties, in particular under a multi period investment setting. Hence, practical risk metrics that are computable based on distribution parameters, rather than a distribution assumption itself, have tremendous implications so long as such metrics are able to achieve risk-return characteristics consistent with investors' attitudes. In this way, more effort can be focused on estimating parameters more accurately, especially given that parameters evolve dynamically. Markowitz's mean-variance framework has the latter computational advantage, however, it often fails to exhibit sufficient risk control (out-of-sample) when implemented within a dynamically rebalanced portfolio environment, see Section 3.
In the sequel, we view portfolio risk in a multi-criterion framework, in the presence of market frictions such as transactions costs of trading and lot-size restrictions. In this context, we propose two modifications on the standard mean-variance portfolio model so that the modified mean-variance model considers more comprehensive portfolio risks in investment decision making. The first modification is "catastrophic risk" (CR), which is used to control portfolio risk due to over-investment in a stock whose volatility is excessive. This form of risk is concerned with the direction of (the future) price of a security being opposite to the sign of the established position in the portfolio. That is, securities in a long portfolio fall in price while the securities in a short portfolio rise in price. Such risk is often the result of error in forecasting the direction of stock price movement. Controlling the portfolio variance does not necessarily counter the effects of catastrophic risk.
The second modification is the concept of "market neutrality", which is used to maintain the portfolio exposure to market risk within specified bands. Portfolio beta is an important metric of portfolio bias relative to the broader market. A balanced investment such that portfolio beta is zero is considered a perfectly beta neutral portfolio and such a strategy is uncorrelated with broader market returns.
The short-term portfolio risk control using the above levers is complemented by a long-term risk mitigation approach based on fundamental analysis-based asset selection. Firm selection based on investment-worthiness is the study often referred to as Fundamental Analysis, which involves subjecting a firm's financial statements to detailed investigation to predict future stock price performance. The dividend discount model, the free cash flow to equity model, and the residual income valuation model from the accounting literature are the standard methods used for this purpose. These DCF models estimate the intrinsic value of firms in an attempt to determine firms whose stocks return true values that exceed their current market values. However, DCF models typically require forecasts of future cash flow and growth rates, which are often prone to error, as well as there is no formal objective mechanism to incorporate influence on firm performance from other firms due to supply and demand competitive forces. We contend that such an absolute intrinsic value of a firm is likely to be a weak metric due

to the absence of relative firm efficiencies. As a remedy, we implement an approach that combines fundamental financial data and the so-called Data Envelopment Analysis (DEA) to determine a metric of relative fundamental (business) strength for a firm that reflects the firm's managerial efficiency in the presence of competing firms. Under this metric, firms can then be discriminated for the purpose of identifying stocks for possible long and short investment.
The chapter is organized as follows. In Section 2, we start with the short-term risk optimization model based on mean-variance optimization supplemented with market dependence risk and catastrophic risk control. Section 3 illustrates how these additional risk metrics improve the standard mean-variance analysis in out-of-sample portfolio performance, using a case study of U.S. market sector investments. Section 4 presents long-term asset selection problem within the context of fundamental analysis. We present the DEA-based stock screening model based on financial statement data. The preceding methodologies are then tested in Section 5 using the Standard and Poors 500 index firms covering the nine major sectors of the U.S. stock market. Using the integrated firm selection model and the risk optimization model, the resulting portfolios are shown to possess better risk profiles in out-of-sample experiments with respect to performance measures such as Sharpe ratio and reward-to-drawdown ratio. Concluding remarks are in Section 6. The required notation is introduced as it becomes necessary.

2. Short-term risk optimization

Portfolio risk management is a broad concept involving various perspectives and it is closely tied with the ability to describe future uncertainty of asset returns. Consequently, risk control becomes a procedure for appropriately shaping the portfolio return distribution (derived according to the return uncertainty) so as to achieve portfolio characteristics consistent with the investors' preferences. The focus here is to specify sufficient degree of control using risk metrics that are efficiently computable under distributional parameters (rather than specific return samples). That is, such risk metrics do not require a distributional assumption or a specific random sample from the distribution, but risk control can be specified through closed-form expressions. One such example is the portfolio variance as considered in the usual mean-variance analysis.
Consider a universe of N (risky) assets at the beginning of an investment period, such as a week or a month, for instance. The investor's initial position (i.e., the number of shares in each asset) is x^0 ($\in \Re^N$) and the initial cash position is C^0. The (market) price of asset j at the current investment epoch is $\$P_j$ per share. At the end of the investment horizon, the rate of return vector is r, which indeed is a random N-vector conditioned upon a particular history of market evolution. Thus, price of security j changes during the investment period to $(1 + r_j)P_j$. Note that $r_j \geq -1$ since the asset prices are nonnegative. Moreover, r is observed only at the end of the investment period; however, trade decisions must be made at the beginning of the period, i.e., revision of portfolio positions from x^0 to x. Then, $x_j - x_j^0$ is the amount of shares purchased if it is positive; and if it is negative, it is the amount of shares sold in asset j. This trade vector is denoted by y and it equals $|x - x^0|$, where $|.|$ indicates the absolute value.
Risk optimization problem is concerned with determining the (portfolio rebalancing) trade vector y such that various risk specifications for the portfolio are met whilst maximizing the portfolio total expected return. The trade vector y is typically *integral*, or in some cases, each y_j must be a multiple of a certain lot size, say L_j. That is, $y_j = kL_j$ where $k = 0, 1, 2, \ldots$.

Furthermore, portfolio rebalancing is generally not *costless*. Usually, portfolio managers face transactions costs in executing the trade vector, y, which leads to reducing the portfolio net return. Placing a trade with a broker for execution entails a direct cost per share traded, as well as a fixed cost independent of the trade size. In addition, there is also a significant cost due to the size of the trading volume y, as well as the broker's ability to place the trading volume on the market. If a significant volume of shares is traded (relative to the market daily traded volume in the security), then the trade execution price may be adversely affected. A large buy order usually lead to trade execution at a price higher than intended and a large sell order leads to an average execution price that is lower than desired. This dilution of the profits of the trade is termed the *market impact loss*, or slippage. This slippage loss generally depends on the price at which the trade is desired, trade size relative to the market daily volume in the security, and other company specifics such as market capitalization, and the beta of the security. See Loeb (1983) and Torre and Ferrari (1999), for instance, for a discussion on market impact costs.

Our trading cost model has two parts: proportional transactions costs and market impact costs. The former cost per unit of trade in asset j is α_{0j}. The latter cost is expressed per unit of trade and it depends directly on the intended execution price as well as the fraction of market daily volume of the asset that is being traded in the portfolio. Denoting the expected daily (market) volume in asset j by V_j shares, and α_{1j} being the constant of proportionality, the market impact cost per unit of trade is

$$\alpha_{1j} P_j \frac{y_j}{V_j}.$$

The constants α_{0j} and α_{1j} are calibrated to the market data. Ignoring the fixed costs of trading, the total transactions and slippage loss function $f_j(y_j)$ is

$$f_j(y_j) := y_j \left(\alpha_{0j} + \alpha_{1j} \frac{P_j y_j}{V_j} \right). \tag{1}$$

Therefore, the (total) loss function in portfolio rebalancing is $F(y) := \sum_{j=1}^{N} f_j(y_j)$. Denoting the cash position accumulated during rebalancing by C, the portfolio wealth satisfies the (self-financing) budget constraint:

$$\sum_{j=1}^{N} P_j (x_j - x_j^0) + F(y) + C = C^0. \tag{2}$$

If the riskfree rate of return for the investment period is κ, the portfolio total gain is given by the random variable,

$$G := \sum_{j=1}^{N} P_j r_j x_j + \kappa C - F(y). \tag{3}$$

The problem of portfolio risk control requires shaping the distribution of this random variable G using an appropriate choice of x. The standard mean-variance (MV) framework requires maximizing the expected portfolio gain, $E[G]$, for an acceptable level of variance risk of the portfolio, $Var[G]$.

2.1 Risk of market dependence

While the MV framework strives to control portfolio's intrinsic variance due to asset correlations with themselves, it fails to capture asset correlations with the broader market. As often is the case, even if portfolio variance is not excessive, by virtue of a strong dependence with the overall market, the portfolio may become overly sensitive (positively or negatively) to market 'moves', especially during market 'down times'. Therefore, it is imperative that the portfolio is rebalanced during certain periods to control this risk of market dependence. A portfolio is said to be perfectly market neutral if the portfolio is uncorrelated with the broader market. Portfolio neutrality is provided by hedging strategies that balance investments among carefully chosen long and short positions. Fund managers use such strategies to buffer the portfolio from severe market swings, for instance, see Nicholas (2000) and Jacobs and Levy (2004).

A prescribed level of imbalance or non-neutrality may be specified in order for the portfolio to maintain a given bias with respect to the market. An important metric of portfolio bias relative to the broader market is the *portfolio beta*. A portfolio strategy is uncorrelated with market return when the portfolio beta is zero, i.e., perfectly beta neutral portfolio. A stock with a beta of 1 moves historically in sync with the market, while a stock with a higher beta tends to be more volatile than the market and a stock with a lower beta can be expected to rise and fall more slowly than the market.

The degree of market-neutrality of the portfolio measures the level of correlation of performance of the portfolio with an underlying broad-market index. Typically, the S&P500 index may be used as the market barometer. Let β_j be the beta of asset j over the investment period. Then, β_j is the covariance of the rates of return between asset j and the chosen market barometer (index), scaled by the variance of the market rate of return. Since r_j is the random variable representing the rate of return of asset j, by denoting the market index rate of return by the random variable R, it follows that

$$\beta_j := \frac{Cov(r_j, R)}{Var(R)}. \tag{4}$$

Proposition 2.1. *Let the portfolio value at the beginning of the investment period be w^0. The portfolio beta, $B(x)$, at the end of the period (after rebalancing) is*

$$B(x) = \frac{1}{w^0}\left(\sum_{j=1}^{N} \beta_j P_j x_j\right). \tag{5}$$

Proof. The portfolio value at the end of the period is given by $w = w^0 + G$, where G is the portfolio gain in (3), and thus, the portfolio rate of return is the random variable $r_P := (w - w^0)/w^0$ and thus, $r_P = G/w^0$. Then,

$$Cov(r_P, R) = \frac{1}{w^0} Cov\,(G, R) = \frac{1}{w^0}\sum_j P_j x_j Cov(r_j, R), \tag{6}$$

since the riskfree rate κ is nonrandom (and thus it has zero correlation with the market). Thus, the result in the proposition follows. ■

To control the portfolio beta at a level $\gamma_0 \pm \gamma_1$, the constraints $\gamma_0 - \gamma_1 \leq B(x) \leq \gamma_0 + \gamma_1$ must be imposed, i.e.,

$$(\gamma_0 - \gamma_1)w^0 \;\leq\; \sum_{j=1}^{N} P_j \beta_j x_j \;\leq\; (\gamma_0 + \gamma_1)w^0, \tag{7}$$

where $w^0 = P'x^0 + C^0$. With $\gamma_1 \approx 0$, rebalancing strives for a portfolio beta of γ_0. In particular, for an almost market-neutral portfolio, one needs to set $\gamma_0 = 0$ and $\gamma_1 \approx 0$. Note that in order to control the risk of market dependence, the asset betas are required. Therefore, one needs to have accurate estimates of asset covariances with the market, as well as the volatility of market itself. Note that this risk expression is free of correlations among the assets themselves.

2.2 Catastrophic risk control

The second form of portfolio risk is concerned with the direction of (the future) price of a given asset being opposite to the sign of the established position in the portfolio. That is, assets in a *long* portfolio fall in price while the assets in a *short* portfolio rise in price. Such risk is often the result of error in forecasting the direction of stock price movement. This would entail observing a drop in price for long assets and an increase in price for shorted assets, a *catastrophic* event for the portfolio. Generally, there is no formal mechanism to safeguard against such events. Controlling the portfolio variance does not necessarily counter the effects of *catastrophic risk*, abbreviated herein as *Cat risk*, see the evidence presented in Section 3.
'Degree-θ Cat risk' is defined as the anticipated total dollar wealth loss, raised to power θ, in the event each stock price moves against its portfolio position by one standard deviation. Denoted by $\mathcal{C}_\theta(x)$, it is given by

$$\mathcal{C}_\theta(x) := \sum_{j=1}^{N} \left(P_j \sigma_j \left| x_j \right| \right)^\theta, \tag{8}$$

where $\theta \geq 1$ is a given constant. By controlling the nonnegative $\mathcal{C}_\theta(x)$ within a pre-specified upper bound, portfolio risk due to over investment in a stock whose volatility is excessive is managed. Note that the Cat risk expression is free of correlations among the assets.

Proposition 2.2. *The following properties hold for the Cat Risk metric.*

i. $\mathcal{C}_\theta(x)$ *is positively homogeneous of degree* θ *in* x.

ii. $\mathcal{C}_\theta(x)$ *is convex in* x *for fixed* $\theta \geq 1$.

iii. For $\theta = 1$, *Cat Risk is an upper bound on the portfolio standard deviation, i.e.,* $\mathcal{C}_1(x) \geq \sigma_P(x)$ *where* $\sigma_P(.)$ *is the portfolio standard deviation.*

Proof. For some $\lambda > 0$, $\mathcal{C}_\theta(\lambda x) = \sum_{j=1}^{N} \left(P_j \sigma_j \left| \lambda x_j \right| \right)^\theta = \lambda^\theta \mathcal{C}_\theta(x)$, thus proving the assertion (i). To show (ii), the first partial derivative of $\mathcal{C}_\theta(x)$ w.r.t. x_j is $\nabla_j \mathcal{C}_\theta(x) = a_j \theta P_j \sigma_j |x_j|^{\theta-1}$, where $a_j = +1$ if $x_j \geq 0$, $a_j = -1$ if $x_j < 0$. Then, the Hessian $\nabla_x^2 \mathcal{C}_\theta(x)$ is a diagonal matrix where the j^{th} diagonal element is $(a_j)^2 \theta(\theta-1) P_j \sigma_j |x_j|^{\theta-2}$, which is nonnegative if $\theta \geq 1$. Thus, $\nabla_x^2 \mathcal{C}_\theta(x)$ is positive semi-definite, implying that $\mathcal{C}_\theta(x)$ is convex in x for $\theta \geq 1$.
To show part (iii), define the random variable $\xi_j := P_j x_j r_j$. Then, the variance of the portfolio is $\sigma_P^2(x) = Var\left(\sum_{j=1}^{N} \xi_j \right)$. Denoting the standard deviation of ξ_j by $\hat{\sigma}_j$ and the correlation

between ξ_i and ξ_j by $\hat{\rho}_{ij}$,

$$\begin{aligned} Var\left(\sum_{j=1}^{N}\xi_j\right) &= \sum_{j=1}^{N}\hat{\sigma}_j^2 + 2\sum_{(i,j),i\neq j}\hat{\sigma}_i\hat{\sigma}_j\hat{\rho}_{ij} \\ &\leq \sum_{j=1}^{N}\hat{\sigma}_j^2 + 2\sum_{(i,j),i\neq j}\hat{\sigma}_i\hat{\sigma}_j \\ &= \left(\sum_{j=1}^{N}\hat{\sigma}_j\right)^2. \end{aligned}$$

Noting that $Var(\xi_j) = \left[P_j x_j \sigma_j\right]^2$, and since the prices are nonnegative, we have $\hat{\sigma}_j = P_j\sigma_j\,|x_j|$. Therefore, the portfolio variance is bounded from above by $[\mathcal{C}_1(x)]^2$. ■

Since portfolio standard deviation is only a lower bound on degree-1 Cat Risk - herein referred to as DOCR -, controlling portfolio variance via mean-variance optimization is not guaranteed to provide adequate protection against catastrophic risk. The two risk metrics, $\mathcal{C}_1(x)$ and $\sigma_P(x)$, however, have distinct characteristics in shaping portfolio positions, as will be demonstrated numerically in the next section. Geometrically, DOCR (as a function of portfolio positions) bounds the portfolio standard deviation by a polyhedral convex cone with apex at the origin.

For pre-specified level of (degree-θ) Cat risk, say cw^0 for some constant c, the following constraint is imposed when determining the rebalanced portfolio:

$$\mathcal{C}_\theta(x) \quad \leq \quad c\,w^0. \tag{9}$$

3. Performance under improved risk control

The focus here is to evaluate the risk control characteristics of the market dependence and catastrophic risk metrics, when applied under the usual mean-variance (Markowitz) framework of portfolio optimization. For this purpose, we consider a portfolio of Exchange-Traded Funds (ETFs) on the U.S. stock market. An ETF is a security that tracks an index, a commodity or a basket of assets like an index fund, but trades like a stock on an exchange. SPDR Trust, which is an ETF that holds all of the S&P 500 index stocks, is used as the market barometer in portfolio rebalancing. SPDR trades under the ticker symbol SPY. The S&P500 stocks belonging to SPY are categorized into nine market sectors, and accordingly, nine separate ETFs are created and traded in the market. These ETFs that track the sector-indices are known by their ticker symbols, as given by XLK (Technology), XLV (HealthCare), SLF (Financials), XLE (Energy), XLU (Utilities), XLY (Consumer Discretionary), XLP (Consumer Staples), XLB (Basic Materials), and XLI (Industrial Goods). In testing the preceding risk metrics, a portfolio of the nine ETFs is formed, whose positions are allowed to be positive or negative, thus, allowing for 'going long or short' in each ETF.

Consider the risk optimization model below for an investment period of one month, comprising the usual mean-variance trade off coupled with market dependence and cat risk constraints, where $\mu_j = E[r_j]$ and $\sigma_{jk} = cov(r_j, r_k)$, the covariance between asset returns. Note

the notation that $\sigma_{jj} = \sigma_j^2$, the variance of return r_j for the investment period.

$$\begin{aligned}
&\max_x \sum_{j=1}^{N} P_j \mu_j x_j - F(y) - \lambda \sum_{j,k=1}^{N} \sigma_{jk} x_j x_k && (10)\\
&\text{s.t. } P'(x - x^0) + F(y) + C = C^0 && \text{(budget)}\\
&\quad F(y) = \sum_{j=1}^{N} y_j \left(\alpha_{0j} + \alpha_{1j} \frac{P_j y_j}{V_j} \right) && \text{(trading costs)}\\
&\quad y = |x - x^0| \,,\ y = \nu L,\ \nu = 0, 1, 2, \ldots && \text{(trade vector)}\\
&\quad (\gamma_0 - \gamma_1) w^0 \ \le\ \sum_{j=1}^{N} P_j \beta_j x_j \ \le\ (\gamma_0 + \gamma_1) w^0 && \text{(market neutrality)}\\
&\quad \sum_{j=1}^{N} P_j \sigma_j \left| x_j \right| \ \le\ c\, w^0 && \text{(degree-1 Cat risk).}
\end{aligned}$$

In (10), a zero risk free rate ($\kappa = 0$) is assumed, trade lot size is L for any asset, and $\theta = 1$ is set for the Cat risk constraint. The portfolio variance risk is controlled by the aversion parameter $\lambda \geq 0$. Note that setting $\gamma_0 = 0$, $\gamma_1 = +\infty$, and $c = +\infty$ yield (10) as the usual mean-variance trade off model, in this case with trading frictions. The latter instance of the model is herein referred to as the MV model. Our computational illustrations compare and contrast the MV model with (10), referred to as the MVX (MV eXtended) model, for the ETF portfolio with $N = 9$ assets.

Computations of all statistical parameters, such as asset return means, standard deviations, asset covariances, and asset betas, use the historical data of the years 2003 and 2004. However, portfolio performance is evaluated in the *out-of-sample* investment horizon, Jan-Jun, 2005. During this horizon, a monthly-rebalancing strategy is applied where portfolio allocations are optimally adjusted at the beginning of each month. Under the monthly rebalancing strategy, parameter estimations are needed at the beginning of each month, conditional upon the data available prior to that point in time. This way, performance is assessed for each month in an out-of-sample style by simulating the dynamically evolving portfolio over the (actual) realized price series.

3.1 Performance of MV and MVX portfolios

The initial positions in all assets at the beginning of Jan 2005 are set to zero, trading cost parameters are $\alpha_{0j} = 2\%$ and $\alpha_{1j} = 1$ for each ETF, trade lot size $L = 50$ shares, initial wealth C^0 =1 million US$. The market barometer SPY index fund has an annualized volatility of roughly 10.55% during the first two quarters of 2005 with an annalized return (loss) of -1.22%. Portfolio comparison of models MV and MVX is based on the following performance metrics, accumulated over the six trading epochs:

1. ARoR (annualized rate of return): the portfolio daily average rate of return, net of trading costs, annualized over 250 days of trading.
2. AStD (annualized standard deviation): the standard deviation of the daily portfolio net rate of return series, annualized over 250 days of trading.
3. maxDD (portfolio maximum drawdown): Portfolio drawdown is defined as the relative equity loss from the highest peak to the lowest valley of a portfolio value decline within

a given time window, in this case from the beginning of January to the end of June, 2005, expressed as a percentage of portfolio value.

4. RTD (reward-to-drawdown) ratio: the ARoR, less the riskfree rate, divided by the maxDD.

The MV model is first applied over the 6-month horizon (with monthly-rebalancing) and its (out-of-sample) efficient frontier is plotted between portfolio ARoR and AStD, see Figure 1. On the same figure, the MVX model is plotted for two portfolio strategies: first, 50% market neutrality with 5% (of wealth) Cat risk, and second, full market neutrality with 5% (of wealth) Cat risk. Observe that the out-of-sample frontier produced by the pure MV model is improved by controlling the Cat risk at 5% with portfolio beta at 0.5, in particular for levels of sufficiently large variance risk. As the portfolio becomes perfectly market neutral (with zero portfolio beta), the MVX frontier improves drammatically as evident from Figure 1.

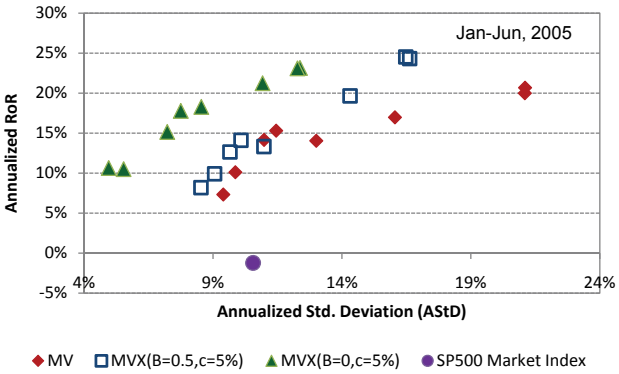


Fig. 1. Out-of-sample efficient frontiers on return vs volatility

An important portfolio performance characteristic is the so-called *drawdown*. Fund managers do not wish to see the value of a portfolio decline considerably over time. A drastic decline in portfolio value may lead to perceptions that the fund is *too risky*; it may even lead to losing important client accounts from the fund. For example, consider a portfolio with value \$5million at the beginning of a year. Suppose it reaches a peak in June to \$8million, and then loses its value to \$6million by the end of the year. Thus, for the period of one year, the fund had a maximum drawdown of $(8-6)/8$, or 25%, while the fund has an annual RoR of $(6-5)/5$, or 20%. Portfolio performance, as measured by the reward-to-drawdown (RTD) ratio, is $20/25=0.8$, whereas RTD of at least 2 is generally considered to be indicative of successful fund management. Figure 2 depicts the RTD ratio for MV and MVX funds, where pure mean-variance model has the weakest performance. But, with Cat risk control and perfectly market neutral portfolios, RTD of 2 or more is easily achievable for the concerned period of investment, while the general market (as evident from the S&P index fund) has logged a negative return with a maxDD of 7%.

The significant performance improvement of the MVX fund, relative to the MV fund or the general market, is due to two reasons: first, while the MV model controls portfolio variance risk, if the long/short asset positions are largely against the actual directions of (out-of-sample) asset returns due to error in the sign of mean forecast, then the portfolio is subject to increased Cat risk. This is evident from the comparison of MV and MVX portfolio

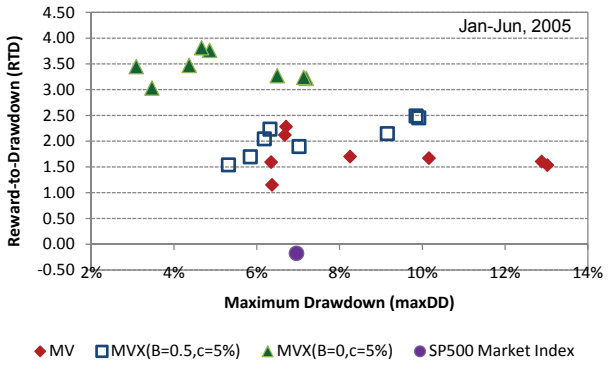


Fig. 2. Out-of-sample efficient frontiers on RTD vs drawdown

risks, see Figure 3. To highlight the effect at an increased level of volatility, the portfolios in Figure 3 are all chosen to have an annualized (out-of-sample) standard deviation of 14.5%. Observe that the maximum (monthly) Cat risk for MV model is about 50% larger than the portfolio maximum standard deviation. However, with Cat risk controlled at 5% (of wealth) in the MVX model, the out-of-sample Cat risk is substantially reduced, for practically no increase in portfolio variance risk.

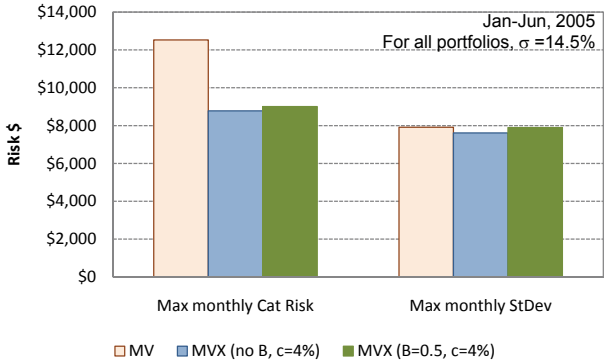


Fig. 3. Realized risks of managed portfolios

The second reason is the improved diversification achieved in MVX. As Figure 4 illustrates, with the presence of controlled market neutrality, long and short positions are created in a balanced manner to hedge the correlations of asset returns with the market.

4. Long-term risk control

The foregoing discussion on risk optimization focused on short term risk-return trade off in a multi-faceted risk framework. However, the asset universe for portfolio optimization was assumed given and no attention was paid to asset fundamentals when assessing the perceived risk of an asset on a longer term basis. In the case when assets are stocks of public firms, risk

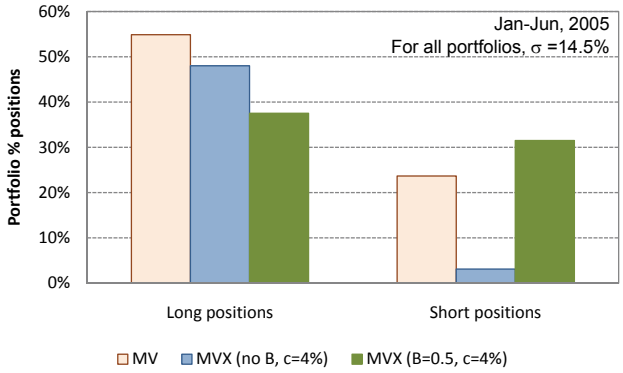


Fig. 4. Portfolio compositions under improved risk metrics

assessments based solely on technicals, such as historical stock prices, fail to account for risks due to firm operational inefficiencies, competition, and impacts from general macroeconomic conditions.

A firm's business strength is related to both its internal productivity as well as its competitive status within the industry it operates in. Internal productivity gains via, say, lower production costs, as well as the firm's effectiveness in dealing with supply competition and in marketing its products and services are reflected in the firm's balance sheets and cash flow statements. The analysis of financial statements of a firm for the purpose of long-term investment selection is referred to as Fundamental Analysis or Valuation (Thomsett, 1998) and it generally involves examining the firm's sales, earnings, growth potential, assets, debt, management, products, and competition. A large body of evidence demonstrates that fundamental financial data of a firm is related to returns in the stock market. For example, Hirschey (2003) concluded that in the long run, trends in stock prices mirror real changes in business prospects as measured by revenues, earnings, dividends, etc. Samaras et al. (2008) developed a multi-criteria decision support system to evaluate the Athens Stock Exchange stocks within a long term horizon based on accounting analysis.

In this section we apply a technique based on Data Envelopment Analysis (DEA) to evaluate and rank a firm's business strength, against other firms within a market sector, in an attempt to identify firms that are potentially less-risky (or more-risky) as long-term holdings. Consequently, the long-term risk exposure of an equity portfolio may be reduced.

4.1 Financial DEA model

Data Envelopment Analysis (DEA) is a nonparametric method for measuring the relative efficiencies of a set of similar decision making units (i.e., firms in a given sector) by relating their outputs to their inputs and categorizing the firms into managerially efficient and inefficient. It originated from the work by Farrell (1957), which was later popularized by Charnes et al. (1978). The CCR ratio model in the latter reference seeks to optimize the ratio of a linear combination of outputs to a linear combination of inputs. The CCR-DEA model necessarily implies a constant returns to scale (CRS) relationship between inputs and outputs. Thus, the resulting DEA score captures not only the productivity inefficiency of a firm at its actual scale size, but also any inefficiency due to its actual scale size being different from the

optimal scale size (Banker, 1984). In long-term risk control by firm selection, the objective is to screen companies within a given market segment based on their financial performance attributes, although these firms may be of different scale sizes. Hence, the CCR model is applied in this section for measuring the underlying (relative) fundamental financial strength of a given firm.

Let $\mathcal{I}$ and $\mathcal{O}$ denote disjoint sets (of indices) of input and output financial parameters, respectively, for computing a firm's efficiency. For a given time period (say, a quarter), the value of financial parameter P_i for firm j is denoted by ξ_{ij}, where $i \in \mathcal{I} \cup \mathcal{O}$, $j = 1, ..., N$, and N is the number of firms under investigation. Then, the DEA-based efficiency for firm k, η_k, relative to the remaining $N-1$ firms, is determined by the linear programming (LP) model,

$$\begin{aligned} \eta_k := \max_{u} \; & \sum_{i\in\mathcal{O}} \xi_{ik} u_i \\ \text{s.t.} \; & \sum_{i\in\mathcal{I}} \xi_{ik} u_i \le 1 \\ & -\sum_{i\in\mathcal{I}} \xi_{ij} u_i + \sum_{i\in\mathcal{O}} \xi_{ij} u_i \le 0, \;\; j = 1, \ldots, N \\ & u_i \ge 0, \; i \in \mathcal{I} \cup \mathcal{O}. \end{aligned} \tag{11}$$

The firm k's efficiency satisfies $0 \le \eta_k \le 1$, which is computed relative to the input/output values of the remaining $N-1$ firms. Since the DEA model is used to measure the (relative) financial strength of a firm, the type of input/output parameters needed are the fundamental financial metrics of firm performance. In general, such financial parameters span various operational perspectives, such as profitability, asset utilization, liquidity, leverage, valuation, and growth, see Edirisinghe & Zhang (2007; 2008). Such data on firms is obtained from the publicly-available financial statements. We consider 4 input parameters and 4 output parameters, as given in Table 1.

Accounts receivables (AR) represent the money yet to be collected for the sales made on credit to purchasers of the firm's products and services, and thus, it is preferable for these short-term debt collections to be small. Long-term Debt (LD) is the loans and financial obligations lasting over a year, and a firm wishes to generate revenues with the least possible LD. Capital expenditure (CAPEX) is used to acquire or upgrade physical assets such as equipment, property, or industrial buildings to generate future benefits, and thus, a firm prefers to use smaller amounts of CAPEX to generate greater benefits. Cost of goods sold (COGS) is the cost of the merchandise that was sold to customers which includes labor, materials, overhead, depreciation, etc. Obviously, a smaller COGS is an indicator of managerial excellence. On the other hand, revenue (RV) and earnings per share (EPS) represent the metrics of profitability of a firm which are necessarily objectives to be maximized. Net income growth rate (NIG) is the sequential rate of growth of the income from period to period, the increase of which is a firm's operational strategy. Price to book (P/B), which is a stock's market value to its book value, is generally larger for a growth company. A lower P/B ratio could mean that the stock is undervalued or possibly there is something fundamentally wrong with the company. Hence, we seek firms in which P/B ratio is large enough by considering it as an output parameter. Accordingly, referring to Table 1, the parameter index sets for the DEA model are $\mathcal{I} = \{1, 2, 3, 4\}$ and $\mathcal{O} = \{5, 6, 7, 8\}$.

It must be noted that the chosen set of firms, N, operate within a particular segment of the economy, for instance, as identified by one of the nine market sectors of the economy, see

i	Financial parameter		Status
1	Accounts Receivables	(AR)	Input
2	Long-term Debt	(LD)	Input
3	Captial Expenditure	(CAPEX)	Input
4	Cost of Goods Sold	(COGS)	Input
5	Revenue	(RV)	Output
6	Earnings per Share	(EPS)	Output
7	Price to Book ratio	(P/B)	Output
8	Net Income Growth Rate	(NIG)	Output

Table 1. Input/Output parameters for fundamental financial strength

Section 3. These sectors have distinct characteristics that make them unique representatives of the overall market. Therefore, evaluation of the efficiency of a firm k must be relative to a representative set of firms of the market sector to which the firm k belongs. Accordingly, the model (11) is applied within each market sector separately. The firm efficiencies so-computed using the above fundamental financial parameters (within a given sector) are herein referred to as Relative Fundamental Strength (RFS) values.

4.2 Asset selection criteria

Having the required financial data for all firms, firm k's fundamental strength (RFS) in period t is evaluated as η_k^t relative to the other firms in the chosen market sector. Those firms with sufficiently large values of η_k^t are deemed strong candidates for long-investment and those with sufficiently small values of η_k^t are deemed strong candidates for short-investment in period t. Such an asset selection criterion based on fundamental firm strengths is supported by the premise of (semi-strong) Efficient Market Hypothesis (EMH), see Fama (1970), which states that all publicly-available information including data reported in a company's financial statements is fully reflected in its current stock price.

However, the practical implementation of such a firm-selection rule at the beginning of period t is impossible because the required financial data ξ is not available until the end of period t. Therefore, a projection of (the unknown) η_k^t must be made at the beginning of period t using historical financial data. That is, a forecast of η_k^t must be made using the computable relative efficiencies η_k^τ of the historical periods $\tau \in [t - t_0, t - 1]$, where t_0 (≥ 1) is a specified historical time window prior to quarter t. Then, if this forecasted efficiency, denoted by $\hat{\eta}_k^t$, is no less than a pre-specified threshold η_L, where $\eta_L \in (0, 1)$, the firm k is declared a potential long-investment with relatively smaller long-term risk. Conversely, if the forecasted efficiency is no larger than a pre-specified threshold η_S, where $\eta_S \in (0, \eta_L)$, the firm k is declared a potential short-investment with relatively smaller long-term risk. Following this Asset Selection Criterion (ASC), the set of N assets is screened for its long term risk propensity and two subsets of assets, N_L and N_S, are determined for long and short investments, respectively, as given below:

$$\text{(ASC)} \ : \quad \begin{cases} N_L := \{k \ : \ \hat{\eta}_k^t \geq \eta_L, \ k = 1, \ldots, N\} \\ N_S := \{k \ : \ \hat{\eta}_k^t \leq \eta_S, \ k = 1, \ldots, N\} \end{cases} \tag{12}$$

To compute the forecast required in ASC, a variety of time series techniques may be employed; however, for the purposes of this chapter, a weighted moving average method is used as

follows:

$$\hat{\eta}_k^t := \sum_{\tau=t-t_0}^{t-1} v_\tau \eta_k^\tau \tag{13}$$

where the convex multipliers v_τ satisfy $v_\tau = 2v_{\tau-1}$, i.e., RFS in period τ is considered to be twice as influential as RFS in period $\tau - 1$, and thus, the convex multipliers v form a geometric progression.

5. Application to portfolio selection

In the application of the short-term risk model (10) in Section 3, sector-based ETFs are used as portfolio assets. In this section, within each such sector, rather than using an ETF, individual firms themselves are selected according to the ASC criterion to mitigate long-term risk exposure within the sector. Subsequently, the model (10) is applied on those screened firms to manage short term risks when rebalancing the portfolio.

The number of firms covered by each sectoral ETF is shown in Table 2, the sum total of which is the set of firms covered by the Standard & Poor's 500 index. However, Financial sector firms are excluded from our analysis following the common practice in many empirical studies in finance. The basic argument is that financial stocks are not only sensitive to the standard business risks of most industrial firms but also to changes in interest rates. In this regard, the famous Fama & French (1992) study also noted: "we exclude financial firms because the high leverage that is normal with these firms probably does not have the same meaning as for nonfinancial firms, where high leverage more likely indicates distress." We also exclude some of the firms in Table 2 due to unavailability of clean data, resulting in the use of only $\hat{N}(h) = 77, 54, 31, 53, 29, 53, 36, 32$ for $h = 1, \ldots, 8$, respectively. Therefore, only 365 out of 416 non-financial sector firms in S&P 500 (or, about 88%) are subjected to fundamental analysis here.

Sector	Technology	Health Care	Basic Mat.	Industrial Goods	Energy
h	1	2	3	4	5
# firms, $N(h)$	86	57	31	53	29
Sector	Consumer Discre.	Consumer Stap.	Utilities	Financial	(**Total**)
h	6	7	8	9	
# firms, $N(h)$	90	38	32	84	(**500**)

Table 2. Market sectors and number of firms

Suppose the investment portfolio selection is desired at the beginning of 2005. By fixing $t_0 = 4$, the ASC criterion requires computing firm efficiencies for the four quarters in 2004. The quarterly financial statement data for parameters in Table 1 for 2004 are used to compute firm efficiencies within a sector. The resulting relative efficiency (strength) scores η_k of 2004 are used to forecast the 2005Q1 fundamental strengths $\hat{\eta}_k$, see (13). Observe that 2004 firm-efficiencies for Q1, Q2, Q3, and Q4 periods are thus weighted by the factors $\frac{1}{15}$, $\frac{2}{15}$, $\frac{4}{15}$, $\frac{8}{15}$, respectively. These efficiency forecasts are plotted in Figure 5 for each sector.

5.1 Firm selections and portfolio optimization

Stocks for possible long-investment are chosen from each sector by using the threshold $\eta_L = 0.80$ while those for likely short-investment use the threshold $\eta_S = 0.2$. All stocks with $0.20 <$

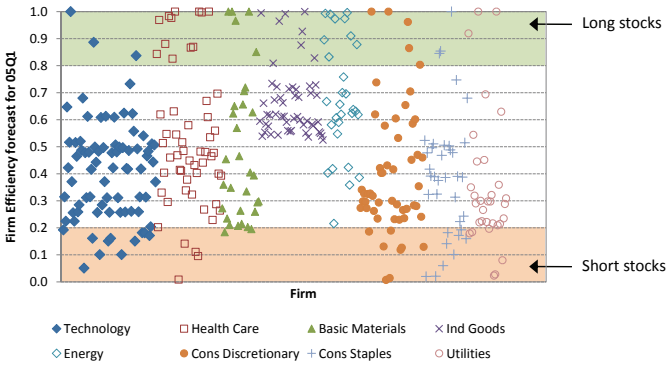


Fig. 5. Firm efficiency 2005Q1 forecasts based on Fundamental Analysis

$\hat{\eta}_k < 0.8$ are not considered for further portfolio analysis due to their (relatively unknown) long term risk potential. This leads to a total of 48 stocks as candidates of long investment and 40 stocks as short candidates. These long (short) number of stocks for each sector h are given by $N_L(N_S)$ = 3(11), 12(4), 5(2), 8(0), 9(0), 5(8), 3(9), 3(6), respectively, for $h = 1, \ldots, 8$. That is, only about 13% and 11% (of the 365 firms) are selected, respectively, for possible long and short investments, whereas the remaining 76% of the firms are labeled "neutral" with respect to investment at the beginning of 2005.

For the 88 stocks selected from the 8 sectors under the RFS metric, a monthly-rebalanced portfolio optimization is carried out from Jan-Jun, 2005, similar to Section 3, using estimates of stock parameters (means, variances, covariances, and stock betas) calculated in the exact same manner using the historical data of 2003 and 2004. In this way, performance of the portfolio of sector ETFs using the risk control model in (10) is directly compared with the long/short portfolio made up of the 88 stocks selected via DEA-based fundamental analysis on stocks within each market sector (albeit the financials). We use the extended-MV (MVX) model with complete market neutrality ($\gamma_0 = 0, \gamma_1 = 0.05$) and Cat risk control at $c = 5\%$. Accordingly, the MVX model allocations using ETFs is refereed to as the ETF portfolio, and the MVX model using RFS-based stock selections is referred to as the RFS portfolio.

For the RFS portfolio model, in (10), we impose the restrictions $x_j \geq 0$ for $j \in N_L$ and $x_j \leq 0$ for $j \in N_S$ to ensure long and short investments are made only in the appropriate RFS-based stock group. The mean-standard deviation efficient frontiers for the two competing sector portfolios are in Figure 6. By virtue of fundamental strength based stock selections within each sector, the RFS-based MVX portfolio is superior to the ETF-based MVX portfolio. In particular, the RFS-based stock discrimination yields significantly better portfolios at higher levels of variance risk. The efficient frontiers using return-to-drawdown and maximum drawdown are in Figure 7. Note the dramatic improvement in RTD at moderate levels of maxDD.

Since the market index (SPY) has an annualized volatility of 10.55% during Jan-Jun, 2005, the risk tolerance parameter λ in (10) is adjusted such that the resulting ETF-based and RFS-based portfolios also will have an out of sample annualized volatility of 10.55% during the same period. These portfolio performances are presented in Figure 8. Quite interestingly, at this level of market volatility, ETF- and RFS-based portfolios track each other closely during the first half of the investment horizon, but it is in the second half that the RFS-based stock

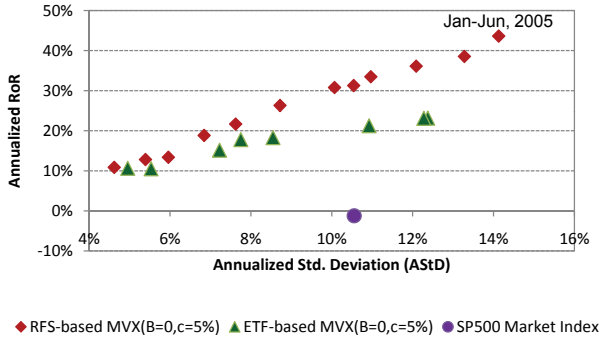


Fig. 6. Efficient frontiers of MVX portfolios using ETF and RFS

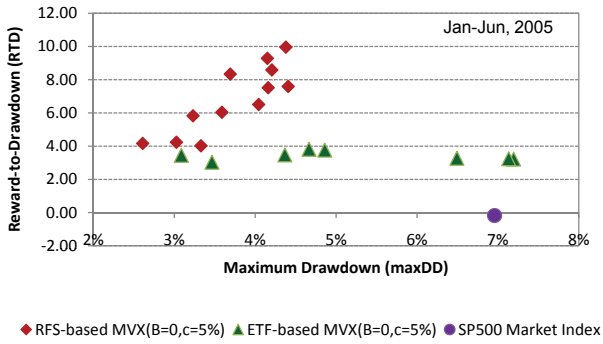


Fig. 7. Drawdown performance of MVX portfolios using ETF and RFS

selections have outperformed the general sector-based ETFs. Monthly long/short exposure for the two portfolios are in Figure 9. It is evident that the fundamental strength based stock screening yields a reduced risk exposure (both long and short) relative to the pure ETF investments.

6. Conclusions

This chapter presents a methodology for risk management in equity portfolios from a long term and short term points of view. Long term view is based on stock selections using the underlying firms' fundamental financial strength, measured relative to the competing firms in a given market sector. The short term risk control is based on an extended mean variance framework where two additional risk metrics are incorporated, namely, portfolio's market dependence and catastrophic risk. It was shown that with portfolios managed with market neutrality under controlled Cat risk, the resulting out-of-sample performance can be orders of magnitude better than the standard MV portfolio. Furthermore, when coupled with the proposed RFS-based stock selection, these MVX portfolios can display outstanding performance, especially during times when the market is expected to have poor performance, such as the investment horizon considered in this chapter.

The work in this chapter can be further improved by considering multi-period short term risk control optimization models, see Edirisinghe (2007). Moreover, applying the RFS methodology based on a rolling horizon basis from quarter-to-quarter, rather than for two quarters at a time as was done in this chapter, it may be possible to achieve further improvement in portfolio performance.

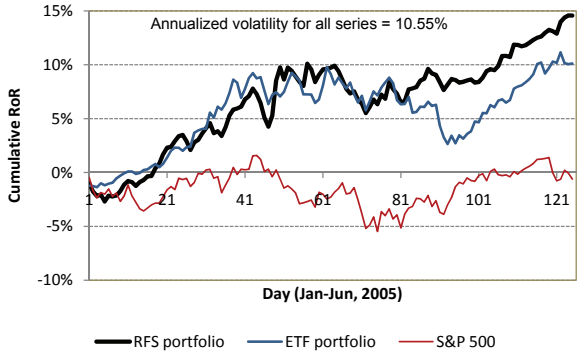


Fig. 8. Out-of-sample performance of ETF and RFS portfolios

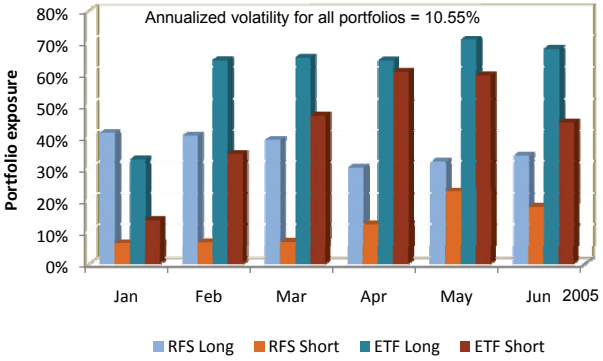


Fig. 9. Long/short risk exposure in ETF and RFS portfolios

7. References

Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent Measures of Risk, *Mathematical Finance* 9: 203–228.

Banker, R. (1984). Estimating most productive scale size using Data Envelopment Analysis, *European Journal of Operational Research* 17: 35–44.

Charnes, A., Cooper, W. & Rhodes, E. (1978). Measuring the efficiency of decision-making units, *European Journal of Operational Research* 2: 429–444.

Chen, Z. & Wang, Y. (2008). Two-sided coherent risk measures and their application in realistic portfolio optimization, *Journal of Banking and Finance* 32(12): 2667–2673.

Edirisinghe, N. (2007). Integrated risk control using stochastic programming ALM models for money management, *in* S. Zenios & W. Ziemba (eds), *Handbook of Asset and Liability Management*, Vol. 2, Elsevier Science BV, chapter 16, pp. 707–750.

Edirisinghe, N. & Zhang, X. (2007). Generalized DEA model of fundamental analysis and its application to portfolio optimization, *Journal of Banking and Finance* 31: 3311–3335.

Edirisinghe, N. & Zhang, X. (2008). Portfolio selection under DEA-based relative financial strength indicators: case of US industries, *Journal of the Operational Research Society* 59: 842–856.

Fama, E. & French, K. (1992). The Cross-Section of Expected Stock Returns, *Journal of Finance* 47: 427–465.

Farrell, M. (1957). The Measurement of Productive Efficiency, *Journal of the Royal Statistical Society* 120: 253–281.

Follmer, H. & Schied, A. (2002). Convex Measures of Risk and Trading Constraints, *Finance and Stochastics* 6(4): 429–447.

Gulpinar, N., Osorio, M. A., Rustem, B. & Settergren, R. (2004). Tax Impact on Multistage Mean-Variance Portfolio Allocation, *International Transactions in Operational Research* 11: 535–554.

Gulpinar, N., Rustem, B. & Settergren, R. (2003). Multistage Stochastic Mean-Variance Portfolio Analysis with Transaction Cost, *Innovations in Financial and Economic Networks*, Edward Elgar Publishers, U.K., pp. 46–63.

Hirschey, M. (2003). Extreme Return Reversal in the Stock Market- Strong support for insightful fundamental analysis, *The Journal of Portfolio Management* 29: 78–90.

Jarrow, R. (2002). Put Option Premiums and Coherent Risk Measures, *Mathematical Finance* 12: 125–134.

Kaut, M., Vladimirou, H. & Wallace, S. (2007). Stability analysis of portfolio management with conditional value-at-risk, *Quantitative Finance* 7(4): 397–409.

Konno, H. & Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market, *Management Science* 37(5): 519–531.

Markowitz, H. (1952). Portfolio selection, *Journal of Finance* 7: 77–91.

Ogryczak, W. & Ruszczynski, A. (1999). From stochastic dominance to mean-risk models: Semideviations as risk measures, *European Journal of Operational Research* 116: 33–50.

Pirvu, T. (2007). Portfolio optimization under the Value-at-Risk constraint, *Quantitative Finance* 7(2): 125–136.

Purnanandam, A., Warachka, M., Zhao, Y. & Ziemba, W. (2006). *Incorporating diversification into risk management*, Palgrave.

Samaras, G., Matsatsinis, N. & Zopounidis, C. (2008). A multi-criteria DSS for stock evaluation using fundamental analysis, *European Journal of Operational Research* 187: 1380–1401.

Thomsett, M. (1998). *Mastering Fundamental Analysis*, Dearborn, Chicago.

von Neumann, J. & Morgenstern, O. (1991). *Theory of games and economic behavior*, Princeton University Press, Princeteon.

Whitmore, G. & Findlay, M. (1978). *Stochastic Dominance: An Approach to Decision Making Under Risk*, Heath, Lexington, MA.

Zhao, Y. & Zeimba, W. (2001). A stochastic programming model using an endogenously determined worst case risk measure for dynamic asset allocation, *Mathematical Programming* 89: 293–309.

Zhu, S. & Fukushima, M. (2009). Worst-Case Conditional Value-at-Risk with Application to Robust Portfolio Management, *Operations Research* 57(5): 1155–1168.

7

Currency Trading Using the Fractal Market Hypothesis

Jonathan Blackledge and Kieran Murphy
Dublin Institute of Technology
Ireland

1. Introduction

We report on a research and development programme in financial modelling and economic security undertaken in the Information and Communications Security Research Group (ICSRG, 2011) which has led to the launch of a new company - Currency Traders Ireland Limited - funded by Enterprise Ireland. Currency Traders Ireland Limited (CTI, 2011) has a fifty year exclusive license to develop a new set of indicators for analysing currency exchange rates (Forex trading). We consider the background to the approach taken and present examples of the results obtained to date. In this 'Introduction', we provide a background to and brief overview of conventional economic models and the problesms associated with them.

1.1 Background to financial time series modelling

The application of mathematical, statistical and computational techniques for analysing financial time series is a well established practice. Computational finance is used every day to help traders understand the dynamic performance of the markets and to have some degree of confidence on the likely future behaviour of the markets. This includes the application of stochastic modelling methods and the use of certain partial differential equations for describing financial systems (e.g. the Black-Scholes equation for financial derivatives). Attempts to develop stochastic models for financial time series, which are essentially digital signals composed of 'tick data'[1] can be traced back to the early Twentieth Century when Louis Bachelier proposed that fluctuations in the prices of stocks and shares (which appeared to be yesterday's price plus some random change) could be viewed in terms of random walks in which price changes were entirely independent of each other. Thus, one of the simplest models for price variation is based on the sum of independent random numbers. This is the basis for Brownian motion (i.e. the random walk motion first observed by the Scottish Botanist Robert Brown) in which the random numbers are considered to conform to a normal of Gaussian distribution. For some financial signal $u(t)$ say (where u is the amplitude - the 'price' - of the signal and t is time), the magnitude of a change in price du tends to depend on the price u itself. We therefore modify the Brownian random walk model to include this observation. In this case, the logarithm of the price change (which is also assumed to conform

[1] Data that provides traders with daily tick-by-tick data - time and sales - of trade price, trade time, and volume traded, for example, at different sampling rates.

to a normal distribution) is given by

$$\frac{du}{u} = \alpha dv + \beta dt \quad \text{or} \quad \frac{d}{dt}\ln u = \beta + \alpha\frac{dv}{dt}$$

where α is the volatility, dv is a sample from a normal distribution and β is a drift term which reflects the average rate of growth of an asset[2]. Here, the relative price change of an asset is equal to a random value plus an underlying trend component. This is the basis for a 'log-normal random walk' model (Copeland et al., 2003), (Martin et al., 1997), (Menton, 1992) and (Watsham and Parramore, 1996).
Brownian motion models have the following basic properties:

- statistical stationarity of price increments in which samples of Brownian motion taken over equal time increments can be superimposed onto each other in a statistical sense;
- scaling of price where samples of Brownian motion corresponding to different time increments can be suitably re-scaled such that they too, can be superimposed onto each other in a statistical sense.

