UiO : University of Oslo

Faculty of Medicine

Institute of Health and Society

Department of Health Management and Health Economics

Acknowledging Patient Heterogeneity in the Era of Personalized Medicine

A Comparison of Conceptual Frameworks used in Economic Evaluation

Author:
Pascale-Renée Cyr

Supervisor:
Eline Aas

Master Thesis

May 2016

Acknowledging Patient Heterogeneity in the Era of Personalized Medicine

A Comparison of Conceptual Frameworks used in Economic Evaluation

2016

Acknowledging Patient Heterogeneity in the Era of Personalized Medicine

Pascale-Renée Cyr

http://www.duo.uio.no/

Print: Reprosentralen, Universitetet i Oslo

Abstract

Personalized medicine is the notion that medical treatments can be adapted to individual patients based on a multitude of personal attributes. The set of personal characteristics that can together explain in part why patients respond differently to treatments is what we call patient heterogeneity. Economic evaluation traditionally uses a population-based approach; treatment recommendations and reimbursement decisions are based on the average outcome measured in an entire population sample. This can mask important sources of patient heterogeneity that could be used to improve decision-making. Instead, patients can be categorized in subgroups based on their personal characteristics and the cost-effectiveness analysis can be done exploring subgroup differences. However, in reality, this is rarely carried out. This is possibly because of the researchers' unfamiliarity with the methods and a lack of clear guidance in economic evaluation guidelines used by manufacturers and health technology assessment agencies. The guidelines published by the Norwegian agencies are vague and unclear on the topic of acknowledging patient heterogeneity and on how to conduct subgroup analyses in economic evaluation.

Therefore, with the intention to ultimately make recommendations to improve the guidance in Norway, this thesis set out two objectives: (1) to describe and compare existing methodology to acknowledge patient heterogeneity and (2) to apply the methodology to the results of an RCT. These two exercises were carried out to allow for the identification of both theoretical and practical strengths and weaknesses of the methods.

Three conceptual frameworks which in order are, Stratified Analysis (SA), Expected Value of Individualized Care (EVIC) and Value of Heterogeneity (VoH) were selected for the exercises. Thoroughly discussing their theoretical foundation pinpointed that even though all three methods are very similar, each present important advantages/disadvantages. Applying the three methods to the results of an RCT showed that there are also several practical differences that needed to be considered before conclusively suggesting a best course of action. Some unexpected technical problems occurred when using RCT results rather than modelling results. However, some solutions were formulated to address these issues. Most importantly, the last exercise made it possible to identify future research questions that builds on the frameworks' concepts and could lead to important practical improvements.

Ultimately, it was concluded that the use of either method alone is sub-optimal. Since the frameworks shared important similarities, it was possible to suggest an integrated approach that uses all three methodologies by playing to their strengths. This approach could serve as a rudimentary better course of action that may be recommended for HTA practices in Norway and from which to build on and improve with future research.

Acknowledgements

The realization of this thesis would not have been possible if not for the support of the remarkable people in my life. These few words are to acknowledge them and to express my sincerest gratitude.

To my supervisor, Professor Eline Aas, who has made herself available for countless hours, listened to me talk about my thesis' problems and offered me invaluable advice. I do not know if anyone could have handled me better than you. Your re-assurance and guidance not only lead to the accomplishment of my project, but also inspired me to want to continue working in the field we both share a passion for.

To my friends at the Faculty who have made this incredible educational journey a much more colourful one. I know I will continue to cherish your friendship beyond our University years and I sincerely hope that our careers will lead to future collaborations in our endeavour to improve health care systems.

To Karl Christian, my wonderful partner in life, who spent time helping me understand the value of computer programming and pushed me to learn beyond what I thought was possible on my own. Your love and understanding on a daily basis and through harder times has made this process much easier.

To Karl Christian's family and friends who along with him have ensured that my experience in Norway would go beyond simple academics. Encouraging me to strap skis on and to explore the beauty this country has to offer helped relieve much of the stress that comes with a student's life. Thank you for making Norway feel like home to me.

Finally, to my parents Jean and Anne, and my brother Maxime, who have encouraged me through my entire decade-long academic career. Whether I was in Ottawa or Oslo, you never let the distance dim the support you provided me. Thank you for always believing in me. No words could ever express my gratitude for having such an incredible family.

Pascale-Renée Cyr

Oslo, May 2016

Contents

List of Tables

List of Figures

List of Abbreviations

C-I	Cost-internalization
CEAC	Cost-effectiveness acceptability curve
CEA	Cost-effectiveness analysis
CUA	Cost-utility analysis
EBM	Evidence-based medicine
EVIC	Expected Value of Individualized Care
EVPI	Expected Value of Perfect Information
HRQoL	Health-related quality of life
HTA	Health Technology Assessment
ICER	Incremental cost-effectiveness ratio
IC	Individualized Care
LUC	Limited-use criteria
NMB	Net monetary benefit
PE	Pharmaco-economic
QALY	Quality-adjusted life year
RCT	Randomized clinical trial
SA	Stratified Analysis
VoH	Value of Heterogeneity
WTP	Willingness-to-pay

Chapter 1

Introduction

In the past few decades, headlines around the world have featured on multiple occasions the notion that we are entering in a new era of "personalized medicine". The hope is that with the abundance of evidence coming out of health research, we will be able to adapt medical treatments to each patient based on a multitude of personal characteristics.

Evidently, despite our shared humanity, the biological make-up of individuals exhibit slight differences when compared to others, but fascinatingly, also when compared to itself at various stages of life. Both genetic and environmental factors are responsible for the uniqueness of individuals and these differences are thought to explain, at least in part, why treatment responses vary between people. This "observable" variability is what has been termed patient heterogeneity (Briggs et al., 2006). While it is impossible, or at the very least impractical, to break down all differences between individuals to a molecular level, some characteristics are already known or can easily be revealed through a battery of diagnostic tests. These can serve to inform clinical decisions, particularly for the purpose of selecting the right treatment.

In this age of information, we are readily becoming more knowledgeable about the workings of our physiology and as such, more will be expected of the care provided through our medical systems. There is an entire new and promising field, pharmacogenomics, which is dedicated to elucidating how genes and pharmaceutical agents interact (Weinstein, 2000; Bala and Zarkin, 2004). It has now become possible to use genetic markers to predict how individuals are likely to respond to and metabolize drugs. This has the potential to allow physicians to "personalize" prescriptions where the most effective option is chosen and the right dosage is adjusted for the individual in question (Weinstein, 2000; Bala and Zarkin, 2004).

It naturally follows that policy makers are starting to ask questions about the impact personalized medicine will have on health care budgets. It is clear that personalized medicine has the potential to range from cost-effective to cost-creating (Davis et al., 2009). Two competing prospects are at play. Will the increase in doctor-patient interaction time and the extra medical tests put an additional burden on the system and raise costs? Or will it be that more successful treatments will decrease waste of medical resources and in turn save money?

It can be argued that a solution will be found in economic analyses. However, even though clinical research is increasingly reporting patient heterogeneity in treatment effects, economic evaluations of medical treatments have tended to neglect doing so (Ramaekers et al., 2013). Perhaps this is why Cohen et al. (2013, p.379) said that there were significant barriers to the successful implementation of pharmacogenomics and pointed directly at regulatory and reimbursement problems. Then again, as Bala and Zarkin (2004, p.496) have put it, "if the driving force behind pharmacogenomics is the concept that 'one drug does not fit all patients', the corresponding concept with pharmaco-economics could be 'one economic analysis does not fit all patients"'.

It is important to first understand that in most countries where health care is publicly funded, it is common place to require a formal health technology assessment (HTA) for new drugs or health devices before they are approved for the market. HTAs not only provide some evidence about clinical effectiveness, but they also serve as an economic evaluation for the purpose of informing decisions about reimbursement. Market approval is important for personalized medicines to reach patients, but as Cohen et al. (2013) explained, it is not sufficient. Reimbursement of therapies is almost equally important, especially in a public health system setting.

Economic evaluations of health measures conventionally use a population approach. The average effects and costs are measured for the entire population sample. If it is found that the average cost-effectiveness is higher for a new therapy when compared to an alternative one, the recommendation will be that the new therapy is offered to all patients. Undeniably, this is an important barrier for personalized medicine. As remarked by Grutters et al. (2013, p.112): "Average population-based economic evaluations [...] can mask important sources of patient heterogeneity within a patient population that may be important to improve population reimbursement decisions". For example, it is easy to imagine a situation where a treatment appears cost effective when looking at the average results, but when they are broken down into gender subgroups, the treatment is highly cost-effective for females and not at all for males. This means that the average population result is in reality attributed to a large effect measured in females. Failing to account for this could lead to a decision in favour of the treatment for everyone and subsequently, cause a waste of resources every time the treatment is given to males. The inverse situation can also be true; if the average population measure turns out not to be cost-effective, it can prevent a subgroup of patient from receiving an effective treatment and a loss of efficiency for society in health benefits forgone (Grutters et al., 2013).

If personalized medicine is to be implemented in a cost-effective manner, it is imperative that economic evaluations serving to inform clinical decisions and those about reim-

bursement consider the financial impacts of patient heterogeneity. It has already been demonstrated by previous researchers, such as by Barbieri et al. (2009), that considering heterogeneity can have a huge pay-off when subgroup specific policies exist. Using an example with medications for osteoporosis, they estimated that, if it was given to all patients diagnosed instead of only a few selected based on on age and risk factors, up to 15 000 000 British pounds could be lost Barbieri et al. (2009)

Unfortunately, the subgroup analyses are often neglected, particularly in economic evaluations. One reason that has been identified for this is pharmaco-economic guidelines do not offer proper guidance on the matter Ramaekers et al. (2013). The lack of consensus on how to analyse patient heterogeneity in economic evaluation may stem from the fact that the methodology to do so has only been developed in the last decade and has not been widely utilized yet.

Currently, in Norway, pharmaco-economic guidelines mention that patient heterogeneity is important to recognize in economic evaluation, however, no clear guidance is provided nor specific methodology is advised. This is slightly problematic because it will most likely be neglected by researchers and manufacturers. As explained by Barbieri et al. (2009), manufacturers have an incentive to neglect subgroup analyses that can show that their drug is not cost-effective for some patients because it could lead to an unfavourable decision and a loss of profits for them. Further, in the chance that they do explore subgroup differences, various methods can be used that may be unfamiliar to decision-makers and may confuse them. Updating the guidelines in Norway will eventually become a necessity if personalized medicine is to be implemented at all. A standardized approach for assessing subgroup differences would prevent that patient heterogeneity is neglected in HTA reports by manufacturers. It would also encourage researchers working for HTA agencies to routinely explore patient differences. Lastly, it would prevent confusing decision-makers with different kinds of results because there is no professional standard and no best course of action to take.

Therefore, this thesis aims to firstly (1) provide a clear description and comparison of the existing methodology to acknowledge patient heterogeneity, then secondly (2) to apply those methods to the same dataset in order to identify their strengths and weaknesses. These two questions are explored in the goal to assess whether a specific course of action can be recommended for HTA practices in Norway.

However, before tackling those two objectives, it is important to have a basic understanding of the purpose of HTA, how it is used and how it can evolve to meet the needs of our time and of personalized medicine. It is also crucial to understand the current method-

ology used in economic evaluations to see how it can be adapted to respond to the same needs. Therefore, Chapter 2 will discuss briefly how HTA came to be used in policy-making around the world and particularly in Norway. Then, the underlying principles of economic evaluation will be presented very shortly to situate the reader before discussing the issue of patient heterogeneity more specifically.

In Chapter 3, patient heterogeneity will be explained and the considerations for defining subgroups in economic evaluation will be discussed at length. Subsequently, in Chapter 4, the most recent developments in methodology to acknowledge patient heterogeneity will be presented. They consist of three conceptual frameworks that in order are, Stratified Analysis (SA) by Coyle et al. (2003), Expected Value of Individualized Care (EVIC) byBasu and Meltzer (2007) and Value of Heterogeneity (VoH) by Espinoza et al. (2014). After their theoretical foundations have been laid out, they will be briefly compared. This is an important exercise in the process of identifying similarities and differences and also some advantages or disadvantages that each method has.

In Chapter 5, all three frameworks have been applied to the same dataset which comes from the results of an RCT in Norway. This was done both to explore the feasibility of using the methodology on patient-level data and to expose the strengths and weaknesses of the three frameworks from a practical viewpoint. Results are presented and findings are discussed within the same chapter. The exercise allowed to detect important technical strengths and weaknesses of each method. It also steered the formulation of interesting research questions that should be explored in the future.

Finally, Chapter 6 concludes the thesis by synthesizing the essence of the findings into a course of action that could be recommended for HTA practices in Norway. The hope is that this work can contribute to opening the door that leads to a cost-effective implementation of personalized medicine in the future for Norway.

Chapter 2

Economic Evaluation in the Context of Health Care Decision-Making

2.1 From evidence-based medicine to health technology assessment

For as far back in time as we can look, medicine has always existed in one form or another. However, considering the long history of medical interventions, effective therapies were surprisingly rare until just recently (Banta, 2003). The health care systems that are experienced today by the citizens of most developed countries only began to organize around a century ago. It is after the Second World War that the idea of publicly funded systems with universal coverage and the goal of "health for all" arose (WHO, 2000). It did not take very long for countries that administered healthcare through tax money to realise that it was a costly endeavour that would necessitate years of reforms and reorganization (WHO, 2000). It is an ongoing and everlasting process. As we expand our medical knowledge and learn to better organize socially, our health systems continue to evolve and improve.

While healthcare systems proved to be dynamic entities, its participants also progressed along with them. With the modernization of our society, information became more accessible and with its abundance, healthcare users became more demanding and providers held to a higher standard of care. The expectation came to be that health workers be lifelong learners that are able to provide the latest and best known treatments available. It was the birth of modern and "evidence-based" medicine, from which eventually stemmed the more recent notion of "personalized" medicine.

Though it existed long before that, the term **"evidence-based medicine"** (EBM) was first used in the 1990s and is defined by researchers as "the conscientious and judicious use of current best evidence from clinical care research in the management of individual patients" (Sackett et al., 1996; Claridge and Fabian, 2005, p.547). In this day and age, medical data is more accessible, research has been multiplied and new solutions can be incorporated into clinical practice much faster than before (Claridge and Fabian, 2005). As such, the practice of EBM can be understood as "integrating individual clinical expertise with the best available external clinical evidence from systematic research" (Banta, 2003, p.124).

At the center of EBM is also another concept fundamental to medical research: **randomized clinical trials** (RCTs). Though the principles of RCTs were defined by Bradford Hill in the 1930s, it only became popular after the Medical Research Council's trial of streptomycin in the treatment of tuberculosis published in 1948 (Banta, 2003). One thing is clear: with the transition into the modern EBM era, the use of RCTs has dramatically increased (Claridge and Fabian, 2005). RCTs are recognized to be "vital [...] for assessing the effectiveness of treatments" and are now common practice in medical research (Claridge and Fabian, 2005, p.552). In most countries today, clinical research and RCTs are conducted in accordance with the high standards set out by the Cochrane Collaboration (Claridge and Fabian, 2005).

With high-quality evidence being produced around the world on the effectiveness of health interventions, a need to synthesise the information to assist clinical decisions emerged. For example, it would be highly impractical for a single physician to read all published results of RCTs and always having to decide by himself what is the best treatment course for a patient with a particular diagnosis. Physicians today agree that they have a duty to continue learning and to be up-to-date with new findings throughout their career. Some tools have been developed to facilitate these tasks for them. Specifically, national clinical guidelines, which suggest to physicians the recommended course of action in treating specific illnesses, are published yearly in most developed countries. However, healthcare managers need assistance in reaching a consensus and designing the right policies for the writing of clinical guidelines.

Consequently, alongside the EBM movement was conceptualised the **health technology assessment** (HTA). It first appeared in the United States when the Congressional Office of Technology Assessment was created in 1972. They saw HTAs as a form of policy research that examined the social consequences of new technology (Banta, 2003). In the US, the emphasis of HTA in the health sphere was placed on the *effectiveness* of treatments since the goal of health care services is, of course, to improve wellness and health (Banta, 2003). However, the birth of HTA would not make sense if not put in the context of the financial pressures experienced by some health care systems. For example, the increasing costs of health provision was thought to be in part due to the emergence of new technologies in the National Health System (NHS) in the UK (Banta, 2003). Cost-effectiveness analysis became an integral part of their HTAs. It helped managers and policy-makers base their decisions on effectiveness but also on *efficiency*, as a way to get better value for money. HTAs evolved predominantly in a cost-containment context in the UK and it largely influenced the way HTAs are conducted in Europe today. It is important to mention that Sweden also had a significant contribution to the development of HTAs.

Unlike in the UK, financial pressures were not felt as strongly there and they had an additional motivation for using HTAs which was to improve *quality* and *equity* (Banta, 2003). Surely, assessing how new technologies may have organizational, legal or ethical consequences can be of importance for decision-makers (Fure et al., 2013).

By the end of the 1990s, most EU member-countries had developed a national agency responsible for HTA. Today, HTA in the healthcare context has multiple functions. The International Network of Agencies for Health Technology Assessment describe it as: "the systematic evaluation of the properties and effects of a health technology, addressing the direct and intended effects of this technology, as well as its indirect and unintended consequences, and aimed mainly at informing decision making regarding health technologies" (International Network of Agencies for Health Technology Assessment (INAHTA), 2016, online).

While HTA has become mainstream in assisting decision-makers in our health care systems, there is always room for improvement. National HTA organizations should not strictly aim to improve the efficiency of health services, but they should also endeavour to improve their methods and the quality of the evidence they produce. This is particularly challenging as we are entering this era of personalized medicine. Because this thesis was motivated by a lack of clear guidance in HTA guidelines in Norway, it is important to also understand how HTA has come to be used in the Norwegian context.

2.2 The use of health technology assessment in Norway

In Norway, HTA-related activities only began to organize in the mid-1990s, much later than in other countries (Fure et al., 2013). A working group concluded that Norway needed a permanent solution to evaluate existing and future technologies in their effectiveness, risks and costs (Mørland, 2009). The Norwegian HTA Centre (SMM) was then established within SINTEF, an independent technological and social science research foundation in the Nordic countries. In its early days, SMM had a small staff of a dozen people and much of their focus was put on *clinical effectiveness* (Mørland, 2009). Research topics were regularly suggested by the ministry, hospitals or medical staff, but the SMM also chose its own topics based on several criteria such as the burden of disease, variations in clinical practice, economic consequences, and relevance for policy-makers (Mørland, 2009). In 2001, HTAs also began to assist decision-making by the Norwegian Medicines Agency (NoMA)(Festøy and Ognøy, 2015). It continues to serve today in the evaluation of new drugs for market authorization, reimbursement by the public system and deciding on information to provide prescribers and the public (The Norwegian Medicines Agency (Statens legemiddelverk), 2016).

Despite the presence of the SMM and regulations in the pharmaceutical sector, health care expenditure continued to rise dramatically in the 2000s. Aside from pharmaceutical products, the introduction of new health technologies in the healthcare system was not regulated in Norway and was perceived as the potential culprit (Mørland et al., 2010). Norway has only recently turned to HTA as a tool to evaluate *cost-effectiveness* rather than solely *clinical effectiveness* (Mørland, 2009).

The SMM permanently moved to the Norwegian Knowledge Centre for Health Services (NOKC) administered under the Directorate of Health in 2004. It continues to conduct independent HTAs and with a much bigger staff, it now provides increasing assistance to both the government and hospitals. In 2013 to the establishment of a System for the Introduction of New Technologies within the Specialised Health Services in Norway (Nasjonalt system for innføring av nye metoder i spesialisthelsetjenesten, 2015) was finally carried out. HTAs are now considered *essential* for the introduction of new health measures. HTA are now nationally conducted by the NOKC (or by manufacturers and evaluated by NoMA in the case of a new drug). Furthermore, mini-HTAs are also being conducted at the local level by Health Trusts (Fure et al., 2013).

What is important is how all of this relates to personalized medical decisions and ultimately the achievement of the cost-containment goal. Mini-HTAs inform decisions taken locally at the hospital-level about the introduction of a new treatment or device. Caregivers must abide by the decisions taken by the Health Trusts (Nasjonalt system for innføring av nye metoder i spesialisthelsetjenesten, 2015). As for the HTAs prepared with the national methodology by the NOKC, they serve the Directorate of Health who advises the government on health policies and also publishes the National Clinical Guidelines (Mørland et al., 2010). The Norwegian Clinical Guidelines are not legally binding, but rather normative by pointing to the desired and recommended courses of action (Helsedirektoratet, 2015). In practice however, it would be highly dubious if a physician had no medical or other valid reason to significantly deviate from those recommendations. This makes the clinical guidelines a powerful instrument to increase efficiency of the health care system by directly influencing clinical decisions. What is recommended should eventually become the clinical practice norm. As Cohen et al. (2013, p.387) reflected in light of diagnostic tests that could help to individualize care: "should medical professional societies incorporate evidence-based testing in their clinical practice guidelines, this may facilitate institutionalization of personalized medicine".

As for the pharmaceutical context, HTAs used by NoMA are not prepared by the agency, but instead by the manufacturers themselves. However, they must follow strictly the

agency's pharmaco-economic guidelines (see The Norwegian Medicines Agency (Statens legemiddelverk) (2012)). NoMA ultimately decides whether the drug is approved for the market, whether it will be reimbursed and what will be recommended for patients. This means that NoMA's recommendations also have the potential to significantly impact the efficiency of pharmaceutical care and reduce the national health expenditure. While discussing catalysts for personalized medicine in the U.S. context, Davis et al. (2009) realised that the adoption of individualized care methods can go much faster when the physicians incentives are aligned with approval and reimbursement decisions. Luckily, in Norway, unlike the Clinical Guidelines, physicians are required to prescribe the "first-choice" drug identified by NoMA and also the cheaper alternative (a generic for example), unless there is a medical reason for not doing so (Ringard et al., 2012).

In the last few years, a large emphasis has been placed on health economic issues and Norway is increasingly concerned with "the best use of resources in the health services" (Mørland, 2009, p.153). Today, as expressed by (Mørland et al., 2010, p.400): "... Norwegian HTA users have spread from the clinical micro level to include managers (meso level) ... the goal of evidence-based work has also been more prominent at the macro level of national policy making".

Because HTA is the key driver for improvements in the health system, it is critical that they are prepared with the best known analytical methods and that the results are of the highest quality possible. Therefore, if Norway is to be ready to enter the era of personalized medicine, ways to acknowledge patient heterogeneity should be reflected in their HTA practices so it can be achieved in a cost-effective manner. The Norwegian Health Directorate published a guide for conducting economic evaluation with the intent to create a common professional standard in 2013 (Helsedirektoratet, 2012). As for the pharmaco-economic guidelines, they date back to 2012. While this appears to have been done recently, it does not mean that they should not be re-visited to ensure they are adapted for the personalized medicine phenomenon that is progressively becoming a reality.

Since this thesis is particularly concerned with the economic evaluation methodology and standards applied in Norway, the next section provides a brief background on the current practices and recommendations in Norway.

2.3 Methodology currently used in Economic Evaluation

Though economic evaluations are an integral part of HTAs (Mathes et al., 2013), as explained by Drummond et al. (2015), they are most useful when they are preceded by

three other types of evaluations: efficacy, effectiveness and availability. However, an extensive discussion of these will not be carried out as it is outside the scope of this thesis. Moreover, economic evaluation is now the dominant feature of HTA helping high-level decision-makers in health care. The scarcity of resources available in our health systems makes the need for prioritizing a necessity and decisions need to be accompanied by a solid base of evidence supporting the goal to get the best value for money.

What is Economic Evaluation?

Economic evaluation has been defined as "the comparative analysis of alternative courses of action in terms of both their *costs* and *consequences*" (Drummond et al., 2015, p.9). "Alternative courses of action" should be interpreted in a wide sense, such as the many ways resources can be used to improve the outcomes of the health care system (Briggs et al., 2006). For example, the most obvious are drugs and medical devices, but one should also think about all other kinds of health interventions such as surgeries, screening and public health programs (Briggs et al., 2006). Because the term "health technology" is often interpreted in a restrictive sense, as a synonym for medical equipment such as an X-ray machine, Norwegian authorities decided that the term "method" ought to be used instead (Fure et al., 2013). They gave a broad definition to new "methods" (to improve health) that includes disease prevention, diagnostics, treatments, rehabilitation and even organizational models (Nasjonalt system for innføring av nye metoder i spesialisthelsetjenesten, 2015).

Though many more approaches to economic evaluation exists, three of the most commonly discussed in the literature are the **cost-benefit analysis** (CBA), **cost-effectiveness analysis** (CEA) and **cost-utility analysis** (CUA).

CBA's origins can be traced to welfare economic theory and the idea that health care programs should be evaluated as any other social programs (Briggs et al., 2006). The question becomes whether a new program or treatment represents a "Pareto improvement in social welfare", which essentially means that its social benefits outweighs the losses (Briggs et al., 2006, p.2). Correspondingly, programs are evaluated in their absolute benefits and a total net benefit that is positive determines that it is a worthwhile investment from a societal perspective, regardless of its relative performance (Drummond et al., 2015). In CBU, health effects are measured in monetary units. This is usually done by eliciting the maximum willingness-to-pay (WTP) or willingness-to-accept (WTA) to estimate the value a health intervention has for an individual. However, this has been criticized in the health care context because values assigned through WTP are often biased by complex factors interacting with the *ability* to pay of respondents (Donaldson, 1999).

An alternative to the standard welfare economic theory is that of the "extra-welfarist" which considers economic evaluation in a decision-making context (Drummond et al., 2015; Briggs et al., 2006). Generally, this approach aims to maximize the benefits derived from health interventions under a budget constraint. CEA and CUA are two methods that have been used to allocate resources under this principle (Briggs et al., 2006). The decision of whether a healthcare program is worthwhile depends on a willingness-to-pay ceiling often pre-determined by the available budget resources (Drummond et al., 2015). In a CEA analysis, the effects are measured in a single unit of effect (or natural units), for example, blood pressure changes, tumour size reduction, number of cases detected, etc. (Drummond et al., 2015). This can be useful when comparing alternatives that have outcomes measurable in similar units such as those within a specific field (Drummond et al., 2015). This method is mostly used for decision-making at a lower level, within hospitals. However, it becomes problematic when different types of healthcare programs need to be compared from a broader perspective at a higher level. Alternatively, in CUA, which can be classified as a type of CEA, the effects of health interventions are quantified in a generic "utility" unit, which is "the preferences individuals or society may have for a particular set of health outcomes" (Drummond et al., 2015, p.14). CUA is particularly useful for decision-makers when they need to allocate a budget between different programs that cannot be easily compared under classic CEA (i.e. deciding between implementing a new screening program or purchasing a new surgical device).

Standardization of Economic Evaluation

As explained by Briggs et al. (2006, p.1), the increasing use of economic evaluation "has placed some very clear requirements on researchers in terms of analytic methods". A large literature now exists on methods to conduct economic analysis and best practices have often been incorporated into guidelines at the national levels, such as pharmaco-economic (PE) guidelines. This is essential for ensuring that the processes are fair and transparent, but it is also to establish a methodology that is of high quality (Mathes et al., 2013). There is still no consensus on the one method that is optimal (Mathes et al., 2013). However, most HTA agencies around the world recommend CUA or CEA as a preferred method (Mathes et al., 2013). This is also the case in Norway where PE guidelines state (The Norwegian Medicines Agency (Statens legemiddelverk), 2012, p.15):

1. *Cost-per-QALY analysis (CUA) is the recommended method of analysis for cost effectiveness evaluations. An important reason for this is based on the benefits measure of such analyses, namely, quality-adjusted life years - QALY [...].*

2. *Cost Benefit Analysis (CBA) [...] is generally not recommended due to the ethical and technical challenges associated with setting a monetary value on health improvements [...].*

3. *Cost-effectiveness Analysis (CEA) is not recommended as the sole analysis method (see point above on CUA).*

Since we are concerned with the Norwegian context, the focus will be on CEA/CUA analysis. This is also most relevant to understand the topic of patient heterogeneity central to this thesis.

Measuring Costs and Effects

Who will make the decision is an important consideration to reflect upon before doing a cost analysis. This is, so that the right perspective is adopted and appropriate items to be included are selected accordingly (Drummond et al., 2015). Most of the time, a societal perspective is taken and it is the recommendation of the PE guidelines in Norway (Drummond et al., 2015; The Norwegian Medicines Agency (Statens legemiddelverk), 2012). However, sometimes, the decision is taken by hospital managers and it may be appropriate they consider the costs only as the portion of money coming out of their budget. This would be the case for mini-HTAs done locally in Norway.

Costs are usually measured by accounting for all the resources used, in what quantities and their unit prices (Drummond et al., 2015). Costs should be all-encompassing, including the medical equipment used, the physician's time, overhead costs, the patient's time and way more. Market prices can be used for most items while non-market items, such as time, are valued using a variety of different methods. Costs are usually adjusted and properly discounted over the relevant time-horizon.[1]

As for health benefits, they can be measured in natural units, as done in CEA. Although useful in certain contexts, it is a narrow take on the definition of "health". It restricts health benefits to physically measurable changes that are occurring in the body. In reality, health has a much broader definition: it is a "state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"(World Health Organization and others, 1950, online). For example, measuring the reduction in size of a tumour assesses a physiological health improvement, but it fails to account for the psychological and social benefits it has for the patient, which are also components of their health. This means that a utility metric such as the health-related quality of life (HRQoL) would offer a more comprehensive estimation of health benefits. As Espinoza (2012, p.18) remarked, "it has been argued that health-related quality of life (HRQoL) is in part due to the objective consequences of the disease [...] and in part to the social participation of individuals". This argument weighs is favour of a CUA over CEA for all levels of decision-making and is likely why Norwegian guidelines prefer the use of CUA

[1]For a thorough introductory discussion on cost-analysis, consult Chapter 4 in Drummond et al. (2015)

whenever possible.

The quality-adjusted life year (QALY) is probably the most common utility measure for health interventions. This is primarily because it takes into account two important dimensions of health: morbidity and mortality. The QALY encompasses into a single unit the improvements of both quality of life and length of life associated with the health interventions under evaluation (Drummond et al., 2015). Health-related quality of life (HRQoL) is estimated with multi-attribute utility-instruments that measures the physical, psychological and social state of the patient with weights that have been calculated from a prior valuation study where preferences for various health states have been measured in the general population (Ramaekers, 2013). QALYs are then derived by combining both the HRQoL and the time spent in that certain state of health. Even though there are several issues related to the methods used to estimate QALYs, an extensive discussion on the topic is outside the scope of this thesis. Researchers using QALYs should be aware of its shortcomings when interpreting results. However, as reasoned by The Norwegian Medicines Agency (Statens legemiddelverk) (2012), for a lack of a better alternative, QALYs are now widely used and accepted as a health benefit metric of choice in economic evaluation.

Decision analytic models

Effects and costs are sometimes both being measured simultaneously when an economic evaluations is done alongside an RCT. However, most often it is necessary to build a model to integrate costs and effects data that come from different sources (The Norwegian Medicines Agency (Statens legemiddelverk), 2012). For example, costs are usually determined with information coming from RCTs, guidelines, administrative data and even expert opinions (Mathes et al., 2013). On the other hand, effects data usually comes from RCTs on efficacy or meta-analysis of RCTs, or sometimes from observational studies (Mathes et al., 2013).

Once both are compiled, they are used in decision analytic models which "use mathematical relationships to define a series of possible consequences that would flow from a set of alternatives options being evaluated" (Briggs et al., 2006, p.6). Modelling uses probabilities of an event occurring to estimate the *expected* mean costs and effects of each alternative being compared Briggs et al. (2006).

Discrete event simulations, Markov models or decisions trees are normally used in economic evaluation. However, PE guidelines are often vague on the specific method or type of model to use, which is probably to leave some flexibility for researchers (Mathes et al., 2013). The Norwegian PE guidelines, state that they will accept different types of mod-

elling techniques as long as the choice is justified and the model properly validated (The Norwegian Medicines Agency (Statens legemiddelverk), 2012). The results synthesized from models serve to inform two important questions (Briggs et al., 2006, see p.165-166):

1. Does the evidence suggest that the new treatment is cost-effective compared to its alternative and should it be adopted considering current information?

2. Should more research be done to collect additional evidence to support the decision?

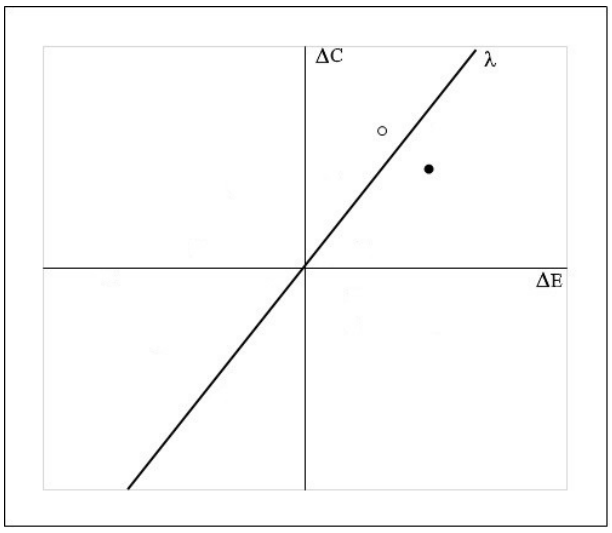
The answer to the first question usually depends on the differences in treatment effects and costs. In CUA, where a budget-ceiling is the basis for decision-making, a maximum willingness-to-pay threshold is used to determine if the new treatment should be adopted or not. Results are usually expressed as an incremental cost-effectiveness ratio (ICER) which is:

$$\text{ICER} = \frac{\Delta C}{\Delta E} \tag{2.1}$$

Where,
ΔC = the expected mean costs of the new treatment minus that of the alternative
ΔE = the expected mean effects of the new treatment minus that of the alternative

The willingness-to-pay (WTP) threshold (λ) is the maximum society will pay for one QALY. Therefore, a treatment is considered cost effective if the ICER $< \lambda$ (Drummond et al., 2015). It is also common to display results in a cost-effectiveness plane to help visualization of the results for decision-makers.

Figure 2.1: Cost-effectiveness plane where • = new treatment is cost-effective and ∘ = new treatment is not cost-effective.



Alternatively, it is also possible to use a net benefit approach, where a treatment is considered cost-effective when it has positive value (Briggs et al., 2006, see p.129)(see also Glick et al. (2014, chap. 7)).

$$\text{Net Monetary Benefit} = (\lambda \times \Delta E) - \Delta C \tag{2.2}$$

$$\text{Net Health Benefit} = \Delta E - (\frac{\Delta C}{\lambda}) \tag{2.3}$$

The second question enumerated above also needs to be answered as decision analysis is also "a systematic approach to decision-making under uncertainty" (Briggs et al., 2006, p.5). Characterizing the uncertainty associated with the decision at hand will help answer whether more research on the topic ought to be conducted.

Uncertainty surrounding decisions

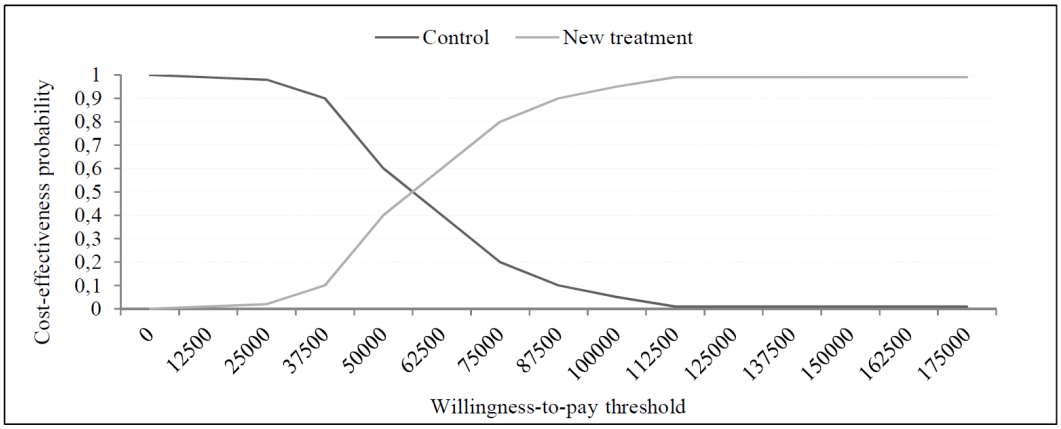
There are two main types of uncertainties related to the model itself: structural uncertainty and parameter uncertainty (Briggs et al., 2006). **Structural uncertainty** refers to the fact that in model building, it is necessary for very complex natural processes underlying treatment outcomes to be simplified. As discussed by Briggs et al. (2006), simplifying requires that assumptions be made and as such the results obtained will never reflect reality perfectly. Briggs et al. (2006) also suggest that a way of dealing with this uncertainty is doing "scenario analyses" which is measuring the expected outcomes under alternative model structures that make different assumptions.

Contrastingly, **parameter uncertainty** refers to the problem that the inputs used in the model, such as the probabilities of events occurring, costs, utilities and treatment effects are *estimations*. It cannot be known in advance how a patient will respond to a treatment nor how much resources will be used. The results of RCTs or other medical studies provide estimates that are imprecise because they usually come from sampled data (Briggs et al., 2006). Therefore, in principle, with the collection of more information, for example a larger sample, uncertainty can be reduced. Probabilistic sensitivity analysis are usually conducted to deal with the issue of parameter uncertainty. It propagates the uncertainty in the model by using probability distributions instead of fixed values as inputs. The ultimate goal is to obtain a measure of expected mean costs and effects with a confidence interval. When the cost and effects data comes directly from an RCT instead of a decision-model, a non-parametric approach called bootstrapping is commonly used to estimate the empirical distribution of mean costs and effects and from which confidence intervals can be estimated (Glick et al., 2014, p.107).

Using the results of sensitivity analyses, **decision uncertainty** is often presented with the

Cost-Effectiveness Acceptability Curve (CEAC). The CEAC essentially plots the probability that a treatment is cost-effective given the current results. Alternatively it can be interpreted at the error probability, which is "1 minus the value of the frontier" that is traced on the CEAC. (Briggs et al., 2006, p.168). This can be seen in Figure 2.2 where for example, at a WTP of 37 500, the control treatment is at a 90% probability of being cost-effective, or the control treatment has a 10% probability of being the wrong decision.

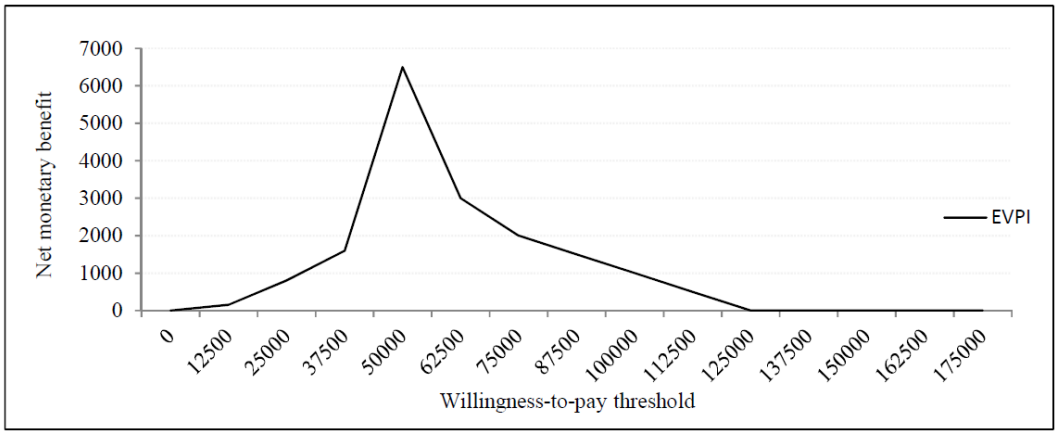
Figure 2.2: Example of a cost-effectiveness acceptability curve



Be that as it may, it has been argued that using confidence intervals with its "classical statistical inference and its Bayesian counterpart is arbitrary and irrelevant to clinical decision making" (Claxton, 1999, p.342). This is because when a new treatment appears cost-effective but is statistically insignificant, unnecessary costs are imposed to society in health benefits forgone if the decision is to not adopt it. This is surely because at least some individuals would have benefited from it. However, as explained by (Briggs et al., 2006, p.170): "this does not mean that adoption decisions can be simply based on little, or poor quality, evidence, as long as the decision to conduct further research to support adoption is made simultaneously". Therefore, the value of information (VoI) analyses are increasingly being used in economic evaluations. The calculation of the "expected value of perfect information" (EVPI) can be interpreted as the expected *cost* of uncertainty Briggs et al. (2006). The net monetary benefit (NMB) approach is used to compute the EVPI. It is derived from the results by considering cost-effectiveness (difference between the two alternatives), the uncertainty (the distribution of net monetary benefits) and the consequence of error given current information (Briggs et al., 2006, see p.170). The goal is to estimate in monetary terms the maximum value that society should be willing to pay to resolve the uncertainty in future research. Seeing the example in Figure 2.3, the EVPI tends to peak when the uncertainty is highest (for example, at a crossing point of two interventions when they are presented on the CEAC as in Figure 2.2). The EVPI has become more relevant than statistical inference simply because it assesses along with the decision to adopt the new technology, whether research to resolve the uncertainty around

the decision is worthwhile.

