Cost Effectiveness Analysis in Healthcare Decision-Making: Stochastic Modeling and Statistical Inference

by

Megan Caroline DeFauw

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2011

Doctoral Committee:

Professor Vijayan N. Nair, Co-Chair
Instructor Joseph W. Norman, Co-Chair
Professor Lawrence M. Seiford
Associate Professor Kerby A. Shedden

© Megan Caroline DeFauw 2011
All Rights Reserved

To my parents, and to Phil

ACKNOWLEDGEMENTS

Many thanks to my dissertation committee. I owe great thanks to Vijay Nair, for taking me on as a student and giving me the latitude to define my own problem, and for spending so much effort developing me as a researcher. Joe Norman always provided me a broader perspective on my work and helping me bridge the medical world and engineering worlds, and for that I am very grateful. Thanks to Larry Seiford for being an invaluable mentor and sounding board, and most especially for helping me develop confidence as a young student in moments when I was ready to quit. And finally, thanks to Kerby Shedden for providing valuable perspective on my work as I concluded. Additionally, I appreciate the time that Allison Rosen spent introducing me to CEA and teaching me about medical decision analysis.

The role of my funding source, the National Science Foundation Graduate Research Fellowship, cannot be understated. I am forever honored and humbled to have been awarded a graduate research fellowship, and am sincerely appreciative of the flexibility it provided me in my studies.

I also owe a debt of gratitude to the excellent staff supporting me in my studies in the Industrial and Operations Engineering Department. A special thanks is due to Tina Blay for all the love she has given me.

I of course owe many thanks to my parents, Robert and Diane, my sister, Nicole, and my brother, Matthew. I considered myself especially blessed and lucky to have attended graduate school so close to my family; football games, dinners around

campus, and their regular presence in my life made this time especially rich for me.

I was also fortunate to have so many friends along for the journey with me, both in Ann Arbor and far away. Some of those deserving special mention for their support are Eric Jankowski, Tia Sutton, Michael Lepech, Steve Vozar, Nan Perez, Katie Tamarelli, Kelene Soltesz, Michelle Lofgren, and Jack Li. I am also grateful I had two roommates navigating graduate school and the IOE program with me: Katrina Appell and Arleigh Waring.

Last, but not least, thank you Phil Gouel for being my partner in all things great and small.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

The field of cost-effectiveness analysis (CEA) deals with the comparison of health interventions based on both costs and effectiveness (ability to improve health). This dissertation makes several methodological contributions to this area.

Part I develops direct methods for computing mean costs and effectiveness, and hence conducting CEA, in multi-state disease processes. The common approach in this case is to use discrete-event simulation techniques to simulate the process to get the associated cost and effectiveness outcomes. However, the setting up and implementation of simulation studies can be time and resource intensive. The dissertation develops analytical expressions for the time-to-failure, reward processes, and their (discounted) expectations for time-homogeneous semi-Markov processes with progressive structure. Direct Monte Carlo methods are proposed for time-varying multi-state processes. The advantages of these direct methods over discrete-event simulation are discussed. Sensitivity analysis to parameter estimation is also considered. The results are demonstrated on illustrative applications.

Part II deals proposes a richer analysis of cost-effectiveness data from discrete event simulation of disease processes. Such simulations generate extensive amounts of data which are rarely examined in detail. The analysis is typically reduced to computing and comparing simple CEA metrics. This part of the dissertation proposes a comprehensive exploratory analysis of the data through graphical techniques. This includes examining both cross-sectional and temporal views of the time-to-failure,

cost, and effectiveness distributions. The concept of a treatment-effect function is discussed, and it leads to generalized versions of two common CEA metrics. The potential for richer analysis is illustrated through various examples.

Part III of the dissertation reviews the common CEA metrics based on means of the cost and effectiveness outcomes and discusses comparisons based on first and second-order stochastic dominance as well as utility functions. It also deals with methods for incorporating statistical uncertainty from estimating the unknown parameters in CEA. Large-sample normal approximations and resampling methods are reviewed. New contributions to the CEA literature include stochastic dominance comparisons in the presence of estimation uncertainty, use of rank methods, and analysis with censored data.

CHAPTER I

Introduction

1.1 Overview of Cost-Effectiveness Analysis

There is growing emphasis in the health care area on containing costs while also improving effectiveness (health). This is the main goal of cost-effectiveness analysis (CEA) which uses metrics based on costs and effectiveness to assess and compare interventions. Cost-effectiveness analysis involves a comparison of two or more treatments for health policy or medical decision making. One or more new interventions are compared against the current or baseline intervention.

CEA was introduced to clinicians in the healthcare literature by Weinstein and Stason (1977). Since that time, it has been widely adopted. Standards for CEA studies were agreed upon in 1996 by the Panel of Cost-Effectiveness in Health and Medicine (Gold et al., 1996). These standards include elements to be included in measuring costs and effectiveness, methods of determining health and effectiveness, and incorporating time preference and the discounting of costs and effects.

There is a large literature on the development and use of various metrics for CEA. Most of them are based on the means of the (random) cost and effectiveness outcomes, where the expected values are taken with respect to some population of subjects of interest. The metric recommended for CEA in (Gold et al., 1996) is the

incremental cost-effectiveness ratio (ICER). The ICER for comparing two treatments 0 and 1 is given by

$$\mu_{ICER} = \frac{\mu_1(C) - \mu_0(C)}{\mu_1(E) - \mu_0(E)} = \frac{\Delta(C)}{\Delta(E)}, \tag{1.1}$$

where $\mu_j(C)$ and $\mu_j(E)$ are the average cost and effectiveness measures (per person) of the $j-$th intervention. Treatment 0 is typically the baseline treatment representing the current standard of practice while 1 is a new treatment under consideration. The decision rule associated with ICER is as follows: let λ be the amount of money that society is willing to pay for a single unit of effectiveness. If the ICER is less than λ, then the treatment is deemed cost-effective. If not, society would not be willing to pay for that treatment, and it is not considered cost-effective.

Two other common metrics which are equivalent to each other are the net health benefit (NHB) and the net monetary benefit (NMB). NHB was introduced by Stinnett and Mullahy (1998). NHB converts costs to effectiveness units using the willingness-to-pay ratio as the conversion factor, and adds the incremental costs (now in effectiveness units) to incremental effectiveness to get an overall benefit in effectiveness units. Similarly, NMB converts effectiveness to cost units before calculating overall benefit. The expression for NMB is

$$\mu_{NMB}(\lambda) = \lambda(\mu_1(E) - \mu_0(E)) - (\mu_1(C) - \mu_0(C)) = \lambda\Delta(E) - \Delta(C). \tag{1.2}$$

The NMB and NHB, while in different units, are equivalent. Positive values of either denote a cost-effective intervention. Otherwise, the treatment is not cost-effective.

The ICER and NMB depend on the average costs and effectiveness for the (new) treatment being considered and the baseline treatment over the lifetime of the analysis. Cost is measured in currency, and typically includes all direct costs of treatment such as the cost of tests, drugs, supplies, doctors, nurses, other health personnel, and

medical facilities (Gold et al., 1996). Cost may also include *direct non-healthcare costs* such as transportation costs to treatment, or costs due to a change in diet. The effectiveness of a treatment can be measured in different ways, including deaths averted or postponed, infections averted, extended life, and quality-adjusted-life-years (QALYs). QALYs are the most commonly used effectiveness measure in CEA and are obtained by weighting the length of life by appropriate weights which are supposed to measure 'quality of life' in the patient's health state. The weight is between 0 and 1, with the value of 1 referring to perfect health.

Per the recommendation in Gold et al. (1996), both costs and QALYs are discounted at some rate to bring the values at different times to a common reference value. Discounting cost is standard in the fields of economics and financial evaluation. Discounting QALYs is a bit less intuitive, but CEA relies on the assumption that QALYs have a monetary value, and also that there is a constant trade-off between costs and QALYs. Failure to discount QALYs while discounting cost results in the Keeler-Cretin paradox, in which any treatment program's cost-effectiveness is improved by postponement (Keeler and Cretin, 1983).

Beyond being an academic pursuit, CEA has important uses in health policy around the globe. The UK, Australia, and Canada require economic evaluations before approving new healthcare technologies (Birch and Gafni, 2004; Hill et al., 2000; Hjelmgren et al., 2001). These are countries with government-run healthcare systems, so a government-imposed limit on the cost-effectiveness of proposed treatments is natural. Even the United States government has expressed support for cost-effectiveness and comparative effectiveness research, in spite of the United States' decentralized medical insurance and payment system, with a broad base of legislation having been introduced in Congress supporting such research (Jacob-

son, 2007). The World Health Organization (WHO) has also developed guidelines for the use of cost-effectiveness analysis, with particular concern on providing cost-effectiveness information when there are limits to the time or resources that can be invested in a study (Murray et al., 2000). Thus, there is a broad base of support for conducting CEA studies. Additionally, there is considerable value for patients, physicians, and policy-makers in understanding the cost, quality of life, and length of life tradeoffs inherent in medical decisions.

In spite of the large literature that exists for CEA, there remain methodological gaps and challenges in estimating cost-effectiveness and quantifying the uncertainty in those estimates. Some of these gaps result from a misunderstanding of statistical issues, due to the fact that as a relatively new and interdisciplinary field, the CEA literature has only recently received the attention of statisticians and operations researchers. Other gaps exist because the field of medical decision-making itself is evolving; for instance, there have recently been calls in the literature to move beyond simple decision analysis for the full population to decision analysis more tailored to individual patients or groups of patients (Basu, 2009; Hayward et al., 2010; Zaric, 2003; Sculpher, 2008).

1.1.1 Methods of Conducting CEA

Two of the main challenges in CEA are obtaining the correct data for the analysis, and using that data to make meaningful inference and conclusions on cost-effectiveness. This dissertation will focus on the latter issue, but the source and quantity of data affects the inferential and modeling techniques available. So in this section, we will discuss the methods of conducting CEA, and the data sources available.

The first source of data is actual studies. Those studies may be clinical trials

or observational studies, but are generally longitudinal. The field data from these studies can be used to estimate ICER, NMB, and other cost-effectiveness metrics. See, for example, Meenan et al. (1998), van Hout et al. (1994), and Kinlay et al. (1996). However, longitudinal studies are expensive to conduct, and it takes several months or even years for the data to become available. Another obvious limitation of these clinical studies, related to the cost, is that they are generally of limited sample size. This may make it more difficult to reach a conclusion on cost-effectiveness, or even just effectiveness, than it would be with a larger sample size. In particular, it may be difficult to identify different effects in heterogeneous groups of people.

With the limitations on clinical data, CEA is also performed by creating models using data from may different sources. These models may be solved for expected value, either using analytical techniques or discrete event simulation. The limitations of this approach are the need to assume a model structure and the assumption that extrapolations of the existing data from studies are valid.

This dissertation will address issues found in both clinical studies and simulated studies. Two of the issues addressed relate to simulation models for CEA. The first issue is in computing mean statistics from cost-effectiveness models in simple manner; CEA has come to rely on discrete event simulation to generate mean values without considering whether there are tractable analytical solutions. The second broad issue is the use of simulation data, if simulation is used to evaluate the model. The time and computational effort invested in discrete event simulation leads to a question of whether and how simulation data can be used to provide individual and policy-level decision-makers with useful information for making decisions. The third issue addressed in this dissertation is decision-making and inference with small sample sizes, and is more applicable in the case of clinical data.

1.2 Research Objectives and Organization of the Dissertation

The goal of this research is to improve the process of making a decision between alternative medical treatments by improving the calculation, characterization, and understanding of differences between them. This goal is accomplished through three different research objectives.

1. Develop computationally simple methods for evaluating the mean cost and effectiveness for disease processes than can be modeled as multi-state processes;

2. Demonstrate the usefulness of graphical methods for analyzing data generated from discrete event simulation more extensively and, and developing key insights into the performance of the treatments and their comparisons;

3. Review and improve upon the literature about incorporating uncertainty into CEA and introduce ideas about evaluating the full distribution of CEA outputs under uncertainty, using utility theory and stochastic dominance.

The first two objectives enhance the understanding of multi-state models used in CEA. The third objective, due to its focus on inference, is more applicable in situations of limited sample sizes, so it is more applicable in the case of conducting CEA alongside a clinical trial or other medical observational study. As a result, the dissertation can be conceptualized as containing two parts, one focused on multi-state models and the other focused on inference. Chapters II, and III, form the first part, and Chapter IV forms the second part. Chapters II, III, and IV have all been written up as individual papers. As each paper is self-contained, there is some repetition in motivation and notation between them.

1.2.1 Multi-State Models for CEA

Models of disease processes are common in the CEA literature. Markov models for medical decision-making were proposed as an alternative to traditional decision-analytic models for understanding medical prognosis by Beck and Pauker (1983). Beck and Pauker (1983) used Markov multi-state models to calculate, via both analytical and simulation methods, the life expectancy of patients as a result of medical decisions. The Markov multi-state model is a specific instance of a more general class of multi-state models. CEA is possible in the context of multi-state models of disease by assigning cost and effectiveness variables to states and transitions within the disease model. In this way, CEA models are specialized examples of reward processes in the operations research and applied probability literature (see, for example, Howard (1971), Janssen and Manca (2006), and Janssen and Manca (2007)). In the case of CEA models, the rewards are cost and effectiveness, as well as measures such as NMB that are functions of the cost and effectiveness rewards. Markov models are popular due to the relative simplicity of inference, but other modeling assumptions are possible. One possible modeling assumption for multi-state models instead of a Markov assumption is a semi-Markov assumption. Semi-Markov CEA models are also present in the literature (Matchar et al., 1997; Castelli et al., 2007), although they are considerably less common than Markov models.

Consider, as a broad example of multi-state modeling, Figure 2.1. Each state in the model represents a distinct health state, and state $K+1$ is the end of the disease process. The end of the process is the same as "failure" in the reliability literature, and is often death, although models exist where $K+1$ could have an alternative meaning, including disease cure. Figure 2.1 shows the case only of progressively worsening diseases, as return to a previous state is not possible. Depending on the

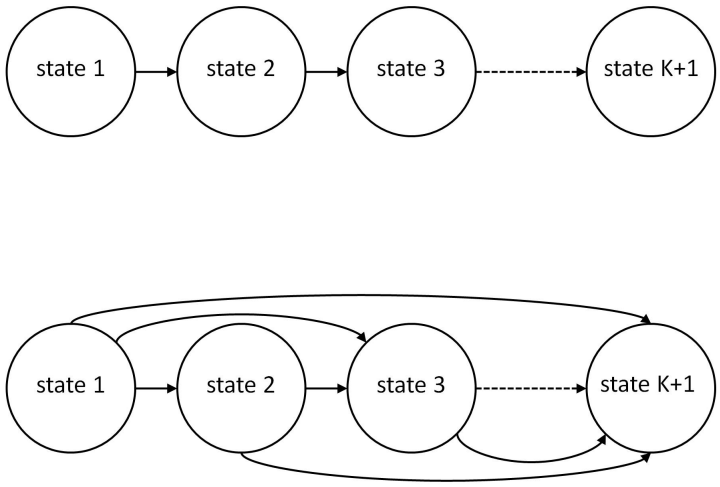

Figure 1.1: Progressive Multi-state Models. *Top:* Single-step *Bottom:* Multi-step.

disease and the definition of the state-space, transitions to previous states may be possible. The top panel is a special case where the subject can move only from a state to its immediate right. This is useful in studying some health situations where the disease progresses in a monotone manner. The multi-state model structure in bottom panel of Figure 2.1 is considerably more general.

Regardless of modeling assumptions, in multi-state models of disease, the patient is in some health state at the beginning of the study, spends an additional random of time in that state, and then moves to another health state according to some transition probability. The subject spends a random amount of time in that state, moves to a different state, and the process repeats until the person reaches an absorbing state.

Even in the case of disease models with the Markov assumption, a common method for evaluating models in the literature is discrete event simulation (DES). The basic idea in DES is to simulate some large number of 'patients' through a disease process,

with patient health states chosen randomly according to the underlying stochastic model. This simulation process creates a stochastic disease process and reward process for each patient. In traditional CEA, when only the sample means are used to calculate the μ_{ICER} and μ_{NMB}, this information is not used, and the simulation may be computationally expensive or difficult to setup. Chapters II and III address the two sides of the issue, with Chapter II focused on easier ways to evaluate multi-state CEA disease models, and Chapter III on using the detailed information from a traditional DES.

Chapter II presents methods for obtaining the estimates of the means of cost-effectiveness studies modeled as progressive multi-state models. The assumption of a progressive model is less limiting than it may seem, because with the creation of additional states, nearly any medical process can be modeled as a progressive process. Particular attention is given to the case where disease progression of both the intervention being studied and the baseline population can be modeled as semi-Markov models. Solutions for each random reward R and expected reward $\mathbb{E}(R)$ for each patient are derived in the case of reward discounting and the case where rewards are not discounted. The solution methods may be calculated analytically in the case of stationary disease models, or models of disease in which the parameters defining the underlying stochastic disease process does not change with time, which is possible for acute diseases. The methods are extended to a simple and direct simulation of sojourn time random variables and attached rewards in the case of time-varying disease processes.

Chapter III demonstrates the usefulness graphical methods for analyzing simulation data more extensively. A simulation of an arbitrarily large number of patients generates rich data on the entire distribution of patient outcomes, with differences

due to a combination of random chance and observable covariates. The simulation also allows the process to be considered as a stochastic process that evolves over time, rather than as a stochastic outcome observed at just one time, which is the standard for CEA. As an example, instead of the summary measures C_j and E_j (where j represents the treatment, and is taken to be 0 or 1 in this dissertation) in traditional CEA there are cumulative reward processes $C_j(t)$ and $E_j(t)$. Across many patients, these processes have a distribution at any time t, and the way the quantiles of those distributions, $Q_j(u;t;C)$ and $Q_j(u;t;E)$, respectively, evolve over time can be explored using graphical techniques. This chapter uses probability density plots and Q-Q plots as starting points for analysis, shows several types of follow-up analysis, and discusses the conclusions a decision-maker could reach in the context of illustrative examples.

1.2.2 Inference for Stochastic Cost-Effectiveness with Limited Sample Size

As noted earlier, determination of cost-effectiveness is sometimes made with data collected directly from a clinical setting. CEA measures, including ICER and NMB must be estimated from data, and the sample size from which to make inference is generally small. The existing CEA literature contains a large discussion about statistical inference for CEA, with a focus on determining confidence intervals for $\widehat{\text{ICER}}$ and $\widehat{\text{NMB}}$. There have been several approaches in the literature to the problem of inference, and they are reviewed in Chapter IV. Included in this review is a discussion of methodological limitations and corrections of errors in the current literature. We also propose the application of rank-based methods to CEA.

The large-sample methods and rank-based methods are for inference about a single point of a distribution; decision-makers with non-risk-neutral utility functions should use information on the full range of CEA outputs. Utility functions can

be used to map from the distribution of outputs in CEA to proper utilities in the decision-theoretic sense of von Neumann and Morgenstern (Von Neumann and Morgenstern, 1947). In general, risk preference is modeled by some utility function $u(x)$, where each random variable x_i is used as the argument is an output of CEA such as QALY, monetary cost, or monetary benefit (for the i-th intervention). A utility function, $u(x)$ specifies the form of a decision-maker's preference to values of an output x; a rational decision-maker will make the choice that maximizes $\mathbb{E}(u(x))$. We demonstrate sensitivity analysis and the creation of preference regions for CEA metrics, with a focus on the summary measure of monetary benefit.

In CEA (and in general), it is difficult to ascertain a decision-maker's exact utility function. If some general information about the utility function is known, however, it can be used. For example, if the decision-maker's utility function for NMB is known to be of the form $u(x) = -\exp(-cx)$, but the value of c is unknown, sensitivity analysis on the parameter c is possible for determining preference regions. If even less is known about the utility function, stochastic dominance can be used to determine which option is the best for an entire class of utility functions. For example, if it is known that the utility function is monotonically increasing and concave (i.e., the decision-maker is risk-averse), then the presence of second-order stochastic dominance indicates the preferred option for all utility functions in that class. If the utility function is restricted even less, to only the class of monotonically increasing utility functions (i.e., the risk-preference of the decision-maker is unknown), then first-order stochastic dominance of one option over the others would indicate the preferred decision. There are very few studies in CEA using stochastic dominance, and all have considered the population distributions to be known, rather than sampled. Chapter IV contributes to the literature by assuming that distributions are esti-

mated, and by extending stochastic dominance inferential techniques and statistical tests into CEA and discussing the implications for decision-makers. It also addresses questions of utility functions and sensitivity analysis relative to utility functions.