Such models fail to predict extreme behaviour in financial time series because of the intrinsic assumption that such time series conform to a normal distribution, i.e. Gaussian processes that are stationary in which the statistics - the standard deviation, for example - do not change with time.
Random walk models, which underpin the so called Efficient Market Hypothesis (EMH) (Fama, 1965)-(Burton, 1987), have been the basis for financial time series analysis since the work of Bachelier in the late Nineteenth Century. Although the Black-Scholes equation (Black& Scholes, 1973), developed in the 1970s for valuing options, is deterministic (one of the first financial models to achieve determinism), it is still based on the EMH, i.e. stationary Gaussian statistics. The EMH is based on the principle that the current price of an asset fully reflects all available information relevant to it and that new information is immediately incorporated into the price. Thus, in an efficient market, the modelling of asset prices is concerned with modelling the arrival of new information. New information must be independent and random, otherwise it would have been anticipated and would not be new. The arrival of new information can send 'shocks' through the market (depending on the significance of the information) as people react to it and then to each other's reactions. The EMH assumes that there is a rational and unique way to use the available information and that all agents possess this knowledge. Further, the EMH assumes that this 'chain reaction' happens effectively instantaneously. These assumptions are clearly questionable at any and all levels of a complex financial system.
The EMH implies independence of price increments and is typically characterised by a normal of Gaussian Probability Density Function (PDF) which is chosen because most price movements are presumed to be an aggregation of smaller ones, the sums of independent random contributions having a Gaussian PDF. However, it has long been known that financial time series do not follow random walks. This is one of the most fundamental underlying problems associated with financial models, in general.

[2] Note that both α and β may very with time.

1.2 The problem with economic models

The principal aim of a financial trader is to attempt to obtain information that can provide some confidence in the immediate future of a stock. This is often based on repeating patterns from the past, patterns that are ultimately based on the interplay between greed and fear. One of the principal components of this aim is based on the observation that there are 'waves within waves' known as Elliot Waves after Ralph Elliot who was among the first to observe this phenomenon on a qualitative basis in 1938. Elliot Waves permeate financial signals when studied with sufficient detail and imagination. It is these repeating patterns that occupy both the financial investor and the financial systems modeler alike and it is clear that although economies have undergone many changes in the last one hundred years, ignoring scale, the dynamics of market behaviour does not appear to have changed significantly.

In modern economies, the distribution of stock returns and anomalies like market crashes emerge as a result of considerable complex interaction. In the analysis of financial time series it is inevitable that assumptions need to be made with regard to developing a suitable model. This is the most vulnerable stage of the process with regard to developing a financial risk management model as over simplistic assumptions lead to unrealistic solutions. However, by considering the global behaviour of the financial markets, they can be modeled statistically provided the 'macroeconomic system' is complex enough in terms of its network of interconnection and interacting components.

Market behaviour results from either a strong theoretical reasoning or from compelling experimental evidence or both. In econometrics, the processes that create time series have many component parts and the interaction of those components is so complex that a deterministic description is simply not possible. When creating models of complex systems, there is a trade-off between simplifying and deriving the statistics we want to compare with reality and simulation. Stochastic simulation allows us to investigate the effect of various traders' behaviour with regard to the global statistics of the market, an approach that provides for a natural interpretation and an understanding of how the amalgamation of certain concepts leads to these statistics and correlations in time over different scales. One cause of correlations in market price changes (and volatility) is mimetic behaviour, known as herding. In general, market crashes happen when large numbers of agents place sell orders simultaneously creating an imbalance to the extent that market makers are unable to absorb the other side without lowering prices substantially. Most of these agents do not communicate with each other, nor do they take orders from a leader. In fact, most of the time they are in disagreement, and submit roughly the same amount of buy and sell orders. This provides a diffusive economy which underlies the Efficient Market Hypothesis (EMH) and financial portfolio rationalization. The EMH is the basis for the Black-Scholes model developed for the Pricing of Options and Corporate Liabilities for which Scholes won the Nobel Prize for economics in 1997. However, there is a fundamental flaw with this model which is that it is based on a hypothesis (the EMH) that assumes price movements, in particular, the log-derivate of a price, is normally distributed and this is simply not the case. Indeed, all economic time series are characterized by long tail distributions which do not conform to Gaussian statistics thereby making financial risk management models such as the Black-Scholes equation redundant.

1.3 What is the fractal market hypothesis?

The economic basis for the Fractal Market Hypothesis (FMH) is as follows:

- The market is stable when it consists of investors covering a large number of investment horizons which ensures that there is ample liquidity for traders;
- information is more related to market sentiment and technical factors in the short term than in the long term - as investment horizons increase and longer term fundamental information dominates;
- if an event occurs that puts the validity of fundamental information in question, long-term investors either withdraw completely or invest on shorter terms (i.e. when the overall investment horizon of the market shrinks to a uniform level, the market becomes unstable);
- prices reflect a combination of short-term technical and long-term fundamental valuation and thus, short-term price movements are likely to be more volatile than long-term trades - they are more likely to be the result of crowd behaviour;
- if a security has no tie to the economic cycle, then there will be no long-term trend and short-term technical information will dominate.

The model associated with the FMH considered in this is is compounded in a fractional dynamic model that is non-stationary and describes diffusive processes that have a directional bias leading to long tail (non-Gaussian) distributions. We consider a Lévy distribution and show the relation between this distribution and the fractional diffusion equation (Section 4.2). Unlike the EMH, the FMH states that information is valued according to the investment horizon of the investor. Because the different investment horizons value information differently, the diffusion of information is uneven. Unlike most complex physical systems, the agents of an economy, and perhaps to some extent the economy itself, have an extra ingredient, an extra degree of complexity. This ingredient is consciousness which is at the heart of all financial risk management strategies and is, indirectly, a governing issue with regard to the fractional dynamic model used to develop the algorithm now being used by Currency Traders Ireland Limited. By computing an index called the Lévy index, the directional bias associated with a future trend can be forecast. In principle, this can be achieved for any financial time series, providing the algorithm has been finely tuned with regard to the interpretation of a particular data stream and the parameter settings upon which the algorithm relies.

2. The black-scholes model

For many years, investment advisers focused on returns with the occasional caveat 'subject to risk'. Modern Portfolio Theory (MPT) is concerned with a trade-off between risk and return. Nearly all MPT assumes the existence of a risk-free investment, e.g. the return from depositing money in a sound financial institute or investing in equities. In order to gain more profit, the investor must accept greater risk. Why should this be so? Suppose the opportunity exists to make a guaranteed return greater than that from a conventional bank deposit say; then, no (rational) investor would invest any money with the bank. Furthermore, if he/she could also borrow money at less than the return on the alternative investment, then the investor would borrow as much money as possible to invest in the higher yielding opportunity. In response to the pressure of supply and demand, the banks would raise their interest rates. This would attract money for investment with the bank and reduce the profit made by investors who have money borrowed from the bank. (Of course, if such opportunities did arise, the banks would probably be the first to invest savings in them.) There is elasticity in the argument because of various 'friction factors' such as transaction costs, differences in borrowing and lending rates,

liquidity laws etc., but on the whole, the principle is sound because the market is saturated with arbitrageurs whose purpose is to seek out and exploit irregularities or miss-pricing.
The concept of successful arbitraging is of great importance in finance. Often loosely stated as, 'there's no such thing as a free lunch', it means that one cannot ever make an instantaneously risk-free profit. More precisely, such opportunities cannot exist for a significant length of time before prices move to eliminate them.

2.1 Financial derivatives

As markets have grown and evolved, new trading contracts have emerged which use various tricks to manipulate risk. Derivatives are deals, the value of which is derived from (although not the same as) some underlying asset or interest rate. There are many kinds of derivatives traded on the markets today. These special deals increase the number of moves that players of the economy have available to ensure that the better players have more chance of winning. To illustrate some of the implications of the introduction of derivatives to the financial markets we consider the most simple and common derivative, namely, the option.

2.1.1 Options

An option is the right (but not the obligation) to buy (call) or sell (put) a financial instrument (such as a stock or currency, known as the 'underlying') at an agreed date in the future and at an agreed price, called the strike price. For example, consider an investor who 'speculates' that the value of an asset at price S will rise. The investor could buy shares at S, and if appropriate, sell them later at a higher price. Alternatively, the investor might buy a call option, the right to buy a share at a later date. If the asset is worth more than the strike price on expiry, the holder will be content to exercise the option, immediately sell the stock at the higher price and generate an automatic profit from the difference. The catch is that if the price is less, the holder must accept the loss of the premium paid for the option (which must be paid for at the opening of the contract). If C denotes the value of a call option and E is the strike price, the option is worth $C(S,t) = \max(S - E, 0)$.
Conversely, suppose the investor speculates that an asset is going to fall, then the investor can sell shares or buy puts. If the investor speculates by selling shares that he/she does not own (which in certain circumstances is perfectly legal in many markets), then he/she is selling 'short' and will profit from a fall in the asset. (The opposite of a short position is a 'long' position.) The principal question is how much should one pay for an option? If the value of the asset rises, then so does the value of a call option and vice versa for put options. But how do we quantify exactly how much this gamble is worth? In previous times (prior to the Black-Scholes model which is discussed later) options were bought and sold for the value that individual traders thought they ought to have. The strike prices of these options were usually the 'forward price', which is just the current price adjusted for interest-rate effects. The value of options rises in active or volatile markets because options are more likely to pay out large amounts of money when they expire if market moves have been large, i.e. potential gains are higher, but loss is always limited to the cost of the premium. This gain through successful 'speculation' is not the only role that options play. Another role is Hedging.

2.1.2 Hedging

Suppose an investor already owns shares as a long-term investment, then he/she may wish to insure against a temporary fall in the share price by buying puts as well. The investor would not want to liquidate holdings only to buy them back again later, possibly at a higher price if

the estimate of the share price is wrong, and certainly having incurred some transaction costs on the deals. If a temporary fall occurs, the investor has the right to sell his/her holdings for a higher than market price. The investor can then immediately buy them back for less, in this way generating a profit and long-term investment then resumes. If the investor is wrong and a temporary fall does not occur, then the premium is lost for the option but at least the stock is retained, which has continued to rise in value. Since the value of a put option rises when the underlying asset value falls, what happens to a portfolio containing both assets and puts? The answer depends on the ratio. There must exist a ratio at which a small unpredictable movement in the asset does not result in any unpredictable movement in the portfolio. This ratio is instantaneously risk free. The reduction of risk by taking advantage of correlations between the option price and the underlying price is called 'hedging'. If a market maker can sell an option and hedge away all the risk for the rest of the options life, then a risk free profit is guaranteed.

Why write options? Options are usually sold by banks to companies to protect themselves against adverse movements in the underlying price, in the same way as holders do. In fact, writers of options are no different to holders; they expect to make a profit by taking a view of the market. The writers of calls are effectively taking a short position in the underlying behaviour of the markets. Known as 'bears', these agents believe the price will fall and are therefore also potential customers for puts. The agents taking the opposite view are called 'bulls'. There is a near balance of bears and bulls because if everyone expected the value of a particular asset to do the same thing, then its market price would stabilise (if a reasonable price were agreed on) or diverge (if everyone thought it would rise). Thus, the psychology and dynamics (which must go hand in hand) of the bear/bull cycle play an important role in financial analysis.

The risk associated with individual securities can be hedged through diversification or 'spread betting' and/or various other ways of taking advantage of correlations between different derivatives of the same underlying asset. However, not all risk can be removed by diversification. To some extent, the fortunes of all companies move with the economy. Changes in the money supply, interest rates, exchange rates, taxation, commodity prices, government spending and overseas economies tend to affect all companies in one way or another. This remaining risk is generally referred to as market risk.

2.2 Black-scholes analysis

The value of an option can be thought of as a function of the underlying asset price S (a Gaussian random variable) and time t denoted by $V(S,t)$. Here, V can denote a call or a put; indeed, V can be the value of a whole portfolio or different options although for simplicity we can think of it as a simple call or put. Any derivative security whose value depends only on the current value S at time t and which is paid for up front, is taken to satisfy the Black-Scholes equation given by (Black& Scholes, 1973)

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0$$

where σ is the volatility and r is the risk. As with other partial differential equations, an equation of this form may have many solutions. The value of an option should be unique; otherwise, again, arbitrage possibilities would arise. Therefore, to identify the appropriate solution, certain initial, final and boundary conditions need to be imposed. Take for example,

a call; here the final condition comes from the arbitrage argument. At $t = T$

$$C(S,t) = \max(S - E, 0)$$

The spatial or asset-price boundary conditions, applied at $S = 0$ and $S \to \infty$ come from the following reasoning: If S is ever zero then dS is zero and will therefore never change. Thus, we have

$$C(0,t) = 0$$

As the asset price increases it becomes more and more likely that the option will be exercised, thus we have

$$C(S,t) \propto S, \qquad S \to \infty$$

Observe, that the Black-Sholes equation has a similarity to the diffusion equation but with additional terms. An appropriate way to solve this equation is to transform it into the diffusion equation for which the solution is well known and, with appropriate Transformations, gives the Black-Scholes formula (Black& Scholes, 1973)

$$C(S,t) = SN(d_1) - Ee^{r(T-t)}N(d_2)$$

where

$$d_1 = \frac{\log(S/E) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}},$$

$$d_2 = \frac{\log(S/E) + (r - \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}}$$

and N is the cumulative normal distribution defined by

$$N(d_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d_1} e^{\frac{1}{2}s^2} ds.$$

The conceptual leap of the Black-Scholes model is to say that traders are not estimating the future price, but are guessing about how volatile the market may be in the future. The model therefore allows banks to define a fair value of an option, because it assumes that the forward price is the mean of the distribution of future market prices. However, this requires a good estimate of the future volatility σ.

The relatively simple and robust way of valuing options using Black-Scholes analysis has rapidly gained in popularity and has universal applications. Black-Scholes analysis for pricing an option is now so closely linked into the markets that the price of an option is usually quoted in option volatilities or 'vols'. However, Black-Scholes analysis is ultimately based on random walk models that assume independent and Gaussian distributed price changes and is thus, based on the EMH.

The theory of modern portfolio management is only valuable if we can be sure that it truly reflects reality for which tests are required. One of the principal issues with regard to this relates to the assumption that the markets are Gaussian distributed. However, it has long been known that financial time series do not adhere to Gaussian statistics. This is the most important of the shortcomings relating to the EMH model (i.e. the failure of the independence and Gaussian distribution of increments assumption) and is fundamental to the inability for EMH-based analysis such as the Black-Scholes equation to explain characteristics of a

financial signal such as clustering, flights and failure to explain events such as 'crashesŠ leading to recession. The limitations associated with the EMH are illustrated in Figure 1 which shows a (discrete) financial signal $u(t)$, the derivative of this signal $du(t)/dt$ and a synthesised (zero-mean) Gaussian distributed random signal. There is a marked difference in the characteristics of a real financial signal and a random Gaussian signal. This simple comparison indicates a failure of the statistical independence assumption which underpins the EMH and the superior nature of the Lévy based model that underpins the Fractal Market Hypothesis.

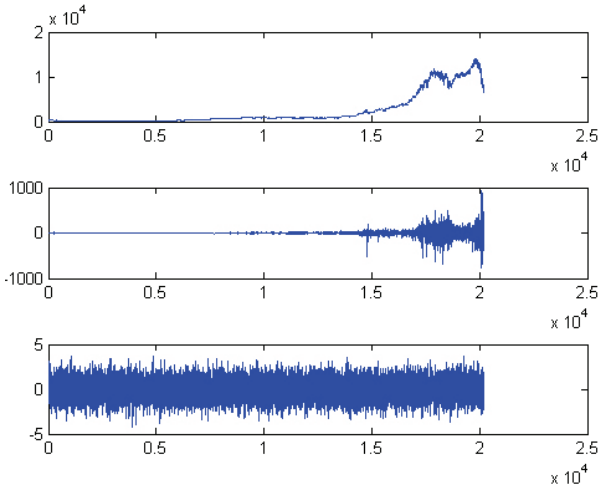


Fig. 1. Financial time series for the Dow-Jones value (close-of-day) from 02-04-1928 to 12-12-2007 (top), the derivative of the same time series (centre) and a zero-mean Gaussian distributed random signal (bottom).

The problems associated with financial modelling using the EMH have prompted a new class of methods for investigating time series obtained from a range of disciplines. For example, Re-scaled Range Analysis (RSRA), e.g. (Hurst, 1951), (Mandelbrot, 1969), which is essentially based on computing and analysing the Hurst exponent (Mandelbrot, 1972), is a useful tool for revealing some well disguised properties of stochastic time series such as persistence (and anti-persistence) characterized by non-periodic cycles. Non-periodic cycles correspond to trends that persist for irregular periods but with a degree of statistical regularity often associated with non-linear dynamical systems. RSRA is particularly valuable because of its robustness in the presence of noise. The principal assumption associated with RSRA is concerned with the self-affine or fractal nature of the statistical character of a time-series rather than the statistical 'signature' itself. Ralph Elliott first reported on the fractal properties of financial data in 1938. He was the first to observe that segments of financial time series data of different sizes could be scaled in such a way that they were statistically the same producing so called Elliot waves. Since then, many different self-affine models for price variation have been developed, often based on (dynamical) Iterated Function Systems (IFS). These models can capture many properties of a financial time series but are not based on any underlying causal theory.

3. Fractal time series and rescaled range analysis

A time series is fractal if the data exhibits statistical self-affinity and has no characteristic scale. The data has no characteristic scale if it has a PDF with an infinite second moment. The data may have an infinite first moment as well; in this case, the data would have no stable mean either. One way to test the financial data for the existence of these moments is to plot them sequentially over increasing time periods to see if they converge. Figure 2 shows that the first moment, the mean, is stable, but that the second moment, the mean square, is not settled. It converges and then suddenly jumps and it is observed that although the variance is not stable, the jumps occur with some statistical regularity. Time series of this type are example of Hurst processes; time series that scale according to the power law,

$$\langle u(t) \rangle_t \propto t^H$$

where H is the Hurst exponent and $\langle u(t) \rangle_t$ denotes the mean value of $u(t)$ at a time t.

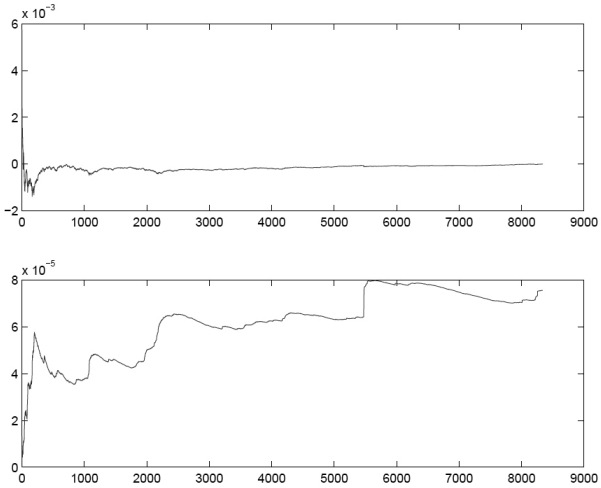


Fig. 2. The first and second moments (top and bottom) of the Dow Jones Industrial Average plotted sequentially.

H. E. Hurst (1900-1978) was an English civil engineer who built dams and worked on the Nile river dam project. He studied the Nile so extensively that some Egyptians reportedly nicknamed him 'the father of the Nile.' The Nile river posed an interesting problem for Hurst as a hydrologist. When designing a dam, hydrologists need to estimate the necessary storage capacity of the resulting reservoir. An influx of water occurs through various natural sources (rainfall, river overflows etc.) and a regulated amount needed to be released for primarily agricultural purposes. The storage capacity of a reservoir is based on the net water flow. Hydrologists usually begin by assuming that the water influx is random, a perfectly reasonable assumption when dealing with a complex ecosystem. Hurst, however, had studied the 847-year record that the Egyptians had kept of the Nile river overflows, from 622 to 1469. Hurst noticed that large overflows tended to be followed by large overflows until abruptly, the system would then change to low overflows, which also tended to be followed by low overflows. There seemed to be cycles, but with no predictable period. Standard statistical

analysis revealed no significant correlations between observations, so Hurst developed his own methodology. Hurst was aware of Einstein's (1905) work on Brownian motion (the erratic path followed by a particle suspended in a fluid) who observed that the distance the particle covers increased with the square root of time, i.e.

$$R \propto \sqrt{t}$$

where R is the range covered, and t is time. This relationship results from the fact that increments are identically and independently distributed random variables. Hurst's idea was to use this property to test the Nile River's overflows for randomness. In short, his method was as follows: Begin with a time series x_i (with $i = 1, 2, ..., n$) which in Hurst's case was annual discharges of the Nile River. (For markets it might be the daily changes in the price of a stock index.) Next, create the adjusted series, $y_i = x_i - \bar{x}$ (where $\bar{x}$ is the mean of x_i). Cumulate this time series to give

$$Y_i = \sum_{j=1}^{i} y_j$$

such that the start and end of the series are both zero and there is some curve in between. (The final value, Y_n has to be zero because the mean is zero.) Then, define the range to be the maximum minus the minimum value of this time series,

$$R_n = \max(Y) - \min(Y).$$

This adjusted range, R_n is the distance the systems travels for the time index n, i.e. the distance covered by a random walker if the data set y_i were the set of steps. If we set $n = t$ we can apply Einstein's equation provided that the time series x_i is independent for increasing values of n. However, Einstein's equation only applies to series that are in Brownian motion. Hurst's contribution was to generalize this equation to

$$(R/S)_n = cn^H$$

where S is the standard deviation for the same n observations and c is a constant. We define a Hurst process to be a process with a (fairly) constant H value and the R/S is referred to as the 'rescaled range' because it has zero mean and is expressed in terms of local standard deviations. In general, the R/S value increases according to a power law value equal to H known as the Hurst exponent. This scaling law behaviour is the first connection between Hurst processes and fractal geometry.

Rescaling the adjusted range was a major innovation. Hurst originally performed this operation to enable him to compare diverse phenomenon. Rescaling, fortunately, also allows us to compare time periods many years apart in financial time series. As discussed previously, it is the relative price change and not the change itself that is of interest. Due to inflationary growth, prices themselves are a significantly higher today than in the past, and although relative price changes may be similar, actual price changes and therefore volatility (standard deviation of returns) are significantly higher. Measuring in standard deviations (units of volatility) allows us to minimize this problem. Rescaled range analysis can also describe time series that have no characteristic scale, another characteristic of fractals. By considering the logarithmic version of Hurst's equation, i.e.

$$\log(R/S)_n = \log(c) + H\log(n)$$

it is clear that the Hurst exponent can be estimated by plotting $\log(R/S)$ against the $\log(n)$ and solving for the gradient with a least squares fit. If the system were independently distributed, then $H = 0.5$. Hurst found that the exponent for the Nile River was $H = 0.91$, i.e. the rescaled range increases at a faster rate than the square root of time. This meant that the system was covering more distance than a random process would, and therefore the annual discharges of the Nile had to be correlated.
It is important to appreciate that this method makes no prior assumptions about any underlying distributions, it simply tells us how the system is scaling with respect to time. So how do we interpret the Hurst exponent? We know that $H = 0.5$ is consistent with an independently distributed system. The range $0.5 < H \leq 1$, implies a persistent time series, and a persistent time series is characterized by positive correlations. Theoretically, what happens today will ultimately have a lasting effect on the future. The range $0 < H \leq 0.5$ indicates anti-persistence which means that the time series covers less ground than a random process. In other words, there are negative correlations. For a system to cover less distance, it must reverse itself more often than a random process.

4. Lévy processes

Lévy processes are random walks whose distribution has infinite moments and 'long tails'. The statistics of (conventional) physical systems are usually concerned with stochastic fields that have PDFs where (at least) the first two moments (the mean and variance) are well defined and finite. Lévy statistics is concerned with statistical systems where all the moments (starting with the mean) are infinite. Many distributions exist where the mean and variance are finite but are not representative of the process, e.g. the tail of the distribution is significant, where rare but extreme events occur. These distributions include Lévy distributions (Sclesinger et al., 1994), (Nonnenmacher, 1990). Lévy's original approach to deriving such distributions is based on the following question: Under what circumstances does the distribution associated with a random walk of a few steps look the same as the distribution after many steps (except for scaling)? This question is effectively the same as asking under what circumstances do we obtain a random walk that is statistically self-affine. The characteristic function $P(k)$ of such a distribution $p(x)$ was first shown by Lévy to be given by (for symmetric distributions only)

$$P(k) = \exp(-a \mid k \mid^{\gamma}), \quad 0 < \gamma \leq 2 \tag{1}$$

where a is a constant and γ is the Lévy index. For $\gamma \geq 2$, the second moment of the Lévy distribution exists and the sums of large numbers of independent trials are Gaussian distributed. For example, if the result were a random walk with a step length distribution governed by $p(x)$, $\gamma \geq 2$, then the result would be normal (Gaussian) diffusion, i.e. a Brownian random walk process. For $\gamma < 2$ the second moment of this PDF (the mean square), diverges and the characteristic scale of the walk is lost. For values of γ between 0 and 2, Lévy's characteristic function corresponds to a PDF of the form

$$p(x) \sim \frac{1}{x^{1+\gamma}}, \quad x \to \infty$$

4.1 Long tails

If we compare this PDF with a Gaussian distribution given by (ignoring scaling normalisation constants)

$$p(x) = \exp(-\beta x^2)$$

which is the case when $\gamma = 2$ then it is clear that a Lévy distribution has a longer tail. This is illustrated in Figure 3. The long tail Lévy distribution represents a stochastic process in

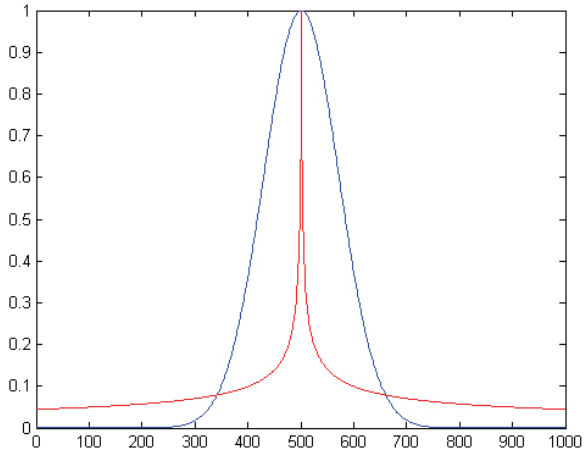


Fig. 3. Comparison between a Gaussian distribution (blue) for $\beta = 0.0001$ and a Lévy distribution (red) for $\gamma = 0.5$ and $p(0) = 1$.

which extreme events are more likely when compared to a Gaussian process. This includes fast moving trends that occur in economic time series analysis. Moreover, the length of the tails of a Lévy distribution is determined by the value of the Lévy index such that the larger the value of the index the shorter the tail becomes. Unlike the Gaussian distribution which has finite statistical moments, the Lévy distribution has infinite moments and 'long tails'.

4.2 Lévy processes and the fractional diffusion equation

Lévy processes are consistent with a fractional diffusion equation (Alea & Thurnerb, 2005) as shall now be shown. Let $p(x)$ denote the Probability Density Function (PDF) associated with the position in a one-dimensional space x where a particle can exist as a result of a 'random walk' generated by a sequence of 'elastic scattering' processes (with other like particles). Also, assume that the random walk takes place over a time scale where the random walk 'environment' does not change (i.e. the statistical processes are 'stationary' and do not change with time). Suppose we consider an infinite concentration of particles at a time $t = 0$ to be located at the origin $x = 0$ and described by a perfect spatial impulse, i.e. a delta function $\delta(x)$. Then the characteristic Impulse Response Function f of the 'random walk system' at a short time later $t = \tau$ is given by

$$f(x,\tau) = \delta(x) \otimes_x p(x) = p(x)$$

where $\otimes_x$ denotes the convolution integral over x. Thus, if $f(x,t)$ denotes a macroscopic field at a time t which describes the concentration of a canonical assemble of particles all undergoing the same random walk process, then the field at $t + \tau$ will be given by

$$f(x,t+\tau) = f(x,t) \otimes_x p(x) \tag{2}$$

In terms of the application considered in this paper $f(0,t)$ represents the time varying price difference of a financial index $u(t)$ such as a currency pair, so that, in general,

$$f(x,t) = \frac{\partial}{\partial} u(x,t) \tag{3}$$

From the convolution theorem, in Fourier space, equation (2) becomes

$$F(k,t+\tau) = F(k,t)P(k)$$

where F and P are the Fourier transforms of f and p, respectively. From equation (1), we note that

$$P(k) = 1 - a \mid k \mid^{\gamma}, \quad a \to 0$$

so that we can write

$$\frac{F(k,t+\tau) - F(k,t)}{\tau} \simeq -\frac{a}{\tau} \mid k \mid^{\gamma} F(k,t)$$

which for $\tau \to 0$ gives the fractional diffusion equation

$$\sigma \frac{\partial}{\partial t} f(x,t) = \frac{\partial^{\gamma}}{\partial x^{\gamma}} f(x,t), \quad \gamma \in (0,2]$$

where $\sigma = \tau/a$ and we have used the result

$$\frac{\partial^{\gamma}}{\partial x^{\gamma}} f(x,t) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} \mid k \mid^{\gamma} F(k,t) \exp(ikx) dk$$

However, from equation (3) we can consider the equation

$$\sigma \frac{\partial}{\partial t} u(x,t) = \frac{\partial^{\gamma}}{\partial x^{\gamma}} u(x,t), \quad \gamma \in (0,2] \tag{4}$$

The solution to this equation with the singular initial condition $v(x,0) = \delta(x)$ is given by

$$v(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(ikx - t \mid k \mid^{\gamma} / \sigma) dk$$

which is itself Lévy distributed. This derivation of the fractional diffusion equation reveals its physical origin in terms of Lévy statistics.
For normalized units $\sigma = 1$ we consider equation (4) for a 'white noise' source function $n(t)$ and a spatial impulse function $-\delta(x)$ so that

$$\frac{\partial^{\gamma}}{\partial x^{\gamma}} u(x,t) - \frac{\partial}{\partial t} u(x,t) = -\delta(x) n(t), \quad \gamma \in (0,2]$$

which, ignoring (complex) scaling constants, has the Green's function solution (**?**)

$$u(t) = \frac{1}{t^{1-1/\gamma}} \otimes_t n(t) \tag{5}$$

where $\otimes_t$ denotes the convolution integral over t and $u(t) \equiv u(0,t)$. The function $u(t)$ has a Power Spectral Density Function (PSDF) given by (for scaling constant c)

$$\mid U(\omega) \mid^2 = \frac{c}{\mid \omega \mid^{2/\gamma}} \tag{6}$$

where

$$U(\omega) = \int_{-\infty}^{\infty} u(t) \exp(-i\omega t) dt$$

and a self-affine scaling relationship

$$\Pr[u(at)] = a^{1/\gamma} \Pr[u(t)]$$

for scaling parameter $a > 0$ where $\Pr[u(t)]$ denotes the PDF of $u(t)$. This scaling relationship means that the statistical characteristics of $u(t)$ are invariant of time except for scaling factor $a^{1/\gamma}$. Thus, if $u(t)$ is taken to be a financial signal as a function of time, then the statistical distribution of this function will be the same over different time scales whether, in practice, it is sampled in hours, minutes or seconds, for example.
Equation (5), provides a solution is also consistent with the solution to the fractional diffusion equation

$$\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^q}{\partial t^q} \right) u(x,t) = -\delta(x) n(t)$$

where $\gamma^{-1} = q/2$ (Blackledge, 2010) and where q - the 'Fourier Dimension' - is related to the Hurst exponent by $q = 2H + 1$. Thus, the Lévy index γ, the Fourier Dimension q and the Hurst exponent H are all simply related to each other. Moreover, these parameters quantify stochastic processes that have long tails and thereby by transcend financial models based on normal distributions such as the Black-Scholes model.

4.3 Computational methods

In this paper, we study the temporal behaviour of q focusing on its predictive power for indicating the likelihood of a future trend in a Forex time series. This is called the 'q-algorithm' and is equivalent to computing time variations in the the Lévy index or the Hurst exponent since $q = 2H + 1 = 2/\gamma$. Given equations (5), for $n(t) = \delta(t)$

$$u(t) = \frac{1}{t^{1-1/\gamma}}$$

and thus

$$\log u(t) = a + \frac{1}{\gamma} \log t$$

where $a = -\log t$. Thus, one way of computing γ is to evaluate the gradient of a plot of $\log u(t)$ against $\log t$. If this is done on a moving window basis then a time series $\gamma(t)$ can be obtained and correlations observed between the behaviour of $\gamma(t)$ and $u(t)$. However, given equation (6), we can also consider the equation

$$\log \mid U(\omega) \mid = b + \frac{1}{\gamma} \log \mid \omega \mid$$

where $b = (\log c)/2$ and evaluate the gradient of a plot of $\log \mid U(\omega) \mid$ against $\log \mid \omega \mid$. In practice this requires the application of a discrete Fourier transform on a moving window basis to compute an estimate of $\gamma(t)$. In this paper, we consider the former (temporal) solution to the problem of computing $q = 2/\gamma$.

5. Application to Forex trading

The Forex or Foreign Exchange market is the largest and most fluid of the global markets involving trades approaching 4 Trillion per day. The market is primarily concerned with trading currency pairs but includes currency futures and options markets. It is similar to other financial markets but the volume of trade is much higher which comes from the nature of the market in terms of its short term profitability. The market determines the relative values of different currencies and most banks contribute to the market as do financial companies, institutions, individual speculators and investors and even import/export companies. The high volume of the Forex market leads to high liquidity and thereby guarantees stable spreads during a working week and contract execution with relatively small slippages even in aggressive price movements. In a typical foreign exchange transaction, a party purchases a quantity of one currency by paying a quantity of another currency.
The Forex is a de-centralised 'over the counter market' meaning that there are no agreed centres or exchanges which an investor needs to be connected to in order to trade. It is the largest world wide network allowing customers trade 24 hours per day usually from Monday to Friday. Traders can trade on Forex without any limitations no matter where they live or the time chosen to enter a trade. The accessibility of the Forex market has made it particularly popular with traders and consequently, a range of Forex trading software has been developed for internet based trading. In this paper, we report on a new indicator based on the interpretation of q computed via the Hurst exponent H that has been designed to optimize Forex trading through integration into the MetaTrader 4 system.

6. MetaTrader 4

MetaTrader 4 is a platform for e-trading that is used by online Forex traders (Metatrader 4, 2011) and provides the user with real time internet access to most of the major currency exchange rates over a range of sampling intervals including 1 min, 5 mins, 1 hour and 1 day. The system includes a built-in editor and compiler with access to a user contributed free library of software, articles and help. The software utilizes a proprietary scripting language, MQL4 (MQL4, 2011) (based on C), which enables traders to develop Expert Advisors, custom indicators and scripts. MetaTrader's popularity largely stems from its support of algorithmic trading. This includes a range of indicators and the focus of the work reported in this paper, i.e. the incorporation of a new indicator based on the approach considered in this paper.

6.1 Basic algorithm - the 'q-algorithm'

Given a stream of Forex data $u_n, \quad n = 1, 2, ..., N$ where N defines the 'look-back' window or 'period', we consider the Hurst model

$$u_n = cn^H$$

which is linearised by taking the logarithmic transform to give

$$\log(u_n) = \log(c) + H \log(n)$$

where c is a constant of proportionality
The basic algorithm is as follows:

1. For a moving window of length N (moved one element at a time) operating on an array of length L, compute $q_j = 1 + 2H_j, \quad j = 1, 2, ..., L - N$ using the Orthogonal Linear Regression Algorithm (Regression, 2011) and plot the result.

2. For a moving window of length M compute the moving average of q_j denoted by $\langle q_j \rangle_i$ and plot the result in the same window as the plot of q_j.
3. Compute the gradient of $\langle q_j \rangle_i$ using a different user defined moving average window of length K and a forward differencing scheme and plot the result.
4. Compute the second gradient of $\langle q_j \rangle_i$ after applying a moving average filter using a centre differencing scheme and plot the result in the same window.

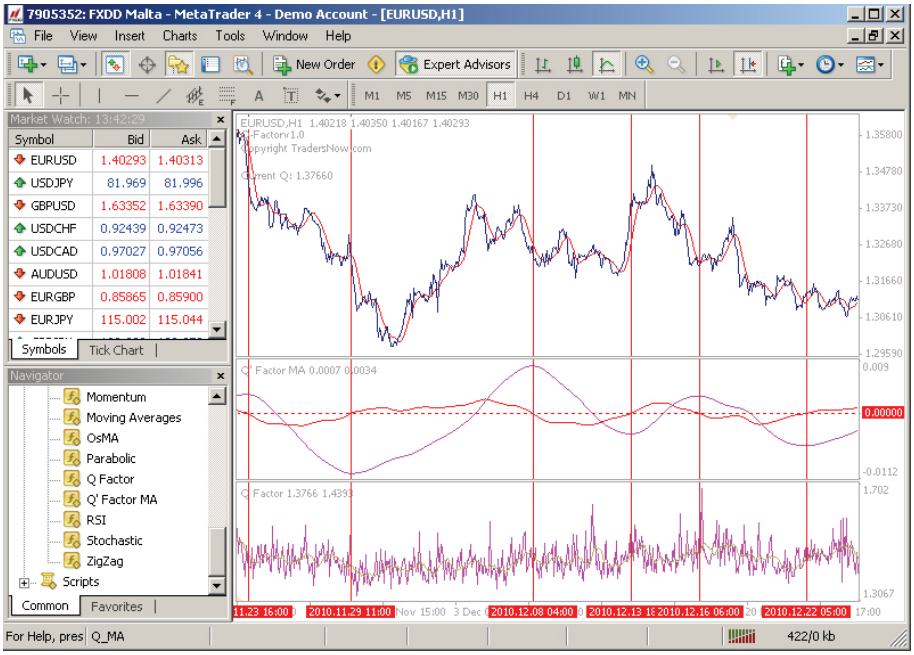


Fig. 4. MetaTrader 4 GUI for new indicators. Top window: Euro-USD exchange rate signal for 1 hour sampled data (blue) and averaged data (red); Centre window: first (red) and second (cyan) gradients of the moving average for $(N, M, K, T) = (512, 10, 100, 0)$. Bottom window: q_j (cyan) and moving average of q_j (Green).

6.2 Fundamental observations

The gradient of $\langle q_j \rangle_i$ denoted by $\langle q_j \rangle_i'$ provides an assessment of the point in time at which a trend is likely to occur, in particular, the points in time at which $\langle q_j \rangle_i'$ crosses zero. The principal characteristic is compounded in the following observation: $\langle q_j \rangle_i' > 0$ tends to

correlates with an upward trend
$\langle q_j \rangle_i' < 0$ tends correlates with a downward trend

where a change in the polarity of $\langle q_j \rangle_i' < 0$ indicates a change in the trend subject to a given tolerance T. A tolerance zone is therefore established $\mid \langle q_j \rangle_i' \mid \in T$ such that if the signal $\langle q_j \rangle_i' > 0$ enters the tolerance zone, then a bar is plotted indicating the end of an upward trend and if $\langle q_j \rangle_i' < 0$ enters the tolerance zone then a bar is plotted indicating the end of a downward trend.

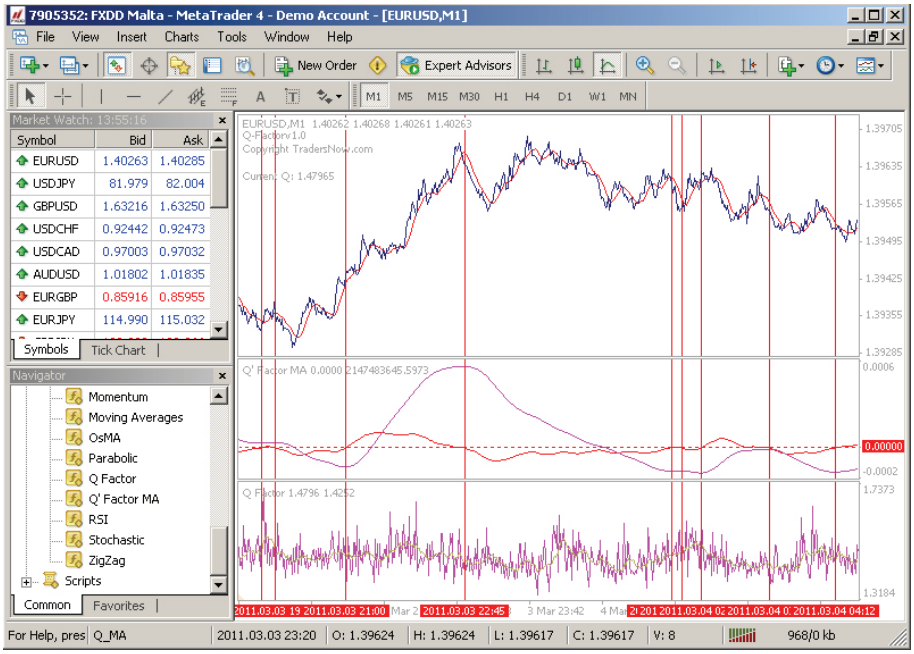


Fig. 5. MetaTrader 4 GUI for new indicators. Top window: Euro-USD exchange rate signal for 1 minute sampled data (blue) and averaged data (red); Centre window: first (red) and second (cyan) gradients of the moving average for $(N, M, K, T) = (512, 10, 100, 0)$. Bottom window: q_j (cyan) and moving average of q_j (Green).

The term 'tends' used above depends on the data and the parameter settings used to process it, in particular, the length of the look-back window used to compute q_j and the size of the window used to compute the moving average. In other words the correlations that are observed are not perfect in all cases and the algorithm therefore needs to be optimised by back-testing and live trading.

The second gradient is computed to provide an estimate of the 'acceleration' associated with the moving average characteristics of q_j denoted by $\langle q_j \rangle_i''$. This parameter tends to correlate with the direction of the trends that occur and therefore provides another indication of the direction in which the markets are moving (the position in time at which the second gradient changes direction occurs at the same point in time at which the first gradient passes through zero). Both the first and second gradients are filtered using a moving average filter to provide a smooth signal.

6.3 Examples results

Figure 4 shows an example of the MetaTrader GUI with the new indicators included operating on the signal for the Euro-USD exchange rate with 1 hour sampled data. The vertical bars clearly indicate the change in a trend for the window of data provided in this example. The parameters settings (N, M, K, T) for this example are $(512, 10, 100, 0)$. Figure 5 shows a sample of results for the Euro-USD exchange rate for 1 minute sampled data with parameter settings

using the same parameter settings In each case, a change in the gradient tends to correlates with a change in the trend of the time series in a way that is reproducible at all scales.

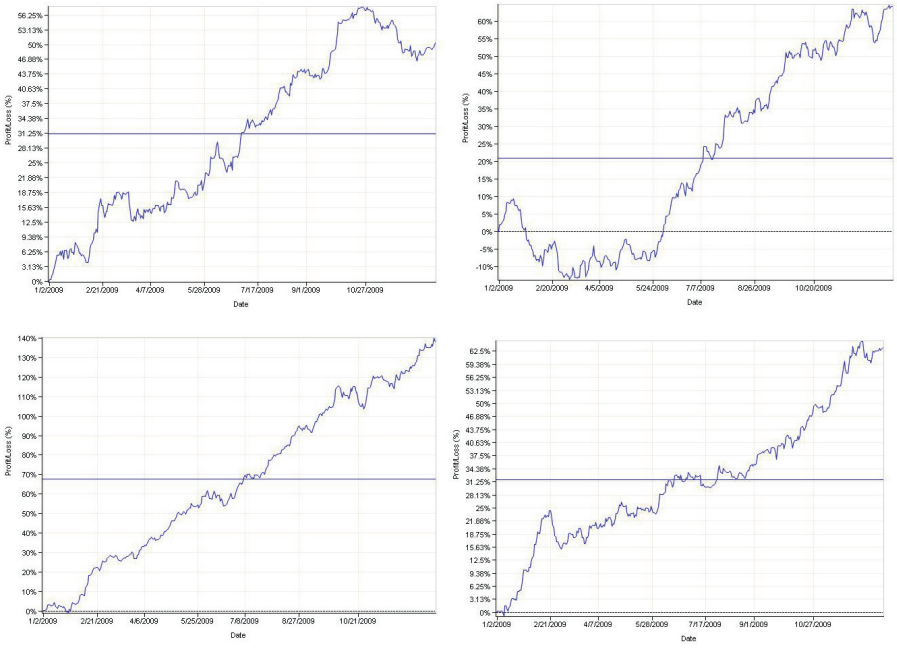


Fig. 6. Example of back-testing the 'q-algorithm'. The plots show Cumulative Profit Reports for four different currency pairs working with 1 hour sampled data from 1/1/2009 - 12/31/2009. Top-left: Euro-USD; Top-right: GGP-JPY; Bottom-left: USD-CAD; Bottom-right: UDSJPY.

Figure 6 shows examples of Cumulative Profit Reports using the 'q-algorithm' based on trading with four different currencies. The profit margins range from 50%-140% which provides evidence for the efficiency of the algorithm based on back-testing examples of this type undertaken to date.

7. Discussion

For Forex data $q(t)$ varies between 1 and 2 as does γ for q in this range since $\gamma^{-1}(t) = q(t)/2$. As the value of q increases, the Lévy index decreases and the tail of the data therefore gets longer. Thus as $q(t)$ increases, so does the likelihood of a trend occurring. In this sense, $q(t)$ provides a measure on the behaviour of an economic time series in terms of a trend (up or down) or otherwise. By applying a moving average filter to $q(t)$ to smooth the data, we obtained a signal $\langle q(t)\rangle(\tau)$ that provides an indication of whether a trend is occurring in the data over a user defined window (the period). This observation reflects a result that is a fundamental kernel of the Fractal Market Hypothesis, namely, that a change in the Lévy index precedes a change in the financial signal from which the index has been computed (from past data). In order to observe this effect more clearly, the gradient $\langle q(t)\rangle'(\tau)$ is taken. This provides the user with a clear indication of a future trend based on the following observation:

if $\langle q(t)\rangle'(\tau) > 0$, the trend is positive; if $\langle q(t)\rangle'(\tau) < 0$, the trend is negative; if $\langle q(t)\rangle'(\tau)$ passes through zero a change in the trend may occur. By establishing a tolerance zone associated with a polarity change in $\langle q(t)\rangle'(\tau)$, the importance of any indication of a change of trend can be regulated in order to optimise a buy or sell order. This is the principle basis and rationale for the 'q-algorithmŠ.

8. Conclusion

The Fractal Market Hypothesis has many conceptual and quantitative advantages over the Efficient Market Hypothesis for modelling and analysing financial data. One of the most important points is that the Fractal Market Hypothesis is consistent with an economic time series that include long tails in which rare but extreme events may occur and, more commonly, trends evolve. In this paper we have focused on the use of the Hypothesis for modelling Forex data and have shown that by computing the Hurst exponent, an algorithm can be designed that appears to accurately predict the upward and downward trends in Forex data over a range of scales subject to appropriate parameter settings and tolerances. The optimisation of these parameters can be undertaken using a range of back-testing trials to develop a strategy for optimising the profitability of Forex trading. In the trials undertaken to date, the system can generate a profitable portfolio over a range of currency exchange rates involving hundreds of Pips[3] and over a range of scales providing the data is consistent and not subject to market shocks generated by entirely unpredictable effects that have a major impact on the markets. This result must be considered in the context that the Forex markets are noisy, especially over smaller time scales, and that the behaviour of these markets can, from time to time, yield a minimal change of Pips when $\langle q(t)\rangle'(\tau)$ is within the tolerance zone establish for a given currency pair exchange rate.

The use of the indicators discussed in this paper for Forex trading is an example of a number of intensive applications and services being developed for financial time series analysis and forecasting. MetaTrader 4 is just one of a range of financial risk management systems that are being used by the wider community for de-centralised market trading, a trend that is set to increase throughout the financial services sector given the current economic environment. The current version of MetaTrader 4 described in this paper is undergoing continuous improvements and assessment, details of which can be obtained from TradersNow.com.

9. Acknowledgment

Professor J M Blackledge is supported by the Science Foundation Ireland and Mr K Murphy is supported by Currency Traders Ireland through Enterprise Ireland. Both authors are grateful to Dublin Institute of Technology and to the Institute's 'Hothouse' for its support with regard to Licensing the Technology and undertaking the arrangements associated with the commercialisation of the Technology to License described in (Hothouse, 2011) and (Video, 2001). The results given in Figure 6 were generated by Shaun Overton, One Step Removed (Custome Programming for Traders) Info@onestepremoved.com

10. References

Information and Communications Security Research Group (2011) `http://eleceng.dit.ie/icsrg`

[3] A Pip (Percentage in point) is the smallest price increment in Forex trading.

Currency Traders Ireland (2011), http://www.tradersnoe.com
Copeland, T. R., Weston, J. F. and Shastri, K. (2003), *Financial Theory and Corporate Policy*, 4th Edition, Addison Wesley.
Martin, J. D., Cox, S. H., McMinn, R. F. and Maminn, R. D. (1997), *The Theory of Finance: Evidence and Applications*, International Thomson Publishing.
Menton, R. C. (1992), *Continuous-Time Finance*, Blackwell Publishers.
Watsham, T. J. and Parramore K., (1996), *Quantitative Methods in Finance*, Thomson Business Press.
Fama, E. (1965), *The Behavior of Stock Market Prices*, Journal of Business Vol. 38, 34-105.
Samuelson, P. (1965), *Proof That Properly Anticipated Prices Fluctuate Randomly*, Industrial Management Review Vol. 6, 41-49.
Fama, E. (1970), *Efficient Capital Markets: A Review of Theory and Empirical Work*, Journal of Finance Vol. 25, 383-417.
Burton, G. M. (1987), *Efficient Market Hypothesis*, The New Palgrave: A Dictionary of Economics, Vol. 2, 120-123.
Black, F. and Scholes M. (1973) *The Pricing of Options and Corporate Liabilities*, Journal of Political Economy, Vol. 81, No. 3, 637-659.
Hurst, H. (1951), *Long-term Storage Capacity of Reservoirs*, Trans. of American Society of Civil Engineers, Vol. 116, 770-808.
Mandelbrot, B. B. and Wallis J. R. (1969), *Robustness of the Rescaled Range R/S in the Measurement of Noncyclic Long Run Statistical Dependence*, Water Resources Research 1969, Vol. 5, No. 5, 967-988.
Mandelbrot, B. B. (1972), *Statistical Methodology for Non-periodic Cycles: From the Covariance to R/S Analysis*, Annals of Economic and Social Measurement, Vol. 1, No. 3, 259-290.
Shlesinger, M. F., Zaslavsky, G. M. and Frisch U. (Eds.), (1994), *Lévy Flights and Related Topics in Physics*, Springer.
Nonnenmacher T. F., *Fractional Integral and Differential Equations for a Class of Lévy-type Probability Densities*, J. Phys. A: Math. Gen., Vol. 23, L697S-L700S.
Abea, S. and Thurnerb, S. (2005), *Anomalous Diffusion in View of EinsteinŠs 1905 Theory of Brownian Motion*, Physica, A(356), Elsevier, 403-407.
Blackledge, J. M., *Application of the Fractional Diffusion Equation for Predicting Market Behaviour*, IAENG International Journal of Applied Mathematics, Vol. 40, No. 3, 130 -158.
MetaTrader 4 Trading Platform, http://www.metaquotes.net/en/metatrader4
MQL4 Documentation, http://docs.mql4.com/
Nonlinear Regression and Curve Fitting: Orthogonal Regression, http://www.nlreg.com/orthogonal.htm
Hothouse, ICT Technologies to License, http://www.dit.ie/hothouse/media/dithothouse/techtolicensepdf/Financial%20Risk%20Management.pdf
Hothouse ICT Video Series, http://www.dit.ie/hothouse/technologiestolicence/videos/ictvideos/

8

Efficient Hedging as Risk-Management Methodology in Equity-Linked Life Insurance

Alexander Melnikov[1] and Victoria Skornyakova[2]
[1]*University of Alberta*
[2]*Workers' Compensation Board-Alberta*
Canada

1. Introduction

Using hedging methodologies for pricing is common in financial mathematics: one has to construct a financial strategy that will exactly replicate the cash flows of a contingent claim and, based on the law of one price[1], the current price of the contingent claim will be equal to the price of the replicating strategy. If the exact replication is not possible, a financial strategy with a payoff "close enough" (in some probabilistic sense) to that of the contingent claim is sought. The presence of budget constraints is one of the examples precluding the exact replication.

There are several approaches used to hedge contingent claims in the most effective way when the exact replication is not possible. The theory of efficient hedging introduced by Fölmer and Leukert (Fölmer & Leukert, 2000) is one of them. The main idea behind it is to find a hedge that will minimize the expected shortfall from replication where the shortfall is weighted by some loss function. In our paper we apply the efficient hedging methodology to equity-linked life insurance contracts to get formulae in terms of the parameters of the initial model of a financial market. As a result risk-management of both types of risks, financial and insurance (mortality), involved in the contracts becomes possible.

Historically, life insurance has been combining two distinct components: an amount of benefit paid and a condition (death or survival of the insured) under which the specified benefit is paid. As opposed to traditional life insurance paying fixed or deterministic benefits, equity-linked life insurance contracts pay stochastic benefits linked to the evolution of a financial market while providing some guarantee (fixed, deterministic or stochastic) which makes their pricing much more complicated. In addition, as opposed to pure financial instruments, the benefits are paid only if certain conditions on death or survival of insureds are met. As a result, the valuation of such contracts represents a challenge to the insurance industry practitioners and academics and alternative valuation techniques are called for. This paper is aimed to make a contribution in this direction.