Figure 2.3: Example of the EVPI plotted over different WTP values



In Norway, the guidelines by the Helsedirektoratet (2012, p.24) mention that uncertainty can be explored with the help of the EVPI, but the The Norwegian Medicines Agency (Statens legemiddelverk) (2012, p.25) guidelines are slightly more comprehensive in their guidance. They specifically point the appropriate literature that should be consulted and expand by saying that the *EVPPI*, which relates to parameter-specific uncertainty, should also be calculated when appropriate.

Uncertainty vs. Variability

It is important to make the distinction between uncertainty and variability which appears to be a great source of confusion. Variability refers to the natural variation in treatment effects measured in individuals. Not every patient responds the same way and most of the time, the extent of the variability is not easy to predict ahead of time (Ramaekers, 2013). Variability can occur simply by chance, for example, an individual patient will not always measure the exact same effect when they receive the same treatment twice. Collecting additional information, such as increasing the sample size, *cannot reduce variability.*

However, variability can also occur because everyone is different in terms of their physiology and in terms of how they value their health. If differences in outcome occurred as a result of a personal characteristic that can be accounted for, such as age or gender for example, some of the variability observed in population results can be explained. Therefore, because personal attributes can explain in part variability and variability contributes to uncertainty surrounding a decision, then collecting additional information on those attributes *can reduce uncertainty.*

Chapter 3

Patient Heterogeneity in Economic Evaluation

Patient heterogeneity is at the center of this thesis and its analysis is the means by which personalized medicine comes to life. This chapter will focus on defining the concepts central to patient heterogeneity. This will serve as an important background before discussing at length the methodology that can be employed to acknowledge patient heterogeneity in economic evaluation.

3.1 What is patient heterogeneity?

Briggs et al. (2006, p.19) formally explained patient heterogeneity as "the extent to which it is possible to explain a proportion of the interpatient variability in a particular measurement on the basis of one or more patient characteristics". Particular measurements can refer to several things in economic evaluation, not just treatment effects. This will be elaborated further in the section on sources of patient heterogeneity.

A distinction needs to be made between "treatment heterogeneity", which is the differences in the types of treatment and the way they are administered, and "patient heterogeneity" (Ramaekers, 2013). Treatment heterogeneity is in a sense the result of research on patient heterogeneity. It integrates into clinical practice the findings of research on patient subgroups that showed a difference in optimal treatment. Treatment heterogeneity is in essence the promise of personalized medicine.

Clinical research already recognizes the importance of understanding that patient differences can be used to anticipate how they will respond to treatments (Ramaekers, 2013). Because of this, medical treatments are becoming more tailored to individual patients, hence more "personalized" (Grutters et al., 2013). However, the personalization of medical care can no longer be considered solely from a clinical angle. As previously mentioned, there are increasing concerns that the additional time and resources spent to provide individual solutions will be costly. Therefore, individualized treatment solutions should be economically evaluated the same way as other treatments.

Minimizing costs through "personalizing" treatment recommendations can be considerably valuable in light of the growing costs of health care services (Ramaekers, 2013). Fortunately, acknowledging patient heterogeneity in economic evaluation has a synergistic effect on achieving the goal of cost-containment. On the one side, by moving away

from the population-based approach and using personal attributes to classify patients in subgroups, optimal treatments can be identified at a higher precision level. On the other side, it ensures that personalized medicine is implemented in a cost-effective manner. Naturally, the gains earned from basing decisions on subgroups will benefit both the health care system, in terms of cost-savings, and the health care users, with improved care solutions designed for them.

3.2 Sources of patient heterogeneity

Because economic evaluation is not only concerned with effectiveness of treatments but also their costs, it is important to understand that the relevant personal characteristics are not strictly those that reflect differences explained by human physiology. Grutters et al. (2013) have identified three important categories of characteristics that are relevant in the realm of cost-effectiveness research and they are presented in Table 3.1.

Table 3.1: Categories of patient heterogeneity relevant in economic evaluation with a non-exhaustive list of examples for each categories. Information is taken from Grutters et al. (2013)

Sources of Patient Heterogeneity		
Demographics	Preferences	Clinical Characteristics
age	*attitudes*	*disease severity*
gender	*beliefs*	*disease history*
income	*risk tolerance*	*genetic profile*

Each of these patient characteristics are important sources of heterogeneity because they can influence the results of economic evaluation generally via four different important input parameters in decision models (Phelps, 1997; Ramaekers, 2013):

1. Treatment effects
2. Baseline risk
3. Health state utility
4. Resource utilization

Because the clinical sector already explores heterogeneity, it is not surprising that differences in **treatment effects** is the most recognized in published studies (Grutters et al., 2013). Differences in treatment effects are acknowledged with inputs in models by the use of relative risks, odds ratios or hazard ratios. For example, different probabilities for a treatment outcome will be used for females and males. Instinctively, one may think

it relates mostly to demographics or clinical characteristics. It is easy to imagine that a treatment given to a 20 year-old compared to a 50 year-old will have a different effect because the ageing process changes the body's physiology. Or, thinking of disease severity, a patient with a stage I cancer compared to one with a stage III cancer will likely respond differently to the same treatment. However, it may be less obvious but patient preferences can also affect the treatment effects. Preferences usually influence the effectiveness "via the impact they have on compliance or adherence with therapy" (Brazier et al., 2009, p.707). Naturally, poor compliance during a treatment will negatively impact its outcome (Brazier et al., 2009). It is also conceivable that beliefs affect the effectiveness of a treatment simply because of an added placebo effect.

As explained by Grutters et al. (2013), **baseline risk** relates more to absolute effects. The baseline risk refers to the risk of outcome for a patient under no treatment conditions (Wang et al., 2009). For example, it is possible that some untreated patients, compared to others, have a much greater risk of dying. Imagine that the risk is 1 in 10 cases for females compared to 1 in 20 males. Once treated, both males and females have reduced their probability of dying to 1 in 25 cases. Therefore, they exhibited no difference in treatment effects as they were both in the same position after having been treated. However, females exhibited a greater absolute risk reduction and seemingly responded much better to treatment than males did.

Health state utility will be influenced by all three categories mentioned in Table 3.1. Health state utility is usually assessed through questionnaires that measure health-related quality of life. For example, the EQ-5D measures 5 dimensions of health: mobility, self-care, usual activity, pain/discomfort and anxiety/depression (Drummond et al., 2015, see p.156). Logically, patients that experienced different treatment effects will likely answer differently. However, sometimes, despite experiencing the same treatment effect, they answer very differently because of their preferences. For instance, there is evidence that some patients value certain dimensions, such as mental health, more highly than physical health (Mukuria et al., 2006; Brazier et al., 2009). Or again, some patients may tolerate pain much better than others, which drives their health utility state value upward. Generally, preferences can affect results in two regards: differences measured between patients within a trial and differences measured between patients and the general population from which the tariffs are taken to derive HRQoL weights (Brazier et al., 2009). Health state utility is also affected by demographics. This is especially obvious when considering the age of patients. The time left to live can have a serious impact on QALY values when evaluating life-prolonging treatments.

Finally, **resource utilization** is related to differences in costs measured for each patients.

This is also influenced by all three categories mentioned in Table 3.1. For example, the same surgery performed on young patients may use less resources because they recover faster and are discharged sooner from the hospital than older patients. Or again, a patient may have a certain gene that makes a drug much more potent in his body compared to the next patient and by so requires smaller doses.

3.3 Selecting heterogeneity parameters from which to define patient subgroups

Selecting patient heterogeneity parameters in economic evaluation necessitates a wide approach because the outcomes can be affected by a greater set of factors compared to when only clinical effectiveness is considered. However, one advantage is that the economic evaluation is usually conducted after the effectiveness research and patient heterogeneity that was relevant clinically will be relevant for the economic evaluation. Regardless, careful consideration will always need to be given to the selection of heterogeneity parameters that are used to define subgroups that will be analysed in economic evaluation. Remembering the three main sources of heterogeneity listed in Table 3.1 above, it is important to choose parameters that are appropriate and relevant to the particular context of the treatments or technologies under evaluation.

Biological and economic plausibility

Sculpher (2008, p.803) cautioned for the importance of a biological plausibility when selecting parameters and explained:

> Uncertainty inevitably arises about whether particular difference between subgroups in cost effectiveness is genuine or simply reflects noise in available data. The chance of identifying spurious subgroups is increased if they are searched for without any clear scientific rationale (so called 'data dredging').

Therefore, it is important that parameters are not explored randomly *post-hoc*, but should instead be selected in advance based on their likelihood of reasonably influencing the treatment effects measured in patients. Sculpher (2008) further suggested that rules for subgroup analyses in the clinical sector already exist (i.e. Oxman and Guyatt (1992)) and should be respected. Clinical experts or previous research can also be of assistance when selecting the appropriate parameters.

A similar rationale can be used when selecting parameters that affect costs. Simple data-mining is equally risky in the context of economic variables.

Operationalizable in practice

Sculpher (2008) suggested that the parameters chosen to define subgroups should be able

to actually assist physicians in making treatment decisions in practice. It therefore requires that the characteristics be "easily observed or routinely measured" (Sculpher, 2008). The same idea was re-taken, though in a different context, and discussed more clearly in a later publication by van Gestel et al. (2012, see p.17), in which they distinguished three different types of parameters:

1. Patient-level attributes known when the treatment decision is made.
2. Patient-level attributes not known but measurable when the treatment decision is made.
3. Patient-level attributes revealed over time.

The attributes that are **known** refer to those easily observable aspects such as age and gender, but also information that is readily available in patients' medical records. These attributes should always be considered in analyses, especially because they require no extra cost in assisting clinical decisions.

On the other hand, attributes that are **unknown but measurable** are those that would require additional inquiry, either through a medical test (i.e. genetic test), or through administering a questionnaire that reveals patient preferences. This means that additional time has to be spent revealing the attributes and will manifestly increase resources used and costs. These parameters should also be explored in subgroup analyses, however their costs should properly be accounted for at the time of decision. If the costs of revealing the attribute is higher than the benefits gained from a subgroup policy, then it should not be implemented.

Finally, attributes that are **revealed over time** are those that can never be known at the time of treatment, but can only be measured retrospectively. van Gestel et al. (2012) use the example of side-effects of a medication, or in their case study, the progression rate of glaucoma. If such a parameter is selected for the analysis, it is usually strictly used for exploratory reasons. As explained by van Gestel et al. (2012), if it turns out that a lot of benefits can be gained by individualizing care based on that parameter, it may indicate that using a predictor for the parameter or research into developing a test that can reveal the parameter is worthwhile. While it has no potential in assisting clinical decisions at the time of diagnosis, the information can shed light on its potential for cost-savings in the future. Further, it may even be of great assistance for treatment continuation decisions (van Gestel et al., 2012).

Equity and ethical constraints

Some important ethical dilemmas or equity concerns may arise in the context of subgroup

analyses. Using socio-demographic attributes, such as age, gender and even more controversial race and income, as a basis for discrimination is often considered inequitable (Sculpher, 2008). For example, if a treatment is *effective* for both males and females, but is *cost-effective* only for females, a decision-maker might find it questionable to discriminate on the basis of financial argument alone (Ramaekers, 2013). On the other hand, in a different situation where gender or race explains a treatment effect difference because of a biological underlying factor such as a gene or something else, then it may be considered more acceptable (Sculpher, 2008; Grutters et al., 2013). Similarly, on the question of age, it is often considered unethical to use in subgroup analyses for life-prolonging treatments. This is because older patients will always inherently have a lower treatment effect when measured in QALYs, simply because they have less years left to live (Grutters et al., 2013). Once again, in a different context, age may be appropriate to use because the treatment is focused on improving the quality of life.

While ethical considerations are often used to discard some parameters, it may be best to proceed with them in the analysis regardless. For instance, the analysis conducted on a questionable attribute can be used to estimate how much monetary gains would be earned and if the ethical concerns should still outweigh those of efficiency. It is important to consider this because the choice to relinquish the monetary gains for the preservation of ethical integrity can impose greater costs on society in health benefits forgone (Grutters et al., 2013). Therefore, ethical constraints should not necessarily be used to discard parameters before an analysis, but instead should be used to provide a value that can help justify the policy decision.

3.4 Subgroup analyses in cost-effectiveness research

After the heterogeneity parameters from which to define subgroups have been selected, patient heterogeneity has to be analysed through economic evaluation methods that can establish whether differences in cost-effectiveness exist between the subgroups. It goes with this idea that patient heterogeneity can be used to identify the optimal strategy in each subgroup and subsequent clinical guidelines and reimbursement decisions can reflect the differences found (Koerkamp et al., 2010).

Different methodology to do so is the main concern of this thesis. Grutters et al. (2013) noticed that the most comprehensive methodology is provided in Briggs et al. (2006); they make use of regression methods to relate input parameters in a model to a patient characteristic. While the methods described in Briggs et al. (2006) are of assistance in building models which synthesize results for analysing subgroup differences, they do not "pronounce on the clinical relevance and relevance with regards to costs of a subgroup for

reimbursement decisions" (Grutters et al., 2013, p.117). For example, regressions can find that age is statistically significant in predicting the cost-effectiveness of a treatment, but it offers no information on which age group should receive what treatment. Further, regression techniques have some disadvantages in their assumptions (i.e. linearity) and even when these can be relaxed using different techniques, they still can lead to false-positive results or fail to detect an association. While it was mentioned previously that the risk of false-positive results can be reduced by specifying in advance the parameters to explore, the risk is not completely eliminated. Regression techniques used to estimate input parameters will not be discussed further in this thesis. The focus will remain on subgroups analysis methods that can be used to assist decision-making. For example, Briggs et al. (2006) recommends that results from modelling be presented for subgroups just as they are for population analysis. Specifically, they recommend that multiple cost-effectiveness acceptability curves be presented, one for each subgroup. However, it is unclear how to proceed exactly to get each subgroup results.

Three frameworks to acknowledge patient heterogeneity in economic evaluations, which are central to this thesis, have been developed recently. Unfortunately, they are rarely used by researchers. This is likely because of a lack of familiarity with the methods, but also because, they are not explicitly mentioned or referenced in HTA guidelines used by national authorities.

Failing to offer appropriate guidance with regards to patient heterogeneity analysis has been pinpointed as one of the major reasons why subgroup differences are under-reported in HTAs (Ramaekers et al., 2013). There is certainly room for improvement in the guidelines used in Norway. For example, the *Guide to Economic Evaluation of Health Measures* (Økonomisk evaluering av helsetiltak – en veileder) published by the Norwegian Health Directorate only mentions: "[Translated] Subgroup analyses may be helpful to document for which subgroups the new measure is most cost-effective" (Helsedirektoratet, 2012, p.24). However, no further details are provided on what approach to take. Similarly, in the pharmaco-economic guidelines published by the Norwegian Medicines Agency, they write: "The impact patient heterogeneity (i.e. differences in patient features) may have on the results should be examined through sub-group analyses, and not through uncertainty analyses" (The Norwegian Medicines Agency (Statens legemiddelverk), 2012, p.25).

Because the guidelines in Norway are so vague, the objective of this thesis is to explore the frameworks that have been developed and are starting to be used by researchers in North America and in the UK. The three will be described and applied to a Norwegian dataset in the hope that recommendations can be made for the improvement of HTA practices in Norway.

Chapter 4

Three Conceptual Frameworks to Acknowledge Patient Heterogeneity

In this Chapter, three conceptual frameworks to acknowledge patient heterogeneity in economic evaluation will be presented in the order of the year they were developed. The methodology will thoroughly be described and some examples are provided for clarification. Finally, the frameworks will be briefly compared and their advantages/disadvantages identified. This will allow for a better understanding when all three frameworks are applied to a dataset in the subsequent chapter. A slightly different notation was used rather than the one in the original papers to avoid confusing readers and facilitate comparison.

4.1 Stratified Analysis (SA) by Coyle et al. (2003)

One of the first group to develop a framework to assess the value of heterogeneity was Coyle et al. (2003), in Canada. Their framework, called "**Stratified Analysis**" (SA), uses a net benefit (NB)[1] approach to compare subgroups in their cost-effectiveness. Stratification, is essentially the process of dividing your population into strata (or subgroups) based on a selected number of heterogeneity parameters. For example in Table 1:

Table 4.1: Stratification based on three heterogeneity parameters

Parameter (θ)	**Stratification**	
Age	$i =$	1 for those aged between 50 to 65 2 for those aged between 65 to 80 3 for those aged 80+
Gender	$j =$	1 for females 2 for males
Genetic Marker	$k =$	1 if the marker is present 2 if the marker is absent

This example in Table 4.1 would give the possibility of a maximum total of 12 subgroups ($3 \times 2 \times 2$). The first step after stratification is to calculate the incremental net monetary benefit (iNMB) for each patients in all the strata.

$$\text{iNMB}_{ijk} = (\Delta e_{ijk} \times \lambda) - \Delta c_{ijk} \tag{4.1}$$

Where,

[1] Net monetary benefit (NMB) or net health benefit (NHB) can be used interchangeably

λ = willingness-to-pay per QALY threshold
Δe_{ijk} = incremental effects of a patient in the ijk^{th} strata
Δc_{ijk} = incremental costs for of a patient in the ijk^{th} strata
i = 1, 2, 3 and j = 1, 2 and k = 1, 2 (considering Table 4.1)

Then, if the stratification basis is the same as described in Table 1 for example, the mean NMB per patient in the ijk^{th} strata (one subgroup) is calculated as follow:

$$\overline{\text{iNMB}}_{ijk} = \frac{\sum \text{iNMB}_{ijk}}{n_{ijk}} \tag{4.2}$$

Where,
n_{ijk} = number of patients in the ijk^{th} strata from the candidate population

Using fictional numbers for demonstrative purposes, Table 4.2 effectively illustrates the results of these calculations.

Table 4.2: Example of Calculation of NMB for each strata

		Control		New Treatment				($\lambda = 500$)
Strata	Patients	Costs	Effects	Costs	Effects	ΔCosts	ΔEffects	NMB_{ijk}
111	n = 4	500	2	300	5	-200	3	1700
		550	2.1	200	4	-350	1.9	1300
		505	2.3	600	5.2	95	2.9	1355
		490	1.9	330	5	-160	3.1	1710
								$\overline{\text{NMB}}_{111}$ = 1516.25
112	n = 5	500	3	515	3.2	15	0.2	85
		480	4.9	680	4.8	200	-0.1	-250
		600	4.3	580	3.8	-20	-0.5	-230
		520	5.9	400	3.9	-120	-2	-880
		485	4	650	4.3	165	0.3	-15
								$\overline{\text{NMB}}_{112}$ = -258
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
ijk	n_{ijk}							
								$\overline{\text{NMB}}_{ijk}$ = ...

However, we are interested in the benefit of using heterogeneity for the purpose of identifying the optimal treatment for each subgroups (based on its cost-effectiveness). Coyle et al. (2003, p.421) termed this concept the "**limited use criteria**" (LUC) policy, which is "to restrict the public subsidy of a medicine to a sub-group of those patients for whom it is licensed with the aim of improved value for money". Coyle et al. (2003) discuss in their article that an efficient policy would be one that provides the new treatment alternative

to the subgroups who have an expected value of mean NMB that is positive. To find the potential gain from using a LUC, the total net benefit of giving the new treatment to only subgroups that show a positive $\overline{\text{NMB}}$ is calculated as follows:

$$\text{TNB}_S = \sum_i \sum_j \sum_k \text{iNMB}_{ijk} \quad \forall_{ijk} \quad where\ \overline{\text{iNMB}}_{ijk} > 0 \tag{4.3}$$

Then the net gain can easily be obtained by subtracting the TNB (sum of *all* incremental NMB) from the TNB_S obtained by using a LUC.

$$\Delta_S\text{TNB} = \text{TNB}_S - \text{TNB} \tag{4.4}$$

If we considered only the first 2 strata depicted in example from Table 4.2 then:

$$\text{TNB} = (1700 + 1300 + 1355 + 1710) + (85 - 250 - 230 - 880 - 15) = 4775$$
$$\text{TNB}_S = (1700 + 1300 + 1355 + 1710) = 6065$$
$$\Delta_S\text{TNB} = 6065 - 4775 = 1290$$
$$\text{NMB gained per patient} = 143{,}33$$

Along with the potential monetary gains from stratification, the percentage of patients that are sub-optimally treated can also be presented. It is obtained by counting the number of patients whom had a $\text{iNMB}_{ijk} < 0$ and dividing by the total number of patients in the strata. This percentage represents the opposite of the cost-effectiveness probability. If 15% of patients are sub-optimally treated, the treatment has an 85% probability of being cost-effective. This calculation is interesting in its own right for the same reasons it helps quantify the uncertainty around a decision-problem at the population level, it does so for individual subgroups. While Coyle et al. (2003) do not make mention of this specifically in their paper, the results of their framework would be easier to interpret when presented along with subgroup cost-effectiveness probability curves (CEACs). They could help decision-makers assess if more efficiency can be gained by further stratification. For example, imagine a scenario where stratification is based only on the parameter gender. Results show that despite a potential gain from a LUC to treat only females, 45 percent would still be treated sub-optimally. This suggests that gender alone is not a very efficient criteria to select the right treatment for females because there is a lot of uncertainty. Additional efficiency could still probably be gained with more information. Then imagine that a genetic marker is added to the stratification process and it shows that females in which it is absent have a positive $\overline{\text{iNMB}}$, whereas those in which it is present have a negative $\overline{\text{iNMB}}$. With a new LUC based on that genetic marker, let's say now, that only 15 percent of females are sub-optimally treated. If the efficiency gained through further stratification outweighs the costs of revealing that genetic marker through a test, then an LUC based on gender and a genetic marker is probably worthwhile from the decision-

makers point of view. If the number of patients sub-optimally treated only changes by little (ex. 1 percent) with further stratification, then a decision-maker might think that it is not worth it, especially if the gained monetary benefits are only marginal.

How much stratification yields the best efficiency gains is an empirical question all researchers face when considering the exploration of patient heterogeneity. Technically, as Coyle et al. (2003) discuss in their paper, the more information you have about a patient, the more efficiency you can gain, providing that the parameters explain the differences measured in effects or costs. Therefore:

$$\text{TNB}_{S(i)} \leq \text{TNB}_{S(ij)} \leq \text{TNB}_{S(ijk)} \leq \text{TNB}_{S(...)} \tag{4.5}$$

However, even though not discussed directly by Coyle et al. (2003) it should always be kept in mind that from a practical point of view, further stratification can make the data much heavier to process, take a much longer time and this for a decreasing efficiency gain. Also, if this method is applied to data from an RCT and not to a model, another important problem arises. The more you divide your original dataset into subgroups, the less predictive power it will have because you are decreasing your sample size quite significantly. Therefore identifying the heterogeneity parameters most likely to have an impact on costs and effect is very important.

Coyle et al. (2003) SA framework offers a particularly good way of analysing the potential loss from various policy when decisions are based on more than just cost-effectiveness. If for ethical reasons, it may not be desirable to use a LUC for a particular subgroup, for example based on age or gender, then it can be calculated how much could have been gained, and if for equity reasons it is justified to forego those benefits. Based on our example, to calculate the cost of an equity policy in which discrimination based on age is not allowed:

$$\Delta_E \text{TNB} = \text{TNB}_{S(ij)} - \text{TNB}_{S(j)} \quad \text{where } i = 1, 2, 3 \text{ and } j = 1, 2 \tag{4.6}$$

When making these calculations, one important assumption is made: the LUC is always followed perfectly for clinical decisions. This is the presumption that physicians would always act as perfect agents for the health system and respect the LUC strictly. However, in reality, physicians are primarily agents for their patients and may wish to exercise some judgement in deciding which treatment to give. For example, an LUC may state that a medicine is only recommended for those under 50 years of age, but given the health state of their 52 year-old patient, physicians may wish to prescribe the medicine anyway. This can have important consequences in the form of a NMB loss if in turns out that the physician made the wrong decision. Coyle et al. (2003) discuss a way to calculate what

they term "leakage" into other cohorts. Leakage can be defined as the associated NMB loss related to a number of patients that might receive the new treatment despite being in a subgroup where it was not optimal.

$$\begin{aligned} \mathrm{TNB}_{S(ijk)} \mid L = & \sum_i \sum_j \sum_k \mathrm{iNMB}_{ijk} \quad \forall_{ijk} \quad where\ \overline{\mathrm{iNMB}}_{ijk} > 0 \\ & + \sum_i \sum_j \sum_k l_{ijk} * \mathrm{iNMB}_{ijk} \quad \forall_{ijk} \quad where\ \overline{\mathrm{iNMB}}_{ijk} < 0 \end{aligned} \tag{4.7}$$

Then,

$$\Delta_L \mathrm{TNB} = \mathrm{TNB}_{S(ijk)} - \mathrm{TNB}_{S(ijk)} \mid L \tag{4.8}$$

Where, l_{ijk} is the probability distribution of receiving the new treatment when it has been identified to not be optimal.

Coyle et al. (2003) clarify that it would be necessary to estimate a probability distribution (l_{ijk}) before making the calculations. Because it might be difficult to estimate, they mention that an assumption can be made instead, such a 10% of the time for example. Coyle et al. (2003) make one further assumption for this calculation when they apply it in their case study, which is that leakage is restricted to neighbouring subgroups. Their reasoning was that non-adherence typically only occurs "for patients who just miss the cutpoint for therapy" (Coyle et al., 2003, p.423). While they do not explicitly explain why, intuitively it makes sense particularly in the context where a continuous parameter, such as age, is being considered. For example, the new treatment is cost-effective for those below 50 years of age and a patient that is 51 years old still received it. However, this idea of neighbouring subgroups seems less relevant if a dichotomous parameter, such as gender or a genetic marker is being considered.

Coyle et al. (2003) explain that leakage can be used to see if there is an alternative stratification basis that would return a higher net benefit. For example, given leakage, is it better to divide the population in strata of 10 years, or of 5 years? Ultimately, what is important is that the loss from leakage does not outweigh the net benefit gained from stratification. If the NMB loss is larger, there is no use at all for recommending a policy with a LUC.

4.2 Expected Value of Individualized Care (EVIC) by Basu and Meltzer (2007)

A few years after Coyle et al. (2003) developed the SA framework in Canada, a research group in the United States, Basu and Meltzer (2007), came out with a similar, yet slightly

different solution to characterize heterogeneity. They called their framework "**Expected Value of Individualized Care**" (EVIC). However, unlike Coyle et al. (2003) who seem to focus on decisions at the subgroup-level, Basu and Meltzer (2007) seem to be more concerned by decisions at the individual level.

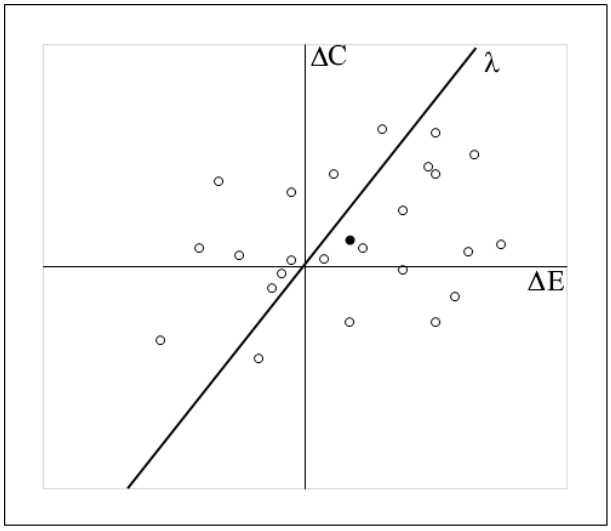
Though the original article described the EVIC framework using net health benefits (NHB), this thesis will use NMB to remain consistent throughout and to make the comparison between methods easier. NHB and NMB can both be used interchangeably without it affecting the conclusions that need to be ultimately drawn from the analysis (Basu and Meltzer, 2007; van Gestel et al., 2012).

The EVIC framework measures the additional gains that can be obtained through individualizing care by comparing it to the paternalistic approach routinely used in CEA until today. Here, a *paternalistic* approach means that the recommendation to treat does not consider the differences between individuals but is based instead on the average cost-effectiveness ratio (ICER) measured for the whole population. Considering Figure 4.1 below, we can see the average population ICER is showing that the new treatment under evaluation is favourable (cost-effective) compared to the alternative. Traditionally, a recommendation would be made that physicians treat <u>all</u> patients with this new treatment. This would lead to a lot of inefficient decisions. Basu and Meltzer (2007) make the distinction between two different types of inefficient decisions. Those taken under an individualized care model with cost-internalization and those taken under an individualized care model without cost-internalization.

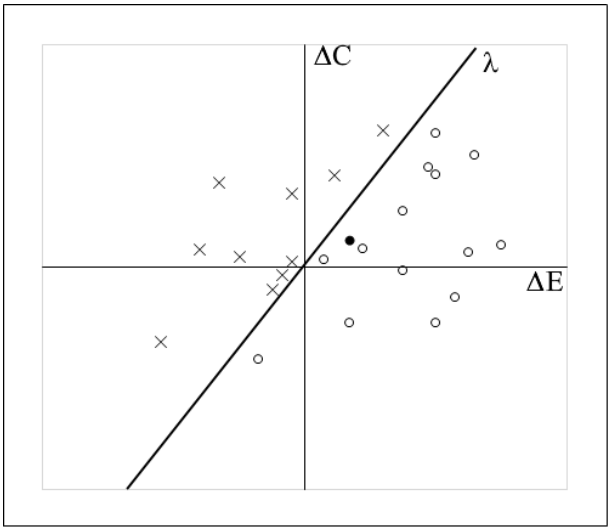
In a model **with cost-internalization** (see Figure 4.1(b)), the decision to treat patients is based on cost-effectiveness. All patients who fall on the left side of the WTP threshold are not treated. On the other hand, when there is **no cost-internalization** (see Figure 4.1(c)), the decision to treat is based on *effectiveness* alone. All the patients who see an increase in effects (those in the upper and lower right quadrants) would receive the treatment. (Basu and Meltzer, 2007) explain that a reason for why a policy-maker would be interested in measuring the gains under a model with no cost-internalization is because in a health system where patients are insured, they have little incentive to consider the costs of treatment and will seek mainly to maximize their health. This is not surprising given that the EVIC methodology was developed in the United States where the health system is not publicly funded and mainly operates on the basis of private health insurance.

Figure 4.1: Different scenarios of the EVIC model presented in a cost-effectiveness plane. λ = WTP threshold, ΔC = incremental costs, ΔE = incremental effects, $\bullet$ = average population ICER, $\circ$ = individual ICER of patients that get the treatment, $\times$ = individual ICER of patients that do not get the treatment

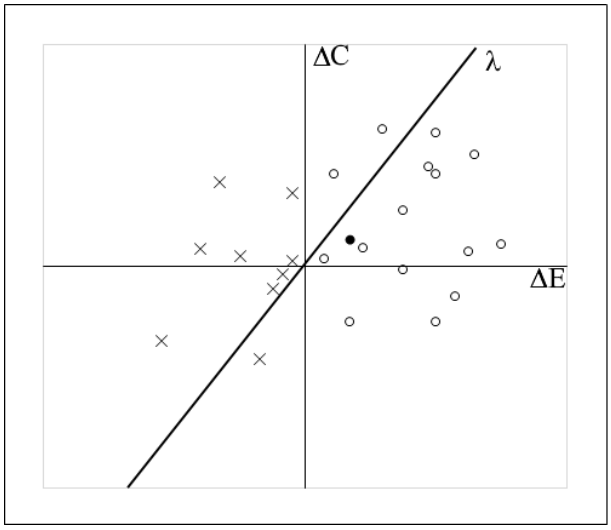
(a) Paternalistic



(b) Individualized care with cost internalization



(c) Individualized care with no cost internalization



In the EVIC framework it is assumed that patients are heterogeneous based on several identifiable parameters forming a vector ($\theta = \theta_i, \theta_j, ...$) that can explain some of the variability measured in costs or treatment effects. There is a joint distribution of θ in the population denoted as $p(\theta)$ (ex. probability of being a male or female and of having a gene or no gene, etc.). In the paternalistic model for decision-making, physicians do not consider individual values of θ and base their decisions on the distribution $p(\theta)$ found in the whole population. The societal value (V) in NMB is calculated as:

$$V_{paternalistic} = max_A \int_{\theta \in \Theta} \text{NMB}(\theta) p(\theta) d\theta \tag{4.9}$$

Where, A = the treatment alternatives under evaluation from which to select the optimal.

In an individualized care model, physicians do consider that individual patients are different based on their vector of parameters (θ) which explains variations in costs or effects. The physicians then base their treatment decisions on the value of those parameters, by choosing the alternative that maximizes the benefits for each patient. The societal value of the individualized care (IC) approach is therefore calculated as follows:

With cost-internalization:

$$V_{IC(\text{with C-I})} = \int_{\theta \in \Theta} max_A \ \text{NMB}(\theta) p(\theta) d\theta \tag{4.10}$$

The EVIC is obtained by the difference between the individualized care model and the paternalistic model:

$$\text{EVIC}_{(\text{with C-I})} = V_{IC(\text{with C-I})} - V_{paternalistic} \tag{4.11}$$

In a scenario with no cost-internalization, things are slightly different. In both the paternalistic and IC model, decision-makers seek to maximize the health benefits rather than the NMB. Therefore,

No cost-internalization:

$$V_{paternalistic(\text{no C-I})} = \left(max_A \int_{\theta \in \Theta} \text{E}(\theta) p(\theta) d\theta) \right) \times \lambda - \int_{\theta \in \Theta} \text{C}(\theta) p(\theta) d\theta \tag{4.12}$$

and

$$V_{IC(\text{no C-I})} = (\int_{\theta \in \Theta} max_A \ \big(\text{E}(\theta) p(\theta) d\theta)\big) \times \lambda - \text{C}(\theta) p(\theta) d\theta \tag{4.13}$$

Where,
E = effects measured for either treatment alternative under consideration that is optimal
C = costs measured for either treatment alternative under consideration that is optimal
λ = WTP threshold

Then EVIC is calculated as:

$$\text{EVIC}_{(\text{no C-I})} = V_{IC(\text{no C-I})} - V_{paternalistic(\text{no C-I})} \tag{4.14}$$

To visualize how these calculations may look like, Table 4.3 below shows an example using the same numbers as the example given for the stratified analyses in the previous section.

Table 4.3: Example of calculation of the EVIC

Control			New treatment				($\lambda = 500$)
Costs	Effects	NMB_C	Costs	Effects	NMB_T	$NMB_{IC(\text{with C-I})}$	$NMB_{IC(\text{no C-I})}$
500	2	500	300	5	2200	2200	2200
550	2.1	500	200	4	1800	1800	1800
505	2.3	645	600	5.2	2000	2000	2000
490	1.9	460	330	5	2170	2170	2170
500	3	1000	515	3.2	1085	1085	1085
480	4.9	1970	680	4.8	1720	1970	1970
600	4.3	1550	580	3.8	1320	1550	1550
520	5.9	2430	400	3.9	1550	2430	2430
485	4	1515	650	4.3	1500	1515	1500
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Mean		1074.44			1705	1857.78	1856.11

$EVIC_{IC} =$	1857.78	-1705	$= 152.78$
$EVIC_{noIC} =$	1856.11	-1705	$= 151.11$

While these calculations can give insight into how much gains can be earned through individualizing care, it does not postulate any specific parameters from the vector (θ) on which the physicians' decisions should be based on. Instead it considers the entire vector of parameters that can explain variations between patients. This means that the calculation of the EVIC can only tell us the value of the efficiency that could *potentially* be gained. Therefore, the EVIC has no practical use to help guide the physician's decision. That is why Basu and Meltzer (2007) take the analysis further by introducing the concept of the **parameter-specific EVIC**.

The parameter-specific EVIC quantifies which portion of the total EVIC as calculated above can be explained by a particular heterogeneity parameter. The goal is to identify the optimal treatment for each individual patient based on the value of the parameter selected (θ_i). The parameter-specific EVIC is calculated as the difference between the total EVIC in the population approach and the EVIC that is calculated by individualizing the treatment based on the attribute of interest. This is shown in the following equation:

$$\text{EVIC}_{\theta_i} = \text{EVIC} - \int_{x \in \theta_i} p_i(x)\text{EVIC}(\theta_c \mid \theta_i = x)d\theta_i \tag{4.15}$$

Where,
θ_i = parameter of interest among the vector of parameters θ
$p_i(x)$ = is the marginal probability distribution of the parameter of interest
θ_c = all remaining parameters

Another research group, van Gestel et al. (2012), clarified this concept in a later publication. They described the parameter-specific EVIC calculation as "a series of simulations consisting of inner loops and outer loops... [where] each inner loop, a cohort of heterogeneous patients is simulated with a fixed value for a [parameter], θ_i, ... [and] all other attributes, θ_c, are randomly drawn for each individual patients" (van Gestel et al., 2012, p.16). They further explained that for "each outer loop, a new value for θ_i is drawn from the $p_i(x)$"(van Gestel et al., 2012, p.16).

This calculation resembles more closely the SA method described in the previous section, though it follows different steps. Describing the second term of equation 4.15 in simpler words, it takes into account the actual identity of a parameter (for example θ_j = gender) to divide the population into subgroups (stratification). The average NMB of both the control and new treatment is calculated for each strata to identify the optimal treatment for that subgroup. This is the subgroup "paternalistic" solution. Then the individualized care value is also calculated for each strata. Then finally, both the paternalistic and IC values are averaged over the whole population according to the probability distribution of that parameter. This process may be best understood by looking at the example in the Table 4.4 below.

Table 4.4: Example of the calculation of a parameter-specific EVIC

Parameter		Control	New treatment	($\lambda = 500$)
$\theta_j = gender$	Patients	NMB_C	NMB_T	NMB_{IC}
$\theta_j = 1$ (female)	n = 4	500	2200	2200
		500	1800	1800
		645	2000	2000
		490	2170	2170
	Mean	533.75	2042.5	2042.5
$\theta_j = 2$ (male)	n = 5	1000	1085	1085
		1970	1720	1970
		1550	1320	1550
		2430	1550	2430
		1515	1500	1515
	Mean	1693	1435	1710
Mean NMB_{pat} =	$\lceil(4/9) \times 2042.5\rceil$	$+\lceil(5/9) \times 1693\rceil$	= 907.78 + 940.56	= 1848.34
Mean NMB_{IC} =	$\lceil(4/9) \times 2042.5\rceil$	$+\lceil(5/9) \times 1710\rceil$	= 907.78 + 950	= 1857.78
EVIC(gender) =	1857.78 − 1848.34	= 9.44		
$EVIC_{\theta_j}$ =	152.78 - 9.44	= 143.34		

Basu and Meltzer (2007) explain that calculating the parameter-specific EVIC is useful for decision-makers as it can rank the parameters in order of magnitude for the added

value they bring to decision-making. Surprisingly, they do not mention the SA methodology developed by Coyle et al. (2003), but it is clear that the parameter-specific EVIC is in essence the same as calculating the gains of stratification described earlier.

However, what is important to retain is that even when the population EVIC is very high, the value of parameter-specific EVIC that identifies the right treatment is usually much lower. These two dimensions of individualizing care were largely the concern of Espinoza et al. (2014) and his colleagues. Their framework builds on to these two methodologies and is described in the next section.