CHAPTER II

Cost-Effectiveness Analysis in Health Policy Decision Making: Direct Approaches for Progressive Multi-State Models

2.1 Introduction

The field of cost-effectiveness analysis (CEA) deals with the comparison of health interventions based on both relative costs and effectiveness (ability to improve health). The interventions being compared may be pharmaceutical treatments, diagnostic tools, surgical treatments, etc. CEA combines the effectiveness of a treatment and cost with the willingness of society to pay for health into a hybrid measure of cost-effectiveness.

Most metrics in CEA are based on the means or expected values of costs and effectiveness outcomes. Perhaps the most common one is the incremental cost-effectiveness ratio (ICER):

$$\mu_{ICER} = \frac{\mu_1(C) - \mu_0(C)}{\mu_1(E) - \mu_0(E)} \tag{2.1}$$

where $\mu_j(C)$ and $\mu_j(E)$ are the average cost and effectiveness measures (per person) of the $j-$th intervention, $j = 0, 1$. Here intervention 0 is the standard or baseline and intervention 1 is the new treatment being considered. The decision rule associated with ICER is as follows. Let λ be the amount of money that society is willing to

pay for a single unit of effectiveness. If the ICER is less than the willingness-to-pay ratio λ, then the new treatment is deemed cost-effective. If not, it is deemed to be not cost-effective.

A second metric that is also common is the net monetary benefit (NMB). It uses the value of λ explicitly in determining cost effectiveness through the expression

$$\mu_{NMB}(\lambda) = \lambda(\mu_1(E) - \mu_0(E)) - (\mu_1(C) - \mu_0(C)). \tag{2.2}$$

Specifically, λ is used to convert effectiveness into its monetary equivalent. If NMB will be positive, treatment 1 is cost-effective; otherwise, it is not.

The emphasis on averages or expected values arises from economic considerations. For example, average costs per person translates directly to the total cost of implementing a medical or health intervention policy (Briggs and Gray, 1998; Thompson and Barber, 2000; O'Hagan and Stevens, 2002). Cost is measured in currency and typically includes all direct costs of treatment such as the cost of tests, drugs, supplies, doctors, nurses, other health personnel, and medical facilities (Gold et al., 1996). Direct non-healthcare costs such as transportation costs to get treatment or costs due to a change in diet may also be included.

The effectiveness of a treatment can be measured in different ways: deaths averted or postponed, infections averted, extended life, and quality-adjusted-life-years (QALYs). QALYs are the most commonly used measure in CEA and are obtained by weighting the lifetime in different states by appropriate weights which are supposed to measure 'quality of life' in that health state. QALYs will be used as the unit of effectiveness in this paper, and we will refer to QALYs and effectiveness interchangeably.

It is common in CEA to discount both costs and QALYs at some rate to bring the figures at different times to a common reference value. Discounting cost is standard

in the fields of economics and finance. Discounting QALYs is less intuitive, but CEA relies on the assumption that QALYs have a monetary value and that there is a constant trade-off between costs and QALYs. Failure to discount QALYs while discounting cost results in the Keeler-Cretin paradox, in which any treatment program's cost-effectiveness is improved by postponement (Keeler and Cretin, 1983).

If the two treatments are compared in longitudinal clinical trials, one can use the field data to estimate ICER, NMB, and other cost-effectiveness metrics. See, for example, Meenan et al. (1998), van Hout et al. (1994), and Kinlay et al. (1996). However, longitudinal studies are expensive to conduct, and it takes several months or even years for the data to become available. Also, potential new interventions may just be contemplated and not actually implemented in a field study. In such cases, one might be interested in conducting multi-scenario analyses about the new intervention, including possible components to include, dosage levels, etc. To accommodate these and other needs, researchers have turned to modeling and simulation techniques, and a particularly useful tool is discrete event simulation (DES).

There are different ways of conducting DES, even within the context of CEA. Regardless of the approach, the basic idea is to simulate some large number of 'patients' through a disease process, with patients' health states chosen randomly according to the underlying stochastic model (see, for example, Sonnenberg and Beck (1993) for a discrete-time Markov model). The multi-state framework is a general way to do this. In this set-up, the subject is in some health state at the beginning of the study, spends an additional random of time in that state, and then moves to another health state according to some transition probability. The subject again spends a random amount of time in the new state, moves to a different state, and the process repeats until the person reaches an absorbing state (failure). The multi-state

process depends on unknown parameters including the transition probabilities, the distributions of state occupancy times, also known as sojourn times, and parameters associated with the effect of (patient) covariates. All of the unknown parameters have to be estimated, usually from past studies. Sometimes these estimates come from a longitudinal study of some relevant population of patients (Castelli et al., 2007; Gardiner et al., 2006). It is more likely that the data come from disparate cross-sectional studies. Examples include Valenstein et al. (2001), Paltiel et al. (2001), Liew (2006), Rosen et al. (2005), and Matchar et al. (2005).

Many software packages are available for conducting DES. If the multi-state model is built as a decision tree, the software TreeAge may be used (Pauker and Wong, 2005); an example of this is the model by Rosen et al. (2005). Other modelers may build their own simulations, using traditional programming languages. The development and implementation of DES for CEA can be time consuming with most real applications. One important issue in this context is that the low-probability paths in the multi-state models may not be sampled even with large simulation samples. Such paths may have unusually high or low rewards (costs or QALYs), so the resulting simulation may lead to higher variability in the estimates of cost-effectiveness metrics. (These issues are discussed in more detail in Section 2.6.2.)

The goal of this paper is to develop direct methods and, when possible, analytical expressions for mean costs and effectiveness and hence cost-effectiveness analysis. We restrict attention to 'progressive' processes (to be defined in the next section) to simplify the problem but the results can be extended to the more general case. There are very few papers in the CEA literature that deal with direct approaches to calculating the mean costs and QALYs associated with patient or policy treatment decisions. Some discussion is available in the operations research and applied prob-

ability literature in the general context of rewards (see, for example, Janssen and Manca (2006) and Janssen and Manca (2007)).

Analytical expressions are developed under a semi-Markov framework for the stationary case. This includes the Markov model as a special case. Direct Monte Carlo methods are described for the cases where either or both of the sojourn time distributions and transition probabilities are non-stationary. These direct approaches have many advantages over discrete event simulation. The benefits of analytical expressions is clear. One does not have to conduct a simulation study, and sensitivity analysis to estimating the parameters in the model can be done rather easily. Even in cases where direct Monte Carlo simulation is used, the simulation effort is considerably less than in setting up a DES study. Further, the simulation approach here samples all possible paths through the multi-state process, weights the results according to the probability, and computes the overall mean cost or effectiveness. In DES, on the other hand, the probability of a path being sampled depends on its probability, so some paths will have a very low probability of being included.

The rest of the paper is organized as follows. In Section 2.2, we discuss progressive multi-state processes with the semi-Markov property. In Sections 2.3 and 2.4 we develop expressions for the lifetime distributions, and the random and expected rewards earned in stationary progressive multi-state semi-Markov models of single patients, and illustrate the methods with small examples. In Section 2.5 we discuss combining patient-level results to achieve population-level results. We then discuss extensions of the method to non-stationary processes and disease processes with multiple causes of death in Section 2.6. Section 2.6.2 compares the methods developed in this paper to more traditional discrete event simulation methods, and discusses the advantages of the methods in this paper, specifically in the context of the extensions

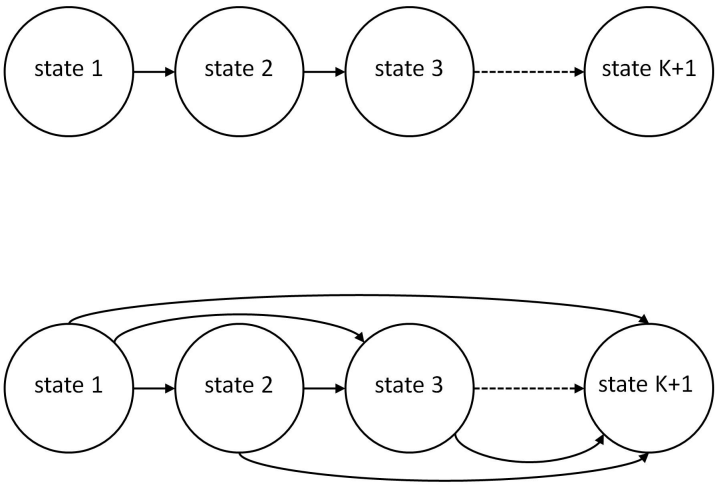

Figure 2.1: Progressive Multi-state Models
Top panel: Single-step. *Bottom panel:* Multi-step.

proposed. Finally, we present an application to illustrate the concepts in this paper in Section 2.7.

2.2 Progressive Multi-state Processes

We will consider the progressive multi-state processes in Figure (2.1). The top panel is a special case where the subject can move only from a state to its immediate right (called single-step). This is useful in studying some health situations where the disease progresses in a straightforward manner. An illustrative application using renal disease is discussed in Section 2.3.3. The bottom panel shows a more general case (multi-step) where the patients can move multiple steps with each transition. As an example of this case, a model of cardiovascular disease is discussed in Section 2.4.3.

Let $\{Y(t),\ t \in \mathcal{T}\}$ denote the state of the subject at time t for $t \in \mathcal{T}$. Time t can be discrete in which case $\mathcal{T}$ is the set of non-negative integers $\{0, 1, 2, ...\}$ or continuous in which case $\mathcal{T}$ is the non-negative half-line $\{t \geq 0\}$. The process $\{Y(t)\}$

takes values in the finite state space $E = \{1, 2, \ldots, (K+1)\}$. We take $(K+1)$ to be an absorbing state or end point of the process, denoted as 'failure.' In the context of CEA applications, this state will typically be death. Throughout this paper, we assume that the states are ordered in some natural way and that the subject moves from left to right only. This is quite natural in health settings. Cases where a patient is ill and then recovers can be handled by adding a new state rather than allowing the patient to return to the initial healthy state. Even though this increases the number of states, it is useful to distinguish a patient who had been in a non-healthy state and has not had health problems for some time from someone who has always been healthy.

The first part of the paper deals with homogeneous multi-state processes. We will be especially interested in semi-Markov processes (SMPs), which are generalizations of Markov processes. It is convenient to define an SMP in terms of its equivalent Markov renewal process (MRP) $\{(J_n, T_n), n = 1, 2, ...\}$ (Janssen and Manca, 2006). The MRP spends a random amount of time in a state $i \in E$ and then jumps to another state $j \in E$. J_n represents the state the process is in before the nth jump, and X_n is the occupancy time in state J_n before the nth jump. Further, $T_n = X_1 + \cdots + X_n$ is the time of the nth jump. The transition probabilities and sojourn times are given as follows:

1. the transition probability $p_{i,j}(t) = P(J_{n+1} = j | J_n = i, T_n = t)$
2. the state occupancy (sojourn) distributions are given by the conditional distributions $F_{i,j}(t_1, t_2) = P(T_{n+1} \leq t_2 | T_n = t_1, J_n = i, J_{n+1} = j)$.

The term $Q_{i,j}(t_1, t_2) = F_{i,j}(t_1, t_2) p_{i,j}(t_1)$ is called the semi-Markov kernel. Note that $\lim_{t_1 \to \infty} Q_{i,j}(t_1, t_2) = p_{i,j}(t_1)$. The corresponding SMP $Y(t)$ can be obtained as

$Y(t) = J_n$ for $t \in (T_{n-1}, T_n)$. For a time-homogeneous SMP, $p_{i,j}(t)$ is constant in time and $F_{i,j}(t_1, t_2) = P(T_{n+1} \leq t_2 | T_n = t_1, J_n = i, J_{n+1} = j)$ does not depend on t_1, the time of the $n-$th jump. The time-homogeneous Markov case is the simplest and most well known example of a multi-state model. In this case, the sojourn distributions $F_{i,j} = F_i$ and are exponential.

2.3 Single-Step Progressive: Stationary Case

This section develops analytical expressions for the lifetime distribution and random and expected rewards for the one-step progressive stationary (time-homogeneous) case (top panel of Figure (2.1)). The results are developed for a single subject who is in state s (with $s \leq K$) at the start of the study and has spent a period time u in that state. The results will be combined later to get expressions for a distribution (population) of subjects. Note that in this one-step case, the transition probability from state g to $g + 1$ is one.

2.3.1 Life Time Distribution

Let $X_{s,s+1}(u)$ be the conditional random variable $[X_{s,s+1} | X_{s,s+1} \geq u]$. Then, we can write $T(s; u)$, the (residual) lifetime of this subject, as a sum of $X_{s,s+1}(u)$ and the subsequent $X_{g,g+1}$'s:

$$T(s; u) = X_{s,s+1}(u) + X_{s+1,s+2} + ... + X_{K,K+1}. \tag{2.3}$$

Recall that the conditional distribution of $X_{s,s+1}(u)$ is the same as that the unconditional distribution of $X_{s,s+1}$ in the exponential case. The special case where the subject enters state 1 at the beginning of the study ($u = 0$) is obtained by taking $s = 1$ and $u = 0$ in the above expression.

Let $\mu_{s,s+1}(u) = \mathbb{E}(X_{s,s+1})$ and $\mu_{g,g+1} = \mathbb{E}(X_{g,g+1})$ for $g > s$. Then, the mean $\mathbb{E}[T(s, u)] = \mu_{s,s+1}(u) + \sum_{g=s+1}^{K} \mu_{g,g+1}$. For an SMP, the $X_{g,g+1}$'s are independent,

so the variance is also the sum of the individual variances. For the more general stationary case (i.e., not an SMP), we have to specify the dependence structure among the $X_{g,g+1}$'s in order to compute the variance.

Special Cases:

Case 1: Consider the Markov case where $F_{i,j} = F_i$ and are exponential with mean η_i for $i = 1, ..., K$. Then, $X_{s,s+1}(u)$ has the same distribution as the unconditional random variable $X_{s,s+1}$. The distribution of $T(s;u)$ is a sum of independent exponentials. This is sometimes called a hypo-exponential distribution (Ross, 2003). The mean is $\sum_{i=s}^{K} \eta_i$ and the variance is $\sum_{i=s}^{K} \eta_i^2$. In the special case where $\eta_i = \eta, i = 1, ..., K$, $T(s;u)$ has a gamma distribution with parameters $(K - s + 1, \eta)$.

Case 2: Suppose we have a semi-Markov process where $F_{i,j}$ is Gamma$(\kappa_{i,j}, \eta)$, so the scale parameter η is the same. Then, $\sum_{g=s+1}^{K} X_{g,g+1}$ is $Z \sim$ Gamma$(\sum_{g=s+1}^{K} \kappa_{g,g+1}, \eta)$, so $T(s;u)$ has the same distribution as $X_{s,s+1}(u) + Z$ where $X_{s,s+1}(u)$ is a conditional Gamma$(\kappa_{s,s+1}, \eta)$ random variable, conditioned on being $\geq u$. If $u = 0$, then the distribution is Gamma$(\sum_{g=s}^{K} \kappa_{g,g+1}, \eta)$. A similar structure holds whenever the sum of $X_{g,g+1}$'s is closed under convolution (the sum belongs to the same family as the original distribution) or more generally, the sum has a closed form expression. Other distributions with this property include the Normal (although this is not usually used to model sojourn times which are positive), Inverse Gaussian with a constraint on the two parameters, and Poisson. However, for many of the common lifetime distributions, such as Weibull and Lognormal, the distribution of the sum does not have a closed form expression even when the components are independent. However, it is relatively easy to simulate these random variables directly and obtain a Monte Carlo approximation to the distribution of the sum in equation (2.3).

2.3.2 Reward Random Variables and Expected Rewards

This section develops expressions for the random rewards and their expectations accrued over a lifetime. The term 'rewards' represent both costs and effectiveness. The results are obtained under the following setup:

1. There is a fixed reward $U_{i,j}$ associated with transitioning from state i to state j. This could, for example, be the cost associated with a one-time treatment;

2. There is a variable reward V_i per unit time associated with the amount of time spent in state i. This could, for example, be the QALYs for that state (sojourn time weighted by quality index) or some treatment cost that depends on the amount of time spent in the state.

No discounting

Consider first the case with no discounting of rewards. Again, the subject is in state s at the beginning of the study and has already spent an amount of time u in that state. The total undiscounted random reward is simply

(2.4)

$$R(s;u) = V_s X_{s,s+1}(u) + U_{s,s+1} + V_{s+1} X_{s+1,s+2} + U_{s+1,s+2} + ... + V_K X_{K,K+1} + U_{K,K+1}.$$

It can be written in a compact form as

$$R(s;u) = \sum_{g=s}^{K} U_{g,g+1} + V_s X_{s,s+1}(u) + \sum_{g=s+1}^{K} V_g X_{g,g+1}. \tag{2.5}$$

Let $\mu_{g,g+1}$ be defined as before as the expectation of $X_{g,g+1}$. Then, the expected reward is

$$\mathbb{E}(R(s;u)) = \sum_{g=s}^{K} U_{g,g+1} + V_s \mu_{s,s+1}(u) + \sum_{g=s+1}^{K} V_g \mu_{g,g+1}. \tag{2.6}$$

The variance of $R(s,u)$ can also be computed from equation (2.5). In the case of an SMP where the $X_{g,g+1}$'s are independent, the variance is also just the sum of the variances:

$$Var(R(s;u)) = V_s^2\sigma_{s,s+1}^2(u) + \sum_{g=s+1}^{K} V_g^2\sigma_{g,g+1}^2,$$

where $\sigma_{s,s+1}^2(u) = Var(X_{s,s+1}(u))$ and $\sigma_{g,g+1}^2 = Var(X_{g,g+1})$.

The more interesting problem is to compute the reward random variable and its moments under discounting.

Discounting

For discrete time, the rewards are discounted by a discount rate r for each unit of time. Throughout, let $\theta = \frac{1}{(1+r)}$, so θ has the simple effect of discounting the reward by a single time unit. For continuous time, the discount rate is ρ, so the discount factor for a time of length x is $\exp(-\rho x)$.

Define

$$W_{g,g+1} = U_{g,g+1} + \alpha(V_{g+1} - V_g) \tag{2.7}$$

with $\alpha = \frac{\theta}{(1-\theta)}$ for discrete time and $\alpha = 1/\rho$ for continuous time. Further, in the discrete time case, the discount factor $Z_{s,s+1} = \theta^{X_{s,s+1}(u)}$ and $Z_{g,g+1} = \theta^{X_{g,g+1}}$ for $g \geq s+1$. In the continuous time case, $Z_{s,s+1} = \exp(-\rho X_{s,s+1}(u))$ and $Z_{g,g+1} = \exp(-\rho X_{g,g+1})$ for $g \geq s+1$.

Proposition 1:

a) The total random discounted reward is

$$R_D(s;u) = \alpha V_s + \sum_{g=s}^{K}\left[W_{g,g+1}\prod_{m=s}^{g} Z_{m,m+1}\right], \tag{2.8}$$

where $V_{K+1} = 0$.

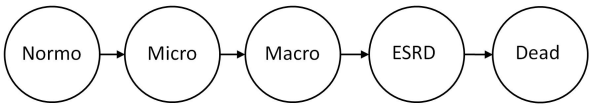

Figure 2.2: Renal Disease Model Structure

b) If the sojourn times are independent (as in the case of an SMP), the total expected discounted reward is

$$\mathbb{E}[R_D(s;u)] = \alpha V_s + \sum_{g=s}^{K} \left[W_{g,g+1} \prod_{m=s}^{g} L_{m,m+1}(\gamma) \right], \tag{2.9}$$

where $L_{s,s+1}(\gamma) = \mathbb{E}(\exp(-\gamma X_{s,s+1}(u))$, the Laplace transform of $X_{s,s+1}(u)$ and $L_{m,m+1}(\gamma)$ is the Laplace transform of $X_{m,m+1}$ for $m > s$. Further, $\gamma = -\log(\theta)$ in the discrete case and $\gamma = \rho$ in the continuous case.

Part (a) of Proposition 1 is derived in Appendix 2.9.1. Part (b) follows from the definition and independence of the $Z_{m,m+1}$'s.

Remark : There are explicit expressions for the Laplace transform of many common distributions. For example, the Laplace transform $L(t)$ for Gamma(κ, η) is $(1+\eta t)^{-\kappa}$. The exponential case corresponds to $\kappa = 1$.