Equity-linked insurance contracts have been studied since their introduction in 1970's. The first papers using options to replicate their payoffs were written by Brennan and Schwartz (Brennan & Schwartz, 1976, 1979) and Boyle and Schwartz (Boyle & Schwartz, 1977). Since

[1] The law of one price is a fundamental concept of financial mathematics stating that two assets with identical future cash flows have the same current price in an arbitrage-free market.

then, it has become a conventional practice to reduce such contracts to a call or put option and apply perfect (Bacinello & Ortu, 1993; Aase & Person, 1994) or mean-variance hedging (Möller, 1998, 2001) to calculate their price. All the authors mentioned above had studied equity-linked pure endowment contracts providing a fixed or deterministic guarantee at maturity for a survived insured. The contracts with different kind of guarantees, fixed and stochastic, were priced by Ekern and Persson (Ekern & Persson, 1996) using a fair price valuation technique.

Our paper is extending the great contributions made by these authors in two directions: we study equity-linked life insurance contracts with a stochastic guarantee[2] and we use an imperfect hedging technique (efficient hedging). Further developments may include an introduction of a stochastic model for interest rates and a systematic mortality risk, a combination of deterministic and stochastic guarantees, surrender options and lapses etc.

We consider equity-linked pure endowment contracts. In our setting a financial market consists of a non-risky asset and two risky assets. The first one, S_t^1, is more risky and profitable and provides possible future gain. The second asset, S_t^2, is less risky and serves as a stochastic guarantee. Note that we restrict our attention to the case when evolutions of the prices of the two risky assets are generated by the same Wiener process, although the model with two different Wiener processes with some correlation coefficient ρ between them, as in Margrabe, 1978, could be considered. There are two reasons for our focus. First of all, equity-linked insurance contracts are typically linked to traditional equities such as traded indices and mutual funds which exhibit a very high positive correlation. Therefore, the case when $\rho = 1$ could be a suitable and convenient approximation. Secondly, although the model with two different Wiener processes seems to be more general, it turns out that the case $\rho = 1$ demands a special consideration and does not follow from the results for the case when $\rho < 1$ (see Melnikov & Romaniuk, 2008; Melnikov, 2011 for more detailed information on a model with two different Wiener processes). The case $\rho = -1$ does not seem to have any practical application although could be reconstructed for the sake of completeness. Note also that our setting with two risky assets generated by the same Wiener process is equivalent to the case of a financial market consisting of one risky asset and a stochastic guarantee being a function of its prices.

We assume that there are no additional expenses such as transaction costs, administrative costs, maintenance expenses etc. The payoff at maturity is equal to $\max\left(S_T^1, S_T^2\right)$. We reduce it to a call option giving its holder the right to exchange one asset for another at maturity. The formula for the price of such options was given in Margrabe, 1978. Since the benefit is paid on survival of a client, the insurance company should also deal with some mortality risk. As a result, the price of the contract will be less than needed to construct a perfect hedge exactly replicating the payoff at maturity. The insurance company is faced with an initial budget constraint precluding it from using perfect hedging. Therefore, we fix the probability of the shortfall arising from a replication and, with a known price of the contract, control of financial and insurance risks for the given contract becomes possible.

[2] Although Ekern & Persson, 1996, consider a number of different contracts including those with a stochastic guarantee, the contracts under our consideration differ: we consider two risky assets driven by the same Wiener process or, equivalently, one risky asset and a stochastic guarantee depending on its price evolution. The motivation for our choice follows below.

The layout of the paper is as follows. Section 2 introduces the financial market and explains the main features of the contracts under consideration. In Section 3 we describe efficient hedging methodology and apply it to pricing of these contracts. Further, Section 4 is devoted to a risk-taking insurance company managing a balance between financial and insurance risks. In addition, we consider how the insurance company can take advantage of diversification of a mortality risk by pooling homogeneous clients together and, as a result of more predictable mortality exposure, reducing prices for a single contract in a cohort. Section 5 illustrates our results with a numerical example.

2. Description of the model

2.1 Financial setting

We consider a financial market consisting of a non-risky asset $B_t = \exp(rt), t \geq 0, r \geq 0$, and two risky assets S^1 and S^2 following the Black-Scholes model:

$$dS_t^i = S_t^i(\mu_i dt + \sigma_i dW_t),\ i = 1,2,\ t \leq T. \tag{1}$$

Here μ_i and σ_i are a rate of return and a volatility of the asset S^i, $W = (W_t)_{t \leq T}$ is a Wiener process defined on a standard stochastic basis $(\Omega, F, \mathbf{F} = (F_t)_{t \leq T}, P)$, T – time to maturity. We assume, for the sake of simplicity, that $r = 0$, and, therefore, $B_t = 1$ for any t. Also, we demand that $\mu_1 > \mu_2, \sigma_1 > \sigma_2$. The last two conditions are necessary since S^2 is assumed to provide a flexible guarantee and, therefore, should be less risky than S^1. The initial values for both assets are supposed to be equal $S_0^1 = S_0^2 = S_0$ and are considered as the initial investment in the financial market.

It can be shown, using the Ito formula, that the model (1) could be presented in the following form:

$$S_t^i = S_0^i \exp\left\{\left(\mu_i - \frac{\sigma_i^2}{2}\right)t + \sigma_i W_t\right\} \tag{2}$$

Let us define a probability measure P^* which has the following density with respect to the initial probability measure P:

$$Z_T = \exp\left\{-\frac{\mu_1}{\sigma_1}W_T - \frac{1}{2}\left(\frac{\mu_1}{\sigma_1}\right)^2 T\right\}. \tag{3}$$

Both processes, S^1 and S^2, are martingales with respect to the measure P^* if the following technical condition is fulfilled:

$$\frac{\mu_1}{\sigma_1} = \frac{\mu_2}{\sigma_2} \tag{4}$$

Therefore, in order to prevent the existence of arbitrage opportunities in the market we suppose that the risky assets we are working with satisfy this technical condition. Further, according to the Girsanov theorem, the process $W_t^* = W_T + \frac{\mu_1}{\sigma_1}t = W_T + \frac{\mu_2}{\sigma_2}t$

is a Wiener process with respect to P^*.

Finally, note the following useful representation of the guarantee S_t^2 by the underlying risky asset S_t^1:

$$\begin{aligned} S_t^2 &= S_0^2 \exp\left\{\sigma_2 W_t + \left(\mu_2 - \frac{\sigma_2^2}{2}\right)t\right\} \\ &= S_0 \exp\left\{\frac{\sigma_2}{\sigma_1}\left(\sigma_1 W_t + \left(\mu_1 - \frac{\sigma_1^2}{2}\right)t\right) - \frac{\sigma_2}{\sigma_1}\left(\mu_1 - \frac{\sigma_1^2}{2}\right)t + \left(\mu_2 - \frac{\sigma_2^2}{2}\right)t\right\} \\ &= \left(S_0\right)^{1-\sigma_2/\sigma_1}\left(S_t^1\right)^{\sigma_2/\sigma_1} \exp\left\{-\frac{\sigma_2}{\sigma_1}\left(\mu_1 - \frac{\sigma_1^2}{2}\right)t + \left(\mu_2 - \frac{\sigma_2^2}{2}\right)t\right\}, \end{aligned}$$

which shows that our setting is equivalent to one with a financial market consisting of a single risky asset and a stochastic guarantee being a function of the price of this asset.

We will call any process $\pi_t = \left(\beta_t, \gamma_t^1, \gamma_t^2\right)_{t\geq 0}$, adapted to the price evolution F_t, a *strategy*. Let us define its value as a sum $X_t^\pi = \beta_t + \gamma_t^1 S_t^1 + \gamma_t^2 S_t^2$. We shall consider only *self-financing* strategies satisfying the following condition $dX_t^\pi = \beta_t + \gamma_t^1 dS_t^1 + \gamma_t^2 dS_t^2$, where all stochastic differentials are well defined. Every F_T-measurable nonnegative random variable H is called a contingent claim. A self-financing strategy π is a *perfect hedge* for H if $X_T^\pi \geq H$ (a.s.). According to the option pricing theory of Black-Scholes-Merton, it does exist, is unique for a given contingent claim, and has an initial value $X_0^\pi = E^* H$.

2.2 Insurance setting

The insurance risk to which the insurance company is exposed when enters into a pure endowment contract includes two components. The first one is based on survival of a client to maturity as at that time the insurance company would be obliged to pay the benefit to the alive insured. We call it a *mortality risk*. The second component depends on a *mortality frequency risk* for a pooled number of similar contracts. A large enough portfolio of life insurance contracts will result in more predictable mortality risk exposure and a reduced mortality frequency risk. In this section we will work with the mortality risk only dealing with the mortality frequency risk in Section 4.

Following actuarial tradition, we use a random variable $T(x)$ on a probability space $\left(\tilde{\Omega}, \tilde{F}, \tilde{P}\right)$ to denote the remaining lifetime of a person of age x. Let ${}_T p_x = \tilde{P}\{T(x) > T\}$ be a survival probability for the next T years of the same insured. It is reasonable to assume that $T(x)$ doesn't depend on the evolution of the financial market and, therefore, we consider (Ω, F, P) and $\left(\tilde{\Omega}, \tilde{F}, \tilde{P}\right)$ as being independent.

We study pure endowment contracts with a flexible stochastic guarantee which make a payment at maturity provided the insured is alive. Due to independency of "financial" and "insurance" parts of the contract we consider the product probability space $\left(\Omega \times \tilde{\Omega}, F \times \tilde{F}, P \times \tilde{P}\right)$ and introduce a contingent claim on it with the following payoff at maturity:

$$H\left(T(x)\right) = \max\left\{S_T^1, S_T^2\right\} \cdot I_{\{T(x)>T\}}. \tag{5}$$

It is obvious that a strategy with the payoff $H = \max\left\{S_T^1, S_T^2\right\}$ at T is a perfect hedge for the contract under our consideration. Its price is equal to E^*H.

2.3 Optimal pricing and hedging

Let us rewrite the financial component of (5) as follows:

$$H = \max\left\{S_T^1, S_T^2\right\} = S_T^2 + \left(S_T^1 - S_T^2\right)^+, \tag{6}$$

where $x^+ = \max(0, x)$, $x \in R^1$. Using (2.6) we reduce the pricing of the claim (5) to the pricing of the call option $\left(S_T^1 - S_T^2\right)^+$ provided $\left\{T(x) > T\right\}$.

According to the well-developed option pricing theory the optimal price is traditionally calculated as an expected present value of cash flows under a risk-neutral probability measure. Note, however, that the "insurance" part of the contract (5) doesn't need to be risk-adjusted since the mortality risk is essentially unsystematic. It means that the mortality risk can be effectively reduced not by hedging but by diversification or by increasing the number of similar insurance policies.

Proposition. The price for the contract (5) is equal to

$$ {}_TU_x = E^* \times \tilde{E}H\left(T(x)\right) = {}_Tp_xE^*\left(S_T^2\right) + {}_Tp_xE^*\left(S_T^1 - S_T^2\right)^+, \tag{7}$$

where $E^* \times \tilde{E}$ is the expectation with respect to $P^* \times \tilde{P}$.

We would like to call (7) as the *Brennan-Schwartz price* (Brennan & Schwartz, 1976).

The insurance company acts as a hedger of H in the financial market. It follows from (7) that the initial price of H is strictly less than that of the perfect hedge since a survival probability is always less than one or

$$ {}_TU_x < E^*\left(S_T^2 + \left(S_T^1 - S_T^2\right)^+\right) = E^*H.$$

Therefore, perfect hedging of H with an initial value of the hedge restricted by the Black-Scholes-Merton price E^*H is not possible and alternative hedging methods should be used. We will look for a strategy π^* with some initial budget constraint such that its value $X_T^{\pi^*}$ at maturity is close to H in some probabilistic sense.

3. Efficient hedging

3.1 Methodology

The main idea behind efficient hedging methodology is the following: we would like to construct a strategy π, with the initial value

$$X_0^\pi \le X_0 < E^*H, \tag{8}$$

that will minimize the expected shortfall from the replication of the payoff H. The shortfall is weighted by some loss function $l: R_+ \to R_+ = [0, \infty)$. We will consider a power loss function $l(x) = const \cdot x^p, p > 0, x \ge 0$ (Fölmer & Leukert, 2000). Since at maturity of the contract X_T^π

should be close to H in some probabilistic sense we will consider $El\left(\left(H - X_T^{\pi}\right)^+\right)$ as a measure of closeness between X_T^{π} and H.

Definition. Let us define a strategy π^* for which the following condition is fulfilled:

$$El\left(\left(H - X_T^{\pi^*}\right)^+\right) = \inf_{\pi} El\left(\left(H - X_T^{\pi}\right)^+\right), \tag{9}$$

where infimum is taken over all self-financing strategies with positive values satisfying the budget restriction (8). The strategy π^* is called the *efficient hedge*.

Ones the efficient hedge is constructed we will set the price of the equity-linked contract (5) being equal to its initial value $X_0^{\pi^*}$ and make conclusions about the appropriate balance between financial and insurance risk exposure.

Although interested readers are recommended to get familiar with the paper on efficient hedging by Fölmer & Leukert, 2000, for the sake of completeness we formulate the results from it that are used in our paper in the following lemma.

Lemma 1. Consider a contingent claim with the payoff (6) at maturity with the shortfall from its replication weighted by a power loss function

$$l(x) = const \cdot x^p, p > 0, x \geq 0. \tag{10}$$

Then the efficient hedge π^* satisfying (9) exists and coincides with a perfect hedge for a modified contingent claim H_p having the following structure:

$$\begin{aligned} H_p &= H - a_p Z_T^{1/(p-1)} \wedge H \quad \text{for } p > 1,\ const = 1/p, \\ H_p &= H \cdot I_{\left\{Z_T^{-1} > a_p H^{1-p}\right\}} \quad \text{for } 0 < p < 1,\ const = 1, \\ H_p &= H \cdot I_{\left\{Z_T^{-1} > a_p\right\}} \quad \text{for } p = 1,\ const = 1, \end{aligned} \tag{11}$$

where a constant a_p is defined from the condition on its initial value $E^* H_p = X_0$.

In other words, we reduce a construction of an efficient hedge for the claim H from (9) to an easier-to-do construction of a perfect hedge for the modified claim (11). In the next section we will apply efficient hedging to equity-linked life insurance contracts.

3.2 Application to equity-linked life insurance contracts

Here we consider a single equity-linked life insurance contract with the payoff (5). Since (6) is true, we will pay our attention to the term $\left(S_T^1 - S_T^2\right)^+ \cdot I_{\{T(x) > T\}}$ associated with a call option. Note the following equality that comes from the definition of perfect and efficient hedging and Lemma 1:

$$X_0 = {}_T p_x E^*\left(S_T^1 - S_T^2\right)^+ = E^*\left(S_T^1 - S_T^2\right)_{p, p>0}^+ \tag{12}$$

where $\left(S_T^1 - S_T^2\right)_p^+$ is defined by (11). Using (12) we can separate insurance and financial components of the contract:

$$ {}_T p_x = \frac{E^*\left(S_T^1 - S_T^2\right)_p^+}{E^*\left(S_T^1 - S_T^2\right)^+}. \tag{13} $$

The left-hand side of (13) is equal to the survival probability of the insured, which is a mortality risk for the insurer, while the right-hand side is related to a pure financial risk as it is connected to the evolution of the financial market. So, the equation (13) can be viewed as a *key balance equation* combining the risks associated with the contract (5).

We use efficient hedging methodology presented in Lemma 1 for a further development of the numerator of the right-hand side of (13) and the Margrabe formula (Margrabe, 1978) for its denominator.

Step 1. Let us first work with the denominator of the right-hand side of (13). We get

$$ E^*\left(S_T^1 - S_T^2\right)^+ = S_0\left\{\Phi\left(b_+(1,1,T)\right) - \Phi\left(b_-(1,1,T)\right)\right\}, \tag{14} $$

where $b_\pm(1,1,T) = \dfrac{\ln 1 \pm \left(\sigma_1 - \sigma_2\right)^2 \dfrac{T}{2}}{\left(\sigma_1 - \sigma_2\right)\sqrt{T}}$, $\Phi(x) = \left(1/\sqrt{2\pi}\right)\int\limits_{-\infty}^{x} \exp(-y^2/2)dy$.

The proof of (14) is given in Appendix. Note that (14) is a variant of the Margrabe formula (Margrabe, 1978) for the case $S_0^1 = S_0^2 = S_0$. It shows the price of the option that gives its holder the right to exchange one risky asset for another at maturity of the contract.

Step 2. To calculate the numerator of the right-hand side of (13), we want to represent it in terms of $Y_T = S_T^1/S_T^2$. Let us rewrite W_T with the help a free parameter γ in the form

$$ \begin{aligned} W_T &= (1+\gamma)W_T - \gamma W_T \\ &= \frac{1+\gamma}{\sigma_1}\left(\sigma_1 W_T + \left(\mu_1 - \frac{\sigma_1^2}{2}\right)T\right) - \frac{\gamma}{\sigma_2}\left(\sigma_2 W_T + \left(\mu_2 - \frac{\sigma_2^2}{2}\right)T\right) \\ &\quad - \frac{1+\gamma}{\sigma_1}\left(\mu_1 - \frac{\sigma_1^2}{2}\right)T + \frac{\gamma}{\sigma_2}\left(\mu_2 - \frac{\sigma_2^2}{2}\right)T. \end{aligned} \tag{15} $$

Using (3) and (15), we obtain the next representation of the density Z_T:

$$ Z_T = G \cdot \left(S_T^1\right)^{-\frac{(1+\gamma)\mu_1}{\sigma_1^2}} \left(S_T^2\right)^{\frac{\gamma\mu_1}{\sigma_1\sigma_2}} \tag{16} $$

where

$$ \begin{aligned} G = &\left(S_0^1\right)^{\frac{(1+\gamma)\mu_1}{\sigma_1^2}} \left(S_0^2\right)^{-\frac{\gamma\mu_1}{\sigma_1\sigma_2}} \\ &\times \exp\left(\frac{(1+\gamma)\mu_1}{\sigma_1^2}\left(\mu_1 - \frac{\sigma_1^2}{2}\right)T - \frac{\gamma\mu_1}{\sigma_1\sigma_2}\left(\mu_2 - \frac{\sigma_2^2}{2}\right)T - \frac{1}{2}\left(\frac{\mu_1}{\sigma_1}\right)^2 T\right). \end{aligned} $$

Now we consider three cases according to (11) and choose appropriate values of the parameter γ for each case (see Appendix for more details). The results are given in the following theorem.

Theorem 1. Consider an insurance company measuring its shortfalls with a power loss function (10) with some parameter $p > 0$. For an equity-linked life insurance contract with the payoff (5) issued by the insurance company, it is possible to balance a survival probability of an insured and a financial risk associated with the contract.

Case 1: $p > 1$

For $p > 1$ we get

$$
\begin{aligned}
{}_T p_x = {} & \frac{\Phi\left(b_+(1,C,T)\right) - \Phi\left(b_-(1,C,T)\right)}{\Phi\left(b_+(1,1,T)\right) - \Phi\left(b_-(1,1,T)\right)} \\
& + \frac{(C-1)^+}{C^{\alpha_p}} \exp\left\{ \alpha_p\left(1-\alpha_p\right)\frac{\left(\sigma_1-\sigma_2\right)^2}{2} T \right\} \times \frac{\Phi\left(b_-(1,C,T) + \alpha_p\left(\sigma_1-\sigma_2\right)\sqrt{T}\right)}{\Phi\left(b_+(1,1,T)\right) - \Phi\left(b_-(1,1,T)\right)},
\end{aligned}
\tag{17}
$$

where C is found from $a_p G^{1/(p-1)} C^{\alpha_p} = C - 1$ and $\alpha_p = -\frac{\mu_1}{\sigma_1(\sigma_1-\sigma_2)(p-1)} - \frac{\sigma_2}{\sigma_1-\sigma_2}$.

Case 2: $0 < p < 1$

Denote $\alpha_p = \frac{\sigma_1\sigma_2(1-p) - \mu_1}{\sigma_1(\sigma_1-\sigma_2)}$.

2.1. If $-\alpha_p \le 1-p$ (or $\frac{\mu_1}{\sigma_1^2} \le 1-p$) then

$$
{}_T p_x = 1 - \frac{\Phi\left(b_+(1,C,T)\right) - \Phi\left(b_-(1,C,T)\right)}{\Phi\left(b_+(1,1,T)\right) - \Phi\left(b_-(1,1,T)\right)},
\tag{18}
$$

where C is found from

$$
C^{-\alpha_p} = a_p \cdot G \cdot \left((C-1)^+\right)^{1-p}.
\tag{19}
$$

2.2. If $-\alpha_p > 1-p$ (or $\frac{\mu_1}{\sigma_1^2} > 1-p$) then

2.2.1. If (19) has no solution then ${}_T p_x = 1$.

2.2.2. If (19) has one solution C, then ${}_T p_x$ is defined by (18).

2.2.3. If (19) has two solutions $C_1 < C_2$ then

$$
{}_T p_x = 1 - \frac{\Phi\left(b_+(1,C_1,T)\right) - \Phi\left(b_-(1,C_1,T)\right)}{\Phi\left(b_+(1,1,T)\right) - \Phi\left(b_-(1,1,T)\right)} + \frac{\Phi\left(b_+(1,C_2,T)\right) - \Phi\left(b_-(1,C_2,T)\right)}{\Phi\left(b_+(1,1,T)\right) - \Phi\left(b_-(1,1,T)\right)}.
\tag{20}
$$

Case 3: $p = 1$

For $p=1$ we have

$$ {}_Tp_x = 1 - \frac{\Phi\left(b_+(1,C,T)\right) - \Phi\left(b_-(1,C,T)\right)}{\Phi\left(b_+(1,1,T)\right) - \Phi\left(b_-(1,1,T)\right)}, \tag{21} $$

where $C = \left(Ga_p\right)^{\frac{\sigma_1(\sigma_1-\sigma_2)}{\mu_1}}$ and $\alpha_p = -\frac{\mu_1}{\sigma_1(\sigma_1-\sigma_2)}$.

The proof of (17), (18), (20), and (21) is given in Appendix.

Remark 1. One can consider another approach to find C (or C_1 and C_2) for (18), (20) and (21). Let us fix a probability of the set $\{Y_T \le C\}$ (or $\{Y_T \le C_1\} \cup \{Y_T > C_2\}$):

$$ P\left(Y_T \le C\right) = 1-\varepsilon,\ \varepsilon > 0, \tag{22} $$

$$ P\left(\{Y_T \le C_1\} \cup \{Y_T > C_2\}\right) = 1-\varepsilon,\ \varepsilon > 0 $$

and calculate C (or C_1 and C_2) using log-normality of Y_T. Note that a set for which (22) is true coincides with $\left\{X_T^{\pi} \ge H\right\}$. The latter set has a nice financial interpretation: fixing its probability at $1-\varepsilon$, we specify the level of a financial risk that the company is ready to take or, in other words, the probability ε that it will not be able to hedge the claim (6) perfectly. We will explore this remark further in the next section.

4. Risk-management for risk-taking insurer

The loss function with $p>1$ corresponds to a company avoiding risk with risk aversion increasing as p grows. The case $0<p<1$ is appropriate for companies that are inclined to take some risk. In this section we show how a risk-taking insurance company could use efficient hedging for management of its financial and insurance risks. For illustrative purposes we consider the extreme case when $p \to 0$. While the effect of a power p close to zero on efficient hedging was pointed out by Föllmer and Leukert (Föllmer & Leukert, 2000), we give it a different interpretation and implementation which are better suited for the purposes of our analysis. In addition, we restrict our attention to a particular case for which the equation (19) has only one solution: that is Case 2.1. This is done for illustrative purposes only since the calculation of constants C, C_1 and C_2 for other cases may involve the use of extensive numerical techniques and lead us well beyond our research purposes.

As was mentioned above, the characteristic equation (19) with $p \le 1+\alpha_p$ (or, equivalently, $p \le 1-\frac{\mu_1}{\sigma_1^2}$) admits only one solution C which is further used for determination of a modified claim (11) as follows

$$ H_p = H \cdot I_{\{Y_T \le C\}} \tag{23} $$

where $H=\left(S_T^1-S_T^2\right)^+$, $Y_T=S_T^1/S_T^2$, and $0<p<1$. Denote an efficient hedge for H and its initial value as π^* and $x=X_0$ respectively. It follows from Lemma 1 that π^* is a perfect hedge for $H_p=\left(S_T^1-S_T^2\right)^+$.
Since the inequality $\left((a-b)^+\right)^p\le a^p$ is true for any positive a and b, we have

$$\begin{aligned}E\left(\left(H-X_T^{\pi^*}(x)\right)^+\right)^p&=E\left[\left(H_p-X_T^{\pi^*}(x)\right)^+\cdot I_{\{Y_T\le C\}}+\left(H-X_T^{\pi^*}(x)\right)^+\cdot I_{\{Y_T>C\}}\right]^p\\&=E\left[\left(H-X_T^{\pi^*}(x)\right)^+\cdot I_{\{Y_T>C\}}\right]^p\\&=E\left[\left(H-X_T^{\pi^*}(x)\right)^+\right]^p\cdot I_{\{Y_T>C\}}\le EH^p\cdot I_{\{Y_T>C\}}.\end{aligned}\tag{24}$$

Taking the limit in (24) as $p\to 0$ and applying the classical dominated convergence theorem, we obtain

$$EH^p\cdot I_{\{Y_T>C\}}\underset{p\to 0}{\to}EI_{\{Y_T>C\}}=P\left(Y_T>C\right)\tag{25}$$

Therefore, we can fix a probability $P\left(Y_Y>C\right)=\varepsilon$ which quantifies a financial risk and is equivalent to the probability of failing to hedge H at maturity.
Note that the same hedge π^* will also be an efficient hedge for the claim $\delta\cdot H$ where δ is some positive constant but its initial value will be $\delta\cdot x$ instead of x. We will use this simple observation for pricing cumulative claims below when we consider the insurance company taking advantage of diversification of a mortality risk and further reducing the price of the contract.
Here, we pool together the homogeneous clients of the same age, life expectancy and investment preferences and consider a cumulative claim $l_{x+T}\cdot H$, where l_{x+T} is the number of insureds alive at time T from the group of size l_x. Let us measure a mortality risk of the pool of the equity-linked life insurance contracts for this group with the help of a parameter $\alpha\in(0,1)$ such that

$$\tilde{P}(l_{x+T}\le n_\alpha)=1-\alpha\,,\tag{26}$$

where n_α is some constant. In other words, α equals the probability that the number of clients alive at maturity will be greater than expected based on the life expectancy of homogeneous clients. Since it follows a frequency distribution, this probability could be calculated with the help of a binomial distribution with parameters ${}_Tp_x$ and l_x where ${}_Tp_x$ is found by fixing the level of the financial risk ε and applying the formulae from Theorem 1.
We can rewrite (26) as follows

$$\tilde{P}\left(\frac{l_{x+T}}{l_x}\le\frac{n_\alpha}{l_x}\right)=\tilde{P}\left(\frac{l_{x+T}}{l_x}\le\delta\right)=1-\alpha\,,$$

where $\delta=n_\alpha/l_x$. Due to the independence of insurance and financial risks, we have

$$\begin{aligned} P\times\tilde{P}(l_x X_T^{\pi}(\delta x)\geq l_{x+T}H) &\geq P\times\tilde{P}(X_T^{\pi}(\delta x)\geq \frac{l_{x+T}}{l_x}H) \\ &\geq P(X_T^{\pi}(\delta x)\geq \delta H)\cdot\tilde{P}(n_\alpha \geq l_{x+T})\geq (1-\varepsilon)(1-\alpha)\geq 1-(\varepsilon+\alpha). \end{aligned} \tag{27}$$

So, using the strategy π^* the insurance company is able to hedge the cumulative claim $l_{x+T}\cdot H$ with the probability at least $1-(\varepsilon+\alpha)$ which combines both financial and insurance risks. The price of a single contract will be further reduced to $\frac{n_\alpha}{l_x}{}_T p_x E^* H$.

5. Numerical example

Using the same reasons as in the previous section, we restrict our attention to the case when $p\to 0$ and the equation (19) has only one solution as is in Case 2.1. Consider the following parameters for the risky assets:

$$\mu_1 = 5\%,\quad \sigma_1 = 23\%,$$

$$\mu_2 = 4\%,\quad \sigma_2 = 19\%.$$

The condition (4) is approximately fulfilled to preclude the existence of arbitrage opportunities. Also, since $1-\mu_1/\sigma_1^2 \cong 0.05$, p should be very small, or $p\leq 0.05$, and we are able to use (25) instead of (19) and exploit (18) from Theorem 1. For survival probabilities we use the Uninsured Pensioner Mortality Table UP94 (Shulman & Kelley, 1999) which is based on best estimate assumptions for mortality. Further, we assume that a single equity-linked life insurance contract has the initial value $S_0 = 100$. We consider contracts with the maturity terms $T = 5,10,15,20,25$ years. The number of homogeneous insureds in a cohort is $l_x = 100$.

Figure 1 represents the offsetting relationships between financial and insurance risks. Note that financial and insurance risks do offset each other. As perfect hedging is impossible, the insurer will be exposed to a financial risk expressed as a probability that it will be unable to hedge the claim (6) with the probability one. At the same time, the insurance company faces a mortality risk or a probability that the insured will be alive at maturity and the payment (6) will be due at that time. Combining both risks together we conclude that if the financial risk is big, the insurance company may prefer to be exposed to a smaller mortality risk. By contrast, if the claim (6) could be hedged with greater probability the insurance company may wish to increase its mortality risk exposure. Therefore, there is an offset between financial and mortality risks the insurer can play with: by fixing one of the risks, the appropriate level of another risk could be calculated.

For Figure 1 we obtained survival probabilities using (18) for different levels of a financial risk ε and found the corresponding ages for clientele using the specified mortality table. Note that whenever the risk that the insurance company will fail to hedge successfully increases, the recommended ages of the clients rise as well. As a result, the company diminishes the insurance component of risk by attracting older and, therefore, safer clientele to compensate for the increasing financial risk. Also observe that with longer contract

maturities, the company can widen its audience to younger clients because a mortality risk, which is a survival probability in our case, is decreasing over time.
Different combinations of a financial risk ε and an insurance risk α give us the range of prices for the equity-linked contracts. The results for the contracts are shown in Figure 2.

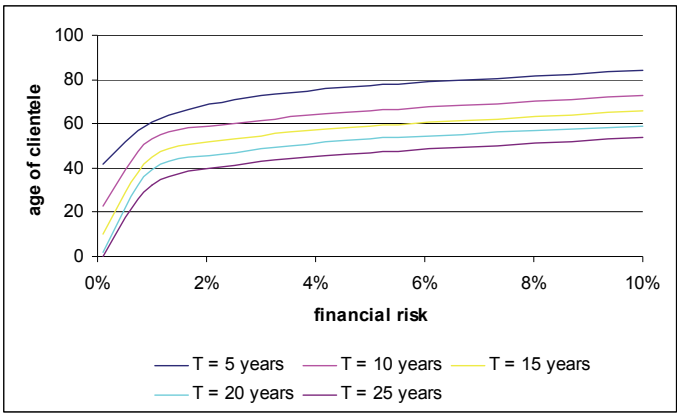


Fig. 1. Offsetting financial and mortality risks

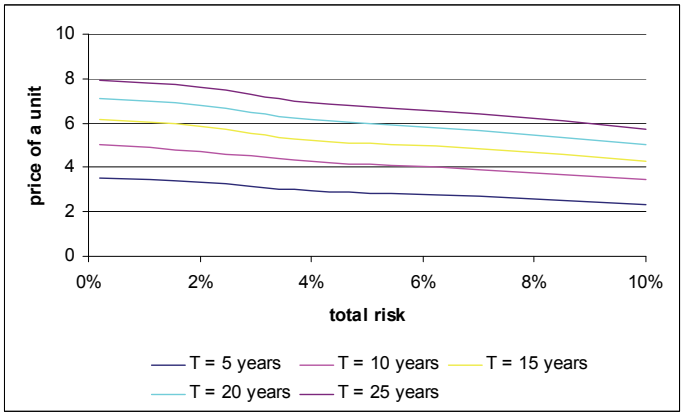


Fig. 2. Prices of $100 invested in equity-linked life insurance contracts

The next step is to construct a grid that enables the insurance company to identify the acceptable level of the financial risk for insureds of any age. We restrict our attention to a group of clients of ages 30, 40, 50, and 60 years. The results are presented in Table 1. The financial risk found reflects the probability of failure to hedge the payoff that will be offset by the mortality risk of the clients of a certain age.

Age of clients	$T=5$	$T=10$	$T=15$	$T=20$	$T=25$
30	0.05%	0.13%	0.25%	0.45%	0.8%
40	0.1%	0.25%	0.55%	1.2%	2.3%
50	0.2%	0.7%	1.8%	3.7%	7%
60	0.8%	2.5%	5.5%	10.5%	18.5%

Table 1. Acceptable Financial Risk Offsetting Mortality Risk of Individual Client

Age of clients	$T=5$	$T=10$	$T=15$	$T=20$	$T=25$
30	3.45	4.86	5.87	6.66	7.22
40	3.45	4.79	5.69	6.25	6.45
50	3.39	4.56	5.11	5.10	4.53
60	3.17	3.84	3.76	2.99	1.70
Margrabe price	3.57	5.04	6.17	7.13	7.97

Table 2. Prices of contracts with cumulative mortality risk $\alpha = 2.5\%$

Prices of the contracts for the same group of clients are given in Table 2. Note that the price of a contract is a function of financial and insurance risks associated with this contract. The level of the insurance risk is chosen to be $\alpha = 2.5\%$. In the last row, the Margrabe prices are compared with reduced prices of equity-linked contracts. The reduction in prices was possible for two reasons: we took into account the mortality risk of an individual client (the probability that the client would not survive to maturity and, therefore, no payment at maturity would be made) and the possibility to diversify the cumulative mortality risk by pooling homogeneous clients together.

6. Appendix

6.1 Proof of (14)

Let $Y_T = S_T^1 / S_T^2$. Then we have

$$\begin{aligned} E^*\left(S_T^1 - S_T^2\right)^+ &= E^*\left(S_T^1 - S_T^2\right)\cdot I_{\{Y_T>1\}} \\ &= E^* S_T^1 - E^* S_T^2 - E^*\left(S_T^1 - S_T^2\right)\cdot I_{\{Y_T\le 1\}}. \end{aligned} \tag{28}$$

Since S^1, S^2 are martingales with respect to P^* , we have $E^* S_T^i = S_0^i = S_0$, i=1,2. For the last term in (28), we get

$$E^* S_T^i \cdot I_{\{Y_T\le 1\}} = E^* \exp\{-\eta_i\}\cdot I_{\{\xi\le \ln 1\}} \tag{29}$$

where $\eta_i = -\ln S_T^i$, $\xi = \ln Y_T$ are Gaussian random variables. Using properties of normal random variables (Melnikov, 2011) we find that

$$E^*\exp\{-\eta_i\}\cdot I_{\{\xi\le\ln 1\}}=\exp\left\{\frac{\sigma_{\eta_i}^2}{2}-\mu_{\eta_i}\right\}\Phi\left(\frac{\ln 1-\left(\mu_\xi-\operatorname{cov}(\xi,\eta_i)\right)}{\sigma_\xi}\right) \tag{30}$$

where $\mu_{\eta_i}=E^*\eta_i,\sigma_{\eta_i}^2=\operatorname{var}(\eta_i),\mu_\xi=E^*\xi,\sigma_\xi^2=\operatorname{var}\xi$, $\Phi(x)=(1/\sqrt{2\pi})\int_{-\infty}^{x}\exp(-y^2/2)dy$.

Using (29), (30), we arrive at (14).

6.2 Proof of (17)

According to (16), we have

$$\begin{aligned}Z_T^{1/(p-1)}\cdot\left(S_T^2\right)^{-1}&=G^{1/(p-1)}\cdot\left(S_T^1\right)^{-\frac{(1+\gamma)\mu_1}{\sigma_1^2(p-1)}}\left(S_T^2\right)^{-1+\frac{\gamma\mu_1}{\sigma_1\sigma_2(p-1)}}\\&=G^{1/(p-1)}\cdot Y_T^{\alpha_p}\end{aligned} \tag{31}$$

with

$$\alpha_p=-\frac{(1+\gamma)\mu_1}{\sigma_1^2(p-1)}=1-\frac{\gamma\mu_1}{\sigma_1\sigma_2(p-1)}. \tag{32}$$

Equation (32) has the unique solution

$$\gamma=\gamma_p=\frac{\sigma_1^2\sigma_2(p-1)+\mu_1\sigma_2}{\mu_1(\sigma_1-\sigma_2)}. \tag{33}$$

It follows from (33) that $\gamma_p>0$ and, therefore, from (32) we conclude that $\alpha_p<0$ and the equation

$$a_pG^{1/(p-1)}y^{\alpha_p}=(y-1)^+,\ y\ge 1 \tag{34}$$

has the unique solution $C=C(p)\ge 1$. Using (31)-(34), we represent $\left(S_T^1-S_T^2\right)_p^+$ as follows

$$\begin{aligned}\left(S_T^1-S_T^2\right)_p^+&=S_T^2(Y_T-1)^+-\left(a_pG^{1/(p-1)}Y_T^{\alpha_p}S_T^2\right)\wedge S_T^2(Y_T-1)^+\\&=S_T^2\left((Y_T-1)^+-\left(a_pG^{1/(p-1)}Y_T^{\alpha_p}\right)\wedge(Y_T-1)^+\right)\\&=S_T^2\left((Y_T-1)^+-(Y_T-1)^+I_{\{Y_T\le C(p)\}}-a_pG^{1/(p-1)}Y_T^{\alpha_p}I_{\{Y_T>C(p)\}}\right).\end{aligned}$$

Taking into account that $I_{\{Y_T>C(p)\}}=1-I_{\{Y_T\le C(p)\}}$, we get

$$\begin{aligned}E^*\left(S_T^1-S_T^2\right)_p^+&=E^*\left(S_T^1-S_T^2\right)^+-E^*\left(S_T^1-S_T^2\right)^+I_{\{Y_T\le C(p)\}}\\&\quad-a_pG^{1/(p-1)}\left(E^*S_T^2Y_T^{\alpha_p}-E^*S_T^2Y_T^{\alpha_p}I_{\{Y_T\le C(p)\}}\right).\end{aligned} \tag{35}$$

Since $C(p) \geq 1$, we have

$$\begin{aligned} E^*\left(S_T^1 - S_T^2\right)^+ - E^*\left(S_T^1 - S_T^2\right)^+ I_{\{Y_T \leq C(p)\}} &= E^*\left(S_T^1 - S_T^2\right) I_{\{Y_T > C(p)\}} \\ &= E^*\left(S_T^1 - S_T^2\right) - E^*\left(S_T^1 - S_T^2\right) I_{\{Y_T \leq C(p)\}}. \end{aligned} \tag{36}$$

Using (36), we can calculate the difference between the first two terms in (35) reproducing exactly the same procedure as in (28)-(30) and arrive at

$$\begin{aligned} E^*\left(S_T^1 - S_T^2\right)^+ - E^*\left(S_T^1 - S_T^2\right)^+ I_{\{Y_T \leq C(p)\}} \\ = S_0\left\{\Phi\left(b_+(1, C, T)\right) - \Phi\left(b_-(1, C, T)\right)\right\}. \end{aligned} \tag{37}$$

To calculate the other two terms in (35), we represent the product $S_T^2 Y_T^{\alpha_p}$ as follows

$$\begin{aligned} S_T^2 Y_T^{\alpha_p} &= S_0 \times \exp\left\{\left(\sigma_1\alpha_p + \sigma_2\left(1-\alpha_p\right)\right)W_T^* - \frac{1}{2}\left(\sigma_1^2\alpha_p + \sigma_2^2\left(1-\alpha_p\right)\right)T\right\} \\ &= S_0 \times \exp\left\{\left(\sigma_1\alpha_p + \sigma_2\left(1-\alpha_p\right)\right)W_T^* - \frac{1}{2}\left(\sigma_1\alpha_p + \sigma_2\left(1-\alpha_p\right)\right)^2 T\right. \\ &\quad \left. + \frac{1}{2}\left(\sigma_1\alpha_p + \sigma_2\left(1-\alpha_p\right)\right)^2 T - \frac{1}{2}\left(\sigma_1^2\alpha_p + \sigma_2^2\left(1-\alpha_p\right)\right)T\right\} \\ &= S_0 \times \exp\left\{\left(\sigma_1\alpha_p + \sigma_2\left(1-\alpha_p\right)\right)W_T^* - \frac{1}{2}\left(\sigma_1\alpha_p + \sigma_2\left(1-\alpha_p\right)\right)^2 T\right. \\ &\quad \left. - \alpha_p\left(1-\alpha_p\right)\left(\sigma_1 - \sigma_2\right)^2 \frac{T}{2}\right\}. \end{aligned} \tag{38}$$

Taking an expected value of (38) with respect to P^*, we find that

$$E^* S_T^2 Y_T^{\alpha_p} = S_0 \exp\left\{-\alpha_p\left(1-\alpha_p\right)\left(\sigma_1 - \sigma_2\right)^2 \frac{T}{2}\right\}. \tag{39}$$

Using (38), (39) and following the same steps as in (28)-(30), we obtain

$$\begin{aligned} -a_p G^{1/(p-1)}\left(E^* S_T^2 Y_T^{\alpha_p} - E^* S_T^2 Y_T^{\alpha_p} I_{\{Y_T \leq C(p)\}}\right) \\ = -a_p G^{1/(p-1)}\left(S_0\right)\exp\left\{-a_p\left(1-a_p\right)\left(\sigma_1 - \sigma_2\right)^2 \frac{T}{2}\right\} \\ \times \Phi\left(b_-(1, C, T) + a_p\left(\sigma_1 - \sigma_2\right)\sqrt{T}\right). \end{aligned} \tag{40}$$

Combining (13), (14), (35), (37), and (40), we arrive at (17). □

6.3 Proof of (18)

Taking into account the structure of $\left(S_T^1 - S_T^2\right)_p^+$ in (11) we represent the product $Z_T\left(S_T^2\right)^{1-p}$ with the help of a free parameter γ (see (15), (16), (31)-(34)) and get

$$Z_T\left(S_T^2\right)^{1-p} = G\left(S_T^1\right)^{-\frac{(1+\gamma)\mu_1}{\sigma_1^2}}\left(S_T^2\right)^{\frac{\gamma\mu_1}{\sigma_1\sigma_2}}\left(S_T^2\right)^{1-p} = GY_T^{\alpha_p}, \tag{41}$$

where $\alpha_p = -\frac{(1+\gamma)\mu_1}{\sigma_1^2} = -\frac{\gamma\mu_1}{\sigma_1\sigma_2} - (1-p)$ and, hence,

$$\begin{aligned} &\gamma = \gamma_p = \frac{\sigma_2\left(\mu_1 - (1-p)\sigma_1^2\right)}{\mu_1(\sigma_1 - \sigma_2)}, \\ &-\alpha_p = \frac{\mu_1}{\sigma_1^2}\left(1 + \frac{\sigma_2\left(\mu_1 - (1-p)\sigma_1^2\right)}{\mu_1(\sigma_1 - \sigma_2)}\right) = \frac{\mu_1}{\sigma_1^2} + \frac{\sigma_2}{(\sigma_1 - \sigma_2)}\left(\frac{\mu_1}{\sigma_1^2} - (1-p)\right). \end{aligned} \tag{42}$$

Consider the following characteristic equation:

$$y^{-\alpha_p} = a_p G\left((y-1)^+\right)^{1-p}, \; y \geq 0. \tag{43}$$

1. If $-\alpha_p > 1-p$, then according to (42)

$$\frac{\mu_1}{\sigma_1^2} + \frac{\sigma_2}{(\sigma_1 - \sigma_2)}\left(\frac{\mu_1}{\sigma_1^2} - (1-p)\right) > 1-p \quad \text{or} \quad \frac{\mu_1}{\sigma_1^2} > 1-p > 0. \tag{44}$$

In this case the equation (43) has zero, one, or two solutions. All these situations can be considered in a similar way as Case 1.

1.1. If (43) has no solution then $I_{\left\{Z_T^1 > a_p H^{1-p}\right\}} \equiv 1$, $H_p = H$ and, therefore, ${}_T p_x = 1$.

1.2. If (43) has one solution $C = C(p)$ then $\left(S_T^1 - S_T^2\right)_p^+ = \left(S_T^1 - S_T^2\right)^+ I_{\{Y_T \leq C(p)\}}$ and, according to (13), we arrive at (18).

1.3. If there are two solutions $C_1(p) < C_2(p)$ to (43) then the structure of a modified claim is $\left(S_T^1 - S_T^2\right)_p^+ = \left(S_T^1 - S_T^2\right)^+ I_{\{Y_T \leq C_1(p)\}} + \left(S_T^1 - S_T^2\right)^+ I_{\{Y_T > C_2(p)\}}$ and we arrive at (20).

2. If $-\alpha_p \leq 1-p$, then $\frac{\mu_1}{\sigma_1^2} \leq 1-p < 1$ and, therefore, the equation (43) has only one solution $C = C(p)$. This is equivalent to 1.2 and, reproducing the same reasons, we arrive at (18).

6.4 Proof of (21)

According to (16), we represent the density Z_T as follows

$$Z_T = G\left(S_T^1\right)^{-\frac{(1+\gamma)\mu_1}{\sigma_1^2}}\left(S_T^2\right)^{\frac{\gamma\mu_1}{\sigma_1\sigma_2}} = GY_T^{\alpha_p} \tag{45}$$

where $\alpha_p = -\frac{(1+\gamma)\mu_1}{\sigma_1^2} = -\frac{\gamma\mu_1}{\sigma_1\sigma_2}$ and, therefore,

$$\gamma = \gamma_p = \frac{\sigma_1\sigma_2}{\sigma_1(\sigma_1-\sigma_2)},$$
$$-\alpha_p = \frac{\mu_1}{\sigma_1\sigma_2}\gamma_p = \frac{\mu_1}{\sigma_1(\sigma_1-\sigma_2)},\ \sigma_1 > \sigma_2. \tag{46}$$

From (16) and (46) we find that

$$\left(S_T^1 - S_T^2\right)^+ I_{\left\{Y_T^{-\alpha_p} > Ga_p\right\}} = \left(S_T^1 - S_T^2\right)^+ I_{\left\{Y_T^{\frac{\mu_1}{\sigma_1(\sigma_1-\sigma_2)}} > Ga_p\right\}} = \left(S_T^1 - S_T^2\right)^+ I_{\{Y_T > C\}} \tag{47}$$

where

$$C = \left(Ga_p\right)^{\frac{\sigma_1(\sigma_1-\sigma_2)}{\mu_1}}. \tag{6.20}$$

Using (13), (14), (36), (37), and (47), we arrive at (21).

7. Conclusion

As financial markets become more and more complicated over time new techniques emerge to help dealing with new types of uncertainties, either not present or not recognized before, or to refine measurements of already existing risks. The insurance industry being a part of the bigger and more dynamic financial industry could benefit from new developments in financial instruments and techniques. These may include introduction of new types of insurance contracts linked to specific sectors of the financial market which were not possible or not thought of before, new ways of hedging already existing types of insurance contracts with the help of financial instruments, more refined measurement of financial or insurance risks existing or emerging in the insurance industry that will improve their management through better hedging or diversification and thus allow insurance companies to take more risk. In any way, the insurance industry should stay attune to new developments in the financial industry.

Stochastic interest rate models, jump-diffusion models for risky assets, a financial market with N $(N > 2)$ correlated risky assets, modeling of transaction costs are few examples of the developments in the financial mathematics which could be incorporated in the financial setting of the model for equity-linked life insurance under our consideration. Some actuarial modeling including lapses, surrender options, the ability of the insured to switch between different benefit options, mortality risk models will also be able to enrich the insurance setting of the model. Methods of hedging/risk management other than efficient hedging could be used as well.

A balanced combination of two approaches to risk-management: risk diversification (pooling homogenous mortality risks, a combination of maturity benefits providing both a guarantee and a potential gain for the insured) and risk hedging (as for hedging maturity benefits with the help of financial market instruments) are going to remain the main focus for combining financial and insurance risk. A third risk management method – risk insurance (reinsurance, insurance of intermediate consumption outflows, insurance of extreme events in the financial market) – could be added for benefits of both the insurance company and the insured.

8. References

Aase, K. & Persson, S. (1994). Pricing of unit-linked insurance policies. *Scandinavian Actuarial Journal,* Vol.1994, No.1, (June 1994), pp. 26-52, ISSN 0346-1238

Bacinello, A.R. & Ortu, F. (1993). Pricing of unit-linked life insurance with endegeneous minimum guarantees. *Insurance: Mathematics and Economics*, Vol.12, No.3, (June 1993), pp. 245-257, ISSN 0167-6687

Boyle, P.P. & Schwartz, E.S. (1977). Equilibrium prices of guarantees under equity-linked contracts. *Journal of Risk and Insurance*, Vol.44, No.4, (December 1977), pp. 639-680, ISSN 0022-4367

Brennan, M.J. & Schwartz, E.S. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee. *Journal of Financial Economics*, Vol.3, No.3, (June 1976), pp. 195-213, ISSN 0304-405X

Brennan, M.J. & Schwartz, E.S. (1979). Alternative investment strategies for the issuers of equity-linked life insurance with an asset value guarantee. *Journal of Business*, Vol.52, No.1, (January 1979), pp. 63-93, ISSN 0021-9398

Ekern, S. & Persson S. (1996). Exotic unit-linked life insurance contracts. *Geneva Papers on Risk and Insurance Theory*, Vol.21, No.1, (June 1996), pp. 35-63, ISSN 0926-4957

Föllmer, H. & Leukert, P. (2000). Efficient hedging: cost versus short-fall risk. *Finance and Stochastics*, Vol.4, No.2, (February 2000), pp. 117-146, ISSN 1432-1122

Margrabe, W. (1978). The value of an option to exchange one asset to another. *Journal of Finance*, Vol.33, No.1, (March 1978), pp. 177-186, ISSN 0022-1082

Melnikov, A. (2011). *Risk analysis in Finance and Insurance* (2nd edition), Chapman&Hall/CRC, ISBN 9781420070521, Boca Raton–London–New York–Washington

Melnikov, A. & Romaniuk, Yu. (2008). Efficient hedging and pricing of equity-linked life insurance contracts on several risky assets. *International Journal of Theoretical and Applied Finance*, Vol.11, No.3, (May 2008), pp. 1-29, ISSN 0219-0249

Möller, T. (1998). Risk-minimizing hedging strategies for unit-linked life-insurance contracts. *Astin Bulletin*, Vol.28, No.1, (May 1998), pp. 17-47, ISSN 0515-0361

Möller, T. (2001). Hedging equity-linked life insurance contracts. *North American Actuarial Journal*, Vol.5, No.2, (April 2001), pp. 79-95, ISSN 1092-0277

Shulman, G.A. & Kelley, D.I. (1999). *Dividing pension in divorce*, Panel Publishers, ISBN 0-7355-0428-8, New York

9

Organizing for Internal Security and Safety in Norway

Peter Lango, Per Lægreid and Lise H. Rykkja
University of Bergen/Uni Rokkan Centre, Uni Research Norway

1. Introduction[1]

Policy-making and political processes imply putting specific societal problems on the agenda, and establishing permanent public organizations to deal with the issue in a systematic and continuous way (Jacobsen, 1964). This chapter analyses the political processes and outcomes within the field of internal security and safety in Norway, examining the development over the last 20 years. We focus on policies and specific crises that have led to changes in procedures as well as organization. We are interested in the question of coordination between public organizations, and more particularly the coordination between the Norwegian Ministry of Justice and other governmental bodies responsible for internal security and safety.

Even though governments work continuously to assess and reduce risks and vulnerabilities, experiences from major disasters and crises have shown that unthinkable and unmanageable situations and crises do occur. They range from completely new and unforeseen crises, to risks that have been anticipated, but not properly assessed. These situations, which cut across administrative levels (central-local government), policy sectors and ministerial responsibility areas, can be defined as *wicked issues* (Harmon & Mayer, 1986). Such complex and fragmented issues do not necessarily fit into the established functional structures and traditional divisions between line ministries, underlying agencies and levels of government. Furthermore, central actors may lack the competence, resources or organizational framework to handle such extreme situations.

This chapter addresses the reorganization of this policy area in Norway over the last 20 years, a period influenced by the end of the Cold War and the realization of new threats related to severe shocks such as the 9/11 terror attack and the tsunami in South-East Asia. Also, domestic polity features, administrative tradition and culture, pre-established routines and an active governmental administrative policy will be taken into account.

A central argument is that risk and crisis management challenges are typically found in the space between policy areas and administrative levels. The policy field of crisis management, internal security and safety typically crosses administrative levels sectors and ministerial areas, creating difficulties for those involved in preparing and securing safety.

The end of the Cold War changed dominant perceptions of risk and threats in many ways, from an attention to Communism and conventional war, to other types of threats such as

[1] This chapter is partly based on Lango & Lægreid (2011).

natural disasters or failures in advanced technological installations (Perrow, 2007; Beck, 1992). Central authorities were forced to redefine their understanding and the content of internal security and safety. A new conception concerned the dividing line between the civil and military defence. In the case of Norway, it included the introduction of new principles for organization, accountability and coordination (Serigstad, 2003).