4.3 Value of Heterogeneity (VoH) by Espinoza et al. (2014)

The latest developments in patient heterogeneity analyses methodology came from a research group in the United Kingdom. Espinoza et al. (2014) recognize that both the SA and EVIC methods sought to understand the value of heterogeneity, but neither framework seemed to fully address how variability due to heterogeneity and uncertainty interact. They also concerned themselves with developing a way to address another important problem: the appropriate level of stratification.

Espinoza et al. (2014) decided to build on the existing frameworks by adding two components. Firstly, they introduced the idea of an efficiency frontier for subgroup analysis. Secondly, they divided the value of heterogeneity into 2 dimensions: (1) the expected health gain because of stratification and (2) the additional value of further research on subgroup-related uncertainty.

Drawing an Efficiency Frontier

The efficiency frontier for subgroup analysis is a concept developed to select the appropriate level of stratification. In RCTs, subgroup analysis is mainly concerned with variations in clinical effectiveness, whereas in CEA subgroup analysis a wider view is taken and other things such as patient preferences and costs can also be considered. In clinical research, heterogeneity parameters are likely selected solely on their biological plausibility. Espinoza et al. (2014) also recognized that it was previously recommended in CEA that parameters should be operationalizable in practice. However, Espinoza et al. (2014) went further by suggesting an additional consideration for defining subgroups, which is the criteria of *efficiency*.

Keeping in mind the SA method described earlier, we know that, with current information (known heterogeneity parameters), the total expected NMB can be estimated for subgroups stratified on that basis. Different combinations and specification of param-

eters (f) can result in a varying number of subgroups (S). The most efficient level of stratification can therefore be obtained by resolving a maximization problem:

$$max_{A,f} E_\theta \text{TNB}_{S,f}(A, \theta) \quad S = 1, 2, ...n; f = 1, 2, ..., F. \tag{4.16}$$

Where,
S = number of subgroups
f = subgroup specification (criteria for stratification)
A = optimal treatment alternative
E_θ = expected value given a vector of parameters (θ)

As Coyle et al. (2003) had explained, the more stratification there is, the closer you get to having perfect information on the individual and the more likely the appropriate decision for treatment can be assessed. However, as Espinoza et al. (2014) remarked, the marginal efficiency gain gets smaller and smaller the more you stratify. Therefore, depicting the efficiency frontier on a graph can help visualize the combinations of parameters and number of subgroups that yields an acceptable societal benefit for the decision-maker.

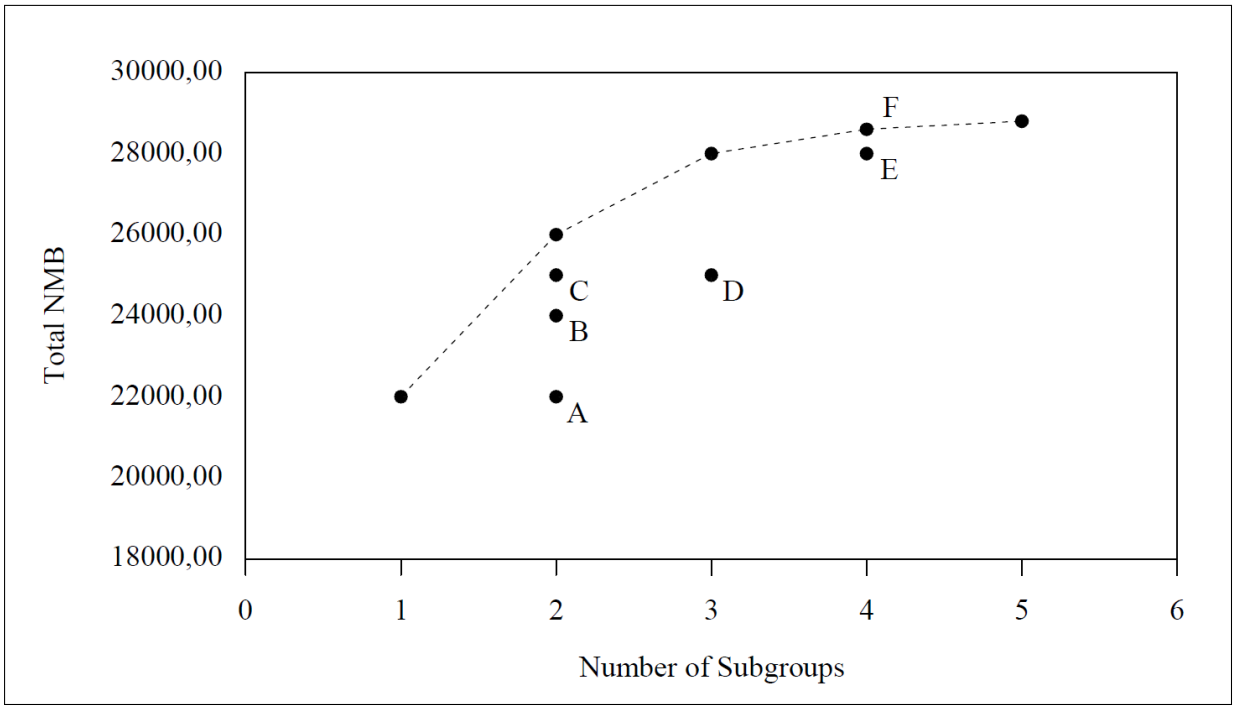


Figure 4.2: Example of an efficiency frontier for a patient heterogeneity analysis. Each points on the graph represent the total NMB gained from stratification based on different parameters and subgroup specifications. Letter markings have been placed as a reference for the discussion of scenarios provided in the main text.

As depicted in Figure 4.2 above, the total NMB will vary depending on (1) the specification of the parameter for stratification (2) the total number of subgroups considered. The efficiency frontier is formed by the specifications that yields the highest total NMB

for each number of subgroups.

Let's imagine that A is a scenario where the two subgroups are divided based on age, (1) aged under 50 and (2) aged over 51. We can see that when compared to the whole population analysis, there are no gains in total NMB. However, this time let's imagine that D represent a scenario where age is also the basis for stratification, but this time the groups are (1) aged 25 to 50, (2) aged 51 to 75 and (3) aged 76 and above. With the same parameter but a different specification, a gain in total NMB could be measured. This can occur when a difference in cost-effectiveness is measured in a small subgroup of patients that would have otherwise been masked if they had been part of a larger subgroup. Therefore, a *proper specification using a single parameter* is crucial.

Parameters chosen to specify the subgroups can also yield different total NMB. For example A, B and C are all scenarios where there are two subgroups, and each measure different total NMB. The scenario in A could be 2 subgroups based on age, B based on gender and C based on a genetic marker. If a decision-maker wanted to write an LUC based on only two patient categories, the genetic marker would return the best value for money. Additionally, it is also possible to use a combination of parameters. This time, we could imagine that scenario E uses both gender and a genetic marker to stratify the population. The gains are higher when parameters are used together than when they are considered alone. This means that *increasing the number of subgroups* by considering more parameters can also yield better results.

The two dimensions of the Value of Heterogeneity (VoH)

The second goal of Espinoza et al. (2014) was to better explain the heterogeneity-related uncertainty. In a CEA, as we discussed earlier, uncertainty arises because of how the model itself is built and because of its parameters. Heterogeneity analysis is used to ascertain which portion of the variability observed between patients can be explained by their characteristics. However, uncertainty can never be fully resolved. There is still a possibility of making a wrong decision and additional evidence is always valuable to inform decisions. The value of information (VoI) can always be quantified by estimating the value of making decisions once all uncertainty is resolved. This concept, termed Expected Value of Perfect Information (EVPI), was discussed earlier and is routinely computed by experts. Perfect information would be a perfect representation of reality rather than just an estimation through a sample size with only a few known parameters.

As Espinoza et al. (2014) explain, the EVIC framework by Basu and Meltzer (2007) is analogous to that of EVPI. If the true value of the vector of heterogeneity parameters θ was known, that is all possible personal characteristics that makes individuals unique,

then the right treatment decision could be made every time. However, the true value of θ is not known, so we estimate the expected value by averaging the maximum NMB of all patients over the joint distribution of θ. That is, as we did in the EVIC framework, calculating a paternalistic approach. Then the approach selecting the alternative that yields the highest benefit for every single patient within each of the subgroups can be compared. Analogous to equation 4.10 above Espinoza et al. (2014) wrote:

$$E_\theta max_A \text{NB}(A, \theta) \tag{4.17}$$

Basu and Meltzer (2007) calculated the EVIC by removing from this the value derived from a paternalistic model where the average NMB for the total population dictates which treatment all patients will receive. It is in the opinion of Espinoza et al. (2014) that there are no distinction between that process and that of calculating EVPI. Therefore they wrote:

$$\text{EVPI} = E_\theta max_A \text{NB}(A, \theta) - max_A E_\theta \text{NB}(A, \theta) \tag{4.18}$$

However, under current information, we do have the value of some parameters that are part of the vector. This means that the optimal treatment for subgroups divided on the basis of those parameters can be identified and the NMB calculated. Espinoza et al. (2014) suggested that the EVPI for each subgroups should be calculated a follows:

$$\text{EVPI}_S = E_\theta max_A \text{NB}_S(A, \theta) - max_A E_\theta \text{NB}_S(A, \theta) \tag{4.19}$$

This is somewhat similar to the second term of the equation 4.15 above. The distinction is that the EVPI is not calculated for the whole population, but rather for one subgroup at a time. Espinoza et al. (2014) explained that the value of subgroup EVPI is useful as it can be considered the maximum investment for further research in resolving the uncertainty left within that particular subgroup. However, it may be of interest to have the value for the whole population as well and it is obtained as such:

$$\text{EVPI}_{(S)} = \sum_{s=1}^{S} \text{EVPI}_S w_S \tag{4.20}$$

Where, w_S is the weight indicating the proportion of the total population represented by the subgroup. The value obtained will be identical as the one obtained from the second term of the parameter-specific EVIC in equation 4.15. See the example in table 4.5 below.

Table 4.5: Example of the calculation of the population EVPI under current information

Parameter		Control	New treatment	($\lambda = 500$)
θ_j	Patients	NMB_C	NMB_T	NMB_{IC}
$\theta_j = 1$ (female)	n = 4	500	2200	2200
		500	1800	1800
		645	2000	2000
		490	2170	2170
	Mean =	533.75	2042.5	2042.5
	$EVPI_S$ =	2042.5 - 2042.5 = 0		
$\theta_j = 2$ (male)	n = 5	1000	1085	1085
		1970	1720	1970
		1550	1320	1550
		2430	1550	2430
		1515	1500	1515
	Mean =	1693	1435	1710
	$EVPI_S$ =	1710 - 1693 = 17		

$EVPI_{(S)}$ =	$\Big(\big[0 \times (4/9)\big]$	$+ \big[17 \times (5/9)\big]\Big)$	= 9.44
Static VoH =	152,78 - 9.44	= 143.34	

The $EVPI_S$ concept presented by Espinoza et al. (2014) has the advantage of providing information on individual subgroups rather than the value of their sum which is used by Basu and Meltzer (2007). Research funds can then be prioritized to resolve where most of the uncertainty lies, more specifically in which subgroup. For example, in the Table 4.5 above, the $EVPI_S$ for the female subgroup is 0 and therefore, future research resolving remaining uncertainty should be focused on the male subgroup where the $EVPI_S$ is 17. van Gestel et al. (2012) suggested that exploring such subgroup EVIC could be useful before calculating the parameter-specific EVIC developed by Basu and Meltzer (2007).

What is important however is that Espinoza et al. (2014) emphasizes that there is a distinction between the benefits that can be gained using current information and the benefits that could potentially be gained with more research. These two dimensions were poorly discussed by both Coyle et al. (2003) and Basu and Meltzer (2007).

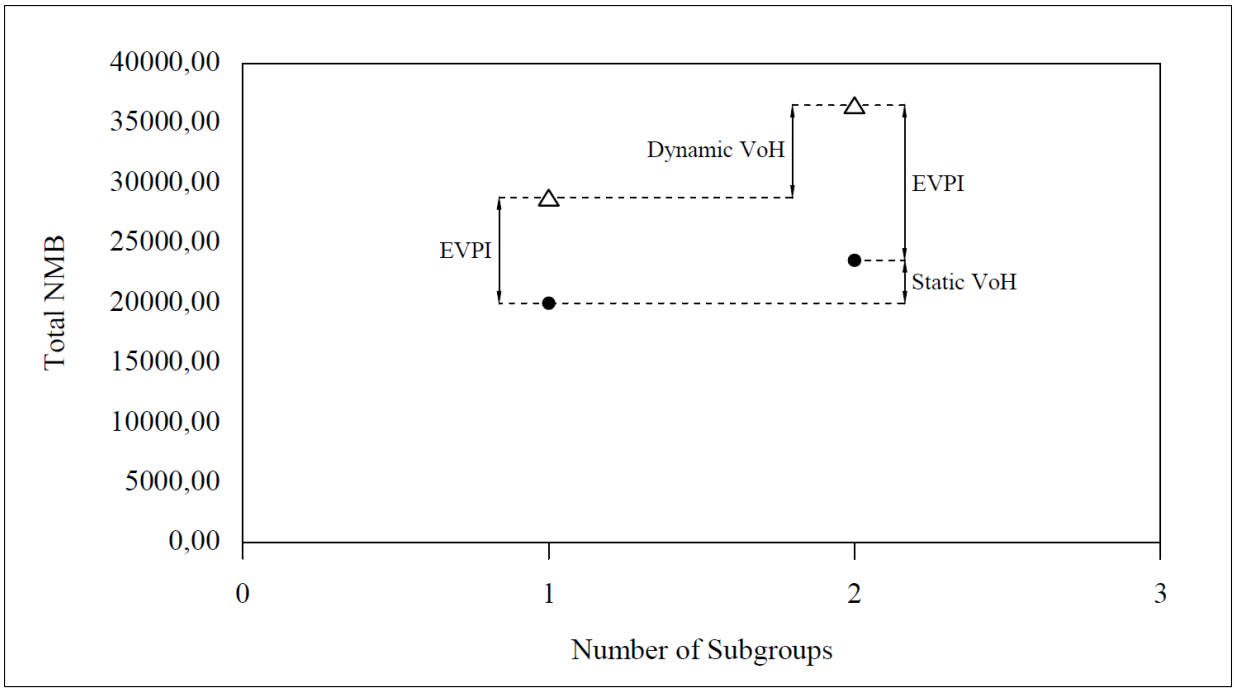


Figure 4.3: Example of the two dimensions of the Value of Heterogeneity (VoH). The total NMB are shown for both a population analysis and a two-subgroup analysis where the • = under current information and △ = under perfect information.

As we can see in Figure 4.3, the total NMB under current information and perfect information are depicted for both the population analysis and a heterogeneity analysis using two subgroups. The difference between the total NMB measured with perfect and current information is the EVPI. However, we can notice that the value under current information increases when there are 2 subgroups. The total NMBs are calculated with equation 4.3 from the SA analysis. The difference between the total NMB measured under current information in the population and in the subgroup analysis is what Espinoza et al. (2014) termed the **static value of heterogeneity**. The concept is equivalent to both the TNB gained (Δ_STNB) developed by Coyle et al. (2003) and parameter-specific EVIC developped by Basu and Meltzer (2007). The second thing to notice on Figure 4.3 is that the total NMB calculated under perfect information has also increased from the population analysis to the subgroup analysis. The difference between the two is what Espinoza et al. (2014) has termed the **dynamic value of heterogeneity**.

It is possible that sometimes there is no static VoH, meaning that the subgroup analysis did not reveal a different optimal treatment for any subgroups than the one that was identified for the whole population in a paternalistic model.

In the scenario depicted in Figure 4.3, it appears that the EVPI measured in the subgroup analysis is smaller than that of the population analysis. This leads to a dynamic VoH that is smaller than the value of the population EVPI. This suggests that the heterogeneity

parameters used were informative about some measured variability and it helped resolve part of the uncertainty in the population analysis. It is also possible that the EVPI measured in the population and subgroup analysis is the same. This would lead to a dynamic VoH equal to the population EVPI. This means that the heterogeneity analysis was not informative about the uncertainty measured in the population analysis. It also indicates that investing in further research might be worthwhile, especially for resolving uncertainty *unrelated* to patient heterogeneity.

Espinoza et al. (2014) recommended that both the dynamic and static value be presented graphically for the alternative specifications of subgroups that were identified in the efficiency frontier analysis. They remarked that in theory, as you use more parameters and your number of subgroup tends towards the number of individuals (n), there will be little residual uncertainty related to heterogeneity left and the decision-maker will be able to allocate resources as efficiently as possible. Uncertainty, however, will never be fully resolved, because "the true value of the individual treatment effect can never be measured, as the counter-factual can never be observed" (Espinoza et al., 2014, p.8). The point of their analysis is to resolve as much as possible the uncertainty explainable by the patients' personal characteristics. The left-over uncertainty will for the most part then be unrelated to patient heterogeneity.

Finally, even though it is mentioned briefly by Basu and Meltzer (2007) as well, Espinoza et al. (2014) highlight the importance of placing the information in the context of the actual population under study and the incidence of the illness or health problem the treatments are aiming to improve. The population that is expected to benefit from the information in the future is:

$$P_{(S)} = \sum_{s=1}^{S} \sum_{t=10}^{T_S} \frac{I_{s,t}}{(1+r)^t} \tag{4.21}$$

Where,
T_S = period of time over which the information could be collected for subgroup S
$I_{s,t}$ is the incidence over period t

Then P_S can be multiplied by the EVPI_S and then the value obtained is the maximum society should spend on future research to resolve the uncertainty. Alternatively, if this was multiplied by the static value of heterogeneity, it is the maximum amount of money that should be spent to reveal the characteristic if some additional costs are involved, for example the cost of a genetic test. It is important that the cost of implementing the subgroup policy does not outweigh the NMB gained.

4.4 SA, EVIC and VoH Compared

All three frameworks seek to characterize the financial gains that can be earned from a subgroup policy compared to the traditional whole-population recommendation. The SA framework developed by Coyle et al. (2003) solely focuses on finding the optimal treatment for subgroups given current information. On the other hand the EVIC framework developed by Basu and Meltzer (2007) tends to focus mostly on the gains that could be earned if care was individualized perfectly. However, the parameter-specific EVIC concept they propose essentially allows to draw the same conclusions found through the SA framework. The framework used by Espinoza et al. (2014) emphasizes that it is important to present these two dimensions of the value of heterogeneity together. It is obvious from this that the biggest drawback of the SA framework is that it fails to properly address the issue of uncertainty. On the other hand, the problem with the Basu and Meltzer (2007) presentation of the EVIC is that it seems to confuse the value of perfect information as something achievable in reality and paints an inflated picture of gains that can be earned under current information. Espinoza et al. (2014) has the advantage of presenting both distinctly and clearly the value of heterogeneity under current and perfect information. This is surely necessary to prevent confusion and to better assist subgroup decision-making.

Coyle et al. (2003) were the only ones that developed the idea of comparing different subgroup approaches to estimate the trade-off between equity and efficiency. This type of equity analysis is, in theory, also possible by using the parameter-specific EVIC. Coyle et al. (2003) were also alone in advancing the idea that policies using LUC may not always be adhered to by physicians 100 percent of the time. This is an important consideration for decision-making. Unfortunately, this time, the EVIC framework would not allow for these calculations to be done easily. It would require additional data manipulations which confers SA an important advantage.

The EVIC framework allows for the valuation of subgroups policies in a context with cost-internalization and one with no cost-internalization. Even though the idea was put forward in the context of privately funded health care, it can also be useful in the context of publicly funded health systems because there may be cases where the approach with no cost-internalization still offers some financial gains when compared to the paternalistic approach. An important weakness of the SA framework comes to light with regards to the approach with no cost-internalization. Because it uses incremental NMB, it would be impossible to calculate. Therefore, only Basu and Meltzer (2007)'s methodology can be used to compute the population and parameter-specific EVIC with no cost-internalization.

Finally, the value of heterogeneity framework by Espinoza et al. (2014) has tried to use the SA framework mainly to trace the efficiency frontier and prioritizing further exploration of heterogeneity on the few subgroup specifications that yield the best value for money given current information. While the considerations for defining subgroups presented in Chapter 3 was to avoid spending time unnecessarily exploring irrelevant parameters, the added criteria of efficiency by Espinoza et al. (2014) is not used for the same purpose. Some part of the analysis (the calculation of the TNB gained) still needs to be done. However, it can save some time for the continuation of the analysis by discarding those parameters that offer little gains. Furthermore, it is mostly useful to prioritize which information to present decision-makers. It is unnecessary to present everything that is explored during the economic evaluation as this can become burdensome to interpret for decision-makers. Therefore, only the most valuable results for the design of policies can be presented.

An interesting point to mention is that Espinoza et al. (2014) were of the opinion that the concept of EVIC is the same as EVPI. Basu and Meltzer (2007) did introduce their conceptual framework as distinct from that of the EVPI and in a later publication van Gestel et al. (2012) did support their view that the EVIC and the EVPI have important differences. They explained that while the EVIC tried to capture the value of optimizing treatment decisions at the level of individual patients, EVPI did so at the population level (van Gestel et al., 2012, Table 1, p.15). Further, they explained that this was because the EVIC captured the source of uncertainty as patient heterogeneity while the EVPI only considered the source of uncertainty as model parameter used in the population analysis (van Gestel et al., 2012, Table 1, p.15). However, since there is no distinction in the non-parametric approach to calculate EVPI when compared to the approach to calculate EVIC, it is hard to disagree with Espinoza et al. (2014). The total EVIC and EVPI is essentially the same. However, if van Gestel et al. (2012) had used the same argument to compare the parameter-specific EVIC to that of parameter-specific EVPI (or EVPPI), it would have made more sense because they both use a different consideration to resolve uncertainty (the EVIC_θ uses a patient attribute while the EVPPI uses an input parameter). Therefore, an enormous advantage of the approach used by Espinoza et al. (2014) is the clarification they brought to this general confusion by dividing the dimensions of heterogeneity into a static and dynamic value. The dynamic value of heterogeneity allowed to identify specifically what part of the uncertainty was resolved by the patient's attributes.

Finally, Espinoza et al. (2014) emphasized the importance of exploring the EVPI not only for the whole population, but also within each subgroup. This can assist decision-makers in establishing whether future research is worthwhile while pinpointing where the efforts

should be focused (in which subgroups specifically). Basu and Meltzer (2007) presented an important weakness in that regard since they only considered the total EVIC.

It is clear that describing the theoretical foundations of each framework makes it possible to identify some strengths and weaknesses. However, it is still hard to decide whether one is better than the other. This is because each seem to have advantages/disadvantages when compared to the other. This is why in the next Chapter, all were applied to the same dataset. This was done to familiarize with the methods, to see if more advantages/disadvantages could be found and to see if one specific course of action is desirable. Such an exercise is crucial before considering making recommendations for HTA practices in Norway.

Chapter 5

Analysing Patient Heterogeneity by Applying the Three Conceptual Frameworks to RCT Results

As stated in the introduction, this thesis sets out two objectives with the intention of eventually improving HTA practices in Norway. More specifically, the improvements are targeting weaknesses in the economic evaluation guidelines related to methodology used to acknowledge patient heterogeneity. The first objective, which was the focus of the previous chapter, was to describe and compare the existing methodology to acknowledge patient heterogeneity in economic evaluation. This chapter will focus on the second objective, which is to apply all three frameworks to the same dataset. Moreover, instead of working with results from a decision analytic model, results from an RCT were used. This was done intentionally and specifically for the purpose of exploring the feasibility of applying these frameworks to RCT data and identifying the challenges in that context. Ultimately, the objective of this exercise is to identify strengths and weaknesses of each method that were not identified while discussing the theoretical foundations in the previous chapter.

5.1 Materials and methods

The data used in this case study came from an economic evaluation done alongside an RCT exploring two alternative treatments for hip fracture surgery in Norway (see Frihagen et al. (2007, 2010); Bjørnelv et al. (2012)).

Inclusion/exclusion criteria

Patients included (n=222) were above 60 years of age, had no previous symptomatic hip pathology and had all suffered a femoral neck fracture. Patients who were found unfit for arthroplasty or waited for more than 4 days before surgery were excluded. Cognitive impairment or failure was not an exclusion criteria and in cases where patients could not give consent, they were only included if their family consented. (Frihagen et al., 2007, 2010).

Treatment alternatives

Patients either received a Charnley-Hastings bipolar cemented hemiarthroplasty (n=112) (referred to as prosthesis from now on) or an internal fixation (IF) with two parallel cannulated screws (n=110) (Frihagen et al., 2007).

Time Horizon

The study time spanned over the course of 2 years after the surgery. Follow-ups were made with patients at 4, 12 and 24 months post-treatment.

Measuring Health Effects

Patients filled out EQ-5D questionnaires and responses were used to assess the health-related quality of life (HRQoL) using published preference scores from a Swedish population (Tidermark et al., 2002). HRQoL along with time was then used to compute QALYs for each patients (n=90). Further, as explained by Waaler (2009); Bjørnelv et al. (2012), cognitive failure made it impossible for some patients to complete the EQ-5D questionnaires at some of the follow-ups. Waaler (2009) estimated the missing values using regression analyses with previous EQ-5D answer and Harris hip scores [1]. Using the same method, QALYs could be computed for a total of more patients (n=127).

In addition to missing answers due to cognitive failure, some patients had either missed the follow-ups or died. In those cases, Waaler (2009) assumed the HRQoL to be the measured mean of patients within the same treatment group at that specific follow-up. For this thesis, this assumption was deemed unacceptable as it would mask important variability between patients that have different personal characteristics. To prevent excluding those patients, who still had valuable information on costs and personal characteristics, they were left in the original sample data with empty cells (95 missing values).

Measuring Costs

An extensive discussion on costs estimation from used resources and their unit prices can be found in the paper by Frihagen et al. (2010) and will not be repeated here. However, relevant is that costs were divided in 3 categories:

1. Hospital costs directly due to the fracture (including re-operations, rehabilitation and out-patient consultations)
2. Hospital costs due to treatment unrelated to the hip fracture.
3. Costs related to a change in living situation, help and assistance in everyday life (ex. home-based care). Frihagen et al. (2010)

Only costs in group 1 were used for the patient heterogeneity analyses. Costs values are given in EURO (€1 = NOK9) and had been adjusted to 2006.

[1] The Harris hip score is given by medical professional to assess hip function on a number of criteria. See Frihagen et al. (2007)

Patient Heterogeneity Parameters

The data on the total population sample (n=222) also contains information on a variety of personal characteristics of which six have been selected for the heterogeneity analyses. Age, gender, dementia and anaemia were selected based on clinical relevance, while the location where the injury occurred and the living situation were selected for exploratory purposes.

Uncertainty Analysis

Non-parametric bootstrapping was used to explore uncertainty as well as to generate the needed results to conduct the patient heterogeneity analysis. Two different sampling mechanisms were used and are different than the traditional one used in CEA. They are described at length in Appendix A. A 1000 repetition of re-sampling with replacement generated equally sized samples (n=222). This resulted in five matrices of 222 by 1000 (two for expected costs, two for expected effects and one for the heterogeneity parameter). The mean of means (taken across the vector of 1000 values) were calculated for costs and effects and the percentile method [2] was used to estimate confidence intervals (C.I.). This process was repeated several time for each independent heterogeneity parameter analyses.

Patient Heterogeneity Analyses

All three frameworks described in Chapter 4 have been applied to the results generated through bootstrapping. Additional or novel calculations will be discussed along with the presentation of results.

A WTP threshold (λ) of €25 000 or €50 000 was chosen for presenting results of the analysis [3].

Material

The patient-level data in this study was first manipulated in Stata/MP 14.1 (StataCorp, 2015) to estimate the QALYs and to compile costs and heterogeneity parameters. Microsoft Excel 2016 was used to conduct all bootstrapping and patient heterogeneity analyses. Macros were written in Visual Basic (Microsoft Corporation, Redmond, WA). Graphs were also prepared with Microsoft Excel 2016.

[2]The percentile method is the most commonly used and is described in Glick et al. (2014). It essentially requires the ordering of the 1000 mean estimates and selecting a cut-off value at the 26th and 975th position.

[3]The *NICE guide to the methods of technology appraisal* suggests a maximum WTP threshold between £20 000 to £30 000 = €25 377 to €38 066 Earnshaw and Lewis (2008)

5.2 Results

5.2.1 Defining subgroups and stratification of the population sample

Before bootstrapping, subgroups needed to be defined and the data stratified. Table 5.1 presents the subgroups when data is stratified on the basis of a single parameter and Table 5.2 when the data is stratified on the basis of 2 parameters.

Table 5.1: Basis of stratification for single parameter analyses

Parameter (θ)	**Subgroup specification**			Int. Fix.	Prosthesis
			n = 222		
Age (2)	$f =$	1 for those aged between 60 to 80	68 (31%)	33 (49%)	35 (51%)
		2 for those aged 81 +	154 (69%)	79 (51%)	75 (49%)
			n = 222		
Age (4)	$g =$	1 for those aged between 60 to 70	19 (9%)	9 (47%)	10 (53%)
		2 for those aged between 71 to 80	49 (22%)	24 (49%)	25 (51%)
		3 for those aged between 81 to 90	128 (58%)	62 (48%)	66 (52%)
		4 for those aged 90+	26 (12%)	17 (65%)	9 (35%)
			n = 222		
Age (7)	$h =$	1 for those aged between 60 to 65	8 (4%)	4 (50%)	4 (50%)
		2 for those aged between 66 to 70	11 (5%)	5 (45%)	6 (55%)
		3 for those aged between 71 to 75	15 (7%)	6 (40%)	9 (60%)
		4 for those aged between 76 to 80	34 (15%)	18 (53%)	16 (47%)
		5 for those aged between 81 to 85	74 (33%)	38 (51%)	36 (49%)
		6 for those aged between 86 to 90	54 (24%)	24 (44%)	30 (56%)
		7 for those aged 90+	26 (12%)	17 (65%)	9 (35%)
			n = 222		
Gender	$i =$	1 for males	57 (26%)	25 (44%)	32 (56%)
		2 for females	165 (74%)	87 (53%)	78 (47%)
			n = 221		
Dementia	$j =$	1 for early signs of dementia	68 (31%)	40 (59%)	28 (41%)
		2 for no early signs of dementia	153 (69%)	72 (47%)	81 (53%)
			n = 219		
Anaemia	$k =$	1 anaemia at the time of hospitalization	88 (40%)	50 (57%)	38 (43%)
		2 no anaemia at the time of hospitalization	131 (60%)	61 (47%)	70 (53%)
			n = 222		
Injury occurred	$l =$	1 outdoors	45 (20%)	24 (53%)	21 (47%)
		2 inside (not at home)	22 (10%)	6 (27%)	16 (73%)
		3 inside at home	104 (47%)	55 (53%)	49 (47%)
		4 nursing home	44 (20%)	26 (59%)	18 (41%)
		5 hospital	7 (3%)	1 (14%)	6 (86%)
			n = 220		
Living	$m =$	1 at home	152 (69%)	78 (51%)	74 (49%)
		2 at a nursing home	47 (21%)	28 (60%)	19 (40%)
		3 at a care home	11 (5%)	4 (36%)	7 (64%)
		4 at a hospital	10 (5%)	1 (10%)	9 (90%)

Table 5.2: Basis of stratification for a two-parameter analysis

Parameter (θ)	**Subgroup specification**		IF	Prosthesis
Age (4) &		n = 221		
Dementia	$gj =$ 1 for those aged between 60 to 70	19 (9%)	9 (47%)	10 (53%)
	2 for those aged between 71 to 80	38 (17%)	18 (47%)	20 (53%)
	3 for those aged between 81 to 90	81 (37%)	35 (43%)	47 (57%)
	4 for those aged 90+	14 (6%)	9 (64%)	5 (36%)
	... with no early signs of dementia			
	5 for those aged between 70 to 80	11 (5%)	6 (55%)	5 (45%)
	6 for those aged between 80 to 90	46 (21%)	27 (59%)	19 (41%)
	7 for those aged 90 +	12 (5%)	8 (67%)	4 (33%)
	... with signs of early dementia			

We can see that there is some imbalances between heterogeneity subgroups for all specifications chosen. Sometimes, there are also imbalances within heterogeneity subgroups (between treatment groups). This is particularly pronounced in the "Injury occured" and "Living" last subgroups where between 85 to 90% are in the prosthesis treatment group.

Because this thesis does not aim to inform actual decision-making, but rather to explore methodology, the selection of heterogeneity parameters was not done with the careful consideration it should be given in reality. As discussed in Chapter 3, it is important to conscientiously select parameters before doing an analysis to avoid simple "data-mining" and false-positive results. Choosing the relevant personal characteristics should be done by consulting experts in the field and whenever possible, supporting the choices with previous research results suggesting there is a reasonable explanation for its relationship with the health or costs outcomes.

In the present case, age, gender, dementia and anaemia were selected because of a reasonable expectation that they may biologically or otherwise affect the effectiveness of treatments. For example:

- Age and gender are typically used as heterogeneity parameters in medical research (Cui et al., 2002)
- Older patients may experience more difficulty healing after a surgical procedure.
- Female patients are more prone to osteoporosis which may affect the success of the surgical treatments under evaluation.
- Anaemia can make patients more susceptible to infections and other complications that may affect the success of post-surgical care (Fowler et al., 2015).
- It has been suggested that older patients may have more difficulty following post-surgery instruction for walking after an internal fixation (IF) surgery. No weight-bearing is allowed

on the operated hip for a period of rehabilitation time. Higher re-operation rates may be as a result of failing to comply with these instructions or because of complications related to reduced mobility (Kos et al., 2011). This can be an important consideration for patients with cognitive problems, such as those with early signs of dementia.

Two heterogeneity parameters, location where the injury occurred and the living situation, were chosen for exploratory purposes only. They cannot assist *clinical* decision-making. However, they can serve as a morbidity indicator, meaning that some personal characteristics related to the state of health of patients depending on the way they live could influence subgroup results. Therefore, if subgroup differences are found, they can inform the decision to do more research and collect information on the presence of other illnesses or health problems likely to affect the success of the treatments under evaluation.

5.2.2 Bootstrapped Results

First, bootstrapping[4] was done on the whole population sample and expected ICER and NMB for both alternatives were estimated with C.I.

Table 5.3: Whole-Population Cost-Effectiveness Results. The mean of means is presented with 95% confidence intervals.

Treatment	**NMB (λ = €25 000)**		
	Mean	lower	upper
Internal Fixation	1 745.86	-1 462.82	4 747.82
Prosthesis	7 611.82	4 880.20	10 015.50
ICER (€/QALY)	-17 224.55	-59 183.58	13 140.31

We can see from Table 5.3 that the prosthesis treatment appears to be cost-effective with a mean NMB much higher than for IF. This effectively identifies the "paternalistic" decision recommended, which would be to treat all patients with prosthesis.

Then bootstrapping was done for the heterogeneity analyses with the sampling mechanism described in Appendix A that accounted for parameters to get subgroup results. Results from the stratification on the basis of age with specification yielding 4 subgroups (θ_g) will be presented throughout this section as a demonstration. Bootstrapped results

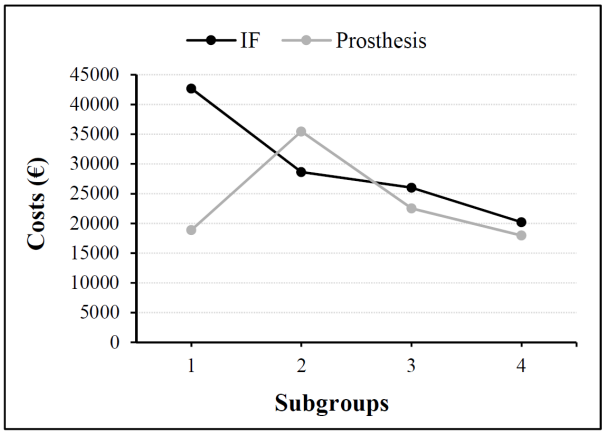
[4]It is advised to consult Appendix A for the modified population sampling protocol to account for missing values in effects data

when the population sample was stratified on the basis of other parameters can be found in Appendix B.

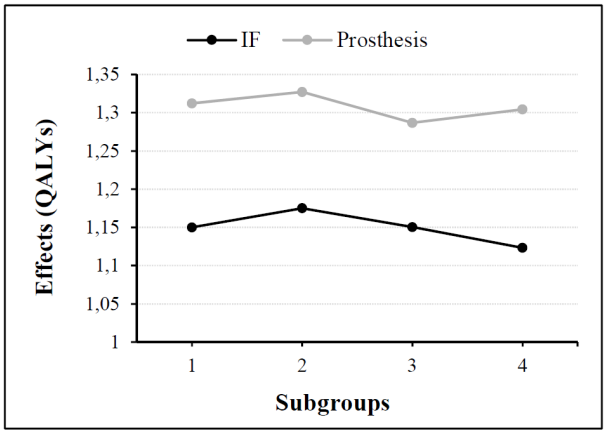
In a traditional CEA conducted on a whole population sample the uncertainty surrounding a decision is usually presented with cost-effectiveness acceptability curve (CEAC), which essentially plots the probability that a treatment is cost-effective at different WTP thresholds. While neither framework discussed in Chapter 4 explicitly said that CEACs should be presented for subgroups, it has been recommended by Briggs et al. (2006). CEACs presented alone can assist decision-making, however they are not as informative as the SA, EVIC or VoH analyses. They are presented in this thesis because they are useful for the interpretation of the results obtained through the application of the other frameworks and they will also be referred to throughout the chapter. Hence, Figure 5.1 and 5.2 are results reported traditionally in CEA, however here they are presented graphically, by subgroups, instead of population means (as presented in Table 5.3).

Figure 5.1: Bootstrapped results of the population sample stratified on the basis of age (θ_g). Subgroups 1 = age 60 to 70, 2 = age 71 to 80, 3 = age 81 to 90 and 4 = age 90 +.