2.3.3 Illustrative Application

Computation of CEA Metrics

We consider the renal-disease component of the model from the application in Rosen et al. (2005) which involves assessing the benefits of angiotensin-converting enzyme (ACE) inhibitors, and construct an illustrative example using the same health states, but different sojourn times and probabilities. ACE-inhibitors are commonly available drugs that are used to treat a variety of health problems, including cardiovascular disease, renal disease, diabetes, high blood pressure, and migraines. They

prevent the production of angiotensin II from angiotensin I. Angiotensin II is a substance in the tissues and blood that narrows blood vessels, resulting in increased blood pressure and stress upon the heart. Generic ACE-inhibitors are widely available, resulting in relatively affordable pricing. The ACE-inhibitor ramipril was shown to be effective in reducing MI, cardiovascular death, other cardiovascular events, and nephropathy (renal disease) in the population aged 55 and older with diabetes (Gerstein et al., 2000). ACE-inhibitors are also an attractive treatment option because they have relatively few side-effects. In this illustrative application, we consider the effects of ACE-inhibitors on renal disease; in later examples, we also consider their effect on cardiovascular disease.

Figure (2.2) shows the one-step progressive model with four states ordered from healthiest to sickest: Normoalbuminuria, Microalbuminuria, Macoalbuminuria, and End Stage Renal Disease (ESRD). The policy decisions being compared are: '1': use of ACE inhibitors (adherence), and '0': no use of ACE-inhibitors (non-adherence). The particular policy decision being considered is full coverage (no co-pay) of the cost of ACE-inhibitors under Medicare for elderly diabetics, and we are specifically interested in the effect on those who change from adherence to non-adherence due to full coverage. In this case, it is assumed that all subjects receiving full reimbursement for ACE-inhibitors will take them (adherence), and all subjects not receiving full reimbursement for ACE-inhibitors will not take them (non-adherence).

We assume that all subjects have diabetes and start in the Normoalbuminuria state. Table 2.1 provides the inputs needed to compute cost and QALYs. All patients incur an annual cost of $\$3,500$ for routine medical treatment. The annual cost of ACE-inhibitors (noted as the cost of treatment adherence) is \$300 per subject. As is usual in the literature, we take the discount rate for costs and QALYs, denoted by

Table 2.1: Model Inputs: Disease Utilities, Costs, and Discount Rate

QALYs	
Diabetes (Baseline Health)	0.88
ESRD	0.61
Annual costs	**$**
Ongoing cost of care	3,500
Diabetes, Normoalbuminuria	1,000
Diabetes with Microalbuminuria	2,000
Diabetes with Macroalbuminuria	4,000
ESRD	80,000
Treatment Adherence	300

Table 2.2: Model Inputs: Disease Prevalence and Progression, Renal Disease Model

Parameter Name	**Non-Adherent**	**Adherent**
Sojourn Distribution Parameters, Gamma Distribution		
Shape Parameters, κ		
Normoalbuminuria to Microalbuminuria ($\kappa_{1,2}$)	2	6
Microalbuminuria to Macoalbuminuria ($\kappa_{2,3}$)	1	4
Microalbuminuria to ESRD ($\kappa_{3,4}$)	1	2
ESRD to Dead ($\kappa_{4,5}$)	0.1	0.1
Scale Parameter, η, (all transitions)	3	3

r, to be 3%(Gold et al., 1996).

The ACE inhibitors affect the sojourn-time distributions in the states, slowing renal disease progression. We illustrate the computations with gamma sojourn time distributions. The model parameters are given in Table 2.2. As an example of the calculations, consider the evaluation of $\mathbb{E}(C_0)$. We have

$$\mathbb{E}(C_0) = \alpha V_1 + W_{1,2}L_{1,2} + W_{2,3}L_{1,2}L_{2,3} + W_{3,4}L_{1,2}L_{2,3}L_{3,4} + W_{4,5}L_{1,2}L_{2,3}L_{3,4}L_{4,5},$$

since all the $U_{i,j}$'s are zero. We can compute each of the above elements: $V_1 = 4500$ (the sum of ongoing cost of care and having diabetes with Normoalbuminuria), $V_2 = 5500$, $V_3 = 7500$, $V_4 = 80000$, and $V_5 = 0$. We take $\rho = 0.03$, so $\alpha = 1/.03$. Further, $L_{m,m+1} = (1+\eta\rho)^{\kappa_{m,m+1}}$, and as an example, $L_{1,2} = (1+3\times0.03)^2$ for the non-adherent case. Form all this, we get $\mathbb{E}(C_0) = \$68,639$. Similar calculations show: $\mathbb{E}(C_1) = \$123,763$, $\mathbb{E}(E_0) = 8.7$, and $\mathbb{E}(E_1) = 19.0$, resulting in $\mu_{ICER} = 5,352\$/\text{QALY}$. The usual comparison value is $\lambda = \$50,000$, so ACE-inhibitors will be considered

cost-effective. Alternatively, μ_{NMB} with $\lambda = \$50,000$ is $\$426,326$; since this is much larger than the reference value of 0, there is a substantial benefit from using the new treatment. We also computed the reward expected values for the case with Weibull sojourn distributions with matching the first two moments matching those of the gamma distributions used here, and the conclusions were similar, suggesting that the results were robust to distributional assumptions in this case. The use of analytical expressions facilitates such comparisons easily.

Parameter Uncertainty and Sensitivity Analysis

The reward estimates are only as reliable as the parameters used to calculate them. We now illustrate how the uncertainty due to parameter estimation can be easily quantified using the analytical expressions in this paper. Specifically, we want to assess how much the conclusions of the CEA analysis depend on the uncertainty in the estimated values of the unknown parameters.

Consider first a sensitivity analysis – how sensitive cost, QALYs, ICER, and NMB are to small perturbations of the estimates of each individual parameter. One way to do this is to compute the derivative of the output as a function of that input parameter. The expression for the expected discounted reward is

(2.10)

$$\mathbb{E}(R_D(1;0)) = \alpha V_1 + W_{1,2}L_{1,2} + W_{2,3}L_{1,2}L_{2,3} + W_{3,4}L_{1,2}L_{2,3}L_{3,4} + L_{4,5}L_{1,2}L_{2,3}L_{3,4}L_{4,5}.$$

If we are interested in assessing the sensitivity to changes in the parameters $\kappa_{1,2}$ and $\kappa_{2,3}$, we can compute the partial derivatives of $\mathbb{E}(R_D(1;0))$ with respect to these parameters and evaluate them at the estimated parameter values. This derivative shows the change in reward due to a one-unit change in the parameter at the estimated values. Doing the calculations, $\partial\mathbb{E}(C_{0,D}(1,0))/\partial\kappa_{2,3} = \$9,429$, and

$\partial\mathbb{E}(C_{1,D}(1,0))/\partial\kappa_{2,3} = \$4,836$. The corresponding values for QALYs are: $\partial\mathbb{E}(E_{0,D}(1,0))/\partial\kappa_{2,3} = 1.8$, and $\partial\mathbb{E}(E_{1,D}(1,0))/\partial\kappa_{2,3} = 0.9$. Thus, the mean values for non-adherent rewards are more sensitive to changes in parameter values. The sensitivity values for all the parameters are given in Table 2.3.

We now discuss computation of the overall standard error by using Taylor series approximation. Let $\widehat{\mathbb{E}}(R(1;0)) = F(\hat{\beta}_1, ..., \hat{\beta}_L)$ where the $\hat{\beta}_j$'s represent the estimated parameters (transition probabilities, parameters of the state-occupancy distributions, etc.). Then, the Taylor series approximation for the variance of $\widehat{\mathbb{E}}(R(1;0))$ is $\text{Var}[\widehat{\mathbb{E}}(R(1;0))] \approx \mathbf{f}^T\mathbf{\Sigma}\,\mathbf{f}$ where f is the vector of partial derivative of F evaluated at the estimated values and $\mathbf{\Sigma}$ is the variance-covariance matrix of the $\hat{\beta}_j$'s.

Suppose we have field data from $n = 100$ patients on the sojourn times in different states. Usually, the field data will be subject to various forms of censoring (right, left, interval, etc.), but for simplicity we assume here there is no censoring. Further, suppose the values in Table 2.2 are the MLEs of the parameters when the data were fitted to gamma distributions. We can use arguments from likelihood theory to get the estimated information matrix and invert it to get the variance-covariance matrix of the parameters. The details are available in standard textbooks and are omitted here (see, for example, Bickel and Doksum (2001)).

The estimated variances and standard errors for costs and effectiveness are also given in Table 2.3. As with the sensitivity analysis on individual parameters, the standard errors for non-adherent rewards are larger than those for adherent rewards.

In practice, one is not likely to get maximum likelihood of the parameter estimates under some parametric model. Rather, the summary data will consist of statistics such as the mean and variance of the distributions and possibly their standard errors. One would then have to estimate the underlying parameters of a parametric model

Table 2.3: Partial Derivatives of Reward Functions

Variable	$\mathbf{C_0}$	$\mathbf{C_1}$	$\mathbf{E_0}$	$\mathbf{E_1}$
$\kappa_{1,2}$	7,012	3,123	1.8	0.9
$\kappa_{2,3}$	9,429	4,836	1.8	0.9
$\kappa_{3,4}$	13,866	7,262	1.8	0.9
$\kappa_{4,5}$	161,404	81,307	1.2	0.6
η	17,073	19397	2.3	3.4
$\hat{\sigma}^2$	1.81562×10^7	3.76968×10^6	0.23	0.09
$\hat{\sigma}$	$4,261$	$1,942$	0.48	0.30

for the sojourn distribution by matching the moments. For a gamma distribution, since the mean is $\kappa\eta$ and the variance is $\kappa\eta^2$, we get $\hat{\eta}$ = (sample variance)/(sample mean) and $\hat{\kappa}$ = (sample mean)^2/(sample variance). We can then use Taylor series approximation to get the standard errors of the estimated parameters.

Similar methods can be used to compute the standard errors of CEA metrics such as ICER and NMB. See, for, example, O'Brien et al. (1994). In this example, the estimated standard error for μ_{ICER} is 245 and that for μ_{NMB} is 28,589. Recall that the estimates for μ_{ICER} and μ_{NMB} at $\lambda = 50,000$ were $5,352$\$/QALY and \$$426,326$ respectively. So, the standard errors due to parameter uncertainty are small in comparison, so the conclusion of cost-effectiveness still holds.

If the computations of the derivatives is difficult, one can use a simple Monte Carlo approach to simulate from the distributions of the input variables (Kennedy and Gentle, 1980; Givens and Hoeting, 2005). These methods are all well known in the statistical literature. They are not necessary in this single-step case, but we will demonstrate their application in the more complicated multi-step case in the next section.

2.4 Multi-Step Progressive: Stationary Case

As before, the results for the multi-step case are developed first for a single subject who is in state s (with $s \leq K$) at the start of the study and has spent a period time

u in that state.

2.4.1 Life Time Distributions

We can see from Figure (2.1) that the subject can reach the absorbing state through many possible paths. Suppose the subject moves through the sequence of states $\{s = J_1 < J_2 < ... < J_k < J_{k+1} = K + 1\}$ where J_i's denote the random indices of the visited states and k is the random number of states visited before failure. Then,

$$T(s;u) = I_{s,J_2} \times I_{J_2,J_3} \cdots \times I_{J_L,K+1} \left(X_{s,J_2}(u) + X_{J_2,J_3} + \cdots + X_{J_L,K+1}\right). \tag{2.11}$$

A discrete-event simulation approach might proceed by mimicking the above process – select J_2 randomly according to the set of transition probabilities, simulate $X_{s,J_2}(u)$, select J_3 randomly according to the transition probabilities, simulate X_{J_2,J_3}, and so on.

In the stationary case, the transition probabilities (and hence the indicator functions in equation (2.11) are independent of time. Thus, we can first condition on the indicator function (the path), compute the sum of the sojourn times for the path, and then sum over all possible paths. This leads to a representation of the total (residual) lifetime as

$$T(s;u) = \sum_j I^{(j)}_{s,i_2,\ldots,i_{k_j},K+1} \left(X^{(j)}_{s,i_2}(u) + X^{(j)}_{i_2,i_3} + \cdots + X^{(j)}_{i_{k_j},K+1}\right), \tag{2.12}$$

where the sum j is over all possible paths $\{s < i_2 < ... < i_{k_j} < K+1\}$ from the initial state s to the absorbing state $K + 1$. Here $I^{(j)}_{s,i_2,\ldots,i_{k_j},K+1} = I_{s,i_2} \times I_{i_2,i_3} \times \cdots I_{i_{k_j},K+1}$. Note that there are $J = 2^{K-s}$ possible such paths. Further, k_j, the number of states visited before absorption, depends on the path j. For simplicity, we suppress the superscript notation j in the rest of the paper. Again, the special case where the

subject just starts in state 1 at the beginning of the study is obtained by taking $s = 1$ and $u = 0$ in the above expression.

As an example, consider the case with 4 states and $s = 1$. There are four possible paths: $1 \to 2 \to 3 \to 4,\ 1 \to 2 \to 4,\ 1 \to 3 \to 4$, and $1 \to 4$, so

$$\begin{aligned} T(1,u) &= I_{\{1,2,3,4\}}(X_{1,2}(u) + X_{2,3} + X_{3,4}) + I_{\{1,2,4\}}(X_{1,2}(u) + X_{2,4}) \\ &+ I_{\{1,3,4\}}(X_{1,3}(u) + X_{3,4}) + I_{\{1,4\}}X_{1,4}(u). \end{aligned} \tag{2.13}$$

The random variables $X_{1,2}(u)$ in the first and second components of the sum in equation (2.13) are independent realizations of the same random variable. The reason for independence is that a given subject can take only one of the 4 possible paths, and different subjects are independent of each other. A similar comment applies to other cases.

The lifetime distribution is a finite mixture with J elements. Its density can be expressed as

$$f_T(t|s;u) = \sum_j p_{s,i_2,\ldots,i_{k_j},K+1} \left(f_{s,i_2}(t;u) * f_{i_2,i_3}(t) * \ldots * f_{i_{k_j},K+1}(t) \right), \tag{2.14}$$

where, as before, the sum j is over all possible paths. Here, $p_{s,i_2,\ldots,i_{k_j},K+1} = E(I_{s,i_2,\ldots,i_{k_j},K+1}) = p_{s,i_2} \times p_{i_2,i_3} \times \cdots \times p_{i_{k_j},K+1}$, $f_{i_j,i_k}(t)$ is the density of X_{i_j,i_k}, $f_{s,i_k}(t;u)$ is the density of $X_{s,i_2}(u)$ and $*$ denotes convolution.

From standard properties of mixture distributions (McLachlan and Peel, 2000), one can express the mean of the life-time distribution as

$$\mathbb{E}(T(s;u)) = \sum_j p_{s,i_2,\ldots,i_{k_j},K+1} \left(\mu_{s,i_2}(u) + \mu_{i_2,i_3} + \ldots + \mu_{i_{k_j},K+1} \right), \tag{2.15}$$

where $\mu_{i,j}$ is the mean of $X_{i,j}$ and $\mu_{s,i_1}(u)$ is the mean of $X_{s,i_1}(u)$. The variance expressions are discussed in Appendix 2.9.3.

Example: Consider the four-state model in equation (2.13). Suppose $s = 1, u = 0$, i.e., the subject starts in state 1 and has not yet spent any time in that state. Further suppose the distributions of $X_{i,j}$ are Gamma$(\kappa_{i,j}, \eta)$ and they are independent. Recall that the sums of independent gamma distributions with the same scale parameter η are also gamma. So,

$$\begin{aligned} f_T(t|1;0) &= p_{1,2,3,4}\ \text{Gamma}(\kappa_{1,2}+\kappa_{2,3}+\kappa_{3,4},\eta) \\ &+ p_{1,2,4}\ \text{Gamma}(\kappa_{1,2}+\kappa_{2,4},\eta) \\ &+ p_{1,3,4}\ \text{Gamma}(\kappa_{1,3}+\kappa_{3,4}\eta) \\ &+ p_{1,4}\ \text{Gamma}(\kappa_{1,4},\eta), \end{aligned}$$

a mixture of gamma distributions. The mean of a Gamma(κ, η) distribution is $\kappa\eta$. This can be used to calculate the mean of the lifetime as

(2.16)

$$\mathbb{E}(T(1;0)) = \eta\left(p_{1,2,3,4}(\kappa_{1,2}+\kappa_{2,3}+\kappa_{3,4}) + p_{1,2,4}(\kappa_{1,2}+\kappa_{2,4}) + p_{1,3,4}(\kappa_{1,3}+\kappa_{3,4}) + p_{1,4}\kappa_{1,4}\right).$$

2.4.2 Reward Random Variables and Expected Rewards

No Discounting

Again, consider first the case where there is no discounting of rewards over time. The total random cost for an individual subject is

$$\begin{aligned} R(s;u) = \sum I_{s,i_2,\ldots,i_{k_j},K+1}\Big(&V_{s,i_2}X^j_{s,i_2}(u) + U_{s,i_2} + V_{i_2,i_3}X^j_{i_2,i_3} + U_{i_2,i_3} \\ &+ \cdots + V_{i_{k_j},K+1}X^j_{i_{k_j},K+1} + U_{i_{k_j},K+1}\Big), \end{aligned} \tag{2.17}$$

which can be written in a compact form as

$$R(s;u) = \sum_j I_{s,i_2,\ldots,i_{k_j},K+1}\left(\sum_{g=1}^{k_j} U_{i_g,i_{g+1}} + V_{i_1}X^j_{i_1,i_2}(u) + \sum_{g=2}^{k} V_{i_g}X^j_{i_g,i_{g+1}}\right), \tag{2.18}$$

where $i_1 = s$ and $i_{k_j} + 1 = K + 1$. The expected reward can be calculated from this as

$$\mathbb{E}(R(s;u)) = \sum_j p_{s,i_2,\ldots,i_{k_j},K+1} \left(\sum_{g=1}^{k_j} U_{i_g,i_{g+1}} + V_{i_1}\mu_{i_1,i_2}(u) + \sum_{g=2}^{k} V_{i_g}\mu_{i_g,i_{g+1}} \right), \tag{2.19}$$

where, as before, $\mu_{i,j}$ is the mean of $X_{i,j}$ and $\mu_{i,j}(u)$ is the mean of $X_{i,j}(u)$. The variance of the expression is a straightforward extension of the single-step case.

Discounting

The case with discounting is of more interest. The results for the multi-step case can be obtained by combining the results for all possible paths. The derivations are given in Appendix 2.9.2. We just provide the main result here.

Recall that in the case with discrete time, the rewards are discounted by a factor r per unit time. Let $\theta = \frac{1}{(1+r)}$. For continuous time, the discount rate is ρ, so the discount for a time of length x is $\exp(-\rho x)$. Let

$$W_{i_g,i_{g+1}} = U_{i_g,i_{g+1}} + \alpha(V_{i_{g+1}} - V_{i_g}), \tag{2.20}$$

for $V_{K+1} = 0$ where $\alpha = \frac{\theta}{(1-\theta)}$ for discrete time and $\alpha = 1/\rho$ for continuous time.

Proposition 2:

a) The total random discounted reward is

$$R_D(s;u) = \sum_j I_{s,i_2,\cdots,i_{k_j},K+1} \left(\alpha V_s + \sum_{g=1}^{k_j} \left[W_{i_g,i_{g+1}} \prod_{m=1}^{g} Z_{i_m,i_{m+1}} \right] \right), \tag{2.21}$$

where $i_1 = s$ and $i_{k_j} + 1 = K + 1$.

b) If the sojourn times are independent (as in the case of an SMP), the total expected discounted reward is

$$\mathbb{E}(R_D(s;u)) = \sum_j p_{s,i_2,\cdots,i_{k_j},K+1} \left(\alpha V_s + \sum_{g=1}^{k_j} \left[W_{i_g,i_{g+1}} \prod_{m=1}^{g} L_{i_m,i_{m+1}}(\gamma) \right] \right), \tag{2.22}$$

where $L_{i_m,i_{m+1}}(\gamma)$ is the Laplace transform of $X_{i_m,i_{m+1}}$ for $m \geq 2$ and of $X_{s,i_2}(u)$ for $m = 1$, and $\gamma = -\log(\theta)$ in the discrete case and $\gamma = \rho$ in the continuous case.