From the early 1990s to the 2000s, several government initiated commissions emphasized the need for a stronger and better coordination within the field in Norway (St.meld. nr. 24 (1992–1993); NOU 2000: 24; NOU 2006: 6). The *Buvik Commission* (1992), the *Vulnerability Commission* (2000), and the *Infrastructure Commission* (2006) proposed a radical reorganization, including the establishment of a new and separate Ministry for internal security, and a new Preparedness Act. However, many of the proposals were, as we shall demonstrate, not followed through.

Internal security in Norway is characterized by an extensive division of responsibility. Proposals for an authoritative and superior coordinating authority has not been carried through. Thus, the field is frequently described as fragmented (Lægreid & Serigstad, 2006; Christensen & Lægreid, 2008). The Commission reports and Government White Papers over the last years leave no doubt that problems related to the fragmentation and lack of coordination is realized by central government and coordinating bodies. Still, there is considerable disagreement on how to solve these problems. Policy proposals have not been transposed into new organizational or comprehensive legal arrangements. At the same time, Norway has been sheltered from large disasters and catastrophes. Combined with the immanent uncertainty of risk management, this policy field is particularly challenging, not the least considering a continuous fight for policy attention and priority.

Focussing on problems of accountability, coordination and specialization within the field of internal security and safety and the reorganization processes in central government is interesting for several reasons. New organizational forms exceedingly confront existing ones as society faces new challenges. New Public Management-based reforms of the 1980s and 1990s encouraging decentralization and structural devolution have increasingly been supplemented by arrangements that emphasize the need for more coordination across sectors and levels, labelled post-NPM, Whole of Government or Joined Up Government (Bogdanor, 2005; Christensen & Lægreid, 2007; Bouckaert, Peters & Verhoest, 2010). At the same time, the awareness of threat related to natural disasters, pandemics and terrorism seems to have increased. This has made the field of internal security an increasingly relevant topic (Christensen et al., 2011).

The data in this paper is based on content analysis of central policy documents, mainly commission reports, government white papers, formal letters of assignment, parliamentary debates and documents, and supervisory reports. Also, a range of qualitative interviews (about 38) with central actors, politicians, commission members and senior civil servants have been carried out. Data collection and analysis was done by participants in the research project "Multilevel Governance and Civil Protection – the tension between sector and territorial specialization". The project was financed by the Norwegian Research Council from 2006–2010. For a more in-depth description of the data base, see Fimreite et al. (2011).

The chapter proceeds in four parts. Firstly, we present our theoretical approach. Next, we lay out central contextual factors, and present crucial principles and organizational arrangements. Thirdly, we describe important developments and central milestones in the efforts to reform the Norwegian internal security and safety policy field over the last 20 years. Then, we analyze and explain the reform process. The chapter closes with a concluding section discussing findings and implications.

2. A transformative theoretical approach

It is impossible to predict and map the probability of all accidents and crises. However, both public and private organizations are frequently faced with a demand to do just that. This means that governments have to handle both uncertainties and risks. Beck (1992) argues that there is an increased perception of a lack of safety in modern societies, related to the development of a *risk society*. In this perspective, advances in technology will potentially lead to disasters of great magnitude. A corresponding increased awareness of such catastrophic events will create even more challenges for responsible authorities. As people become more aware of risks and adverse consequences, the perceived level of risk will rise. Failure to manage risks might generate a potential for an even wider crisis. (Smith, 2006).

A general theory on how and why crises happen, and how they best can be managed does not exist. However, the work of for instance Perrow (Natural Accident Theory) (Perrow, 1984) and La Porte (High Reliability Theory) (La Porte, 1996) are highly valued and have been widely discussed. In different ways they try to to explain how crises arise and might be avoided. In many cases, a crisis can be traced back to organizational failure or poor risk management. Our theoretical point of departure is that different types of coordination and specialization will have important consequences for actors within public bodies, for the public bodies themselves, and for the policy field affected (March & Olsen, 1989; Egeberg, 2003; Egeberg, 2004). We argue that the organizational lay-out of the internal security and safety field is of crucial importance to risk management (Fimreite et al., 2011). Organizational forms affect which issues get attention and which are ignored, how the issues are grouped together and how they are separated. Organizational arrangements will therefore have vital importance for risk management. Furthermore, external pressures, developments and shocks (crises), may well be influential and result in new perceptions, and organizational or procedural changes.

Following this argument, organizational structure and operation can neither be viewed merely as a response to externally motivated influences and shocks, nor merely as a result of deliberate political choices. The organization of public administration, the relationship between the state and local level and between different sectors is not a mere technincal-neutral question. Organizations, seen as *institutions*, have an autonomous and influential authority based in established structures, rules, procedures, culture, traditions and dynamics (Olsen, 2010). Public authorities are characterized by a complex interaction, between political and administrative steering, design, negotiation, diverse interests, cultural bindings, and adaption to external pressures and influences (Christensen et al., 2004). In order to fully understand the content of public politics and policies, we need to analyze organizational structures and organizations as well as policy content. The scope of action and notions of approriate behavior for civil servants is affected by organizational affiliation. This also affects the content of policies (March & Olsen, 2006). The understanding of problems, instruments, consequences and behavioural patterns are affected by internal features of an organization, and the relationships and connections to other organizations. Based on this, we assume that different forms of specialization and coordination are decisive for organizational behaviour as well as policy outcomes (Fimreite, Lægreid & Rykkja, 2011).

In order to explain the development of the internal security field in Norway, we adopt a transformative approach, combining an instrumental and institutional perspective on public policy reform (Christensen & Lægreid, 2007). The organizations within the field are not

merely seen as instruments, but as institutional actors that not necessarily will adapt to new signals from political executives or to shifting demands in the environment (March & Olsen, 1983). The institutional dynamics of reforms can best be interpreted as a complex mixture of environmental pressure, polity features and historical institutional context.

The instrumental perspective directs our attention towards formal arrangements, while the institutional perspective focuses on informal norms, values and practices that have developed over time. Informal social roles are seen as impersonal, they exist independently of people within the organizations at any given point of time (Christensen & Lægreid, 2002). The institutional perspective takes as a starting point that organizations are carriers of values, and have a distinct identity (Selznick, 1957).

The instrumental perspective perceives organizations as disposable tools for the leaders involved. Within this perspective, rationality is related to the formal organizational structures, and creates limitations on actors' options. The institutional perspective, on the other hand, opens up for a perception where organizations incorporate routines, rules and values that independently influence actors and their behaviour.

Within an instrumental perspective, the underlying behavioural logic is a logic of consequence, based on rational actors that are assumed to be able to accurately predict consequences of choices, and find the means to reach their goals. Change is perceived as rational adaption to new goals or changing external demands. The concept of rationality has been modified somewhat by Simon (1976), who launched the concept of bounded rationality. This concept emphasizes limitations to the abilities to account for all possible choices and outcomes.

Within the instrumental perspective we distinguish between a hierarchically oriented variant, where the leaders' control and analytical-rational calculation is central, and a negotiation-based variant, which allows for the articulation of interests and for compromise and negotiation between organizations and actors whose goals and interests are partially conflicting (Christensen et al., 2004). An institutional perspective is, on the other hand, based on a logic of appropriateness (March & Olsen, 1989). Here, human action is driven by rules of appropriate or exemplary behaviour, organized into institutions. This implies that action is based on experience, and on what is perceived as reasonable and acceptable within the specific context. Goals and means are discovered, explored and developed within the specific organization, and can be interpreted differently from formally established goals and ends. Thus, intrinsic organizational values may obstruct fundamental change. Organizations are seen more robust and change is usually incremental. Moderate changes will meet less resistance than major reforms. Frequent and extensive changes will generate extensive transaction costs, referred to as *historical inefficiency* (Ibid.). However, the possibilities for change are greater if reform proposals are in accordance with the existing organizational traditions and established culture (Brunsson & Olsen, 1993). These processes can be understood as *path dependent*, where former choices constrain later options (Krasner, 1988).

A third perspective sees organizational structures mainly as a response to external pressures (Olsen, 1992; DiMaggio & Powell, 1983; Meyer & Rowan, 1977). This implies adaptation to established norms, beliefs and prevailing doctrines within a wider community, the incorporation of NPM values being one relevant example. It may also imply adaptation to a changing technical environment or to challenges and vulnerabilities created or revealed by external shocks and/or crises.

We use these theoretical perspectives in a supplementary manner (Roness, 1997), and argue that the organizational processes within the field of internal security and safety can neither be viewed one-sidedly as a result of instrumental processes and leader strategies, nor merely as a product of history, existent informal norms, or adaption to external pressure. Processes of policy formation and change are characterized by complex interaction between different factors. This is vital when one wants to understand the organization and development of risk management.

3. Context

3.1 Ministerial responsibility, strong line ministries and autonomous municipalities

Individual ministerial responsibility is a core concept within the Norwegian central government. The Minister, as head of a given Ministry, bears the ultimate responsibility for actions within that Ministry, including those of subordinate agencies. Ministerial responsibility in the Norwegian case implies strong sector ministries and a strong vertical coordination, resulting in a corresponding weaker horizontal coordination between policy areas and sectors (Christensen & Lægreid, 1998). Specialization by sector or purpose/tasks is a dominant principle, making it difficult to establish coordinative arrangements across traditional sectors. Consequently, sector ministries have been substantially stronger than ministries responsible for sector-crossing activities and coordination. This indicates that ministries operate as separate 'silos' with limited ability to apprehend cross-cutting policy issues (Bouckaert, Ormond & Peters, 2000).

Another central feature of the Norwegian polity is the concept of local self government. Local democracy and authority is a relatively strong value (Fimreite et al., 2002; Flo, 2004). Following the expansion of the welfare state after World War II, local authorities became responsible for providing a broad range of services. Greater municipal responsibility also meant a closer integration across government levels, and, at least until 1992, a sectorized organization mirroring central government institutions (Tranvik & Fimreite, 2006). A series of reforms aimed at municipal devolution was implemented from the 1980s and culminated with the Municipal Act of 1992. The new legislation aimed at joined-up (non-sectoral) government structures at the municipal level in order to counter the strong sectorization of Norwegian public government and the centralizing forces that allegedly reduced local government autonomy. Whether or not the reforms succeeded, is still debated (Ibid.)

3.2 Internal security and safety

The attention towards internal security and safety is quite new, both in academia and politics. The concept covers terms like 'domestic security', 'civil defence', 'homeland security', 'societal security' and 'civil emergency' (Lægreid & Serigstad, 2006), and has gained currency since the beginning of the 1990s – even more so the 9/11 terrorist attacks in the US in 2001 (Kettl, 2004).

At present, there exists no agreed-on international definition. The term 'societal security' (*samfunnssikkerhet*) is a specific Norwegian term. It is a rather broad concept, and does not differentiate between safety and security, or between natural disasters beyond human control and conscious destructive acts (Burgess & Mouhleg, 2007). Thus, it straddles the rather blurred boundary between civil society and internal affairs on the one hand, and the military and defence sector on the other. According to Olsen et al. (2007, pp. 71), a viable

definition of the concept of 'societal security' could be "*The society's ability to maintain critical social functions, to protect the life and health of the citizens and to meet the citizens' basic requirements in a variety of stress situations*". It comprises all categories of actions intended to hinder unwanted events or conditions and to reduce the consequences should these occur, covering both preventive and proactive actions pursued in order to reduce adverse effects (St.meld nr. 17 (2001–2002), pp. 3). This covers both extraordinary events (e.g. hurricanes, terrorism, etc.) and more 'ordinary' events (e.g. traffic accidents, fires, etc.), and includes both internal security and civil protection and safety. It further communicates that national security is more than military defence and border control, and thus finds outlet for the Norwegian conception of 'total defence' (Serigstad, 2003).

In this chapter, we have settled on the term 'internal security and safety'. This has mainly a practical reasonIt is a conception that is more common and easily accessible for the international reader. By focussing on both security and safety, we emphasize that the outlook is not delimited to mere security issues, or to mere safety issues. Although the typical Norwegian approach covers both extraordinary and more 'ordinary' accidents, we will focus more on the importance of extraordinary events and risks.

Inherent in this definition, and in our focus, is an explicit attention to the question of how government manages risk. At its heart, the policy field of internal security and safety concerns risk management and 'the politics of uncertainty' (Power, 2004). The problem both public and private organizations face, is that of responding to both anticipated and unanticipated risks. A particular difficult question is how to prepare for low-probability and high-impact events (Baldwin & Cave, 1999). Organizations and organizational arrangements play a crucial role in the prevention of and response to risk (Lægreid & Serigstad, 2006, pp. 1379). Within this framework, reorganization can also be seen as means for managing risk.

3.3 Central principles: liability, decentralization and conformity

In Norway, three crucial principles for internal security and safety guide authorities involved in risk and crisis management. These are a principle of liability, a principle of decentralization and a principle of conformity (St.meld. nr. 22 (2007–2008)).

The liability principle implies that every ministry and authority has responsibility for internal security and safety within its own sector. It is closely related to the doctrine of individual ministerial responsibility, emphasizing strong sector ministries. According to our empirical material, this has made the Ministry of Justice's responsibility for horizontal coordination more difficult. Furthermore, the principle of liability is modified by extensive civil-military cooperation with the Ministry of Defence and its subordinate bodies.

The decentralization (or subsidiarity) principle emphasizes that a crisis should be managed at the lowest operational level possible. This corresponds to the dominant doctrine of local self-government and authority. Consequently, the County governors and municipalities are given an important function in risk assessment and crisis management. At the regional level, the County governors operate as mediators between sector interests as well as state and local level administration (Rykkja, 2011). Traditionally, the Norwegian municipalities have enjoyed widespread autonomy within the field of civil protection. Territory, or geography, is therefore an important additional organizing concept Herein lays an important (organizational) paradox: the principle of liability implies strong coordination within specific sectors, but weak coordination across them. The decentralization principle, on the

other hand, implies strong coordination across sectors on a low level and hence less coordination between horizontal levels of government.
The principle of conformity (or similarity) stresses that the organization forms under a crisis or a crisis-like situation should be as similar to the daily organizational forms as possible. This can be particularly difficult to maintain during an 'extraordinary' crisis. As current literature on crisis management emphasizes, when a major disaster happens, the necessity of supplementing existing formal organizations with improvisation and temporary organizations becomes crucial (Czarniawska, 2009).
These three principles comprise a central fundament for the organization of internal security and safety in Norway. Nevertheless, several small organizational and policy changes beyond these principles have taken place over the last years. This is the topic of the following sections.

4. The reform process: Reorganization of the central administration for security and safety

4.1 From the Cold War to a 'vulnerable society'

In 1946, a government appointed defence commission established the concept of '*total defence*'. This implied an integration of the Norwegian military and civil defence, with a primary task to protect Norwegian territory, citizens, national values and sovereignty. At the time, the orientation of the 'total defence' was mainly towards the threat of war, not crises or more delimited accidents (Serigstad, 2003). However, changes in the perception of threats and risks in the public sphere gradually led to the adaption of existing arrangements to encompass solutions that were more suitable to a new situation. The end of the Cold War created a new situation for both the military and the civil defence. Without a unitary and stable enemy, assessing risks and threats became more complex. This resulted in the adoption of a broader concept, embraced by both the Ministry of Justice and the Ministry of Defence. 'Total defence' came to imply mutual support and cooperation between the military services and the civil society, covering war-like situations as well as more delimited crises affecting the civil society. Today, it involves both contingency planning and more operative matters in all types of crises (Høydal, 2007).
Internal security and safety was first conceptualized in a government White Paper presented by the Ministry of Justice in 1993 (St.meld. Nr 24 (1992-1993)), and further recognized by the Vulnerability Commission in 2000 (NOU 2000: 24). With these statements, the Government signalized a broader definition of the field, with less emphasis on the military dimension and more on the civil dimension and on crises that arise in peace-time (Lægreid & Serigstad, 2006; Olsen et al., 2007). This new conceptualization implied a transfer of responsibilities from the Ministry of Defence to the Ministry of Justice.
Internal security and safety is a relatively new task for the Ministry of Justice. The construction of a Norwegian Civil defence started after the experiences from World War II. In 1970, a Directorate for Civil Protection was set up under the auspices of the Ministry of Justice. At the same time, a corresponding preparedness office within the Ministry was created. Over time, there have been several attempts to strengthen the Ministry of Justices' coordinating role (Høydal, 2007).
The White Paper presented in 1993 (St.meld. nr. 24 (1992–1993)) clearly articulated a need for a coordinating ministry. However, the principle of responsibility was not abandoned. Ultimately, responsibility was to be placed with whichever ministry had the administrative

and sector responsibility, depending on the type of crisis. Constitutional ministerial responsibility would still lie with the relevant Minister. Nevertheless, the interest of assigning the over-arching responsibility to a single ministry was strong. The coordinating ministry would take cooperative initiatives on behalf of other involved ministries in order to ensure better coordination of resources in both peace and war.
Several candidate ministries were envisaged having a coordinative role. The Prime Ministers Office and the Ministry of Justice were the most prominent (Høydal, 2007). The Buvik Commission (1992) recommended a leading role for the Ministry of Justice, and this was supported by the government. A central argument was that the Ministry of Justice already was responsible for the Directorate for Civil Protection and Civil Defence, and for the Police and rescue services (St.meld. nr. 24 (1992–1993)). In 1994, the Ministry was formally assigned the task to coordinate civil preparedness across sectors (Serigstad, 2003).

4.2 The Vulnerability Commission and the Ministry of Justices' responsibility for coordination

In 1999, the Ministry of Justice initiated a project with a vision to enhance the attention to the area of internal security and safety. This led to the establishment of a public commission led by a former distinguished politician and Prime Minister, Kåre Willoch. The Commission on the Vulnerability of Society (also called the Vulnerability Commission) presented a broad range of proposals in order to improve efforts to reduce vulnerability and ensure safety, security and civil protection for the Norwegian society (Serigstad, 2003; Lægreid & Serigstad, 2006; NOU 2000: 24).
The Commission identified several problems concerning civil defence and internal security. One of its central conclusions was that the policy area was highly fragmented, lacked superior organizing principles, and was to a large extent organized in an ad hoc manner, responding to specific crises or accidents (Høydal, 2007). Allegedly, this resulted in ambiguity and serious liability concerns. A central argument was that the Ministry of Justice did not execute its superior and coordinative functions within the area very well. Civil protection and crisis management was mainly executed by a small unit with limited resources, and was not adequately prioritized. Overall, internal security and safety was seen as a policy area that had been systematically under-prioritized for quite a while. Furthermore, the Ministry's coordinative responsibilities were vaguely defined, and therefore largely ignored by other relevant ministries and departments.
The Vulnerability Commission concluded by advocating a higher concentration of responsibility, competence and resources, in order to give the area a stronger political foothold and ensure better coordination. One central recommendation was to establish a new Ministry for Internal Security and Safety, incorporating responsibility for assessing national threats and vulnerabilities, establishing main goals and standards within the field, coordinate efforts to handle terrorism and sabotage, as well as existing emergency departments and the civil defence. This implied a total restructuring of the Ministry of Justice, and a transfer of central administrative responsibilities from other sectors in order to ensure a stronger and more autonomous role for the new ministry.
The Commission report was followed by public hearing and a government White Paper on the Safety and Security of Society (St.meld. nr. 17 (2001–2002)). The decision-making process prior to the White Paper was characterized by defensive institutional arguments and major conflicts of interest, especially between the justice and defence/military sector (Serigstad, 2003). The hearing did not provide any major changes to the original proposal. In the middle of this

process came 9/11, and the following organizational changes in the US administration for Homeland Security. The situation led to a delay and reassessment of the Commission's work, but in the end it did not have any major impact on its conclusions (Ibid.).

The proposal to establish a new Ministry for Internal Security and Safety turned out too controversial. Instead, the White Paper proposed a further strengthening of existing structures, by merging existing units and agencies, and by establishing new ones. Consequently, the proposals confirmed existing principles and doctrines of public organization and management within the field. The existing principles of liability, decentralization and conformity were maintained. The result is a rather ambiguous and hybrid organizational model (Lægreid & Serigstad, 2006).

The White Paper proposed the reorganization of several existing agencies and the following establishment of two (partly) new agencies; the Directorate for Civil Protection and Emergency Planning (DSB), and the Norwegian National Security Authority (NSM) (St.meld. nr. 17 (2001–2002)). DSB was organized as an agency under the responsibility of the Ministry of Justice. It supports the Ministry's coordinative activities within the field, and consists of the former Directorate for Civil Protection and the former Directorate for Fire and Electrical Safety Inspection. DSB is responsible for overall emergency planning and crisis management, providing information and advice as well as supervising responsible ministries, county governors and municipalities.

The National Security Authority (NSM) is responsible for protective supervision and the security of vital national interests, primarily countering threats of espionage, sabotage and acts of terrorism (NOU 2006: 6, Act of 20 March 1998 on Protective Security Services). Initially, the Vulnerability Commission wanted to establish NSM as an agency under the proposed new Ministry. This would mean a transfer of the agency from the Ministry of Defence. This resulted in a conflict of interest between the two Ministries involved, and between the two corresponding parliamentary committees.

On one side, the Ministry of Justice and the Parliamentary Standing Committee on Justice argued for a broader definition of the field and the inclusion of civil protection and safety within its realms. On the other side, the Ministry of Defence and the Parliamentary Standing Committee for Defence wanted to keep a focus on (military) security issues, and therefore also retain the administrative responsibility for NSM. The solution was a compromise. NSM was administratively placed under the responsibility for the Ministry of Defence, but would also report to the Ministry of Justice in matters concerning civil protection (Lægreid & Serigstad, 2006).

The developments within the field illustrates how difficult it can be to restructure established arrangements, and transfer responsibility between ministries, even in situations where existing problems are recognized. The Norwegian case exposed a fundamental conflict concerning the framing of the field. Should internal security and safety be defined as a responsibility alongside many other equally important tasks, or should it rather be defined as a particular policy field, characterized by distinct and more vital problems and challenges? Was this mainly a security issue, and therefore a military defence matter, or was it rather a safety issue, and therfore a problem concerning civil protection and defence? Discussions on the degree of integration between civil and military protection and defence, and safety and security issues, continued. In the end, the White Paper was discussed jointly by the two Parliamentary committees. This indicates a shift towards safety and civil protection, since earlier these issues were mainly discussed in the Standing Committee for Defence alone.

The basic conflicts and challenges portrayed here have had significant consequences for the perception of relevant problems, policy solutions, of relevant actors and participants in the process. The process can be perceived not only as a clear-cut decision-making process, but also as a process of meaning-making, concerning the definition, interpretation and development of a common understanding, and as a process of constructing a certain political reality and negotiation ground for those best suited to implement the tasks at hand (Baumgartner & Jones, 1993; Rochefort & Cobb, 1994; Kettl, 2004).

4.3 The Cabinet Crisis Council and the Crisis Support Group

The same reluctance to establish a stronger coordinative and authoritative role for the Ministry of Justice through more permanent organizational arrangements can be observed when analyzing the processes leading to the establishment of a Cabinet Crisis Council and a Crisis Support Group in 2006. The Indian Ocean earthquake and the following tsunami on Boxing day in South-East Asia in 2004 were crucial for the establishment of these organizations.

Although the tsunami disaster hit abroad, it had important consequences for Norway. At the time, about 4000 Norwegian citizens were in the area. Most of them were on vacation. 84 Norwegian citizens were killed. Because it happened abroad, and following from the established principle of liability, the situation was handled by the Ministry of Foreign Affairs. However, the Ministry was not very well prepared for a situation like this, and was quickly criticized for their efforts to coordinate activities and responses, both within the Ministry itself and across other involved ministries (Brygård, 2006; Jaffery & Lango, 2011).

After the tsunami, the Government presented a White Paper on Central Crisis Management, referring directly to the tsunami disaster (St.meld. nr. 37 (2004–2005)). It continued the discussion concerning the demarcation and responsibility lines between the different ministries, authorities and administrative levels involved in the crisis, and presented several measures to improve coordination and crisis management at central governmental level. This included an effort to clarify responsibilities for crisis management. More importantly, in a crisis the lead was to be placed with the Ministry mostly affected. The intention was to stall potential conflicts of competence and responsibility one had experienced on earlier occasions. This principle further emphasized the principle of liability.

The White Paper also proposed the establishment of a Cabinet Crisis Council, and a strengthening of the administrative support through the setting up of a Crisis Support Group. The initial proposal was to organize the Cabinet Crisis Council permanently to the Prime Minister's Office. However, this was turned down, and the result was a more ad hoc organization. If and when a complex crisis that demands coordination at Ministerial level hits, any affected Ministry may summon the Council. It consists of the Permanent Secretary's from the Prime Ministers Office, the Ministry of Justice, the Ministry of Defence, the Ministry of Health and the Ministry of Foreign Affairs. When summoned, the Council functions as the superior administrative coordinating body, and is responsible for coordinating measures across the relevant ministries. However, the constitutional and ministerial responsibility still rests within each Ministry.

The Crisis Support Group may be called upon in certain demanding crisis situations by the leading Ministry. It is formally organized under the Ministry of Justice, but can be called upon by any responsible Ministry and be expanded upon need. It is mainly an administrative resource supporting whichever Ministry takes the lead. The establishment of

such a group was originally proposed by the Vulnerability Commission, but not followed through until a larger crisis hits Norway, then.
Corresponding to earlier policy documents within the field, the White Paper following the tsunami disaster emphasised the importance of the Ministry of Justice's leading role in crisis situations. However, the principle of liability was not to be altered. A consequence may be an even more fragmented organization, whereas the Cabinet Crisis Councils functions may counteract the recently established leading role for the Ministry of Justice in certain exceptionally demanding disasters.

4.4 The Infrastructure Commission

A few months before the tsunami disaster hit, the Norwegian government set up a public commission to report on the security of critical infrastructure. This resulted in a report on the 'Protection of critical infrastructures and critical societal functions in Norway (NOU 2006: 6). Four issues were central: A discussion concerning the extent of public ownership, a discussion on the coordinative role of the Ministry of Justice, a proposal for a statutory obligation for preparedness in local authorities, and a proposal for a new, overarching and sector-crossing Preparedness Act.
The commission presented several proposals concerning the Ministry of Justice's coordinative role, from having a superior yet advisory role, taking initiative and organizing collaboration and information, to being a national junction and reference point in crises that demanded international operations. Further, it argued for better coordination of relevant agencies (NSM, DSB and the Police Security Service, PST). However, the proposal of establishing a new and more authoritative Ministry for internal security and civil safety launched by the Vulnerability Commission was not followed up.
The argument for a new Preparedness Act was that it would secure the integration of risk and vulnerability analysis, operational and supply safety, preparedness planning, information sharing, cooperation control and sanctions against both businesses and public authorities responsible for critical infrastructure (Bjørgum, 2010; NOU 2006: 6). The commission argued that this new legislation would give the Ministry of Justice stronger coordination powers. Nevertheless, the report did not provide a detailed review of existing legislation. The result was more a call for attention on the question, and it did not present any draft legislation. It was implicit that more research and analysis was needed before a new law could be enacted.
The Commission report was followed by a new White Paper (St.meld. nr. 22 (2007–2008)). Here, the coordinative role of the Ministry of Justice was defined as a responsibility for securing a general and coordinated preparedness. A sectoral approach to relevant agencies, County governors and Joint Rescue Coordination Centres across the country was emphasized. Hence, no radical reforms were proposed. The proposal to establish a new Preparedness Act was not followed up, although the White Paper recommended a more detailed examination of existing legislation in order to determine relevant priorities and problem areas. The most significant proposal was the establishment of a statutory obligation for local authorities to provide adequate preparedness. This was implemented when a revised Civil Defence Act was adopted in 2010.

4.5 Critique from the Office of the Auditor General

In 2008, the Norwegian Office of the Auditor General presented a report on the Ministry of Justice's coordination responsibility within the field of internal safety and security

(Riksrevisjonen, 2008). The report was rather critical. A central finding was that several responsible Ministries did not perform adequate risk and vulnerability analyses, and thus did not prioritize risk management. It also pointed out failings in the Ministry of Justice's audit of other ministries, and in its dialogue with them. Adequate coordination would demand the Ministry to take a more active and deliberate role towards its coordinative responsibilities within the field. The other Ministries found the coordinative responsibility of the Ministry of Justice to be unclear. A main conclusion was that, despite evident changes within the field, important challenges concerning accountability and coordination remained. The Ministry's response to this critique was to establish a Ministries' Coordination Consulting Group for internal security and safety. This is a common inter-ministerial arena for exchange of information and experiences, and for the discussion of general rules concerning preparedness. The arrangement followed the existing organizational policy, with rather weak network arrangements that do not threat the power of line ministries.
The same findings pointing to an apparent lack of coordination can be found in the Office of the Auditors Generals report on goal achievements and efficiency in the County governors' offices (Riksrevisjonen, 2007). The County governors have important coordinative responsibilities at regional level, and are responsible for preparedness, risk and crisis management within their region (Rykkja 2011). The report from 2007 pointed out that risk and crisis management was largely under-prioritized. Furthermore, there existed certain ambiguities concerning the County governors' coordinating role, and that the coordination vis-à-vis municipalities and other state authorities within the region was characterized as rather ineffective.

4.6 Summing up the reform process – a reluctant reformer

The most important changes in the Norwegian policy for internal security and safety since the Cold War, have been the introduction of the three central steering principles (the principle of liability, decentralization and conformity), a development and clarification of the Ministry of Justices' authority and coordination responsibilities, and the establishment of new directorates, agencies and more ad hoc organizational arrangements under the Ministry. This includes the Cabinet Crisis Council and the Crisis Support Group. Furthermore, responsibilities between central and local government have been spelled out more clearly through the establishment of a municipal preparedness duty.
Our analysis reveals that the principles of ministerial superiority and autonomous local government have set distinct limitations on legislative and organizational proposals, on how they are formed, followed up on, and carried through. In general, established organizational forms are strengthened, resulting in a somewhat cautious adaptation to a new situation following the end of the Cold War. Although it took a very long time to realize, the establishment of a statutory obligation for preparedness within the local authorities indicates that the principle of autonomous local government may be easier to shift than the principle of ministerial superiority securing strong sector hegemony.
A central tension has been the relationship between the military and civil sector (Serigstad, 2003; Dyndal, 2010). The concept of 'total defence' signalized increased focus on civil issues. Over the years we see a shift from the military towards the civil sector and issues concerning internal security and safety (NOU 2006: 6). This has resulted in new relations between the military and the civil sector. An example is the establishment of the National Security Authority (NSM), subordinate to the Ministry of Justice in matters of civil concerns,

and subordinate to the Ministry of Defence in matters of military concern. This joint arrangement may result in tensions between ministries, concerning allocation of resources, establishment of central goals and priorities, and adequate steering and adequate steering.
In the same period, there has been an effort to strengthen coordinating authorities within the field through the clarification the responsibilities of the Ministry of Justice, the Directorate for Civil Protection and Emergency Planning (DSB), the National Security Authority (NSM), the County governors and the municipalities. However, our analysis reveals that the principle of liability has not been surrendered, and still stands strong. This continues to create tensions between organizational units, sectors and administrative levels. An indication of the complex relationship is that the different ministries are required to perform internal control and system audit within their respective sectors, while at the same time the Ministry of Justice and DSB audits the individual ministries. Høydal (2007) reports that it is especially difficult to get this arrangement to work.
Experiences with certain crises have revealed that the responsible authorities are not always well prepared. This is also documented in the general literature on disasters and crisis management (Fimreite et al., 2011). A particular relevant example in our case is the handling of the tsunami disaster in South East Asia in 2004. This crisis revealed serious challenges related to the coordination and specialization between responsible ministries (Jaffery & Lango, 2011).
Organizational changes after the Cold War have been rather discrete. Parallel processes have been influential, and major proposals have been countered by strong sector interests. The experiences after the tsunami led to some reorganization in the central administration, but not to the establishment of completely new arrangements. This seems to follow a rather common pattern, where Norway has been labelled a reluctant reformer compared to other countries (Christensen & Lægreid, 2007). Other events, such as the Mad Cow Disease (BSE), also led to changes that to a large extent followed existing lines of responsibilities (Rykkja, 2008).
Our analysis shows that coordination and strengthening of central government in this policy area seems difficult, mainly due to strong line ministries with different interests, and also influenced by a strong preference of decentralized solutions following the doctrine of local self government. The government is reluctant to build up strong permanent core organizations in central government, with adequate capacity and resources, within this policy field. The Ministry of Justice remains the central coordinating body, but is still characterized as rather weak. Attempts to build a strong overarching coordinating ministry failed, largely due to the strength of the principle of ministerial responsibility. Crises have not resulted in radical changes. Instead, our analysis reveals incremental processes.

5. Understanding processes and outcomes

In accordance with an instrumental perspective, the development of a new coordination policy for internal security and safety can be seen as a process of deliberate and strategic choices following an interest in strengthening the Ministry of Justices' coordinating functions. In this perspective, policy-making and reorganization are important tools, utilized to improve practices and results within the field. However, the process exhibits examples of bounded rationality, and local rationality seems linked to the relevant organizations' apprehension of problems and solutions (Cyert & March, 1963). Changes in practices can partly be seen as relevant measures for ensuring the unity of the policy field and a strengthening of horizontal coordination. Our study has nevertheless shown that

conflicts of interests between central actors have limited horizontal coordination and the development of a coherent coordination policy.

Institutionalized opinions and the persistence of institutions are central elements in an institutional perspective (Krasner, 1988). Our research reveals that the institutionalized tradition of ministerial responsibility continues to stand strong within the Norwegian polity, limiting the efforts to strengthen horizontal coordination. The processes can well be seen as a result of path dependency, that at least partly may explain why the changes within the field are incremental. This points to processes where former routines and established practices lead to selective and inconsistent attention to new problems (Cyert & March, 1963).

The end of the Cold War represents an important contextual factor that might explain the gradual shift from a military focus to a stronger attention to civil protection and safety. It may also be regarded as a manifestation of historical inefficiency, where old organization patterns and understandings still linger and hinder major reforms and reorganizations, despite obvious changes in the context (March & Olsen, 1989).

The events of 9/11 2001 happened at the same time as the Norwegian Vulnerability Commission worked with their report, and was definitely an external shock with important consequences for the American government (Kettl, 2004). However, we find little evidence that it had a decisive influence on the Norwegian reorganization process. The White Paper was put on hold for a while, but no radical proposals were put forward (St.meld. nr. 17 (2001–2002)). Furthermore, many of the proposals of the Vulnerability Commission were not followed up. The incidents of 9/11 led to changes in terror legislation in many countries, including Norway (Meyer, 2009; Rykkja, Fimreite & Lægreid, 2011). The Norwegian government was obviously aware of a new kind of threat after 9/11, but it does not seem that the organizational changes and establishment of Department of Homeland Security in the US (described as the largest reorganization in the American central administration in newer history) had any major impact.

The tsunami disaster in 2004 seems to have had a stronger direct impact on the reorganization processes. It hit Norwegian citizens more directly, and also created notable difficulties for the responsible authorities. Problems related to ambiguous responsibility lines and competences, and a corresponding lack of coordination between relevant ministries was revealed. This promoted the enactment of organizational changes within the central administration (the Cabinet Crisis Council and the Crisis Support Group).

To some extent, Norway represents a special case, whereas it has largely been spared direct encounters with path-breaking and devastating disasters. The reform processes in the Norwegian approach to internal security and safety resulted in a reorganization of agencies reaffirming a network based model (Lægreid & Serigstad, 2006). Our study reveals a process and a policy field characterized by complex interactions between mutually influential factors. Rather than holding different explanatory factors up against each other as competing or alternative, we argue that they complement each other. External shocks and incidents have had an impact, but changes within the field do not always follow predictable patterns. Individual choices and behaviour can, at least to some extent, be determined by examining characteristics and changes in the organizations' scope of action. However, we emphasize that this is not a deterministic relationship. Some scope of action for deliberate interference and proactive behaviour by political and administrative leaders remain.

Firstly, our analysis of the reform process within the field of internal security and safety in Norway bear witness that reforms are to a large extent formed through established

organizational arrangements, doctrines and principles that set limits to the scope of action. Secondly, we find that the process and outcome cannot be characterized as a result of rational planning alone but has clear negotiation based features. The reform initiatives involved actors with partly conflicting interests, and the organizational pattern appears to be a result of turf wars, conflicts of interests, compromises, and the establishment of certain winning coalitions. Thirdly, the established arrangements and institutions seem infused by traditions, organizational culture, established routines and rules, and informal norms, values and identities. This has a vital independent influence on how work within the field is carried out, and more particularly on how crises are prevented and managed. Central actors do not necessarily easily adapt, neither to new signals from political or administrative leaders, nor to alterations in the external environment.

Crisis and risk management typically takes place under uncertain and ambiguous conditions. In these situations, the prevalence of rational choices characterized by clear, stable and consistent goals, a fair understanding of available goals and means, and an apparent centre of authority and power, is not realistic. More likely, central goals will be rather unclear, ambiguous and partly conflicting, technological constraints may be uncertain, and there will be difficulties concerning the prediction of events and effects of relevant choices. In these situations, a flexible political and administrative coordination based on institutionally fixed rules, routines and roles may be a reasonable alternative to action based on calculated planning (Olsen, 1989). The extent to which such coordination processes succeed, is related to the existence of mutual trust and dependence between political and administrative leaders, between different professions, between central and local governments, between sectors, and between the population and government.

In Norway, such high levels of mutual trust are relatively well established (Rothstein & Stolle, 2003). This is the case both between public organizations, and between the population and the government. It has facilitated the introduction of trust-based regulatory arrangements, such as internal audit and control. It can also explain a relatively high acceptance of strong measures in the fight against terror (Rykkja, Fimreite & Lægreid, 2011). An explanation for the high level of trust in the Norwegian community is that there have been few disasters or path-breaking crises that have challenged these trust relationships. However, there are also examples of crises that are largely the result of poor internal audit and lack of control, for example the management of an explosion accident that spread toxic gases in western Norway in 2007 (Lervåg, 2010).

A discussion of adequate risk management raises the question of accountability. When risks with potentially large adverse effects are identified, the next step is to ask who is responsible, or who should take the blame? A natural institutional response is often to evade responsibility and try to avoid blame. If responsibility is not placed, the result is often damaging blame games. The tsunami disaster in 2004 may be a telling example. The reorganization process we have followed in this chapter also touches upon these issues.

Accountability relationships in crisis situations are complex, and illustrate the 'wickedness' of the field of internal security and safety. A main a question is to whom account is to be rendered. Bovens has framed this as '*the question of the many eyes*' (Bovens, 2007). Here, there is a central dividing line between administrative or managerial accountability, and political accountability. Fimreite, Lango, Lægreid & Rykkja (2011) show, by analyzing several different cases of crisis management in Norway, that managerial accountability is more

often discussed than political accountability. This is especially the case in complex crisis situations and in situations where problems cross traditional sector lines. Judging from these case-studies, it seems that politicians have a tendency avoid blame, while administrative leaders leave office. Because the decentralization principle dictates that a crisis should be handled at the lowest possible level, it can be seen to emphasize managerial accountability. The question of political accountability is a difficult one. Causes, as well as responsibility lines are often diffuse and difficult to trace. Politicians may frame policies, but administrators and civil servants implement them and have the operational responsibility, thereby creating potentially influential policy outcomes. Furthermore, the framing of the crisis is important, but may change over time, creating even more accountability problems.

Lastly, professional accountability is often an important dimension. In situations of uncertainly, and often when 'new' risks arise, there are frequently disagreements concerning who has the relevant type of expertise, which type of (scientific) knowledge is the most reliable, and questions as to whether experts are independent (Bovens & 't Hart, 1996; Jasanoff, 1990). Conflicts between different groups of experts are not uncommon. In these kinds of situations the question on professional accountability is especially difficult to handle.

6. Conclusions

Our chapter has revealed that organization for internal security and safety is a struggle concerning both attention to and definition of relevant problems and organizational solutions. A central theme has been the establishment of permanent organizations that can address the wicked inter-organizational issues in a coordinated and continuous manner, with enough resources and capacity. We have also seen that the definition of problems and solutions varies over time, and is affected by executive design, historical-institutional constraints, negotiations and external pressures.

The established principles concerning liability, decentralization and conformity represent different forms of specialization, but have little to offer when it comes to coordination, a particular pressing problem within the field. The principle of conformity seems particularly difficult to practice. The expectation that the organization model in extreme crisis situations is to be similar to a normal situation is difficult to live up to. Crises and disasters are in their being unexpected and surprising situations, where established organizational forms often prove inadequate. Generally, there is an urgent need for improvisation, rapid and flexible response. Often, established hierarchical structures, lines of command and competence areas are overstepped.

The principle of liability is also problematic, whereas crises and disasters are exceedingly 'wicked', in the sense that they typically cross established organizational borders. Increasingly, successful crisis management has to be performed at the interface between organizations and levels of administration. The principle of liability establishes responsibility within single organizations, but represents an obstruction for the coordination problems in a larger crisis situation. The principle of decentralization may also represent a problem, whereas crisis situations often demand a balancing of the need for a flexible scope of action at the local level, and the need for central control and leadership. A central

question is therefore how to handle the demand for more hierarchical control and leadership when major crises, disasters or risks threat society.

Several studies have shown that military command and control methods in crisis management can be problematic (Boin et al., 2008). In crises of a larger magnitude, and where the issue crosses traditional institutional borders, crisis management may take place in the shadows of formal hierarchy, especially since there often is a need for flexibility, improvisation and network cooperation. Traditional risk management is often focused on single issues, neglecting the risk of complex and wicked issues.

The two main doctrines in the Norwegian government system, the principle of ministerial responsibility and the principle of local self government, have quite clearly set limitations to the efforts to solve cross-cutting coordination problems. Our analysis shows that these doctrines stand firm. Attempts to establish a strong coordination at the centre over the last years have largely failed. Instead, we have seen a rather cautious upgrading of the over-arching and coordinative responsibilities of the Ministry of Justice and its underlying agencies. New organizational arrangements have been set up as more ad hoc or virtual organizations without permanent resources attached.

Organizing for internal security and safety is a constant struggle to gain sufficient attention. It concerns definitions of what are the relevant problems and organizational solutions, and has been concentrated on discussions on the establishment of permanent bodies that may work continuously within the field and with sufficient resources. Definitions of and solutions vary across time, and are affected by traditions and interest of the different actors involved. What we see is a mixture of rational design, negotiations, administrative cultural constraints and adaptation to external changes and pressure. There is no simple explanation to the reorganization processes. Our argument is that the reforms in this policy field are based on a combination of different driving forces (Fimreite, Lægreid & Rykkja, 2011). Both instrumental and institutional approaches contribute to the understanding of the reorganization process. Single-factor explanations face considerable problems when their claims are confronted with empirical data (Lægreid & Verhoest, 2010). The organization of this policy field is becoming increasingly complex with different organizational principles resulting from multiple factors working together in a complex mix. It is an example of a compound administrative reform process, which is multi-dimensional and represents 'mixed' orders, and a combination of competing, inconsistent and partly contradictory organizational principles and structures that co-exist and balances different interests and values (Olsen 2010).

The field of internal security and safety in Norway has developed and changed since the end of the Cold War, and has gained new focus and attention. The development of a new coordination policy is characterized by the shift from a military to stronger civil focus. But despite the radical proposals from the Vulnerability Commission and Infrastructure Commission, changes have been small and incremental. The tsunami disaster in 2004 resulted in a new effort in the reorganization process, but largely only led to slight changes to the established responsibility relationships. The development on this field is distinct to other reforms in Norway, for instance the hospital reform and the Norwegian Labour and Welfare Service reform, where organizational changes have been more radical (Christensen & Lægreid, 2010). In these cases the government managed to implement large structural reforms. An important difference is that these reforms did not raise cross-sector issues to the

same degree. Both reforms happened largely within closed sectors. This confirms the strengths of the separate policy sectors in the Norwegian case.

A strengthening of the central governmental core within the internal security and safety area has proved difficult in the Norwegian case. However, risk management can not be based on an excessive belief in hierarchy, command and control. To manage internal security issues by establishing central meta-organizations, such as the Department of Homeland Security in the US in 2003, do not necessarily reduce risk-levels (Peters, 2004).

In spite of the reluctant organizational changes, our study illustrates that the field is still very fragmented. Responsibility is divided among several actors at different levels, at different levels, and within different sectors and organizational settings. There is a mismatch between the strong specialization by sector, administrative apparatus, and a policy field that does not follow traditional sector lines. A strengthening of the horizontal coordination between the ministries might be necessary to handle the need for better integration and horizontal coordination. The dominant steering principles and in the Norwegian system, are still fundamental cornerstones, and have not been altered. Thus, any change within the field will take time, or may possibly require the effects of a major and path-breaking disaster hitting Norway more directly.

7. References

Baldwin, R. & Cave, M. (1999). *Understanding Regulation,* Oxford, Oxford University Press.

Baumgartner, F.R. & Jones, B.D. (1993). *Agendas and Instability in American Politics,* Chicago, IL: University of Chicago Press.

Beck, U. (1992). *Risk Society. Towards a New Modernity,* London, SAGE Publications.

Bjørgum, L. (2010). *Samordning og samvirke for samfunnssikkerhet – en studie av prosessen rundt St.meld. nr. 22 (2007– 2008) Samfunnssikkerhet, samvirke og samordning*, Master thesis, Institutt for administrasjon og organisasjonsvitenskap, Universitetet i Bergen.

Bogdanor, V. (2004). *Joined up Government*, Oxford: Oxford University Press.

Boin, A.; McConnell, A. & 't Hart, P. (eds.) (2008). *Governing after Crisis: The Politics of Investigation, Accountability and Learning,* Cambridge, Cambridge University Press.

Bouckaert, G.; Ormond, D. & Peters, B.G. (2000). *A Potential Governance Agenda for Finland*, Helsinki, Finansdepartementet.

Bouckaert, G.; Peters, B.G. & Verhoest, K. (2010). *The Coordination of Public Sector Organizations. Shifting Patterns of Public Management*, Houndmills, Basingstoke: Palgrave Macmillan.

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework, *European Law Journal,* Vol.13, No.4**,** pp. 447–468.

Bovens, M. & 't Hart, P. (1996). *Understanding Policy Fiascoes.* New Brunswick, N.J.: Transaction Publishers.

Burgess, J.P. & Mouhleb, N. (2007). *Societal Security. Definitions and Scope for the Norwegian Setting*, Policy Brief No. 2/2007, Oslo, PRIO.

Brunsson, M. & Olsen, J.P. (1993). *The Reforming Organization,* London, Routledge.

Brygård, M.C. (2006). *Ny struktur for krisehåndtering – en analyse av beslutningsprosessen som ligger til grunn for opprettelsen av Regjeringens kriseråd og Krisestøtteenheten*, Master thesis, Institutt for Statsvitenskap, Universitetet i Oslo.

Christensen, T. & Lægreid, P. (1998). Administrative Reform Policy: The Case of Norway, *International Review of Administrative Sciences*, Vol.64, No.3, pp. 457–475.

Christensen, T. & Lægreid, P. (2002). New Public Management: Puzzles of Democracy and the Influence of Citizens, *The Journal of Political Philosophy,* Vol.10, No.3**,** pp. 167–295.

Christensen, T. & Lægreid, P. (eds.) (2007). *Transcending New Public Management,* Aldershot, Ashgate.

Christensen, T. & Lægreid, P. (2008). The Challenge of Coordination in Central Government Organizations: The Norwegian Case, *Public Organization Review,* Vol.8, No.2**,** pp. 97–116.

Christensen, T. & Lægreid, P. (2010). Increased complexity in public organizations – the challenges of combining NPM and post-NPM features, In: Lægreid, P. & Verhoest, K. (eds.) *Governance of Public Sector Organizations. Proliferation, autonomy and performance,* London, Palgrave Macmillan.

Christensen, T.; Fimreite, A.L. & Lægreid, P. (2011). Crisis Management – The Case of Internal Security in Norway, *Administration and Society*. Forthcoming.

Christensen, T.; Lægreid, P.; Roness, P.G. & Røvik, K.A. (2004). *Organisasjonsteori for offentlig sektor,* Oslo, Universitetsforlaget.

Cyert, R.M. & March, J.G. (1963). *A Behavioral Theory of the Firm,* Engelwood Cliffs, N.J., Prentice Hall.

Czarniawska, B. (ed.) (2009). *Organizing in the Face of Risk and Threat,* Cheltenham, Edward Elgar.

DiMaggio, P. & Powell, W.W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields, *American Sociological Review,* Vol. 48, No.2**,** pp. 147–160.

Dyndal, G.L. (ed.) (2010). *Strategisk ledelse i krise og krig,* Bergen, Fagbokforlaget.

Egeberg, M. (2003). How Bureaucratic Structure Matters: An Organizational Perspective, In: Peters, B.G. & Pierre, J. (eds.) *Handbook of Public Administration*, London, Sage.

Egeberg, M. (2004). An organizational approach to European integration. Outline of a complementary approach, *European Journal of Political Research,* Vol.43, No.2, pp. 199–214.

Fimreite, A.L.; Flo, Y. & Tranvik, T. (2002). *Lokalt handlingsrom og nasjonal integrasjon: kommuneideologiske brytninger i Norge i et historisk perspektiv*, Oslo, Unipub.

Fimreite, A.L.; Lango, P.; Lægreid, P. & Rykkja, L.H. (eds.) (2011). *Organisering, samfunnssikkerhet og krisehåndtering,* Oslo, Universitetsforlaget.

Fimreite, A.L.; Lægreid, P. & Rykkja, L.H. (2011). Organisering for samfunnssikkerhet og krisehåndtering, In: Fimreite, A.L.; Lango, P.; Lægreid, P. & Rykkja, L.H. (eds.), *Organisering, samfunnssikkerhet og krisehåndtering*, Oslo, Universitetsforlaget.

Flo, Y. (2004). *Staten og sjølvstyret. Ideologier og strategiar knytt til det lokale og regionale styringsverket etter 1900,* Bergen, Universitetet i Bergen.

Harmon, M.M. & Mayer, R.T. (1986). *Organization Theory of Public Administration,* Glenview, IL, Scott, Foresman & Co.

Høydal, H.R. (2007). Samordning av samfunnsikkerhet i norsk sentralforvaltning, *Notat*, 7/2007, Bergen, Uni Rokkansenteret.

Jacobsen, K.D. (1964). *Teknisk hjelp og politisk struktur,* Oslo, Universitetsforlaget.

Jaffery, L. & Lango, P. (2011). Flodbølgekatastrofen, In: Fimreite, A.L.; Lango, P.; Lægreid, P. & Rykkja, L.H. (eds.), *Organisering, samfunnssikkerhet og krisehåndtering*, Oslo, Universitetsforlaget.

Jasanoff, S. (1990). *The Fifth Branch: Science Advisors as Policymakers,* Cambridge, Mass., Harvard University Press.

Kettl, D.F. (2004). *System under stress, homeland security and American politics,* Washington, DC, CQ Press.

Krasner, S.D. (1988). Sovereginity – An institutional perspective, *Comparative Political Studies,* Vol.12, No.1**,** pp. 66–94.

La Porte, T.R. (1996). High reliability organizations: unlikely, demanding and at risk, *Journal of Contingencies and Crisis Management,* Vol.4, No.2**,** pp. 61–71.

Lango, P. & Lægreid P. (2011). Samordning for samfunnssikkerhet, In: Fimreite, A.L.; Lango, P.; Lægreid, P. & Rykkja, L.H. (eds.), *Organisering, samfunnssikkerhet og krisehåndtering*, Oslo, Universitetsforlaget.

Lægreid, P. & Serigstad, S. (2006). Framing the Field of Homeland Security: The Case of Norway, *Journal of Management Studies,* Vol.43, No.6**,** pp. 1395–1413.

Lægreid, P. & Verhoest, K. (eds.) (2010). *Governance of public sector organizations. Proliferation, autonomy and performance*, Basingstoke, Palgrave Macmillan.

Lervåg, K. (2010). *Tilsyn uten ansvar – en studie av offentlige myndigheters regulering av Vest Tank før og etter eksplosjonsulykken ved bedriften den 24. mai 2007*, Master thesis Institutt for administrasjon og organisasjonsvitenskap, Universitetet i Bergen.

March, J.G. & Olsen, J.P. (1983). Organizing Political Life: What Administrative reorganization Tells Us About Governance, *American Political Science Review,* Vol.77, No.2**,** pp. 281–296.

March, J.G. & Olsen, J.P. (1989). *Rediscovering Institutions. The Organizational Basis of Politics,* New York, The Free Press.

March, J.G. & Olsen, J.P. (2006). The Logic of Appropriateness, In: Moran, M.; Rein, M. & Goodin, R.E. (eds.) *The Oxford Handbook of Public Policy,* Oxford, Oxford University Press.

Meyer, J. & Rowan, B. (1977). Institutional Organizations: Formal Structure as Myth and Cermony, *American Journal of Sociology,* Vol.83, No.2**,** pp. 340–363.

Meyer, S. (2009). *Ingen fare for den med god samvittighet? Terrorlovgivningens dilemmaer,* Oslo, Cappelen Damm.