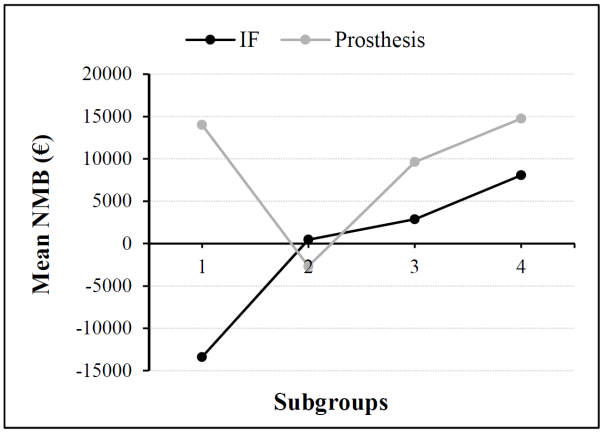
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

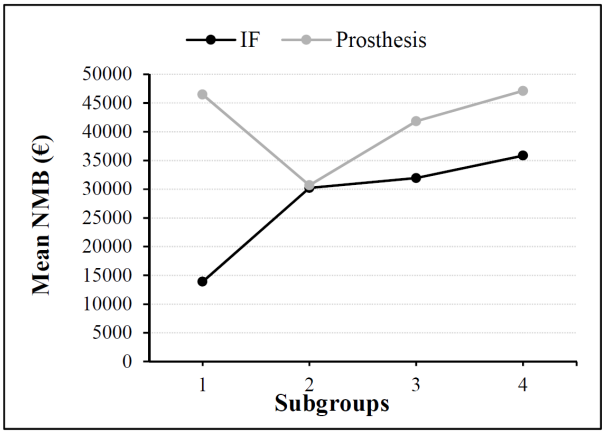
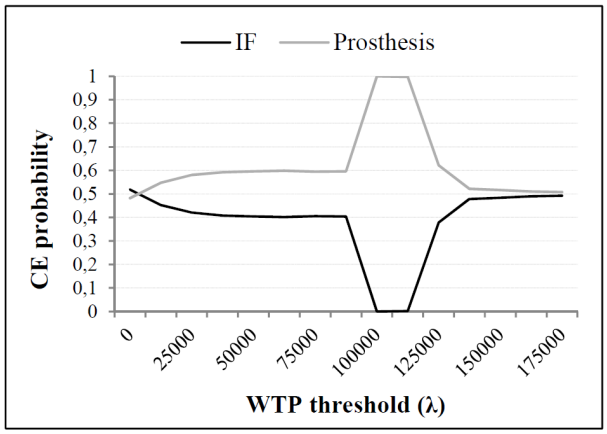
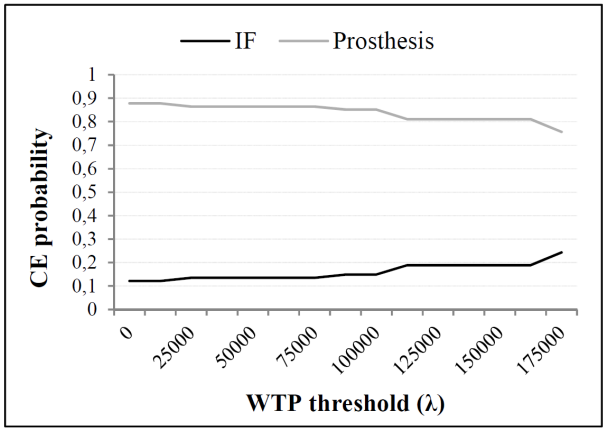


Figure 5.2: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of age (θ_g). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

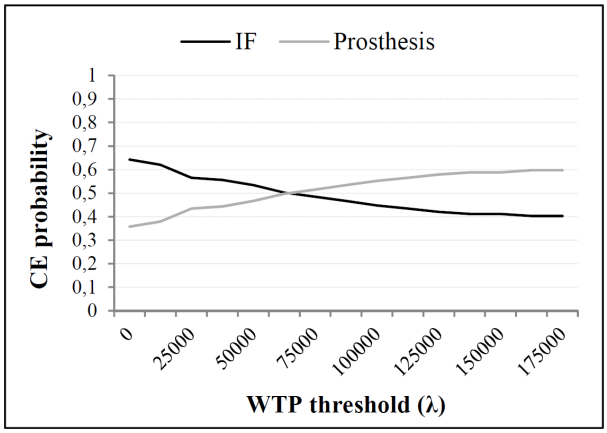
(a) Population CEAC



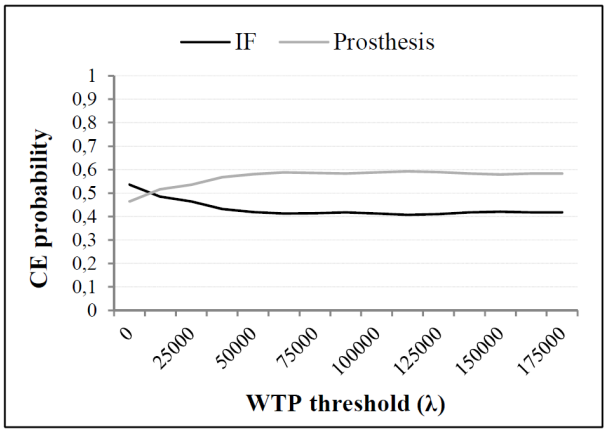
(b) Subgroup 1 = Aged between 60 and 70



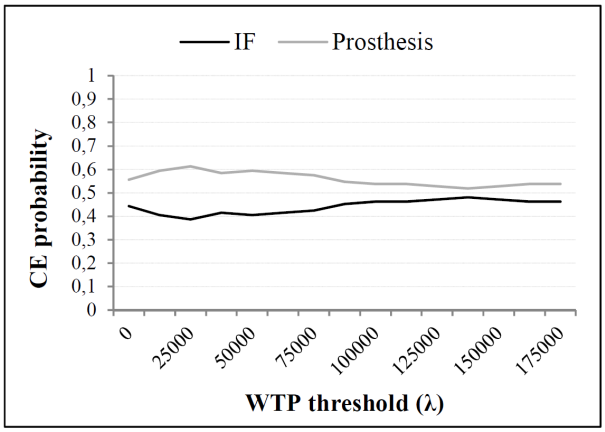
(c) Subgroup 2 = Aged between 71 and 80



(d) Subgroup 3 = Aged between 81 and 90



(e) Subroup 4 = Aged 90+



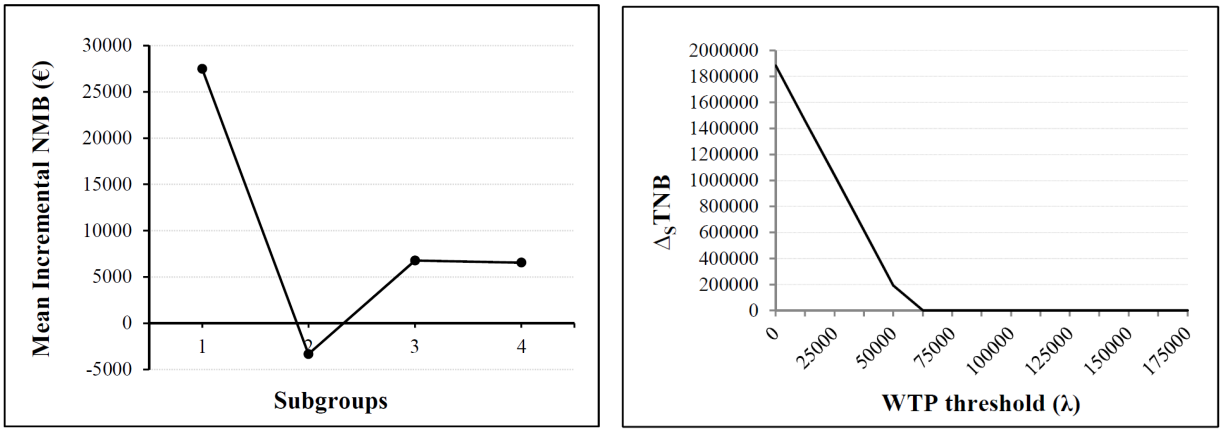
We can see from Figure 5.1 (c) that prosthesis is the optimal treatment choice for all subgroups except subgroup 2 (aged between 71 to 80). By increasing the WTP (Figure 5.1 (d)), the mean NMB for IF in subgroup 2 is not increasing as fast and the optimal treatment is unclear. Considering the CEAC in Figure 5.2(c) it can be seen that, although

very uncertain, eventually, prosthesis becomes more cost-effective for subgroup 2 as well.

5.2.3 Applying the Stratified Analysis framework

Figure 5.3: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of age (θ_g). Subgroups 1 = age 60 to 70, 2 = age 71 to 80, 3 = age 81 to 90 and 4 = age 90 +. The figure in (b) is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000) **(b)** Total NMB gained (Δ_STNB) at different WTP



The mean incremental NMB presented graphically in Figure 5.3 (a) shows again, that prosthesis is the right approach for all subgroups except subgroup 2. However, in SA, the graphical results might be easier on the eyes with only one curve and intuitively easier to interpret. Optimal subgroups (where prosthesis is cost-effective) are identified as those that have an incremental mean NMB larger than 0.

The optimal subgroups identified in the other analyses, when the population was stratified on the basis of other parameters, are presented in Table 5.4 and graphical presentation of results are found in Appendix B.

Table 5.4: Optimal subgroups identified through Stratified Analyses (SA) conducted independently on the basis of different subgroup specifications with a WTP of €25 000.

Stratification basis	**Optimal subgroups**		
Age (θ_f)	f =	all	out of 2
Age (θ_g)	g =	1, 3 and 4	out of 4
Age (θ_h)	h =	1, 2, 3, 5, 6 and 7	out of 7
Gender (θ_i)	i =	all	out of 2
Dementia (θ_j)	j =	2	out of 2
Anaemia (θ_k)	k =	all	out of 2
Injury occured* (θ_l)	l =	1, 2 and 3	out of 5
Living* (θ_m)	m =	1 and 3	out of 4
Age & dementia (θ_{gj})	gj =	1, 3, 4 and 5	out of 7

* Exploratory analysis only

Figure 5.3 (b) shows that the TNB gained tends to decrease and eventually reaches a value of 0. The same was observed when the population sample was stratified on the basis of age (θ_f), (θ_h) and anaemia (θ_k). This is because the mean incremental NMB tends to increase for all subgroups as the WTP value increases. Eventually, all the subgroups show a positive mean incremental NMB. At that point, the population (paternalistic) approach offers the same solution (everyone gets prosthesis) and there is no more benefit from stratifying. This is easier to interpret looking at the subgroup CEACs. In all cases where the TNB tends to decrease and reach a value of zero, the CEAC for individual subgroups all showed that eventually, prosthesis becomes more cost-effective.

Conversely, when the population sample is stratified on the basis of dementia (θ_j), injury occurred (θ_l) and living situation (θ_m), the TNB gained tends to increase with the WTP. This occurs when it is found that the optimal treatment is not the same for all subgroups (IF is more cost-effective for some, while prosthesis is more cost-effective for others) and as the WTP increases, there are no changes in optimal treatments. Once again, the CEACs come in handy for the interpretation and we can see that none show a cross-over situation.

An interesting case occurs when the population sample is stratified on the basis of age & dementia (θ_{gj}). The TNB gained decreases first and then suddenly kinks and starts to increase. This is because, as the WTP increases, there are changes in the optimal strategy in some subgroups (switching from IF to prosthesis and vice-versa). This can be seen also by looking at the subgroup CEACs where different changes in optimal strategies occur.

Lastly, when the sample was stratified on the basis of gender (θ_i), we can see that the TNB gained is always 0. It would be expected that the CEACs show prosthesis always being the most cost-effective for all subgroups, but surprisingly, we observe that IF is more cost effective for males at lower WTP threshold, but eventually changes to prosthesis. One might have anticipated that there would be some TNB gained at first before falling to 0. The reason for this is because the decision is actually based on the mean values of the incremental NMB, not the probability, and prosthesis is on average more cost-effective. This scenario can occur when there are a lot of individual incremental NMB values in favour of internal fixation but very close to 0. They are so small that it is not enough to bring the mean value below 0 and overtake the larger values measured in favour of prosthesis. Therefore, this suggests that there may be an advantage in presenting the CEACs along with results as the uncertainty can be better understood.

The findings of the SA analysis are resumed in Table 5.5 below with WTP threshold that are of similar values as those that may be used to inform decision-making.

Table 5.5: Results of Stratified Analyses (SA) conducted independently on the basis of different subgroup specifications. The total NMB gained (Δ_STNB) has been adjusted to reflect the NMB gained per patient treated. Results are ranked by order of magnitude considering a WTP of €25 000.

Stratification basis	**NMB gained per patient (€)**	
	λ = €25 000	λ = €50 000
Injury occured* (θ_l)	2 319.83	3 034.78
Living* (θ_m)	2 717.80	3 640.52
Age & Dementia (θ_{gj})	2 081.54	1 055.00
Dementia (θ_j)	1 366.49	2 101.99
Age (θ_h)	1 081.13	780.55
Age (θ_g)	811.04	459.53
Age (θ_f)	0	0
Gender (θ_i)	0	0
Anaemia (θ_k)	0	0

* Exploratory analysis only

Results show that the most benefits can be gained when considering the location where the injury occurred or the living situation. However, since those were used solely for exploratory purposes, they would not help in the writing of a policy with a LUC. However, it shows researchers that perhaps another underlying factor that lead to some being unable to live independently, could also influence the success of the treatment alternatives under study. Because we used it for exploratory purposes assuming it was a morbidity indicator, other health problems to define subgroups should be explored in future research.

Results on the biologically plausible parameters show that early signs of dementia or age can help gain some NMB. These are two personal characteristics that could be used in the writing of a policy using a LUC.

It should also be noted that the three different subgroup specifications based on age exhibited large differences. Dividing the population sample in two subgroups (θ_f) failed to capture differences between individuals that were measurable when the population sample was divided by four (θ_g) and seven (θ_h) instead. This suggests that the NMB gained can be more specifically attributed to those aged between 76 and 80 years.

Age is often not considered in economic evaluation for ethical reasons and also some might show reluctance using mental illness as a mean for discriminating between patients. Therefore, an equity analysis was done to assess the trade-off with efficiency in the case where decision-makers preferred to not recommend an LUC.

Table 5.6: Equity analysis using the SA Framework with a WTP of €25 000.

Stratification basis	Optimal subgroups	% sub-optimally treated	Δ_STNB per patient	Δ_ETNB per patient
Age and Dementia(θ_{gj})	$gj = 1, 3, 4$ and 5	0%	2 081.54	-
Only Dementia (θ_j)	$j = 2$	∼22.2%	1 366.49	715.05
Only Age (θ_g)	$g = 1, 2, 3, 5, 6$ and 7	∼31.2%	811.04	1 270.50
None	All	∼43.4%	0	2 081.54

If discrimination on the basis of dementia was considered unethical for the writing of an LUC, up to €715.05 would be lost per patient and approximately 22.2% of patients would be sub-optimally treated given current information. This effect is even larger if discrimination on the basis of age was considered unethical, with a loss of €1270.50 per patient and with around 31.2% of patients sub-optimally treated. When both are not considered, the loss is quite significant and the number of patients sub-optimally treated grows over 40%.

Finally, the concept of "leakage" developed by Coyle et al. (2003) was used to evaluate the NMB loss if there was non-adherence to the recommended LUC based on age (θ_g) where IF is recommended for those aged between 71 and 80. When the LUC is not adhered to 10% of the time, the loss would be approximately €81,10 per patient. This is well below the gains from applying the LUC and shows that it would still be worthwhile despite this small loss.

5.2.4 Applying the Expected Value of Individualized Care framework

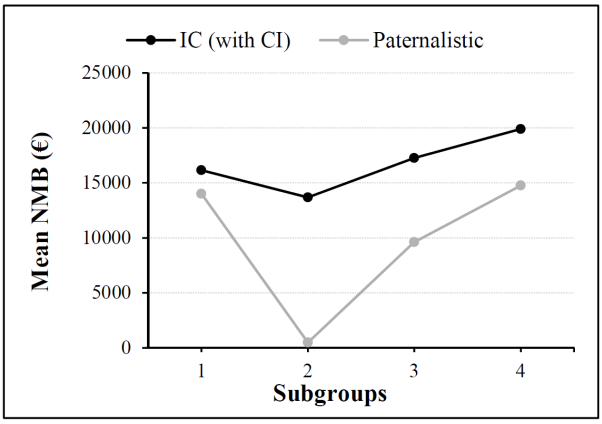
Results from applying the EVIC framework to the population sample stratified on the basis of age (θ_g) are presented graphically in Figure 5.4. Figure 5.4 (a) and (b) demonstrate that, consistent with the theory, when the calculations are done with cost-internalization, the mean NMB per patient in all subgroups is higher than when calculated with no-cost internalization. The same was observed for all the analyses done with the sample stratified on the basis of other parameters (see Appendix B).

In the case with no cost-internalization, maximizing health benefits in subgroup 1 (aged between 60 and 70) lead to a lower mean NMB than with a paternalistic approach. Similar situations occurred when stratified on the basis of age (θ_h) for subgroups 1, 2 and 4, on the basis of where the injury occurred (θ_l) for subgroups 2, 3 and 4, on the basis of the living situation (θ_m) for subgroups 3 and 4 and finally on the basis of age and dementia (θ_{gj}) for subgroups 1, 2, 3, 4 and 5. (See Appendix B for graphical results). This means that depending on the monetary gains from the other subgroups, individualizing care is

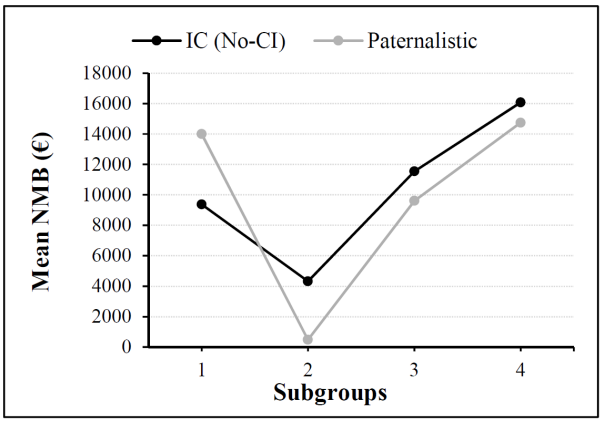
potentially overall less beneficial than the paternalistic approach.

Figure 5.4: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of age (θ_g). Subgroups 1 = age 60 to 70, 2 = age 71 to 80, 3 = age 81 to 90 and 4 = age 90 +. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

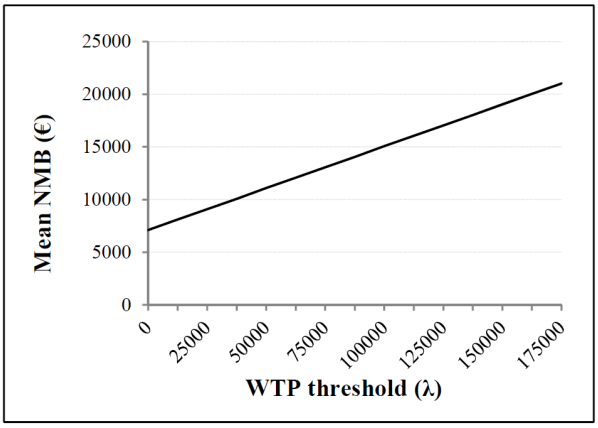
(a) Mean NMB by Subgroup (λ= €50 000)



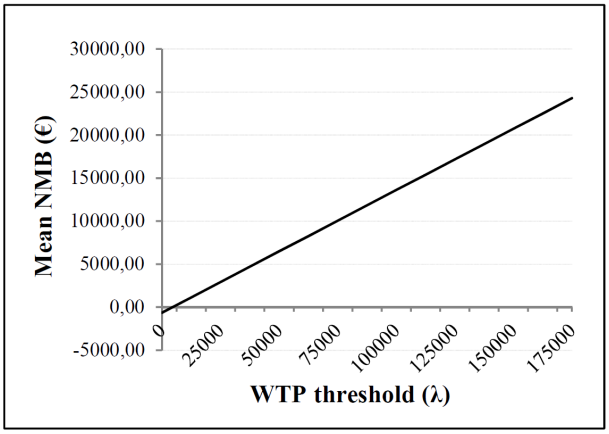
(b) Mean NMB by Subgroup (λ= €50 000)



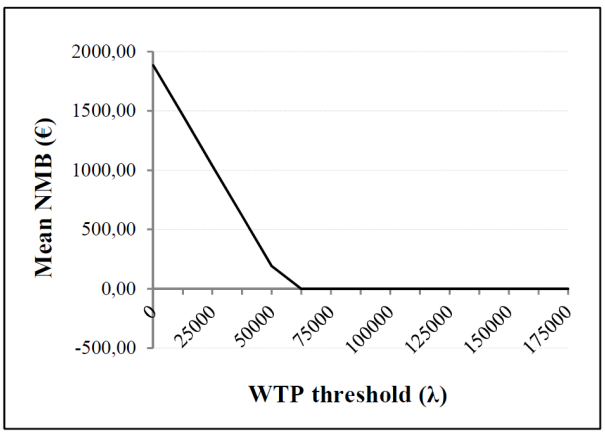
(c) Mean EVIC with CI at different WTP



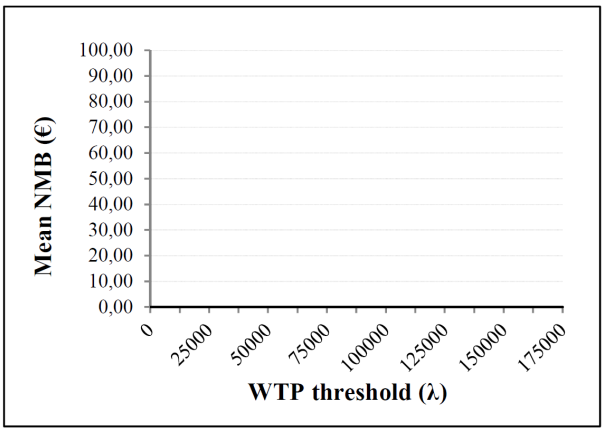
(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP



Once again consistent with the theory, Figure 5.4 (c) and (d) show that the EVIC with cost-internalization is greater than with no cost-internalization. Further, it shows that both tend to increase with higher WTP. This is because when considering the *whole population sample*, the probability that prosthesis is cost-effective remains above that of IF over all values of WTP considered (see the population CEAC in figure 5.2(a)). Also, as the theory would have predicted, Figure 5.4 (e) shows that the parameter-specific EVIC with cost-internalization is identical the TNB gained measured in the SA analysis when it is adjusted per patient. The same was observed for all the analyses done with the sample stratified on the basis of other parameters (see Appendix B).

However, something is different when the parameter-specific EVIC is calculated under no cost-internalization. Figure 5.4 (f) depicts that in the case of stratification based on age (θ_g), its value is always 0. Similar results were found for stratification on the basis of age (θ_f), gender (θ_i) and anaemia (θ_k). These results can be explained by the fact that the strategy of maximizing health effects identified prosthesis as the optimal treatment for all subgroups, which is the same as the paternalistic approach. Therefore, there are no NMB gained. Differently, when stratified on the basis of age (θ_h), dementia (θ_j), injury occured (θ_l), living (θ_m) and age & dementia (θ_{gj}), the parameter-specific EVIC (no C-I) tends to increase. This means that the strategy of maximizing health benefits does not always choose prosthesis as the optimal treatment in all subgroups. If the curve shows a negative value, it means that individualizing care by maximizing health effects is overall less cost-effective than the paternalistic approach. On the other hand, when it shows positive values, individualizing care by maximizing health effects is more cost-effective than the paternalistic approach (see Appendix B for examples).

Results from the EVIC analyses done on the population sample stratified on the basis of different parameters are resumed in Table 5.7 below. It is important to note that the population EVIC is presented for all parameters. This is because it slightly varies (between €7 808 to €9 187 per patient) as a results of the bootstrap sampling mechanism employed and the fact that results for single parameters were obtained through independent analyses. This is problematic when trying to compare between parameters because the baseline (population mean) is not the same.

It is clear from the results that the calculation of the parameter-specific EVIC with cost-internalization offers no new information if conducted after the SA analysis. On the other hand, when the EVIC is calculated with no cost-internalization it shows that for dementia (θ_j), location where the injury occurred (θ_l) and the living situation (θ_m) that even with the approach that maximizes the patient's health benefits rather than cost-effectiveness, some gains are still obtained when compared to the paternalistic approach. This means

that for an increase in efficiency, there is no need to sacrifice health benefits at all. The trade-off in efficiency gains and health benefits can easily be calculated by the difference between the two parameter-specific EVIC. For example, €927.96 would be lost per patient if the policy using a LUC based on the living situation was written with the goal of maximizing the health returns for patients. However, €1789.84 is still gained when compared to the paternalistic approach. Therefore, the decision-maker would have to balance the trade-off between efficiency gains and health gains when making a decision. The idea of considering the trade-off between the two approaches never seems to have been explicitly suggested before.

Table 5.7: Results of the EVIC for all parameters and subgroup specifications. The analysis is done with a maximum WTP of €25 000. Results are given in the form of NMB(€) per patient.

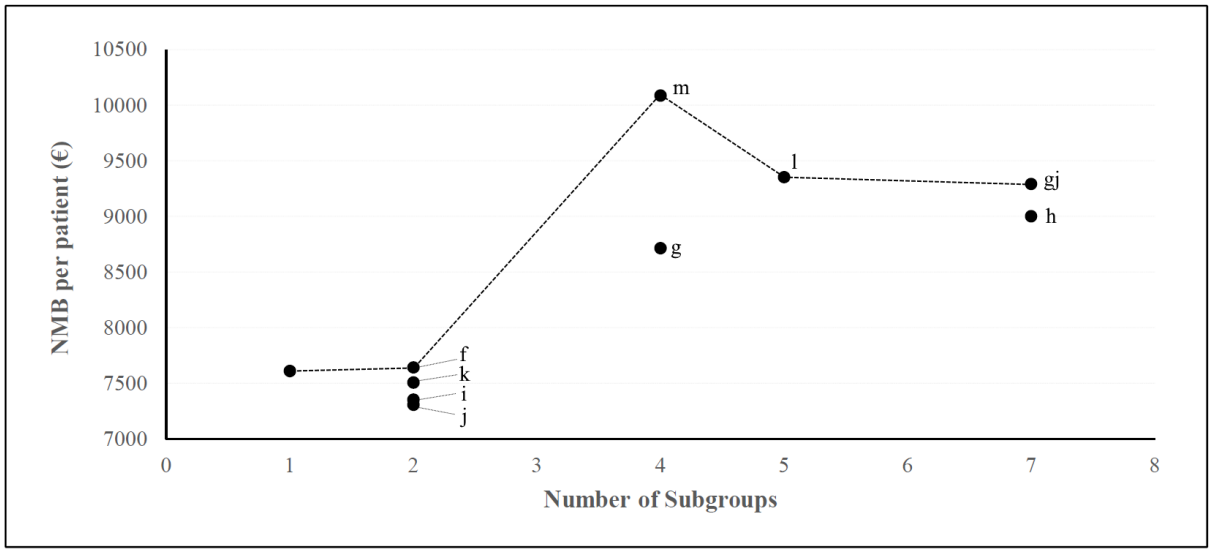
	With Cost-Internalization			**No Cost-Internalization**		
Strat. Basis	EVIC(pop)[a]	EVIC(θ)[b]	EVIC$_\theta$[c]	EVIC(pop)[d]	EVIC(θ)[b]	EVIC$_\theta$[c]
None	8 800.55	-	-	2 522.63	-	-
Age (θ_f)	8 699.38	8 699.38	0	2 572.27	2 572.27	0
Age (θ_g)	8 765.61	7 954.57	811.04	2 382.21	2 382.21	0
Age (θ_h)	8 879.44	7 798.31	1 081.13	1 858.25	3 051.72	-1 193.47
Gender (θ_i)	9 187.73	9 187.73	0	2 945.52	2 945.52	0
Dementia (θ_j)	9 056.41	7 689.92	1 366.49	3 484.33	2 117.84	1 366.49
Anaemia (θ_k)	8 863.02	8 863.02	0	2 572.06	2 572.069	0
Injury occurred* (θ_l)	7 962.36	5 642.53	2 319.83	1 802.12	347.38	1 454.74
Living* (θ_m)	8 109.76	5 391.96	2 717.80	3 496.15	1 706.31	1 789.84
Age + Dementia (θ_{gj})	7 808.69	5 727.15	2 081.54	1 531.74	1 808.55	-276.81

[a]Expected value of individualized care calculated for the whole population using equation 4.9 through 4.11
[b]Expected value of individualized care calculated by subgroups. Representing the 2nd term of equation 4.15
[c] Parameter-specific EVIC calculated with equation 4.15
[d]Expected value of individualized care calculated for the whole population using equation 4.12 through 4.14
*Exploratory analysis only

5.2.5 Applying the Value of Heterogeneity framework

First, the efficiency frontier was estimated using the total net monetary benefit gained as proposed by Espinoza et al. (2014). Considering the maximization problem articulated in equation 4.16 from chapter 4, the frontier appears to be formed by the parameters age (θ_f), living (θ_m), injury occurred (θ_l) and age & dementia (θ_{gj}). The subgroups and specification providing the highest gains is those based on the parameter living (θ_m). Results are presented in Figure 5.5.

Figure 5.5: Efficiency frontier traced by maximizing with current information the NMB per patient using a WTP of €25 000. All subgroup specifications are presented on the graph where f = age (2), g = age (4), h = age (7), i = gender, j = dementia, k = anaemia, l = injury occurred, m = living and gj = age & dementia.



The shape of the frontier is not convex as depicted by the theory. That is simply because not all parameters were considered in multiplication with each other and with all specification configurations feasible. This was not possible simply because the dataset used was too small to allow for this type of analysis. Problems related to this will be elaborated in the discussion section. Regardless, it does not prevent the identification of the efficiency frontier given the parameters considered.

Then using the method proposed by Espinoza et al. (2014), the maximum mean NMB under current and perfect information was computed for both a scenario with cost-internalization and no cost-internalization. From this, it was possible to calculate the static VoH and the EVPI for each of the bootstrapped samples stratified based on different parameters. Results are presented in Table 5.8 below.

The first thing to notice is that when looking at the clusters of subgroup numbers, the highest mean NMB calculated under current information did not necessarily provide the highest static VoH. This was the case in the cluster where the subgroup number is two. We can see that age (θ_f) offered the highest mean NMB under current information while dementia (θ_j) offered the highest static value.

Table 5.8: Results of the Value of Information (VoH) analysis for all parameters and subgroup specifications. The analysis is done with a maximum WTP of €25 000. Results are given in the form of NMB(€) per patient and have been ranked by their number of subgroups (S) and value under current information.

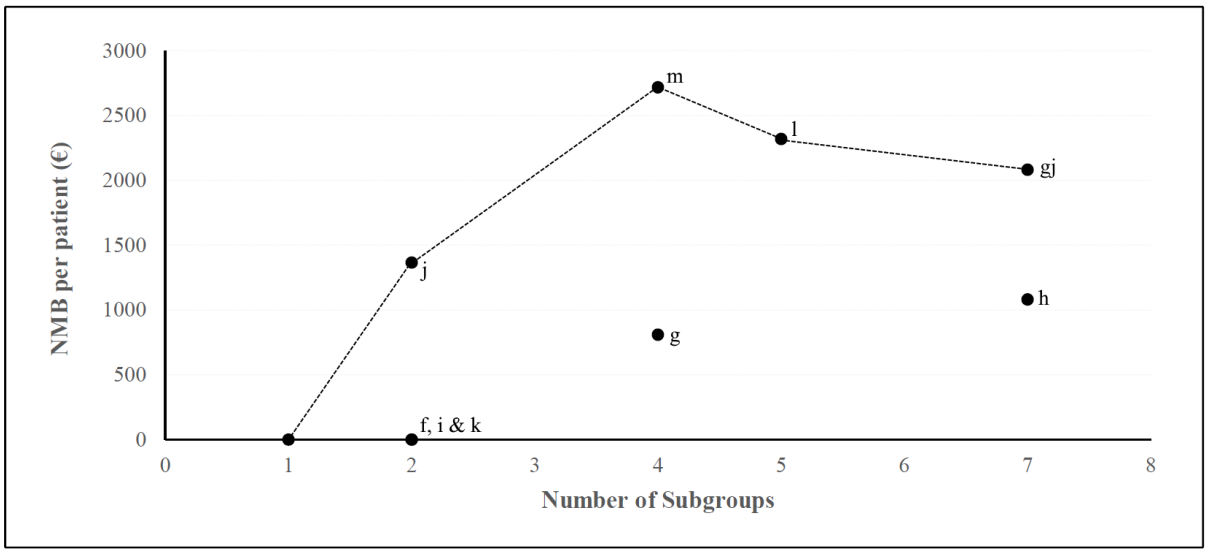
		With Cost-Internalization				
S	Stratification Basis	Paternalistic[a]	current info.[b]	perfect info.[c]	Static VoH[d]	EVPI[e]
1	None	7 611.82	7 611.82	16 412.37	0	8 805.55
2	Age (θ_f)	7 643.84	7 643.80	16 343.22	0	8 699.38
2	Anaemia (θ_k)	7 509.65	7 509.65	16 372.67	0	8 863.02
2	Gender (θ_i)	7 354.57	7 354.57	16 542.30	0	9 187.73
2	Dementia (θ_j)	5 942.53	7 309.02	14 998.94	1 366.49	7 689.92
4	Living* (θ_m)	7 371.88	10 089.68	15 481.64	2 717.80	5 391.96
4	Age (θ_g)	7 906.63	8 717.67	16 672.24	811.04	7 954.57
5	Injury occurred* (θ_l)	7 035.76	9 355.59	14 998.12	2 319.83	5 642.53
7	Age + Dementia (θ_{gj})	7 211.58	9 293.12	15 020.27	2 081.54	5 727.15
7	Age (θ_h)	7 922.52	9 003.65	16 801.96	1 081.13	7 798.31
		No Cost-Internalization				
1	None	7 611.82	7 611.82	10 134.45	0	2 522.63
2	Age (θ_f)	7 643.84	7 643.80	10 216.11	0	2 572.27
2	Anaemia (θ_k)	7 509.65	7 509.65	10 081.71	0	2 572.06
2	Gender (θ_i)	7 354.57	7 354.57	10 300.09	0	2 945.52
2	Dementia (θ_j)	5 942.53	7 309.02	9 426.86	1 366.49	2 117.84
4	Living* (θ_m)	7 371.88	9 161.72	10 868.03	1 789.84	1 706.31
4	Age (θ_g)	7 906.63	7 906.63	10 288.84	0	2 382.21
5	Injury occurred* (θ_l)	7 035.76	8 490.50	8 837.88	1 454.74	347.38
7	Age + Dementia (θ_{gj})	7 211.58	6 934.77	8 743.32	-276.81	1 808.55
7	Age (θ_h)	7 922.52	6 729.05	9 780.77	-1 193.47	3 051.72

[a]Alternative with the highest mean NMB
[b]Maximum mean NMB with current info. is calculated from equation 4.16 and adjusted per patient.
[c] Maximum mean NMB with perfect info is calculated with equation 4.17.
[d]Static VoH is calculated with equation 4.20
[e]EVPI is calculated by removing the value of current info. to that of perfect info.
*Exploratory analysis only

The fact that the mean NMB under current information and static VoH do not match in their ranking order is because of the bootstrap sampling mechanism used and the fact that the analyses were conducted independently. This resulted in a baseline value (paternalistic) that is not the same for all parameters. Consequently, it can be misleading when using the mean NMB under current information to trace the efficiency frontier. It would be best to use relative values, hence the static VoH. In that case, the frontier is only slightly changed and is now formed by dementia (θ_j), living (θ_m), injury occurred (θ_l)

and age & dementia (θ_{gj}). The new efficiency frontier using the static VoH is presented in Figure 5.6 below.

Figure 5.6: Efficiency frontier traced with the static value of heterogeneity (VoH) using a WTP of €25 000. All subgroup specifications are presented on the graph where f = age (2), g = age (4), h = age (7), i = gender, j = dementia, k = anaemia, l = injury occurred, m = living and gj = age & dementia.



Results presented in table 5.8 can then be used to plot visually the alternatives offering the best gains under current information along with their associated EVPI. Espinoza et al. (2014) suggested they should be presented this way, probably to simplify the passing on of information to decision-makers and avoid confusion by presenting too many results.

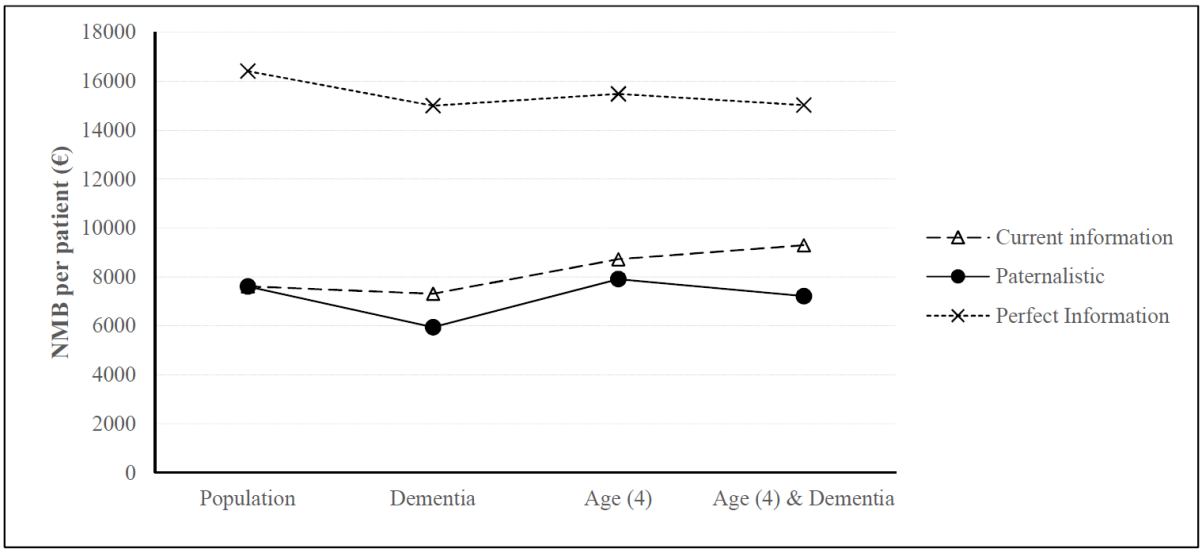
Distinctively, in this case study, it was decided to also plot the paternalistic approach as a reference because the baseline results were different between parameter analyses. Figure 5.7 presents the scenario with cost-internalization and with no cost-internalization. The parameters that were exploratory in nature (location where injury occurred and living situation) have been omitted because they would not serve to inform a decision for the writing of a policy using a LUC.

As results suggested in Table 5.8, it is now easy to visualize that the EVPI calculated with cost-internalization is much higher than the one calculated with no cost-internalization.

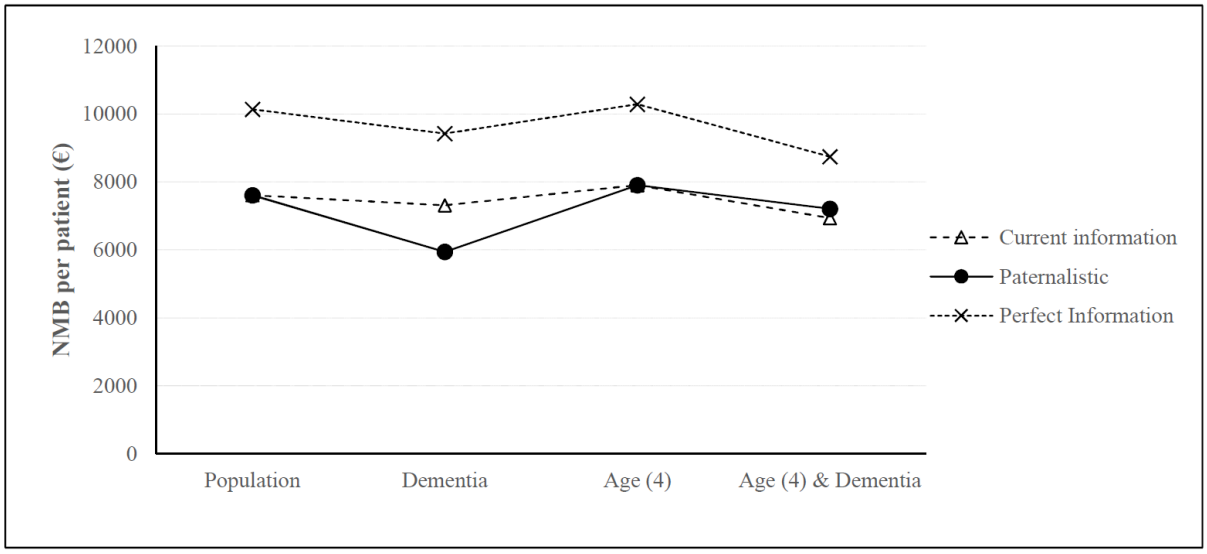
The static VoH is higher when the analysis is done assuming cost-internalization. Based on numbers from Table 5.8, we also know that the static VoH under cost-internalization is identical to total NMB gained (Δ_STNB) and the parameter-specific EVIC. This is not surprising since the theory discussed in chapter 4 predicted that although different steps were taken to get there, they were essentially the same concept.

Figure 5.7: Dimension of the VoH calculated with a WTP of €25 000

(a) With Cost-Internalization



(b) No Cost-Internalization



Moreover, something that was not anticipated at first occurred as a result of using RCT data. The dynamic VoH proved to be impossible to estimate because of the fact that analyses were conducted independently for each parameter. It was mentioned earlier that, as a result, the baseline (paternalistic) value is different for all parameters. The dynamic VoH would normally be calculated by subtracting the mean NMB calculated under perfect information for the entire population sample to that of the one calculated under a sample stratified on the basis of a parameter for example. As Espinoza et al. (2014, p.12) explained it: "the dynamic value, [...] is the additional value of resolving second-order uncertainty in the future when we compare 2 adjacent levels of disaggregation". In this

thesis, because the baseline values are different, using this operation would be wrong from a theoretical viewpoint. Therefore, it cannot be known if heterogeneity was informative or not and resolved some of the uncertainty found in the population analysis.

The fact that the dynamic VoH cannot be calculated is a drawback from using the bootstrap sampling mechanism used. However, it does not prevent the analysis from informing which parameters are best to gain a better value for money. It also does not prevent the calculation of the subgroup EVPIs as suggested by Espinoza et al. (2014). Therefore, decision-makers could still find out where the largest part of the subgroup-related uncertainty lies.