Example: Consider again the 4-state model in equation (2.13). Taking $u = 0$ and $s = 1$,

$$\begin{aligned} R(1;0) &= I_{1,2,3,4}(\alpha V_1 + W_{1,2}Z_{1,2} + W_{2,3}Z_{1,2}Z_{2,3} + W_{3,4}Z_{1,2}Z_{2,3}Z_{3,4}) \\ &+ I_{1,2,4}(\alpha V_1 + W_{1,2}Z_{1,2} + W_{2,4}Z_{1,2}Z_{2,4}) \\ &+ I_{1,3,4}(\alpha V_1 + W_{1,2}Z_{1,3} + W_{3,4}Z_{1,2}Z_{2,4}) \\ &+ I_{1,4}(\alpha V_1 + W_{1,4}Z_{1,4}). \end{aligned}$$

If it is a semi-Markov model, the expected reward is obtained by replacing $Z_{i,j}$ in the above expression with its Laplace transform. Suppose the state-occupancy times are gamma; i.e., $F_{i,j}$ =Gamma(κ, η). Its Laplace transform is $(1 + t\eta)^{-\kappa}$. So, the expected reward for a gamma semi-Markov process can be written as

$$\begin{aligned} \mathbb{E}(R(1;0)) = {} & p_{1,2}p_{2,3}p_{3,4}(\alpha V_1 + W_{1,2}(1+t\eta)^{-\kappa_{1,2}} + W_{2,3}(1+t\eta)^{-\kappa_{1,2}}(1+t\eta)^{-\kappa_{2,3}} \\ & + W_{3,4}(1+t\eta)^{-\kappa_{1,2}}(1+t\eta)^{-\kappa_{2,3}}(1+t\eta)^{-\kappa_{3,4}}) \\ & + p_{1,2}p_{2,4}(\alpha V_1 + W_{1,2}(1+t\eta)^{-\kappa_{1,2}} + W_{2,4}(1+t\eta)^{-\kappa_{1,2}}(1+t\eta)^{-\kappa_{2,4}}) \\ & + p_{1,3}p_{3,4}(\alpha V_1 + W_{1,3}(1+t\eta)^{-\kappa_{1,3}} + W_{3,4}(1+t\eta)^{-\kappa_{1,3}}(1+t\eta)^{-\kappa_{3,4}}) \\ & + p_{1,4}(\alpha V_1 + W_{1,4}(1+t\eta)^{-\kappa_{1,4}}). \end{aligned}$$

The exponential distribution is a special case with $\kappa_{i,j} = 1$.

2.4.3 Illustrative Application

We now consider the cardiovascular disease (CVD) component of the application in Rosen et al. (2005). The structure, shown in Figure (2.3), is a multi-step process with four states: healthy (no history of MI), a history of one MI (which was survived), a history of two MIs (both of which were survived), and death due to MI. We make

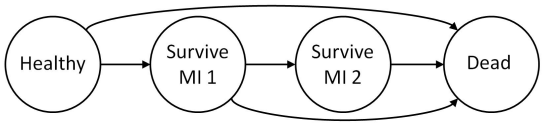

Figure 2.3: Cardiovascular Disease Model Structure

Table 2.4: Model Inputs: Disease Utilities, Costs, and Discount Rate

QALYs	
Diabetes (Baseline Health)	0.88
MI	0.88
Annual costs	**$**
Ongoing cost of care	3,500
History of MI	2,500
Acute MI (impulse cost)	35,000
Treatment Adherence	300

a simplifying assumption that a third MI would automatically be fatal. In addition, we restrict attention to deaths due to MI only. We assume that the subject starts with no history of MI. Again, the intervention in question is full cost-coverage of ACE inhibitors in diabetics.

The sojourn and impulse costs and QALYs are in Table 2.4. The transition probabilities and sojourn time distributions are in Table 2.5. Note that the state-transition probabilities do not vary with treatment, but the sojourn time distributions do.

The possible paths through this MI disease process are:

1. Healthy $\rightarrow$ 1 MI $\rightarrow$ 2 MI $\rightarrow$ Dead
2. Healthy $\rightarrow$ 1 MI $\rightarrow$ Dead
3. Healthy $\rightarrow$ Dead

For notational simplicity, we will refer to the healthy state as state 1, survival of a single MI as state 2, survival of 2 MI as state 3, and death as state 4.

This application differs from the earlier on renal disease in two ways. First, there

Table 2.5: Model Inputs: Disease Prevalence and Progression, MI Model

Parameter Name	**Non-Adherent**	**Adherent**
Transition Probability		
Healthy to MI ($p_{1,2}$)	0.7	0.7
Single MI to second MI ($p_{2,3}$)	0.7	0.7
Healthy to Dead ($p_{1,4}$)	0.3	0.3
Single MI to Dead ($p_{2,4}$)	0.3	0.3
Second MI to Dead ($p_{3,4}$)	1	1
Sojourn Distribution Parameters, Gamma Distribution		
Shape Parameter, κ		
Healthy to MI ($\kappa_{1,2}$)	3	4
Single MI to second MI ($\kappa_{2,3}$)	1	2
Healthy to Dead ($\kappa_{1,4}$)	8	10
Single MI to Dead ($\kappa_{2,4}$)	3	4
Second MI to Dead ($\kappa_{3,4}$)	1	1
Scale Parameter, η, (all transitions)	3	3

are multiple paths for which probabilities are needed. We can calculate the probability of each path using the parameters in Table 2.5. For instance, the probability of path $1 \to 2 \to 3 \to$ is $p_{1,2}p_{2,3}p_{3,4} = 0.7 \times 0.7 \times 1 = 0.49$. Second, an MI has an impulse cost, whereas the renal disease model had only sojourn rewards. So, for example, the non-adherent cost $W_{1,2}$ is calculated per Equation (2.20) as

$$W_{1,2}^{C_0} = 35000 + \frac{1}{.03}\left((3500 + 1000 + 2500) - (1000 + 3500)\right) = 118,333.\overline{3}.$$

The expected rewards can be calculated using the transition probabilities and gamma sojourn distribution parameters in Table 2.5 and Equation (2.22). They are: $\mathbb{E}(E_0) = 11.6$, $\mathbb{E}(E_1) = 14.2$, $\mathbb{E}(C_0) = \$117,829$, $\mathbb{E}(C_1) = \$133,766$. So we have $\mu_{ICER} = 5,595\$/\text{QALY}$, or equivalently with $\lambda = 50,000$, an $\mu_{NMB} = \$118,540$. As in the renal case, the treatment will be considered cost-effective. We also did the analysis with Weibull distributions with the first two moments matched to the gamma sojourn distributions, and the results were very similar.

We now turn to an examination of parameter uncertainty and demonstrate the use of Monte Carlo simulation. Specifically, we simulate the distribution of mean rewards as the parameter values in the models change. Since the parameter values

for the gamma distributions have to be positive, we assumed they were distributed according to a lognormal distribution. The transition probabilities were simulated according to a beta distribution. The parameters of these distributions were chosen to reflect the values in Table 2.5.

We simulated $10,000$ patients per path. The densities of $\mathbb{E}(E_1) - \mathbb{E}(E_0)$, $\mathbb{E}(C_1) - \mathbb{E}(C_0)$, μ_{ICER}, and μ_{NMB} (calculated at $\lambda = 50,000$) are shown in Figure (2.4). These distributions capture the uncertainty due to parameter estimation. We see that the estimated densities for the differences in mean costs and effectiveness do not contain 0, providing unambiguous evidence for the presence of a treatment effect for all the reward measures. Also, the support of the density of μ_{NMB} is above 0 and the support of the density of μ_{NMB} is below the willingness-to-pay of $\$50,000$/QALY.

2.5 Combining Subject-Level Results to Population Level

The reward expressions and computations presented so far are for a single subject who starts in a given state and has already spent a certain amount of time there. To obtain population-level results, suppose there are N subjects in the population, and N_s subjects are in state s at the start of the study, $s = 1, 2, ..., K$. Further, suppose the N_s subjects have spent $u_{s,1}, u_{s,2}, ..., u_{s,N_s}$ units of time in state s, for $s = 1, ..., K$. Then, the total (random) reward for the population is just

$$TRP = \sum_{s=1}^{K} \sum_{n=1}^{N_s} R_D(s; u_{s,n}),$$

where $R_D(s; u_{s,n})$ is the discounted reward for an individual subject. The average random reward per subject is

$$\mathbb{E}(TRP) = \sum_{s=1}^{K} \sum_{n=1}^{N_s} W_s R_D(s; u_{s,n}),$$

where $W_s = N_s/N$, the proportion of subjects who are in state s.

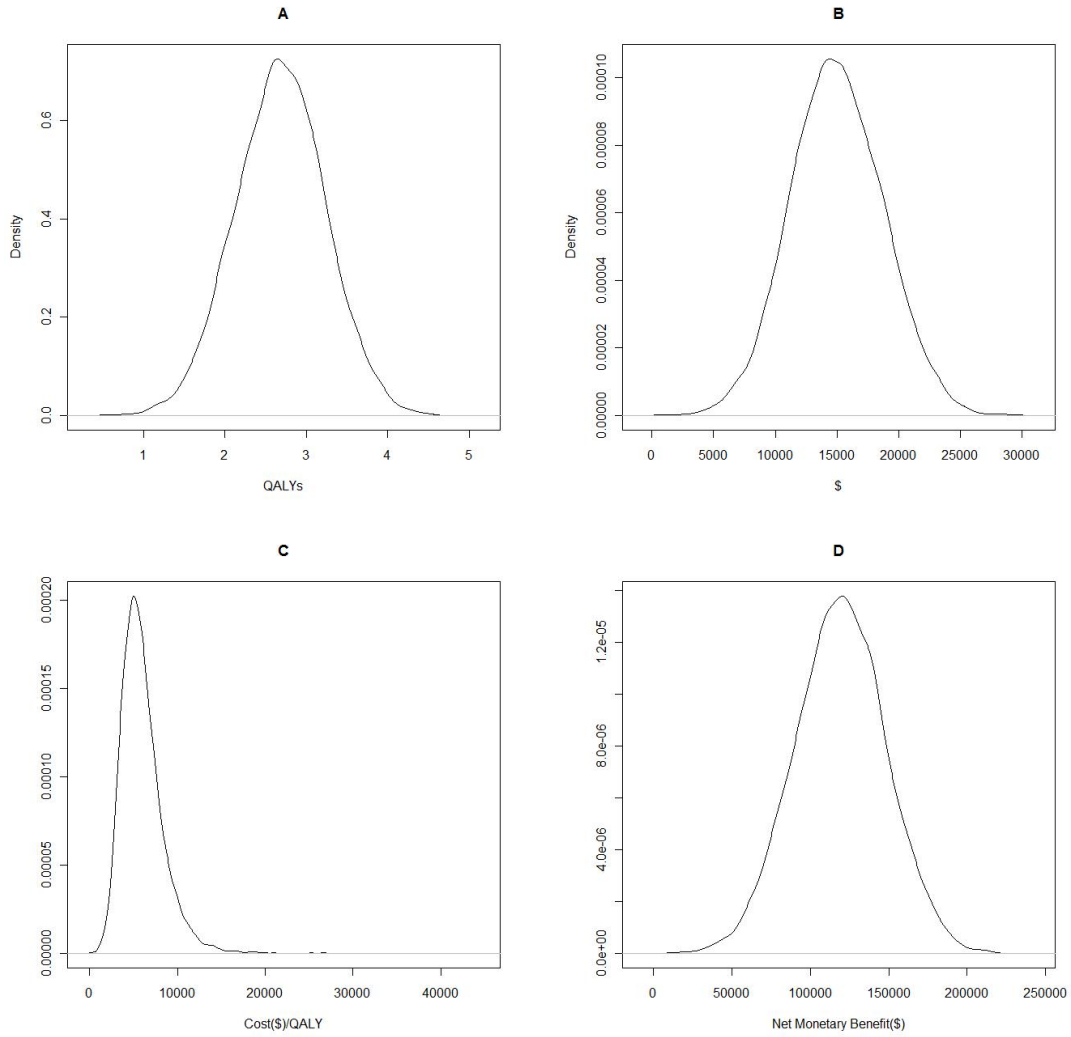

Figure 2.4: Densities of Mean Estimates. Figure A: $\mathbb{E}(E_1) - \mathbb{E}(E_0) = \mu_1(E) - \mu_0(E)$. Figure B: $\mathbb{E}(C_1) - \mathbb{E}(C_0) = \mu_1(C) - \mu_0(C)$. Figure C: μ_{ICER}. Figure D: μ_{NMB}.

In DES, the population of interest in a conceptual one, so one has to select the values for N_s and $u_{s,n}$. The values of N_s are typically selected to represent the profiles of the actual population of interest. The values of $u_{s,n}$ can be simulated using the information about state-occupancy distributions $X_{s,g}; g > s$.

The expected reward expressions can be used to compute average costs and QALYs for the various policy decisions and compute the CEA metrics such as ICER and NMB. As discussed, the unknown parameters for the transition probabilities and

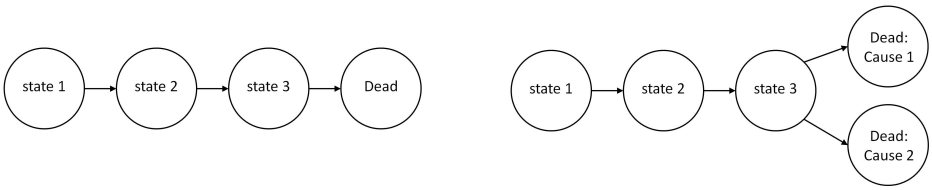

Figure 2.5: Four-state Model with Multiple Causes of Death. *Left panel:* Original 4-state model. *Right panel:* Model with separate causes of death.

expected values of the sojourn times have to be estimated, typically from historical data.

2.6 Extensions

Thus far, we have focused on stationary semi-Markov models. The results extend easily to the case with static covariates (covariates that do not change with time such as gender). Other covariates such as age, weight, and body mass index will vary with time. Including them in the analysis will introduce non-stationarity which is more difficult to handle. We discuss the nonstationary case in this section, along with an extension to competing causes of death.

2.6.1 Multiple Causes of Death: Competing Risks

One is generally interested in analyzing policy decisions in the presence of multiple causes of death. If the costs and QALYs associated with all the causes of death are the same, we can collapse them into one state and use the analysis discussed thus far. In practice, however, this will not be the case, and we have to treat them separately. In fact, there may be other reasons for analyzing the different types of death separately, such as preserving information about failure modes for analysis.

Analysis of multiple causes of death does not pose serious complications for the analysis. The additional causes of death have to added as new states. Then, one has to enumerate the different paths to absorption and use the results in the same way as

before. Some simplifications can be obtained in special cases, such as the single-step case in the right panel of Figure (2.5). Then, only the last component of the lifetime and reward expressions change. Suppose there is one additional cause of death. Then, for the example in Figure (2.5), the life-time expression in Equation (2.3) changes to

$$T(s;u) = \sum_{g=s}^{K-1} (X_{g,g+1}) + \sum_{d=1}^{2} (I_{K,K+d} X_{K,K+d}) \,, \tag{2.23}$$

and the total discounted random reward from equation (2.8) to

$$R_D(s;j) = \alpha V_s + \sum_{g=s}^{K-1} \left[W_{g,g+1} \prod_{m=s}^{g} Z_{m,m+1} \right] + \sum_{d=1}^{2} I_{K,K+d} \left[W_{K,K+d} Z_{K,K+d} \prod_{m=s}^{K-1} Z_{m,m+1} \right] . \tag{2.24}$$

2.6.2 Non-stationary cases

The analysis discussed thus far has been restricted to stationary cases. Non-stationarity can arise from several sources: non-stationary transition probabilities, sojourn distributions, time-varying covariates, or a combination. For example, the sojourn time distributions and the transition probabilities can depend on (calendar) time or 'age' of the patient.

The formulation of semi-Markov processes are general enough to handle these. Recall the general definition of the semi-Markov kernel in Section 2 with $Q_{i,j}(t_1, t_2) = F_{i,j}(t_1, t_2) p_{i,j}(t_1)$ as the semi-Markov kernel. Here t_1 denotes the time at which the subject transitioned from state i to state j. However, it is not possible to obtain analytical expressions for non-stationary cases in general. We discuss alternatives in this section.

Non-stationary sojourn distributions with stationary transition probabilities

Semi-Markov models with non-stationary sojourn distributions but stationary transition probabilities have been referred to as additive semi-Markov processes (Janssen and Manca, 2006, 2007). When the transition probabilities are stationary, the general expressions in Section 2.4 still hold. In particular, the (residual) time-to-failure can continue to be expressed as

$$T(s;u) = \sum_j I_{s,i_2,\ldots,i_{k_j},K+1} \left(X_{s,i_2}(u) + X_{i_2,i_3} + \cdots + X_{i_{k_j},K+1} \right), \tag{2.25}$$

and the expressions for the random rewards are also the same. However, the random variables $X_{i_g,i_{g+1}}$'s are no longer independent; in particular, their distributions depend on the times at which the subject transitioned from state i_g to i_{g+1}. To be specific, consider a path $\{s < i_2 < ... < i_{k_j} < K+1\}$, and let

$$S_{i_m}(u) = X_{s,i_2}(u) + \cdots + X_{i_{m-1},i_m} \tag{2.26}$$

be the time at which the subject transitioned into state i_m. Then, the distribution of $X_{i_m,i_{m+1}}$ depends on $S_{i_m}(u)$. For example, the distributions could be exponential with mean $\exp(\beta_{i_m,i_{m+1}} S_{i_m})$.

Because of the dependence structure, the expectations of the lifetime and rewards cannot be derived in analytical form in general. But it is easy to get these values through direct Monte Carlo simulation. For each possible path, one first simulates $X_{s,i_2}(u)$, then the conditional random variables $[X_{i_2,i_3}|X_{s,i_2}(u)]$, $[X_{i_3,i_4}|X_{s,i_2}(u), X_{i_2,i_3}]$ etc. This process is repeated N times to get a Monte Carlo average, repeated for all possible paths and weighted using the transition probabilities to get the overall expected rewards. This is still much easier than setting up a discrete-event simulation framework. It also has the advantage that all possible paths are sampled and weighted appropriately in the computation of the CEA metrics.

General non-homogeneous case

We turn now to the more general situation where both the transition probabilities and the sojourn time distributions are non-stationary. Situations with time-varying covariates (such as health status, age, etc.) also fall into this category.

We can write the time-to-failure as in Section 2.4,

$$T(s;u) = I_{s,J_2} \times I_{J_2,J_3} \cdots \times I_{J_L,K+1} \left(X_{s,J_2}(u) + X_{J_2,J_3} + \cdots + X_{J_L,K+1} \right), \tag{2.27}$$

but now both the indicator functions for the transitions and the sojourn times depend on the $S_{i_m}(u)$'s in equation (2.26). We can take the expectation of $T(s,u)$ by first conditioning on the sojourn times and then taking the expectation with respect to the distributions of the sojourn times. Doing this allows us to represent the expectation as

(2.28)

$$\begin{aligned}\mathbb{E}[T(s;u)|\text{sojourn times}] = \sum_j & p_{s,i_2}(S_s(u))p(i_2,i_3)(S_{i_2}(u)) \times \cdots \times p_{i_{k_j},K+1}(S_i k_j) \\ & \times \left(X_{s,i_2}(u) + X_{i_2,i_3} + \cdots + X_{i_{k_j},K+1} \right),\end{aligned}$$

where $p(i_2,i_3)(S_{i_2}(u)) = \mathbb{E}[I_{i_2,i_3}(S_{i_2}(u))|\text{sojourn times}]$. A corresponding expression will hold for the random rewards. Note that the sum is still over all possible paths. These conditional transition probabilities have explicit forms as they are given by the multi-state model. For example, one possible form is $p(i_2,i_3)(t) = 1 - \exp(t/\eta_{i_1,i_2})$, indicating that the probability of transition increases with time.

Again, it is not possible to obtain analytical expressions for the expected lifetime and rewards in general. But the direct Monte Carlo method outlined in the last subsection extends readily to this situation also.

Comparison with Discrete-Event Simulation

The advantages of the approaches developed in the paper are clear when there are analytical expressions for the expected rewards. There is no need to set up a discrete event simulation framework and simulate sample paths through the multi-state. The advantages are less evident when analytical expressions are not available and one still has to use Monte Carlo methods to compute the expectations. Nevertheless, the direct approaches discussed for the non-stationary cases are easier to work with and implement than simulating a large-scale multi-state model. The only computational challenge is in enumerating all the possible paths through the system. This is not a difficult problem.

A more important advantage is that the approaches outlined here evaluate the random variables for each possible path, weight them according to the probability of using that path and obtain an overall average. In DES, on the other hand, not all possible paths will be used; rather the paths are sampled in each simulation replication according to their likelihood, so paths with very small probabilities have a low probability of being sampled. This may be fine for the purposes of getting expected lifetimes but not for computing rewards. Small probability paths may have high costs.