NOU (2000: 24). *Et sårbart samfunn – Utfordringer for sikkerhets- og beredskapsarbeidet i samfunnet*, Official Norwegian Reports, Oslo, Justis- og politidepartementet.

NOU (2006: 6). *Når sikkerheten er viktigst – Beskyttelse av landets kritiske infrastrukturer og kritiske samfunnsfunksjoner*, Official Norwegian Reports, Oslo, Justis- og politidepartementet.

Olsen, J.P. (1989). *Petroleum og politikk,* Oslo, TANO.

Olsen, J.P. (1992). Analyzing institutional dynamics, *Staatswissenchaften und Staatspraxis*, Vol.3, pp. 247–271.

Olsen, J.P. (2010). *Governing through institution bulilding*, Oxford, Oxford University Press.

Olsen, O.E.; Kruke, B.I. & Hovden, J. (2007). Societal Safety: Concept, Borders and Dilemmas, *Journal of Contingencies and Crisis Management,* Vol.15, No.2**,** pp. 69–79.

Perrow, C. (1984). *Normal Accidents: Living with High-Risk Technologies,* New York, Basic Books.

Perrow, C. (2007). *The next catastrophe: reducing our vulnerabilities to natural, industrial, and terrorist disasters,* Princeton, Princeton University Press.

Peters, B.G. (2004). Are we safer today?: Organizational responses to terrorism, In: Groty, W. (ed.) *The Politics of terror. The US response to 9/11,* Boston, Northeastern University Press.

Power, M. (2004). *The Risk Management of Everything,* London, Demos.

Riksrevisjonen (2007). *Riksrevisjonens undersøkelse av måloppnåelse og effektivitet ved fylkesmannsembetene.* Dokument nr. 3:14 (2006–2007), The Office of the Auditor General's investigation, Oslo, Riksrevisjonen.

Riksrevisjonen (2008). *Riksrevisjonens undersøkelse av Justisdepartementets samordningsansvar for samfunnssikkerhet*. Dokument 3:4 (2007–2008), The Office of the Auditor General's investigation, Oslo, Riksrevisjonen.

Rochefort, D.A. & Cobb R.W. (1994). *The Politics of Problem Definition*, Lawrence, University Press of Kansas.

Roness, P.G. (1997). *Organisasjonsendringar: Teoriar og strategiar for studier av endringsprosessar,* Bergen, Fagbokforlaget.

Rothstein, B. & Stolle, D. (2003). Introduction: Social Capital in Scandinavia, *Scandinavian Political Studies*, Vol.26, No.1, pp. 1–26.

Rykkja, L.H. (2008). *Matkontroll i Europa: – en studie av regulering i fem europeiske land og EU*, Institutt for administrasjon og organisasjonsvitenskap, Universitetet i Bergen.

Rykkja, L.H. (2011). Fylkesmannen som samordningsinstans, In: Fimreite, A.L.; Lango, P.; Lægreid, P. & Rykkja, L.H. (eds.), *Organisering, samfunnssikkerhet og krisehåndtering*, Oslo, Universitetsforlaget.

Rykkja, L.H.; Lægreid, P. & Fimreite, A.L. (2011). Attitudes towards Anti-terror Measures: The Role of Trust, Political Orientation and Civil Liberties Support, *Critical Studies on Terrorism*, Vol.4, No.2. Forthcoming.

Selznick, P. (1957). *Leadership in Administration,* New York, Harper and Row.

Serigstad, S. (2003). Samordning og samfunnstryggleik: Ein studie av den sentrale tryggleiks-og beredskapsforvaltninga i Norge i perioden 1999–2002, *Rapport*, 16/2003, Bergen, Uni Rokkansenteret.

Simon, H.A. (1976). *Administrative Behavior: A Study of Decision-making Processes in Administrative Organization,* New York, Macmilliam.

Smith, D. (2006). Crisis Management – Practice in Search of a Paradigm, In: Smith, D. & Elliott, D. (eds.) *Key Readings in Crisis Management – Systems and Structures for Prevention and Recovery* London/New York, Routledge.

St.meld. nr. 24 (1992–1993). *Det fremtidige sivile beredskap*, White Paper, Oslo.

St.meld. nr. 17 (2001–2002). *Samfunnssikkerhet – Veien til et mindre sårbart samfunn*, White Paper, Oslo, Justis- og politidepartementet.

St.meld. nr. 37 (2004–2005). *Flodbølgekatastrofen i Sør-Asia og sentral krisehåndtering,* White Paper, Oslo, Justis- og politidepartementet.

St.meld. nr. 22 (2007–2008). *Samfunnssikkerhet*, White Paper, Oslo, Justis- og politidepartementet.

Tranvik, T. & Fimreite, A. L. (2006). Reform failure: The processes of devolution and centralization in Norway, *Local Government Studies*, Vol. 32, No.1, pp. 89–107.

10

System Building for Safe Medication

Hui-Po Wang, Jang-Feng Lian and Chun-Li Wang
School of Pharmacy, College of Pharmacy, Taipei Medical University,Taiwan, Republic of China

1. Introduction

This article aims to report (1) the scientific aspects of system biology that governs the mechanism of xenobiotics-host interaction; (2) the beauty and the odds of xenobiotics in the biological system; (3) integrative risk-benefit assessment on using xenobiotics for medication purpose; (4) global trend of conceptual change in risk management from product-oriented pharmacovigilance to proactive pharmacovigilance planning for risk minimization; (5) summary of public information regarding to potential risk underlying co-medication of licensed drugs with complementary/alternative medicine (CAM), traditional Chinese medicine (TCM) and nutraceuticals; (6) epidemiological aspects in co-medication of licensed drugs with herbal medicine; and (7) opinion on system building for safe medication in societies where irrational medication and co-medication is prevalent.

2. System biology

2.1 The biological system

The biological system is full of mechanisms in manipulating the action and the destination of xenobiotics, i. e. drugs and food, in the body. Mechanisms governing the xenobiotic-host interaction include absorption, distribution, metabolism and excretion (ADME, Fig. 1). Typical examples associated with xenobiotic-host interaction are the change of drug efficacy due to the competition of drugs and food in intestinal absorption, the interference of drugs or food in the rate and the profile of metabolism, modification of drug distribution by food or other drugs, the change of renal clearance due to the competition of food and drugs for excretion transporters in the kidney, and the occurrence of drug resistance due to the modification of ADME process (Wishart, 2007).

2.2 Evidence-based medicine

The biological activity, i. e. the pharmacodynamic outcome, is used to be the major concern in conventional drug research and development. Pharmacokinetic (PK) evaluation, the descriptor of drug-host interaction, is usually conducted at the later stage of drug development. However, the disposition of the biological active substances in the body system determines the success of these substances to become therapeutic agents. As a consequence, the successful rate of bringing chemical entities from preclinical to clinical stage was rather low, estimated to be 1/2000 (Nassar-1, Nassar-2, 2004). The failure in most cases is due to the unsatisfactory PK after the chemical entities enter the biological system (Fig. 2) (Grossman, 2009).

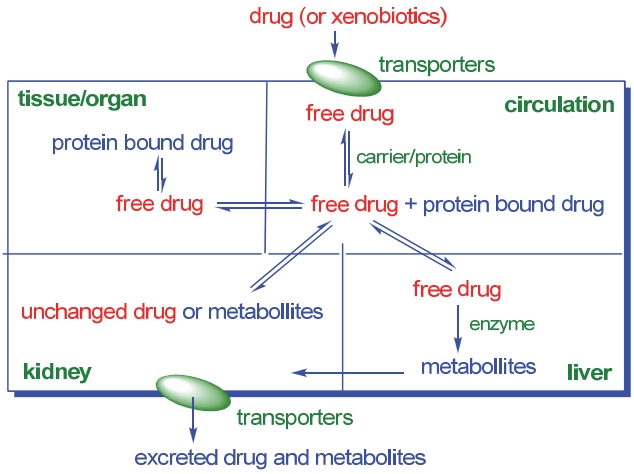


Fig. 1. ADME determines the destination of xenobiotics in biological system.

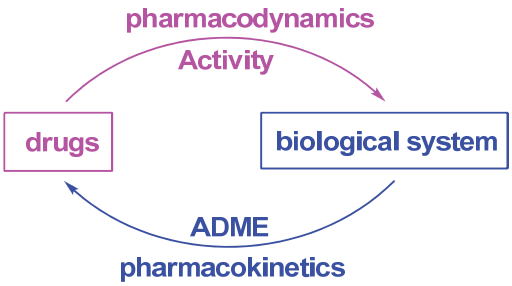


Fig. 2. Integrative pharmacodynamic and pharmacokinetic outcome determines the therapeutic efficacy of drugs.

3. The beauty and odds of xenobiotics in the biological system

Biological processing of xenobiotics via ADME determines the feasibility of medicinal substances to become effective therapeutic agents (Eddershaw et al., 2000; Ekins et al., 2010; Lombardo & Waters, 2011; Ruiz-Garcia et al., 2008). Factors affecting the fate of xenobiotics may exist anywhere along the ADME process and may lead to a change of well designed and documented pharmacokinetic profiles of registered pharmaceuticals (Harris et al., 2003; Yang C. Y. et al., 2006). Risk and benefit assessment is thus not only on the medicinal substances *per se*, but also on factors affecting the biological processing of these substances.

3.1 The sites and the mechanisms of xenobiotic–host Interaction

Scientific evidences regarding to the sites and mechanisms of xenobiotic–host interaction are emerging. It is well documented that transporters in the intestine, liver, kidney and brain are involved in the uptake and the efflux of chemical substances like food and drugs (Brandsch et al., 2008; Oostendorp et al., 2009; Rubio-Aliaga & Daniel, 2008; Yang et al., 2006; Zhou, 2008). The pharmacological effect and the disposition of drugs are thus highly influenced by the function of transporters located in specific tissues (Ayrton & Morgan, 2008; Calcagno et al., 2007; Türk & Szakács, 2009; Yuan et al., 2008). Evidence also supported the consequence of the involvement of transport proteins in the pharmacokinetic variability and the safety of drugs in human use (Tsuji, 2006).

3.2 Drug-drug and drug-food interaction along biological processing of xenobiotics

Reports demonstrated that transporters in the intestine for absorption and in the kidney for excretion showed characteristics of broad substrate specificity, indicating the possibility of drug-drug and drug-food interactions. The pitfalls of transporter-mediated drug-drug, drug-food or drug-herbal interaction is thus an important issue to be elaborated for drug safety concern (Huang & Lesko, 2004; Pal & Mitra, 2006; Ward, 2008). Kidney, for example, is one of the important sites of drug-drug and drug-food interaction. The competition of renal transporter between drugs and food may change the bioavailability of drugs due to the change of renal clearance rate (Bachmakov et al, 2009; Kindla et al., 2009; Li et al., 2006; Tsuda et al., 2009; Wojcikowski, 2004). Thus a predictable ADME-toxicity modulation is important in the process along drug development (Szakács et al., 2008).

The metabolic system processing the biotransformation of xenobiotics provides another pitfalls for drug-drug and drug-food interaction (Tirona & Bailey, 2006). Reports indicated that hepatotoxicity (Brazier & Levine, 2003; Furbee et al., 2006; Holt & Ju, 2006; Schiano, 2003; Tang, 2007; Wang et al., 2006) and renal toxicity (Wojcikowski et al., 2004) of xenobiotics are associated with the formation of reactive metabolites no matter they are from synthetic or herbal resources (Venkatakrishnan & Obach, 2007; Zhou et al., 2007).

3.3 Risk-benefit assessment of pharmaceutical products

As potential risks in relation to the administration of xenobiotics are frequently reported, the biological activity is not the only criteria for the justification of medicinal substances for therapeutic use. The integrative judgment of medicinal substance-host interaction based on the quality, safety and efficacy is essential for risk-benefit assessment in drug approval. In order to increase the successful rate, strategy in new drug development is thus evolved from the conventional sequential involvement of chemistry, pharmacodynamics (PD), toxicity (tox) and pharmacokinetics (ADME/PK) (Fig. 3a) to parallel PD/PK assessment (Fig. 3b) for optimizing drug efficacy. Novel approaches are using biological ADME mechanism for new drug design at early stage of drug discovery (Fig. 3c) (Dingemanse & Appel-Dingemanse, 2007). Evidence-based justification of drug-drug and drug-food interaction also becomes a standard procedure for safety evaluation of new drug application by pharmaceutical regulatory bodies (Hartford et al., 2006; Zhang et al., 2008).

3.4 Pharmacovigilance

Genetic and culture differences such as food and nutritional intake are among the factors that influence the therapeutic outcome of drugs. Therefore, safety evaluation of marketed

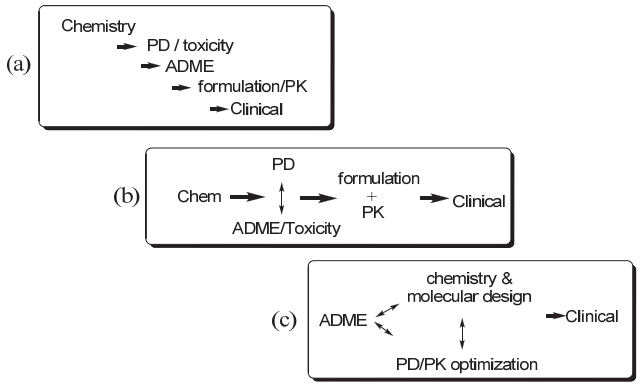


Fig. 3. The evolution of strategies in drug development from (a) sequential involvement of PD, ADME /PK to (b) PD and ADME /PK abreast and to (c) ADME for new drug design.

drugs should be based on good quality of evidence of the growing population that take the drug after a reasonably long period of time (Laupacis et al, 2002). In order to overcome the fragmentation of information, pharmacovigilance requires comprehensive risk-benefit assessment based on the accumulated data of the population using the individual pharmaceutical product (McFarlance, 2002).

4. Potential risk from co-medication

4.1 Polypharmacy

Polypharmacy is widespread in the general population, especially in the elderly. Besides registered medicine, the population of CAM users is growing, especially in the aged and in patients with chronic disease (Chung et al., 2009; Desai & Grossberg, 2003; Kennedy, 2005; McKenna & Killoury, 2010; Miller et al., 2008; Nowack et al., 2009; Ohama et al., 2006; Ramage-Morin, 2009). The most prevalent use of CAM are for treating cardiovascular disease, pain healing, cancer adjuvant therapy and obesity (Izzo, 2005). According to a questionnaire-based survey research on CAM use, 55% of the 356 patients registered in hospital emergency department have tried at least one CAM therapy within the past 12 months, 17% have tried CAM for their presenting medical problem (Li et al., 2004).

A considerable large portion of patients take CAM with registered medicines without notification to professionals. Therefore, standard tools for regular monitoring of pharmacovigilance have its limitation. Safety threat as a result of drug-CAM interaction emerges from various scientific and pharmacoepidemiological reports (Anastasi et al., 2011; Balbino & Dias, 2010; Chiang et al., 2005; Cockayne et al., 2005; Sim & Levine, 2010; Smith et al., 2011; Tarirai et al., 2010). As it is not evidence-based, risk from polypharmacy especially from co-medication of prescribed drugs with CAM is inevitable. A UK perspective report raised an increasing awareness of herbal use and the need to develop pharmacovigilance practice (Barnes, 2003).

4.2 Social aspects in relation to the risk of medication and polypharmacy

Polypharmacy implies a potential risk of pharmacovigilance in societies where co-medication is prevalent. Taiwan for example is known for its outstanding national health

insurance program which benefits 99% of the population. The welfare-like program rendered Taiwanese a potential overuse of the healthcare system, as indicated by the high physician's visit per person and the large number of drug items per prescription (Table 1) (Department of Health, 2008; Gau, 2007; Huang, & Lai, 2006; Hsu et al., 2004). Moreover, most of the prescriptions are massively dispensed in hospitals, with a released rate of 0.41%(year 2008) to community pharmacies on refills for patients with chronic disease (Bureau of National Health Insurance, Department of Health, 2011). The imbalanced distribution of pharmacy service between hospitals, clinics and community pharmacies further reflects the lack of mechanism for risk prevention on medication (Table 2).

	Taiwan	OECD countries
physician's visits (no. of visits/person/year)	15.2	5.9
Drug items per prescription	4.2	1.9
Drug expenditure to total national health insurance cost	25%	~15%

Table 1. Statistics of medication profile in Taiwan. Data of year 2008 are from National Health Insurance Database.

	Number of prescriptions	Number of pharmacists	Prescriptions dispensed /pharmacist/day
Medical Center	31,172,000	725	154
Regional Hospital	34,368,000	880	139
Local hospital	35,137,000	770	160
Clinics	217,052,000	8,404	91
Community Pharmacy	31,290,000	3,348	33

Table 2. Distribution of prescriptions to pharmacy for dispensing in Taiwan. Data of year 2008 are from National Health Insurance Database.

4.3 Regulatory aspects in relation to the risk of polypharmacy

CAM are marketed without license in most of the developed countries. Claims for therapeutic efficacy of CAM are thus prohibited or limited to authorized indications (World Health Organization, 2001 & 2004; Ziker, 2005). However, traditional Chinese medicine (TCM) are classified as licensed drugs in oriental societies. For example, TCM are separately registered from conventional pharmaceutical products via bilateral regulatory systems in Taiwan. Drug adverse events are managed via bilateral reporting systems as well. With the requirement of good manufacturing practice (GMP), the number of license issued to conventional medicine decreased drastically. The number of TCM license, on the other hand, increased with a significantly high growth rate (Table 3). The separation of regulatory and administrative management on conventional medicine and TCM leads to the fragmentation of information regarding to polypharmacy. Patients and consumers are thus facing an unknown risk from irrational co-medication.

	Conventional pharmaceutical products		TCM products	
year	prescription	over-the-counter	Prescription	over-the-counter
1995	14718	7152	2394	4663
Total in 1995	21,870		7,075	
2006	4235	1385	4663	6444
Total in 2006	5,620		11,107	

Table 3. Licenses issued for conventional pharmaceutical products and TCM in Taiwan.

4.4 Pharmaco-epidemiological aspects in relation to the risk of polypharmacy

Herbal medicine includes TCM, CAM and nutraceuticals. With the prevalence of CAM use, inappropriate commercial advertisements in the media are also prevalent. According to a report of survey study in Taiwan, the identified illegal advertisement of products with therapeutic claims on cable TV counts for 12% of total healthcare related advertisements (183 out of 1591 cases), of which 41% goes to food and nutraceuticals and 15% goes to TCM (Fig. 4a). The illegal advertisement rate is even higher on radio, with TCM ranked the top (53%) followed by nutraceuticals (31%) (Fig. 4b). Most of the advertisements are claims for weight reduction and for the treatment of erectile dysfunction while are lack of evidence.

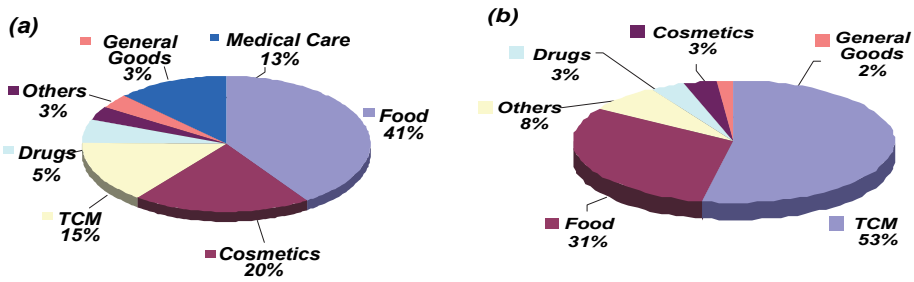


Fig. 4. Identified illegal advertisement of medicinal products in year 2004 on cable TV (a) and (b) radio in Taiwan (data are from Taiwan Drug Relief Foundation).

The incidence rate of end-stage renal disease (ESRD) of Taiwan ranked the top among the world (Fig. 5) (United States Renal Data System, 2006). The prevalence rate of ESRD in Taiwan raised from 1 per 2999 population in year 1991 to 1 per 498 population in 2006 (Fig. 6) (National Kidney Foundation, 2006). Reports indicated that herbal therapy was positively associated with chronic kidney disease (Bagnis et al., 2004; Chang et al., 2001; Chang et al., 2007; Guh et al., 2007; Nowack, 2008; Zhou et al., 2007). Safety issue in relation to polypharmacy becomes a challenge to the authority and the medical society.

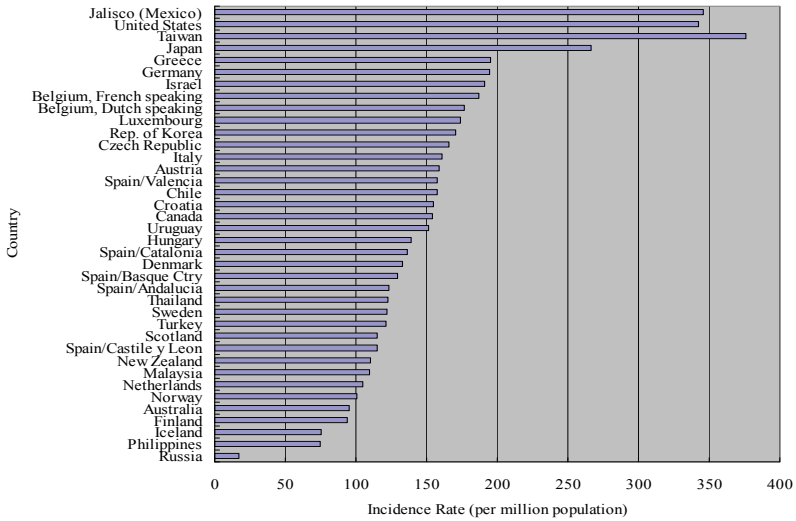


Fig. 5. The statistics of global incidence rate of end-stage renal disease (ESRD).

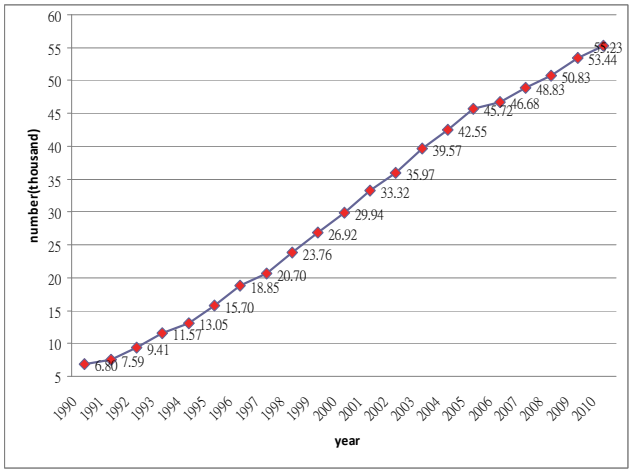


Fig. 6. The prevalence of end stage renal dialysis (ESRD) in Taiwan. Data are from the Bureau of National Health Insurance, Department of Health.

5. Risk management of medication

5.1 Global trend on risk management of pharmaceutical products

Two conceptual aspects regarding to risk management on medication were introduced by International Conference on Harmonization (ICH) (Bahri & Tsintis, 2005; Moseley, 2004; Tsintis & La Mache, 2004). Pharmacovigilance Specification (PV) addressed the evidence-based justification of drug safety throughout the life cycle of individual pharmaceutical

product from preclinical development to post-market use. Pharmacovigilance Planning (PVP) emphasizes risk prevention and minimization of medication use (Callréus, T. 2006; Cappe et al., 2006).

5.2 From pharmacovigilance to pharmacovigilance planning

Following the conceptual initiation of PVP, the Council for International Organizations of Medical Sciences (CIOMS) and ICH developed and published Topic ICH E2E Guidance in 2005 as an action to implement PVP (International Conference on Harmonization, 2005). The guidance addresses the identification of all possible signals of risk regarding to drug use. Evidence-based approaches to risk assessment, such as genetic/racial and cultural factors (food and nutrition), are included. Pharmaco-epidemiological study becomes important for risk analysis (Fig. 7).

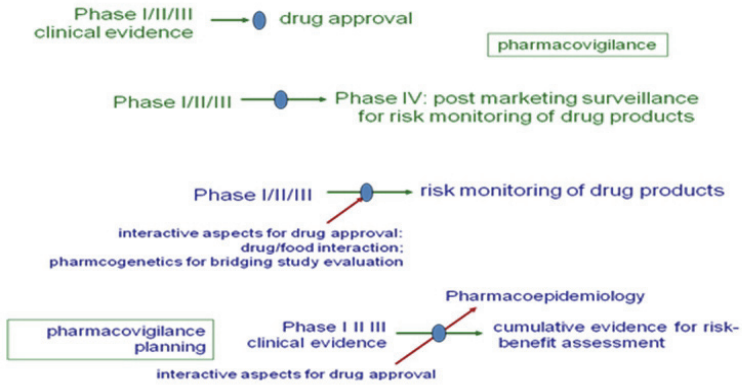


Fig. 7. The evolution of risk management of medication from product-oriented pharmacovigilance to risk management in pharmacovigilance planning.

5.3 System building for safe medication

The change from PV to PVP indicated the evolution from product-oriented risk management on individual medicine to a proactive risk prevention and minimization of medication. However, the risk management for pharmacovigilance initiated by ICH is essentially based on the refinement of safety-signal identification of registered pharmaceutical products. What is less addressed is the medicinal-type products without drug license. Risk prevention and minimization is thus difficult to be implemented in societies where patients tend to take conventional medicine and CAM without evidence-based justification in mind.

There is urgent need to call for public attention for the system building of safe medication. Risk and benefit assessment should be conducted on subjects who take all kinds of medicinal products via an un-biased integrative justification process. Humanity-based medication thus should be justified by the quality, safety and efficacy of medicines, no matter they are from synthetic, biological, biotechnological or herbal resources.

5.4 GDDP is essential for implementing pharmacovigilance planning

Following the guideline of Good Dispensing Practice (GDP), safe medication is fundamentally guaranteed for patients taking licensed pharmaceutical products. However, besides professional pharmacists, stakeholders involved in product and information delivery, namely product providers, medical professionals, the third party drug payers, media, patients and consumers, and policy makers in charge of food and drug administration, should also be responsible for the system building of safe medication. The concept of Good Dispensing and Delivery Practice (GDDP) is thus proposed. In this aspect, good practice in the delivery of medicinal products as well as medication information is equally important to good dispensing practice (Fig. 8). This is especially important in societies where the due process of safe medication is not properly implemented by the authority. For example, due to the lack of a due process in the separation of prescription from dispensing in Taiwan, irrational co-medication is common. A study on risk factor analysis of co-medication of cisapride and erythromycin identified that the major risk came from the mal-prescription of medical professionals (Gau et al., 2007).

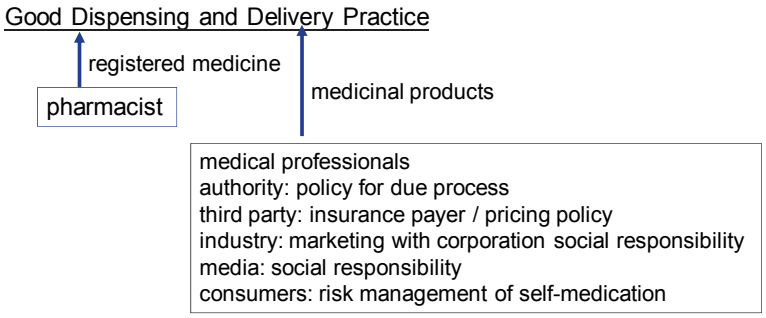


Fig. 8. Good Dispensing and Delivery Practice is essential for the system building of safe medication.

6. Conclusion

Risk of medication not only comes from registered drugs but also from irrational use and co-use of all types of products claiming therapeutic effect. Evidence-based medication is thus important for the system building of safe medication. The use of medicinal products needs to be evolved from pharmacovigilance of individual products to humanity-based integrative risk-benefit assessment for risk minimization. Although challenging the culture in societies prevalent of irrational medication and co-medication is most likely unwelcome, mechanism for consumer protection on system building for risk minimization need to be continuously addressed, proactively designed and pragmatically implemented.

7. Acknowledgement

This report comes from a study supported by the Department of Health, The Republic of China (grant no. DOH98-TD-D-113).

8. References

Anastasi, J. K., Chang, M. & Capili, B. (2011). Herbal Supplements: Talking with your Patients. *J. Nurse Pract.*, Vol.7, No.1, pp. 29-35.

Ayrton, A. & Morgan, P. (2008). Role of transport proteins in drug discovery and development: pharmaceutical perspective. *Xenobiotica*, Vol.38, No.7-8, pp. 676-708.

Bachmakov, I., Glaeser, H., Endress, B., Mörl, F., König, J. & Fromm M. F. (2009). Interaction of beta-blockers with the renal uptake transporter OCT2. *Diabetes Obes. Metab.*, Vol.11, No.11, pp. 1080-1083.

Bagnis, C. I., Deray, G., Baumelou, A., Le Quintrec, M. & Vanherweghem, J. L. (2004). Herbs and the kidney. *Am. J. Kidney Dis.*, Vol.44, No.1, pp. 1-11.

Bahri, P. & Tsintis, P. (2005). Pharmacovigilance-related topics at the level of the International Conference on Harmonisation (ICH). *Pharmacoepidemiol. Drug Saf.*, Vol.14, No.6, pp. 377-387.

Balbino, E. E. & Dias, M. F. (2010). Pharmacovigilance: A step towards the rational use of herbs and herbal *medicines. Brazilian J. Pharmacognosy*, Vol.20, No.6, pp. 992-1000.

Barnes, J. (2003). Pharmacovigilance of herbal medicines: A UK perspective. *Drug Saf.*, Vol.26, No.12, pp. 829-851.

Brandsch, M., Knütter, I. & Bosse-Doenecke E. (2008). Pharmaceutical and pharmacological importance of peptide transporters. *J. Pharm. Pharmacol.*, Vol.60, No.5, pp. 543-585.

Brazier, N. C. & Levine, M. A. (2003). Drug-herb interaction among commonly used conventional medicines: A compendium for health care professionals. *Am. J. Ther.*, Vol.10, No.3, pp. 163-169.

Bureau of National Health Insurance, Department of Health (2011). Available from: http://www.nhi.gov.tw/webdata/webdata.aspx?menu=&menu_id=&wd_id=&webdata_id=3812

Calcagno, A. M., Kim, I. W., Wu, C. P., Shukla, S. & Ambudkar, S. V. (2007). ABC drug transporters as molecular targets for the prevention of multidrug resistance and drug-drug interactions. *Curr. Drug Del.,* Vol.4, No.4, pp. 324-333.

Callréus, T. (2006). Use of the dose, time, susceptibility (DoTS) classification scheme for adverse drug reactions in pharmacovigilance planning. *Drug Saf.*, Vol.29, No.7, pp. 557-566.

Cappe, S., Blackburn, S., Rosch, S. & Tsintis, P. (2006). Proactive planning in pharmacovigilance. *Good Clin. Practice J.*, Vol.13, No.6, pp. 14-17.

Chang, C. H., Wang, Y. M., Yang, A. H. & Chiang, S. S. (2001). Rapidly progressive interstitial renal fibrosis associated with Chinese herbal medications. *Am. J. Nephrol.*, Vol.21, No.6, pp. 441-448.

Chang, C. H., Yang, C. M. & Yang, A. H. (2007). Renal diagnosis of chronic hemodialysis patients with urinary tract transitional cell carcinoma in Taiwan. *Cancer*, Vol.109, No.8, pp. 1487-1492.

Chiang, H. M., Fang, S. H., Wen, K. C., Hsiu, S. L.,Tsai, S. Y., Hou, Y. C., Chi, Y. C. & Chao, P, D. (2005). Life-threatening interaction between the root extract of Pueraria lobata and methotrexate in rats. *Toxicol. Appl. Pharmacol.*, Vol.209, No.3, pp. 263-268.

Chung, V. C., Lau, C. H., Yeoh, E. K. & Griffiths, S. M. (2009). Age, chronic non-communicable disease and choice of traditional Chinese and western medicine outpatient services in a Chinese population. *BMC Health Serv. Res.,* Vol.9, No.207.

Cockayne, N. L., Duguid, M. & Shenfield, G. M. (2005). Health professionals rarely record history of complementary and alternative medicines. *Br. J. Clin. Pharmacol.*, Vol.59, No.2, pp. 254-258.

Department of Health, Executive Yuan database (2010). Available from: http://www.doh.gov.tw/CHT2006/DM/DM2_2_p02.aspx?class_no=440&now_fod_list_no=11468&level_no=1&doc_no=77184

Desai, A. K. & Grossberg, G. T. (2003). Herbals and botanicals in geriatric psychiatry. *Am. J. Geriatr. Psychiatry*, Vol.11, No.5, pp. 498-506.

Dingemanse, J. & Appel-Dingemanse, S. (2007). Integrated pharmacokinetics and pharmacodynamics in drug development. *Clin. Pharmacokinet.,* Vol.46, No.9, pp. 713-737.

Eddershaw, P. J., Beresford, A. P. & Bayliss, M. K. (2000). ADME/PK as part of a rational approach to drug discovery. *Drug Discov. Today*, Vol.5, No.9, pp. 409-414.

Ekins, S., Honeycutt, J. D. & Metz, J. T. (2010). Evolving molecules using multi-objective optimization: Applying to ADME/Tox. *Drug Discov. Today*, Vol.15, No.11-12, pp. 451-460.

Furbee, R. B., Barlotta, K. S., Allen, M. K. & Holstege, C. P. (2006). Hepatotoxicity associated with herbal products. *Clin. Lab. Med.*, Vol.26, No.1, pp. 227-241.

Gau, C. S., Chang, I. S., Wu, F. L. L., Yu, H. T., Huang, Y. W., Chi, C. L., Chien, S. Y., Lin, K.M., Liu, M.Y., Wang, H.P. (2007). Usage of the claim database of national health insurance programme for analysis of cisapride-erythromycin co-medication in Taiwan. *Pharmacoepidemiol. Drug Saf.* Vol.16, No.1, pp. 86-95.

Grossman, I. (2009). ADME pharmacogenetics: Current practices and future outlook. *Expert Opin. Drug Metab. Toxicol.,* Vol.5, No.5, pp. 449-462.

Guh, J. Y., Chen, H. C., Tsai, J. F. & Chuang, L. Y. (2007). Herbal therapy is associated with the risk of CKD in adults not using analgesics in Taiwan. *Am. J. Kidney Dis.*, Vol.49, No.5, pp. 626-633.

Harris, R. Z., Jang, G. R. & Tsunoda, S. (2003). Dietary effects on drug metabolism and transport. *Clin. Pharmacokinet.*, Vol.42, No.13, pp. 1071-1088.

Hartford, C. G., Petchel, K. S., Mickail, H., PerezGutthann, S., McHale, M., Grana, J. M. & Marquez, P. (2006). Pharmacovigilance during the pre-approval phases: An evolving pharmaceutical industry model in response to ICH E2E, CIOMS VI, FDA and EMEA/CHMP risk-management guidelines. *Drug Saf.*, Vol.29, No.8, pp. 657-673.

Holt, M. P. & Ju, C. (2006). Mechanisms of drug-induced liver injury. AAPS J. Vol.8, No.6, pp. E48-E54.

Hsu, Y. C., Huang W. F. & Cheng S. H. (2004). Inappropriate prescribing of non-narcotic analgesics in Taiwan NHI ambulatory visits. *Chin. Pharm. J.*, Vol.56, No.36, pp. 111-120.

Huang, S. M. & Lesko, L. J. (2004). Drug-drug, drug-dietary supplement, and drug-citrus fruit and other food interactions: What have we learned? *J. Clin. Pharmacol.*, Vol.44, No.6, pp. 559-569.

Huang, W. F. & Lai, I. C. (2006). Potentially inappropriate prescribing for insomnia in elderly outpatients in Taiwan. *Int. J. Clin. Pharmacol. Ther.*, Vol.44, No.7, pp. 335-342.

International Conference on Harmonization. (2005). Guidance on E2E pharmacovigilance planning; availability. Notice. *Fed. Regist.* Vol.70, No.62, pp. 16827-16828.

Izzo A. A. (2005). Herb-drug interactions: An overview of the clinical evidence. *Fundam. Clin. Pharmacol.*, Vol.19, No.1, pp. 1-16.

Kennedy, J. (2005). Herb and supplement use in the US adult population. *Clin. Ther.*, Vol.27, No.11, pp. 1847-1858.

Kindla, J., Fromm, M. F. & König, J. (2009). In vitro evidence for the role of OATP and OCT uptake transporters in drug-drug interactions. *Expert Opin. Drug Metab. Toxicol.*, Vol.5, No.5, pp. 489-500.

Laupacis, A., Anderson, G. & O'Brien, B. (2002). Drug policy: making effective drugs available without bankrupting the healthcare system. *Healthc Pap,* Vol.3, No.1, pp. 12-30.

Li, J. Z., Quinn, J. V., McCulloch, C. E., Jacobs, B. P. & Chan, P. V. (2004). Patterns of complementary and alternative medicine use in ED patients and its association with health care utilization. *Am. J. Emerg. Med.*, Vol.22, No.3, pp. 187-191.

Li, M., Anderson, G. & Wang, J. (2006). Drug-drug interactions involving membrane transporters in the human kidney. *Expert Opin. Drug Metab. Toxicol.,* Vol.2, No.4, pp. 505-532.

Lombardo, F. & Waters, N. J. (2011). Drug design from the ADME/PK perspective: Does chemical intuition suffice in multifaceted drug discovery? *Curr. Top. Med. Chem.,* Vol.11, No.4, pp. 331-333.

McFarlance, A. (2002). Drug policy: the challenge is to overcome fragmentation. *Healthc Pap,* Vol.3, No.1, pp. 38-42.

McKenna, F. & Killoury, F. (2010). An investigation into the use of complementary and alternative medicine in an urban general practice. *Ir. Med. J.*, Vol.103, No.7, pp. 205-208.

Miller, M. F., Bellizzi, K. M., Sufian, M., Ambs, A. H., Goldstein, M. S., Ballard-Barbash, R. (2008). Dietary Supplement Use in Individuals Living with Cancer and Other Chronic Conditions: A Population-Based Study. *J. Am. Diet. Assoc.,* Vol.108, No.3, pp. 483-494.

Moseley, J. N. S. (2004). Risk management: A European regulatory perspective. *Drug Saf.*, Vol.27, No.8, 499-508.

Nassar, A. E. F., Kamel, A. M. & Clarimont, C. (2004). Improving the decision-making process in the structural modification of drug candidates: Enhancing metabolic stability. *Drug Discov. Today*, Vol.9, No.23, pp. 1020-1028.

Nassar, A. E. F., Kamel, A. M. & Clarimont, C. (2004). Improving the decision-making process in structural modification of drug candidates: Reducing toxicity. *Drug Discov. Today*, Vol.9, No.24, pp. 1055-1064.

Nowack, R. (2008). Herb-drug interactions in nephrology: Documented and theoretical. *Clin. Nephrol.,* Vol.69, No.5, pp. 319-325.

Nowack, R., Ballé, C., Birnkammer, F., Koch, W., Sessler, R. & Birck, R. (2009). Complementary and Alternative Medications Consumed by Renal Patients in Southern Germany. *J. Ren. Nutr.,* Vol.19, No.3, pp. 211-219.

Ohama, H., Ikeda, H. & Moriyama, H. (2006). Health foods and foods with health claims in Japan. *Toxicology*, Vol.221, No.1, pp. 95-111.

Oostendorp, R. L., Beijnen, J. H. & Schellens, J. H. M. (2009). The biological and clinical role of drug transporters at the intestinal barrier. *Cancer Treat. Rev.,* Vol.35, No.2, pp. 137-147.

Pal, D. & Mitra, A. K. (2006). MDR- and CYP3A4-mediated drug-herbal interactions. *Life Sci.*, Vol.78, No.18, pp. 2131-2145.

Ramage-Morin, P. L. (2009). Medication use among senior Canadians. *Health reports / Statistics Canada, Canadian Centre for Health Information,* Vol.20, No.1, pp. 37-44.

Rubio-Aliaga, I. & Daniel, H. (2008). Peptide transporters and their roles in physiological processes and drug disposition. *Xenobiotica*, Vol.38, No.78, pp. 1022-1042.

Ruiz-Garcia, A., Bermejo, M., Moss, A. & Casabo, V. G. (2008). Pharmacokinetics in drug discovery. *J. Pharm. Sci.,* Vol.97, No.2, pp. 654-690.

Schiano, T. D. (2003). Hepatotoxicity and complementary and alternative medicines. *Clin. Liver Dis.*, Vol.7, No.2, pp. 453-473.

Sim, S. N. & Levine, M. A. H. (2010). An evaluation of pharmacist and health food store retailer's knowledge regarding potentional drug interactions associated with St. John's wort. *Can. J. Clin. Pharmacol.,* Vol.17, No.1, pp. E57-E63.

Smith, C. A., Priest, R., Carmady, B., Bourchier, S. & Bensoussan, A. (2011). The ethics of traditional Chinese and western herbal medicine research: Views of researchers and human ethics committees in Australia. *Evid. Based Complement. Alternat. Med.,* Vol.2011, No.256915.

Szakács, G., Váradi, A., ÖzvegyLaczka, C. & Sarkadi, B. (2008). The role of ABC transporters in drug absorption, distribution, metabolism, excretion and toxicity (ADME-Tox). *Drug Discov. Today*, Vol.13, No.910, pp. 379-393.

Tang, W. (2007). Drug metabolite profiling and elucidation of drug-induced hepatotoxicity. *Expert Opin. Drug Metab. Toxicol.,* Vol.3, No.3, pp. 407-420.

Tarirai, C., Viljoen, A. M. & Hamman, J. H. (2010). Herb-drug pharmacokinetic interactions reviewed. *Expert Opin. Drug Metab. Toxicol.,* Vol.6, No.12, pp. 1515-1538.

Tirona, R. G. & Bailey, D. G. (2006). Herbal product-drug interactions mediated by induction. *Br. J. Clin. Pharmacol.*, Vol.61, No.6, pp. 677-681.

Tsintis, P. & La Mache, E. (2004). CIOMS and ICH initiatives in pharmacovigilance and risk management: overview and implications. *Drug Saf.*, Vol.27, No.8, pp. 509-517.

Tsuda, M., Terada, T., Ueba, M, Sato, T., Masuda, S., Katsura, T. & Inui, K. I. (2009). Involvement of human multidrug and toxin extrusion 1 in the drug interaction between cimetidine and metformin in renal epithelial cells. *J. Pharmacol. Exp. Ther.*, Vol.329, No.1, pp. 185-191.

Tsuji, A. (2006). Impact of transporter-mediated drug absorption, distribution, elimination and drug interactions in antimicrobial chemotherapy. *J. Infect. Chemother.*, Vol.12, No.5, pp. 241-250.

Türk, D. & Szakács, G. (2009). Relevance of multidrug resistance in the age of targeted therapy. *Curr. Opin.Drug Disc. Devel.,* Vol.12, No.2, pp. 246-252.

United States Renal Data System. (2006). Annual Data Report (ADR) Volume I: ATLAS of Chronic Kidney disease and end-stage renal disease in the United States, Available from http://www.usrds.org/

Venkatakrishnan, K. & Obach, R. S. (2007). Drug-drug Interactions via Mechanism-Based Cytochrome P450 Inactivation: Points to consider for risk assessment from In vitro data and clinical pharmacologic evaluation. *Curr. Drug Metab.*, Vol.8, No.5, pp. 449-462.

Wang, K., Mendy, A. J., Dai, G., He, L. & Wan, Y. J. Y. (2006). Retinoids activate the RXR/SXT-mediated pathway and induce the endogenous CYP3A4 activity in Huh7 human hepatoma cells. *Toxicol. Sci.*, Vol.92, No.1, pp. 51-60.

Ward, P. (2008). Importance of drug transporters in pharmacokinetics and drug safety. *Toxicol. Mech. Methods*, Vol.18, No.1, pp. 1-10.

Wishart, D. S. (2007). Improving early drug discovery through ADME modelling: An overview. *Drugs in R and D*, Vol.8, No.6, pp. 349-362.

Wojcikowski, K., Johnson, D. W. & Gobe G. (2004). Medicinal herbal extracts - Renal friend or foe? part one: The toxicities of medicinal herbs. *Nephrology*, Vol.9, No.5, pp. 313–318.

World Health Organization. (2001). Legal Status of Traditional medicine and Complementary/Alternative Medicine: a Worldwide Review. Geneva, document reference WHO/EDM/TRM/2001. 2.

World Health Organization. (2004). WHO guidelines for governments and consumers regarding the use of alternative therapies. *Pan Am. J. Public Health,* Vol.16, No.3, pp. 218-221.

Yang, C. Y., Chao, P. D. L., Hou. Y. C., Tsai, S. Y., Wen, K. C. & Hsiu, S. L. (2006). Marked decrease of cyclosporin bioavailability caused by coadministration of ginkgo and onion in rats. *Food Chem. Toxicol.*, Vol.44, No.9, pp. 1572-1578.

Yang, H. Y., Lin, J. L., Chen, K. H., Yu, C. C., Hsu, P. Y. & Lin, C. L. (2006). Aristolochic acid-related nephropathy associated with the popular Chinese herb Xi Xin. *J. Nephrol.*, Vol.19, No.1, pp. 111-114.

Yuan, H., Li, X., Wu, J., Li, J., Qu, X., Xu, W. & Tang, W. (2008). Strategies to overcome or circumvent P-Glycoprotein mediated multidrug resistance. *Curr. Med. Chem.*, Vol.15, No.5, pp. 470-476.

Zhang, L., Zhang, Y., Strong, J. M., Reynolds, K. S. & Huang, S. M. (2008). A regulatory viewpoint on transporter-based drug interactions. *Xenobiotica*, Vol.38, No.,78, pp. 709-724.

Zhou, S. F. (2008). Structure, function and regulation of P-glycoprotein and its clinical relevance in drug disposition. *Xenobiotica*, Vol.38, No.78, pp. 802-832.

Zhou, S. F., Xue, C. C., Yu, X. Q. & Wang, G. (2007). Metabolic activation of herbal and dietary constituents and its clinical and toxicological implications: An update. *Curr. Drug Metab.*, Vol.8, No.6, pp. 526-553.

Ziker, D. (2005). What lies beneath: an examination of the underpinnings of dietary supplement safety regulation? *Am. J. Law Med.*, Vol.31, No.23, pp. 269-284.

11

Mental Fatigue Measurement Using EEG

Shyh-Yueh Cheng [1] and Hong-Te Hsu[2]
[1] *Department of Occupational Safety and Hygiene, Chia-Nan University of Pharmacy and Science,Tainan,*
[2] *Institute of Engineering Science and Technology, National Kaohsiung First University of Science and Technology,Kaohsiung,*
[1,2]*Taiwan,ROC*

1. Introduction

1.1 Background

We live in a highly technological and information-oriented society. The use of computers in modern society is omnipresent. They are used for innumerable applications in various sizes and forms. Over the past 20 years, the personal computer has become widely used, both in the office and at home. People use computers to write documents, maintain databases, manage finances, draw diagrams and graphics, make presentations, compile mailing lists, search computer databases, write application programs, use the Internet, and myriad other tasks. Since such work requires prolonged vigilance and mental activity with sedentary work, fatigue caused from visual display terminal(VDT) tasks frequently occurs in the workplace.

Fatigue is a major, but usually neglected, factor that increases the occurrence of performance errors and lapses. Fatigue, especially mental fatigue, is inevitable for office workers and in life in general. Fatigue is usually related to a loss of efficiency and disinclination to effort. It is also possible that cumulative mental fatigue leads to decreased productivity in the workplace and induces critical errors in the worst cases. Many experimental studies have demonstrated that mental fatigue induces deterioration in cognitive functions. Responses become slower, more variable, and more error prone after mental fatigue (Scheffers et al., 1999; Dorrian et al., 2000; Smith et al., 2002). The importance of adequate fatigue monitoring could be demonstrated by the Exxon Valdez oil tanker accident. The direct cause of this, America's worst oil spill, was a human performance error, which had been observed and cautioned about before; however, the warning had arrived too late in order to remedy the situation because the severely fatigued mate did not immediately respond to the warning (Dement and Vaughan, 1999). Deficits in perceptual processes after extended wakefulness are responsible for performance deficits.

Mental fatigue refers to the effects that people may experience after or during prolonged periods of cognitive activity. In this sense, it is a very common phenomenon in everyday modern life. Therefore, the management of mental fatigue is important from the viewpoint of occupational risk management, productivity, and occupational health.

1.2 Motivation

Until now, very little has been known about the psychophysiological mechanisms underlying mental fatigue. Here, this study was in an attempt to gain more insight in the

mechanisms that are central to mental fatigue and in arousal level and the cognitive functions that are most affected by mental fatigue. The assessment of mental fatigue should be conducted based on physiological evidences using both arousal level from EEG (Okogbaa et al., 1994) and cognitive information processing from ERP (Murata et al., 2005). These measures would provide reliable and effective evaluation of mental fatigue.

1.3 The objectives of this study

This study aimed to assess mental fatigue by using electroencephalographic measures and response tests in visual display terminal (VDT) tasks. The experimental design used by Murata et al. (2005) was adopted herein to evaluate mental fatigue using ERP. The combination of indices based on arousal level (EEG) and cognitive information processing (ERP) were employed to evaluate mental fatigue in this study. The objects of this study were included as following:

1. To explore the arousal level and cognitive function for mental fatigue in VDT tasks by using electroencephalographic measures.
2. To compare the behavior response (RT, ER) and physiological response (EEG, ERP) to mental fatigue in VDT tasks.
3. To examine the recovery state from mental fatigue after 180 min experimental tasks with 60 min period of rest.

2. EEG and ERP

2.1 Cerebrum

The cerebrum is the part of the brain that most people think of when the term brain is mentioned. Anatomically, the brain can be divided into three parts: the forebrain, midbrain, and hindbrain; the forebrain includes the several lobes of the cerebral cortex that control higher functions. The cerebrum has two cerebral hemispheres. A cerebral hemisphere (hemispherium cerebrale) is defined as one of the two regions of the brain that are delineated by the body's median plane. The brain can thus be described as being divided into left and right cerebral hemispheres. The cerebral cortex includes the frontal, temporal, occipital, and parietal lobes and the central sulcus (as depicted in Figure 2.1). The frontal lobes are positioned in front of (anterior to) the parietal lobes. The temporal lobes are located beneath and behind the frontal lobes. The occipital lobes located in the rearmost portion of the skull and behind the parietal lobes are the smallest of four true lobes in the human brain. The central sulcus separates the parietal lobe from the frontal lobe (Seeley et al., 2003).

1. The frontal lobe

The frontal lobe is an area in the brain of mammals located at the front of each cerebral hemisphere. In the human brain, the precentral gyrus and the related cortical tissue that folds into the central sulcus comprise the primary motor cortex, which controls voluntary movements of specific body parts associated with areas of the gyrus. The frontal lobes have been found to play a part in impulse control, judgment, language production, working memory, motor function, problem solving, sexual behavior, socialization, and spontaneity. The frontal lobes assist in planning, coordinating, controlling, and executing behavior. The so-called executive functions of the frontal lobes involve the ability to recognize future consequences resulting from current actions, to choose between good and bad actions (or better and best), override and suppress unacceptable social responses, and determine similarities and differences between things or events.

2. The parietal lobe

The parietal lobe integrates sensory information from different modalities, particularly determining spatial locations of objects. For example, it comprises somatosensory cortex and the

dorsal stream of the visual system. This enables regions of the parietal cortex to map objects perceived visually into body coordinate positions. The parietal lobe plays important roles in integrating sensory information from various parts of the body, knowledge of numbers and their relations, and in the manipulation of objects. Portions of the parietal lobe are involved with visuospatial processing. Much less is known about this lobe than the other three in the cerebrum.

3. The temporal lobes

The temporal lobes are part of the cerebrum. They lie at the sides of the brain, beneath the lateral or Sylvian fissure. The temporal lobes are where the thumbs would be. The temporal lobe is involved in auditory processing and is home to the primary auditory cortex. It is also heavily involved in semantics both in speech and vision. The temporal lobe contains the hippocampus and is therefore involved in memory formation as well. The functions of the left temporal lobe are not limited to low-level perception but extend to comprehension, naming, verbal memory and other language functions.

4. The occipital lobe

The occipital lobe is the visual processing center of the mammalian brain, containing most of the anatomical region of the visual cortex. There are many extrastriate regions, and these are specialized for different visual tasks, such as visuospatial processing, color discrimination and motion perception. Retinal sensors convey stimuli through the optic tracts to the lateral geniculate bodies, where optic radiations continue to the visual cortex. Each visual cortex receives raw sensory information from the outside half of the retina on the same side of the head and from the inside half of the retina on the other side of the head.

5. Central sulcus

The central sulcus is a fold in the cerebral cortex of brains in vertebrates. Also called the central fissure, it was originally called the fissure of Rolando or the Rolandic fissure, after Luigi Rolando. The central sulcus is a prominent landmark of the brain, separating the parietal lobe from the frontal lobe and the primary motor cortex from the primary somatosensory cortex. Also included is the somatomotor system (complex and multifaceted) which controls the skeletal musculature. It interacts with primary sensory systems and the cerebellum, which also has important interactions with the sensory systems.