For example, using Table 5.9 below, when the sample stratified basis of age (θ_g) the largest part of the uncertainty is found in the second subgroup. Considering that it is the only subgroup that was found to not be optimal (see Table 5.4), then a LUC for subgroup 2 would maybe have to wait or perhaps revised after the results of future research.

Table 5.9: Subgroup EVPI calculated when the population sample is stratified on the basis of age (θ_g). Results are given in the form of NMB(€) per patient.

Subgroup	$EVPI_S$
1 (aged between 60 and 70)	2 130,65
2 (aged between 71 and 80)	13 195,76
3 (aged between 81 and 90)	7 629,62
4 (aged 90 +)	5 145,75

Finally, interpreting the results altogether from Figure 5.7, it appears that if this analysis was done to inform actual decision-making, the best strategy to gain better value for money would be to recommend a LUC based on both age and dementia together. Since no necessary tests are needed, no costs are involved in implementing this policy. A ceiling of €5 727.15 per patient should be placed for additional research to resolve the remaining uncertainty.

There is no parameter evaluated under the approach with no cost-internalization that offered a solution that could benefit society other than if dementia was considered alone. However, it is irrelevant because the subgroup LUC solution with dementia with or without cost-internalization is identical.

Lastly, given that the living situation and the location where the injury occurred offered the potential to gain more efficiency, future research in elucidating what personal characteristics at play could explain these differences is worthwhile if no more than €5 642.53 per patient diagnosed every year is spent.

5.3 Discussion

Applying the three frameworks to the RCT data was an enlightening exercise for many reasons. Firstly, it demonstrated that the frameworks can also be applied to RCT results. However, in that context, some important challenges materialized that are not a concern when using modelling results. This is primarily due to the size of the original dataset and the bootstrap sampling mechanism employed. Problems in relation to this will be elaborated in the next section.

Secondly, the exercise allowed contemplating the results in details. Presenting the results graphically illustrated that they can take different forms and have different trends. Previous papers have not discussed at length how the trends can be interpreted. Since this was discussed throughout the result section, it will not be re-explained here.

Finally, the exercise also helped to elucidate some more practical advantages/disadvantages that had not been identified when comparing the methods in their theoretical foundations. It also made it easier to pinpoint some areas within the frameworks that could be improved with future research. The strengths of all three methods and their area for improvements will be reflected upon as well in this discussion section. This will help in the identification of a better course of action that can be suggested for HTA practices in Norway.

5.3.1 Challenges posed by using RCT results and the bootstrap method when analysing patient heterogeneity

Important shortcomings posed by using patient-level data and by the bootstrap method became apparent during this case study. Working with patient-level data is challenging in the context of patient heterogeneity analyses for two related reasons:

1. The stratification of a population sample by heterogeneity parameters may lead to serious imbalances *between* and *within* subgroups.

2. The more stratification a dataset is subject to, the smaller the sample (n) of the subgroup gets and the more uncertain the results will be (loss of statistical power).

It is clear that modelling offers some advantages to conduct patient heterogeneity analyses. Mean costs and effects are not calculated directly from a resulting data set, but instead are obtained through a discrete event simulation model where outcomes are given based on the probability distributions of various input parameters. Before propagating a cohort into a model, each individual patients are assigned a different vector of heterogeneity parameters (θ) resulting in a given set of subgroups that can be properly balanced

beforehand. Then, based on their vector, the input probabilities in the model will estimate the proper outputs (costs and effects) for each patient. The synthetic results generated contains, for each data point, the treatment received, costs and effects but also the *entire vector of heterogeneity parameters* (θ) associated with each patient.

When patient-level data such as the results of an RCT is used, things are not so simple. Looking at the results in Table 5.1, we can see that even with stratification based on just one parameter, there can be imbalances between and within subgroups.

Imbalances *between* subgroups makes the bootstrapped results of a sample stratified the basis of a single parameter difficult to compare to the results of another (ex. comparing age alone to gender alone). As such, when subgroup differences do exist, the re-sampling mechanism used during bootstrapping (consult Appendix A) can bias the resulting mean costs and effects in either direction. It can be explained by the fact that costs and effects pairs for each treatment alternative cannot be measured in the same patient. Bootstrapping generates, in a sense, much like a model would, a synthetic dataset where each data point has an assigned costs and effects pairs for each treatment alternatives. Using the traditional sampling mechanism for a population analysis (no stratification), the pairs from each treatment are picked at random in the entire sample. Then the optimal treatment for the population is found by the one presenting the highest mean NMB. On the other hand, when bootstrapping is done with the sampling mechanism used for subgroup analyses, the selection of pairs for each treatment is not done at random, but is instead based on the value of the heterogeneity parameter. Therefore, when the NMB of each treatment is averaged over the entire population sample, the resulting value can be different compared to if the sampling had been done at random.

It could be argued that a sampling mechanism dependent on the distribution of a heterogeneity parameter is a good approach anyway because it allows for the comparison of similar patients in subgroups. However, it is only true provided that the patient characteristics can reasonably explain differences measured in patient outcomes. This emphasizes the importance of carefully choosing the heterogeneity parameters *ad hoc*. Imbalances between subgroups is not a problem for interpreting results of a heterogeneity parameter analysis when it is considered on its own because the identification of the optimal treatment for each subgroups is still possible. It only becomes a problem when trying to compare multiple parameters whose bootstrapped results were obtained independently. It would be the same problem as trying to compare results of two discrete event simulation models that used different input probabilities. This is why baseline population results were different and why it impossible to calculate the dynamic VoH.

Imbalances *within* subgroups, also termed **selection bias**, is another complication that is much more important. This is when proper randomization between treatment groups is not achieved and it makes the results incomparable and the identification of an optimal treatment impossible or questionable at best. This randomization problem is usually more pronounced in subgroups where the sample size is small. For example, selection bias was most pronounced in the last subgroup of the sample stratified on the basis of the location where the injury occurred(θ_l) and the living situation (θ_m), with only one patient in the IF treatment group. While this thesis proceeded with the analysis anyway, in reality, this would not be considered acceptable if a decision had to be made based on the results.

While larger sample sizes can make the randomization problem easier to circumvent, a **small sample size** is a problem in itself when using patient-level data. Stratifying the population sample using the entire known vector of heterogeneity parameters would solve the problem of imbalances between subgroups because it would then be unnecessary to conduct analyses for each parameter independently. However, in reality it is often not feasible because the more stratification the original data is subject to, the smaller the subgroup sizes get and statistical power is lost. For example, looking at Table 5.1, when the population sample is stratified based on age into seven subgroups, subgroup 1 is left with only four patients in each treatment alternatives. While there is no selection bias, calculating mean costs and effects from only four values will have a large associated uncertainty and can increase the chance of type II errors. However, small subgroup sample sizes become particularly dangerous for the interpretation of bootstrapped results. This is because re-sampling from a small subgroup will consistently re-pick the same few values and will mistakenly make the empirical estimation of NMBs appear to have a small confidence interval. This can increase the chance of type I errors. This is a relevant consideration if the guidelines require formal statistical testing to establish subgroup differences.

Therefore, because the probability distribution of costs and effects within each subgroup is unknown, the subgroup sample should be of an appropriate size so that it can be reasonably estimated. Researchers should always pay close attention to the resulting subgroup sample sizes after stratification.

While van Gestel et al. (2012, p.16) believe that data from RCTs is not "suitable" for heterogeneity analysis because "they divide the study population into separate study arms", this thesis proves that it is not entirely true. Bootstrapping can provide an effective solution. However, as van Gestel et al. (2012, p.16) had suggested, "[p]atient data for EVIC analysis must therefore be retrieved from studies with special designs ...". This means that when economic evaluations are done alongside RCTs and there is a desire to

explore patient heterogeneity, careful planning by researchers will be required. This was also pointed out by (Cui et al., 2002, p.348-350):

> Subgroup analyses may be specified in a study protocol prior to the initiation of the trial. Accordingly, the original randomization may be stratified by the disease prognostic factors or other characteristics defining the subgroups. If there is no stratified randomization, the validity of randomization within each subgroup may be in question, particularly, for subgroup analyses conducted in a post-hoc fashion.
> [...]
> In general, a sufficiently large sample size and randomization stratified for the subgroup are basic means to reduce the chance of group incomparability, and consequently reduce the chance of the presence of selection bias. In case that randomization is not stratified within the subgroup, the chance of treatment group incomparability can still be reduced if the size of the subgroup is sufficiently large.

Cui et al. (2002) also pointed out that planning for subgroup analyses is rarely done in practice. While this has important implications for using patient-level data for heterogeneity analyses in CEA, further discussion related to the topic of RCT planning and design is not the main focus of this thesis. However, it is extremely important to keep these statistical issues in mind when interpreting results from the application of the three frameworks used in this thesis.

5.3.2 Reflecting on the Stratified Analysis framework

The SA framework by Coyle et al. (2003) is by far the easiest method to use when trying to see whether decisions at the subgroup-level ought to be taken. This is because little manipulation of the results are needed to identify the optimal treatment for each subgroup. However, one important drawback of the method is that it serves strictly to inform decisions under current information and fails to provide information on the second important question in economic evaluation: whether more research is worthwhile.

An interesting advantage of the SA framework is the practicality of plotting the incremental NMB per subgroups calculated. Visually, it is immediately clear what are the optimal treatments for each of the subgroups. Those who measure a mean incremental NMB below zero should receive the control treatment and those who measure above 0 should receive the new treatment. However, because the uncertainty is not addressed directly, it was found that it may be better to present results along with subgroup CEACs as was the original suggestion by Briggs et al. (2006).

Another unforeseen advantage of plotting the incremental NMB by subgroups is that it can help visualize when difference between subgroups are very large or very small. This

can be very interesting especially in the context of non-adherence and the concept of "leakage" developed by Coyle et al. (2003). This gave the idea that the concept could be elaborated further and used to assist decision-makers in determining whether a policy with a LUC should be applied strictly, or whether some room for discretion should be allowed.

One major disadvantage in using the equations 4.16 and 4.17 is that it assumes that leakage occurs equally throughout the entire subgroup. However, when considering a continuous parameter such as age, and as already pointed out by Coyle et al. (2003), leakage is more likely to occur at the values close to the cut-off for treatment. This is why Coyle et al. (2003) decided to use only the "neighbouring" subgroups. However, this does not solve the problem related to the fact that the leakage probability also likely varies for patients *within* the subgroup. Using the example of the sample stratified on the basis of four age groups (see Figure 5.3 (a)), the treatment was not optimal for ages between 71 and 80. However, a doctor with a patient that is 71 or 72 years of age, may think his health resembles that of a 60 year-old and decides to proceed with the treatment anyway. The same can also occur at the upper cut-off value. For example, when a patient is 79 years old and is frail, the physician might believe it is better to treat the patient as if he was older than 80 years of age and proceed with the treatment. Therefore, it is much less likely that leakage would occur in the mid-subgroup values such as at ages between 74 and 76.

One solution that could be explored in the future is that while classifying patients in different subgroups (either through bootstrapping or in a model), the actual value of the age could also be kept. Then, in their respective subgroups, patients could be ranked by age. The mean NMB loss due to leakage could then be estimated by considering a desired percentage of patients right next to the lower and upper cut-off values. It is very likely that they will be different in magnitude, so the losses for either side will not be the same. This is visible on Figure 5.3 (a) where the slope between subgroups 1 and 2 is much steeper than the one between 2 and 3. This suggests that there can be potentially more losses when the treatment is wrongfully given to those close to 71 years of age, compared to those closer to 80 years of age. Therefore, the policy using the LUC could recommend, for example, that it be followed strictly for those above the 71 year old cut-off value, but once approaching 80 years of age, more discretion could be given to the physician and patient when deciding which alternative to choose. The decision to allow the physician and patient to exercise their judgement when exploring which treatment to use should be based on the willingness to bear the financial risk related to making the wrong decision. While this calculation was not done in this thesis, using this method should be explored in future research.

The "leakage" calculation can equally be useful for some dichotomous parameters as well in another context not mentioned by Coyle et al. (2003). Sometimes, identifying in which subgroup a patient belongs to is not a certainty. This may be because a diagnostic test is needed to determine the presence or not of a certain prognosis factor. If a diagnostic test is needed, it will likely have a sensitivity and specificity (likelihood of false positive or negative). This also means there will be a probability that a patient is wrongfully classified in a subgroup and receives a treatment when he should not. The potential losses can be calculated by using the same "leakage" concept. Therefore, the probability of having a false positive (or negative) result could be multiplied the mean NMB found in the strata in which the patient would have been wrongfully classified. Then the sum of the losses can inform whether the LUC based on that prognosis factor is worthwhile. It could happen that the losses are greater than the actual benefits gained from the stratification. There is then no point in using the prognosis factor for the writing of a LUC as the diagnostic test is not accurate enough. On the other hand, the loss before leakage can shed light into the value it could potentially have and whether it should be considered to invest in research to ameliorate the diagnostic test's accuracy.

What is clear about the SA framework, is that it is easy to manipulate and all different kinds of calculations such as the one discussed above can easily be done to better assist decision-makers. However, it cannot be used on its own because it does not assess uncertainty and the value of perfect information.

5.3.3 Reflecting on the Expected Value of Individualized Care framework

Compared to the SA Framework, the graphical presentation of the EVIC does not offer a clear visual on the optimal treatment that ought to be given in each subgroup. It necessitates further data manipulation and the calculation of the parameter-specific EVIC. The method employed by Basu and Meltzer (2007) seems unnecessarily long and complicated after having used the SA framework. On the other hand, the EVIC framework had major advantages when compared to the SA framework. It is largely focused on the second question in economic evaluation, which is whether more research should be conducted to resolve the uncertainty. Putting a value on uncertainty and establishing a ceiling for research-related spending in the future is essential to inform decision-making.

In addition, Basu and Meltzer (2007) were the only ones to develop the idea that treatment decisions can also be taken in a context with no cost-internalization. While this is certainly important for the private health insurance market, it also has value for publicly-funded health care. As we have seen, there are occasions when a strategy that maximizes

health benefits is financially more efficient than one which uses the traditional paternalistic approach. While most of the time it will not increase the value for money as highly as the approach that uses cost-internalization, it may be a better solution overall for society by leaving less patients sub-optimally treated. It was advanced in this thesis that the trade-off between efficiency gains and health gains can be calculated by the difference between the two approaches. One further avenue that could be explored in future research would be to calculate the total percentage of patients that would be sub-optimally treated in the approach with cost-internalization and the approach with no cost-internalization. While the total NMB gained from cost-internalization and the number of patients sub-optimally treated are likely to be related, it can still put the results in a perspective that is easier to understand for the decision-maker. For example, using fictional numbers, results could be presented as: with cost-internalization €5000 is gained, but 40% of patients are sub-optimally treated, whereas with no cost-internalization, only €3000 is gained but only 27% of patients are sub-optimally treated. Sometimes, the approach with cost-internalization will leave a large amount of patients sub-optimally treated simply because despite the new treatment being more effective, they just missed the cost-effective cut-off. In those situations, it is likely that using an approach with no cost-internalization is beneficial for society. Reluctance to opt for the strategy with no cost-internalization is related to the fact that additional money saved with cost-internalization could be spent in other health programs where the benefits could be greater. This thesis cannot pronounce which solution is more appropriate, but can only suggests a way to help policy-makers in reaching a decision.

On a more critical note, Basu and Meltzer (2007) were highly focused on individual decisions rather than subgroup decisions. The enthusiasm of the research team to improve clinical decision-making for the individual is even more apparent when reading their later publications (i.e. Basu (2009)). While this objective is not wrong per say, it has perhaps lead to some confusion in the literature as to the actual usefulness of the EVIC framework in the context of subgroup analyses. It seems that Basu and Meltzer (2007) failed to properly explain the information encompassed in the EVIC. They presented its value as though potentially achievable, whereas in reality, it is not. As Grutters et al. (2013, 120) remarked: "In practice, this upper bound may not be reached, as due to variability we may not always be able to predict the optimal treatment for an individual patient (EVIC), just like we will never reach complete certainty (EVPI)". While this argument is true, there is another reason why the EVIC inflates the value of individualizing care: it does not account for the value of information *unrelated* to patient heterogeneity that it also contains. This is a considerable weakness of the EVIC that needed to be addressed. Fortunately, Espinoza et al. (2014) did exactly that.

5.3.4 Reflecting on the Value of Heterogeneity framework

Noticeably, Espinoza et al. (2014) tried to build on both the SA and EVIC frameworks' strengths and also tried to clarify some confusion between the two by distinguishing between the value of heterogeneity under current information and perfect information. The only advantage in terms of new calculations proposed by Espinoza et al. (2014) is the suggestion that subgroup EVPIs be calculated. This was to better understand how the uncertainty is distributed in the entire population sample as reported by the total EVIC suggested by Basu and Meltzer (2007). This only requires a small change in the order of steps taken in the EVIC framework and the rest is otherwise similar.

Their suggested way of estimating the efficiency frontier to pick the right level of stratification proved to be a small problem when RCT results are used. However, this problem was easily circumvented by using the static VoH instead. Another problem related to using RCT results was that the dynamic value of heterogeneity was impossible to compute. While their idea is interesting from a theoretical point of view, it is not a complete necessity to assist decision-making. Clearly it is a disadvantage because it cannot be detected if the patient heterogeneity helped resolve some of the uncertainty compared to when it was not considered. Therefore it is hard to say if future research spending should prioritize resolving uncertainty that is *related* or *unrelated* to patient heterogeneity. However, that is only partly true. The subgroup EVPIs can still be calculated and this can prevent unnecessary research on subgroups where the value of information is very low. Therefore, in a way, prioritization can still be accomplished to some degree without knowing the dynamic value of heterogeneity. It is also primordial to understand that this is a very small issue when considering the bigger picture. This is never a problem when decision analytic models are used and it also would not have been a problem if the RCT had been carefully planned and only one bootstrap exercised had been used for all parameters.

Probably the biggest practical contribution by Espinoza et al. (2014) is that they realised that results obtained through both SA and EVIC frameworks are very informative but are very inconvenient to be interpreted by decision-makers when presented in many graphs and many tables. Plotting the value under current and perfect information for the parameters that yield the best efficiency is a great way to resume results and prevent confusing decision-makers with too much information. They ensure that the decision-makers' attention is directed to the most important information retrieved from the analyses. This way the decision-making process is facilitated and priorities can be established in a forthright and timely manner.

Chapter 6

Conclusion

The two main objectives of this thesis were to (1) describe the existing methodology to analyse patient heterogeneity and (2) to apply the conceptual frameworks to one dataset. Repeatedly throughout this thesis, it was mentioned that the purpose was to identify the strengths and weaknesses of the approaches so that eventually recommendations could be given to the Norwegian Medicine's Agency and the Norwegian Health Directorate to review and renew their guidelines with proper guidance on how to assess patient heterogeneity. This is becoming imperative as we enter the era of personalized medicine. Economic evaluations that recognizes patient heterogeneity can give treatment recommendations and advise reimbursement decisions that reflect a more individualized approach and ensure that personalized medicine is implemented in a cost-effective manner.

This thesis successfully identified some important advantages conferred by all three methods and also some drawbacks, particularly when it is used on RCT results. In Chapter 4, the theoretical foundations of the three conceptual frameworks were laid out and the similarities and differences between them were highlighted. However, it was hard to conclusively say that one method was better than the other. Then in Chapter 5, the three frameworks were applied to the same dataset and allowed to get more familiar with practical issues. Fortunately, it was also possible to come up with some useful solutions. Further, some interesting future research questions could be proposed with regards to the issue of non-adherence to LUC and the choice between the approach with cost-internalization and no cost-internalization. This suggests that there is still room for improving the methods and more work remains to be done.

Most importantly is that it cannot be concluded that any single method should be preferred over another. It is clear that all three complement each other. However, some similarities between the methods were pinpointed so that in the future, different calculations leading to the same concept and result are not duplicated.

Some authors in the past have suggested that an important barrier to patient heterogeneity analyses is the fact that they are computationally intensive and very time-consuming (van Gestel et al., 2012). This served as an argument to not spend too much time exploring at length subgroup differences. However, with the advent of technology and programming, that line of reasoning no longer holds. It is true that programming all three methods for the first time can take some time. However, all methods are very similar and only require

slight changes to the protocol to obtain the results needed. The macros written in visual basics for this thesis were done so that the code was flexible and could be re-used for any size datasets. This made it possible to obtain patient heterogeneity analyses results in a matter of a day. Accordingly, it appears realistic to use results derived from all three frameworks when conducting routine economic evaluations.

The final conclusion for this thesis is therefore that an integrated approach using solutions from all three frameworks evaluated is feasible and a better solution than using either method alone. While future research is still necessary to better the methodology, based on the findings of this thesis at least a rudimentary better course of action can be recommended for HTA practices in Norway.

Recommendations for HTA practices in Norway

Firstly, it should be recommended that the desire to explore patient heterogeneity be decided prior to conducting the economic evaluation. This is especially important in the case of RCTs that require careful planning to accommodate these types of analyses. The parameters chosen should first be selected on the basis of their biological/economic plausibility and practical operationalizability.

Secondly, the personal characteristics selected should go through one additional filter, namely efficiency, as suggested by Espinoza et al. (2014). Because patient heterogeneity analyses are time-consuming, only the combination of parameters that returns the highest net benefits should be explored further. Therefore, the Coyle et al. (2003) SA framework should be recommended to explore this question. It is quick and provides all the needed information to trace the efficiency frontier suggested by Espinoza et al. (2014). The efficiency frontier should be traced with the static value of heterogeneity when the results for different parameter combinations have been obtained through independent bootstrapped simulations. On the other hand, the total NMB gained is also acceptable to use if the results are obtained through a model simulation in which the entire vector of parameter is preserved together.

Thirdly, the SA framework developed by Coyle et al. (2003) offers a quick way of identifying the optimal subgroups and can assist in making decisions at the subgroup-level that leads to significant cost-savings. Most guidelines that do suggest a specific method to analyse heterogeneity explicitly advise the use of stratified analyses (Ramaekers et al., 2013). It would therefore be consistent for Norway to recommend the same in its guidelines.

Stopping there is not sufficient. Once the most efficient parameter combinations have

been identified, the Basu and Meltzer (2007) framework should used to compute the EVIC which assesses uncertainty and the value of doing further research. Because the SA analysis has already provided the value of heterogeneity under current information, there is no need to waste time computing the parameter-specific EVIC with cost-internalization, as it is the same thing. However, if researchers are interested, they could compute one with no cost-internalization. This would be in order to determine if a strategy that maximizes health benefits can still offer some efficiency gains. When it does, the trade-off between the strategy that maximizes cost-effectiveness and the strategy that maximizes health benefits should be calculated. Decision-makers could then opt for the strategy they deem appropriate given their priority and the resources they have.

Results should be presented with graphs plotting both the value of heterogeneity under current information and under perfect information as suggested by Espinoza et al. (2014). The value under current information which comes from the SA analysis can assist in the writing of clinical guidelines or recommendations for drug prescriptions and use. The leakage calculations that quantify the losses related to non-adherence to a LUC avails to be useful in that context. It could determine when a LUC should be followed strictly or when some discretion should be left in the hands of physicians and their patients. The findings under current information should also be used to quantify the trade-off between efficiency and equity when policy-makers are reluctant on using certain heterogeneity parameters for ethical reasons. As for the value of heterogeneity under perfect information, it should be used to determine if it is worthwhile to invest in future research. More precisely, the subgroup EVPI should be estimated, as suggested by Espinoza et al. (2014), to ensure that if research funds are allocated, they are spent by prioritizing the subgroups where the largest uncertainty lies. Finally, whenever possible, the dynamic value of heterogeneity should be calculated to see if the research funds should instead prioritize resolving uncertainty that is *unrelated* to patient heterogeneity.

While this suggested integrated approach uses the three frameworks explored to their advantages, improvements and adjustments are still likely going to be necessary. As the methods become more common in economic evaluation, better solutions can be developed. Once patient heterogeneity analyses becomes part of the economic evaluation routine, it will also be interesting to see if the methods translate into concrete decision-making and whether it leads to actual cost-savings. Given current evidence, it is probable that at least proceeding with this suggested approach is much better than the status quo. Since personalized medicine is already here, policy-makers in Norway will need a strategy to ensure that it is implemented efficiently.

Bibliography

Bala, M. V. and G. A. Zarkin

2004. Pharmacogenomics and the evolution of healthcare: Is it time for cost-effectiveness analysis at the individual level? *Pharmacoeconomics*, 22(8):495–498.

Banta, D.

2003. The development of health technology assessment. *Health Policy*, 63:121–132.

Barbieri, M., N. Hawkins, and M. Sculpher

2009. Who does the numbers? The role of third-party technology assessment to inform health systems' decision-making about the funding of health technologies. *Value in Health*, 12(2):193–201.

Basu, A.

2009. Individualization at the heart of comparative effectiveness research: the time for i-cer has come. *Medical Decision Making*, 29(6):NP9–NP11.

Basu, A. and D. Meltzer

2007. Value of information on preference heterogeneity and individualized care. *Medical Decision Making*, 27:112–127.

Bjørnelv, G. W., F. Frihagen, J. Madsen, L. Nordsletten, and E. Aas

2012. Hemiarthroplasty compared to internal fixation with percutaneous cannulated screws as treatment of displaced femoral neck fractures in the elderly: cost-utility analysis performed alongside a randomized, controlled trial. *Osteoporosis International*, 23(6):1711–1719.

Brazier, J. E., S. Dixon, and J. Ratcliffe

2009. The role of patient preferences in cost-effectiveness analysis. *Pharmacoeconomics*, 27(9):705–712.

Briggs, A., T. Clark, J. Wolstenholme, and P. Clarke

2003. Missing.... presumed at random: cost-analysis of incomplete data. *Health economics*, 12(5):377–392.

Briggs, A., M. Sculpher, and K. Claxton

2006. *Decision Modelling For Health Economic Evaluation*, 1st edition. Oxford University Press.

Briggs, A. H., D. E. Wonderling, and C. Z. Mooney

1997. Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Health economics*, 6(4):327–340.

Claridge, J. A. and T. C. Fabian

2005. History and development of evidence-based medicine. *World Journal of Surgery*, 29:547–553.

Claxton, K.

1999. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of health economics*, 18(3):341–364.

Cohen, J., A. Wilson, and K. Manzolillo

2013. Clinical and economic challenges facing pharmacogenomics. *The pharmacogenomics journal*, 13(4):378–388.

Coyle, D., M. J. Buxton, and B. J. O'Brien

2003. Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. *Health Economics*, 12:421–427.

Cui, L., H. James Hung, S. J. Wang, and Y. Tsong

2002. Issues related to subgroup analysis in clinical trials. *Journal of biopharmaceutical statistics*, 12(3):347–358.

Davis, J. C., L. Furstenthal, A. A. Desai, T. Norris, S. Sutaria, E. Fleming, and P. Ma

2009. The microeconomics of personalized medicine: today's challenge and tomorrow's promise. *Nature reviews Drug discovery*, 8(4):279–286.

Donaldson, C.

1999. Valuing the benefits of publicly-provided health care: does 'ability to pay'preclude the use of 'willingness to pay'? *Social Science & Medicine*, 49(4):551–563.

Drummond, M. F., M. J. Sculpher, K. Claxton, G. L. Stoddart, and G. W. Torrance

2015. *Methods for the economic evaluation of health care programmes*. Oxford university press.

Earnshaw, J. and G. Lewis

2008. Nice guide to the methods of technology appraisal. *Pharmacoeconomics*, 26(9):725–727.

Efron, B. and R. J. Tibshirani

1994. *An introduction to the bootstrap*. CRC press.

Espinoza, M. A.

2012. Heterogeneity in cost-effectiveness analysis: Methods to explore the value of subgroups and individualized care in a collectively funded health system.

Espinoza, M. A., A. Manca, K. Claxton, and M. J. Sculpher
2014. The value of heterogeneity for cost-effectiveness subgroup analysis conceptual framework and application. *Medical Decision Making*, P. 0272989X14538705.

Festøy, H. and A. H. Ognøy
2015. Pharmaceutical pricing and reimbursement information (PPRI)- pharma profile Norway. *WHO Collaborating Centre for Pharmaceutical Pricing and Reimbursement Policies*.

Fowler, A., T. Ahmad, M. Phull, S. Allard, M. Gillies, and R. Pearse
2015. Meta-analysis of the association between preoperative anaemia and mortality after surgery. *British Journal of Surgery*, 102(11):1314–1324.

Frihagen, F., L. Nordsletten, and J. E. Madsen
2007. Hemiarthroplasty or internal fixation for intracapsular displaced femoral neck fractures: randomised controlled trial. *Bmj*, 335(7632):1251–1254.

Frihagen, F., G. M. Waaler, J. E. Madsen, L. Nordsletten, S. Aspaas, and E. Aas
2010. The cost of hemiarthroplasty compared to that of internal fixation for femoral neck fractures: 2-year results involving 222 patients based on a randomized controlled trial. *Acta orthopaedica*, 81(4):446–452.

Fure, B., V. Lauvrak, H. Arentx-Hansen, A. Skår, S. S. Ormstad, V. Jusnes Vang, and K. Bjørnebek Frønsdal
2013. Metodevurderinger: Kunnskapsbasert beslutningsstøtte på overordnet nivå i helsetjenesten. *Norsk Epidemiologi*, 23(2):165–169.

Glick, H. A., J. A. Doshi, S. S. Sonnad, and D. Polsky
2014. *Economic evaluation in clinical trials*. OUP Oxford.

Grutters, J. P., M. Sculpher, A. H. Briggs, J. L. Severens, M. J. Candel, J. E. Stahl, D. De Ruysscher, A. Boer, B. L. Ramaekers, and M. A. Joore
2013. Acknowledging patient heterogeneity in economic evaluation. *Pharmacoeconomics*, 31(2):111–123.

Helsedirektoratet
2012. Økonomisk evaluering av helsetiltak – en veileder.

Helsedirektoratet
2015. Om nasjonale faglige retningslinjer. `https://helsedirektoratet.no/nfr/om-nasjonale-faglige-retningslinjer`.

International Network of Agencies for Health Technology Assessment (INAHTA)
2016. What is health technology assessment (HTA)? `http://www.inahta.org/`.

Koerkamp, B. G., M. C. Weinstein, T. Stijnen, M. H. Heijenbrok-Kal, and M. M. Hunink
2010. Uncertainty and patient heterogeneity in medical decision models. *Medical Decision Making*.

Kos, N., H. Burger, and G. Vidmar
2011. Mobility and functional outcomes after femoral neck fracture surgery in elderly patients: a comparison between hemiarthroplasty and internal fixation. *Disability and rehabilitation*, 33(22-23):2264–2271.

Mathes, T., E. Jacobs, J.-C. Morfeld, and D. Pieper
2013. Methods of international health technology assessment agencies for economic evaluations-a comparative analysis. *BMC health services research*, 13(1):371.

Mørland, B.
2009. The history of health technology assessment in Norway. *International Journal of Technology Assessment in Health Care*, 25(sup. 1):148–155.

Mørland, B., Å. Ringard, and J.-A. Røttingen
2010. Supporting tough decisions in Norway: A healthcare system approach. *International Journal of Technology Assessment in Health Care*, 26(4):398–404.

Mukuria, C., J. Brazier, and A. Tsuchiya
2006. Exploring the relationship between health and happiness: a comparison across studies of different conditions using the sf-6d and eq-5d. In *International Society of Quality of Life Research meeting abstracts*.

Nasjonalt system for innføring av nye metoder i spesialisthelsetjenesten
2015. Om systemet (english). `https://nyemetoder.no/`.

Oxman, A. D. and G. H. Guyatt
1992. A consumer's guide to subgroup analyses. *Annals of internal medicine*, 116(1):78–84.

Phelps, C. E.
1997. Good technologies gone bad how and why the cost-effectiveness of a medical intervention changes for different populations. *Medical Decision Making*, 17(1):107–117.

Ramaekers, B.
2013. Acknowledging patient heterogeneity in health technology assessment: towards personalized decisions in innovative radiotherapy treatments.

Ramaekers, B. L., M. J. Joore, and J. P. Grutters
2013. How should we deal with patient heterogeneity in economic evaluation: A systematic review of national pharmacoeconomic guidelines. *Value in Health*, 16.

Ringard, Å., A. Sagan, S. I. Sperre, and A. Lindahl
2012. Norway: health system review. *Health systems in transition*, 15(8):1–162.

Sackett, D. L., W. M. Rosenberg, J. A. Muir Gray, R. B. Hayes, and W. S. Richardson
1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312:71–72.

Sculpher, M.
2008. Subgroups and heterogeneity in cost-effectiveness analysis. *Pharmacoeconomics*, 26(9):799–806.

StataCorp
2015. *Stata Statistical Software: Release 14. College Station*.

The Norwegian Medicines Agency (Statens legemiddelverk)
2012. Guidelines on how to conduct pharmacoeconomic analyses.

The Norwegian Medicines Agency (Statens legemiddelverk)
2016. The Norwegian health care system and pharmaceutical system. `http://www.legemiddelverket.no/English/the-norwegian-health-care-system-and-pharmaceutical-system/Sider/default.aspx`.

Tidermark, J., N. Zethraeus, O. Svensson, H. Törnkvist, and S. Ponzer
2002. Femoral neck fractures in the elderly: functional outcome and quality of life according to euroqol. *Quality of life research*, 11(5):473–481.

van Gestel, A., J. Grutters, J. Schouten, C. Webers, H. Beckers, M. Joore, and J. Severens
2012. The role of the expected value of individualized care in cost-effectiveness analyses and decision making. *Value in Health*, 15(1):13–21.

Waaler, G. M.
2009. Is hemiarthroplasty a cost effective treatment? A cost utility analysis comparing hemiarthroplasty and internal fixation in displaced femoral neck fractures.

Wang, H., J.-P. Boissel, and P. Nony
2009. Revisiting the relationship between baseline risk and risk under treatment. *Emerging Themes in Epidemiology*, 6(1):1.

Weinstein, J. N.
2000. Pharmacogenomics: Teaching old drugs new tricks. *The New England Journal of Medicine*, 3(19):1408–1409.

WHO

2000. *The World Health Report 2000: Health System: Improving Performance*. World Health Organization. Geneva, Switzerland.

World Health Organization and others

1950. The preamble of the constitution of the World Health Organization.

Appendices

Appendix A

Bootstrap Sampling Mechanisms

Non-parametric bootstrapping is used to explore uncertainty around the results. The method is ideal in the case of RCTs because it makes no assumption about the underlying distribution of costs and effects in the population (Drummond et al., 2015). With the bootstrapping method, re-samples are taken from the original dataset, with replacement, to build an empirical estimate of the sampling distribution of the ICER (Drummond et al., 2015).

The traditional sampling mechanism used in CEA described by ((Briggs et al., 1997) follows 5 steps:

1. Costs & effects pairs (measured in the same patient) from the *treatment* group are sampled at random, with replacement to simulate a new sample of size n_T. Expected mean costs $\overline{C}_T^*$ and effects $\overline{E}_T^*$ are calculated as bootstrap replicates of $\overline{C}_T$ and $\overline{E}_T$.

2. Similarly, costs/effects pairs from the *control* group are sampled at random with replacement to simulate a new sample of size n_C. Expected mean costs $\overline{C}_C^*$ and effects $\overline{E}_C^*$ are calculated as bootstrap replicates of $\overline{C}_C$ and $\overline{E}_C$.

3. The bootstrap estimate of the ICER is then calculated as:

$$\hat{R}^* = \frac{\overline{C}_T^* - \overline{C}_C^*}{\overline{E}_T^* - \overline{E}_C^*} \tag{A.1}$$

4. Then steps 1 through 3 are repeated a number of times B to obtain a vector of bootstraps estimates $(\hat{R}_1^*, \hat{R}_2^*, ...\hat{R}_B^*)$ which is the empirical sampling distribution of the ICER denoted by:

$$\overline{R}^* = \frac{1}{B}\sum_{b=1}^{B} R_b^* \tag{A.2}$$

5. Finally different methods exist to calculate confidence intervals from this estimate (Briggs et al., 1997; Drummond et al., 2015; Glick et al., 2014).

The traditional sampling mechanism described above is not used in this thesis. The sampling mechanism used for the whole-population analysis is very similar to the traditional one except that step 1 and 2 is slightly modified. This is to circumvent the problem of missing *health effects* data.

Therefore, the sampling for step 1 and 2 is modified as such:

1. Costs from the *treatment* group C_T is sampled at random and

   (a) if the effects value was observed E_T the pair is kept intact(C_T & E_T)

   (b) if the effects value is missing, one is sampled at random to form a new pair ($C_{T'}$ & $E_{T'}$), where *prime* ($'$) symbolizes a "matched" pair due to missing data.

   The sampling is repeated with replacement to simulate a new sample of size n_T. Expected mean costs ($\overline{C}_T^*$) and effects ($\overline{E}_T^*$) are calculated as bootstrap replicates of ($\overline{C}_T$) and ($\overline{C}_T$).

2. Follow the same procedure as in step 1 to sample from the *control* group to form a new sample of size n_C and calculate expected mean costs and effects.

This sampling mechanism is slightly problematic because the dissociation of costs and effects pairs to form new ones may damage the statistical integrity of the bootstrap method. Briggs et al. (1997) formally explained:

> Suppose a particular population has a real but unobserved probability distribution F from which a random sample x of n independent observations is taken and the statistic of interest s(x) is calculated. [...] The bootstrapping approach treats the observed random sample as an empirical estimate of the probability distribution of F by weighting each observation in x by the probability 1/n.

Therefore, in the case of ICER or NMB estimations, the probability distribution refers to that of the costs and effects that are *together associated* to individual patients that form the population. Further, and as pointed out by Briggs et al. (1997), the creators of the bootstrapping method, Efron and Tibshirani, advocated that the re-sampling mechanism should mirror as best as possible the way the original data was obtained (Efron and Tibshirani, 1994). Accordingly, it would not be recommended to use the sampling mechanism chosen in this thesis when the purpose of bootstrapping is to inform decision-making.

On the other hand, missing data is a problem that is very common when using RCT results. There are many ways to address the problem and Briggs et al. (2003) discuss at length various approaches for estimating values of missing data. When there is an appropriate amount of information permitting, an estimation of values is probably best than the sampling mechanism described above.

However, in the present case, it was decided acceptable to use the mechanism because the results simply serve to explore different methodology, not inform any decision. Further,

the analysis of heterogeneity would have been impossible if the choice to exclude patients with missing data had been taken instead.

In order to obtain results necessary for a heterogeneity analysis, the bootstrap sampling mechanism needed to be modified further. This time, along with costs and effects, the personal characteristics associated to patients needed to be considered. Therefore a new bootstrap sampling mechanism was applied to the data:

1. For iteration i:

   (a) A costs value from the *treatment* group is sampled at random C_T^i together with its associated heterogeneity parameter value θ_j^i (where $j = 1, 2, ...J$). Then,

      i. If the effects value is observable E_T^i the costs & effects pair is kept intact (C_T^i & E_T^i).

      ii. If the effects value is missing, one is sampled at random until its associated heterogeneity parameter value is equal to the one sampled in (a) for the same iteration and a new costs & effects pair is formed ($C_{T'}^i$ & $E_{T'}^i$).

   (b) A costs value from the *control* group is sampled at random C_C^i until its associated heterogeneity parameter value is equal to the one sampled in (a) for the same iteration. Then,

      i. If the effects value is observable E_C^i the costs & effects pair is kept intact (C_T^i & E_T^i).

      ii. If the effects value is missing, one is sampled at random until its associated heterogeneity parameter value is equal to the one sampled in (a) for the same iteration and a new costs & effects pair is formed ($C_{C'}^i$ & $E_{C'}^i$).

2. Step 1 is repeated for I iterations ($i = 1, 2, ...I$) to form a new sample of size $n_T + n_C$. Expected mean costs and effects $\overline{C}_T^*$ & $\overline{E}_T^*$ and $\overline{C}_C^*$ & $\overline{E}_C^*$ are calculated as bootstrap replicates of $\overline{C}_T$ & $\overline{E}_T$ and $\overline{C}_C$ & $\overline{E}_C$. Expected mean costs and effects are also calculated for each subgroups dependent on the heterogeneity parameter values ($\overline{C}_T^*$ & $\overline{E}_T^* \mid \theta_j$) and ($\overline{C}_C^*$ & $\overline{E}_C^* \mid \theta_j$).