Also, the approaches described here examine each of these paths, giving us direct information about all possible elements that go into the average calculations. The comparison between DES and the approaches here are similar to the comparison between simple random and stratified sampling. The latter sample from each stratum (mixture component or path in our set up) and hence provide detailed information about the population of interest as well as global information about the average.

2.7 Application

This section combines the renal disease model and the CVD model previously discussed so that we can illustrate the results on a more complex application. This example is very similar to the application in Rosen et al. (2005), but it has been modified to be progressive, stationary, and with sojourn times that are distributed according to the gamma distribution. Further, only deaths due to renal disease or MI are considered; competing risk deaths, which vary with time, are not included.

Figure (2.6) shows part of the multi-step process. The mode of patient death determines the cost of the death event; in particular, there is an impulse cost associated with dying of MI, as reflected in Table 2.7. As a result, the two modes of death are represented as separate health states, as shown in Figure (2.6), and we use the competing risk extension discussed in Section 2.6. All patients enter the process in the Normoalbuminuria (healthy) state. To simplify the figure, we have not shown included additional paths from the relatively healthy states to death caused by MI; but any patient may die due to MI from any health state. Not all transitions between states in the model are possible; for instance, a patient cannot progress directly from Normoalbuminuria to ESRD. Thus, while in theory there are $2^{14-2} = 4096$ patient paths, there are only 49 paths with positive probability. Even 4096 can be handled with relative ease if we have analytical solutions. The full set of probabilities and the shape parameters for the gamma sojourn-time distributions, $\kappa_{i,j}$, for all transitions in the full model with positive probability are given in Table 2.6. All transitions are assumed to have scale parameter $\eta = 3$, and all patients are assumed to start with Normoalbuminuria and no history of MI. In the table, we abbreviate the Normoalbuminuria state as Normo, Microalbuminuria as Micro, and Macroalbuminuria

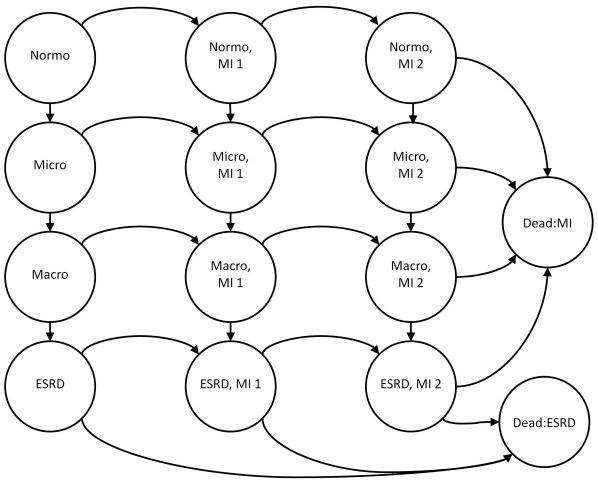

Figure 2.6: Application Structure.

as Macro. Additionally, MI_1 and MI_2 denote surviving 1 and 2 MIs, respectively. Finally, D_{MI} denotes death due to MI, and D_{ESRD} denotes death due to ESRD.

The probabilities and sojourn time distribution parameters are more involved in this bigger model. One reason is that the progression of renal disease affects the progression of cardiovascular disease; notably, those with Macroalbuminuria progress more quickly to an MI. Additionally, in the earlier cases with just renal disease or CVD, medication adherence affected only the sojourn time distribution and not the transition probabilities; in this combined model, transition probabilities also change as a result of medication adherence. In general, adherence to medication increases the time to any event, and decreases the probability of fatal cardiovascular disease events. Adherence also increases the probability that the next state transition will be into the next renal disease state, rather than into the next cardiovascular disease state.

Table 2.6: Model Inputs: Disease Prevalence and Transition Probabilities, Full Model

	Transition Probability		**Scale Parameter, κ**	
Parameter Name	**Non-Adherent**	**Adherent**	**Non-Adherent**	**Adherent**
Normo, MI_1	0.32	0.4	3	4
Normo to D_{MI}	0.14	0.20	8	10
Normo to Micro	0.55	0.34	2	6
Normo, MI_1 to Normo, MI_1	0.48	0.59	1	2
Normo, MI_1 to D_{MI}	0.20	0.25	3	4
Normo, MI_1 to Micro, MI_1	0.32	0.17	2	6
Normo, MI_2 to D_{MI}	0.68	0.83	1	1
Normo, MI_2 to Micro, MI_2	0.32	0.17	2	6
Micro to Micro, MI_1	0.19	0.37	3	4
Micro to D_{MI}	0.08	0.16	8	10
Micro to Macro	0.73	0.47	1	4
Micro, MI_1 to Micro, MI_2	0.34	0.52	1	2
Micro, MI_1 to D_{MI}	0.15	0.22	3	4
Micro, MI_1 to Macro, MI_1	0.52	0.52	1	4
Micro, MI_2 to D_{MI}	0.49	0.74	1	1
Micro, MI_2 to Macro, MI_2	0.52	0.26	1	4
Macro to Macro, MI_1	0.41	0.44	1	2
Macro to D_{MI}	0.17	0.19	3	4
Macro to ESRD	0.42	0.37	1	2
Macro, MI_1 to Macro, MI_2	0.54	0.57	1	1
Macro, MI_1 to D_{MI}	0.23	0.24	1	1
Macro, MI_1 to ESRD, MI_1	0.23	0.19	1	2
Macro, $MI_2$to D_{MI}	0.77	0.85	0.3	1
Macro, MI_2 to ESRD, MI_2	0.23	0.15	1	2
ESRD to ESRD, MI_1	0.17	0.13	1	2
ESRD to D_{MI}	0.07	0.06	3	4
ESRD to D_{ESRD}	0.76	0.81	0.3	0.3
ESRD, MI_1 to ESRD, MI_2	0.31	0.25	1	1
ESRD, MI_1 to D_{MI}	0.13	0.14	1	1
ESRD, MI_1 to D_{ESRD}	0.56	0.60	0.3	0.3
ESRD, MI_2 to D_{MI}	0.44	0.56	0.3	1
ESRD, MI_2 to D_{ESRD}	0.38	0.62	0.3	0.3

The reward structure is shown in Table 2.7. There is one complication in the reward structure for the combined model because a patient's health state is defined by a combination of the renal disease state and the CVD state. As a result, both costs and QALYs are dependent upon the rewards for each disease. QALYs are multiplicative across the two disease (so an individual with Normoalbuminuria and a history MI would have an overall utility of 0.88×0.88), and costs are additive, so the same individual would have a sojourn cost per unit time of $1000 + 3500 + 2500$.

Table 2.7: Model Inputs: Disease Utilities, Costs, and Discount Rate

QALYs	
Diabetes (Baseline Health)	0.88
ESRD	0.61
MI	0.88
Annual costs	**$**
Treatment Adherence	300
Ongoing cost of care	3500
Diabetes, Normoalbuminuria	1000
Diabetes with Microalbuminuria	2000
Diabetes with Macroalbuminuria	4000
ESRD	80000
History of MI	2500
Acute MI (transition cost)	35000
Discount Rate, %	3

The expected reward and probability of each path can be calculated in a manner similar to in the CVD case, but for the full set of 49 paths this time. The result is $\mathbb{E}(E_0) = 12.1$, $\mathbb{E}(E_1) = 17.1$, $\mathbb{E}(C_0) = \$151,307$, $\mathbb{E}(C_1) = \$153,180$, so $\mu_{ICER} =$ 375 \$/QALY, and $\mu_{NMB} = \$243,868$, resulting in the intervention being very cost-effective even for the combined model.

We can use the Monte Carlo simulation of the Taylor Series method introduced in the multi-step section to determine the uncertainty due to parameter estimation. The set up is the same as in the CVD case. The densities of $\mathbb{E}(E_1)-\mathbb{E}(E_0)$, $\mathbb{E}(C_1)-\mathbb{E}(C_0)$, μ_{ICER}, and μ_{NMB} (calculated at $\lambda = 50,000$) are shown in Figure (2.7). The density for $\mathbb{E}(C_1) - \mathbb{E}(C_0)$ contains 0, so the evidence on whether the intervention reduces cost is a bit ambiguous. This is also reflected in the distribution for μ_{ICER}. However, the support of the density for μ_{NMB} is quite a bit to the right of 0, so based on this metric, one would conclude that the treatment is effective even after taking into account parameter uncertainty.

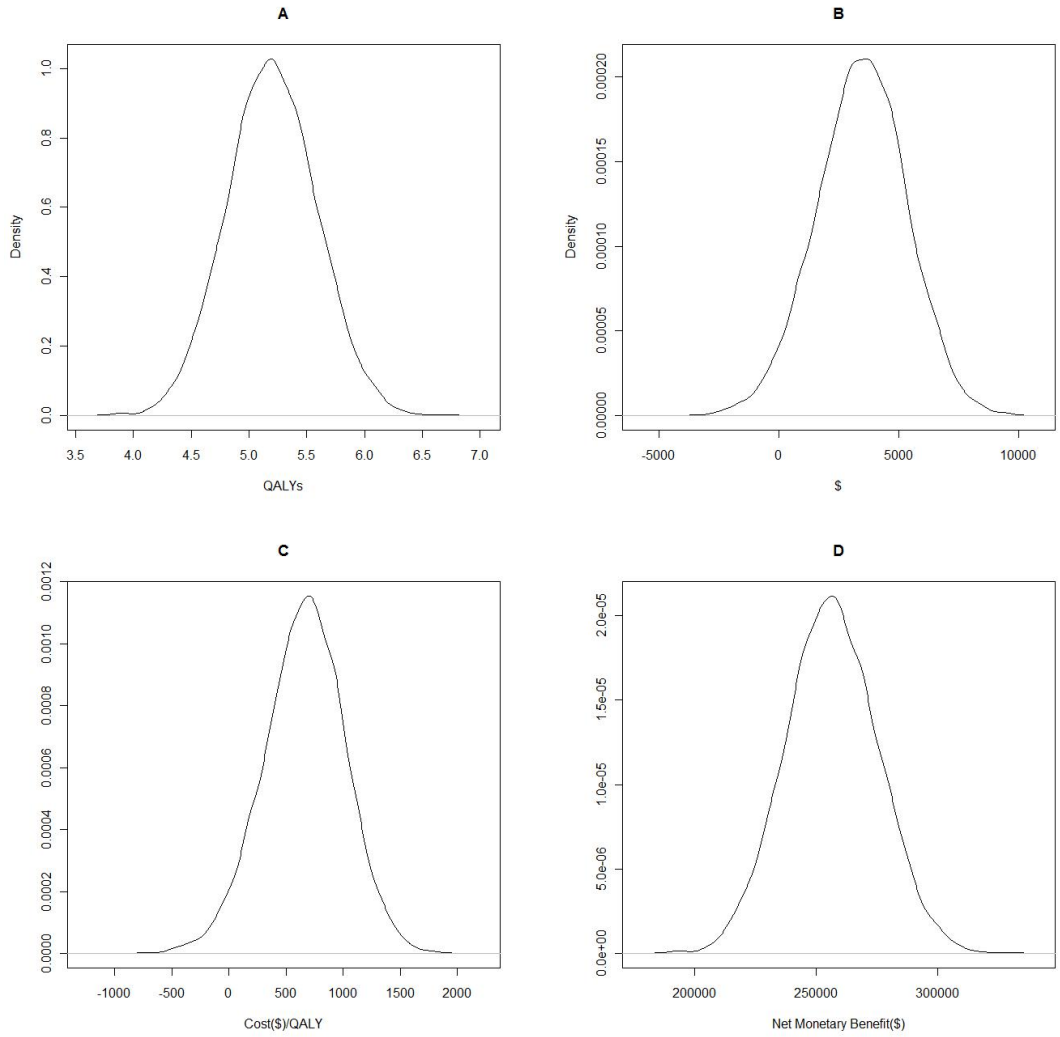

Figure 2.7: Densities of Mean Estimates. Figure A: $\mathbb{E}(E_1) - \mathbb{E}(E_0) = \mu_1(E) - \mu_0(E)$. Figure B: $\mathbb{E}(C_1) - \mathbb{E}(C_0) = \mu_1(C) - \mu_0(C)$. Figure C: μ_{ICER}. Figure D: μ_{NMB}.

2.8 Conclusion

This paper has developed analytical methods for computing cost-effectiveness metrics with stationary semi-Markov processes. For non-stationary cases, direct Monte-Carlo simulation has been proposed as an alternative to discrete event simulation. These methods are considerably easier to implement and use with CEA. Additional advantages have been noted in the paper. Common aspects of CEA, including sensitivity analysis, are also easily performed within this framework, and

were demonstrated in this paper. Some common extensions, to multiple causes of death and to non-stationary models, were also demonstrated.

While the results in the paper have been restricted to progressive models, they can be extended to non-progressive (recurrent) models. As a practical matter, however, nearly all disease models can be built to be progressive through the addition of more states. While analysts have previously been reluctant to expand the state-space due to computational concerns in traditional DES, state-space expansion is less of a concern with the methods in this paper.

2.9 Appendix

2.9.1 Proof of Proposition 1

Single-step Discrete case

We consider the discrete case first. Let $f_{g,g+1}(x),\ s < g \leq K$, be the probability mass function (pmf) of $X_{g,g+1}$ and $f_{s,s+1}(x;u)$ be the conditional pmf of $[X_{s,s+1}|X_{s,s+1} \geq u]$.

We compute the total rewards associated with the fixed components, $U_{s,s+1}$ and $U_{g,g+1},\ \ s < g \leq K$, that are incurred at the time of transition. The subject starts in state s and spends an amount of time $X_{s,s+1}(u)$ in state s before moving to state $s+1$, at which time s/he incurs a reward of $U_{s,s+1}$. So this amount has to be discounted by $\theta^{X_{s,s+1}(u)}$ (recall that $\theta = \frac{1}{1+r}$). The subject then spends $X_{s+1,s+2}$ time units in state $s+1$ before moving to state $s+2$. So the reward $U_{s+1,s+2}$ has to be discounted by $\theta^{X_{s,s+1}(u)+X_{s+1,s+2}}$. Repeating this argument, we get the fixed reward (reward due to transitions) over the remaining life of a subject as

$$\begin{aligned} FR_D(s;u) = U_{s,s+1}\theta^{X_{s,s+1}(u)} + U_{s+1,s+2}\theta^{(X_{s,s+1}(u)+X_{s+1,s+2})} + ... \\ + U_{K,K+1}\theta^{(X_{s,s+1}(u)+X_{s+1,s+2}+...+X_{K,K+1})}. \end{aligned} \tag{2.29}$$

Define $Z_{s,s+1} = \theta^{X_{s,s+1}(u)}$ and $Z_{m,m+1} = \theta^{X_{m,m+1}}$ for $m \geq s+1$. Then, we can write the discounted fixed reward $FR_D(s;u)$ in a compact form as

$$FR_D(s;u) = \sum_{g=s}^{K} \left(U_{g,g+1} \prod_{m=s}^{g} Z_{m,m+1} \right). \tag{2.30}$$

Consider now the variable reward that depends on the amount of time spent in the different states (sojourn time). Again, the subject starts in state s and spends an additional $X_{s,s+1}(u)$ time units in state s. So the total discounted reward for this period is $V_s(\theta^1 + ... + \theta^{X_{s,s+1}(u)})$, which can be expressed as $V_s\theta(1-\theta^{X_{s,s+1}(u)})/(1-\theta)$. The subject then moves to state $s+1$ and spends an additional $X_{s+1,s+2}$ time units. So the total discounted reward for this period is $V_{s+1}\theta^{X_{s,s+1}(u)}(\theta^1 + \cdots + \theta^{X_{s+1,s+2}(u)})$ which can be represented as $V_{s+1}\theta^{X_{s,s+1}(u)}(1-\theta^{X_{s+1,s+2}(u)})/(1-\theta)$. So the total variable reward over the remaining life of the subject is

(2.31)

$$\begin{aligned} VR_D(s;u) &= \frac{\theta}{(1-\theta)} \left(V_s(1-\theta^{X_{s,s+1}(u)}) + (V_{s+1}\theta^{X_{s,s+1}(u)})(1-\theta^{X_{s+1,s+2}}) + \cdots \right. \\ &\quad \left. + (V_K\theta^{(X_{s,s+1}(u)+X_{s+1,s+2}+...+X_{K-1,K})}(1-\theta^{X_{K,K+1}})) \right) \\ &= \frac{\theta}{(1-\theta)} \left(V_s + (V_{s+1}-V_s)Z_{s,s+1} + (V_{s+2}-V_{s+1})Z_{s,s+1}Z_{s+1,s+2} + \cdots \right. \qquad (2.32) \\ &\quad \left. + (V_K - V_{K-1})Z_{s,s+1}...Z_{K-1,K} - (V_K Z_{s,s+1}...Z_{K,K+1}) \right) \\ &= \frac{\theta}{(1-\theta)} \left(V_s + \sum_{g=s}^{K} (V_{g+1}-V_g) \prod_{m=s}^{g} Z_{m,m+1}] \right) \qquad (2.33) \end{aligned}$$

where $V_{K+1} = 0$.

The fixed and variable costs can be combined to get the total (random) reward as follows. Let

$$W_{g,g+1} = U_{g,g+1} + \frac{\theta}{(1-\theta)}(V_{g+1} - V_g).$$

Then, the total discounted random reward $R_D^u(s)$ can be written in a compact form

as

$$R_D(s;u) = \frac{\theta}{(1-\theta)} V_s + \sum_{g=s}^{K} [W_{g,g+1} \prod_{m=s}^{g} Z_{m,m+1}]. \tag{2.34}$$

This proves part (a) of Proposition 1 for the discrete case.

From the independence of the $X_{m,m+1}$'s, we get the expected discounted reward as

$$\mathbb{E}(R_D(s;u)) = \frac{\theta}{(1-\theta)} V_s + \sum_{g=s}^{K} [W_{g,g+1} \prod_{m=s}^{g} \mathbb{E}(Z_{m,m+1})]. \tag{2.35}$$

Let $L_{m,m+1}(\gamma)$ be the Laplace transform of $X_{m,m+1}$ for $m > s$ and of $X_{s,s+1}(u)$ for $m = s$. Then, $\mathbb{E}(Z_{i,j}) = \mathbb{E}(\theta^{X_{m,m+1}}) = L_{m,m+1}(-\log(\theta))$. This proves part (b) of Proposition 1 for the discrete case.

Single-step Continuous Case

The proofs for the continuous case are very similar to the discrete case. For completeness, we give some of the details below. Let $f_{g,g+1}(x)$ be the probability density function (pdf) of $X_{i,j}$, $s < g \leq K$, and $f_{s,s+1}(u)$ be the conditional pdf of $[X_{s,s+1}|X_{s,s+1} \geq u]$. The main difference between the discrete and continuous cases is in the discounting of the rewards. The discounting factor in continuous time is $\exp(-\rho x)$, for a time of length x with continuous discount rate ρ. Again, let the rewards associated with the fixed components be $U_{g,g+1}$ and $U_{s,s+1}$, incurred at transition. As before, the subject starts at state s and spends $X_{s,s+1}(u)$ time units in state s before moving to state $s+1$. $U_{s,s+1}$ is discounted at the time of the transition, to get $U_{s,s+1} \exp(-\rho X_{s,s+1}(u))$. The subject then spends $X_{s+1,s+2}$ time units in state $s+1$ before moving to state $s+2$, so the resulting total discounted reward from the second transition is $U_{s+1,s+2} \exp(-\rho(X_{s,s+1}(u) + X_{s+1,s+2}))$. We see that the form of the discounting is the same as the discrete case except that θ^x is now replaced by $\exp(-\rho x)$.

Define $Z_{m,m+1} = \exp(-\rho X_{m,m+1})$ for $m > s$ and $Z_{s,s+1} = \exp(-\rho X_{s,s+1}(u))$. Then, as in the discrete case, we can write the discounted fixed reward $FR_D(s;u)$ compactly as

$$FR_D(s;u) = \sum_{g=s}^{K} \left(U_{g,g+1} \prod_{m=s}^{g} Z_{m,m+1} \right). \tag{2.36}$$

Following the same arguments as in the discrete case, we can write total variable reward over the remaining life of the subject compactly as

$$VR_D(s;u) = \frac{1}{\rho} \left(V_s + \sum_{g=s}^{K} [(V_{g+1} - V_g) \prod_{m=s}^{g} Z_{m,m+1}] \right) \tag{2.37}$$

where $V_{K+1} = 0$.