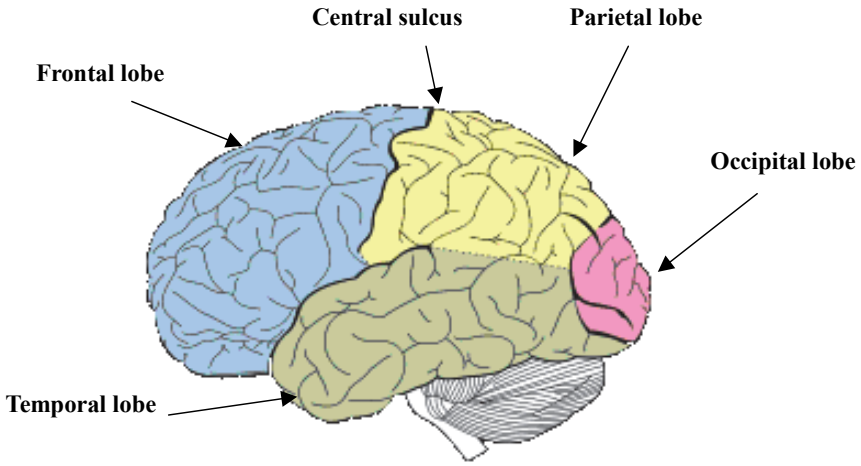


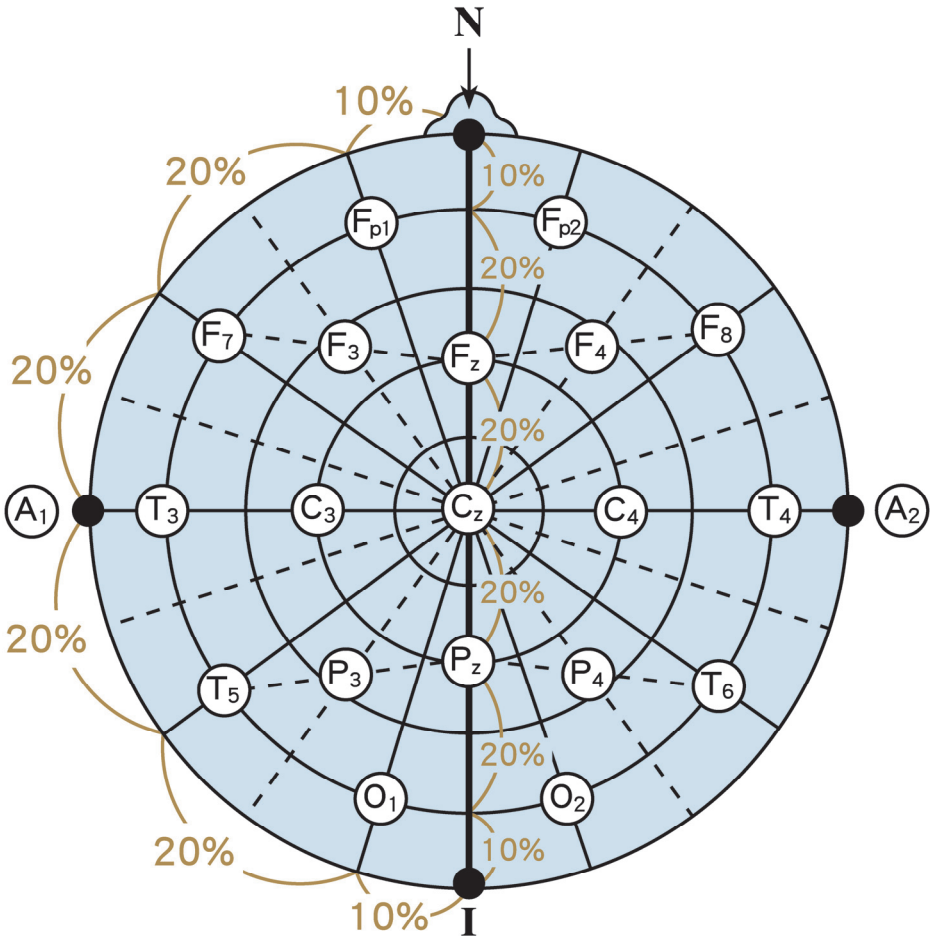
Fig. 2.1. The cerebral cortex include the frontal, temporal, occipital, parietal lobes and central sulcus.

2.2 EEG

Hans Berger (1873~1941), the discoverer of the human EEG, was a neuropsychiatrist. Electroencephalography is the neurophysiologic measurement of the electrical activity of the brain by recording from electrodes placed on the scalp or, in special cases, subdurally or in the cerebral cortex. Electrode placement is accomplished by measuring the scalp. Electrode locations and names are specified by the International 10–20 system, as depicted in Figure 2.2 (Andreassi, 2000). This system ensures a system of placement that is reliable and reproducible. The resulting traces are known as an electroencephalogram (EEG) and represent an electrical signal (postsynaptic potentials) from a large number of neurons. These are sometimes called brainwaves, though this use is discouraged (Cobb, 1983), because the brain does not broadcast electrical waves. The EEG is a brain function test, but in clinical use it is a "gross correlate of brain activity" (Ebersole, 2002). Electrical currents are not measured, but rather voltage differences between different parts of the brain. "EEGs" are frequently used in experimentation because the process is non-invasive to the research subject. The subject does not need to make a decision or behavioral action in order to log data, and it can detect covert responses to stimuli, such as reading. The EEG is capable of detecting changes in electrical activity in the brain on a millisecond-level.

Four major types of continuous rhythmic sinusoidal EEG activity are recognized (alpha, beta, delta and theta), as depicted in Figure 2.3 (Fisch, 1991). There is no precise agreement on the frequency ranges for each type. Delta is the frequency range up to 4 Hz and is often associated with the very young and certain encephalopathies and underlying lesions. It is seen in stage 3 and 4 sleep. Theta is the frequency range from 4 Hz to 8 Hz and is associated with drowsiness, childhood, adolescence and young adulthood. This EEG frequency can sometimes be produced by hyperventilation. Theta waves can be seen during hypnagogic states such as trances, hypnosis, deep day dreams, lucid dreaming and light sleep and the preconscious state just upon waking, and just before falling asleep. Alpha is the frequency range from 8 Hz to 13 Hz. It comes from the occipital (visual) and parietal cortex and is characteristic of a relaxed, alert state of consciousness. For alpha rhythms to arise, usually the eyes need to be closed. Alpha attenuates with extreme sleepiness or with open eyes and increased visual flow. Beta is the frequency range above 13 Hz. Low amplitude beta with multiple and varying frequencies is often associated with active, busy or anxious thinking and active concentration.

When people become fatigued, they usually report difficulties in concentrating and focusing their attention on the tasks they are required to perform (Boksem et al., 2005). Various aspects of EEG, including power distribution and event-related potential (ERP), have been employed to assess specific mental tasks, e.g. arousal level (Eoh et al., 2005; Waard and Brookhuis, 1991) and cognitive depth (Boksem et al., 2005; Murata et al., 2005). One of the common findings of EEG studies on a drop in arousal level is that the EEG shifts from fast and low amplitude waves to slow and high amplitude ones (Klimesch, 1999; Lafrance and Dumont, 2000). More specifically, under decreased alertness, there is a progressive increase in low-frequency alpha and theta activity (Klimesch, 1999; Lafrance and Dumont, 2000; Oken and Salinsky, 1992), probably reflecting a decrease in cortical activation (Cook et al., 1998; Laufs et al., 2003). Therefore, the amount of alpha and theta power provides an adequate index of the level of fatigue that subjects experience (Boksem et al., 2005).



F_p1, F_p2 : prefrontal
F3, F4 : frontal
C3, C4 : central
P3, P4 : parietal
O1, O2 : occipital
F7, F8 : anterior temporal
N : Nasion
I : Inion

T3, T4 : mid-temporal
T5, T6 : posterior temporal
A1, A2 : ear (or mastoid)
Fz : frontal midline
Cz : central vertex
Pz : parietal midline
(note: z = zero)

Fig. 2.2. International 10–20 system, electrode positions are determined by measurements from landmarks on the head.

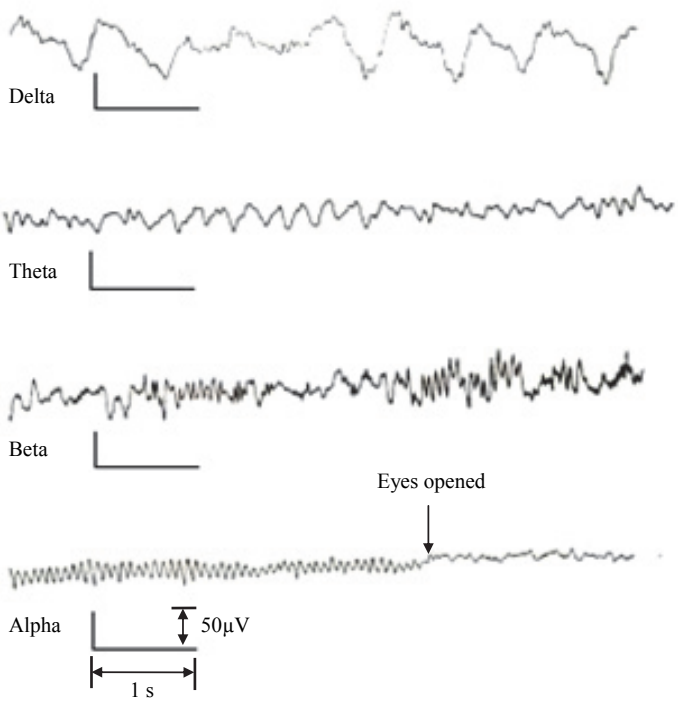


Fig. 2.3. Basic EEG waveform. First row: delta rhythm frequency at 0.5–4 Hz; second row: theta rhythm frequency at 4–8 Hz; third row: beta rhythm frequency at 13–20 Hz; fourth row: alpha rhythm frequency at 8–13 Hz.

2.3 ERP

The event-related potential (ERP) is a transient series of voltage oscillations in the brain recorded from scalp EEG following a discrete event. An ERP is a stereotyped electrophysiological response to an internal or external stimulus. More simply, it is a measured brain response as a result of a thought or perception. ERPs can be reliably measured using electroencephalography (EEG), a measure of brain electrical activity from the skull and scalp. As the EEG reflects thousands of simultaneously ongoing brain processes, the brain response to a specific stimulus or event of interest is usually masked with direct EEG measurement. One of the most robust features of the ERP response is a response to unpredictable stimuli. In actual recording situations, it is difficult to see an ERP after the presentation of a single stimulus. Rather, the ERPs become visible, when many dozens or hundreds of individual presentations are averaged together (as depicted in Figure 2.4). This technique cancels out noise in the data, and only the voltage response in relation to the stimulus is mathematically enhanced. While evoked potentials reflect the processing of the physical stimulus, event-related potentials are caused by the higher processes, that might involve memory, expectation, attention, or changes in the mental state, among others. Description of the scalp or surface cortical ERP distribution is the starting point for identifying the ERP generators, involving the topographic mapping of the ERP waveform

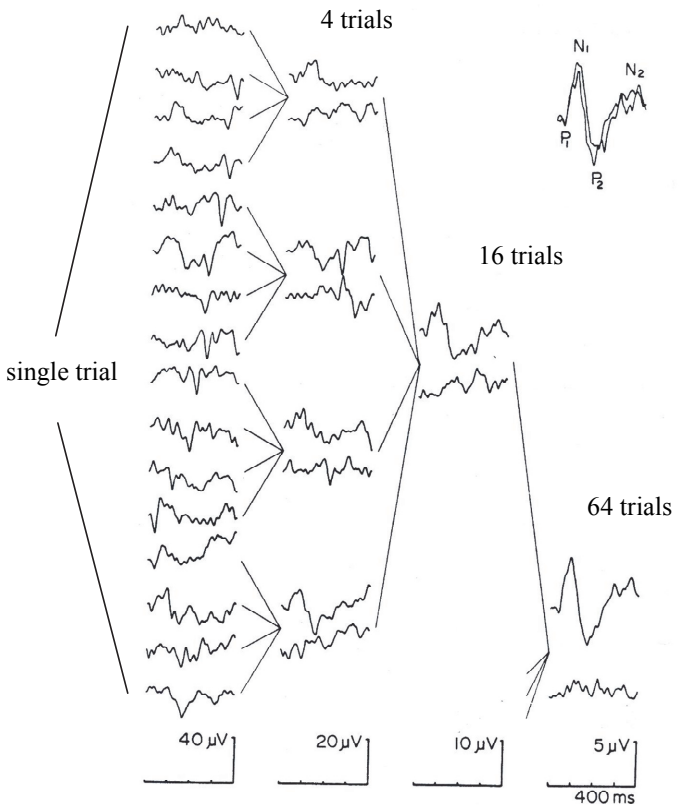


Fig. 2.4. Averaged waveform of ERP. Leftmost column: 16 single trial ERPs. Second column from left: average ERPs computed across 4 trials (upper waveform of each pair) and an estimate of the noise residual (lower waveform of each pair). Second column from right: average ERP computed across 16 trials (upper waveform) and noise residual (lower waveform). Rightmost column: average ERP computed across 64 trials and noise residual. (From Picton, 1980)

over time according to the International 10-20 system (Figure 2.2). It is often convenient as a first approximation to identify ERP peaks and troughs as positive or negative "components," as is the standard practice in the analysis of human scalp-recorded ERP (Picton, 1988; Niedermeyer et al., 1993). The ERP has been traditionally partitioned into a number of separate components. The most consistent finding is a modulation of the posterior P100 (peaking between 100 and 160 ms after stimulus presentation) and N100 (160–210 ms) components by attention (Eason, 1981; Rugg et al., 1987; Wijers et al., 1989a, b). When a particular location is attended, the exogenous P100 and N100 waves elicited by stimuli at that location are enlarged (Hillyard and Münte, 1984; Mangun and Hillyard, 1988, 1990), an effect that has been interpreted as a sign of attentional modulation of sensory processing in the visual pathways (Mangun et al., 1993). This has been viewed as a representation of a "sensory gain" mechanism (Hillyard et al., 1990): as a result of biasing

the information processing system, the responsivity to stimuli presented at attended locations is amplified, and further processing of these stimuli will therefore be enhanced. A later component, starting at approximately 200–250 ms post stimulus, consisting of negativity at central electrodes, with a maximum at Cz, has been labeled the N200 component. This ERP component has been found to reflect the further processing of relevant information (i.e. stimuli that require a response) (Lange et al., 1998; Okita et al., 1985; Wijers et al, 1989a, b). In the stimulus-locked ERP, the P300 was defined as the most positive peak in a window between 200 and 500 milliseconds. The latency of each ERP component was defined as the time between the onset of the arrow array and the time when the peak value appeared for stimulus-locked ERP. (Ullsperger et al., 1986, 1988). The P300 component is useful to identify the depth of cognitive information processing. It has been reported that the P300 amplitude elicited by mental task loading decreases with the increase in the perceptual/cognitive difficulty of the task (Donchin, 1979; Isreal et al., 1980a, b; Kramer et al., 1983, 1985; Mangun and Hillyard, 1987; Ullsperger et al., 1986, 1988). Thus, the P300 amplitude mainly reflects the depth or degree of cognitively processing the stimulus. In other words, it is highly related to the level of attention. In addition to magnitude aspect, the P300 latency was prolonged when the stimulus was cognitively difficult to process (Murata et al., 2005). Uetake and Murata (2000) reported that the P300 amplitude and latency could be employed to assess mental fatigue induced during a VDT task. They indicated that the P300 latency was prolonged and the P300 amplitude decreased with cumulative mental fatigue.

3. Methods

3.1 Subjects

Twenty-three university male students with a mean age 22.0 ± 1.3 years participated as volunteer subjects. They had normal hearing and normal or corrected-to-normal vision (via medical tests). Each participant met all the inclusion criteria: no medical, psychiatric, or head injury, and not using any medications or drugs. However, three participants were terminated by the experimenter due to excessive movement artifacts in the EEG during the test. Thus, complete data sets were collected from twenty participants who were right handed by self-report. An informed written consent form was obtained from all the participants after the procedure of the study was explained and the laboratory facilities were introduced to them. They were paid for their participation in the study.

3.2 Experimental procedures

The participants were instructed to avoid alcohol and caffeine in the 24 hours before the test. On the test day, the experimental task started at 8 AM. Participants performed the task alone in a dimly lit, sound-attenuated, electrically shielded test room. The experiment task was clearly explained first, and participants were allowed to practice until they felt familiar with it. The subject was required to record the EEG and measure the ERPs before starting the experimental session. The EEG was measured at rest condition for five min, and then a modified Eriksen flanker task was performed (Eriksen and Eriksen, 1974) under the experimenter's instruction.

After the measurement of the ERPs was finished, the subject conducted an experimental task for 180 min. The experimental task was to mentally add two three-digit numbers that were displayed on the LCD and enter the answer using a keyboard for 30 min. There was no time constraint for the mental addition trial and the task was self-paced. The task was

programmed on a personal computer using C language. The illumination on the LCD was about 300 lx. The viewing distance was about 80 cm. The response time and the error trial, if any, were recorded on a hard disk data file. After mental arithmetic, the subjects performed data entry for 2 h, and then underwent mental arithmetic for 30 min. The experimental procedure is shown in Figure 3.1.

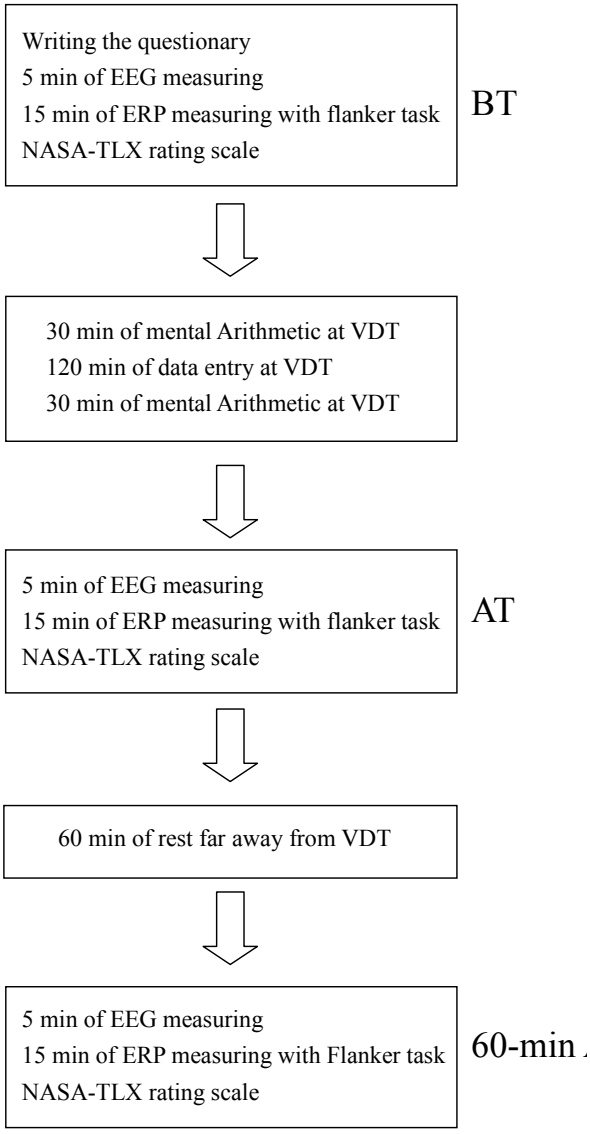


Fig. 3.1. The flow chart of the experimental procedure, including three measuring sessions (BT, AT, and 60-min AT), 120 min of VDT tasks, and 60 min of rest.

Similar EEG recordings were conducted immediately after the completion of the 180-min experimental task. After 60 min rest, the participants repeated the EEG measurement mentioned above, and then finished the whole test. At the end of each EEG measurement, self-report assessments of task loading were obtained by using the NASA-Task Load Index (TLX) rating scale (Hart and Staveland, 1988). The NASA-Task Load Index (NASA-TLX) consists of six component scales. An average of these six scales, weighted to reflect the contribution of each factor to the workload of a specific activity from the perspective of the rater, is proposed as an integrated measure of overall workload (referred to Appendix).

3.3 Behavior response tasks

A modified Eriksen flanker task with word stimuli replaced by arrow stimuli was adopted in this study. The stimuli were presented on a computer screen (15 inches) with a dark background and with a viewing distance of 80 cm (as shown in Figure 3.2(a)). The participants wore an elastic cap and comfortable clothing and sat in front of the computer monitor, as shown in Figure 3.2(b). A participant was required to press a designated button on a control panel (with reference to Adam et al. 1996, as depicted in Figure 3.2(c)) connected with the computer in response to the target stimulus. Designed buttons on the control panel were applied to orient the position between the start and control points of participant's moving finger.

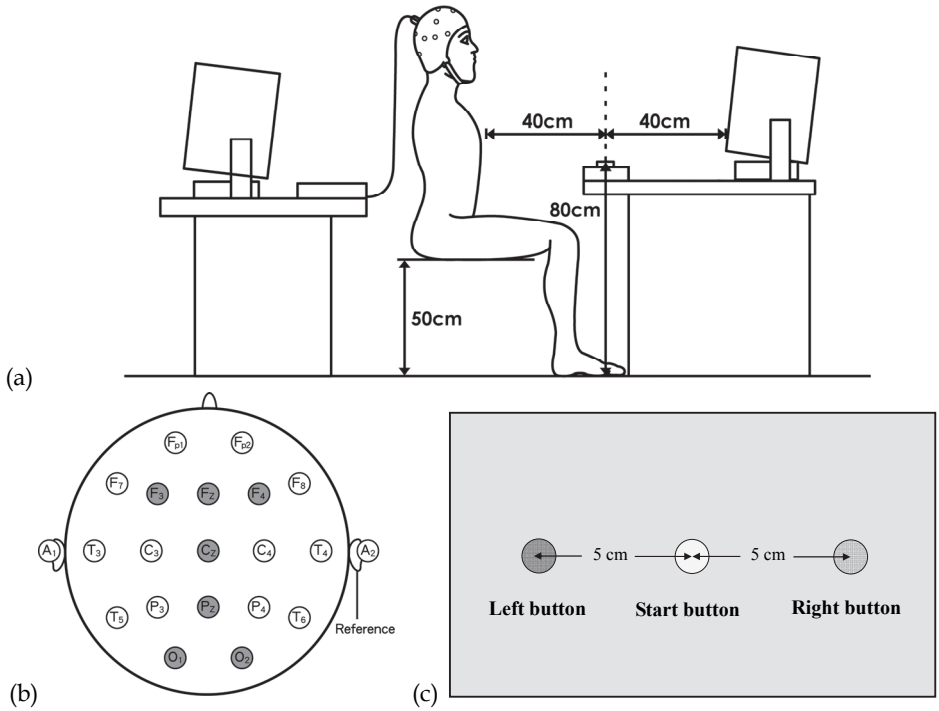


Fig. 3.2. (a) The layout and the position of the test device related to the participant wearing an EEG cap with scalp electrodes in international 10–20 montage. (b) Reference electrode is located on the right earlobe. (c) The self-made panel of control buttons was connected with the test device

Fig. 3.3. A participant wearing an EEG cap with scalp electrodes performed the modified flanker task

A participant was asked to focus on the arrow in the center of a visual array of five arrows on a computer screen, designated as target, and to respond with the right index finger to press the left or right button depending on the direction of the target arrow. The target arrow was flanked by four other arrows, two pointing to the left and two to the right, pointing in the same direction as the target (congruent) or in the opposite direction (incongruent) (as delineated in Figure 3.4). Congruent and incongruent trials were presented with equal probabilities. The left- and right-button responses signaled by target arrows occurred equally as often as well.

When the experimental task started, target arrows appeared at one time for each test trial. As soon as the target arrows were presented, the participant withdrew the right index finger from the start button to press a corresponding button and then returned the finger to the start button and finished a test trial. Trials were presented in pseudorandom order to limit the consecutive number of trials with same arrow arrays below five. Participants pressed the left button in response to a target arrow pointing to the left and the right button to a right-

pointing target. Each trial started with the presentation of a central fixation white cross "+", which lasted for 1 second. The arrow array appeared 200 milliseconds later after the fixation cross disappeared and it lasted for 50 milliseconds. The target arrow was in dark gray. The flanker arrows immediately surrounding the target arrow were in light gray and larger than the target, while the farthest flankers were in white and larger than the adjacent flankers. The inter-trial interval started from 2 seconds (maximum time interval for response) or when a button was pressed within 2 seconds after the presentation of the arrow array and was randomized between 1200 and 2000 milliseconds.

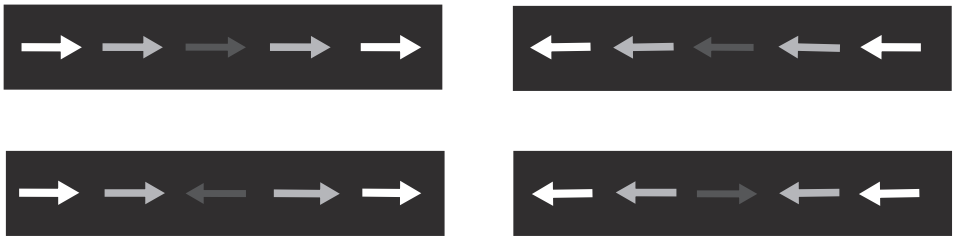


Fig. 3.4. The target arrow was flanked by four other arrows, two left arrows and two right ones, pointing in the same direction as the target (upper rows as congruent) or in the opposite direction (lower rows as incongruent).

Participants were initially trained with 50 trials and without feedback. Subsequently, they completed 3 blocks of 200 trials at the three test sessions (BT, AT, and 60-min AT). One test session took about 15 to 20 minutes. The whole experiment including EEG measures and experimental task lasted about 5 h, including pauses, placement and removal of EEG electrodes.

In this study, RT was measured as the time between the onset of the arrow array and the control button press. Trials with RTs longer or shorter than twice the value of the standard deviation for RT were excluded from calculation for mean RT. ER was calculated as the percentage of miss or erroneous responses.

3.4 EEG recording and data analysis

During the task performance, EEG was recorded by using an electrode cap (Quick-Cap, Compumedics NeuroScan, El Paso, Texas) with Ag/AgCl electrodes placed at F3, Fz, F4, Cz, Pz, O1, and O2 in the International 10–20 montage with an electronically linked mastoids reference as shown in Figure 3.2(b) (Andreassi, 2000). Two Ag/AgCl electrodes were placed 2 cm above and 2 cm below the left eye to record vertical electrooculogram (EOG). Two electrodes were positioned at 1 cm external to the outer canthus of each eye for horizontal EOG recording. A ground electrode was placed on the forehead. Electrode impedances were kept below 10 kΩ. The EEG and EOG were amplified by SYNAMPS amplifiers (Neuroscan, Inc.) and sampled at 500 Hz. The EEG epochs were then corrected by eye movement by using the Ocular Artifact Reduction (Semlitsch et al., 1986) command of SCAN 4.3 (Neuroscan, Inc.) and then underwent movement-artifact detection by using the Artifact Rejection command.

For measuring the background EEG pattern of participant, EEG spectral analysis was performed only for the 5-min rest condition. The recorded EEG during 5-min rest condition was subsequently transformed from time into frequency domains by fast Fourier transform (FFT) using a 5-s Hanning windowing function.
For ERP analysis, the EEG data were further digitally high-pass filtered at 1 Hz (–12 dB/octave) and were then segmented into stimulus-locked EEG epochs from 200 milliseconds before and 800 milliseconds after the onset of displaying the arrow array of flank test. The stimulus-locked EEG signals were baseline corrected between –100 milliseconds before the onset of stimulus. The averaged waveforms (i.e. ERPs) for stimulus-locked EEG epochs were band-pass filtered at 1 to 10 Hz prior to subsequent analyses.The amplitude and latency measures for P300 were derived from the stimulus-locked ERP recorded at F3, Fz, F4, Cz, Pz, O1, and O2 electrodes, respectively. It is noted that the EEG epochs of the trials with omitted responses or with RTs longer or shorter than twice the value of the standard deviation for RT were not included in the stimulus-locked ERP.
We analyzed the relationship between EEG power of θ, α, β, θ/α, β/α and (α+θ)/β indices (Brookhuis and Waard, 1993; Eoh et al., 2005; Ryu and Myung, 2005) as well as the amplitudes and latencies of the P300 component (Murata et al., 2005). The basic index means the relative power of the EEG θ, α and β bands. The δ band was not included in our analysis, since it happens in a deep sleep state and usually overlaps with artifacts. The relative power equation of the θ, α, and β bands are represented respectively as:

$$\text{Relative power of } \theta = (\text{power of } \theta) \;/\; (\text{power of } \theta + \text{ power of } \alpha + \text{ power of } \beta) \quad (1)$$

$$\text{Relative power of} \alpha = (\text{power of } \alpha) \;/\; (\text{power of } \theta + \text{ power of } \alpha + \text{ power of } \beta) \quad (2)$$

$$\text{Relative power of } \beta = (\text{power of } \beta) \;/\; (\text{power of } \theta + \text{ power of } \alpha + \text{ power of } \beta) \quad (3)$$

Since the basic indices have a tendency to "contradict each other", the ratio indices were calculated to amplify the difference. The known ratio indices β/α, θ/α, and (θ+α)/β were analyzed in previous studies (Brookhuis and Waard, 1993; Pyun and Kim, 2000; Ryu and Myung, 2005)
EEG power and ERP measured at recording sites F3, Fz, F4, Cz, Pz, O1, and O2, were analyzed by means of separate repeated-measures analyses of variance (ANOVA) with the within-subjects factors "session" including before (BT), immediately after (AT), and 60 min after (60-min AT) tasks, and "electrode" (F3, Fz, F4, Cz, Pz, O1, and O2). Where appropriate, differences from, sessions, electrodes, or electrode-by-session interactions were further evaluated with Fisher LSD post hoc tests (nominal level of alpha: $P<0.05$). ANOVA test, an inferential statistical procedure, examines the variation and tests whether the between group variance is greater than the within group variance. The larger the F ratio (the larger the variation between the groups) is, the greater the probability (the smaller p value) of rejecting a multiple group situations are the same. A one-way ANOVA ($p<0.05$) is used to determine if there is a difference between the groups.

4. Results

4.1 Performance and psychological evaluation of fatigue

All false responses on a modified Eriksen flanker task were calculated as ER. The mean RTs for each trial and the ERs were obtained at the three test sessions. The RT and ER results are

summarized in Appendix. A one-way (session: BT, AT, and 60-min AT) ANOVA carried out on the RT revealed no significant main effect of the session, whereas a one-way ANOVA conducted on the ER revealed a predominant difference between BT and AT ($F(2,38) = 6.371$, $p < 0.05$), while no significant difference was found between BT and 60-min AT. Figure 4.1 depicts the comparison of RTs on the modified Eriksen flanker task among three sessions (BT, AT, and 60-min AT). The RT tended to be prolonged at the post-task measurement. As a result of a similar one-way ANOVA carried out on the RT, no significant main effect of the measurement epoch was detected. The mean rating scale of mental fatigue tended to increase immediately after the completion of the task. At 60 min after the completion of the experimental task, the rating scale decreased and was nearly equal to the value in the BT session (as shown in Figure 4.2). A one-way ANOVA conducted on the rating scale revealed a pronounced difference between BT and AT ($F(2,38) = 5.23$, $p < 0.05$).

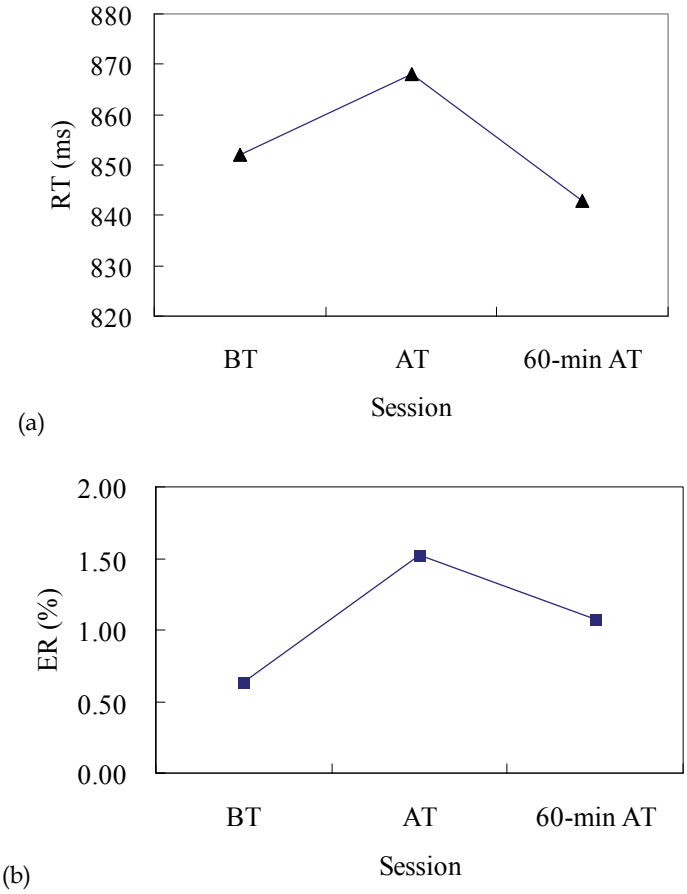


Fig. 4.1. Comparison of (a) RT and (b) ER on modified Eriksen flanker task among three sessions. BT: before task, AT: immediately after task

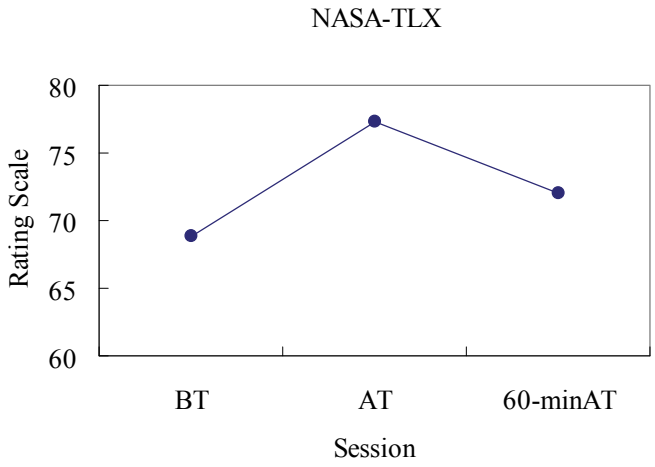


Fig. 4.2. Comparison of NASA-Task Load Index (TLX) rating scale on mental fatigue among three sessions. BT: before task, AT: immediately after task

4.2 EEG power spectra

The EEG indices, classified into two groups—the basic index and the ratio index, were derived from the reorganized data. Since the basic indices have a tendency to "contradict each other", the ratio indices were calculated to amplify the differences. The known ratio indices β/α, θ/α, and $(\theta+\alpha)/\beta$ were analyzed in previous studies (Brookhuis and Waard, 1993; Pyun and Kim, 2000; Ryu and Myung, 2005). The ANOVA results of EEG measured at the three sessions (BT, AT, and 60-min AT) are summarized in Table 1. All indices showed significant differences in location, and all indices except β and β/α showed significant differences in session. (see Table 4.1). Student-Newman-Keuls (SNK) post hoc analysis for the factor of location showed that the frontal (F3, Fz, F4), centro-parietal (Cz, Pz) and occipital (O1, O2) were separated into statistically different groups ($\alpha = 0.05$). In the post hoc analysis for the factor of the session, BT and AT revealed significantly different. No indices showed a significant difference of interaction effect. The ANOVA for 3 basic indices and 3 ratio indices are shown in Table 4.2 ~ 4.7.

Index	Location	Session	Interaction
θ	<0.01**	<0.01**	0.997
α	<0.01**	<0.01**	0.880
β	<0.01**	0.718	0.819
θ/α	<0.01**	<0.01**	0.070
β/α	<0.01**	0.224	0.574
$(\alpha+\theta)/\beta$	<0.01**	<0.01**	0.171

*Significant at $\alpha = 0.05$, **Significant at $\alpha = 0.01$.

Table 4.1. ANOVA summary for EEG measurement.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	8.216	16.038	16.752	13.641	7.391	6.414	5.155
P value for (1–2)	**0.010**	**0.001**	**0.001**	**0.002**	**0.014**	**0.020**	**0.035**
F for (1–3)	2.794	5.846	4.205	3.392	0.982	1.984	1.418
P value for (1–3)	0.111	**0.026**	0.054	0.081	0.334	0.175	0.248

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.2. ANOVA of basic index θ.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	11.594	13.728	14.935	16.059	4.532	6.962	5.998
P value for (1–2)	**0.003**	**0.002**	**0.001**	**0.001**	**0.047**	**0.016**	**0.024**
F for (1–3)	8.298	8.505	5.442	5.068	0.741	2.381	2.292
P value for (1–3)	**0.010**	**0.009**	**0.031**	**0.036**	0.400	0.139	0.147

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.3. ANOVA of basic index α.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	0.106	0.138	0.034	0.319	0.584	0.641	0.002
P value for (1–2)	0.748	0.714	0.856	0.579	0.454	0.433	0.965
F for (1–3)	2.419	1.097	2.688	1.139	0.068	0.153	0.021
P value for (1–3)	0.136	0.308	0.118	0.299	0.797	0.700	0.885

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.4. ANOVA of basic index β.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	0.106	0.138	0.034	0.319	0.584	0.641	0.002
P value for (1–2)	0.748	0.714	0.856	0.579	0.454	0.433	0.965
F for (1–3)	2.419	1.097	2.688	1.139	0.068	0.153	0.021
P value for (1–3)	0.136	0.308	0.118	0.299	0.797	0.700	0.885

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.5. ANOVA of ratio index β/α.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	11.701	22.309	14.945	14.560	7.245	6.734	4.928
P value for (1–2)	**0.003**	**0.000**	**0.001**	**0.001**	**0.014**	**0.018**	**0.039**
F for (1–3)	4.337	5.299	5.568	4.097	1.996	2.506	1.825
P value for (1–3)	0.051	**0.033**	**0.029**	0.057	0.174	0.130	0.193

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.6. ANOVA of ratio index θ/α.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	0.905	0.309	0.857	0.164	0.106	4.509	4.416
P value for (1–2)	0.353	0.585	0.366	0.690	0.748	**0.046**	**0.049**
F for (1–3)	5.058	4.519	5.330	3.561	0.456	1.077	0.623
P value for (1–3)	**0.037**	**0.047**	**0.032**	0.075	0.508	0.312	0.440

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.7. ANOVA of ratio index (α+θ)/β.

The basic indices θ and α at all recording sites tended to increase and decrease respectively, immediately after the completion of an experimental task. At 60 min after the experimental task was completed, the basic indices θ and α decreased and increased respectively, and recovered to the level in the BT session, as depicted in Figure 4.3(a) and 7(b). As shown in Figure 4.3 (e), the ratio indices θ/α revealed significantly increased immediately after the completion of an experimental task than those before the task at all electrode sites. However, (θ+α)/β showed a significant decrease after the completion of an experimental task. Only the value at the occipital increased to the level in the BT session at 60 min after the experimental task was completed, as shown in Figure 4.3 (f).

4.3 P 300 component of ERP

The ANOVA results of P300 component of ERP measured at the three sessions (BT, AT, and 60-min AT) are summarized in Table 4.8. The ANOVA for P300 latency and amplitude are shown in Table 4.9 and 4.10 respectively. The amplitude and latency showed significant differences in location, and the latency showed significant difference in session, while the amplitude revealed no significant differences in session. The P300 latency at Pz tended to decrease immediately after the completion of an experimental task. It increased at 60 min after the experimental task was completed but failed to recover to the level in the BT session, as illustrated in Figure 4.4 (e). A one-way (measurement epoch) ANOVA conducted on the P300 latency revealed a significant main effect at recording site of Pz, $F(2,38) = 5.684$, $p<0.05$. The P300 amplitude, in accordance with this, tended to decrease at the post-task measurement and failed to recover to the pre-task level at 60 min after the completion of the experimental task. A similar ANOVA conducted on the P300 amplitude revealed a significant main effect of the

measurement epoch at recording sites of Pz ((F(2,38) = 4.575, p<0.05), O1 (F(2,38) = 9.182, p<0.01) and O2 (F(2,38) = 4.694, p<0.05) (as illustrated in Figures 4.4 (e) ~ (g)).

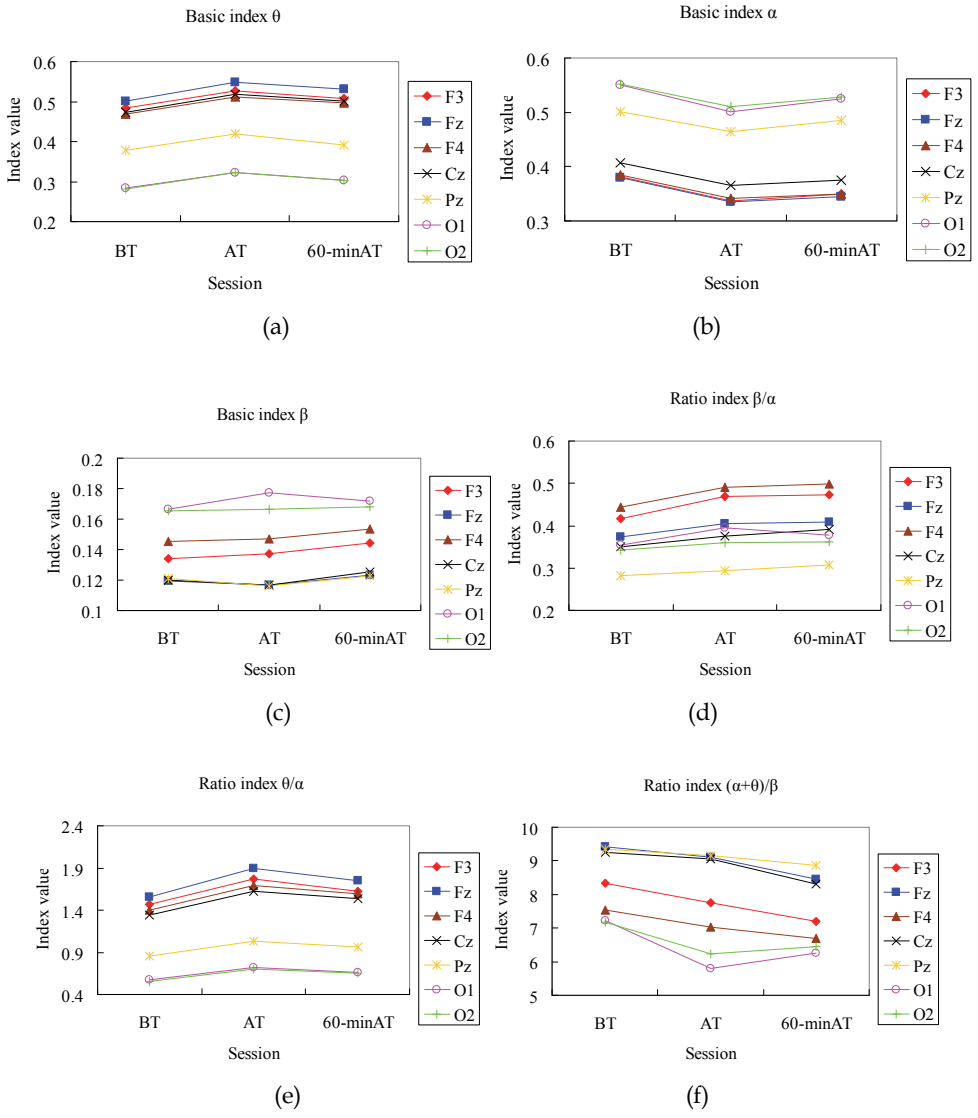


Fig. 4.3. Comparison of EEG indices among three sessions. BT: before task, AT: immediately after task. (a), (b), and (c) are basic indices θ, β, and α. (d), (e), and (f) are ratio indices β/α, θ/α and (θ+α)/β

Component	Location	Session	Interaction
Amplitude	<0.01**	0.077	0.243
Latency	<0.01**	<0.05*	0.286

*Significant at $\alpha = 0.05$, **Significant at $\alpha = 0.01$.
Table 4.8. ANOVA summary for P300 component of ERP measurement.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	0.533	2.152	3.308	0.533	5.684	1.556	0.748
P value for (1–2)	0.137	0.159	0.085	0.474	**0.028**	0.227	0.398
F for (1–3)	2.365	1.752	0.022	0.311	2.967	0.776	3.142
P value for (1–3)	0.141	0.201	0.884	0.584	0.101	0.389	0.092

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.9. ANOVA of P300 latency.

Electrode	F3	Fz	F4	Cz	Pz	O1	O2
F for (1–2)	2.525	0.757	1.072	0.499	4.575	9.182	4.694
P value for (1–2)	0.129	0.395	0.313	0.488	**0.041**	**0.007**	**0.043**
F for (1–3)	0.043	1.375	1.262	4.404	0.009	2.476	1.823
P value for (1–3)	0.837	0.256	0.275	**0.049**	0.924	0.132	0.193

Note: 1 denoted session BT; 2 denoted session AT; 3 denoted session 60-min AT
Table 4.10. ANOVA of P300 amplitude

5. Discussion and conclusion

5.1 Discussion

Based on the RT and ER on the modified Eriksen flanker task (see Figure 4.1) and the psychological rating scale of fatigue (see Figure 4.2) in the experimental task, it can be judged that the experimental task induced a tendency toward mental fatigue in the subjects. The effects of mental fatigue resulted in the level of attention to decrease (Boksem et al., 2005). The decreased level of attention caused a significant increase in ER and an increased tendency in RT. At 60 min after the completion of the experimental task, the RT and ER on Eriksen flanker task decreased, which indicated that the state of fatigue had improved during the 60-min rest, but did not recover to the original state.

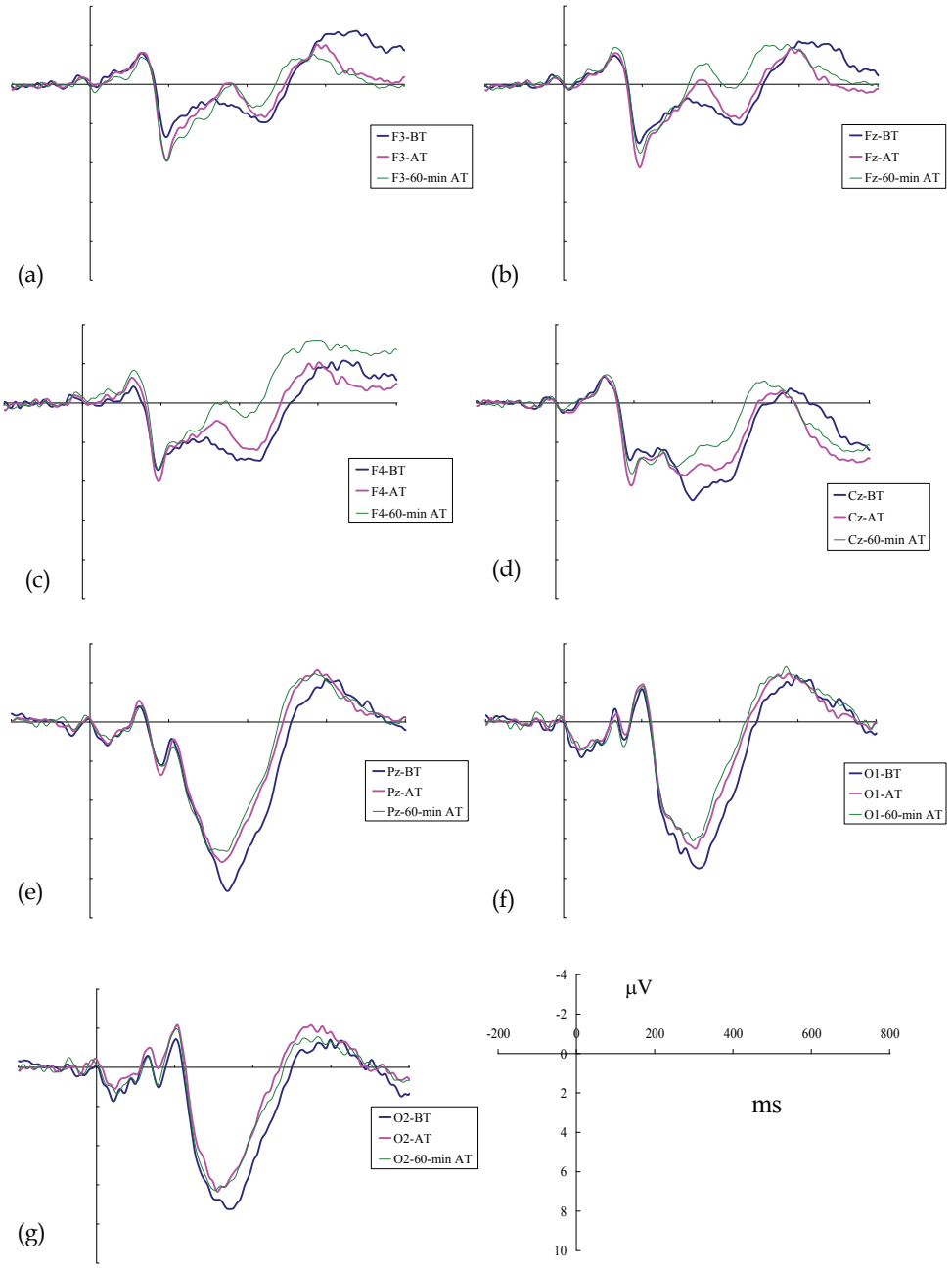


Fig. 4.4. Grand-averaged P300 waveform at seven recording sites F3, Fz, F4, Cz, Pz, O1, and O2 among three sessions. BT: before task, AT: immediately after task.

Research has shown that β waves are associated with increased alertness and arousal, α waves occur during relaxed conditions, at decreased attention levels and in a drowsy but wakeful state, and θ waves mainly occur at sleep state one (Grandjean, 1988; Okogbaa et al., 1994; Rains and Penzien, 2003). Among the EEG power spectra observed in this study, the α and θ waves showed significant changes while completing the experiment task in this study. The changes were consistent with those of previous studies (Åkerstedt et al., 1991; Lal and Craig, 2001). Changes in EEG with vigilance have generally shown that deterioration in performance is associated with increased θ wave and changes in α intensity (Davies, 1965; Morrel, 1966; Gale et al., 1977). Makeig and Jung (1995) also found that changes in α and θ waves were related to reduced performance and fatigue. In this study, we found that the index θ increased and the index α decreased after 3 h of VDT task. The subjects revealed some extent of mental fatigue, but their alertness level was increased after the experimental task. Mental arithmetic was a secondary task in this study. Hitch (1978) argued that working memory plays a major role in the task. On the other hand, it is also recognized that it includes the processes required to recognize numbers in their Arabic form, those required to comprehend verbal representation of numbers, those that assign magnitudes to numerical quantities, those that report a numerical sum, and so forth (Dahaene et al., 1999). Therefore, mental arithmetic seems to engage memory processes in respect to retrieval of arithmetic facts from long-term memory. During the experimental task, the processes required to recognize numbers is also needed. Mental effort that requires memory processes is known to suppress EEG alpha activity (Klimesch, 1997).

Among the ratio indices, index θ/α and index (α+θ)/β were statistically significant in this experiment. Index θ/α of session BT discriminated session AT which no other indices could do and it recovered to closely original state at 60-min AT (see Figure 4.3). Index (α+θ)/β showed different statistical characteristics compared to index θ/α due to the mutual addition effect of α waves and θ waves during the repetitive phase transition between wakefulness and microsleep. Because the changes of basic indices α and θ (increased θ and decreased α) showed different direction, the mutual addition effect of α and θ waves counteracted each other and reduced the index value, while θ/α accelerated the increase of the index valve due to the amplification of division. After the completion of an experimental task, the ratio index (θ+α)/β was decreased significantly at the occipital— visual dominating area. It revealed the main fatigue induced from the VDT task was in the visual area. The index value of (θ+α)/β increased at 60-min AT manifested the fatigue had improved, but did not recover to original state, except for visual sensory, after 60 min of rest. In this study, we found that index θ/α was more available than the other two ratio indices for assessment of mental fatigue in VDT task.

On the other hand, the P300 component of ERP indicated that the mentally and physically fatigued state could be explained by decreased activity of the central nervous system (CNS). This phenomenon revealed a decreased depth of cognitive information processing and a decreased level of attention. It has been pointed out that the increase of P300 latency is related to the temporal aspect due to the difficulty in cognitive information processing (Ullsperger et al., 1986, 1988; Donchin, 1979; Neuman et al., 1986). These findings were applied to the evaluation of mental fatigue. Uetake and Murata (2000) indicated that the appearance of mental fatigue is reflected more strongly in the two P300 components of

amplitude and latency. The delayed cognitive information processing (the prolonged P300 latency) and the decreased activity of cognitive information processing (decrease of the P300 amplitude) were found to be effective measures of mental fatigue. Traditionally, the assessment of mental fatigue is conducted by using the decreased arousal level (EEG). Okogbaa et al. (1994) investigated the relationship between EEG and mental fatigue. Although θ, α, β, and θ/α indexes were calculated to assess mental fatigue, these indices did not necessarily decrease with time and correlate with the appearance of mental fatigue or decrease of performance. The indices based on EEG measurement, in general, show the arousal level in the brain, but do not necessarily reflect the cognitive aspects such as the depth of cognitive information processing and the delay of processing. As clarified in this study, mental fatigue seems to be more strongly related to the declining the depth of cognitive information processing, i.e. the cognitive function. However, we did not find the delay of information processing due to the decrease of P300 latency after the experimental task. The possible reason was the mental arithmetic task improved the information processing capability. It revealed that the amplitude of P300 had better discrimination than latency of P300 for assessment of mental fatigue.