3. The bootstrap estimates of subgroup NMB are then calculated as:

$$\hat{\text{NMB}}_T^* = (\overline{E}_T^* \times \lambda) - \overline{C}_T^* \mid \theta_j \tag{A.3}$$

$$\hat{\text{NMB}}_C^* = (\overline{E}_C^* \times \lambda) - \overline{C}_C^* \mid \theta_j \tag{A.4}$$

   Where λ is the selected maximum WTP threshold.

4. Then the steps 1 and 2 are repeated B number of times to obtain a vector of bootstrap estimates of NMB for both the treatment and control groups.

5. The mean of means are then calculated and the C.I. computed using the percentile methods (see Glick et al. (2014)).

Unlike in stratified bootstrapping, where the parameter of interest values are pre-defined for a number of iterations, the sampling mechanism described above allows for randomly picking a value from its distribution in the original population sample.

Alternatively, clusters of iterations could have been pre-assigned a parameter value based on its proportion in the population sample. For example, if the original dataset has 125 females and a 200 males, then the first 125 iterations could have picked costs and effects of each alternatives from only females, and for the other 200 iterations from only males.

The random selection method is chosen instead because it is consistent with the statistical principles of bootstrapping. It is also in agreement with the argument made by Efron and Tibshirani stated earlier, that the mechanism should reflect as best as possible the way the original data was obtained (Briggs et al., 1997; Efron and Tibshirani, 1994).

Assuming that the heterogeneity parameters selected can reasonably be expected to predict the effects and costs values measured in individual patients, replacing the missing values with the bootstrapping mechanism described above is an acceptable solution.

Appendix B

Bootstrapped Results of Patient Heterogeneity Analyses

In this Appendix, the bootstrapped results of independent patient heterogeneity analyses are presented. A total of 9 analyses were done and presented in the order:

- Age (θ_f) = 2 subgroups
- Age (θ_g)= 4 subgroups
- Age (θ_h) = 7 subgroups
- Gender (θ_i) = 2 subgroups
- Dementia (θ_j) = 2 subgroups
- Anaemia (θ_k) = 2 subgroups
- Location where the injury occurred (θ_l) = 5 subgroups
- Living situation (θ_m) = 4 subgroups
- Age and Dementia (θ_{gj}) = 7 subgroups

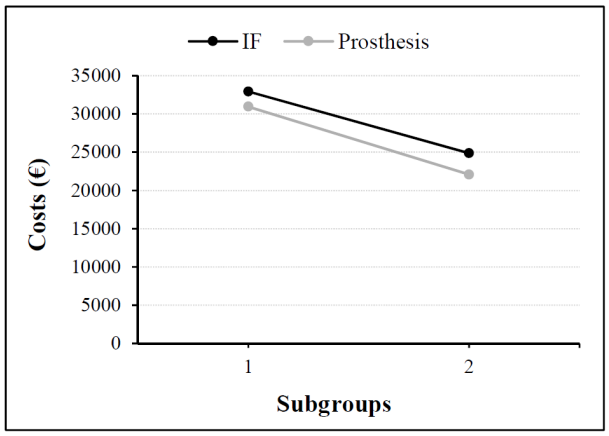
Results obtained by applying the patient heterogeneity conceptual frameworks are presented in the following order:

1. Mean Costs, effects and NMB by subgroups
2. Results of the Stratified Analysis framework
3. Results of the Expected Value of Individualized Care framework
4. Population and subgroups CEACs

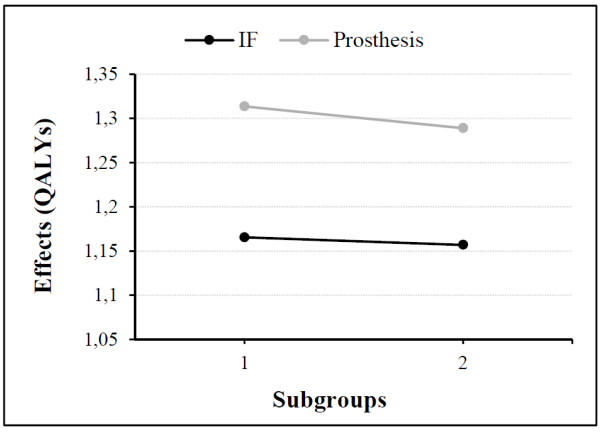
B.1 Age (θ_f)

Figure B.1: Bootstrapped results of the population sample stratified on the basis of age (θ_f). Subgroups 1 = age 60 to 80 and 2 = age 81 +.

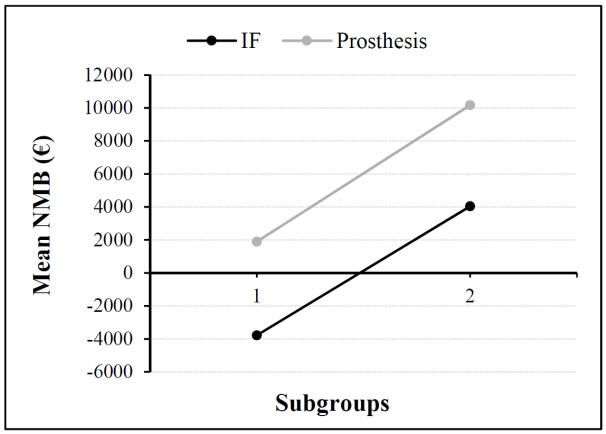
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

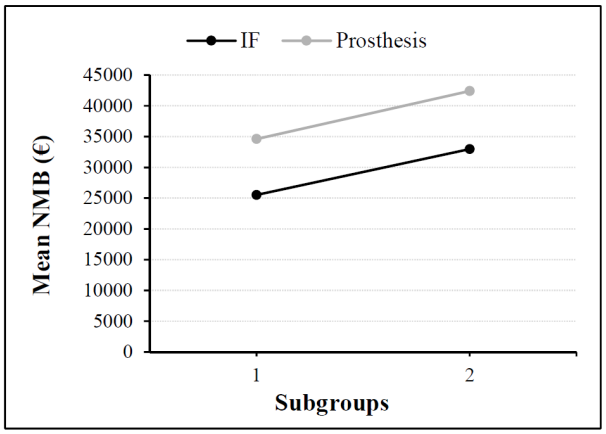
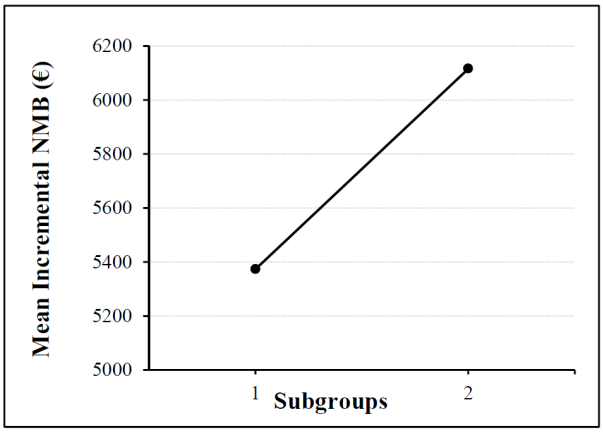


Figure B.2: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of age (θ_f). Subgroups 1 = age 60 to 80 and 2 = age 81 +. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

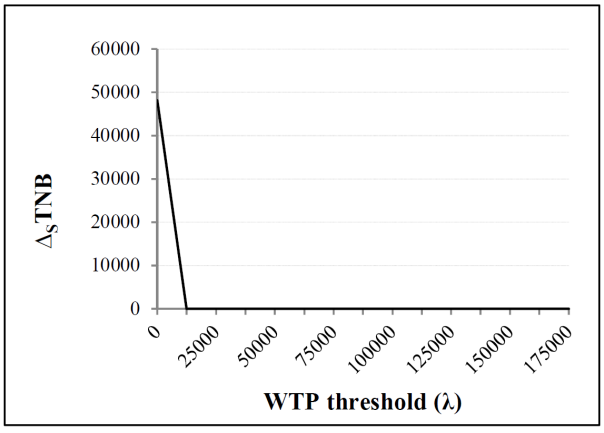
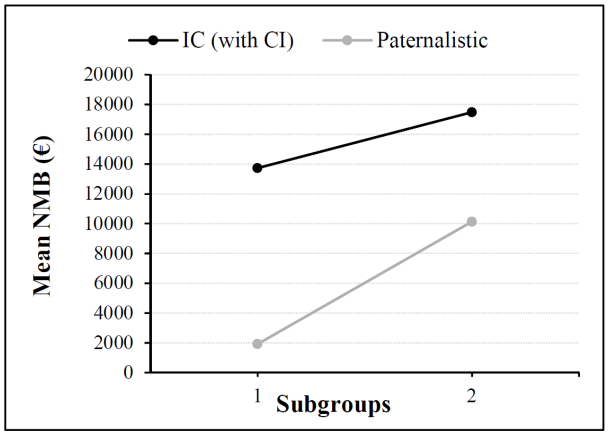
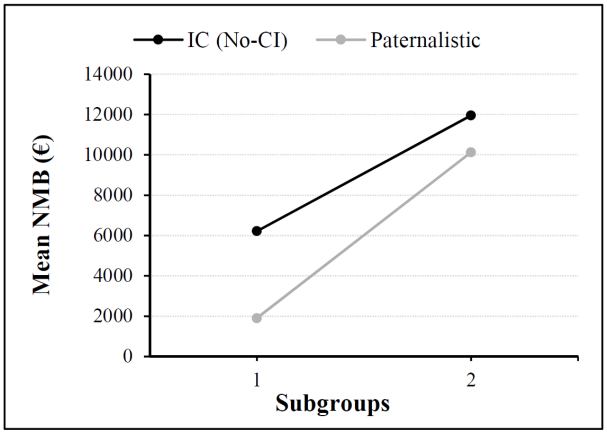


Figure B.3: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of age (θ_f). Subgroups 1 = age 60 to 80 and 2 = age 81 +. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

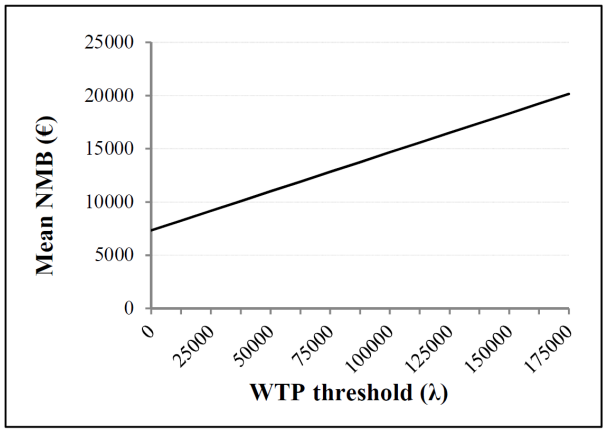
(a) Mean NMB by Subgroup (λ= €25 000)



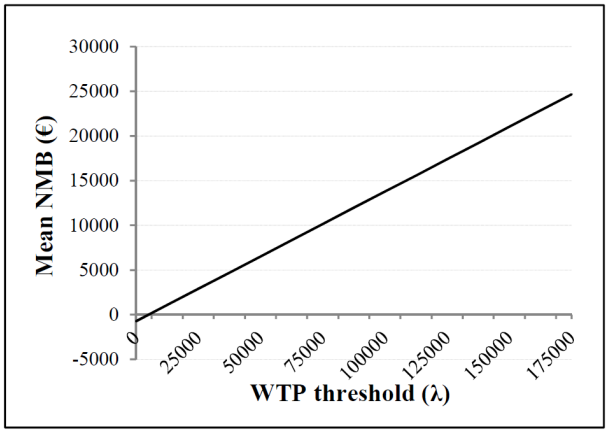
(b) Mean NMB by Subgroup (λ= €25 000)



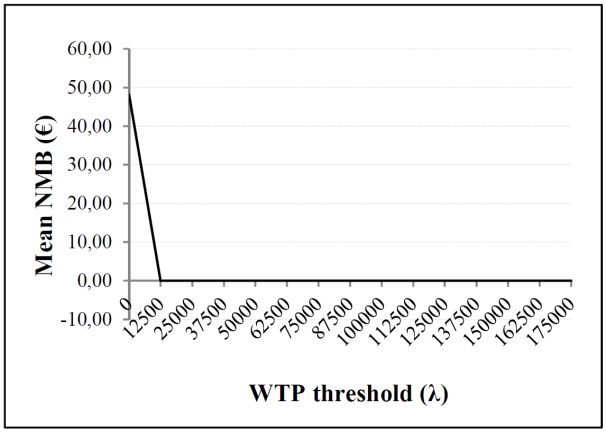
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

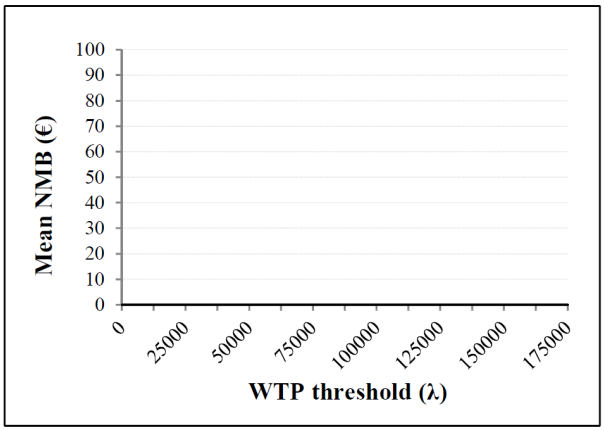
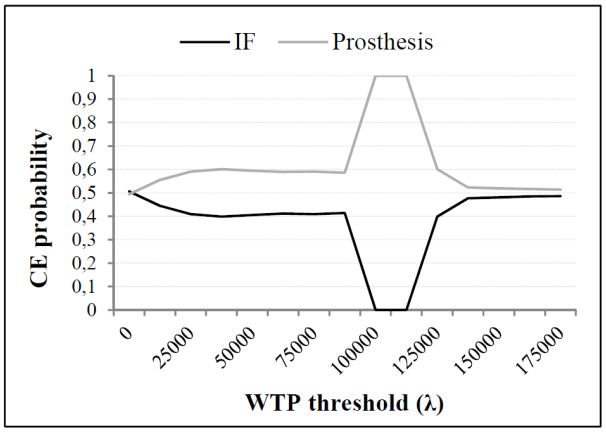
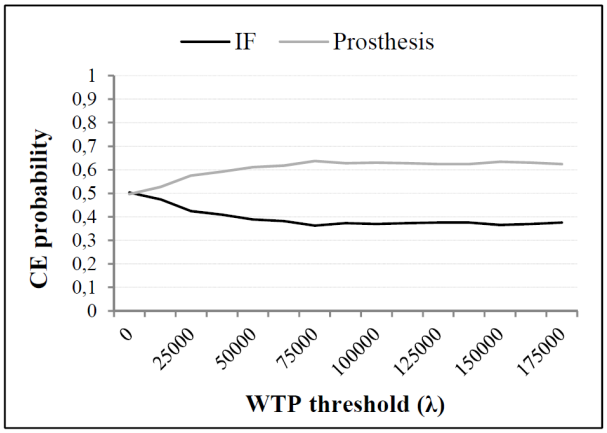


Figure B.4: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of age (θ_l). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

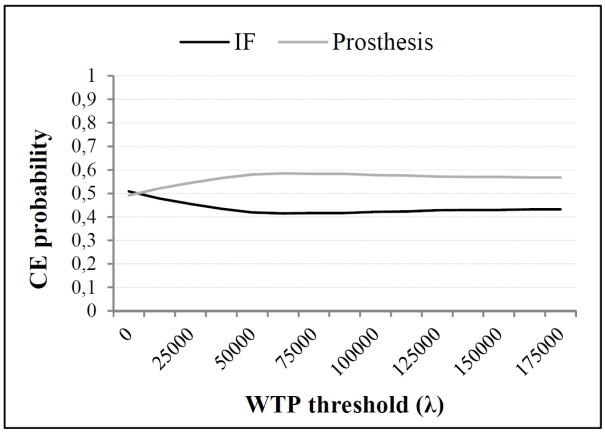
(a) Population CEAC



(b) Subgroup 1 = Aged between 60 and 80



(c) Subgroup 2 = Aged 81 +



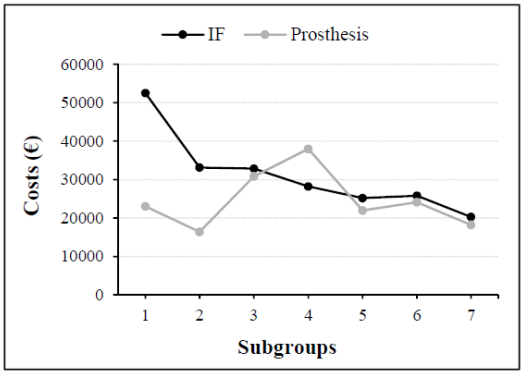
B.2 Age (θ_g)

Results presented in Chapter 5

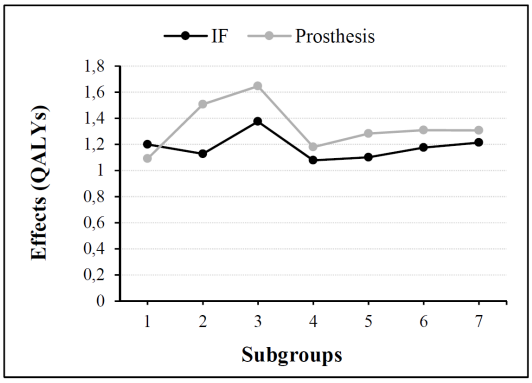
B.3 Age (θ_h)

Figure B.5: Bootstrapped results of the population sample stratified on the basis of age (θ_h). Subgroups 1 = age 60 to 65, 2 = age 66 to 70, 3 = 71 to 75, 4 = 76 to 80, 5 = 81 to 85, 6 = 86 to 90 and 7 = 91+.

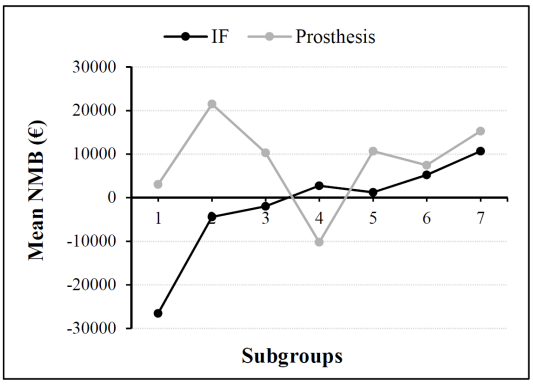
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

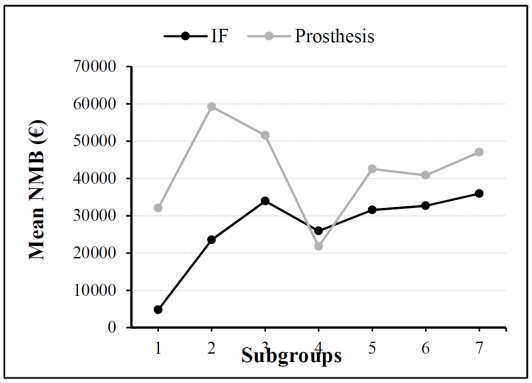
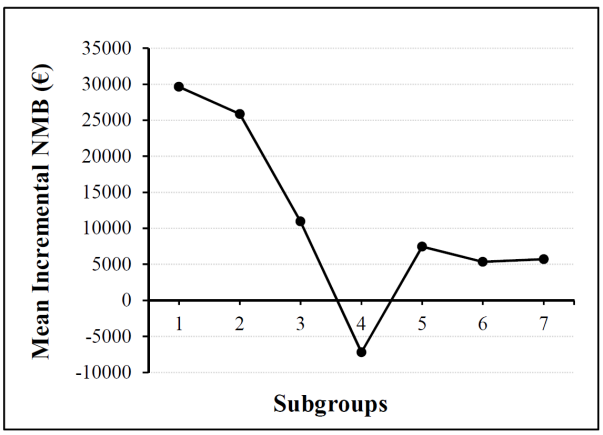


Figure B.6: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of age (θ_h). Subgroups 1 = age 60 to 65, 2 = age 66 to 70, 3 = 71 to 75, 4 = 76 to 80, 5 = 81 to 85, 6 = 86 to 90 and 7 = 91+. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

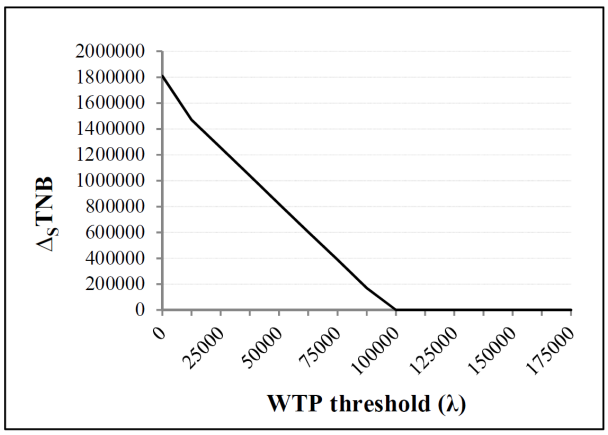
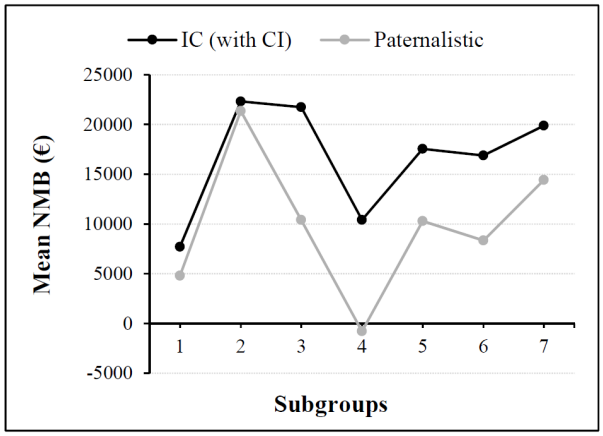
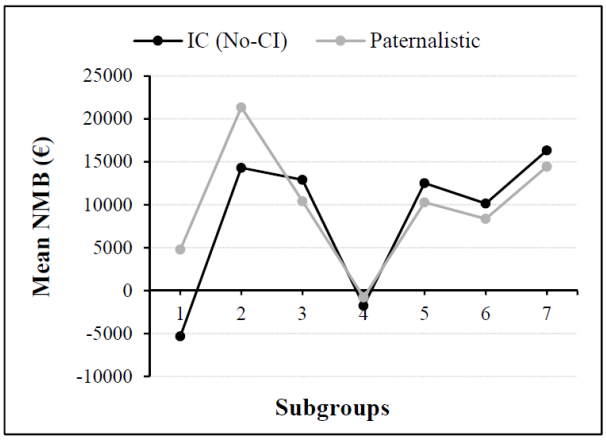


Figure B.7: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of age (θ_h). Subgroups 1 = age 60 to 65, 2 = age 66 to 70, 3 = 71 to 75, 4 = 76 to 80, 5 = 81 to 85, 6 = 86 to 90 and 7 = 91+. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

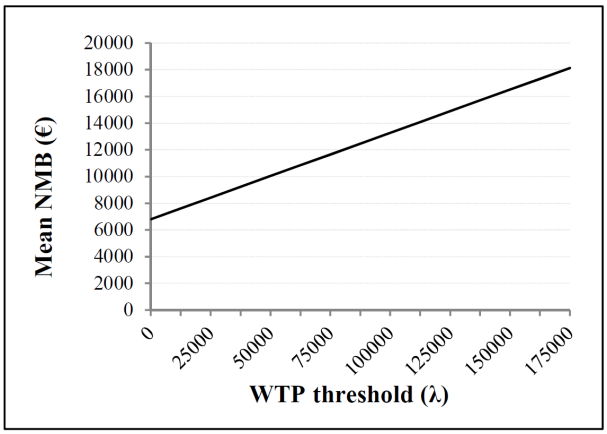
(a) Mean NMB by Subgroup (λ= €25 000)



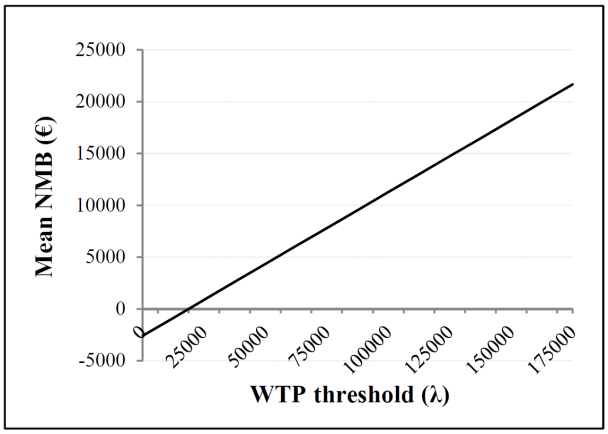
(b) Mean NMB by Subgroup (λ= €25 000)



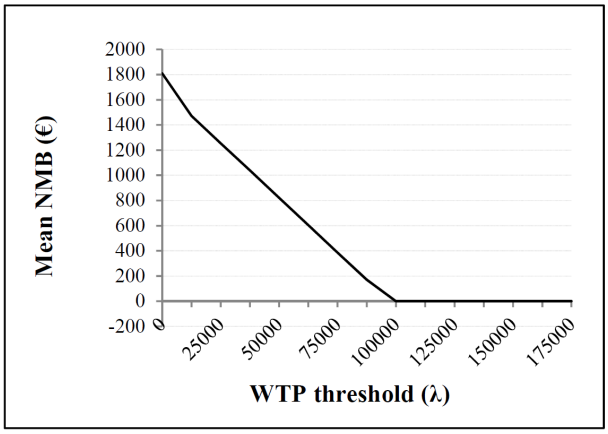
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

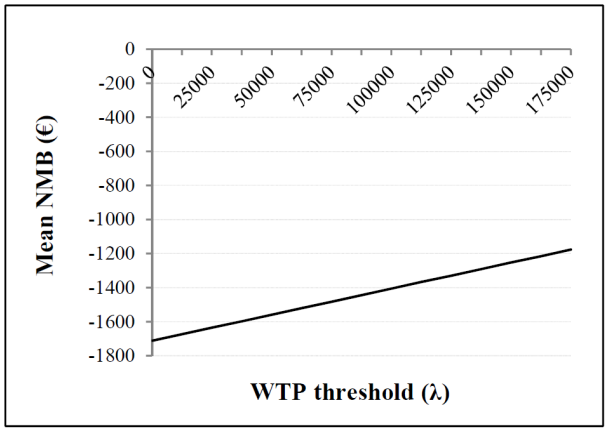
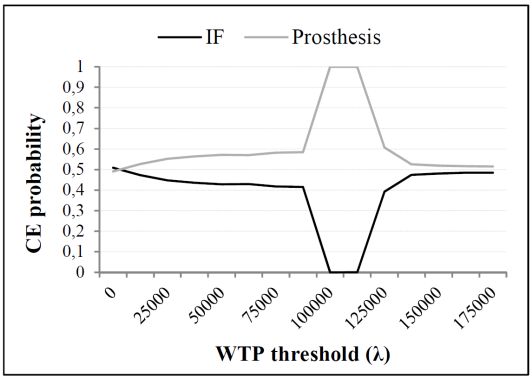
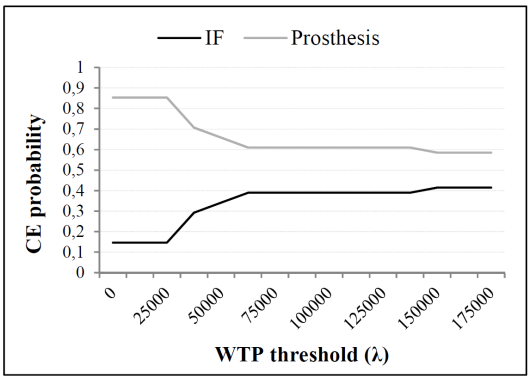


Figure B.8: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of age (θ_h). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

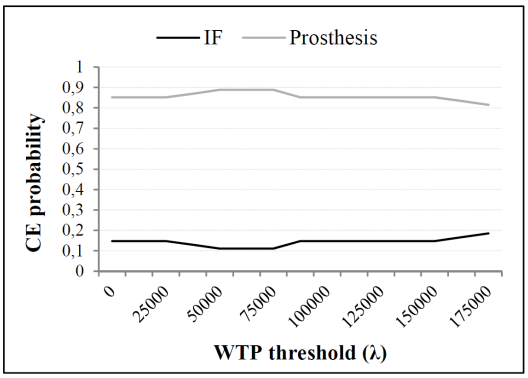
(a) Population CEAC



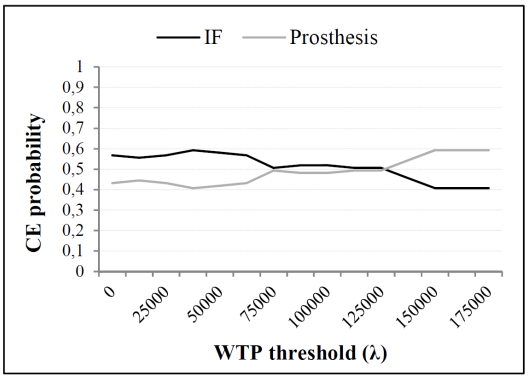
(b) Subgroup 1 = aged between 60 and 65



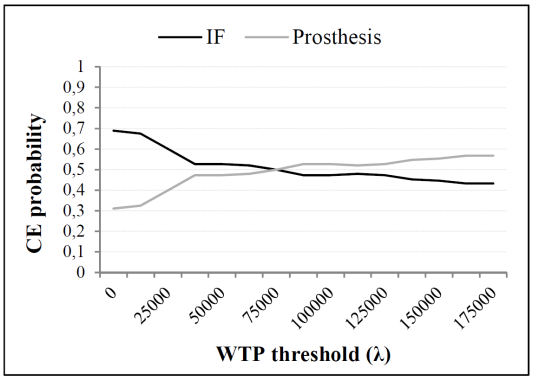
(c) Subgroup 2 = aged between 66 and 70



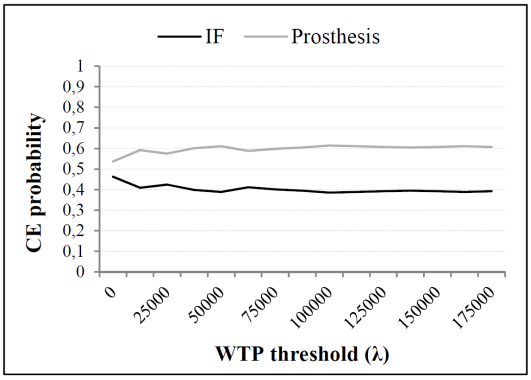
(d) Subgroup 3 = aged between 71 and 75



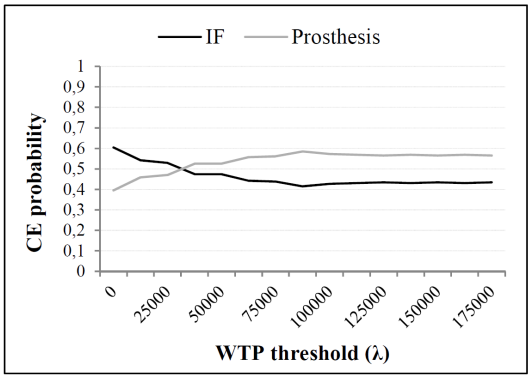
(e) Subgroup 4 = aged between 76 and 80



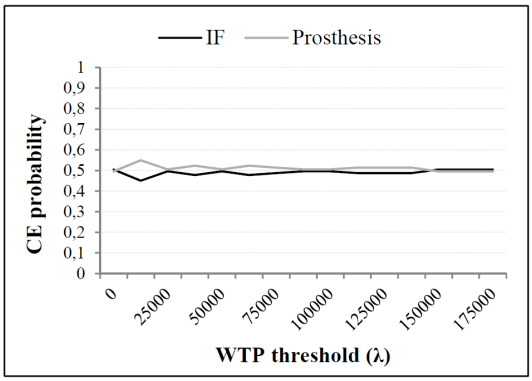
(f) Subgroup 5 = aged between 81 and 85



(g) Subgroup 6 = aged between 86 and 90



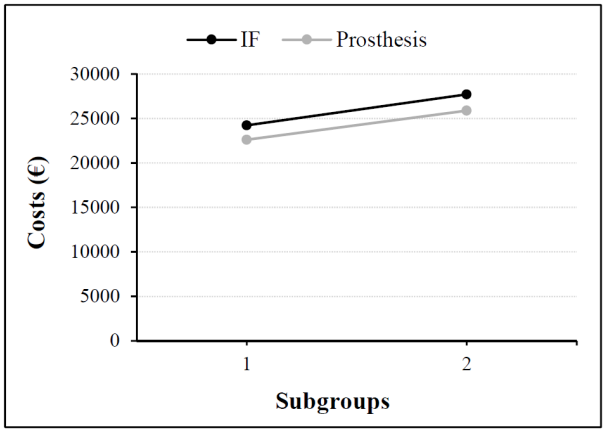
(h) Subgroup 7 = aged 90 +



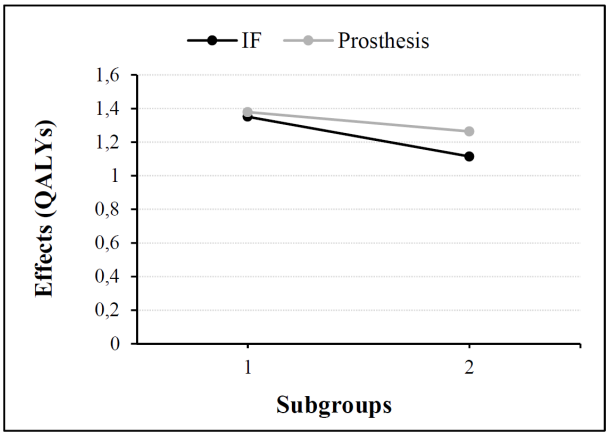
B.4 Gender (θ_i)

Figure B.9: Bootstrapped results of the population sample stratified on the basis of gender (θ_i). Subgroups 1 = males and 2 = females.

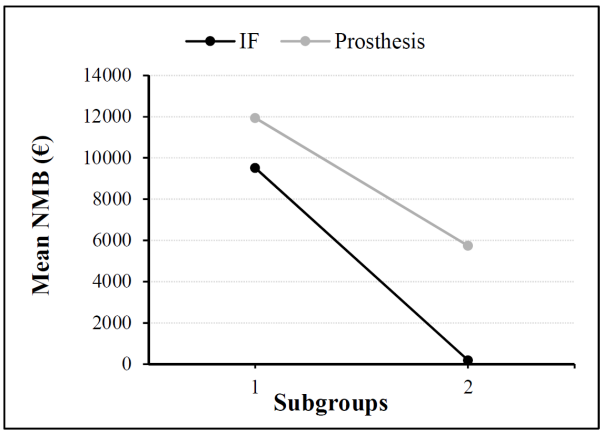
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

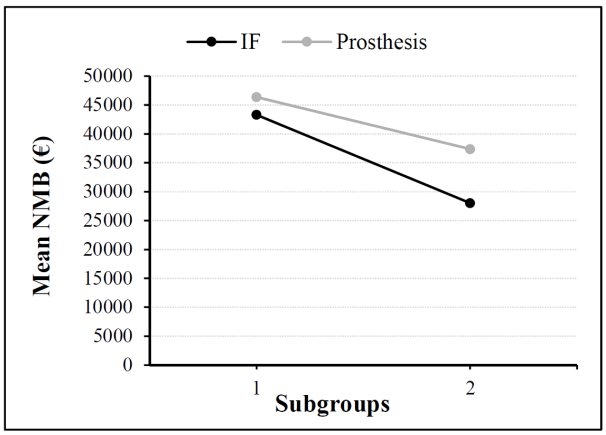
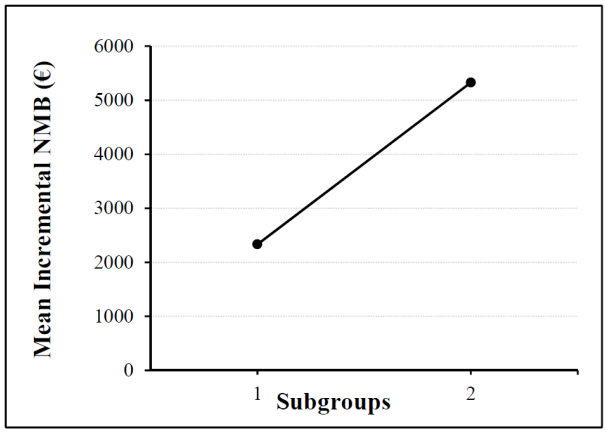


Figure B.10: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of gender (θ_i). Subgroups 1 = males and 2 = females. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

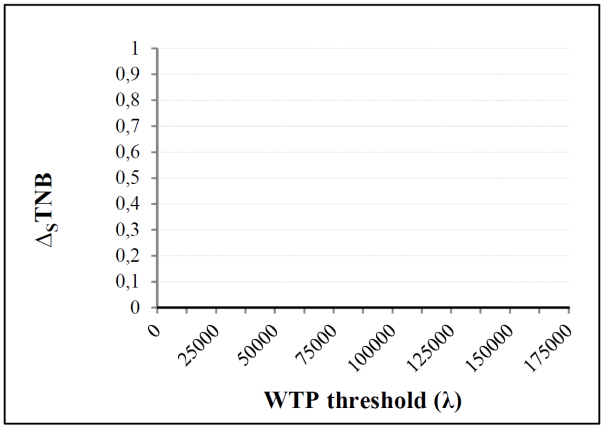
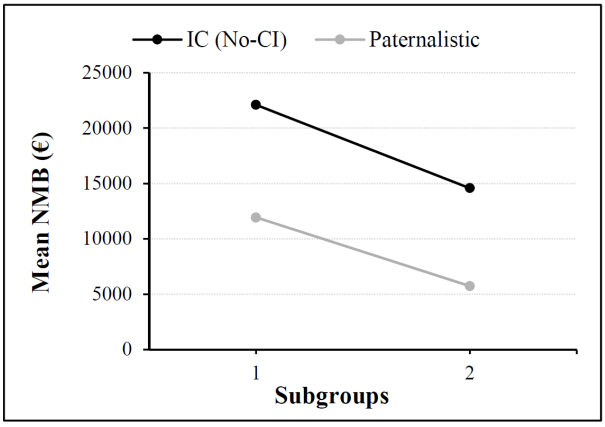
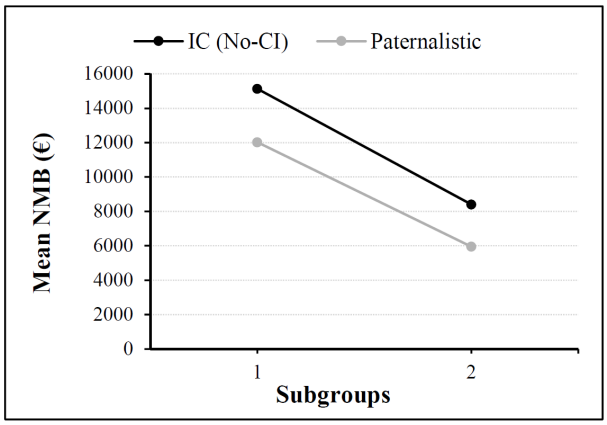


Figure B.11: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of gender (θ_i). Subgroups 1 = males and 2 = females. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

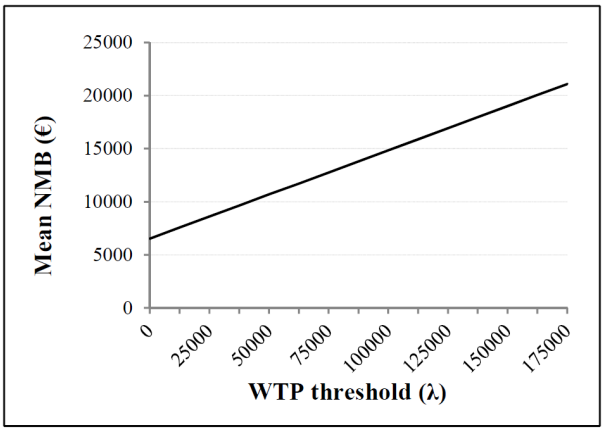
(a) Mean NMB by Subgroup (λ= €25 000)