Let

$$W_{g,g+1} = U_{g,g+1} + \frac{1}{\rho}(V_{g+1} - V_g).$$

The fixed and variable costs can be combined to get the total (random) reward $R_D(s;u)$ as

$$R_D(s;u) = \frac{1}{\rho} V_s + \sum_{g=s}^{K} [W_{g,g+1} \prod_{m=s}^{g} Z_{m,m+1}]. \tag{2.38}$$

This proves part (a) of Proposition 1 for the continuous case.

From this, we see that expected total discounted reward in the continuous case has the same form as the discrete case:

$$\mathbb{E}(R_D(s;u)) = \frac{1}{\rho} V_s + \sum_{g=s}^{K} [W_{g,g+1} \prod_{m=s}^{g} \mathbb{E}(Z_{m,m+1})]. \tag{2.39}$$

Let $L_{m,m+1}(\gamma)$ be the Laplace transform of $X_{m,m+1}$ for $m > s$ and of $X_{m,m+1}(u)$ for $m = s$. Then, $\mathbb{E}(Z_{m,m+1}) = \mathbb{E}(\exp(-\rho X_{m,m+1})) = L_{m,m+1}(\rho)$. This proves part (b) of Proposition 1 for the continuous case.

2.9.2 Proof of Proposition 2

Multi-step Discrete Case

Again, we start with the discrete case. Let $f_{i_g,i_{g+1}}(x)$ be the pdf of X_{i_g,i_g+1} $2 \leq g \leq k_j$ and $f_{s,i_2}(x;u)$ is the conditional pdf of $[X_{s,i_2}|X_{s,i_2} \geq u]$. Due to the sojourn time distribution being discrete, $Z_{i_m,i_{m+1}} = \theta^{X_{i_m,i_{m+1}}}$ for $m > 2$ and $Z_{s,i_2} = \theta^{X_{s,i_2}(u)}$ for $m = 1$.

Take any of the possible sample path: $\{s, i_2, ..., i_{k_j}, K+1\}$. The discounted reward random variable can be calculated in exactly the same manner as in the single-step case. Combining the results for the different paths, we get the overall fixed reward as

$$\begin{aligned} FR_D(s;u) = \sum_j I^{(j)}_{s,i_2,\cdots,i_{k_j},K+1} U_{s,i_2} \left(Z_{s,i_2} + U_{i_2,i_3} Z_{s,i_2} Z_{i_2,i_3} + \cdots \right. \\ \left. + U_{i_{k_j},K+1}(Z_{s,i_2} Z_{i_2,i_3} \cdots Z_{i_{k_j},K+1})\right). \end{aligned} \tag{2.40}$$

The same arguments can be used to get the variable reward over the remaining life of the subject as

$$VR_D(s;u) = \sum_j I^{(j)}_{s,i_2,\cdots,i_{k_j},K+1} \frac{\theta}{1-\theta} \left(V_s + \sum_{g=1}^{k} (V_{i_g} - V_{i_{g+1}}) \prod_{m=1}^{g} Z_{i_m,i_{m+1}}\right), \tag{2.41}$$

where $i_1 = s$ and $i_{k+1} = K+1$.

Let $W_{i_g,i_{g+1}} = U_{i_g,i_{g+1}} + \frac{\theta}{(1-\theta)}(V_{i_{g+1}} - V_{i_g})$. Combining the fixed and variable rewards, we get the total discounted random reward as

$$R_D(s;u) = \sum_j I^{(j)}_{s,i_2,\cdots,i_{k_j},K+1} \left(V_s + \sum_{g=1}^{k} [W_{i_g,i_{g+1}} \prod_{m=1}^{g} Z_{i_m,i_{m+1}}]\right), \tag{2.42}$$

where $i_1 = s$ and $i_{k+1} = K+1$. This proves part (a) of Proposition 2 for the discrete case.

From the independence of the $X_{i,j}$'s, we get the expected discounted reward as

$$\mathbb{E}(R_D(s;u)) = \sum_j p_{s,i_2,\cdots,i_{k_j},K+1} \left(V_s + \sum_{g=1}^{k} [W_{i_g,i_{g+1}} \prod_{m=1}^{g} E(Z_{i_m,i_{m+1}})]\right). \tag{2.43}$$

Let $L_{i_m,i_{m+1}}(\gamma)$ be the Laplace transform of $X_{i_m,i_{m+1}}$ for $m \geq 2$ and of $X_{s,i_2}(u)$ for $m = 1$. Then, $\mathbb{E}(Z_{i_m,i_{m+1}}) = \mathbb{E}(\theta^{X_{i_m,i_{m+1}}}) = L_{i_m,i_{m+1}}(-\log(\theta))$. This proves part (b) of Proposition 2 for the discrete case.

Multi-step Continuous Case

The results for the continuous case follow easily from the discrete multi-step case. Recall that $f_{i_g,i_{g+1}}(x)$ is the pdf of X_{i_g,i_g+1} $2 \leq g \leq k_j$ and $f_{s,i_2}(x;u)$ is the conditional pdf of $[X_{s,i_2}|X_{s,i_2} \geq u]$. Due to the sojourn time distribution being continuous, $Z_{i_m,i_{m+1}} = \exp(-\rho X_{i_m,i_{m+1}})$ for $m > 2$ and $Z_{s,i_2} = \exp(-\rho X_{s,i_2}(u))$ for $m = 1$.

The discounted random variable for the fixed rewards along a path j has the same form as in the case with the discrete-time sojourn distribution:

$$FR_D^u(s) = \sum_j I^{(j)}_{s,i_2,\cdots,i_{k_j},K+1} \left(U_{s,i_2}Z_{s,i_2} + U_{i_2,i_3}Z_{s,i_2}Z_{i_2,i_3} + \cdots \right.$$
$$\left. + U_{i_{k_j},K+1}(Z_{s,i_2}Z_{i_2,i_3}\cdots Z_{i_{k_j},K+1})\right). \tag{2.44}$$

For the variable rewards, the random variable is

$$VR_D^u(s) = \sum_j I^{(j)}_{s,i_2,\cdots,i_{k_j},K+1} \frac{1}{\rho}\left(V_s + \sum_{g=1}^{k}(V_{i_g} - V_{i_{g+1}}) \prod_{m=1}^{g} Z_{i_m,i_{m+1}}\right), \tag{2.45}$$

where $i_1 = s$ and $i_{k+1} = K + 1$.

Now, let $W_{i_g,i_{g+1}} = U_{i_g,i_{g+1}} + \frac{1}{\rho}(V_{i_{g+1}} - V_{i_g})$. The total discounted random reward is thus

$$R_D^u(s) = \sum_j I^{(j)}_{s,i_2,\cdots,i_{k_j},K+1} \left(V_s + \sum_{g=1}^{k}[W_{i_g,i_{g+1}} \prod_{m=1}^{g} Z_{i_m,i_{m+1}}]\right), \tag{2.46}$$

This proves part (a) of Proposition 2 for the continuous case.

The expected discounted reward as

$$\mathbb{E}(R_D^u(s)) = \sum_j p^{(j)}_{s,i_2,\cdots,i_{k_j},K+1} \left(V_s + \sum_{g=1}^{k}[W_{i_g,i_{g+1}} \prod_{m=1}^{g} E(Z_{i_m,i_{m+1}})]\right). \tag{2.47}$$

Let $L_{i_m,i_{m+1}}(\gamma)$ be the Laplace transform of $X_{i_m,i_{m+1}}$ for $m \geq 2$ and of $X_{s,i_2}(u)$ for $m = 1$. Then, $\mathbb{E}(Z_{i,j}) = \mathbb{E}(\exp(-\rho X_{i_m,i_{m+1}})) = L_{i_m,i_{m+1}}(\rho)$. This proves part (b) of Proposition 2 for the continuous case.

2.9.3 Variances of the Discounted Rewards

We have thus far ignored the calculation of the discounted reward variance, since expected value is the normal decision-making criterion in CEA. Nevertheless, the variance of the discounted rewards can also be calculated analytically, and the general results are in this section.

One-step Case

Recall that in the case with discrete time, the rewards are discounted by a factor r per unit time, and $\theta = \frac{1}{(1+r)}$. For continuous time, the discount rate is ρ, so the discount for a time of length x is $\exp(-\rho x)$. As before, let $W_{g,g+1} = U_{g,g+1} + \alpha(V_{g+1} - V_g)$, where $\alpha = \frac{\theta}{(1-\theta)}$ for discrete time and $\alpha = 1/\rho$ for continuous time. Further, in the discrete case, $Z_{s,s+1} = \theta^{X_{s,s+1}(u)}$ and $Z_{m,m+1} = \theta^{X_{m,m+1}}$ for $m \geq s+1$. In the continuous case, $Z_{s,s+1} = \exp(-\rho X_{s,s+1}(u))$ and $Z_{m,m+1} = \exp(-\rho X_{m,m+1})$ for $m \geq s+1$.

After some algebraic manipulations, we get the variance of $R_D^u(s)$ as

$$
\begin{aligned}
Var(R_D(s;u)) = & \sum_{g=s}^{K} W_{g,g+1}^2 \prod_{m=s}^{g} \mathbb{E}(Z_{m,m+1}^2) \\
& + 2 \sum_{g=s}^{K} \sum_{n=g+1}^{K} W_{g,g+1} W_{n,n+1} \prod_{m=s}^{g} \mathbb{E}(Z_{m,m+1}^2) \prod_{\ell=g+1}^{n} \mathbb{E}(Z_{\ell,\ell+1}) \\
& - \left(\sum_{g=s}^{K} [W_{g,g+1} \prod_{m=s}^{g} \mathbb{E}(Z_{m,m+1}) \right)^2 .
\end{aligned} \tag{2.48}
$$

Let $L_{m,m+1}(\gamma)$ be the Laplace transform of $X_{m,m+1}$ for $m > s$ and of $X_{s,s+1}(u)$ for $m = s$. Then, $\mathbb{E}(Z_{m,m+1}) = \mathbb{E}(\theta^{X_{m,m+1}}) = L_{m,m+1}(-\log(\theta))$, $\mathbb{E}(Z_{m,m+1}^2) =$

$L_{m,m+1}(-2\log(\theta))$. For the continuous case, $\mathbb{E}(Z_{m,m+1}) = L_{m,m+1}(\rho)$ and $\mathbb{E}(Z^2_{m,m+1}) = L_{m,m+1}(2\rho)$.

Multi-step Case

Call R_j the reward earned on path j and T_j the time-to failure on path j in the multi-step case, and p_j the probability of path j, and suppose there are J total disease paths. From standard properties of mixture distributions (McLachlan and Peel, 2000),

$$Var(R(s;u)) = \sum_{j=1}^{J} p_j \left(Var(T_j) + (\mathbb{E}(T_j))^2\right) - \left(\sum_{j=1}^{J} p_j \mathbb{E}(T_j)\right)^2, \tag{2.49}$$

where $Var(T_j)$ and $\mathbb{E}(T_j)$ are calculated as in the single-step case. Additionally

$$Var(R(s;u)) = \sum_{j=1}^{J} p_j \left(Var(R_j) + (\mathbb{E}(R_j))^2\right) - \left(\sum_{j=1}^{J} p_j \mathbb{E}(R_j)\right)^2. \tag{2.50}$$

The variance of $R(s;u)$ can be calculated using this equation and the independence of the $X_{i_g,i_{g+1}}$'s and is given by

$$\begin{aligned} Var[R^u_D(s)] = \sum_j p_{s,i_2,\ldots i_{k_j},K+1} \Bigg(& \sum_{g=1}^{K} W^2_{i_g,i_{g+1}} \prod_{m=1}^{g} \mathbb{E}(Z^2_{i_m,i_{m+1}}) \\ & + 2\sum_{g=1}^{K}\sum_{n=g+1}^{K} \left[W_{i_g,i_{g+1}} W_{i_n,i_{n+1}} \prod_{m=1}^{g} \mathbb{E}(Z^2_{i_m,i_{m+1}}) \prod_{\ell=g+1}^{n} \mathbb{E}(Z_{i_\ell,i_{\ell+1}}) \right] \\ & + \left(\alpha V_s + \sum_{g=s}^{K} [W_{i_g,i_{g+1}} \prod_{m=1}^{g} \mathbb{E}(Z_{i_m,i_{m+1}}) \right)^2 \Bigg) \\ & - \left(\sum_j p_{s,i_2,\ldots i_{k_j},K+1} \left(\alpha V_s + \sum_{g=s}^{K} [W_{i_g,i_{g+1}} \prod_{m=1}^{g} \mathbb{E}(Z_{i_m,i_{m+1}}) \right) \right)^2. \end{aligned} \tag{2.51}$$

Where again, let $L_{i_m,i_{m+1}}(\gamma)$ be the Laplace transform of $X_{i_m,i_{m+1}}$ for $m \geq 2$ and of $X_{s,i_2}(u)$ for $m = 1$. Then, $\mathbb{E}(Z_{i_m,i_{m+1}}) = \mathbb{E}(\theta^{X_{i_m,i_{m+1}}}) = L_{i_m,i_{m+1}}(-\log(\theta))$, $\mathbb{E}(Z^2_{i_m,i_{m+1}}) = L_{i_m,i_{m+1}}(-2\log(\theta))$. For the continuous case, $\mathbb{E}(Z_{i_m,i_{m+1}}) = L_{i_m,i_{m+1}}(\rho)$ and $\mathbb{E}(Z^2_{i_m,i_{m+1}}) = M_{i_m,i_{m+1}}(2\rho)$.

CHAPTER III

Modeling Cost-Effectiveness Data for Medical Decision Making: A Statistical Framework

3.1 Introduction

Cost-effectiveness analysis (CEA) is part of the broad field of comparative effectiveness analysis for health/medical decision making. CEA is designed to assess the comparative value of expenditures on different health interventions (Gold et al., 1996). The conceptual basis is that, for a given level of available resources, the interventions that provide the greatest health value should be selected. As the name implies, CEA requires a comparison two types of health outputs: cost and effectiveness.

Cost-effectiveness analyses have been conducted for a variety of treatments in a range of population for diseases including depression (Valenstein et al., 2001), asthma (Paltiel et al., 2001), HIV (Long et al., 2006), cardiovascular disease (Liew, 2006; Rosen et al., 2005), stroke (Matchar et al., 2005), and colorectal cancer (Castelli et al., 2007). The UK, Australia, and Canada require economic evaluations before approving new healthcare technologies (Birch and Gafni, 2004; Hill et al., 2000; Hjelmgren et al., 2001), although CEA is not required before the introduction of a treatment in the United States. Clearly, there is considerable value for patients, physicians, and policy-makers in understanding the cost, quality of life, and length

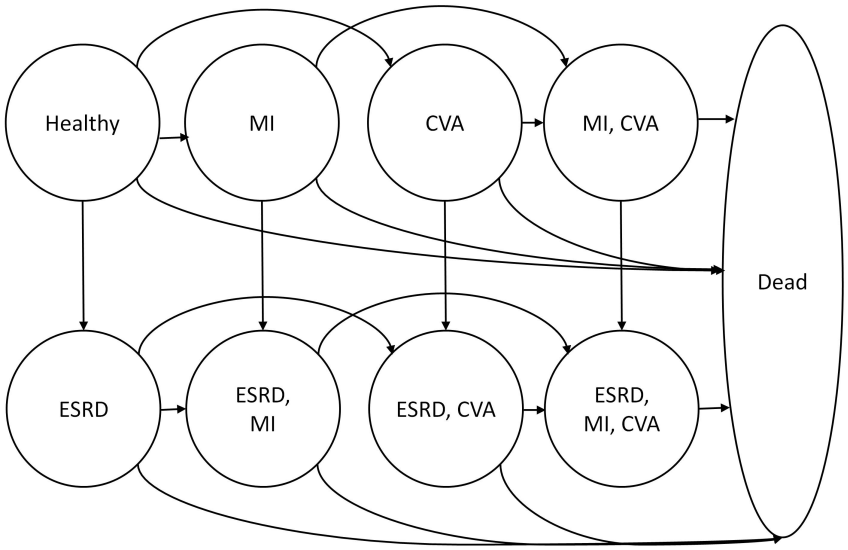

Figure 3.1: Example Model Disease States and Transitions.

of life tradeoffs inherent in medical decisions. The cost outcome is the cumulative amount of dollars spent on the subject's health care due to the intervention, usually discounted to present-day dollars at start of intervention. Quality and length of life outcomes are combined to get a single measure called quality-adjusted-life-years (QALYs). Quality is a weight between 0 and 1 that is assigned to the subject depending on the health state (such as normal, renal disease, cardiovascular diseases, etc.). The weight is multiplied by the life-time in that state and accumulated to get a QALY value, which is also usually discounted to get a present value.

Data on cost and effectiveness of the treatment group and a baseline comparison group may be obtained from through clinical trials, observational studies, or simulation. In this paper, we focus on data from discrete event simulation models of multi-state disease processes. We will use the context of the following application to illustrate the methods and the usefulness of the analysis in the rest of the paper.

Figure (3.1) represents a simplified version of the model developed in Rosen et al. (2005) to assess the benefits of angiotensin-converting-enzyme-inhibitors (ACE- in-

hibitors) for a high-risk group of patients. Rosen et al. (2005) studied the problem from Medicare's perspective, with the intervention (treatment 1) being coverage for beneficiaries with diabetes, and the standard (treatment 0) was standard care, requiring an out-of-pocket co-pay. Figure (3.1) is the model structure for examining combined effects of the intervention on cardiovascular disease (CVD) and renal disease. The number of states has been reduced for illustrative purposes. There are two states of renal disease (mild or no renal disease, and end-stage renal disease (ESRD)) and three CVD states (myocardial infarction or heart attack (MI), cerebrovascular disease or stroke (CVA), or both MI and CVA). Finally, there is an absorbing state representing death. There are four possible causes of death in this set up: MI, CVA, ESRD, and other (independent) causes. Figure (3.1) shows the state space and possible transitions for this hypothetical example. We will use different sets of illustrative data to demonstrate the benefits of the various data analysis strategies.

Discrete event simulation (DES) simulates the disease process for a subject from the population of interest, starting with an initial state and progressing through the different states over time until death. The costs and QALYs are computed as the subject evolves through the multi-state process until death. This is done for both the baseline care as well as the proposed new intervention. The simulation will be repeated for a large number of subjects with characteristics chosen to represent the population of interest and with covariates that can vary over time, such as those representing the health status of the subjects. Throughout this paper, we assume that the simulation size is large enough to treat the estimated processes and distributions as continuous.

Most of the commonly used metrics for CEA summarize this extensive amount of simulation data into simple mean-based measures. Let $\mu_j(C)$ be the average

(discounted) cost and $\mu_j(E)$ be the average (discounted) effectiveness (QALYs) for treatment j, $j = 0, 1$. Here treatment 0 is the baseline and treatment 1 is the new treatment. (There can be more than one new treatment being compared, but we restrict to just one for simplicity). The cost and effectiveness are usually computed over the entire remaining life span of the subjects or some fixed long horizon such as 30 years. Then, the metrics used for comparison are based on the differences in means: $\mu_1(C) - \mu_0(C)$ and $\mu_1(E) - \mu_0(E)$. The most common one-dimensional measure is the incremental cost-effectiveness ratio (ICER)

$$\mu_{ICER} = \frac{\mu_1(C) - \mu_0(C)}{\mu_1(E) - \mu_0(E)}. \tag{3.1}$$

Another commonly used metric is the Net Monetary Benefit (NMB)

$$\mu_{NMB} = \lambda \left[\mu_1(E) - \mu_0(E)\right] - \left[\mu_1(C) - \mu_0(C)\right], \tag{3.2}$$

where λ measures society's willingness to pay for one QALY. These and other commonly used metrics for CEA are traditionally based on expected values or averages. One justification for using average cost is that, from a decision-making perspective, it is the total cost of the intervention that matters. But the analogous justification for using average effectiveness is less compelling.

Such simple metrics are popular because they are simple and easy to understand by high-level policy decision makers. However, like all such summaries, they are overly simplistic and do not provide sufficient insights into the complex data, which is time-varying and heterogeneous, and into the trade-offs between cost and benefits. In fact, there has been debate in the literature to the extent to which cost-effectiveness ratios are appropriate for informing medical decisions (Birch and Gafni, 1992; Garber and Phelps, 1997; Sendi et al., 2003; Weinstein et al., 2001).

Our goal in this paper is to demonstrate the usefulness of using graphical methods

to analyze the simulation data more extensively and develop key insights into the performance of the treatments and their comparisons. The salient aspects of the simulation data can be summarized easily in simple cases such as Markov models. In more realistic situations, such as non-homogeneous processes, the data can have complex structure even though they are just realizations of simulated processes with known inputs.