It is impossible to continuously measure ERP, therefore, the assessment of mental fatigue should be conducted with more confidence from the viewpoints of both arousal level (EEG) and cognitive information processing (ERP). The finding that mental fatigue was reflected in the decreased cognitive function such as the P300 component, would be significant and useful to promote the assessment of mental fatigue by means of multiple psychophysiological indices. None of these measures alone was a particularly powerful signal or warning of mental fatigue. Systematically taking these measures into account would lead to an effective evaluation method. The method proposed in this study is potentially applicable to the evaluation of the fatigued state of workers and to the management of mental fatigue from the viewpoints of occupational risk management, productivity, and occupational health.

5.2 Conclusion

The assessment of mental fatigue induced from 3 h of VDT task was undertaken by using indices of EEG bands and P300 component of ERP. In the EEG analysis for the VDT task, basic indices α and θ, ratio indices θ/α and (α+θ)/β were found to be statistically significant. It revealed the main fatigue induced from the VDT task was in the visual area. After 60 min of rest, the participants' fatigue did not diminish to the original state except visual sensory. After the experimental task, the amplitude significantly decreased, and the latency of P300 significantly shortened due to the mental arithmetic task improving the information processing capability. It revealed that index θ/α was more available than the other two ratio indices and the amplitude of P300 had better discrimination than latency of P300 for assessment of mental fatigue in VDT tasks. The P300 component of ERP indicated the possibility that one aspect of the mentally fatigued state could be explained by the decreased activity of CNS. This phenomenon is related to a decreased depth of cognitive information processing and a decreased level of attention. The method proposed in this study is potentially applicable to the evaluation of the fatigued state of workers and to the management of mental fatigue from the viewpoint of occupational risk management.

6. References

Åkerstedt, T., Kecklund, G., Knuttsson, A., 1991, "Manifest sleepiness and the EEG spectral content during night work", Sleep, vol. 14, pp. 221–225.

Adam, J.J., Paas, F.G.W.C., Buekers, M.J., Wuyts, I.J., Spijkers, W.A.C., Wallmeyer, P., 1996, "Perception-action coupling in choice reaction time tasks", Human Movement Science, vol. 15, pp. 511–519.

Andreassi, J.L., 2000, Psychophysiology: Human Behavior and Physiological Response, fourth ed. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Boksem, M.A.S., Meijman, T.F., Lorist, M.M., 2005, "Effects of mental fatigue on attention: An ERP study", Cognitive Brain Research, vol. 25, pp. 107 – 116.

Brookhuis, K.A., Waard, D., 1993, "The use of psychophysiology to assess driver status", Ergonomics, vol. 39(9), pp. 1099–1110.

Cobb, W.A., 1983, Recommendations for the practice of clinical neurophysiology, Amsterdam: Elsevier.

Cook, I.A., O'Hara, R., Uijtdehaage, S.H.J., Mandelkern, M., Leuchter, A.F., 1998, "Assessing the accuracy of topographic EEG mapping for determining local brain function", Electroencephalography and Clinical Neurophysiology, vol. 107, pp. 408–414.

Davies, D.R., 1965, "Skin conductance, alpha activity and vigilance", American Journal of Psychology, vol. 78 (2), pp. 304–306.

Dahaene, S., Spelke, E., Pinel, P., Stanescu, R., Tsivkin, S., 1999, "Sources of mathematical thinking: behavioral and brain-imaging evidence", Science, vol. 284, pp. 970–974.

Dement, W.C., Vaughan, C., 1999, The Promise of Sleep. New York, DTP..

Donchin, E., 1979, Event-related brain potentials: a tool in the study of human information processing. In: Begleite, H. (Ed.), Evoked Brain Potentials and Behavior. Plenum Press, New York.

Dorrian, J., Lamond, N., Dawson, D., 2000, "The ability to self-monitor performancewhen fatigued", J. Sleep Res., vol. 9, pp. 137–44.

Eason, R.G., 1981, "Visual evoked potential correlates of early neural filtering during selective attention", Bull. Psychon. Soc. , vol. 18, pp. 203– 206.

Ebersole J.S., 2002, Current Practice of Clinical Electroencephalography. Lippincott, Williams & Wilkins.

Eoh, H.J., Chung, M.K., Kim, S.H., 2005, "Electroencephalographic study of drowsiness in simulated driving with sleep deprivation", International Journal of Industrial Ergonomics, vol. 35, pp. 307–320.

Eriksen, B.A., Eriksen, C.W., 1974, "Effects of noise letters upon the identification of a target letter in a nonsearch task", Percept Psychophysiology, vol. 16, pp. 143–149.

Fisch, B.J., 1991, Sphelmann's EEG Primer, 2nd ed. Elsevier Science BV, Amsterdam, The Netherlands.

Gale, A., Davies, R., Smallbone, A., 1977, "EEG correlates of signal rate time in task and individual differences in reaction time during a five-stage sustained attention task", Ergonomics, vol. 20, pp. 363–376.

Grandjean, E., 1988, Fitting the Task to the Man, Taylor & Francis, London.

Hart, S.G., Staveland, L.E., 1988, Development of NASA-TLX (Task Load Index): results of experimental and theoretical research. In: Hancock, P.A., Meshakati, N. (Eds.), Human Mental Workload. North-Holland, Amsterdam, pp. 39–183.

Hillyard, S.A., Münte, T.F., 1984, "Selective attention to colour and location: an analysis with event-related brain potentials", Perception Psychophysiology, vol. 36, pp. 185–198.
Hillyard, S.A., Mangun, G.R., Luck, S. J., Heinze, H.J., 1990, Electrophysiology of visual attention, in: E.R. John, T. Harmony, L.S. Pricheb, M. Valdez, P. Valdez (Eds.), Machinery of Mind, Birkhauser, Boston, MA.
Hitch, G.J., 1978, "The role of short-term working memory in mental arithmetic", Cognitive Psychology, vol. 10, pp. 302–323.
Isreal, J.B., Chesney, G.L., Wickens, C.D., Donchin, E., 1980, "P300 and tracking difficulty: evidence for multiple resources in dual-task performance", Psychology, vol. 17, pp. 259–273.
Isreal, J.B., Wickens, C.D., Chesney, G.L., Donchin, E., 1980, "The event-related brain potential as an index of display monitoring workload", Human Factors, vol. 22, pp. 211–224.
Klimesch, W., 1997, "EEG alpha rhythms and memory processes", International Journal of Psychophysiology, vol. 26, pp. 319–340.
Klimesch, W., 1999, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis", Brain Research Reviews, vol. 29, pp. 169– 195.
Kramer, A.F., Wickens, C.D., Donchin, E., 1983, "An analysis of the processing demands of a complex perceptual-motor task", Human Factors, vol. 25, pp. 597–622.
Kramer, A.F., Wickens, C.D., Donchin, E., 1985, "Processing of stimulus properties: evidence for dual-task integrality", Journal of Experimental Psychology, Human Perception and Performance, vol. 11, pp. 393–408.
Lafrance, C., Dumont, M., 2000, "Diurnal variations in the waking EEG: comparison with sleep latencies and subjective alertness", Journal of Sleep Research, vol. 9, pp. 243– 248.
Lal, S.K.L., Craig, A., 2001, "A critical review of the psychophysiology of driver fatigue", Biological Psychology, vol. 55, pp. 173–194.
Lange, J.J., Wijers A.A., Mulder, L.J., Mulder, G., 1998, "Color selection and location selection in ERPs: differences, similarities and Fneural specificity", Biol. Psychol., vol. 48 (2), pp. 153–182.
Laufs, H., Kleinschmidt, A., Beyerle, A., Eger, E., Salek-Haddadi, A., Preibisch, C., Krakow, K., 2003, "EEG-correlated fMRI of human alpha activity", Neuroimage, vol. 19, pp. 1463–1476.
Makeig, S., Jung, T., 1995, "Changes in alertness are a principal component of variance in the EEG spectrum", Neuroreport, vol. 7, pp. 213–216.
Mangun, G.R., Hillyard, S. A., 1987, "The spatial allocation of visual attention as indexed by event-related brain potentials", Human Factors, vol. 29, pp. 195–211.
Mangun, G.R., Hillyard, S. A., 1988, "Spatial gradients of visual attention: behavioral and electrophysiological evidence", Electroencephalogr. Clin. Neurophysiol., vol. 70, pp. 417–428.
Mangun, G.R., Hillyard, S.A., 1990, "Allocation of visual attention to spatial locations: trade-off functions for event-related brain potentials and detection performance", Percept. Psychophys., vol. 47, pp. 532–550.
Mangun, G.R., Hillyard, S.A., Luck, S.J., 1993, Electrocortical substrates of visual selective attention, in: S. Kornbloum, D.E. Meyer (Eds.), Atten. Perform., vol. XIV, Erlbaum, Hillsdale, NJ, pp. 219–243.
Morrel, L.K., 1966, "EEG frequency and reaction time– a sequential analysis", Neuropsychol., vol. 4, pp. 41–48.

Murata, A., Uetake, A., Takasawa, Y., 2005, "Evaluation of mental fatigue using feature parameter extracted from event-related potential", International Journal of Industrial Ergonomics, vol. 35, pp. 761–770.

Muscio, B., 1921, "Is a fatigue test possible?", British Journal of Psychology, vol. 12, pp. 31–46.

Neuman, U., Ullsperger, P., Gille, H.-G., Erdman, U., 1986, "Effects of graduated processing difficulty on P300 component of the event-related potential", Zeitschrift fur Psychology, vol. 194, pp. 25–37.

Niedermeyer, E., Da Silva, F.L., 1993, Electroencephalography: Basic principles, clinical applications, and related fields, 3rd ed. Williams & Wilkins, Maryland.

Oken, B.S., Salinsky, M., 1992, "Alertness and attention: basic science and electrophysiologic correlates", Journal of Clinical Neurophysiology, vol. 9 (4), pp. 480–494.

Okita, T., Wijers, A.A., Mulder G., Mulder, L. J. M., 1985, "Memory search and visual spatial attention: an event-related brain potential analysis", Acta Psychol., vol. 60, pp. 263–292.

Okogbaa, O.G., Shell, R.L., Filipusic, D., 1994, "On the investigation of the neurophysiological correlates of knowledge worker mental fatigue using the EEG signal", Applied Ergonomics, vol. 25 (6), pp. 355–365.

Picton, T.W. and Stuss, D., 1980, The component structure of the human event-related potentials, In: Kornhuber, H. H. and Deecke, L. (Eds.), Progress in Brain Research, Motivation, Motor and sensory processes of the Brain: Electrical Potentials, Behavior and Clinical Use, Vol. 54, Elsevier, Amsterdam, pp. 17–49.

Picton, T.W., 1988, Handbook of electroencephalography and clinical neurophysiology. In: Herbert, G., Vaughan, J.R., Joseph, C.A., (Eds.), The Neural Basics of Event-Related Potentials, Vol. 3, Elsevier, Amsterdam, pp. 45–96.

Pyun, H.G., Kim, J.R., 2000, "A study on the effect of emotion-evoking advertisement with EEG analysis", Proceedings of 2000 Joint Conference of KIIE and KORMS, KIIE and KORMS, Seoul, pp. 413–416.

Rains, J.C., Penzien, D.B., 2003, "Sleep and chronic pain challenges to the a-EEG sleep pattern as a pain specific sleep anomaly", Journal of Psychosomatic Research, vol. 54, pp. 77–83.

Rugg, M.D., Milner, A.D., Lines, C.R., Phalp, R., 1987, "Modulation of visual event-related potentials by spatial and non-spatial visual selective attention", Neuropsychologia, vol. 25, pp. 85–96.

Ryu, K, Myung, R., 2005, "Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic", International Journal of Industrial Ergonomics, vol. 35, pp. 991–1009.

Scheffers M.K., Humphrey D.G., Stanny, R.R., Kramer, A.F., Coles, M.G., 1999, "Error-related processing during a period of extended wakefulness", Psychophysiology, vol. 36, pp. 149–57.

Seeley, R.R., Stephens, T.D., and Tate, P., 2003, Anatomy and physiology, Boston, Mass: McGraw-Hill.

Semlitsch, H.V., Anderer, P., Schuster, P., Presslich, O., 1986, "A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP", Psychophysiology, vol. 23, pp. 695–703.

Smith, M.E., McEvoy, L.K., Gevins, A., 2002, "The impact of moderate sleep loss on neurophysiologic signals during working-memory task performance", Sleep, vol. 25, pp. 784–94.

Uetake, A., Murata, A., 2000, "Assessment of mental fatigue during VDT task using event-related potential (P300)", Proceedings of the 2000 IEEE International Workshop on Robot and Human Interactive Communication, pp. 235–240.

Ullsperger, P., Neuman, U., Gille, H.-G., Pietschann, M., 1986, P300 component of the ERP as an index of processing difficulty. In: Flix, F., Hagendor, H. (Eds.), Human Memory and Cognitive Capabilities. North-Holland, Amsterdam.

Ullsperger, P., Metz, A.-M., Gille, H.G., 1988, "The P300 component of the event-related brain potential and mental effort", Ergonomics, vol. 31, pp. 1127–1137.

Waard, D., Brookhuis, K.A., 1991, "Assessing driver status: a demonstration experiment on the road", Accident Analysis and Prevention, vol. 23 (4), pp. 297–307.

Wijers, A.A., Lamain, W., Slopsema, S., Mulder, G., Mulder, L. J.M., 1989, "An electrophysiological investigation of the spatial distribution of attention to coloured stimuli in focussed and divided attention conditions", Biol. Psychol., vol. 29, pp. 213–245.

Wijers, A.A., Mulder, G., Okita, T., Mulder, L.J.M., Scheffers, M.K., 1989, "Attention to colour: an ERP-analysis of selection, controlled search, and motor activation", Psychophysiology, vol. 26 (1), pp. 89–109.

12

Risk Management in the Development of New Products in the Pharmaceutical Industry

Ewa J. Kleczyk
Advanced Analytics,TargetRx,Horsham, Pa
USA

1. Introduction

1.1 Trends in R&D spending and production of new drugs

Due to the excessive new product research opportunities and limited financial resources, deciding which new pharmaceutical products to develop can be a challenging and lengthy process for many pharmaceutical companies. The returns on investment are attractive, but they vary considerably between drugs. New pharmaceutical products usually undergo costly and time-consuming testing before receiving government approvals for distribution to patients (Congressional Budget Office, 2006).

Only about one percent of researched chemical molecules withstands the three phases of clinical trials, the scrutiny of the Food and Drug Agency (FDA), and becomes available for patient use. In addition, research and development (R&D) costs can reach more then $800 million to develop and test a potential drug, so the selected product must return at least the accrued financial expenditures over its lifecycle (Nelson, 2009). With such high development costs and low probability of product success, project-prioritization and new product-portfolio selection are of high importance to pharmaceutical managers. Trading off available resources and investment opportunities helps identify drugs worthwhile to bring to market (Ogawa & Piller, 2006).

In the pharmaceutical industry, the risk management problem includes deciding which new products to develop, continue to research, terminate, and invest in. These decisions include trading off risks, returns, and time horizons for future payoffs. In theory, such tradeoffs are easily undertaken by optimization problems; however, the complexity and uncertainty of the new drug development process can make the optimal solution hard to obtain, and may result in employing less complicated, and therefore, less precise methods of new product identification (Gino & Pisano, 2006).

This chapter will focus on assessing the different risk management methods employed by pharmaceutical executives in the new product portfolio evaluation and consequently, which pharmaceutical products to bring to market. First, however, a review of the product development process, as well as the costs associated with research and development of new drugs in the U.S. will be presented. The process of product development will be described as it happens in the United States, although the R&D approach is not that different between the U.S. and the European countries. After the short R&D process summary, a description of the

risk assessment methods will follow. Pharmaceutical executives will find this chapter useful in making their product portfolio investment decision, as it will list several well known and widely used techniques of risk evaluation, as well as provide guidance on how they compare to each other, how they differ, and when should they be used.

2.The cost of developing a new drug

Over the past 20 years, the total costs associated with research and development (R&D) of new drugs has tripled. In 1980, U.S. pharmaceutical companies spent a total of $5.5 billion on research and development of pharmaceutical products, while in 2003, these costs increased to more than 17 billion (NSF, 2005). Continued growth in the R&D spending, however, has just a small effect on the pace at which new drugs have been developed in the past 20 years, as the number of innovative molecules in research has steadily increased (Congressional Budget Office, 2006).

On average, it is estimated that R&D of new innovative pharmaceutical products costs nearly $802 million, and takes about 12 years for a pharmaceutical company to bring a drug to market (DiMasi et al., 2003). The R&D cost estimates include the actual accrued R&D expenditures that are estimated at $403 million, as well as expenditures of failed projects and the value of foregone alternative investments, which in total amount to $399 million (DiMasi et al., 2003; Rawlins, 2004).

R&D costs for new drugs are highly variable, and depend on several factors that include the type of drug being developed, whether the drug is based on either a new molecular entity (NME) or it is an incremental modification of an existing product, the likelihood of product failures and government agency (i.e. Food and Drug Agency) approvals, and finally the expected revenues associated with product sales (Congressional Budget Office, 2006). In the next few sections, these topics will be described, as well as their impact on driving the R&D costs and development decisions of new and innovative pharmaceutical products.

2.1 Types of pharmaceutical products in development

2.1.1 Acute illness vs. Chronic disease product research

Until the late 1980s, pharmaceutical companies invested mostly in treating acute illnesses[1], such as common colds, flu, and headaches, as these products are usually cheaper to develop, and can provide a quick return on investment (Congressional Budget Office, 2006). In the past 30 years, the industry's developmental efforts have grown to include therapeutic classes, such as diabetes, cancer, and cardiovascular diseases. These diseases are referred to as chronic illness[2], and tend to develop slowly over time. They can never be cured, and require advanced treatment (Congressional Budget Office, 2006).

The shift in the product development type is associated with the changing population demographics. For example, today in the U.S., there are almost 100 million adults that are 50

[1] An Acute Illness typically starts suddenly and is short lived. Two common examples are colds and the flu. Acute illnesses, caused by viruses, may go away by themselves, while others can be cured either with antibiotics or other medical treatment (Carlson, 2008).

[2] A Chronic Illness develops slowly over time and lasts a long time. Examples of common chronic illnesses include diabetes, arthritis, congestive heart failure, and Alzheimer's disease. Chronic conditions are typically caused by multiple factors including family history, diet, stress levels, and surrounding environment. Some chronic diseases will never be cured, and require advanced treatment (Carlson, 2008).

years old or older, and every year more than 3.5 million of Americans join this age group (Pirkl, 2009). As a result of the changing demographics, the need for treating chronic conditions has been also increasing. It is estimated that a 1% growth, in the potential market for a category of drugs treating chronic disease, leads to an increase of roughly 4% in the entry of new drugs in that category (Lanjouw & Cockburn, 2001). In addition, growth in sales revenue for these types of drugs has provided the financial opportunities for additional research and development in this area, resulting in an increase in the number of targets in development from 500 to more than 3,000 in recent years (Congressional Budget Office, 2006).
As the pharmaceutical products treating elderly population grows over time, the number of new drugs in therapeutic areas associated with treatment of young people, such as attention deficit hyperactivity disorder (ADHD), juvenile diabetes, pediatric vaccines, has declined in the recent years. The decrease is associated with the continuously declining number of births (in 2009, the U.S. birth rate was 14%), and consequently, lower expected returns on R&D investment related to products for treatment of children and teenagers (Center for Disease Control, 2010).

2.1.2 New molecular entity vs. Incremental modification of an existing product

The cost of R&D typically depends on the type of developmental drugs being pursued by pharmaceutical companies. There are two types of products developed in the pharmaceutical industry: new molecular entity and an incremental modification of an existing product. New molecular entity (NME) is defined as a drug that contains no active molecule previously approved by the FDA. In addition, an NME can also represent a 'me-to-drug', which is still an innovative entity, but works in a similar way to an NME already available for patient use (Congressional Budget Office, 2006).
The other drug category is an incremental modification of an existing product. The product modifications can include changes in drug delivery system, dosing scheme, as well as obtaining additional treatment and indication approvals ('new label'). Most pharmaceutical companies pursue testing of current products to identify opportunities for patent extension for other product uses (Congressional Budget Office, 2006).
On average, it is more costly to develop an NME compared to the incremental modification product, due to a longer time frame for development and testing, a higher probability for a clinical trial failure, and a more restrictive FDA approval system. On average, R&D costs of an NME are between $300 and $500 million higher compared to the product extension research costs (Piturro, 2006; Rawlins, 2004).

2.2 The likelihood of product failures in clinical trials

Research and development of a new pharmaceutical product relies heavily on clinical research and product testing. Product clinical trials are conducted to collect safety and efficacy data on the researched molecule. On average, only 1% of tested products successfully completes the three phases of testing, and can accrue more then $450 millions in R&D expenses (Rawlins, 2004). If a drug successfully passes through clinical trials, it usually is approved by the Food and Drug Agency (FDA) (or a European equivalent) for use in the general population (Pocock, 2004).
Phase I clinical trials are the first stage of molecule testing in patients, and require from 20 to 100 healthy volunteers. The test phase takes about 1 year to 2 years, and includes trials designed to assess safety and tolerability of a potential drug (Pocock, 2004). Once the initial

safety level of the studied drug has been confirmed in Phase I trials, Phase II clinical trials are performed on a larger group of about 20-300 patients to assess how well the drug works. They usually take anywhere from 2 to 4 years. Phase III clinical trials are performed on large patient groups, usually of 300–3,000 patients, depending upon the disease and medical condition studied. These experiments examine how effective the drug is in comparison with current treatments. Due to the size and comparatively long duration, even up to 6 years for chronic illness trials, Phase III trials are the most expensive, time-consuming, and difficult tests to design and successfully run (Pocock, 2004).

2.3 Food and drug agency new drug approvals

Once the new pharmaceutical product positively tests in clinical trials, the next step is to obtain an approval from a federal agency responsible for regulating pharmaceutical products available for patient use. In the U.S., this agency is called the Food and Drug Administration (FDA), while in the European Union it is called the European Medicines Agency (EMEA).

In the U.S., the Food and Drug Administration (FDA) is the federal agency responsible for protecting and promoting public health through the regulation and supervision of food, pharmaceutical, and healthcare products. The group responsible for the pharmaceutical product evaluation is the Center for Drug Evaluation and Research (CDER). The center evaluates new drugs before their availability for patient use, while ensuring that potential drug candidates work correctly, and their health benefits to patients outweigh their known risks. The review process can take up to two and a half years, and the obtained approvals allow the approved product to be sold only with a prescription (FDA, 2010).

On average, one-third of new drugs applications submitted to the FDA are for new molecular entities (NMEs). Most of the rest are for incremental modification of existing drugs, which include the additional health conditions, for which an existing drug can be prescribed. In the past 15 years, the FDA approval rate declined, and the total number of NMEs approved each year fell from 53 in 1996 to 20 in 2005. The drop in approvals might be a result of longer research and development (R&D) cycles, and increased scrutiny of new pharmaceutical products by federal agencies (Congressional Budget Office, 2006).

2.4 Sales, revenue, advertisement, and patent impact on development of new drugs

The current and future R&D expenditures are also associated with the expected sales and revenue trends from launching new drugs to market. Usually, the potential product will not be investigated if it is expected not to recover accrued R&D costs.

2.4.1 Sales and revenues of new drugs

In the past 20 years, the U.S. profit growth has been about the same every year during that time period. The average yearly return on revenue is about 17%. The high and consistent growth places the pharmaceutical industry as the third most profitable of all industries in the U.S., and second best industry to invest in (Fortune 500, 2009). In the past decade, retail sales of prescription drugs has increased by 250% from $72 billion to $250 billion, while the average price of prescriptions has more than doubled from $30 to $68 (Census Bureau, 2008).

The continued profit growth is partially related to the drug exclusivity rights, ranging from 3 years to 20 years after product approval for patient use. Patent protection enables the pharmaceutical companies to recover the costs of research and development through high profit margins for their drugs. When the patent protection for the pharmaceutical product

expires, a generic drug is usually developed and sold by a competing company (Kaufman, 2005).

2.4.2 Managed care and formulary status impact on new drug success

Besides product's exclusivity rights, managed care system and formulary status of the pharmaceutical drugs also impact the future profitability levels of the pharmaceutical industry. Private insurance (i.e. Keystone and Aetna) or public health bodies (i.e. Medicare and Medicaid) can restrict the drug access to patients through the use of formularies and required out-of-pocket expenses (Shih & Sleath, 2004).

Government agencies also impact the prices and availability of pharmaceutical products by passing laws and bills enabling a greater access to healthcare services and drugs. For example, in 2010, the U.S. Congress passed a Health Care Bill, mandating all American citizens to purchase either a privately owned or government provided insurance plan to improve a public access to healthcare services and providers, as well as pharmaceutical products (Tumulty, 2010).

2.4.3 Pharmaceutical brand advertising impact on new drug success

The last factor, impacting sales, revenue, and profitability of pharmaceutical companies, is advertising of pharmaceutical products already available for patient use. In the U.S., pharmaceutical companies spend nearly $19 billion a year on pharmaceutical product promotion to impact sales numbers and profitability margins of their products (Moynihan, 2003).

Product advertising is common in healthcare journals, as well as through more mainstream media routes, such as radio and TV (Moynihan, 2003). Pharmaceutical companies also promote directly to healthcare providers via employing sales representatives. Every year more then $5 billion is spent to support this type of promotion (Robinson, 2003). Finally, with the technological development of computers and handheld devises, such as Smart Phones and iPads, pharmaceutical brand advertising has also moved into the digital arena. Brand specific websites, as well as electronic detailing to physicians have become a popular venue of pharmaceutical promoting in the last 5 years (Howie & Kleczyk, 2011b).

2.5 Summary of the new pharmaceutical product R&D process and associated costs

New pharmaceutical products usually undergo costly and time-consuming testing before receiving government approvals for distribution to patients (Congressional Budget Office, 2006). Only about one percent of researched chemical molecules can withstand the three phases of clinical tests, the scrutiny of the Food and Drug Agency (FDA), and becomes available for patient use. Research and development (R&D) costs have reached more then $800 million, and the product development process takes 12 years to complete. As a result, the selected pharmaceutical molecule must return at least the accrued financial expenditures over its lifecycle (Nelson, 2009).

With the changes in demographic population, as well as enhancements in technology, more emphasis is placed on chronic illness product development, instead of acute illness product development. Although these drugs are more expensive and require more time to develop, they have the opportunity to return not only the invested financial capital, but also increase significantly net profits of the pharmaceutical companies, due to the changing population demographics towards a higher proportion of elderly citizens. With the continued high

spending allocated to advertising of pharmaceutical products, as well as increased use of internet and digital media to inform healthcare providers and patient population of their treatment options, expected sales and revenues can be increased even more. The only barrier in the entire process is the rate of FDA approvals and the formulary status of the new products, which tend to slow down the speed at which products are brought to market, as well as their affordability and access to the patient population.

3. Risk management evaluation methods

Deciding which new products to develop is a major challenge for many pharmaceutical companies with an excess of opportunities, but limited resources. Project prioritization and new product-portfolio selection has long been the domain of the new product arm of the corporation (Blau et al., 2000). Pharmaceutical product development, as any other management task, requires important decisions about the tradeoffs between the available resources, as managers decide which drugs to bring to market (Ogawa & Piller, 2006).
Assuming a fixed research and development budget, the management problem includes deciding which new products to develop, continue to research, terminate, and invest in. In making these decisions, managers face tradeoffs between risks, returns, and time horizons for future payoffs. In theory, such tradeoffs are easily tackled by optimization problems; however, the complexity and uncertainty of the new drug development process can make the solution hard to obtain, and result in employment of less complicated, and therefore, less precise methods of new product identification (Gino & Pisano, 2006).
Currently in the pharmaceutical industry, there is no one recommended method of risk assessment for evaluation of investment opportunities. There are, however, a variety of methods cited that can help mangers in making these decisions. Depending on the needed precision, complexity, and objectives of the analysis, the pharmaceutical managers can choose between different risk management methods to meet their study goals. Due to the importance of selecting the right approach of risk evaluation, and making the right decisions in selecting products for investment, several of the currently utilized methods will be reviewed and evaluated in this part of the chapter (Howie & Kleczyk, 2011a).
There are two types of risk management methods: Net Present Value (NPV) methods and Consumer Theory based approaches. These NPV based methods include Net Present Value of Income analysis, Capacity Constrained NPV approach, and Stochastic Dominance. All of the above methods account for the financial impact of the chosen alternatives (Grabowski & Vernon, 1998; Blau et al., 2000; Smit & Trigeorgis, 2006). The Consumer Theory based approaches do not take into account the financial aspects of the new product development and analyze consumers' preferences for different product alternatives instead. These models usually involve Conjoint Analysis / Discrete Choice models, determining the most preferred product attribute mix (Dakin et al., 2006).
The above methods will be compared to each other on the basis of the inputted information (i.e. R&D expenditures, future drug prices, and probability of FDA approvals, etc.), complexity of the theoretical model (i.e. mathematical simulations, econometric and statistical analyses), as well as the precision and reliability of the theoretical frameworks, in selecting product portfolios with the highest return on investment.

3.1 Net present value (NPV) based risk assessment methods

There are several NPV (otherwise known as a payoff) based methods of the new product development identification process. These approaches include Net Present Value of Income

(NPV) analysis, Capacity Constrained NPV approach, and Stochastic Dominance analysis. All of these methods account for the financial impact of chosen alternatives, but differ by their complexity level, precision, and reliability of product selection (Grabowski & Vernon, 2000; Blau et al., 2000; Smit & Trigeorgis, 2006).

3.1.1 Net present value of income analysis

Until the late 1980s, cash flows, expected returns, and net present value of income were the key variables in the decision-making process of the new drug development and investment. The relationship between investment and cash-flow statements provided pharmaceutical managers with a working framework for resource-allocation decisions (Grabowski & Vernon, 1998). This most widely used framework, called the Net Present Value (NPV) of Income, is described by cash inflows and outflows being discounted back to their present value (PV), and then being summed together. As a result, NPV of Income is the sum of all of following terms:

$$NPV = \sum R_t / (1+i)^t, \quad (1)$$

where t is the time of the cash flow, i is the discount rate defined as the rate of return that could be earned on an investment in the financial markets with similar risk, and R_t is the net cash flow (the amount of cash inflow minus cash outflow) at time t (Khan, 1993). This analysis is performed for every potential product, and the drug with highest NPV of income is usually selected for pharmaceutical investment, and future market availability and sales.

Although the NPV of Income framework provides a very simple and clean approach of investment profitability, as potential product investments can be ranked by their NPV amount, it is still the subject to change, and depends on a range of prices and operating costs associated with the investment and development of new pharmaceutical products. Demand, drug prices, as well as development and operating costs are the source of uncertainty within the framework. Modeling this uncertainty is the primary struggle observed within this approach (Grabowski & Vernon, 1998).

3.1.2 Capacity constrained NPV approach

In the early 1990s, pharmaceutical managers started leveraging a Capacity Constrained NPV approach to evaluate new potential pharmaceutical products. This method includes analysis of capacity planning and development management. This approach not only focuses on the cash flows and NPV framework, but also on the rate of FDA approvals and success rate of clinical trials. As a result, the new additions to the model account for the uncertainty associated with the dynamics of the pharmaceutical market (Rogers et al., 2004).

In 2000, Blau et al. introduced a probabilistic simulation model of a pharmaceutical product development into this framework to prioritize candidate drugs, based on their risk-reward ratios. Their approach captures the complexity of the new pharmaceutical product research and development process, by incorporating probability of clinical trial success into the NPV concept (Blau et al., 2000; Lave et al., 2007). As a result, this model helps select innovative product candidates that provide an acceptable exposure level to risk, while also providing adequate financial returns.

The chosen risk level depends on the risk attitude of the management and stakeholders, as well as the status of the current commercial products and the characteristics of new drug candidates already in the development pipeline. A risk-averse management might prefer

molecules with high technical success ratio and low resource requirement, while a risk-taking management might be willing to push molecules with greater returns at a greater risk (Blau et al., 2000). Most of the R&D drug information is generally available from researchers and engineers developing these products, while sales and marketing executives can provide estimates for expected sales upon marketplace launch (Blau et al., 2000).

A simulation model, using data representing R&D related variables, as well as expected sales and revenues, is usually used to analyze the different investment options, while incorporating the risk-reward analysis and the probability distribution of production success. Once a portfolio of molecules is selected, the next issue is the speed at which these molecules can be pushed through the developmental and production pipeline, without violating the resource constraints, and therefore maximizing the net present value (NPV) of the selected portfolio. This is a 'resource constrained scheduling problem under uncertainty' and involves use of linear mathematical programming for the analysis (Blau et al., 2000). The problem is usually described as maximizing a NPV function subject to multiple constraints (financial and human capital resources):

$$\text{Maximize a NPV function: } F(x_1; x_2; \dots ; x_n) \tag{2}$$

Subject to the following constraints:

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,n}x_n &\le b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,n}x_n &\le b_2 \\ a_{3,1}x_1 + a_{3,2}x_3 + \dots + a_{3,n}x_n &\le b_3 \\ &\dots \\ a_{n,1}x_1 + a_{n,2}x_2 + \dots + a_{n,n}x_n &\le b_n, \end{aligned} \tag{3}$$

where $(x_1 \dots x_n)$ define the inputs for the NPV optimization, and b_n represents the constraint values. Constraints are usually defined as human and financial capital, time frame of product development, expected sales and revenues, and any other important to the management variables that should be controlled for when deciding, which potential molecules to invest in and bring to market (Champ et al., 2003).

The constrained problems can range from a single project optimization with no resource constraints (Schmidt & Grossmann, 1996) to a more complicated problems defined by sequencing and scheduling of multiple testing tasks under resource constraints for a fixed set of products (Jain & Grossmann, 1999). To estimate the latter approach, a mixed-integer linear programming (MILP) model is usually utilized, and maximum resource availability is employed to enforce resource constraints (Honkomp, 1998). In 2003, Submarinian et al. even further extended the framework by formulating a simulation-optimization problem that combines mathematical programming with discrete choice simulation to also account for planning and scheduling uncertainty.

Although these models account for the high level of complexity regarding new product development, they tend to be time-consuming, and are not easily executable by pharmaceutical managers. Consequently, not many pharmaceutical executives actively use this type of product portfolio optimization methodology, and tend to turn to easier and more practical ways of deciding, which products to develop and bring to market (Baker, 2002).

3.1.3 Stochastic dominance method

As the Capacity Constrained NPV approach tends to be time-consuming, and rather difficult to implement by pharmaceutical management, Kleczyk (2008) applied a Stochastic Dominance methodology to eliminate the complexity in the decision of new chemical molecule investment. Stochastic Dominance evaluates the pharmaceutical product development process and chemical molecule prioritization via accounting for not only the uncertainty in drug prices, but also for development and operating costs related to product research and development (R&D). In addition, it is an intuitive and easily implemented tool that is uniquely suited to the objectives of new product development selection process (Kleczyk, 2008).

Stochastic Dominance is usually employed in the analysis of financial portfolio optimization, which attempts to maximize financial portfolio's expected return for a given amount of risk, or equivalently to minimize risk for a given level of expected return, by carefully choosing the appropriate investment choices (Edwin et. al, 1997). The basis for this method is not only how each potential product performs on their own (i.e. NPV), but also how each potential product changes its expected revenues relative to other products' changes in their expected revenues too (Edwin et al, 1997). The analysis includes trading off risk and expected returns. For example, for a given amount of risk, the method describes how to select a potential product with the highest possible expected return; and for a given expected return, how to select a drug with the lowest possible risk (Markowitz, 1952).

The framework makes many assumptions about pharmaceutical managers and drug companies, including the use of Normal Distribution function[3] to model expected returns, the utility maximization framework[4], unlimited credit availability to the pharmaceutical companies, and no transaction costs or federal and state taxes. Unfortunately, in reality, some of these assumptions, such as no transaction fees and unlimited credit amount available for lending, are relaxed to better represent the current environment, and provide realistic estimates of potential chemical molecules' payoffs. As a result, more complex versions of the financial portfolio model can take into account a more sophisticated view of the world, such as one with non-normal distributions and taxes (Markowitz, 1959; Shleifer, 2000).

There are two types of Stochastic Dominance methods that can be employed in the analysis of potential pharmaceutical products for market use: First and Second Degree Stochastic Dominance. The First Degree Stochastic Dominance (FSD) informs which potential product's NPV distribution dominates all other choices. For example, if a decision maker prefers NPV distribution for molecule 1, which is mathematically presented as $[f(x_i)]$, to NPV distribution for molecule 2, which is mathematically presented as $[g(x_j)]$, then $f(x_i)$ dominates $g(x_j)$ by FSD:

$$f(x_i) \geq g(x_j) \text{ by FSD.} \quad (4)$$

As a result, the cumulative probability distribution function[5] of NPV for molecule 1, $[F(x_i)]$, is less or equal to cumulative probability distribution function of NPV for molecule 2, $[G(x_j)]$, (Kleczyk, 2008):

---

[3] Normal Distribution Function describes real-valued random variables that tend to cluster around a single mean value (Varian, 1992).

[4] The Utility Maximization Framework represents maximization of a utility function based on a specified pharmaceutical company's financial resource constraint requirement (Varian, 1992).

[5] Cumulative Probability Distribution Function represents the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x (Varian, 1992).

$$F(x_i) \leq G(x_j). \quad (5)$$

Furthermore, when molecule 1 dominates molecule 2, the expected value of the payoff for molecule 1, [NPVf(x_i)], will be greater than the expected value of the payoff for molecule 2, [NPV(g(x_j)] (Kleczyk, 2008):

$$NPV(f(x_i)) \geq NPV(g(x_j)). \quad (6)$$

The other commonly used type of Stochastic Dominance is the Second Degree Stochastic Dominance (SSD). For two chemical molecules 1 and 2, molecule 1 has second-order stochastic dominance over product 2, if the former is more predictable, involves less risk, and has at least as high of a mean. All risk-averse expected-utility maximizers prefer a second-order stochastically dominant potential product to a dominated product (Kleczyk, 2008). In terms of cumulative probability distribution functions: [$F(x_i)$] of NPV and [$G(x_j)$] of NPV, chemical molecule 1 is second-order stochastically dominant over molecule 2, if and only if, the area under [$F(x_j)$] of NPV is less than or equal to that under [$G(x_j)$] for all real numbers [R] x, with strict inequality at some x:

$$\int_{-\infty}^{x} F(x_i) \leq \int_{-\infty}^{x} G(x_j), \text{ for } x \in R. \quad (7)$$

The analysis typically assumes that all managers are risk-averse, and therefore, no investor would choose a potential molecule that is second-order stochastically dominated by some other molecule (Kleczyk, 2008).

The inputs needed to perform this type of analysis are similar to those used in the NPV of Income and Capacity Constrained approaches, and include the excepted sales, revenues, potential product sales price point, operating costs, taxes, as well as the probability of passing the clinical trial and being approved by the FDA. The risk level can be adjusted depending on the management risk aversion level and product portfolio in the company's pipeline. The above analyses are usually conducted using Monte Carlo simulation and sensitivity analysis to indentify, which potential products are worth of pharmaceutical companies to invest in. The final stage of the analysis involves comparing between NPV values for each molecule, and choosing a molecule with the highest NPV, as a recommendation for pharmaceutical company's investment. The model can be extended to incorporate a linear programming approach, in order to add workforce and planning constraints into the model. The extension, however, implies that the analysis may again morph into a complex and time-consuming framework that might be difficult for pharmaceutical executives to execute on (Kleczyk, 2008).

3.1.4 Summary of the NPV based risk assessment methods

In summary, the basic NPV of Income analysis and Stochastic Dominance are simple methods to implement by pharmaceutical management, when prioritizing portfolio of chemical molecules. The NPV framework has been used for more than 80 years in the decision of resource allocation by estimating the net present value of the expected future revenues and expenses. Stochastic Dominance allows for executing against the financial and strategic company goals, while controlling for the important factors of the FDA approvals rate, the clinical trial success rate, expected sales and revenues, operating costs, and the tax base. Both of the above approaches allow for comparing NPVs of all potential products, and choosing the molecule with the highest expected NPV.

The items not accounted for in these two approaches are the production capacity constraints that include production scheduling, human capital availability, and currently produced and available for patient use drugs. The Capacity Constrained NPV approach accounts for these constraints by including them in the maximization of the NPV function. Both of the Stochastic Dominance and NPV methods can be extended to include the additional assumptions; however, the optimization process may become more complicated, time-consuming, and therefore might be not easily understood by pharmaceutical executives.

3.2 Consumer theory based risk assessment methods

There are also other risk management methods that help in deciding, which new chemical molecules to invest in and bring to market. They do not necessarily take into account the financial aspects of new product development, but rather analyze healthcare providers' preferences for different potential product alternatives. These models are usually based on the Consumer Theory, and involve employment of Conjoint Analysis (CA) (otherwise known as Discrete Choice Analysis) models, determining the most preferred new product attribute mix (Dakin et al., 2006).

The Conjoint Analysis (CA) framework has been applied successfully to several marketing decisions, including designing of new products, targeting market selections, pricing of new products, and studying competitive reactions. One of the advantages of CA is its ability to answer various 'what if' questions when employed for analysis of hypothetical and /or real choice alternatives (Rao et al., 2008). This approach, however, can be very lengthy and complicated, due to the multiple steps required to design and complete the research. The required steps include, but are not limited to: a development of survey instruments, development of product stimuli based on a number of potential pharmaceutical product attributes in consideration, interviews of healthcare professionals and patients, an econometric and statistical analysis, mathematical simulations, and employment of the estimates in tackling any managerial problems, such as new product forecasting.

The Conjoint Analysis (CA) is a stated-preference study that uses a survey instrument, as well as an experimental design to elicit pharmaceutical costumers' preferences for pharmaceutical goods. Pharmaceutical customers are usually represented by healthcare providers and patients. They participate in market research studies to provide their responses to survey questions, regarding alternatives of pharmaceutical products, varying in attribute levels to inform their preferences for multiple states of a potential drug. The introduction of the expected drug price and / or formulary status of the potential product, as an attribute, extends the application of the method into welfare analysis. Based on the preference function knowledge, simulation and optimization algorithms aid the process of determining the preference level for each product-attribute combination (Champ et al., 2003; Rao, 2007).

3.2.1 Theoretical framework: random utility maximization theory

The theoretical model, guiding the CA preference elicitation methodology, is the Random Utility Maximization (RUM). RUM is based on consumers' (i.e. healthcare providers and patients) choices from a set of competing alternatives of potential pharmaceutical products. Each survey respondent chooses the most preferred alternative from a set of drug alternatives, while at the same time making tradeoffs between attributes of each alternative. Each respondent is trying to select an alternative that would provide them with the highest satisfaction, otherwise known as utility (Champ et al., 2003).

The basic problem of utility or preference maximization represents the set of all pharmaceutical chemicals (alternatives) satisfying financial resource constraints. The financial constraint can include the financial restrictions of pharmaceutical companies, healthcare providers, and patients. The company's primary end-users' (i.e. physicians, nurses, patients, etc.) are assumed to have preferences for each potential new product within a developmental product set X. As a result, the preference maximization problem is defined as maximization of a utility function based on a specified financial resource constraint requirement (Varian, 1992):

$$\text{Maximize utility function: } u(x) \tag{8}$$

$$\text{Subject to: } px \leq m, \text{ where } x \text{ is in } X, \tag{9}$$

where u(x) represents the utility function, and $px \leq m$ represents the financial resource constraint, with m being the fixed amount of money available to a company for product R&D, as well as healthcare providers' and patients' available funds for medical and healthcare needs (Champ et al., 2003).

3.2.2 Survey development process

In order to forecast a product potential, a survey instrument has to be developed first. Healthcare providers are usually invited to participate in questionnaires created to elicit their preferences and attitudes regarding a set of potential products. The survey format varies from a paper version to an internet based exposure. The collected data is then analyzed via employing econometric and statistical tools. A Conjoint Analysis depends on the design of stimuli, which describes potential pharmaceutical product profiles. Employment of experimental design allows generating a set of potential drug profiles for review when surveying respondents (Champ et al., 2003; Rao, 2007).

The survey information collected from healthcare providers include preference rating data of selected product alternatives, ranking of product profiles, and choosing the preferred product over another option. In case of preference rating surveys, ratings are collected from respondents using attribute based pharmaceutical product profiles. The rating scale questions appeal to many researchers, due to the simplicity of the econometric analysis, and the ease with which respondents answer rating questions. The rating scales values can rage from 1 thru 5 values, where 1 is the least preferred, and 5 is the most preferred, to 1 thru 9 values, implying the same preference scheme, but a larger response variability (Champ et al., 2003).

In the choice-based surveys, respondents rank a set of product profiles from most to least preferred. For example, a preferred hypertension product is selected from among multiple alternatives, described by set of product attributes, such as product efficacy and safety (Champ et al., 2003). The framework assumes the most preferred profile to be chosen first from the choice set, followed by the second ranked alternative chosen from the remaining choice set, and so forth. The participants might get fatigued, while proceeding through the sequence of choices, which in turn might result in unreliable analysis, and imprecise potential product forecast (Champ et al., 2003).

In addition to selecting their preferred pharmaceutical product, healthcare providers are asked to present their anticipated use of the chosen alternatives when it becomes available on the market, as well as indentify the change in the use of the current treatment options, as

a result of the new entrant. Based on their responses, market share forecast for new products are developed to help in the decision-making process. The forecasts might vary from one single data point to monthly 1 to 5 year forecasts, depending on the need and confidence in the product potential predication (Howie & Kleczyk, 2011a).

3.2.3 Econometric and statistical analysis

The most commonly used econometric and statistical models, employed in estimation of healthcare providers' preferences and attitudes for new products, include logit and probit models. Depending on the type of data collected, either binary or multinomial logit and probit models are employed. Binary choice models relate to either selecting or not selecting a presented product alternative; while multinomial models relate to choosing a product alternative from a provided set (ranking or rating exercise). Multinomial probit and logit models are more often selected for the analysis of choosing the right product attribute combination, as pharmaceutical managers are mostly faced with evaluation of multiple chemical molecule alternatives at one time.

Both probit and logit models are based on the utility maximization framework of a healthcare provider choosing a particular pharmaceutical alternative over another. As a result, the respondent also maximizes the probability of a potential product being chosen from a presented set of alternatives. The probability specification is expressed as a function of observed variables, relating to the pharmaceutical product alternatives and the respondent. In its general form, the probability [P_{ni}] that a person n chooses a molecule alternative i is expressed as follows:

$$P_{ni} = P(x_{ni}, x_{nj}; s_n, \beta) \text{ for } j \neq i, \tag{10}$$

where x_{ni} is a vector of attributes of molecule i faced by a healthcare provider n, x_{nj} is a vector of attributes of the other alternatives (other than i) faced by person n, s_n is a vector of characteristics of person n, and β is a set of parameters that relate variables to probabilities, which are estimated statistically. A vector of respondent characteristics includes the type of treatment expertise, age, gender, geographic area of practice, and healthcare provider's function at the medical office (i.e. nurse, physician, etc).

Today's econometric and statistical software include both of the above models in their analysis menu, so the preferred pharmaceutical molecule evaluation is easily executable. The econometric model estimates the coefficients for each product attributes [β], which then allow identifying the preferred product alternative with the greatest market potential. Mathematical optimizations and simulations are usually employed to compute forecasts for the different product alternatives and attribute combinations (Champ et al., 2003).

3.2.4 Product potential analysis and forecasting

The described econometric and statistical models allow indentifying the preferred chemical molecule for research and development. In addition, the results are used to simulate and compute market potential of each alternative, in terms of product market share, as well as expected sales and revenues. The market potential forecast can span from just one data point at the end of a defined time period to monthly predictions, spanning from 1 year to even to up to 5 years after product availability for patient use. These data points are analyzed by pharmaceutical mangers to inform their investment decisions into new chemical molecules. Alternatives with the greatest market potential are usually considered for the R&D investment (Champ et al., 2003; Howie & Kleczyk, 2011a).

While surveying healthcare providers, to learn their preferred treatment options, is the appropriate approach to learning the new product potential, the problem with this approach is respondents' ability to correctly identify the 'future drug use,' once approved for patient use. These product 'future use' estimates can be unreliable and overstate the future prescribing behavior of the pharmaceutical drugs. As a result, the overstatement leads to an unreliable forecast for the product potential (Howie & Kleczyk, 2011a).
While experience may provide some guidance, as to how to correct for this overstatement, every product is unique, and the appropriate 'correction factor' is itself highly unreliable. Depending on the brand and treatment categories, the estimated product market shares are adjusted without employing a methodologically sound approach. For example, some pharmaceutical managers employ a rule of lowering these estimates by half and then by third to adjust for physician drug future prescribing overestimation (Howie & Kleczyk, 2011a).
In 2011, Howie and Kleczyk analyzed healthcare provider level data for 75 product uptakes after their market availability combined with respondents' pre-launch stated product uptake series. Based on their analysis, they developed a unique 'correction factor' for each service provider and each drug profile. Consequently, they determined the various levels of each drug's expected performance. The 'correction factor' is derived based on individual respondents' answers to questions of product use and thoughts about the product profile. In addition, questions regarding the speed of new product adoption, perceptions of the product over the current product treatments, level of knowledge about the new product, and intended use (in either first or second line of therapy) are also employed to adjust each provider's estimates. Their new approach of product potential estimation has been shown to be highly predictive of the actual future prescribing behavior of each healthcare provider. Their unique approach increases the forecast accuracy from R-square[6] of 0.233 to 0.796 (Howie & Kleczyk, 2011a), which can further help inform decision-making process when evaluating several chemical molecules for investment.

3.2.5 Summary of the consumer theory based risk assessment methods

As presented above, the Consumer Theory based framework is yet another way of managing risk when deciding which new product to develop. This approach can be based on healthcare providers' preferences (as well as patients, pharmacists, and other healthcare decision makers, depending on the study objective) and their perceived needs for new patient treatment options. Based on their potential product preferences and predicted 'future use' upon product availability, pharmaceutical managers are able to make informed investment choices of new chemical molecules. The output allows for analysis of future sales and revenues, and also for analyzing whether the proposed product meets the current market needs (product attribute evaluation). In addition, the improvements to the product market potential forecast, introduced by Howie and Kleczyk (2011a), allow pharmaceutical mangers to make even more informed decisions, regarding future product investment, due to increased reliability and precision of the data analysis.
The above approach can also be inputted, as expected product sales / revenues, into the NPV based approaches. The combined analysis effort increases managements' confidence in the potential product forecast results, as well as ensures that all aspects, related to product development, such as R&D costs, as well as clinical trials success and FDA approvals rates,

[6] R-square refers to the fraction of variance explained by a model (Champ et al., 2003).

are accounted for in the process of choosing the new chemical molecule. Consequently, the management ensures that investment in the selected product profile will return pharmaceutical company the financial capital expenses accrued during the drug research and development process.

3.3 Comparison of risk management methods

As discussed in the above sections, there are several risk management methods that differ from each other with regards to inputs, complexity, and the precision of product potential evaluation. As there is no one recommended approach for risk assessment in the pharmaceutical industry, it is important to understand how these methods differ and which situations should be used in. This section compares the NPV and Consumer Theory approaches based on the following criteria: 1) theoretical model; 2) inputs (i.e. R&D expenditures, future drug prices, and probability of FDA approvals, etc.); 3) analysis (i.e. mathematical simulations, econometric and statistical analyses); 4) complexity of the framework; 5) the recommended use in the product selection process. The presented analysis can assist pharmaceutical managers in deciding on the appropriate risk management method for their product assessment process.

As shown in Table 1 below, the Net Present Value of Income analysis is the simplest method of risk assessment in the pharmaceutical industry. It is based on the investigation of cash flows, requires fewest inputs, and is simple to compute, as well as to employ into the decision-making process. For example, this approach can be used for preliminary evaluation of the pharmaceutical products in development to indentify the potential candidates for investment, while requiring only limited information on future costs and revenues. In comparison to other methods, however, the NPV of Income analysis is the least reliable and precise in predicting the market potential of an evaluated product, and therefore recommending products for R&D (Grabowski & Vernon, 1998).

When more precise forecast is required, but a moderately complicated approach is preferred, Stochastic Dominance is usually employed. As shown in Table 1, this framework is still somewhat simple, but provides more reliable investment recommendations. The required inputs include cash flows and rate of return, as well as the FDA approvals rate, clinical trial success rate, and financial resource information (i.e. tax rate, operating costs, etc). These additional variables improve the precision of the analysis, and help better guide the decision-making process (Kleczyk, 2008).