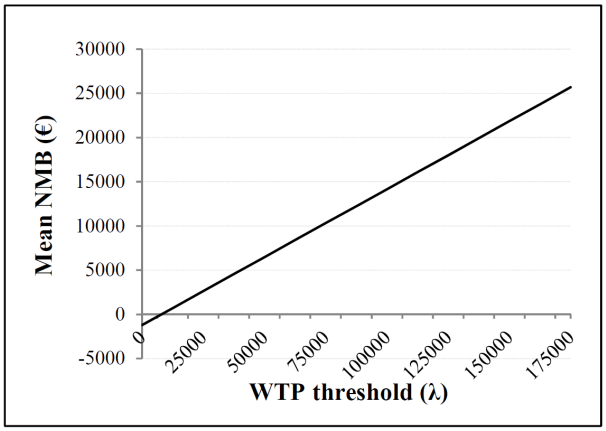
(b) Mean NMB by Subgroup (λ= €25 000)



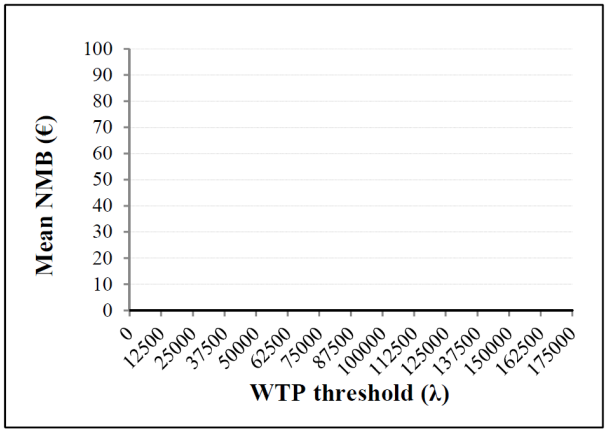
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

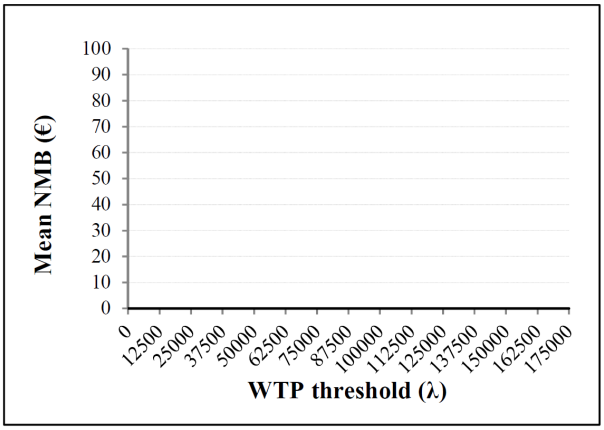
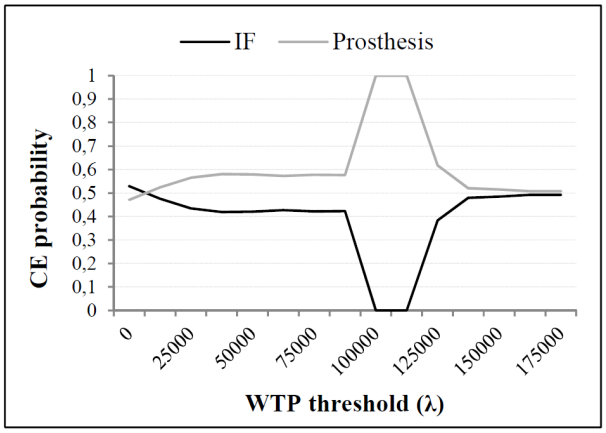
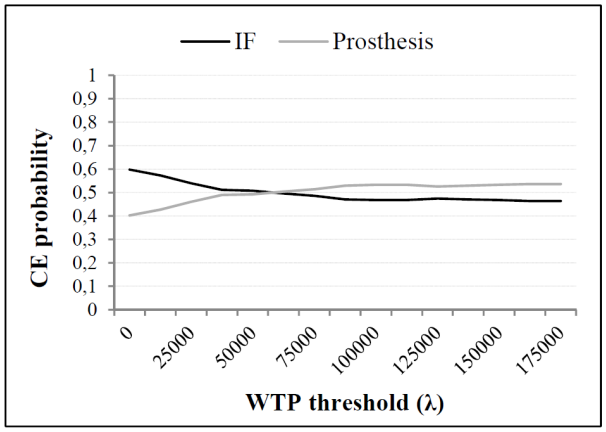


Figure B.12: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of gender (θ_i). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

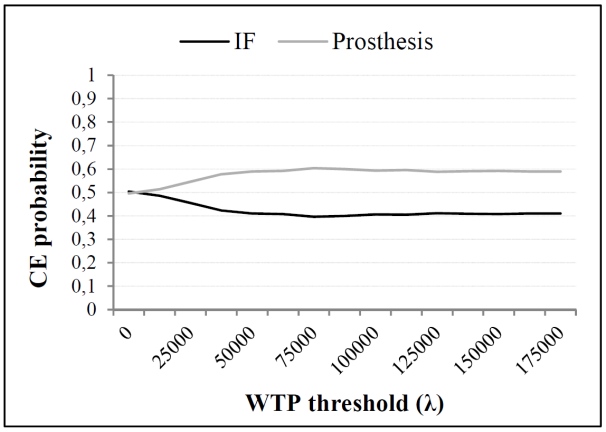
(a) Population CEAC



(b) Subgroup 1 = males



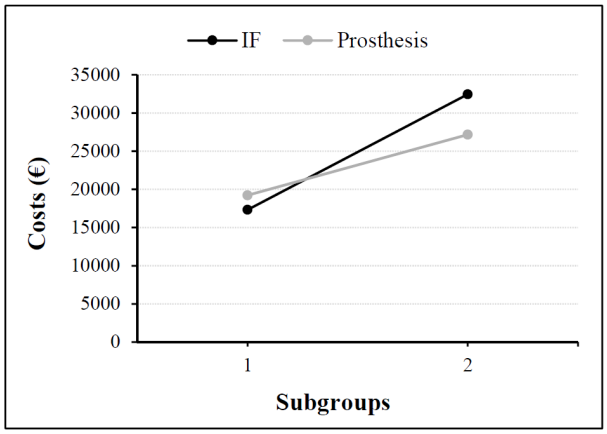
(c) Subgroup 2 = females



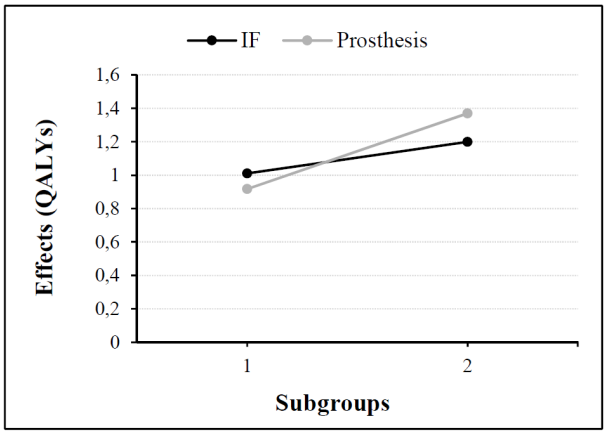
B.5 Dementia (θ_j)

Figure B.13: Bootstrapped results of the population sample stratified on the basis dementia (θ_j). Subgroups 1 = with dementia and 2 = no dementia.

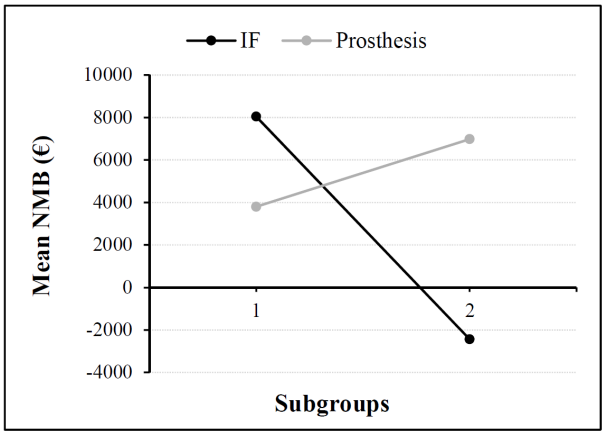
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

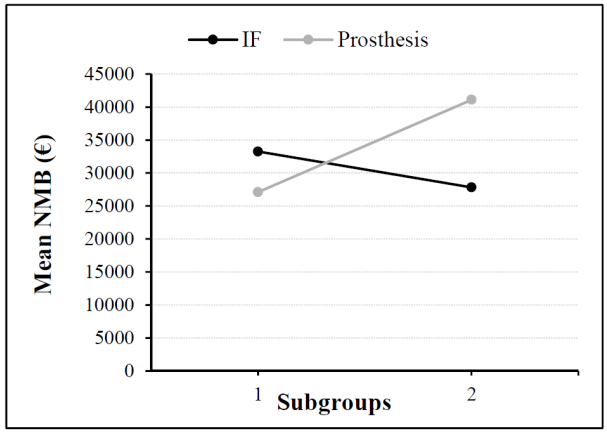
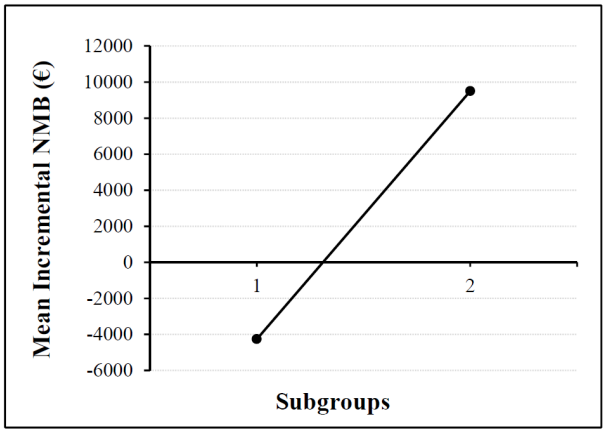


Figure B.14: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of dementia (θ_j). Subgroups 1 = with dementia and 2 = no dementia. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

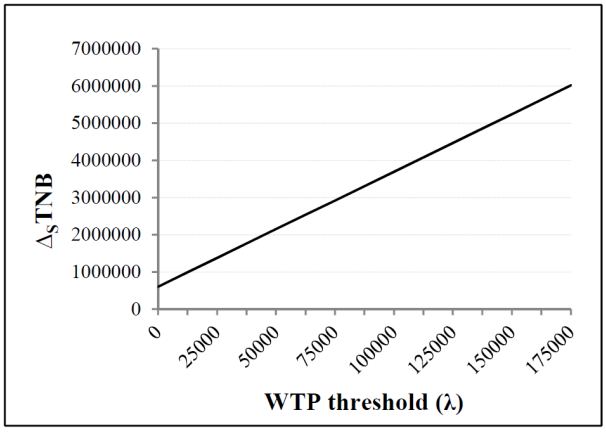
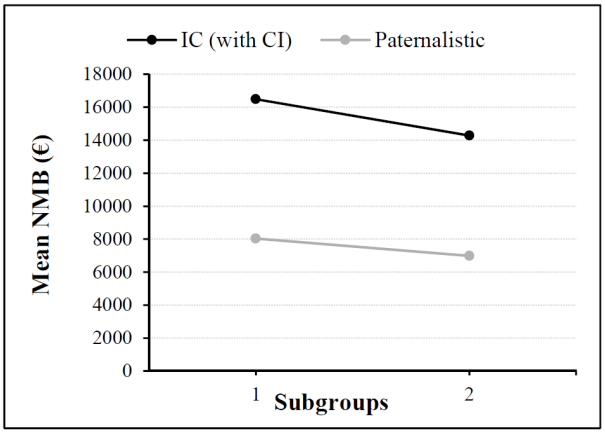
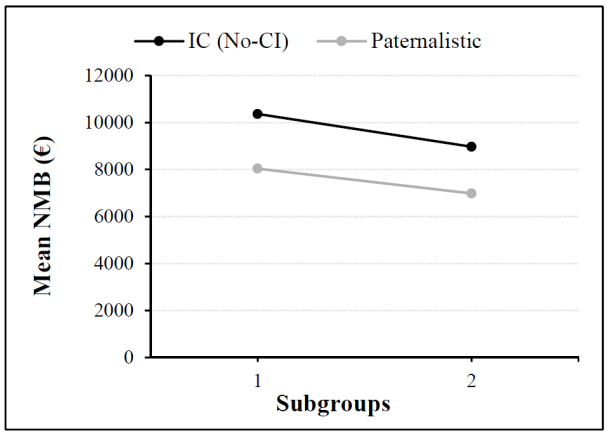


Figure B.15: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of dementia (θ_j). Subgroups 1 = with dementia and 2 = no dementia. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

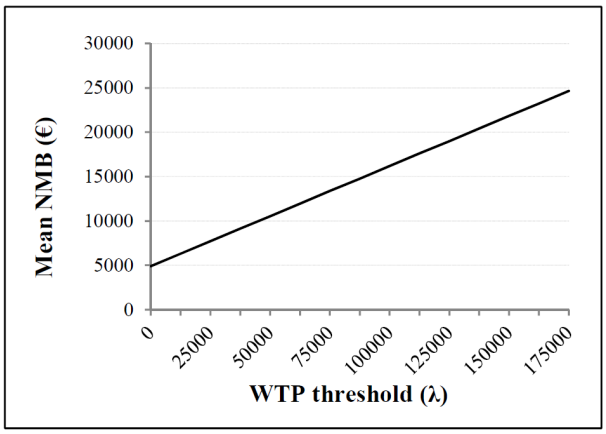
(a) Mean NMB by Subgroup (λ= €25 000)



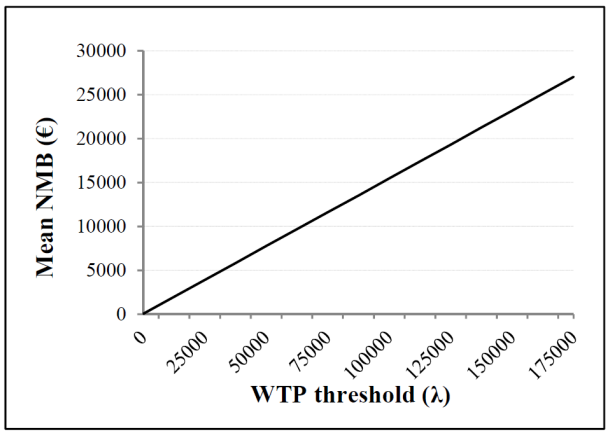
(b) Mean NMB by Subgroup (λ= €25 000)



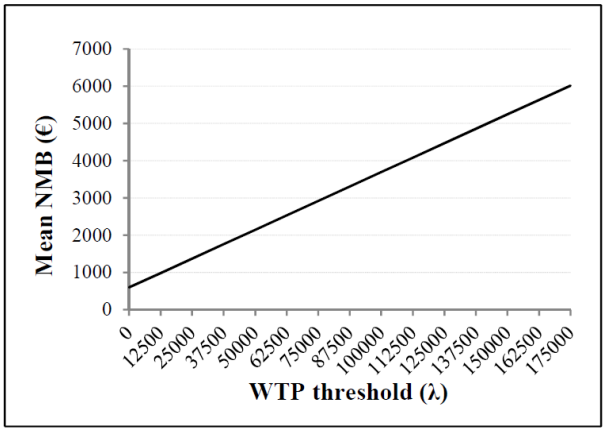
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

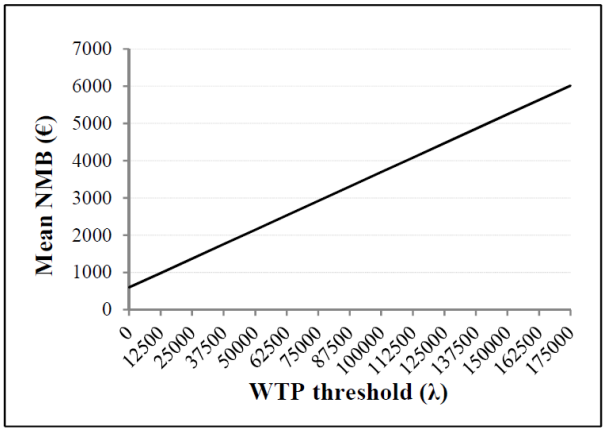
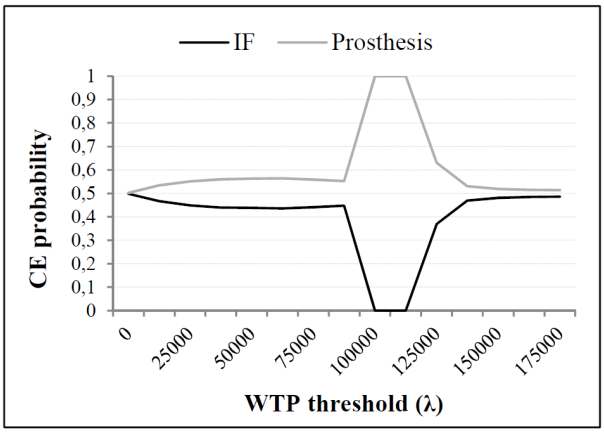
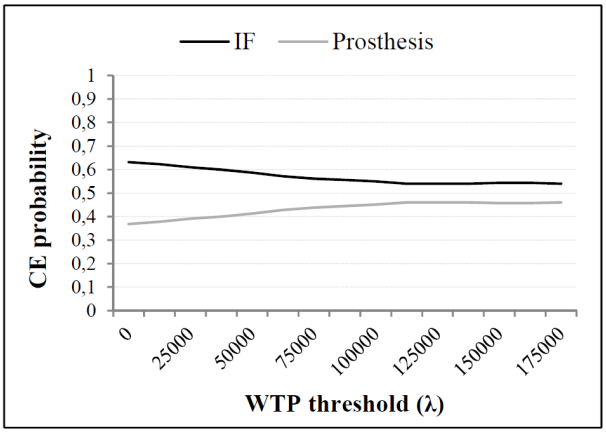


Figure B.16: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of dementia (θ_j). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

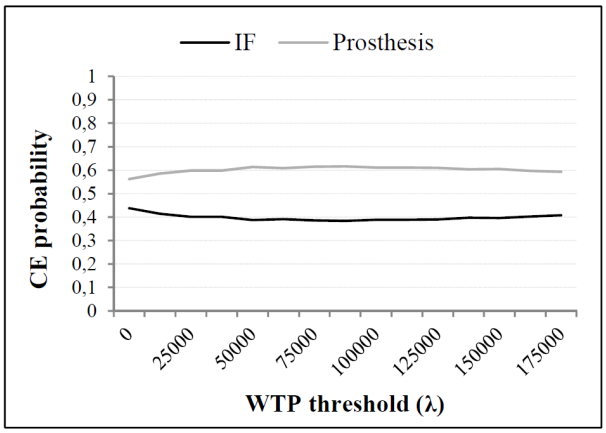
(a) Population CEAC



(b) Subgroup 1 = with dementia



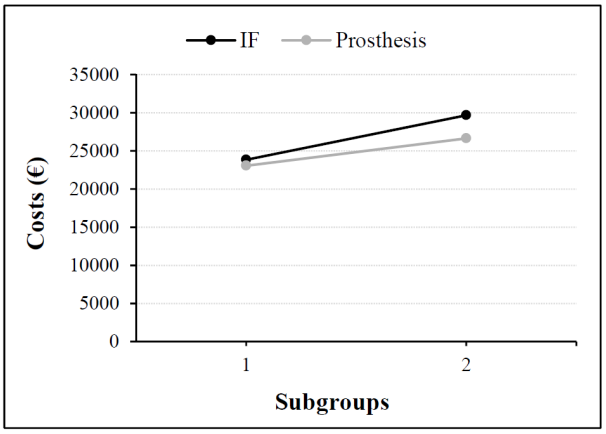
(c) Subgroup 2 = no dementia



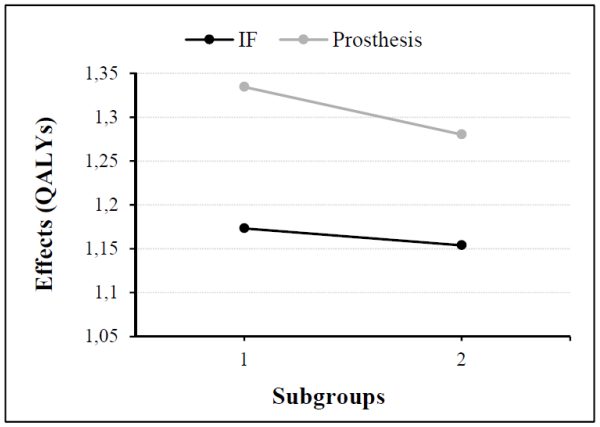
B.6 Anaemia (θ_k)

Figure B.17: Bootstrapped results of the population sample stratified on the basis of anaemia (θ_k). Subgroups 1 = with anaemia and 2 = no anaemia.

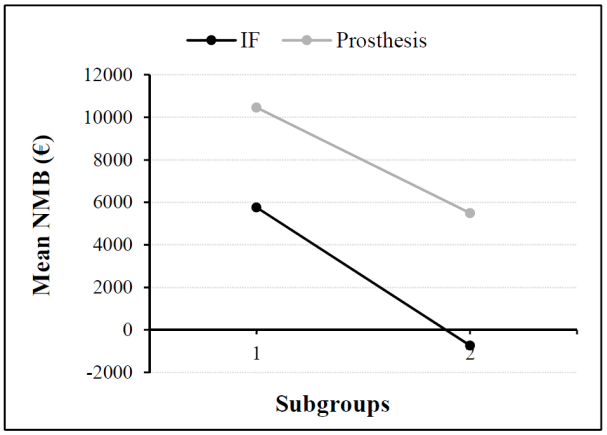
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

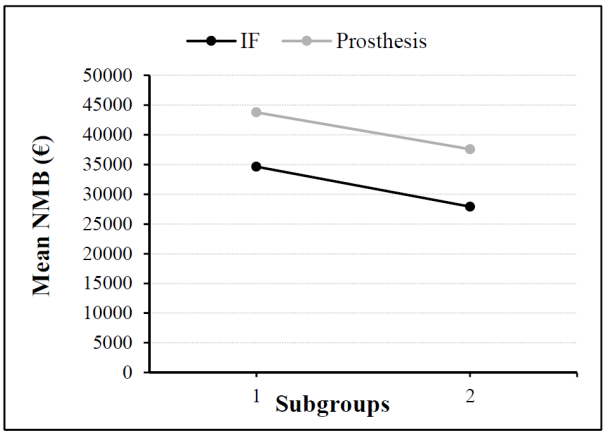
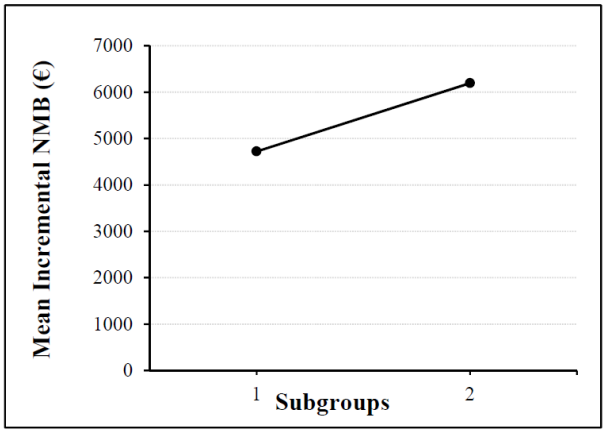


Figure B.18: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of anaemia (θ_k). Subgroups 1 = with anaemia and 2 = no anaemia. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

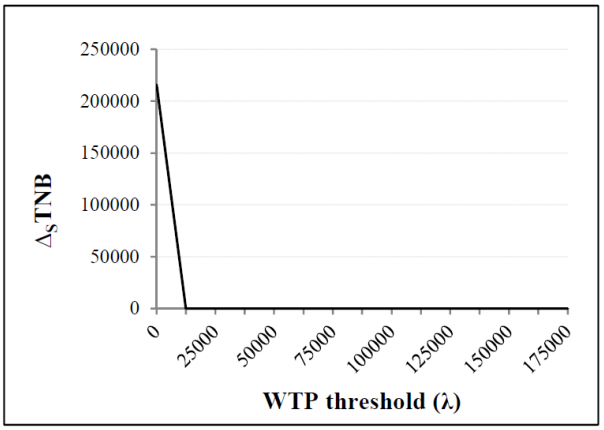
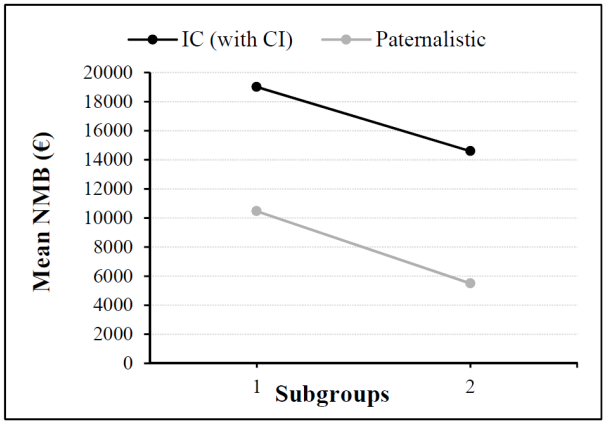
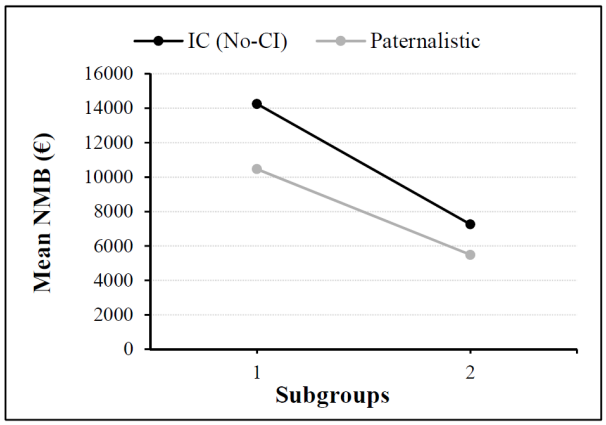


Figure B.19: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of anaemia (θ_k). Subgroups 1 = with anaemia and 2 = no anaemia. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

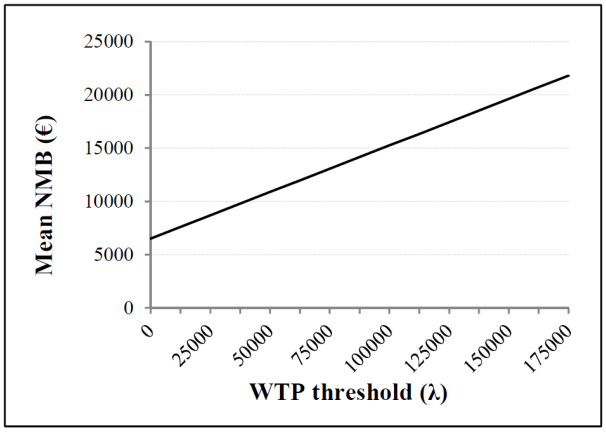
(a) Mean NMB by Subgroup (λ= €25 000)



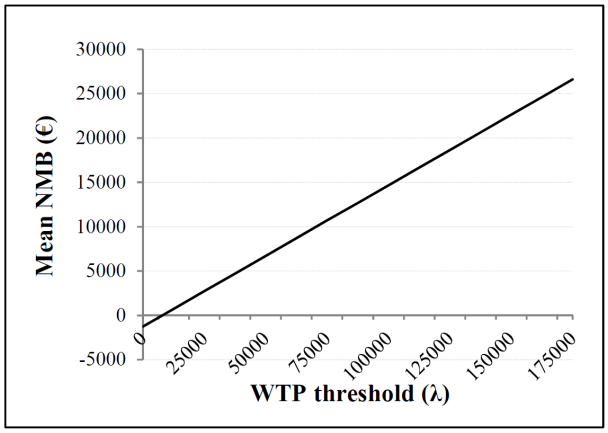
(b) Mean NMB by Subgroup (λ= €25 000)



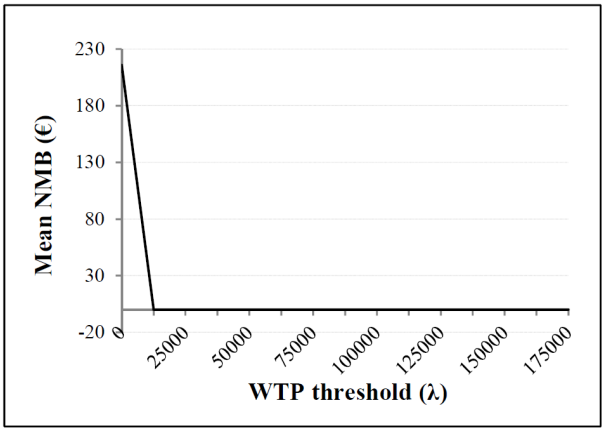
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

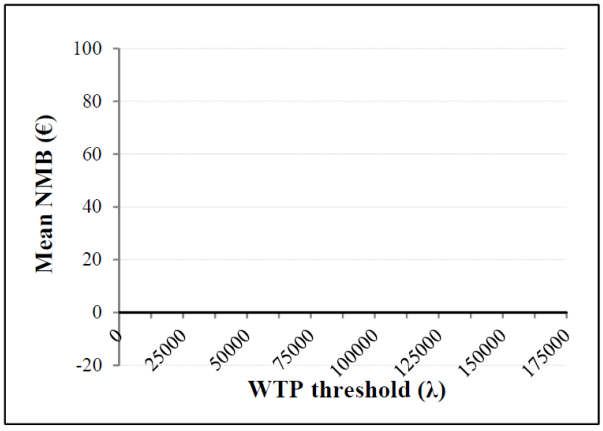
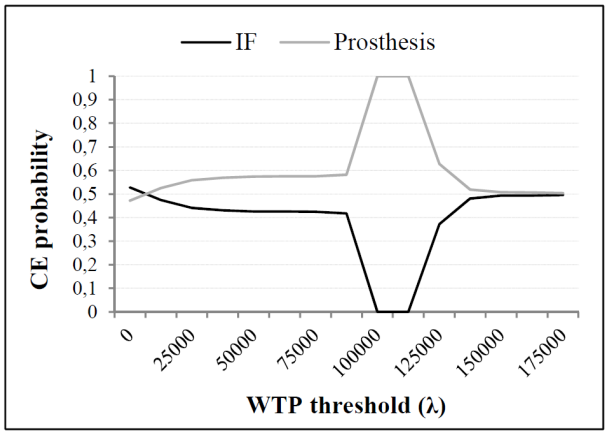
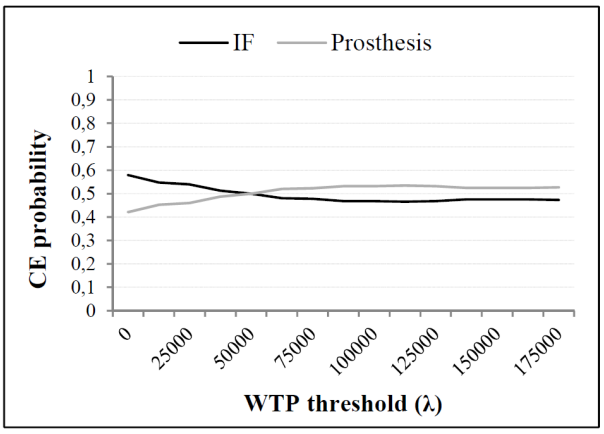


Figure B.20: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of anaemia (θ_k). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

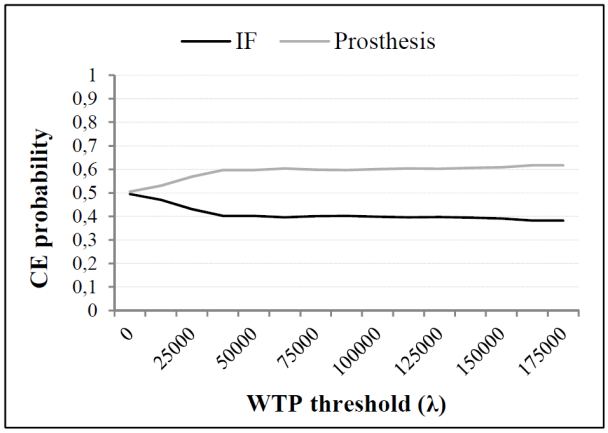
(a) Population CEAC



(b) Subgroup 1 = with anaemia



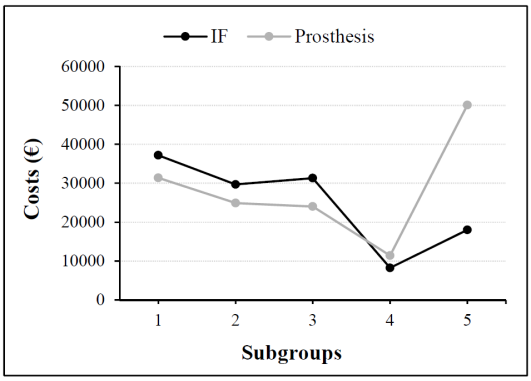
(c) Subgroup 2 = no anaemia



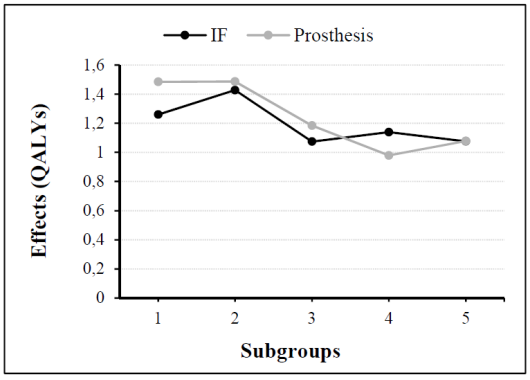
B.7 Where the injury occurred (θ_l)

Figure B.21: Bootstrapped results of the population sample stratified on the basis injury occurred (θ_l). Subgroups 1 = outdoors, 2 = inside (not home), 3 = inside home, 4 = nursing home and 5 = hospital.

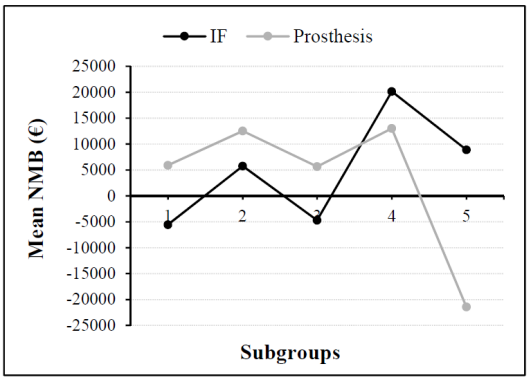
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

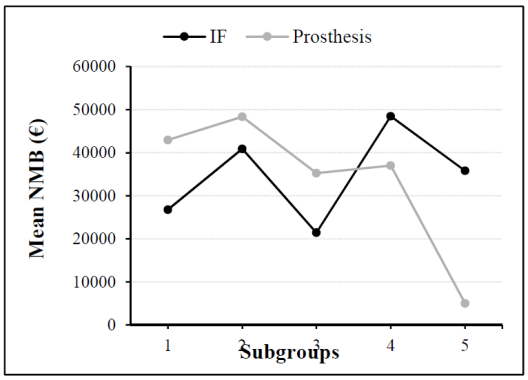
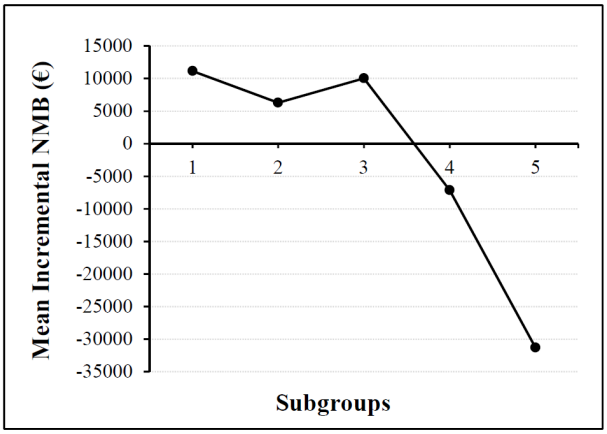


Figure B.22: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of injury occurred (θ_l). Subgroups 1 = outdoors, 2 = inside (not home), 3 = inside home, 4 = nursing home and 5 = hospital. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

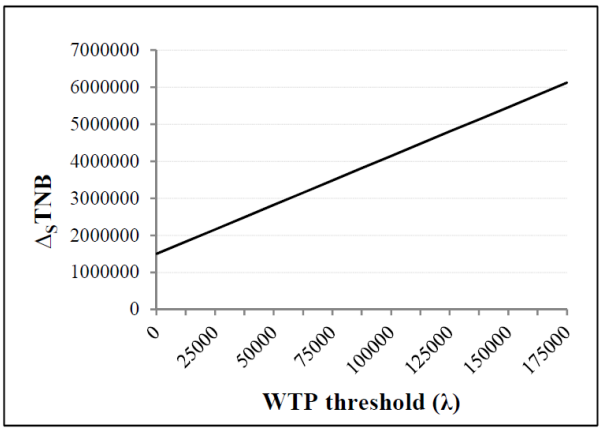
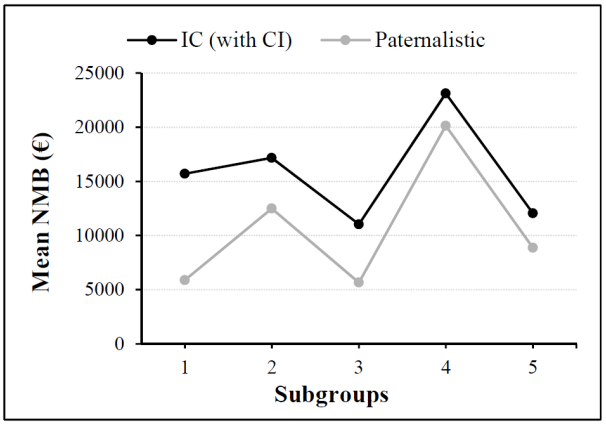
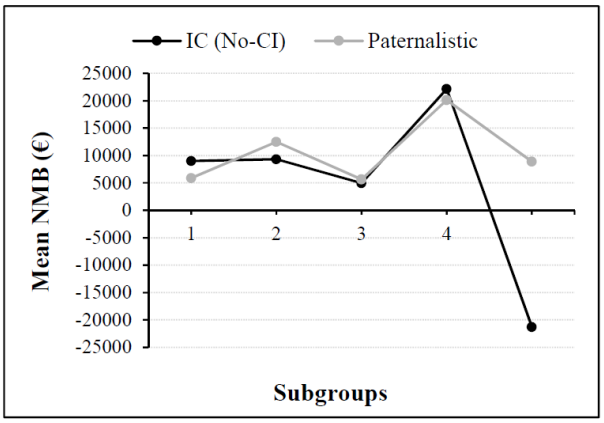


Figure B.23: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of injury occured (θ_l). Subgroups 1 = outdoors, 2 = inside (not home), 3 = inside home, 4 = nursing home and 5 = hospital. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

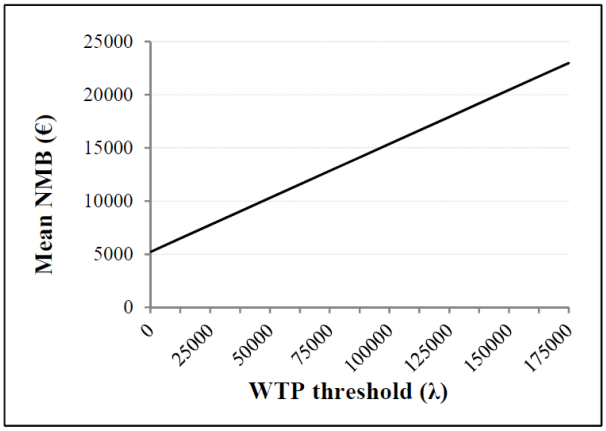
(a) Mean NMB by Subgroup (λ= €25 000)