The distribution of costs and benefits associated with an intervention is a result of randomness and differences in the underlying population, both in terms of clinical indicators and personal preferences (Stevens and Normand, 2004; Sculpher and Gafni, 2001). McMahon et al. (2005) commented on the differences in outcomes and CE ratios in heterogeneous patients, and how clinical trials often show the most benefits from interventions are concentrated in the highest-risk patients (Vijan et al., 1997; Kent et al., 2002). There is a lot of information to be gleaned from examining the entire distribution, examining it for heterogeneity, multi-modes, tail behavior, etc. These are all important sources of information for the research investigator about the behavior of the treatment and its comparison to the baseline. Before a final policy decision is made, it is incumbent upon the researcher to systematically scrutinize the underlying data, examine different features both visually and analytically, develop insights about the properties of the intervention over time and across groups of patients, and use all the information before making overall recommendations and decisions. Of course, experienced clinicians and researchers will have considerable insights even before the simulation study is conducted, but even they will get new information from the analysis. Of course, there will be others, including the decision makers, for whom the analysis will be very useful in understanding the reasons for and the trade-offs involved in the ultimate decision.

The paper is organized as follows. Section 3.2 introduces and discusses the stochastic disease process. Section 3.2.2 discusses the stochastic reward processes derived from the disease process. Section 3.3 applies the stochastic process framework and applies graphical methods for exploratory data analysis to hypothetical CEA simulation results.

3.2 Disease Progression Model

In this section, we develop some terminology and notation to characterize the underlying stochastic processes that are being simulated.

3.2.1 Disease Process

We will use a multi-state model (as in Figure (3.1)) to characterize the disease process: how the subject evolves through the various disease states until death. Let $\{Y_j(t),\ t \in \mathcal{T}\}$ denote the state of the subject at time t for $t \in \mathcal{T}$ for treatments $j = 0, 1$. Time t can be discrete in which case $\mathcal{T}$ is the set of nonnegative integers $\{0, 1, 2, ...K+1\}$ or continuous in which case $\mathcal{T}$ is the non-negative half-line $\{t \geq 0\}$. The process $\{Y_j(t)\}$ takes values in the finite state space $E = \{1, 2, \ldots, (K+1)\}$, which are the various states of health status. We take $(K+1)$ to be an absorbing state or end point of the process, denoted as 'failure.' In the context of CEA problems, this state will typically be death. Subjects can start the simulation in different initial states and they will have been in the state for different periods of time, and may have different ages. There are also important covariates that will distinguish different subjects. We will ignore these important points for the time being come back to them later in the discussion.

There has to be some specification of the underlying probabilistic mechanism that generates the transitions and the sojourn times. For our purposes here, we assume

that they are specified in setting up the simulation study and reasonable estimates of the underlying parameters are available from the literature of field studies. We will also not focus on sensitivity to the estimated values, as this problem has received substantial attention in the literature (Hunink et al., 1998; Halpern et al., 2000).

There are analytical approaches to studying the behavior of multi-state processes in simple cases: Markov processes and semi-Markov processes and more general time-homogeneous processes (see, for example, Howard (1971) and Janssen and Manca (2006)). We will assume that the processes being studied here are more complex and so a discrete event simulation study is used.

There are several random variables of interest associated with the disease process. Perhaps the most important one is the time-to-failure T_j, the time t when the process $Y_j(t)$ reaches the absorbing state $K+1$. Throughout, we denote the cumulative distribution function (CDF) of T_j as $F_j(t;T), t \in \mathcal{T}$ and the corresponding quantile functions as $Q_j(u;T) = F_j^{-1}(u), u \in (0,1)$. Further, the mean or expected time-to-failure will be denoted as $\mu_j(T)$ and the hazard rate as $h_j(t;T)$ for $j = 0, 1$. In general, the F_j's are a mixture distribution and their expressions cannot be obtained analytically, even for Markov and semi-Markov processes (see Aalen (1995), DeFauw et al. (2011) and Yang and Nair (2011)). But the simulation data will allow us to study their properties and make comparisons, based on means, quantiles, tail behavior, hazard rate, etc.

3.2.2 Cost, Effectiveness and Other Derived Processes

As subjects evolve through the disease process, they accumulate costs and QALYs. The particular structure of these, especially costs, will depend on the application. For example, there can be an immediate cost associated with a transition from state i to state j (for example, surgery for an MI) and there can be costs per unit time

spent in state j (such as cost of long term medical care). These issues are discussed in general terms under reward processes in Howard (1971) and Janssen and Manca (2006). See also DeFauw et al. (2011).

We will use the notation $\{C_j(t),\ t \in \mathcal{T}\}$ and $\{E_j(t),\ t \in \mathcal{T}\}$ to denote the cumulative cost and reward processes: so, $C_j(t)$ is the cumulative cost and $E_j(t)$ is the cumulative effectiveness for treatment j up to time t, for $j = 0, 1$. These processes may or may not have been discounted, depending on the application. Further, we let $F_j(x; t; C) = P(C_j(t) \leq x)$ denote the CDF of the cost process at time t and its quantile function as $Q_j(u; t; C) = F_j^{-1}(x; t; C)$ with similar definitions for effectiveness: $F_j(u; t; E)$ and $Q_j(u; t; E)$. The means of the corresponding distributions will be denoted as $\mu_j(t; C)$ and $\mu_j(t; E)$. The end of the study will correspond to $t = \infty$ (when all the subjects have died) or more likely, a large value of t. Analytical expressions for the cost and effectiveness processes and their means were developed in DeFauw et al. (2011) for the case of progressive, time-homogeneous semi-Markov processes. Reference Janssen and Manca (2006) and Janssen and Manca (2007) for a general discussion of reward processes. Again, the simulation data will allow us to study the behavior of these quantities and interpret their behavior from CEA perspective, as to be discussed.

Sometimes, the notions of incremental cost and effectiveness processes are useful. For example, for chronic diseases where patients live with for many years, studying incremental annual costs and seeing how they vary over time makes sense. In the case of more short-term, acute, or rapidly advancing diseases, weekly or daily representation of incremental processes may be more useful. We will define incremental cost as $c_j(t) = C_j(t) - C_j(t-\delta)$ and incremental effectiveness $e_j(t) = E_j(t) - E_j(t-\delta)$ for some suitable time period δ. For discrete processes, δ will typically be 1.

Another derived process of interest is the cumulative monetary benefit (MB) process which is defined for treatment j at time t as $B_j(t) = \lambda E_j(t) + C_j(t)$ for $j = 0, 1$. We let $F_j(u; t; B) = P(B_j(t) \leq x)$ be the CDF of the MB process at time t and its quantile function as $Q_j(u; t; B) = F_j^{-1}(x; t; B)$ The means of the corresponding distributions will be denoted as $\mu_j(t; B)$ and $\mu_j(t; E)$. Recall that the usual NMB is given by $\mu_1(\infty, B) - \mu_0(\infty, B)$. As with the other cases, we define incremental monetary benefit as $b_j(t) = B_j(t) - B_j(t - \delta)$ for some suitable time period δ.

3.3 Richer Analysis: Illustrative Examples

This section will rely primarily on graphical methods to illustrate the potential in exploring the simulation data in more detail. There is an extensive literature on graphical methods, and some good references include Chambers (1983) and Cleveland (1993). See also Doksum and Sievers (1976) for a discussion of treatment-effect functions. Throughout, we are assuming that the number of simulations is very large so that, for all practical purposes, the distributions are assumed to be known.

3.3.1 Exploratory Analysis for Interesting Features

There are many graphical techniques for visually examining the data when the sample sizes are small to moderate. These include boxplots, stem-and-leaf-diagrams and histograms. Sample sizes in the kinds of simulation studies considered here will be quite large, and we use a density plot (smoothed histogram) to examine features of the data.

Consider first cost and effectiveness outcomes at the end of the study. The left panel in Figure (3.2) shows illustrative densities for treatment and standard: effectiveness data in the right panel and cost data in the left panel. Both cost distributions are skewed with heavy right tails. This is typical of cost outcomes (just as is the case

with income distributions), indicating that a small proportion of the population contributes most to the total cost. The tails will have an undue influence on the mean, so mean-based comparisons can be unduly affected. More importantly, relying on just simple metrics for comparison can result in the loss of very useful information. The cost distribution for the standard treatment has a heavier right tail than the new treatment, and it would be of interest to explore the reasons for this difference. Suppose we took a deeper look at the percentage of costs incurred by patients in the highest 20% of each cost distribution, leading to Figure (3.3). We see that a disproportionate amount of the costs in the right tail for the standard treatment are incurred by patients in the ESRD state. The treatment is very effective in reducing this part of the costs.

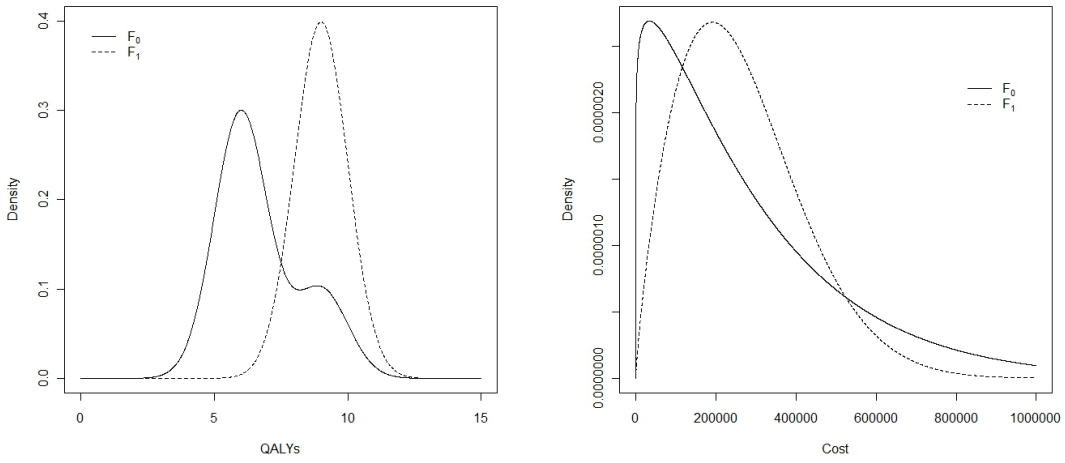

Figure 3.2: *Left:* Effectiveness Distributions. *Right*: Cost Distributions.

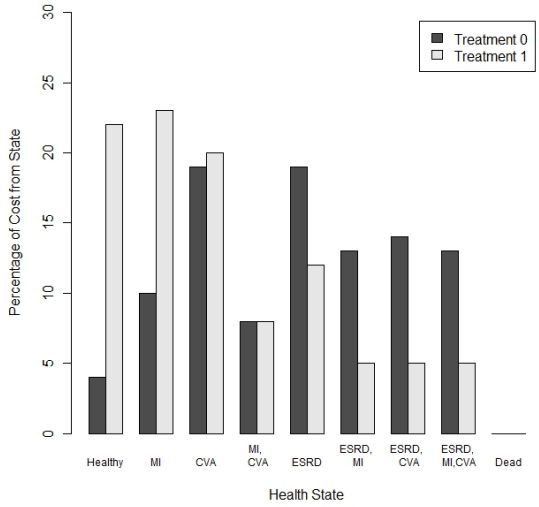

Figure 3.3: Percentage of Cost from Each State.

Consider now the left panel in Figure (3.2) which shows effectiveness distributions. We see that the density under treatment 0 is bimodal, with a smaller mode in the right tail. This is not the case for treatment 1. Multi-modality is often an indication of a mixture phenomenon, suggesting differential effects on the population. Suppose we drill down into the distribution to examine the reasons for the bimodality and find that the reason is due to differences in income levels of the patients (socioeconomic status, or SES). Specifically, a significant portion of the high-income patients were already taking ACE-inhibitors, so the policy of covering them under Medicare did not affect their health status. This is confirmed in Figure (3.4) which shows a substantial difference between the two populations under treatment 0, and shows that the patients in the high SES group have a distribution similar to the distribution of the entire population under treatment 1. This group formed only about 25% of the overall population, so the difference between the groups showed up only slightly in Figure (3.2). Of course, in a simulation study, this information is part of the input parameters and so is already known. However, the effects of different input

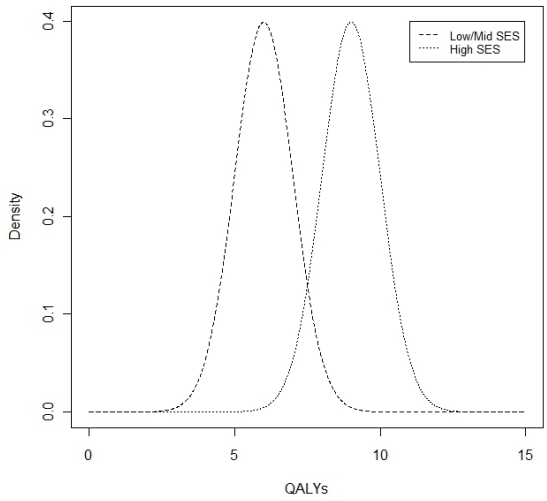

Figure 3.4: *Left:* Effectiveness Distribution of Treatment 0, Stratified by SES.

parameters may not be easily discerned up front. Moreover, the effects could be offset by other factors, so there is considerable value in examining the data to determine the final outcomes and dig down to understand the reasons. In this example, there are several possible consequences to the finding: a) one could consider the option of not providing coverage for high-income patients if this is feasible; and b) the effectiveness of the treatment in the remaining population is much bigger than a simple mean-based comparison would suggest.

3.3.2 Comparing the Difference in Distributions Using Q-Q plots

The densities are useful in highlighting certain aspects of the distributions (center, spread, tails, unusual features, etc.) but are not as effective in comparing the differences between distributions. In this section, we consider comparisons of CDFs. The case where the effect of the treatment is to simply shift the response by a constant value Δ is shown in the left panel of Figure (3.5). This corresponds to shift distributions where $F_1(x) = F_0(x - \Delta)$, or equivalently $Q_1(u) - Q_0(u) = \Delta$, i.e., the difference in every $u - th$ quantile is constant. Since $\int_0^1 Q_j(u)\ du = \mu_j$, the mean,

it also implies that the differences in means equal Δ. In this case, Δ measures the treatment effect in a reasonable sense.

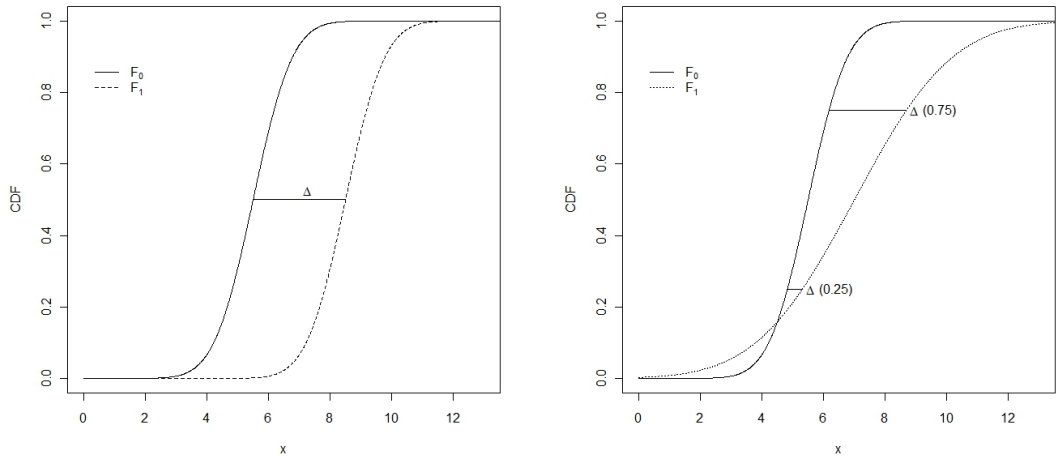

Figure 3.5: *Left:* Shift distribution treatment effect. *Right*: Uneven treatment effect.

Suppose F_1 is entirely below F_0, i.e., $F_1(x) \leq F_0(x)$ for all x (with strict inequality for some x), as is the case in the left panel of Figure (3.5). Then treatment 1 is said to stochastically dominate the other in terms of first order (first-order stochastic dominance or FSD). Note that $F_1(x) \leq F_0(x)$ means that larger values are likely under F_1 compared to F_0. Perhaps an easier way of thinking about this is in terms of the quantile functions: $Q_1(u) \geq Q_0(u)$. If the outcome is effectiveness, then treatment 1 is preferable under FSD, but the opposite is true for cost.

But the treatment effect is likely to be more complex in practice, and we need ways to interpret them. In general, the distributions will cross (as in the right panel of Figure (3.5)), indicating that treatment 1 is better than treatment 0 in one region but not the other. It is difficult to assess the magnitude of the differences between CDFs as both are curves. For this reason, statisticians have proposed plotting the quantiles against each other: $Q_0(u)$ on the horizontal axis and $Q_1(u)$ on the vertical axis (see Figure (3.6)). This is called the quantile-quantile or Q-Q plot in the statistics

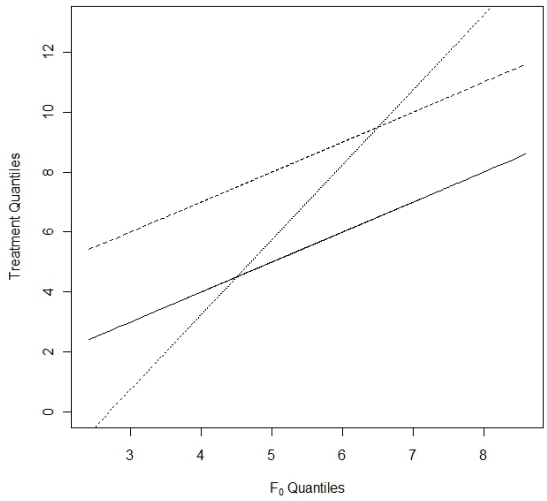

Figure 3.6: Quantile-Quantile Plot.

literature. The effect of this plot is to transform the y-axis from a probability scale to a quantile scale.

There are a number of visual advantages with the Q-Q plot. If the two distributions are identical, the plot will fall on the identity line. If the location model $F_1(x) = F_0(x-\Delta)$ holds, (effect of the treatment is to just change the outcome by Δ at all values), the plot will be linear with slope 1 and intercept Δ. It is much easier to determine such an effect from this graph than from the plot of the two CDFs in Figure (3.5). More generally, if a location-scale model holds, i.e., $F_1(x) = F_0([x-a]/b)$ for some constants a and $b > 0$, the plot will still be linear, but with slope b and intercept a. It is unlikely that a location-scale model will hold for outcomes such as failure times, costs and effectiveness, which are constrained to be non-negative. However, it is common in statistical analysis to take the logarithm of the data and then do a Q-Q plot to compare the distributions. For example, distributions such as log-normal and Weibull have a location-scale property in the log-scale.

Let $\Delta(u) = Q_1(u) - Q_0(u)$ be the difference of the Q-Q plot $(Q_0(u), Q_1(u))$ from

the identity function. We will refer to this as the treatment-effect function. (Doksum and Sievers (1976) referred to this as a response function but used a slightly different form: $\theta(x) = Q_1(F_0(x)) - x$, which equals $\Delta(F_0(x))$.) This treatment-effect function captures the difference in the treatments in different regions (quantiles) of the distributions. For example, the difference in medians is $\Delta(.5)$ and $\int_0^1 \Delta(u)\ du = \int_0^1 [Q_1(u) - Q_0(u)]\ du = \mu_1 - \mu_0$, the difference in means.

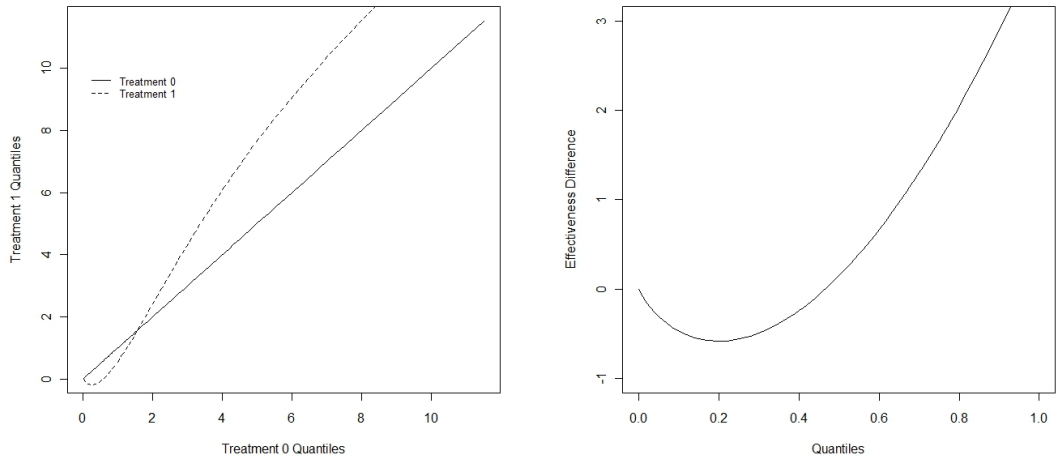

Figure 3.7: *Left:* Q-Q Plot of Effectiveness Distributions. *Right*: The treatment-effect, $\Delta(u)$, for Effectiveness.