When precision and reliability of the forecast are an issue, as well as capacity constraints (i.e. financial, workload, resource planning) are an important input into the analysis, the Capacity Constrained NPV approach is recommended. This method provides highly reliable and precise product return on investment prediction, due to accounting for the many variables important in the development, production, and sales of pharmaceutical products. As mentioned previously, however, the main problem, with the Capacity Constrained NPV, is the complexity level of the analysis, as it requires a high level of linear mathematical programming knowledge, as well as utilization of mathematical optimization and simulation. The method is recommended for use after the initial assessment of the potential products for investment is completed, in order to aid resource allocation and management during the product research and development process (Blau et al., 2000).

The last approach described in Table 1, the Conjoint Analysis method, can be also somewhat complex and time-consuming to employ, due to the multiple steps required to execute this

framework successfully (i.e. survey development, healthcare providers study recruitment, analysis, etc). Differently from the former methods, the Conjoint Analysis (CA) is based on the Consumer Theory instead of NPV, and analyzes end-users preferences for potential products. It is usually utilized by marketing mangers to help in forecasting product market potential (i.e. market share, revenue, sales), deciding order of product release to the market, as well as in the development of marketing strategies (i.e. market positioning, defining / confirming clinical end-point, etc). As it can be a reliable and precise forecasting tool, it is also employed as an input into the NPV based approaches to even further improve these methods' precision and reliability. Differently from the NPV based methods, the required inputs include product profile information and development of a survey instrument. To account for the financial aspect of the analysis, the expected drug prices and the formulary status of the product might be included. The mathematical analysis can become somewhat complex, and usually involves use of econometric and statistical tools, such as regression analysis, as well as mathematical optimization and simulation, to help identify the best product / product profile / attribute mix that will also result in the highest expected return on investment (Champ et al., 2003).

As the different methods of risk management vary with regards to the inputs, analysis, complexity of the framework, and the recommended use in the product selection process, pharmaceutical managers should consider the following criteria, when choosing the right approach for evaluation of potential pharmaceutical products: 1) the stage of the product development; 2) the preferred precision of the output; 3) the complicity of the model; and finally 4) the objective of the study. If quick and simple study of product assessment is needed, either the NPV of Income or Stochastic Dominance analysis should be an adequate tool to complete the task. If human and financial capital constraints are considered in the investment evaluation, then the Capacity Constrained NPV would be the preferred model. Finally, when product market share forecast still needs to be defined and / or preferred product attributes of a potential investment confirmed, the Consumer Theory based approach might be the best choice to pursue.

4. Concluding remarks

In this chapter, the research and development (R&D) process of new drugs, as well as methods of evaluating potential risks related to this procedure were discussed. As presented, new pharmaceutical products usually undergo costly and time-consuming testing, before receiving government approvals for distribution to patients. At the end, only about one percent of researched chemical molecules withstands the clinical trials, the scrutiny of the Food and Drug Agency (FDA), and becomes available for patient use. The costs associated with the R&D process has reached more then $800 million in recent years, and they are continually increasing. The average time length of product development is now 12 years, and will increase to even a longer time frame, due to the shift from the acute illness product development to chronic illness product development (Nelson, 2009). The shift is associated with a greater percent of elderly population, which is more prone to develop chronic diseases. Development of drugs that help either slow down or cure these types of diseases requires a longer time frame of clinical trials, as well as greater amounts of financial investments. Although these pharmaceutical products are more expensive and require more time to research and develop, they may return not only the invested financial capital, but also increase significantly net profits of pharmaceutical companies.

	Net Present Value of Income	**Capacity Constrained NPV**	**Stochastic Dominance**	**Conjoint Analysis**
Theoretical Framework	Net Present Value	Net Present Value	Net Present Value	Consumer Theory
Inputs	Expected R&D expenses; expected revenues; rate of return	Expected R&D expenses; expected revenues; rate of return; company financial information (i.e. taxes base; operating costs; financial resource constraints); FDA approval rate; clinical trial success rate; workload & planning constraints; currently manufactured product portfolio	Expected R&D expenses; expected revenues; rate of return; company financial information (i.e. taxes base; operating costs; financial resource constraints); FDA approval rate; clinical trial success rate;	Product attributes & clinical end-points; expected price point & expected formulary status *Survey instrument development
Analysis	Simple mathematical computation	Linear mathematical programming; mathematical optimization & simulation	Mathematical optimization & simulation; sensitivity analysis	Regression analysis (probit / logit); mathematical simulation & optimization
Complexity	Low	High	Moderate	Moderately High
Precision / Reliability	Low level of precision & reliability	High level of precision & reliability	Moderate level of precision & reliability	High level of precision & reliability
Recommended Use	Quick / preliminary product investment selection	Selection of highest potential product that meets all pre-specified criteria and constraints	Selection of highest potential product with financial, clinical trials, and FDA approvals constraints	Forecast of product sales / market potential; development of marketing / product positioning strategies; an input into the NPV based approaches (expected sales)

Table 1. Comparison of Risk Management Methods

The promotional efforts of pharmaceutical products also impact the risk management and future investment returns of the new drugs. With the continued high spending allocated to the advertising of these products (Direct-to-Consumer advertising and personal promotion to healthcare providers), as well as increased use of internet and digital media to inform healthcare providers and patient population of their treatment options, the expected sales and revenues can be increased even more in the near future. The increased financial revenues provide return on the past R&D investments, but also develop a stream of financial capital for future developmental projects.

The only, but quite a large, barrier in the entire product research and development process is the rate of FDA approvals and the managed care and / or formulary status of the new products. These agencies tend to slow down the speed at which products are brought to the market, as well as impact their affordability and access to the patient population. The FDA standards for product approvals are increasingly more stringent, while new drugs receive the lowest formulary status (a highest copay amount to be paid by patients) when launched to market for patient use, in comparison to more mature brands and generic competitors. To overcome the access and affordability issues, many pharmaceutical companies are assisting patients with their out-of-pocket costs to lower the financial burden, and to increase their product use. However, with an expected increase in the number of generic products on the market, even if this strategy might not benefit pharmaceutical companies in the long run (Alazraki, 2011).

In order to ensure recovering the high R&D costs invested by pharmaceutical companies, choosing the appropriate products for research and development requires important decisions about the tradeoffs between the available resources, as well as risk levels, returns, and time horizons for future payoffs. In theory, such tradeoffs are easily tackled by optimization problems; however, as discussed in this chapter, the complexity and uncertainty of the new drug development processes can make the solution hard to obtain, and might require employment of less complicated, and therefore, less precise methods of new product identification (Gino & Pisano, 2006).

Most risk management methods employed in the pharmaceutical industry include two types of methods: NPV and Consumer Theory based, to solve the new product research and development problem. As these method types differ on the basis of the analysis, inputs, precision and reliability of the approaches, as well as recommendations, knowing and understanding the differences between the various theoretical frameworks can help in selecting the right evaluation process of product selection and investment.

However, it is also important to know that each approach investigates a different angle of the risk management problem, and multiple analyses are usually recommended to ensure making informed investment decisions. Starting with a quick and simple NPV of Income analysis of potential product, and extending it to Stochastic Dominance, followed by the Capacity Constrained NPV approach for increased forecast precision, should help in predicting the success probability of brining the product to market, and the required resources for development and production. The Consumer Theory based methods can help further define the best product attributes, product positioning, and estimate the true product uptake when available for patient use. Knowing possible challenges, as well as benefits of the product of choice can help managers to avoid potential product failures, and recommend a product or set of products that will maximize investment returns for the pharmaceutical company in the future. In addition, it is recommended not to limit the risk

management analysis only to performing the computation internally, but rather to inquire for a third party / expert opinion. Having an outside expert, provide an unbiased opinion on the product investment options, can only strengthen the decision-making process, and help guide successful investment choice for the company.

In conclusion, the process of new pharmaceutical product research and development is complicated and requires a large financial and time investment. Since the financial and time costs are extensive, having the right tools in making investment decisions is vital in ensuring successful product selection. Currently available methods of risk management can help define the potential investment opportunities, and guide the selection process. These methods will grow to be even more important, as the R&D resources become in even more scares, and the product development costs increase.

5. Acknowledgment

The author would like to thank James R. Strout for providing comments, and editing earlier versions of this chapter. In addition, suggestions for changes, provided by the Book Review Board as well as InTech Editors, are also much appreciated. These editorial comments helped to shape this chapter into a well-written and thought-through manuscript.

6. References

Alazraki, M. 2011. The 10 Biggest-Selling Drugs That Are About to Lose Their Patent, *Daily Finance*, *An AOL Money & Finance Site*. Accessed: March 2011, Available Online from: www.aol.com.

Baker, T. 2002. A System Engineering Approach to Requirements Validation Product Development Risks Can Be Reduced by Validating Project Requirements Before the Design Process Begins, *Medical Device and Diagnostic Industry Magazine*. Accessed: October 2010, Available Online from: http://www.mddionline.com.

Blau, G., Metha, B., Bose, S., Pekny, J. F., Sinclair, G., Kuenker, K., and Bunch, P. R. 2000. Risk Management in the Development of New Products in Highly Regulated Industries, *Computers and Chemical Engineering*. Vol. 24, No. 9-10, pp. 659-664, ISSN 0098-1354.

Champ, P., Boyle, K., and Brown, T. 2003. *A Primer on Nonmarket Valuation*, Boston: Kluwer Academic Publishers, ISBN 1402-014457.

Census Bureau. 2008. Retail Prescription Drug Sales 1995-2006. Accessed: January 2011, Available Online from: www.cencus.gov .

Center for Disease Control. 2010. Births, Marriages, Divorces, and Deaths: Provisional Data for 2009, *NVSR Monthly Provisional Report*. Accessed: January 2010, Available Online from: http://www.cdc.gov/nchs.

Congressional Budget Office. 2006. Research and Development in the Pharmaceutical Industry, In: *A CBO Study*, Publication No. 2589. Accessed: December 2010, Available Online from: http://www.cbo.gov .

Dakin, H., Devlin, N., and Odeyemi. I. 2006. "Yes", "No" or "Yes, but"? Multinomial Modeling of NICE Decision-making. *Health Policy,* Vol. 77, No.3, pp. 352 – 367, ISSN 0168-8510.

DiMasi, J. A., Hansen, R. W., and Grabowski, H. G. 2003. The Price of Innovation: New Estimates of Drug Development Costs. *Journal of Health Economics,* Vol. 22, No. 2, pp. 151-185, ISSN 0167-6296.

Edwin, J.E. and Gruber, M.J. 1997. Modern Portfolio Theory, 1950 to Date. *Journal of Banking & Finance,* Vol. 21, pp. 1743-1759, ISSN 0378-4266.

Fortune 500. 2009. Fortune 500 Annual Ranking of America's Best Corporations. Accessed: January 2011, Available Online from: http://money.cnn.com/magazines/fortune/fortune500/2009/performers/industries/profits/ .

Food and Drug Agency. 2010. About FDA, In: *U.S. Department of Health and Human Services*. Accessed: January 2011, Available Online from: http://www.fda.gov/AboutFDA.

Carlson, G. 2008. Health: Quick Answers: What's the difference between a chronic illness and an acute illness? Continuing Medical Education, School of Medicine, University of Missouri-Columbia. Accessed: January 2011, Available Online from: http://missourifamilies.org/quick/healthqa/healthqa15.htm .

Gino, F. and Pisano, G.. 2006. Do Managers' Heuristics Affect R&D Performance Volatility? A Simulation Informed by the Pharmaceutical Industry, In: *Harvard Business School Division of Faculty Research and Development*. Accessed: October 2010, Available Online from: http://www.hbs.edu/research/pdf/05-015.pdf.

Grabowski, H. and Vernon, J. 2000. The Determinants of Pharmaceutical Research and Development Expenditures, *Journal of Evolutionary Economics,* Vol. 10, No. 1, pp. 201-215, ISSN 1432-1386.

Howie, P.J. and Kleczyk, E. J. 2011a. Accurately Predicting Product Market Potential, *Proceedings of the 4th Annual Advanced Pharma Resource Planning and Portfolio Management Conference,* Philadelphia, PA., February 28 – March 1, 2011.

Howie, P.J. and Kleczyk, E. J. 2011b. Alternative Technologies to Health Care Providers: What Digital Channel to Use, What Message to Deliver, and Who to Target?, *Proceedings of Pharmaceutical Market Research Group 2011 Annual National Conference,* Phoenix, AZ, March 27-29, 2011.

Honkomp, S. J. 1998. Solving Mathematical Programming Planning Models Subject to Stochastic Task Success, PhD Thesis, School of Chemical Engineering, Purdue Univ., West Lafayette.

Jain, V. and Grossmann, I. E. 1999. Resource-Constrained Scheduling of Tests in New Product Development, *Industrial Engineering Chemical Resources,* Vol. 38, pp. 3013-3036, ISSN 0888-5885.

Kaufman, M. 2005. Merck CEO Resigns as Drug Probe Continues. *The Washington Post*. Accessed: January 2011, Available Online from: http://www.washingtonpost.com .

Khan, M.Y. 1993. *Theory & Problems in Financial Management*, Boston: McGraw Hill Higher Education. ISBN 9780-0746-3683-1.

Kleczyk, E.J. 2008. Risk Management in the Development of New Products in the Pharmaceutical Industry. *African Journal of Business Management*, Vol. 2, No. 10, ISSN 1993-8233.

Lanjouw, J. O. and Cockburn, I. 2001. New Pills for Poor People?: Empirical Evidence. After GATT, *World Development*, Vol. 29, No. 2, pp. 265-89.

Lave, N., Parrott, H.P., Grimm, A., Fleury, M., Reddy, A. 2007. Challenges and Opportunities with Modeling and Simulation in Drug Discovery and Drug Development, *Informa Healthcare*, Vol. 10, pp. 1295 – 13010, ISSN 1744-7666.

Markowitz, H.M. 1952. Portfolio Selection, *The Journal of Finance,* Vol. 7, No. 1, pp. 77–91, ISSN 0022-1082.

Markowitz, H.M. 1959. *Portfolio Selection: Efficient Diversification of Investments*, New York: John Wiley & Sons. Reprinted by Yale University Press, 1970, ISBN 9780-3000-1372-6; 2nd ed. Basil Blackwell, 1991, ISBN 9781-5578-6108-5.

Moynihan, R. 2003. Who Pays For the Pizza? Redefining the Relationships between Doctors and Drug Companies, *British Medical Journal*, Vol. 326, No. 7400, pp. 1193–1196, ISSN 0959-8138.

Nelson, T.J. 2009. Why Do So Many Drugs Failed in Clinical Testing? *Science Notes*. Accessed: October 2010, Available Online from: http://brneurosci.org/drug-failures.html.

National Science Foundation. 2005. *Increase in U.S. Industrial R&D Expenditures Reported for 2003 Makes Up for Earlier Decline.* Accessed: January 2011, Available Online from: www.nsf.gov/statistics/infbrief/nsf06305/nsf06305.pdf.

Ogawa, S. and F.T. Piller. 2006. Reducing the Risks of New Product Development, *MIT Sloan Management Review,* Vol. 47, No. 2, pp. 65-71, ISSN 1532-9194.

Pirkl, J. 2009. The Demographics of Aging, In: *Transgenerational Design Matter.* Accessed: January 2011, Available Online from: http://www.transgenerational.org/aging/demographics.htm.

Piturro, M. 2006. When It Comes to Drugs Development, What Do Our Dollars Buy? *ENT Today.* Accessed: January 2011, Available Online from: http://www.enttoday.org/.

Pocock S.J. 2004. *Clinical Trials: A Practical Approach*, John Wiley & Sons, ISBN 0471-9015-55.

Rao, R. S., Kumar, G.C., Prakasham, R.S. and Hobbs, P.J. 2008. The Taguchi Methodology as a Statistical Tool for Biotechnological Applications: A Critical Appraisal, *Biotechnology Journal*. Vol. 3, No. 4, pp.510 – 523, ISSN 1860-7314.

Rao, V. R. 2007. Development in Conjoint Analysis, In B. Wierenga (Ed.), Marketing Decision Making. Working Paper. Cornell University. Johnson Graduate School of Management.

Rawlins, M.D. 2004. Cutting the Cost of Drug Development? *Perspectives,* Vol. 3, pp. 360-364, ISSN 0743-7021. Accessed: January 2011, Available Online from: www.nature.com/reviews/drugdisc .

Robinson, J. T. 2003. Changing the Face of Detailing by Motivating Physicians to See Pharmaceutical Sales Reps, *Health Banks.* Accessed: January 2010, Available Online from: http://www.healthbanks.com/PatientPortal/Public/support_documents/PMT_Robinson.pdf.

Rogers, S. H., Seager, T. P., and Gardner, K. H. 2004. Combining Expert Judgment and Stakeholder Values with PROMETHEE: A Case Study in Contaminated Sediments Management, In: *Comparative Risk Assessment and Environmental Decision Making, E*ditors: I. Linkov, AB. Ramadan, pp. 305–322, Boston (MA): Kluwer Academic, IBSN 1402-0189-59.

Schmidt, C.W. and Grossmann, I.E. 1996. Optimization of Industrial Scale Scheduling Problems in New Product Development, *Computers and Chemical Engineering*, Vol. 2, Sp. 1, pp. S1027-S1030, ISSN 0098-1354.

Shih, Y. C. T. and Sleath, B. L. 2004. Health Care Provider Knowledge of Drug Formulary Status in Ambulatory Care Settings, *American Journal of Health-System Pharmacy,* Vol. 61, No. 24, pp. 2657-2663, ISSN 1535-2900.

Shleifer, A. 2000. Inefficient Markets: An Introduction to Behavioral Finance, *Clarendon Lectures in Economics*.

Smit, H. T. and L. Trigeorgis. 2006. Strategic Planning: Valuing and Managing Portfolios of Real, *Options. R&D Management*, Vol. 36, No. 4, pp. 403 – 419, ISSN 1467-9310.

Tumulty, K. 2010. Making History: House Passes Health Care Reform, *Time Magazine,* Publisher: Time, Inc. Accessed: February 2011, Available Online from: http://www.time.com/time/politics/article .

Varian, H. R. 1992. Microeconomics Analysis, (3rd ed.), New York: W.W Norton and Company, IBSN 0393-9573-57.

13

Risk Management Plan and Pharmacovigilance System - Biopharmaceuticals: Biosimilars

Begoña Calvo and Leyre Zúñiga
Pharmaceutical Technology Department . Faculty of Pharmacy,
University of the Basque Country,
Spain

1. Introduction

The chapter addresses similar biological medicinal products (biosimilars) safety monitoring and describes the activities that should be developed in their risk minimisation plan. This is an issue that has aroused great interest with the recent expiration of biotech drugs patents and the advent of biosimilar products on the market.

2. Risk management

A medicinal product is authorised on the basis that in the specified indication(s), at the time of authorisation, the risk-benefit is judged positive for the target population. However, not all actual or potential risks will have been identified when an initial authorisation is sought. In addition, there may be subsets of patients for whom the risk is greater than that for the target population as a whole.

The management of a single risk can be considered as having four steps, *risk detection*, *risk assessment, risk minimisation and risk communication* which are summarized at table 1. However, a typical individual medicinal product will have multiple risks attached to it and individual risks will vary in terms of severity, and individual patient and public health impact. Therefore, the concept of risk management should also consider the combination of information on multiple risks with the aim of ensuring that the benefits exceed the risks by the greatest possible margin both for the individual patient and at the population level. Meanwhile Table 1 explains the management of a single risk, Figure 1 goes further and describes a complete risk management system, the so-called "Risk Management Plan" (EU-RMP) which contains two parts: *pharmacovigilance* and *risk minimization*. It covers how the safety of a product will be monitored and measured to reduce risk.

This chapter focuses on the activities that should be developed in the risk minimisation plan to be applied to biopharmaceuticals and more specifically to biosimilars (medicines similar but not identical to a biological medicine approved once patent lifetime for the original biotherapeutic has expired). Biopharmaceuticals often exhibit safety issues such as immunotoxicity that may lead to a loss of efficacy and/or to side effects (Giezen et al., 2009;

Stanulovic et al., 2011). The CHMP guidelines on biosimilars states that data from pre-authorisation clinical studies normally are insufficient to identify all potential differences with the reference product (Giezen et al., 2008). The main regulatory basis related to risk management are listed on Table 2.

DIFFERENT STEPS OF RISK MANAGEMENT		
RISK DETECTION AND ASSESSMENT	*Identify the risks*	Preclinical studies
		Harms identified in clinical trials & meta-analyses
		Formal mortality and morbidity studies
	Understand the risk	Rigorous case definition
		Case series analysis
		Clear description in label
	Monitor the risk	Post marketing surveillance
		Database analyses
		Prospective cohort studies and registries (to study potentially rare but important risks where risk identification or product attribution is difficult)
RISK MINIMISATION AND COMMUNICATION	*Communicate the risk*	Advice in label (not enough to communicate specific risk minimisation activities or change behaviours)
		Partnership with regulators
		Education of physicians, patients, company staff
	Act to reduce the risk	Limited distribution
		Limited prescribing rights
		Contra-indicate for certain groups, indications, routes of administration
		Advice for high risk groups
	Measure outcome of interventions	

Table 1. Risk Management steps

REGULATORY FOCUS ON RISK MANAGEMENT	
ICH E2E	Pharmacovigilance Planning (Nov 2004)
EMA	The Guideline on Risk Management Systems for Medicinal Products for Human Use (EMEA/CHMP/96268/2005). The Guideline has been included as chapter I.3 of Volume 9A. Annex C: Template for EU Risk Management Plan (EMEA/192632/2006)
GMP	ANNEX 20 Quality Risk Management (Feb 2008)

Table 2. Risk Management Legal Framework

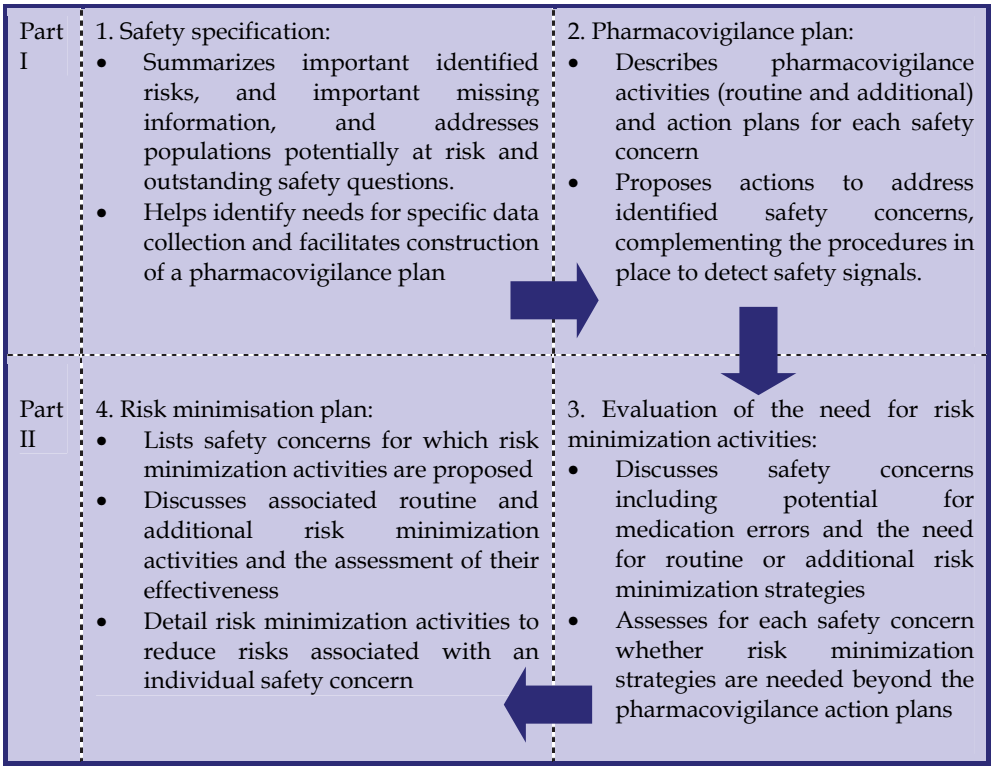


Fig. 1. Risk Management Plan development

2.1 Risk identification and safety specification

This is a summary of the important specified risks of a medicinal product, important potential risks, and important missing information. It also addresses the populations potentially at risk and outstanding safety questions, which warrant further investigation to refine understanding of the benefit-risk profile during the post-authorisation period. Table 3 explains the different considerations to take in mind when collecting safety data during the non-clinical and clinical development of a biosimilar medicinal drugs.

The safety issues identified in the safety specification should be based on the information related to the safety of the product included in the Common Technical Document (CTD), especially the overview of safety, benefits and risks conclusions and the summary of clinical safety (Zúñiga & Calvo, 2010a). The safety specification can be a stand-alone document, usually in conjunction with the pharmacovigilance plan, but elements can also be incorporated into the CTD.

Clinical safety of similar biological medicinal products must be monitored closely on an ongoing basis during the post-approval phase including continued risk-benefit assessment.

Even if the efficacy is shown to be comparable, the biosimilar product can exhibit a different safety profile in terms of nature, seriousness, or incidence of adverse reactions. Marketing Authorisation Holder (MAH) should provide safety data prior to marketing authorisation,

but also post-marketing as possible differences might become evident later, even though comparability with regard to efficacy has been shown. It is important to compare adverse reactions in terms of type, severity and frequency between biosimilar and reference medicinal product. Attention should be paid to immunogenicity and potential rare serious adverse events, focusing on patients with chronic treatments. The risk management plans for biosimilars should focus on:

- Heightened pharmacovigilance measures
- Conduct antibody testing
- Implement special post-marketing surveillance

For the marketing authorisation application a risk management program / pharmacovigilance plan is required. This includes a risk specification describing the possible safety issues caused by the differences (i.e. hostcells, manufacturing, purification, excipients etc.) of the biosimilar to the reference product.

ELEMENTS OF THE SAFETY SPECIFICATION	
Non-Clinical	
Non-clinical safety findings that have not been adequately addressed by clinical data	• Toxicity • General pharmacology • Drug interactions • Other toxicity-related information and data
	If the product is intended for use in special populations, consideration should be given to wether specific non-clinical data needs exist.
Clinical	
Limitations of the human safety database	• Discussion of the implications of the database limitations with respect to predicting the safety of the product in the marketplace • Reference to the populations likely to be exposed during the intended or expected use of the product in medical practice. • Discussion of the world-wide experience: - The extent of the world-wide exposure - Any new or different safety issues identified - Any regulatory actions related to safety • Detail the size of the study population using both numbers of patients and patient time exposed to the drug. This should be stratified by relevant population categories. • Detail the frequencies of adverse drug reactions detectable given the size of the database. • Detail suspected long-term adverse reactions when it is unlikely that exposure data is of sufficient duration and latency.
Populations not studied in the pre-authorisation phase	• Discussion of which populations have not been studied or have only been studied to a limited degree in the pre-authorisation phase and the implications of this with respect to predicting the

	safety of the product in the marketplace: - Children - The elderly - Pregnant or lactating women - Patients with relevant co-morbidity such as hepatic or renal disorders - Patients with disease severity different from that studied in clinical trials - Sub-populations carrying known and relevant genetic polymorphism - Patients of different racial and/or ethnic origins • Reference the relevance of inclusion and exclusion criteria in relation to the target population
Adverse events/ adverse reactions	The risk data should be presented according to the specific format described in section 3.6.2.c) of the Volume 9A The rules governing medicinal products in the EU (March 2007)
	• List the important identified and potential risks that require further characterization or evaluation (*identified* or *potential* risks)
	Identified Risks (an untoward occurrence for which there is adequate evidence of an association with the medicinal products of interest). • Include more detailed information on the most important identified adverse events/ adverse reactions (serios, frequent and/or with an impact on the balance of benefits and risks of the medicinal product). • Include evidence bearing on a casual relationship, severity, seriousness, frequency, reversibility and at-risk groups, if available. • Discussion of risk factors and potential mechanisms
	Potential risks (an untoward occurrence for which there is some basis for suspicion of an association with the medicinal product of interest but where this association has not been confirmed). • Description of important potential risks with the evidence that led to the conclusion that there was a such a type of risk
Identified and potential interactions including food-drug and drug-drug interactions	• Discussion of identified and potential pharmacokinetic and pharmacodynamic interactions • Summary of the evidence supporting the interaction and the possible mechanism • Discussion of the potential health risks posed for the different indications and in the different populations • Statement listing the interactions that require further investigation

Epidemiology	• Discussion of the epidemiology of the indications including incidence, prevalence, mortality and relevant co-morbidity (take into account stratification by age, sex and racial/ethnic origin) • Discussion of the epidemiology in the different regions with emphasis on Europe • Review the incidence rate of the important adverse events that require further investigation among patients in whom the medicinal product is indicated • Include information on risks factors for an adverse events
Pharmacological class effects	• Identify risks believed to be common to the pharmacological class (justified those risks common to the pharmacological class but not thought to be a safety concern)
Additional EU requirements	• Discussion of the following topics: - Potential for overdose - Potential for transmission of infectious agents - Potential for misuse for illegal purposes - Potential for off-label use - Potential for off-label paediatric use
Summary	
• Important identified risks • Important potential risks • Important missing information	

Table 3. Elements of the risk identification and safety specification (EMA, 2006)

2.2 Pharmacovigilance plan

The pharmacovigilance plan should be based on the safety specification and propose actions to address the safety concerns identified (relevant identified risks, potential risks and missing information). An action plan model can be found on Table 4. Only a proportion of risks are likely to be foreseeable and the pharmacovigilance plan will not replace but rather complement the procedures currently used to detect safety signals.

Safety concern	**Planned action (s)**
Important identified risks	<> List
Important potential risks	<> List
Important missing information	<> List

Table 4. Summary of safety concern and planned pharmacovigilance actions (EMA, 2006)

The plan can be discussed with regulators during product development, prior to approval of the new product or when safety concerns arise during the post-marketing period. It can be a stand-alone document but elements could also be incorporated into the CTD (table 5) (Zúñiga & Calvo, 2010b).

ROUTINE PHARMACOVIGILANCE	ADDITTIONAL PHARMACOVIGILANCE ACTIVITIES
• For medicinal products where no special concerns have arisen	• For medicinal products with important identified risks, important potential risks or important missing information • The activities will be different depending on the safety concern to be addressed

Table 5. Pharmacovigilance activities

The ***action plan*** for each safety concern should be presented and justified according to the following structure:

- Safety concern
- Objective of proposed actions
- Actions proposed
- Rationale for proposed actions
- Monitoring by the MAH for safety concern and proposed actions
- Milestones for evaluation and reporting

Protocols for any formal studies should be provided. Details of the monitoring for the safety concern in the clinical trial will include stopping rules, information on the drug safety monitoring board and when interim analyses will be carried out.

The outcome of the proposed actions will be the basis for the decision making process that needs to be explained in the EU-RMP.

CHMP biosimilars guidelines emphasise need for particular attention to pharmacovigilance, especially to detect rare but serious side effects.

Important issues include:

- Pharmacovigilance systems should differenciate between originator and biosimilar products (so that effects of biosimilars are not lost in background of reports on reference products).
- Ensure Traceability (importance of the international nonproprietary name, INN).

2.3 Evaluation of the need for risk minimisation activities

For each safety concern, the Applicant/Marketing Authorisation Holder should assess whether any risk minimisation activities are needed. Some safety concerns may be adequately addressed by the proposed actions in the Pharmacovigilance Plan, but for others the risk may be of a particular nature and seriousness that risk minimisation activities are needed. It is possible that the risk minimisation activities may be limited to ensuring that suitable warnings are included in the product information or by the careful use of labelling and packaging, i.e. routine risk minimisation activities. If an Applicant/Marketing Authorisation Holder is of the opinion that no additional risk minimisation activities beyond these are warranted, this should be discussed and, where appropriate, supporting evidence provided.

However, for some risks, routine risk minimisation activities will not be sufficient and additional risk minimisation activities will be necessary. If these are required, they should be described in the risk minimisation plan which should be included in Part II of the EU-RMP.

Within the evaluation of the need for risk minimisation activities, the Applicant/Marketing Authorisation Holder should also address the potential for medication errors (some examples are listed on Table 6) and state how this has been reduced in the final design of the pharmaceutical form, product information, packaging and, where appropriate, device.

POTENTIAL REASONS FOR MEDICATION ERRORS	
Naming	Taking into account the Guideline on the Acceptability of Invented Names for Human Medicinal Products Processed through the Centralised Procedure. CPMP/328/98 Rev 5, Dec 2007.
Presentation	Size, shape and colouring of the pharmaceutical form and packaging
Instructions for use	Regarding reconstitution, parenteral routes of administration, dose calculation
Labelling	

Table 6. Potential reasons for medication errors that the applicant needs to take into account

Applicants/Marketing Authorisation Holders should always consider the need for risk minimisation activities whenever the Safety Specification is updated in the light of new safety information on the medicinal product.

2.4 The risk minimization plan

The risk minimisation plan details the risk minimisation activities which will be taken to reduce the risks associated with an individual safety concern. When a risk minimisation plan is provided within an EU-RMP, the risk minimisation plan should include both routine and additional risk minimisation activities. A safety concern may have more than one risk minimisation activity attached to an objective.

The risk minimisation plan should list the safety concerns for which risk minimisation activities are proposed. The risk minimisation activities, i.e. both routine and additional, related to that safety concern should be discussed. In addition, for each proposed additional risk minimisation activity, a section should be included detailing how the effectiveness of it as a measure to reduce risk will be assessed. Table 7 shows how to approach the risk minimisation plan.

3. Postmarketing pharmacovigilance

MAHs should ensure that all information relevant to a medicinal product´s balance of benefits and risks is fully and promptly reported to the Competent Authorities; for centrally authorised products, data also should be reported to EMA. The MAH must have a qualified person responsible for pharmacovigilance available permanently and continuously.

3.1 Legal framework

The legal framework for pharmacovigilance of medicinal products for human use in the European Union (EU) is given in Regulation (EC) No 726/2004 and Directive 2001/83/EC

(Title IX) on the Community code relating to medicinal products for human use, as last amended by Directive 2004/24/EC and by Directive 2004/27/EC (EudraLex, 2007).

Safety concern	
Routine risk minimisation activities (i.e. product information, labelling and packaging)	<*short description of what will be put in the Summary of Product Characteristics (SPC), labelling etc to minimize risk e.g. warning in 4.4 (special warnings and precautions for use), that caution should be used in patients with cardiac failure, etc*>
Additional risk minimisation activity 1 (e.g. educational material or training programmes for prescribers, pharmacists and patients, restricted access programmes)	*Objective and rationale*
	Proposed actions
	Criteria to be used to verify the success of proposed risk minimisation activity
	Proposed review period
Additional risk minimisation activity 2, etc	*Objective and rationale*
	Proposed actions
	Criteria to be used to verify the success of proposed risk minimisation activity
	Proposed review period

Table 7. Information required for each important identified or potential risk for which additional risk minimisation measures are planned

For the biosimilar medicinal drugs approved in the Community through the centralised procedure, legal provisions are set forth in *Regulation (EC) No. 726/2004* (Title II, Chapter 3) (European Commission, 2004) and Commission *Regulation (EC) No. 540/95* (reporting of nonserious unexpected adverse reactions). The legal texts are supported by a series of guidelines, some of which have been compiled into Eudralex (Volume 9-Pharmacovigilance) (EudraLex, 2004). The requirements explained in these guidelines are based on the International Conference on Harmonisation (ICH) guidelines but may be further specified or contain additional request in line with Community legislation.
The obligations concerned with the monitoring of adverse reactions occurring in clinical trials do not fall within the scope of pharmacovigilance activities. The legal framework for such obligations is Directive 2001/20/EC. However, Part III of Volume 9A deals with

technical aspects relating to adverse reaction/event reporting for pre- and post-authorisation phases.
Pharmacovigilance activities are within the scope of quality, safety and efficacy criteria, because new information is accumulated on the normal use of medicinal products in the EU marketplace. Pharmacovigilance obligations apply to all authorised medicinal products, including those authorised before 1 January 1995 (Fruijtier, 2006), whatever procedure was used for their authorisation.
At approval there is limited clinical experience. Accurate pharmacovigilance and correct attribution of adverse events is vital.
Pharmacovigilance has been defined by the Worl Health Organization as the science and activities relating to the detection, assessment, unsderstanding and prevention of adverse effects or any other medicine-related problem (EudraLex, 2007).
The three main goals in Pharmacovigilance are:

- Protect the patients
- Protect the Pharmaceutical Company
- Comply with regulatory Requirements

3.2 Pharmacovigilance for centrally authorised productsreporting of adverse reactions and other safety-related information

Pre-Authorisation Phase

Once an application for a marketing authorisation is submitted to the Agency, in the pre-authorisation phase, information relevant to the risk-benefit evaluation may become available from the Applicant or Member States where the product is already in use on a compassionate basis, or from third countries where the product is already marketed. Since it is essential for this information to be included in the assessment carried out by the (Co-)Rapporteur(s) assessment teams, the Applicant is responsible for informing immediately the Agency and the (Co-) Rapporteur(s).
In the period between the CHMP reaching a final Opinion and the Commission Decision there need to be procedures in place to deal with information relevant to the risk-benefit balance of centrally authorised products, which were not known at the time of the Opinion. It is essential for this information to be sent to the Agency and (Co-)Rapporteur(s) so that it can be rapidly evaluated to an agreed timetable and considered by the Committee for Medicinal Products for Human Use (CHMP) to assess what impact, if any, it may have on the Opinion. The Opinion may need to be amended as a consequence.

Post-Authorisation Phase

Suspected adverse reactions related to centrally authorised products are reported directly by Healthcare Professionals, to each Member State. Marketing Authorisation Holders report serious suspected adverse reactions to the Member State in which the reactions occurred, within 15 calendar days of receipt. Each Member State is responsible for following up the Individual Case Safety Reports it receives to obtain further information as necessary.
The Member States should forward to the Agency serious suspected adverse reactions occurring within their territories.
The Agency and all Member States should receive directly from the Marketing Authorisation Holders suspected serious and unexpected adverse reactions that occur in a country outside of the EU.

The Agency should ensure that all relevant information about suspected serious unexpected adverse reactions from outside the EU are entered into the EudraVigilance database, and Member States should ensure that data on suspected serious adverse reactions occurring in their territory are uploaded into the EudraVigilance database.
Table 8 shows the main aspects to be considered relating biosimilar drugs safety during pre-autorisation and post-authorisation phase. The table highlights the additional reporting requirements for biosimilars when comparing to general safety reporting.

REPORTING OF ADVERSE REACTIONS AND OTHER SAFETY-RELATED INFORMATION		
	GENERAL REPORTING (Scharinger, 2007)	BIOSIMILARS REPORTING
PRE-AUTHORISATION PHASE	• All Suspected Unexpected Serious Adverse Reactions (SUSARs) • Sponsors to report to: - Concerned Member States (paper or electronically) - Concerned Ethics Committees (on paper) - EudraVigilance Trial Module (EVCTM) at the EMA (electronically) • Legal basis: Volume 10 of EudraLex- Clinical Trials guidelines	• Clinical safety data always required, even if efficacy is shown to be comparable • Sufficient number of patients to compare common Adverse Drug Reactions (ADRs) between referenced and claimed biosimilar products (type, severity, frequency) • Risk specification and pharmacovigilance plan part of the application dossier, as per EU legislation and guidelines • Pharmacovigilance systems/ procedures should be in place (traceability as per current EU guidelines)
POST-AUTHORISATION PHASE	• Adverse Drug Reactions/ Individual Case safety Reports (ICSRs) • Electronic reporting: - Mandatory e-reporting of ICSRs - Definition of exceptional circumstances that prevent electronic reporting (mechanical, program, electronic or communication failure) - Fall-back procedures to maintain expedited reporting compliance are established • Periodic Safety Update Reports (PSURs) from MAH to the Competent Authorities • Legal basis: Volume 9A of EudraLex- Pharmacovigilance	• Benefit-risk assessment on an ongoing basis. Importance of clinical experience with biologics: 2-3 years after market approval to adequately validate risk/benefit profile. • Risk management programme may be required if rare but serious adverse reactions.

Table 8. Biosimilars: pre and post-authorisation safety concerns

3.3 Monitoring of the safety profile

Signal Identification

It is likely that many potential signals will emerge in the early stages of marketing and it will be important for these to be effectively evaluated.
A signal of possible unexpected hazards or changes in severity, characteristics or frequency of expected adverse effects may be identified by:

- the Marketing Authorisation Holders;
- the Rapporteur;
- the Member States;
- the Agency in agreement with the Rapporteur

It is the responsibility of each Member State to identify signals from information arising in their territory. However, it will be important for the Rapporteur and the Agency to have the totality of information on serious adverse reactions occurring inside and outside the EU in order to have an overall view of the experience gathered with the concerned centrally authorised product.
As a matter of routine, the Rapporteur should continually evaluate the adverse reactions included in the EudraVigilance system and all other information relevant to risk-benefit balance in the context of information already available on the product, to determine the emerging adverse reactions profile. Additional information should be requested from the Marketing Authorisation Holder and Member States as necessary, in liaison with the Agency.
When a Member State other than the Rapporteur wishes to request information from the Marketing Authorisation Holder (apart from routine follow-up of cases occurring on their own territory) for the purposes of signal identification, the request should be made in agreement with the Rapporteur and the Agency.
Member States will inform the Rapporteur(s) and the Agency when performing class-reviews of safety issues which include centrally authorised products.
The Pharmacovigilance Working Party (PhVWP) should regularly review emerging safety issues which will be tracked through the *Drug Monitor* system.

Signal Evaluation

As signals of possible unexpected adverse reactions or changes in the severity, characteristics or frequency of expected adverse reactions may emerge from many different sources of data (see above), the relevant information needs to be brought together for effective evaluation, over a time scale appropriate to the importance and likely impact of the signal.
Irrespective of who identified the signal, a signal evaluation should be carried out by:

- the Rapporteur; or
- the Member State where a signal originated.

The Rapporteur should work closely with the identifier of the signal to evaluate the issue. Agreement needs to be reached in each case on the responsibility for the Assessment Report on the risk-benefit balance, by the Rapporteur or the Member State where the signal originated from, or jointly.
A Member State other than that of the Rapporteur should not start a full evaluation prior to having contacted the Agency and the Rapporteur, in order to prevent any unnecessary duplication of effort.
At request of the CHMP, the PhVWP evaluates signals arising from any source and keeps any potential safety issues under close monitoring.

Evaluation of Periodic Safety Update Reports

The Marketing Authorisation Holder is required to provide Periodic Safety Update Reports (PSURs) to all the Member States and the Agency. It is the responsibility of the Agency to ensure that the Marketing Authorisation Holder meets the deadlines.
The Marketing Authorisation Holder should submit any consequential variations simultaneously with the PSUR at the time of its submission, in order to prevent any unnecessary duplication of effort. Variations may, however, also be requested subsequently by the Rapporteur, after agreement by the CHMP.
It is the responsibility of the Rapporteur to evaluate and provide a report in accordance with the agreed timetable and to determine what issues if any need to be referred to the PhVWP and CHMP.
Actions required following the evaluation of a PSUR will be determined by the Rapporteur and the Marketing Authorisation Holder will be informed by the Agency, after agreement by the CHMP.
Where changes to the marketing authorisation are required, the CHMP will adopt an Opinion which will be forwarded to the European Commission for preparation of a Decision (Ebbers et al., 2010).

Evaluation of Post-Authorisation Studies, Worldwide Literature and Other Information

Final and interim reports of Marketing Authorisation Holder sponsored post-authorisation studies and any other studies, and other relevant information, may emerge from the Marketing Authorisation Holder, the Member States or other countries at times in between PSURs.
The Rapporteur should receive and assess any relevant information and provide an Assessment Report where necessary.
As above, the Rapporteur should determine what issues if any need to be referred to the PhVWP and CHMP.
The actions required following an evaluation will be determined by the Rapporteur and the Marketing Authorisation Holder will be informed by the Agency, after agreement by the CHMP.
Where changes to the marketing authorisation are required, the CHMP will adopt an Opinion which will be forwarded to the European Commission for preparation of a Decision.
The Marketing Authorisation Holder should submit any consequential variations simultaneously with the data, in order to prevent any unnecessary duplication of effort. Variations may, however, also be requested subsequently by the Rapporteur, after agreement by the CHMP.

Evaluation of Post-Authorisation Commitments

It is the responsibility of the Agency to ensure that the Marketing Authorisation Holder meets the deadlines for the fulfilment of specific obligations and follow-up measures, and that the information provided is available to the Rapporteur and the CHMP.
The Marketing Authorisation Holder should submit any consequential variations simultaneously with the requested information for the fulfilment of specific obligations/follow-up measures, in order to prevent any unnecessary duplication of effort. Variations may, however, also be requested subsequently by the Rapporteur, after agreement by the CHMP.
For marketing authorisations granted under exceptional circumstances, specific obligations will be set out in Annex II.C of the CHMP Opinion. Specific obligations should be reviewed by

the Rapporteur, at the interval indicated in the Marketing Authorisation and at the longest annually, and should be subsequently agreed by the CHMP. As above, the Rapporteur should determine what issues if any need to be referred to the PhVWP and CHMP.
For marketing authorisations granted under exceptional circumstances, the annual review will include a re-assessment of the risk-benefit balance. The annual review will in all cases lead to the adoption of an Opinion which will be forwarded to the European Commission for preparation of a Decision.
For all marketing authorisations (whether or not the authorisation is granted under exceptional circumstances) follow-up measures may be established, which are annexed to the CHMP Assessment Report. These will be reviewed by the Rapporteur, and will be considered by PhVWP and CHMP at the Rapporteur's request.
Where changes to the marketing authorisation are required, the CHMP will adopt an Opinion which will be forwarded to the European Commission for preparation of a Decision.
In the case of non-fulfilment of specific obligations or follow-up measures, the CHMP will have to consider the possibility of recommending a variation, suspension, or withdrawal of the marketing authorisation.
Table 9 shows the Omnitrope® Risk Management Plan Summary published by EMA.

Safety issue	**Proposed pharmacovigilance activities**	**Proposed risk minimisation activities**
Diabetogenic potential of rhGH therapy in short children born SGA	Phase IV prospective, single arm clinical trial in short children born SGA (part of registry reviewing patients' demographics, long term safety and immunogenicity).	Warning regarding diabetic potential in Section 4.4 of SPC*. Rare cases of type II diabetes mellitus in Section 4.8 of SPC.
Occurrence and clinical implications of anti-rhGH antibodies	Phase IV prospective, single arm clinical trial in short children born SGA measuring immunogenicity. Prolongation of ongoing Phase III study EP2K-02-PhIIIlyo to provide long-term immunogenicity data Immunogenicity testing for children enrolled in registry as appropriate (e.g. loss of efficacy).	Development of antibodies included in Section 4.8 of SPC.
Occurrence of malignancies in rhGH treated patients	Registry of patients reviewing patients' demographics, long term safety including malignancy and other safety issues.	Warning in Section 4.4 regarding reoccurrence of malignancy. Leukaemia mentioned as a very rare adverse effect in Section 4.8.
Risks of rhGH treatment in PWS patients	Registry expected to include patients with PWS and will record demographics, long term safety as well as other safety issues in this group.	Warnings on use of rhGH in PWS in Section 4.4. • Respiratory impairment and infection • Sleep apnoea • Severe obesity scoliosis

* SPC Summary of Product Characteristics
Table 9. Omnitrope® Risk Management summary (EMA, 2008)

3.4 Handling of safety concerns

Safety Concerns in the Pre-Authorisation Phase

Following the receipt of Individual Case Safety Reports or other information relevant to the risk-benefit balance of a product by the Agency and the (Co-)Rapporteur(s), the latter should assess these pharmacovigilance data. The outcome of the evaluation should be discussed at the CHMP for consideration in the Opinion.
If pharmacovigilance findings emerge following an Opinion but prior to the Decision, a revised
Opinion, if appropriate, should be immediately forwarded to the European Commission to be taken into account before preparation of a Decision.

Safety Concerns in the Post-Authorisation Phase

A Drug Monitor, including centrally authorised products, is in place as a tracking system for safety concerns and is reviewed on a regular basis by the PhVWP at its meetings. This summary document also records relevant actions that have emerged from PSURs, specific obligations, follow-up measures and safety variations.
Following the identification of a signal the relevant information needs to be brought together for effective evaluation, over a time scale appropriate to the importance and likely impact of the signal:

- Non-urgent safety concerns
- Urgent safety concerns

3.5 Information to healthcare professionals and the public

The management of the risks associated with the use of biosimilars demands close and effective collaboration between the key players in the field of pharmacovigilance. Sustained commitment to such collaboration is vital if the future challenges in pharmacovigilance are to be met. Those responsible must jointly anticipate, describe and respond to the continually increasing demands and expectations of the public, health administrators, policy officials, politicians and health professionals. However, there is little prospect of this happening in the absence of sound and comprehensive systems for biosimilars which make such collaboration possible. Understanding and tackling these are an essential prerequisite for future development of the biosimilars.
Healthcare Professionals (and the public if applicable) need to be informed consistently in all Member States about safety issues relevant to centrally authorised biosimilar, in addition to the information provided in Product Information. If there is such a requirement the Rapporteur or the Marketing Authorisation Holder in cooperation with the Rapporteur should propose the content of information for consideration by the PhVWP and subsequent discussion and adoption by the CHMP. The agreed information may be distributed in Member States. The text and timing for release of such information should be agreed by all parties prior to their despatch. The Marketing Authorisation Holder should notify, at his own initiative, the Agency at an early stage of any information he intends to make public, in order to facilitate consideration by the PhVWP and adoption by the CHMP as well as agreement about timing for release, in accordance with the degree of urgency. Marketing Authorisation Holders are reminded of their legal obligations under *Article 24(5) of Regulation (EC) No 726/2004* to not communicate information relating to pharmacovigilance concerns to the public without notification to the Competent Authorities/Agency (European Commission, 2004).

4. References

Ebbers, H.C.; Mantel-Teeuwisse, A,K.; Tabatabaei, F.A.S.; Moors, E. H. M.; Schellekens, H. & Leufkens, H. G. M. (2010). The contribution of Periodic Safety Reports (PSURs) to safety related regulatory actions of biopharmaceuticals. *Drug Safety*, Vol. 33, No. 10, pp. 919, ISSN 0114-5916

EudraLex [Homepage]. (2004). Date of access : February, 3, 2011. Available from: http://ec.europa.eu/enterprise/pharmaceuticals/eudralex/index.htm

European Commission. (2004). Regulation (EC) no 726/2004 of the European Parliament and of the Council of 31 March 2004 laying down Community procedures for the authorisation and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency . Official Journal No L 136, 30.04.2004; pp. 1 – 33

EMA, European Medicines Agency. (2006). EMEA/192632/06 . Template for EU Risk Management Plan (EU-RMP). Date of access : February, 14, 2011, Available from <http://www.ema.europa.eu/>

EMA, European Medicines Agency [Homepage]. (2008). EPARs for authorised medicinal products for human use. Date of access : February, 21, 2011, Available from: <http://www.ema.europa.eu>

EudraLex [Homepage]. (2007). Volume 9A Guidelines on Pharmacovigilance for Medicinal Products for Human Use. En: The rules governing medicinal products in the European Union. Date of access : February, 3, 2011. Available from: <http://ec.europa.eu/>

Fruijtier, A. (2006). Pharmaceutical Postmarketing and Compliance with the Marketing Authorisation. In: Fundamentals of EU Regulatory Affairs. Michor S, Rowland K. pp. 137-144. Ed. RAPS Regulatory Affairs Professionals Society. ISBN 0978700619 Rockville

Giezen,T.J. ; Mantel-Teeuwisse, A.K. ; Straus, Sabine M. J. ; Schellekens, H. ; Leufkens,H. G. & Egberts, A.C.G. (2008). Safety-related regulatory actions for biologicals approved in the United States and the European Union . *JAMA-Journal of the American Medical Association*, Vol. 300, No. 16, (October, 2008), pp. 1887-1896, ISSN 0098-7484

Giezen,T.J. ; Mantel-Teeuwisse, A.K. & Leufkens, H. G. (2009). Pharmacovigilance of biopharmaceuticals: challenges remain. *Drug Safety*, Vol. 32, No. 10, (January, 2009), pp. 811-817, ISSN 0114-5916

Scharinger, R. (2007). Pharmacovigilance: safety monitoring of medicines from an European perspective. AEMH (European Association of Senior Hospital Physicians) Conference; Viena. Date of access : February, 21, 2011, Available from: <www.aemh.org>

Stanulovic, V. ; Zelko R. & Kerpel-Fronius, S. (2011). Predictability of serious adverse reaction alerts for monoclonal antibodies. *International Journal of Clinical Pharmacology and Therapeutics*, Vol. 49, No. 3, (March, 2011), pp. 185-190, ISSN 0946-1965

Zúñiga, L & Calvo, B. (2010). Biosimilars approval process. *Regulatory Toxicology and Pharmacology*, Vol. 56, No. 3, (April, 2010), pp. 374-377, ISSN 0273-2300

Zúñiga, L & Calvo, B. (2010). Biosimilars: pharmacovigilance and risk management. *Pharmacoepidemiology and Drug Safety*, Vol. 19, No. 7, (July, 2010), pp. 661-669, ISSN 1099-1557