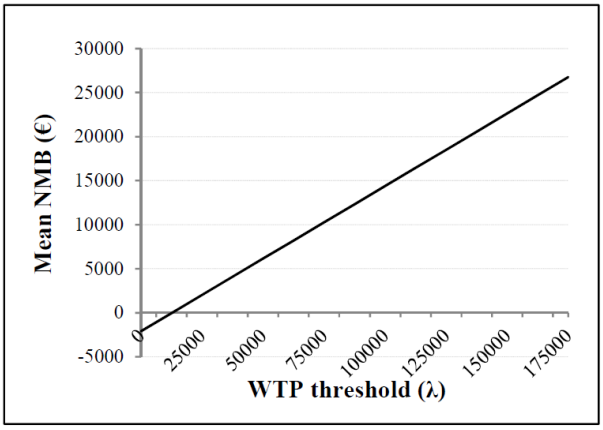
(b) Mean NMB by Subgroup (λ= €25 000)



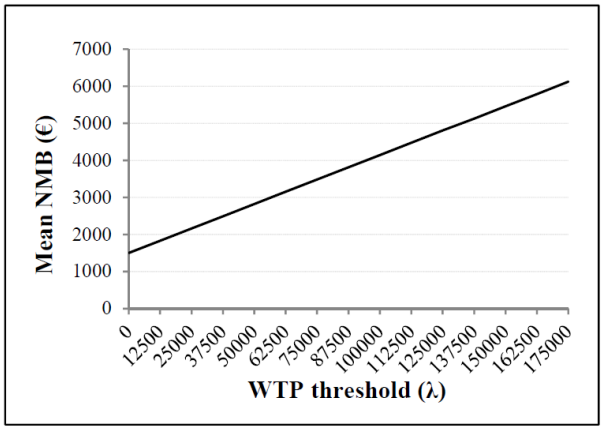
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

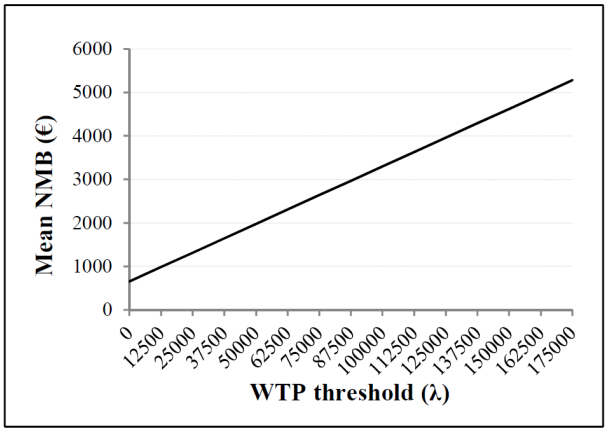
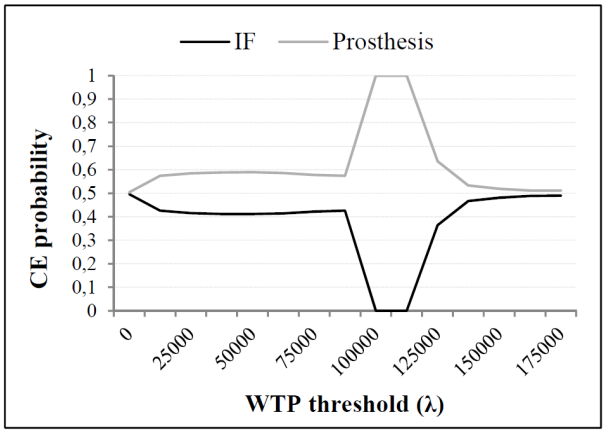
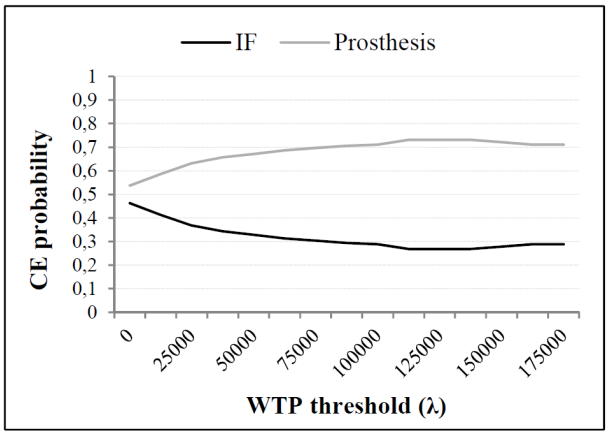


Figure B.24: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of injury occurred (θ_l). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

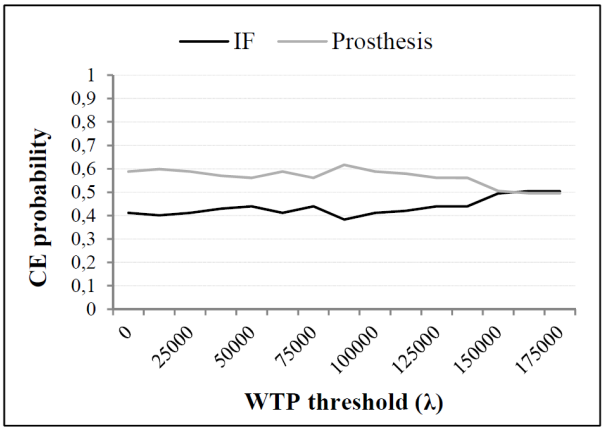
(a) Population CEAC



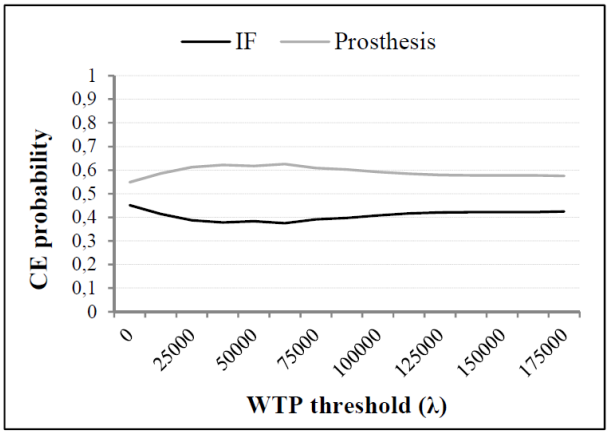
(b) Subgroup 1 = outdoors



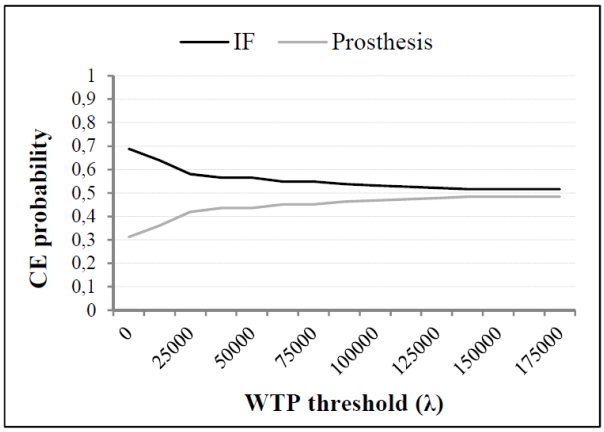
(c) Subgroup 2 = inside (not home=



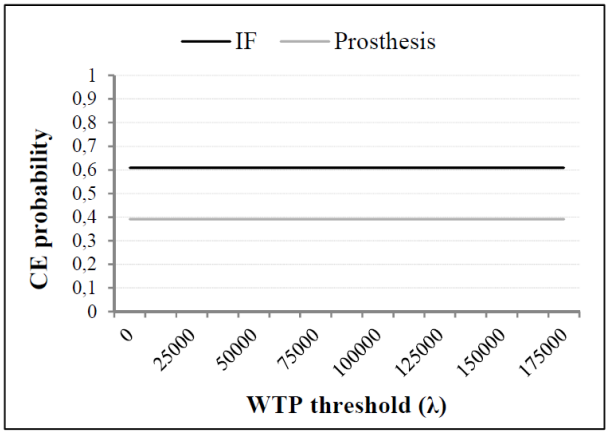
(d) Subgroup 3 = inside home



(e) Subgroup 4 = nursing home



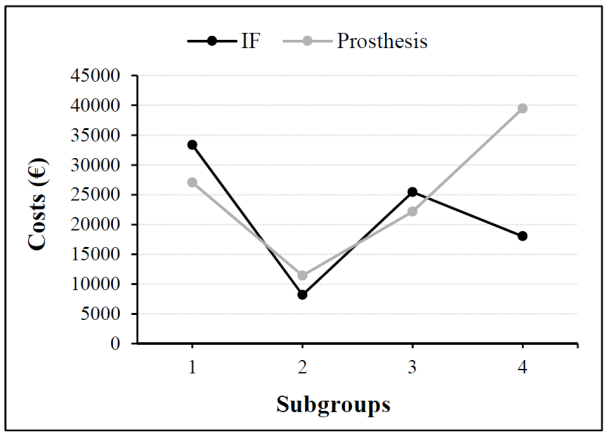
(f) Subgroup 5 = hospital



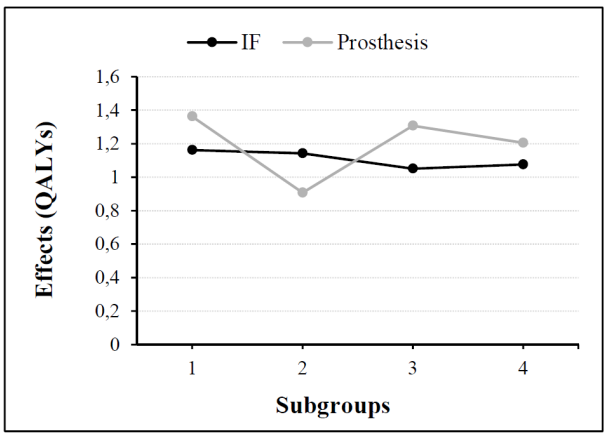
B.8 Living (θ_m)

Figure B.25: Bootstrapped results of the population sample stratified on the basis of living (θ_m). Subgroups 1 = home, 2 = nursing home, 3 = care home, 4 = hospital.

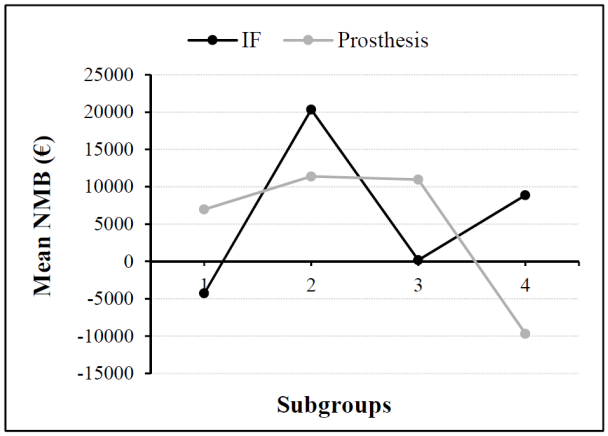
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

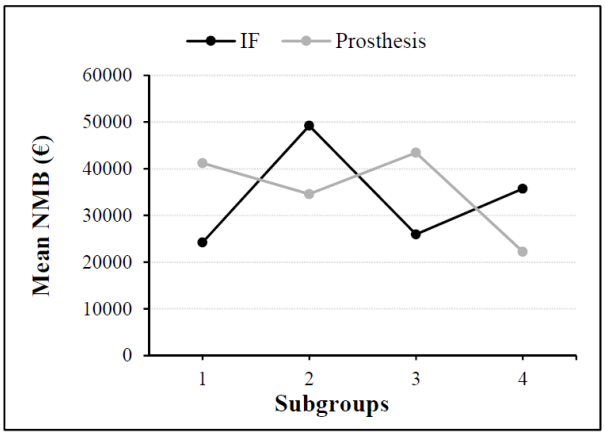
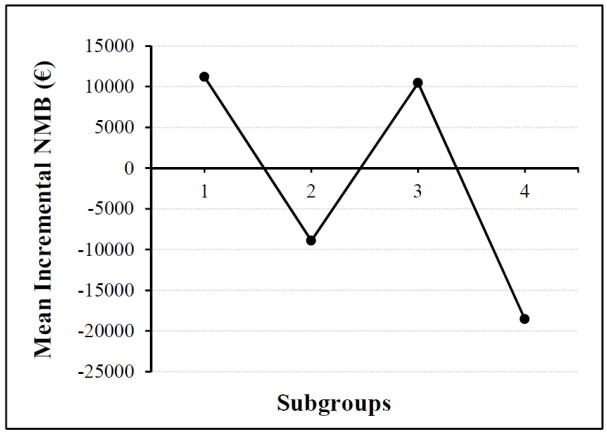


Figure B.26: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of living (θ_m). Subgroups 1 = home, 2 = nursing home, 3 = care home, 4 = hospital. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

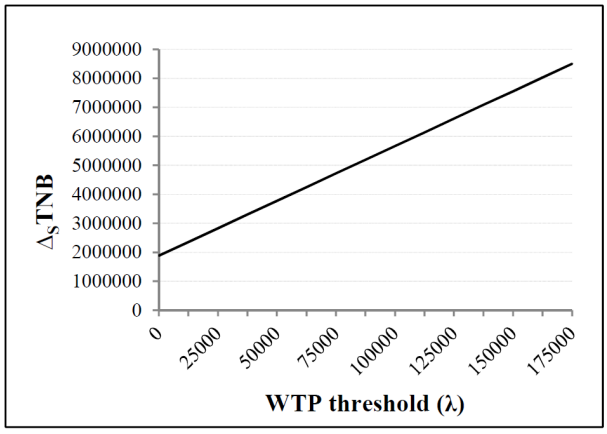
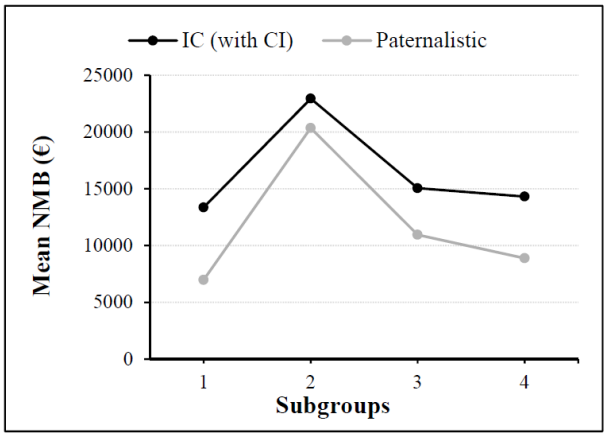
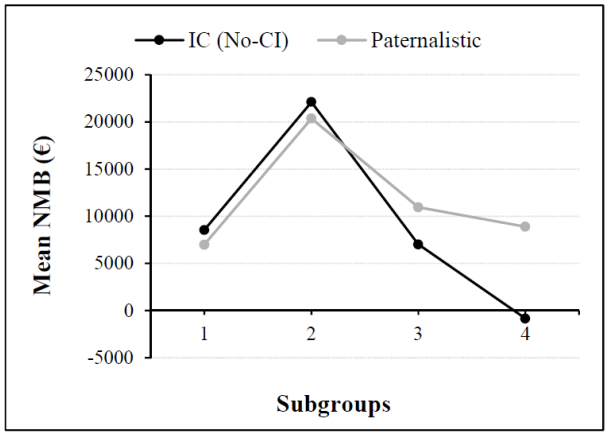


Figure B.27: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of living situation (θ_m). Subgroups 1 = home, 2 = nursing home, 3 = care home, 4 = hospital. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

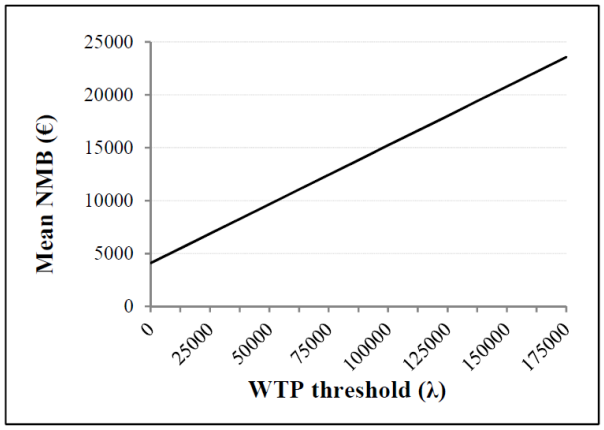
(a) Mean NMB by Subgroup (λ= €25 000)



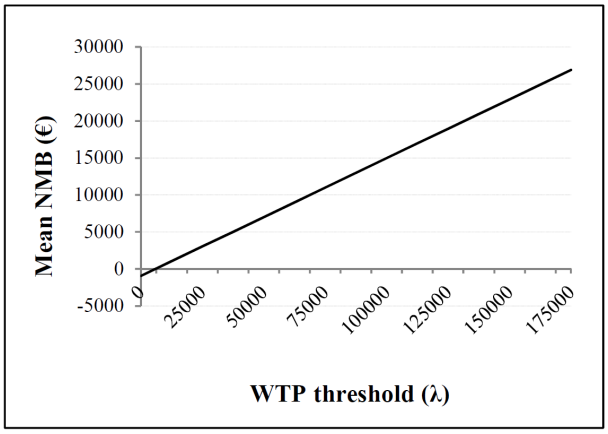
(b) Mean NMB by Subgroup (λ= €25 000)



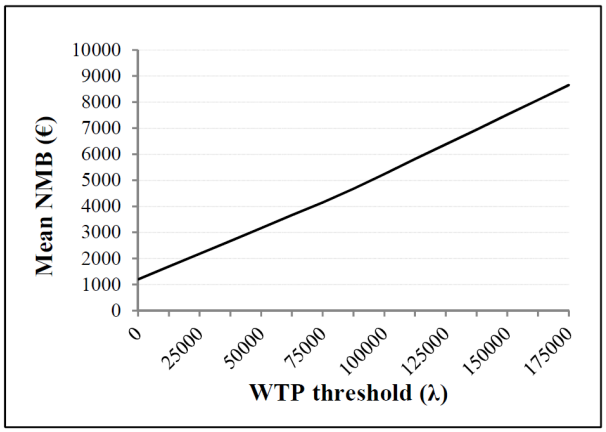
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

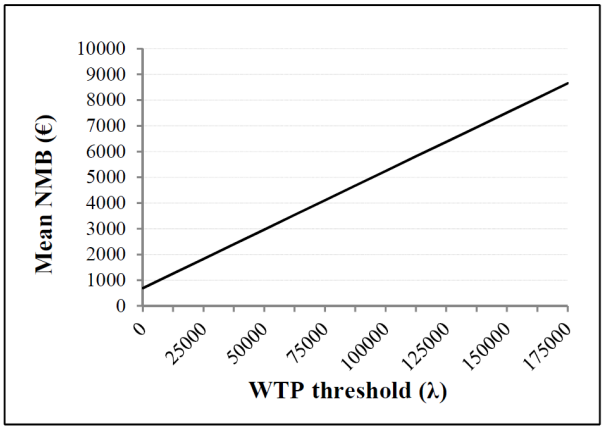
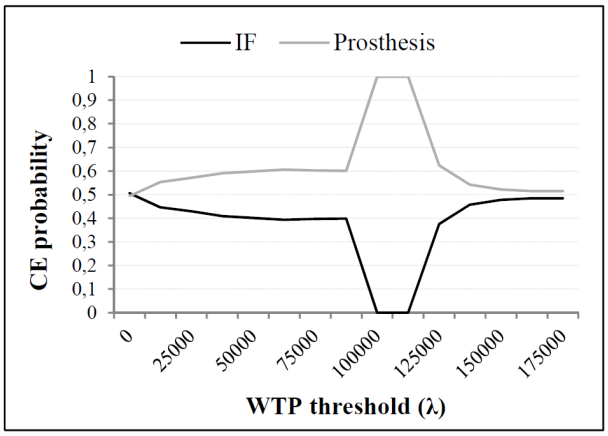
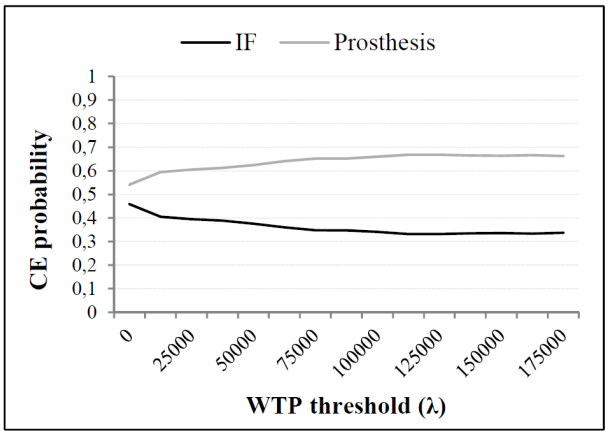


Figure B.28: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of living situation (θ_m). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

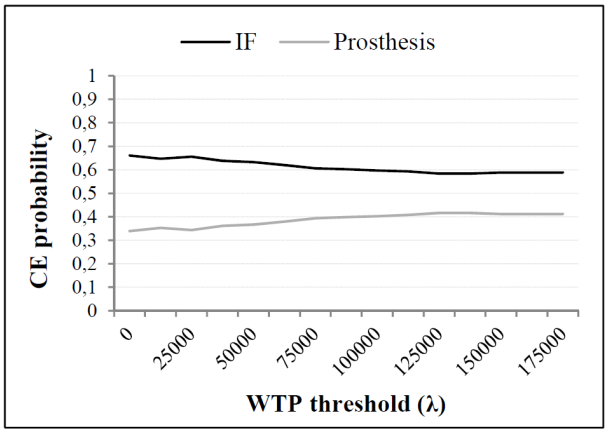
(a) Population CEAC



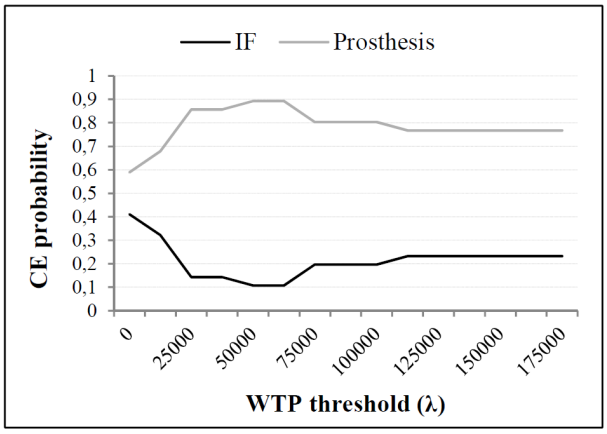
(b) Subgroup 1 = home



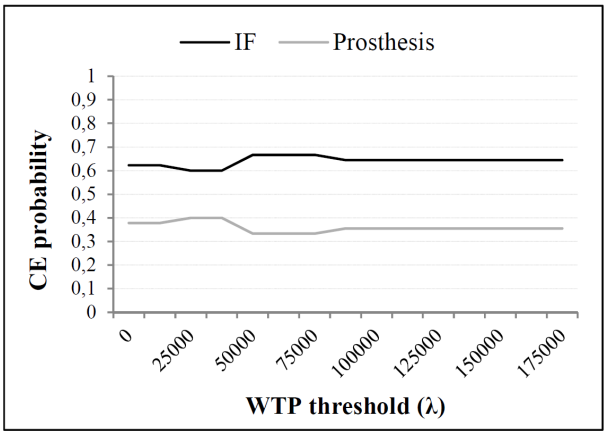
(c) Subgroup 2 = nursing home



(d) Subgroup 3 = care home



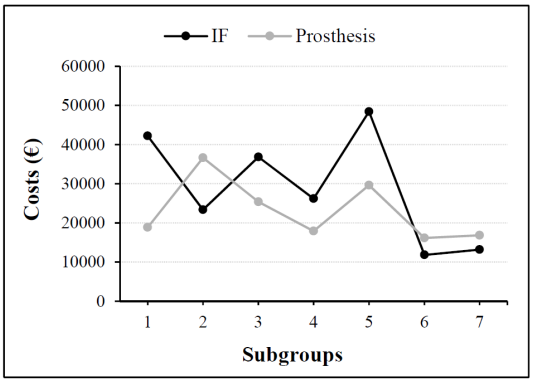
(e) Subgroup 4 = hospital



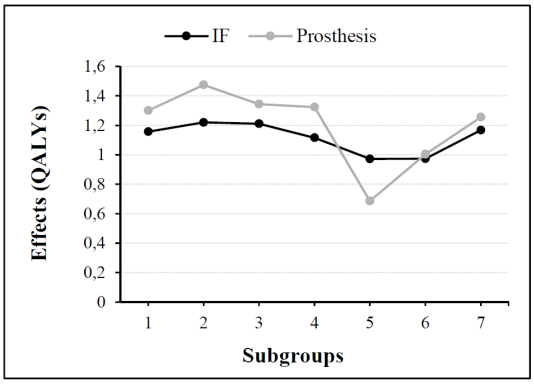
B.9 Age and Dementia (θ_{gj})

Figure B.29: Bootstrapped results of the population sample stratified on the basis of age & dem. (θ_{gj}). Subgroups 1 = age 60-70 no dem., 2 = age 71-80 no dem., 3 = age 81-90 no dem., 4 = age 90+ no dem., 5 = age 71-80 with dem., 6 = age 81-90 with dem. and 7 = age 91+ with dem.

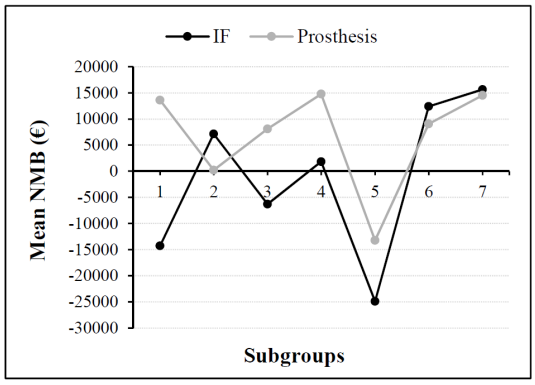
(a) Mean Costs by Subgroup



(b) Mean Effects by Subgroup



(c) Mean NMB by Subgroup (λ=€25 000)



(d) Mean NMB by Subgroup (λ=€50 000)

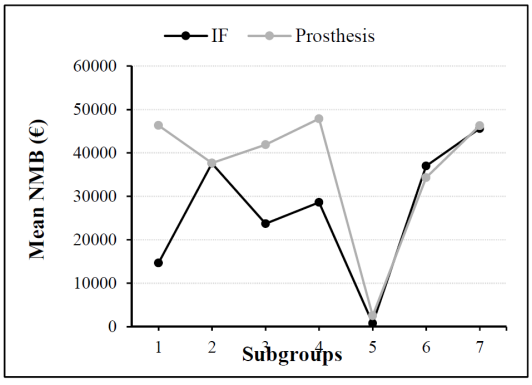
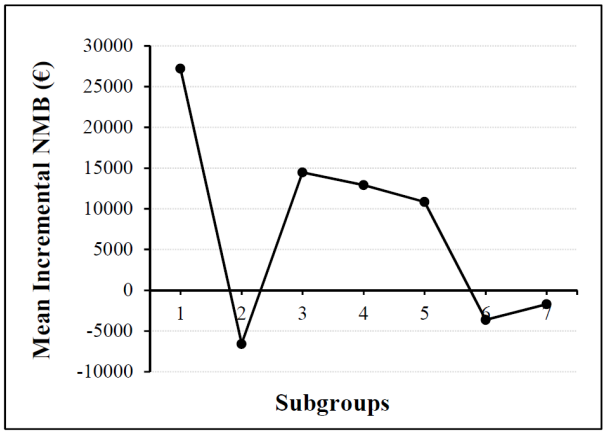


Figure B.30: Bootstrapped results of a Stratified Analysis (SA) when the population sample is stratified on the basis of age & dem. (θ_{gj}). Subgroups 1 = age 60-70 no dem., 2 = age 71-80 no dem., 3 = age 81-90 no dem., 4 = age 90+ no dem., 5 = age 71-80 with dem., 6 = age 81-90 with dem. and 7 = age 91+ with dem. The (b) figure is obtained from only one bootstrapped re-sample of a 1000 iterations.

(a) Mean incr. NMB by Subgroup (λ= €25 000)



(b) Total NMB gained (Δ_STNB) at different WTP

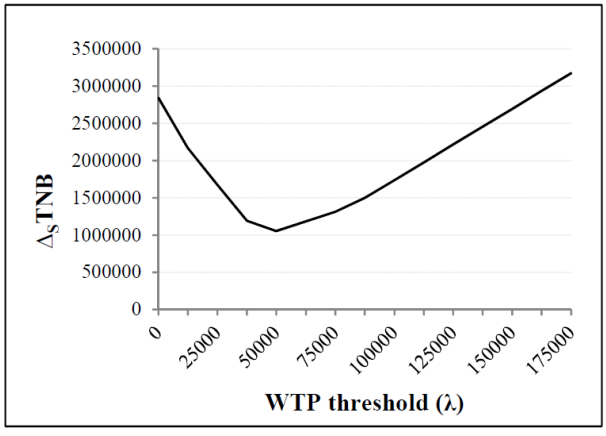
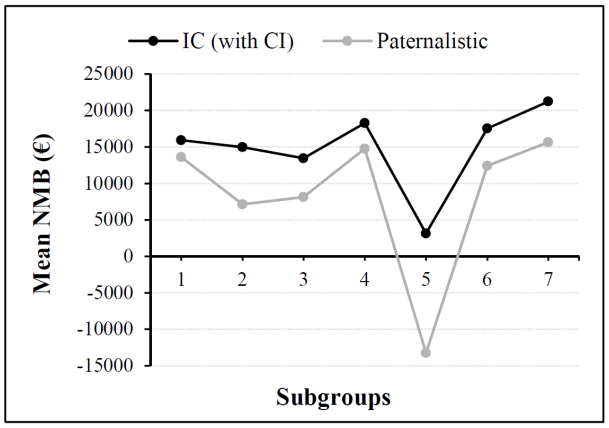
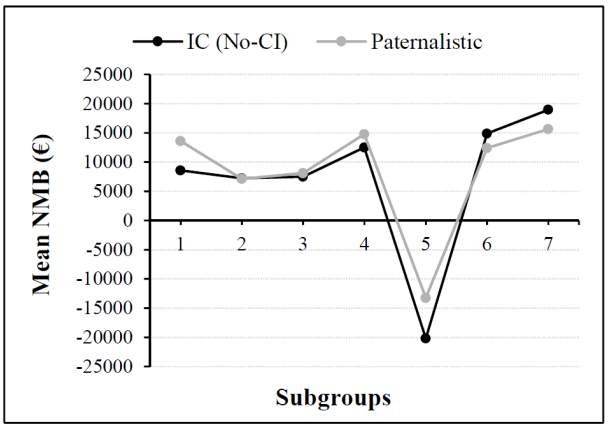


Figure B.31: Bootstrapped results of the EVIC analysis when the population sample is stratified on the basis of age and dementia (θ_{gj}). Subgroups 1 = age 60-70 no dem., 2 = age 71-80 no dem., 3 = age 81-90 no dem., 4 = age 90+ no dem., 5 = age 71-80 with dem., 6 = age 81-90 with dem. and 7 = age 91+ with dem. The results with cost-internalization are presented to the left and those with no cost-internalization are presented to the right. Figures (c) to (f) are obtained from only one bootstrapped re-sample of a 1000 iterations.

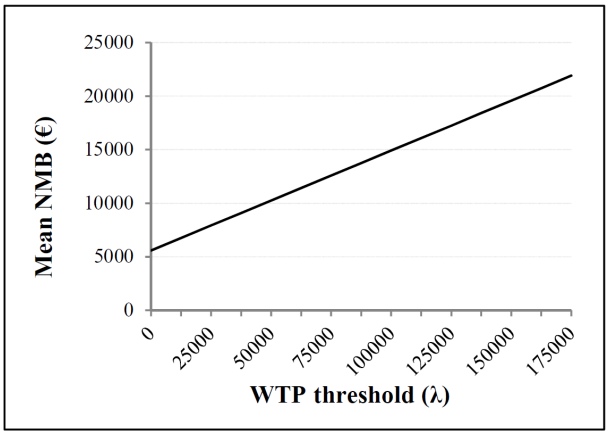
(a) Mean NMB by Subgroup (λ= €25 000)



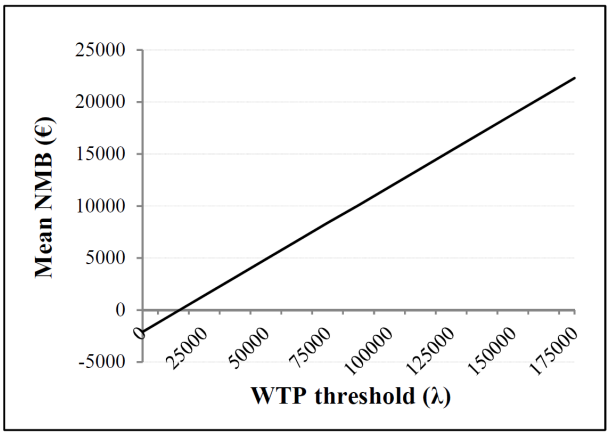
(b) Mean NMB by Subgroup (λ= €25 000)



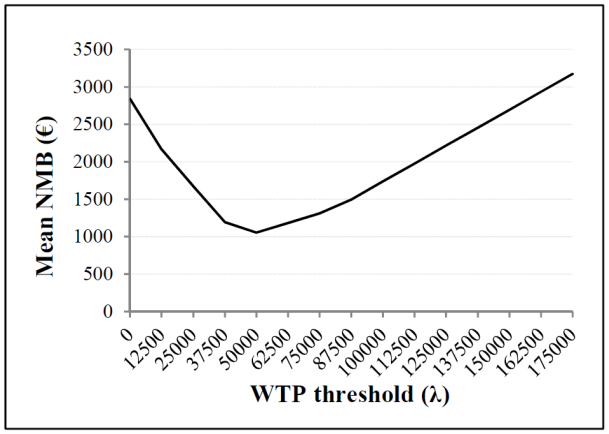
(c) Mean EVIC with CI at different WTP



(d) Mean EVIC with no CI at different WTP



(e) Mean parameter-specific EVIC with CI at different WTP



(f) Mean parameter-specific EVIC with no CI at different WTP

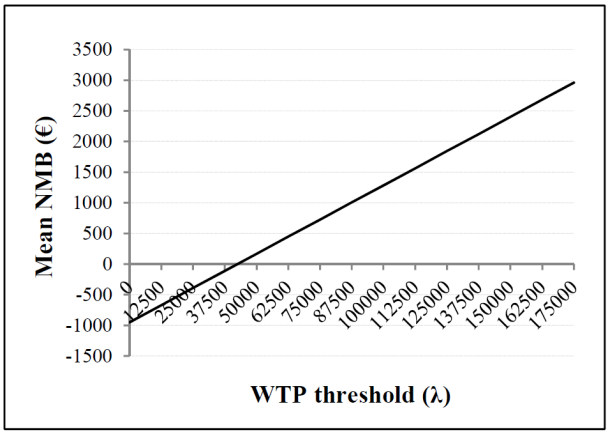
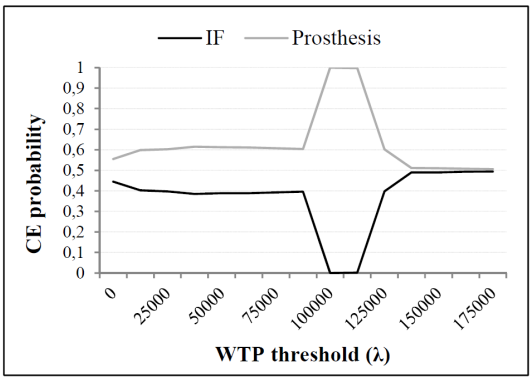
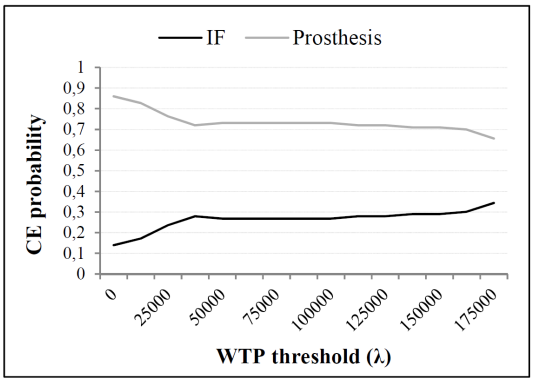


Figure B.32: Cost-effectiveness acceptability curves (CEACs) for the population and subgroups, when the population sample is stratified on the basis of age and dementia (θ_{gj}). Results are obtained from only one bootstrapped re-sample of a 1000 iterations.

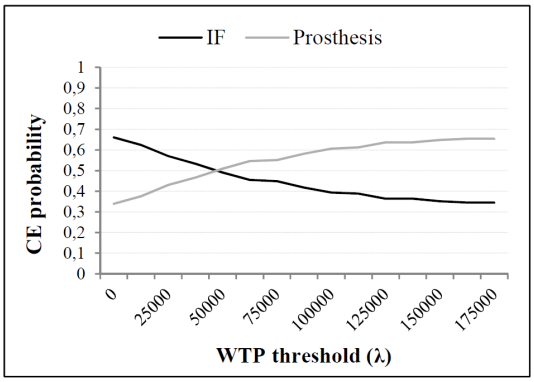
(a) Population CEAC



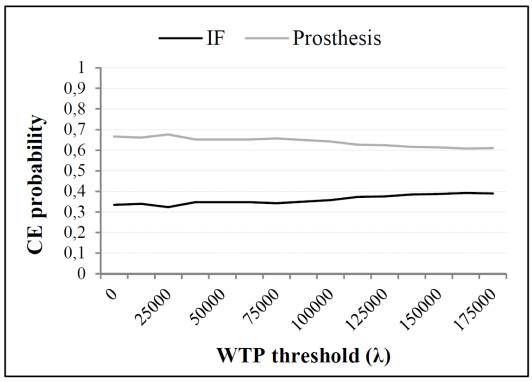
(b) Subgroup 1 = Age 60 to 70 no dementia



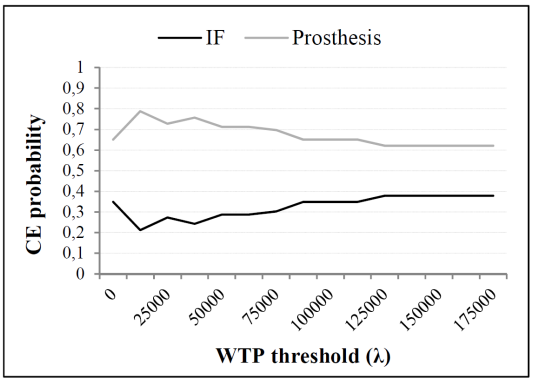
(c) Subgroup 2 = Age 71 to 80 no dementia



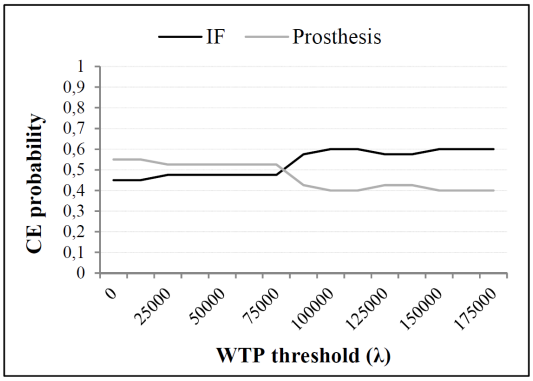
(d) Subgroup 3 = Age 81 to 90 no dementia



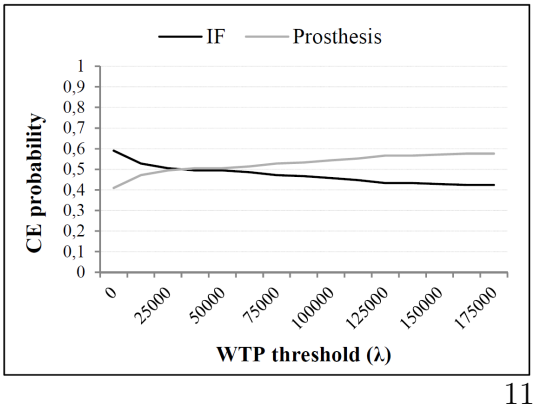
(e) Subgroup 4 = Age 90 + no dementia



(f) Subgroup 5 = Age 70 to 80 with dementia



(g) Subgroup 6 = Age 81 to 90 with dementia



(h) Subgroup 7 = Age 90 + with dementia