Figure (3.7) shows an illustrative example for a (hypothetical) intervention for cardiovascular and heart disease, different from ACE-inhibitor example already used. The left panel is a Q-Q plot of effectiveness distributions and the right panel shows the treatment-effect function $\Delta(u; E)$. An examination of this function shows that the treatment is slightly harmful, compared to the standard, in the lower tail but is quite effective in the other regions. The natural next step is to dig into possible reasons for this behavior. Figure (3.8) shows a bar plot of the patients in the lower 30% of the treatment 1 effectiveness distribution and the overall population, classified by the two relevant racial groups for this example: Caucasians and African-Americans.

We see that there is a disproportionately higher portion of African-Americans in the lower tail of treatment 1, suggesting that the treatment is harmful to this group. Again, such a finding will have important consequences for the decision makers and would not have been discovered in a simple mean-based comparison.

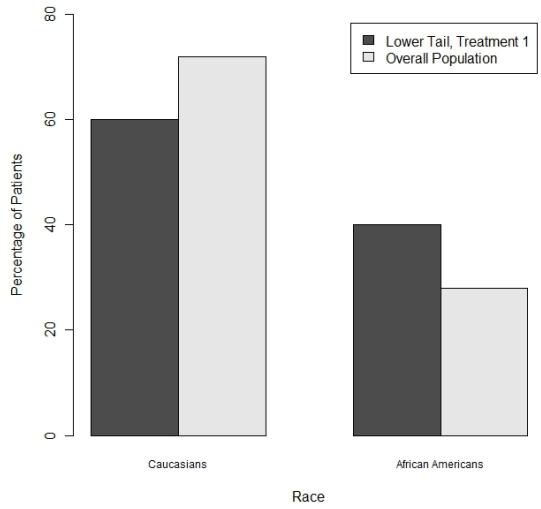

Figure 3.8: Weight Classification of Patients in Lower Quantiles of Effectiveness Distributions.

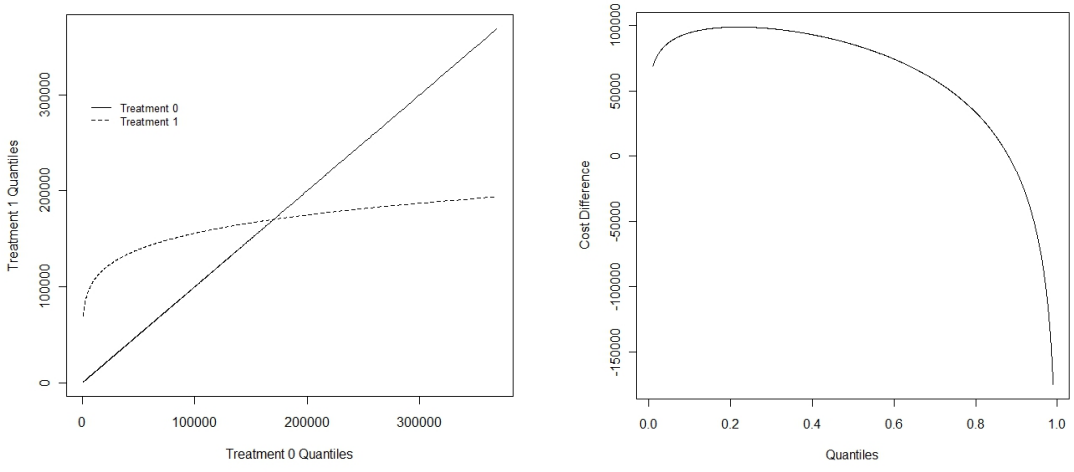

Figure 3.9: *Left:* Q-Q Plot of Cost Distributions. *Right*: $\Delta(u)$ of the Cost Distributions.

Consider a different example. Figure (3.9) compares the cost distributions of two hypothetical treatments, and we see that $\Delta(u; C) > 0$ at the low quantiles with

the situation reversed at the higher quantiles. Suppose further analysis to identify the reasons for the differences leads to Figure 3.10: the left panel shows Q-Q plot for patients who were healthy at the start of the study and right panel shows Q-Q plot for patients whose initial state was ESRD. The plot in the left panel looks very similar to that of the overall population in Figure (3.9). On the other hand, $\Delta(u; C)$ is always negative in the right panel, with the function increasing towards the right tail. This is another example of effect heterogeneity, this time caused by differences in initial states. Again, this type of information may be obvious to an experienced investigator, who may know that the analysis should be separated by initial states. But there are many possible ways to stratify the data, and an analysis such as the one proposed in the paper suggests ways to systematically explore the data to identify differences and then dig deeper to determine possible reasons.

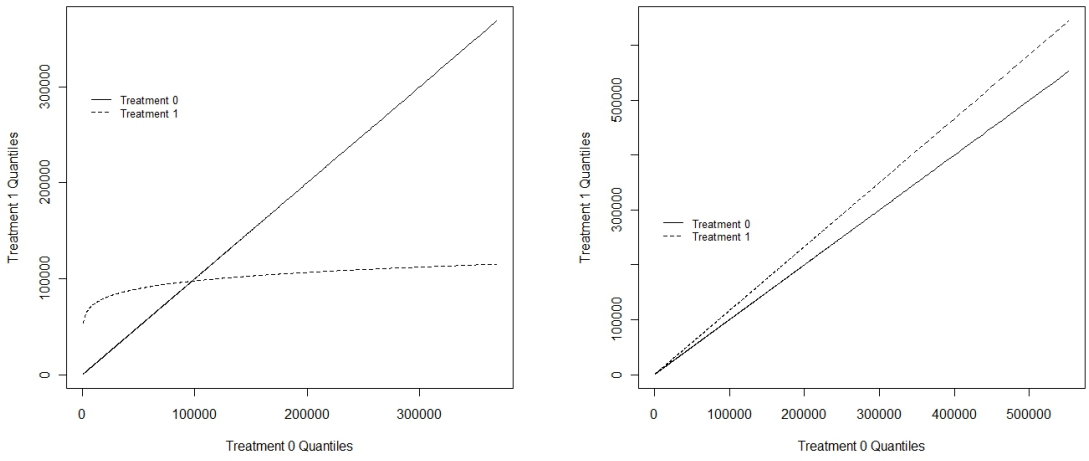

Figure 3.10: *Left:* Q-Q Plot of Cost Distributions for Patients Starting in Healthy State. *Right*: Q-Q Plot of Cost Distributions for Patients Starting with ESRD.

Continuing with the same example, one might want to examine the effect of the stratification for the corresponding effectiveness distributions. Figure 3.11 shows an illustrative situation. The right panel suggests that the difference in effectiveness

distributions is negligible for subjects starting in the ESRD initial state. The left panel indicates that the difference is much more substantial for subjects starting in the healthy state. We can now close the loop on this story: the treatment is not effective for patients who already had ESRD and the added cost is due to cost of the treatment. One obvious conclusion is to not use this treatment for patients with ESRD. Again, we reiterate that such a conclusion may already be known to (or at least suspected by) clinicians before the study, but we are using a simple example such as this to illustrate that similar benefits may arise in more subtle situations.

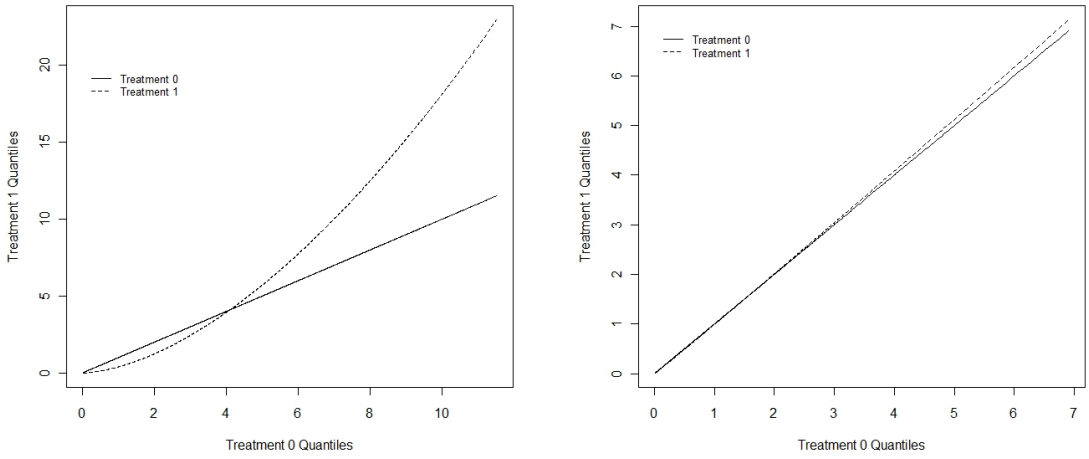

Figure 3.11: *Left:* Q-Q Plot of Effectiveness Distributions for Patients Starting in Healthy State. *Right*: Q-Q Plot of Effectiveness Distributions for Patients Starting with ESRD.

3.3.3 Comparisons Over Time

We have focused thus far on analyzing data at the end of the study – either when all subjects in the simulation have died or after a fixed period. This is also the focus of CEA. Nevertheless, it is of interest to see how the cost and effectiveness distributions evolve over time, for budgetary or ethical reasons, or to develop further insights into the treatment effect.

We revisit the examples in Figure (3.2) where the effectiveness distribution at the

end of the study was bimodal. Figure (3.12) shows this distribution at three different time points: $t = 5$, $t = 20$ and at the end of the study. it shows that the distribution was unimodal at an early stage but started to develop the bimodal behavior by $t = 20$ and it became more pronounced by the end of the study. Depending on the particular intervention, this could again have policy implications or affect the delivery of the intervention.

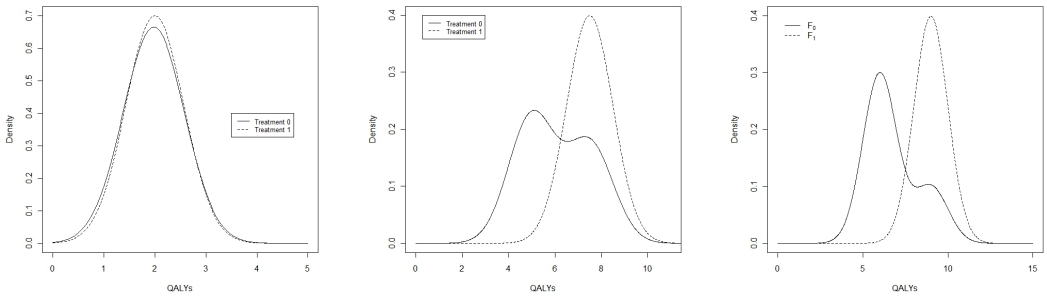

Figure 3.12: *Left:* Effectiveness Distribution at t=5. *Center:* Effectiveness Distribution at t=20. *Right:* Effectiveness Distribution at t=30.

It can be tedious to compare the density functions or Q-Q plots at many different times; further, it may be difficult to spot small differences. An alternative approach to analyzing the evolution of reward over time is to plot selected quantiles of the reward (cost or effectiveness) distributions against time. Figure (3.13) shows the evolution of the quantiles over time for a pair of hypothetical treatments 0 and 1. It shows clearly that the variability of the rewards is low in the early part of the study for both treatments. As time progresses, the variability in the reward gets higher for treatment 1 than treatment 0, and the pattern is consistent. Figure (3.14) shows how the treatment-effect function, $\Delta(u;t) = Q_1(u;t) - Q_0(u;t)$ varies with time. In contrast, Figure (3.15) shows a different shape for $\Delta(u;t)$, where the treatment-effect is zero for large values of u and positive for smaller values.

We make an important, cautionary note in trying to interpret the treatment-

effect function. The characteristics of the subjects who are in the lower (or upper) quantiles of Treatment 0 may not be the same as those who are in the corresponding quantiles of Treatment 1, and these could also change over time. This complicates the interpretation, and one has to dig deeper into the data to understand possible reasons in terms of patient characteristics.

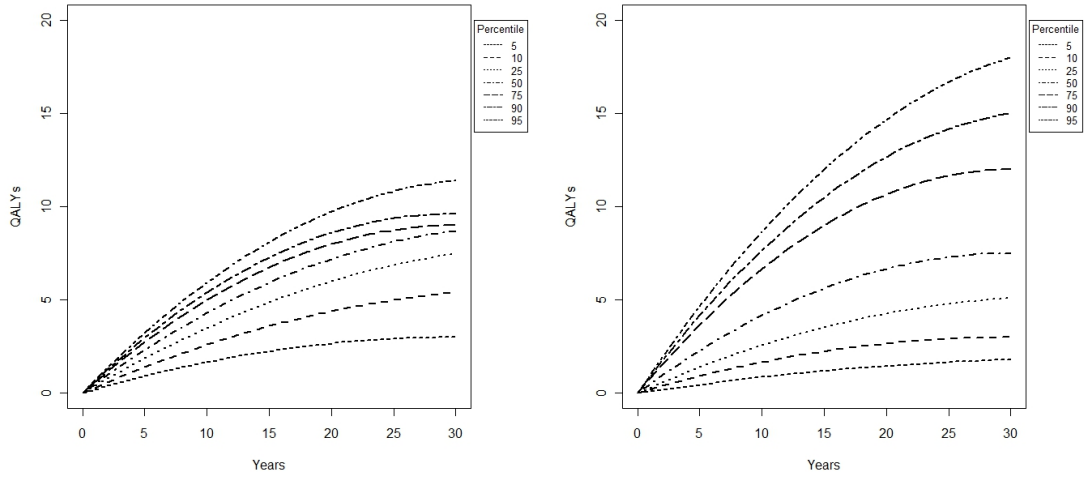

Figure 3.13: *Left:* Quantile functions over time for Treatment 0; Quantiles. *Right*: Quantile functions over time for Treatment 1.

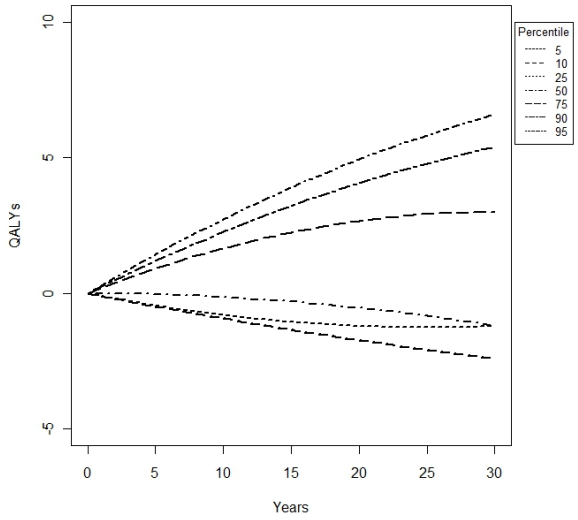

Figure 3.14: Treatment Effect Function Over Time,$\Delta(u;t)Q_1(p;E;t) - Q_0(p;E;t)$.

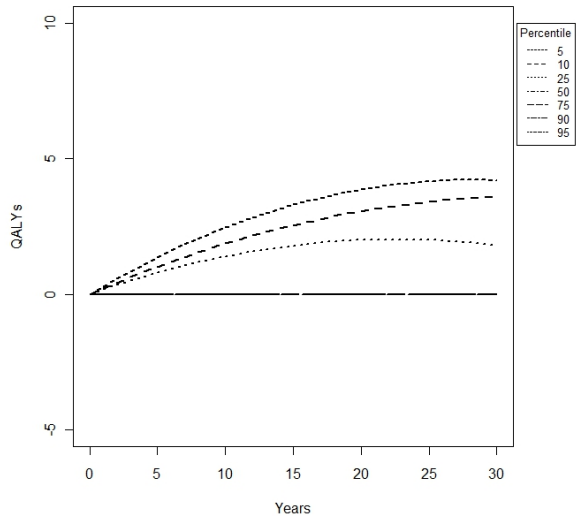

Figure 3.15: Treatment Effect Function Over Time, $Q_1(u;E;t) - Q_0(u;E;t)$.

3.3.4 Generalized ICER and NMB Metrics

The notion of the treatment-effect function suggests an obvious way to extend the definitions of ICER and NMB metrics and conduct a more detailed CEA. Define the ICER function

$$ICER(u;t) = \frac{\Delta(u;t;C)}{\Delta(u;t;E)}, \tag{3.3}$$

the ratio of the treatment-effectiveness functions for cost and effectiveness. This metric allows one to examine how the ICER metric varies with the quantiles (for fixed t) and varies with time (for fixed u). Similarly, we can define an NMB function as

$$NMB(u;t) = \lambda\Delta(u;t;E) - \Delta(u;t;C). \tag{3.4}$$

Figure 3.16 shows an illustrative example with the ICER function on the left and NMB function on the right. Notice that there is huge variability in the ICER function early on but it stabilizes over time. Also, the quantiles cross, showing an inconsistent behavior over time.

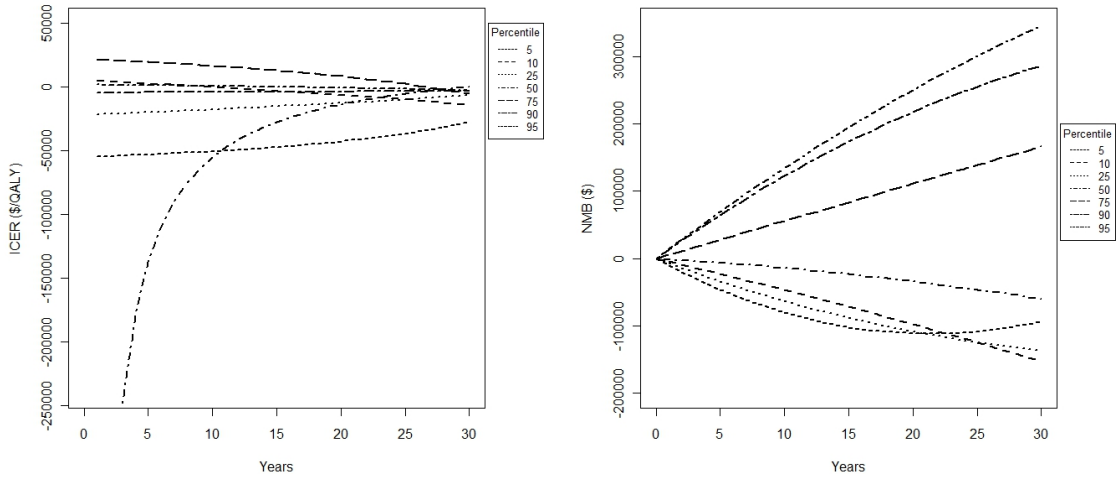

Figure 3.16: *Left:* $ICER(u;t)$. *Right*: $NMB(u;t)$.

3.3.5 Analysis of Time-to-Failure Data

While the reward processes are of primary interest in CEA, analysis of the time-to-failure data would also be of interest. We provide some illustrative examples here.

If an intervention is considered effective, it would normally be expected to increase the length of life, on average. The increased lifespan my arise in a number of ways. First, a treatment may not change disease progression (the health states a patient experiences), but slow it down (increase the time spent in each health state). If this is true, the treatment effect on the time in health states may be the same for all health states (i.e., treatment increases the time in each health state by 25%). Alternatively, it may slow disease progression unequally in each health state. A second way a treatment might act is my changing the disease progression entirely. In the case of a multi-state model, this would mean that the treatment would actually alter the probabilities of a patient experiencing or dying from a given medical event. As an example, a treatment may lower the lifetime probability of a patient experiencing an

MI.

In a multi-state disease process, understanding the time spent in each state can also provide insight into the drivers of the difference between the two treatments. Figure (3.17) shows this comparison for the Medicare coverage of ACE-inhibitors intervention (treatment 1), versus the baseline (treatment 0). The side-by-side barplot shows that patients receiving treatment 1 live longer, and spend much more time in the healthy state. Additionally, patients receiving treatment 0 spend more time in all states containing ESRD. This tells the decision-maker that on average, treatment 1 helps patients avoid time in ESRD states and remain healthier, longer, than those receiving treatment 0.

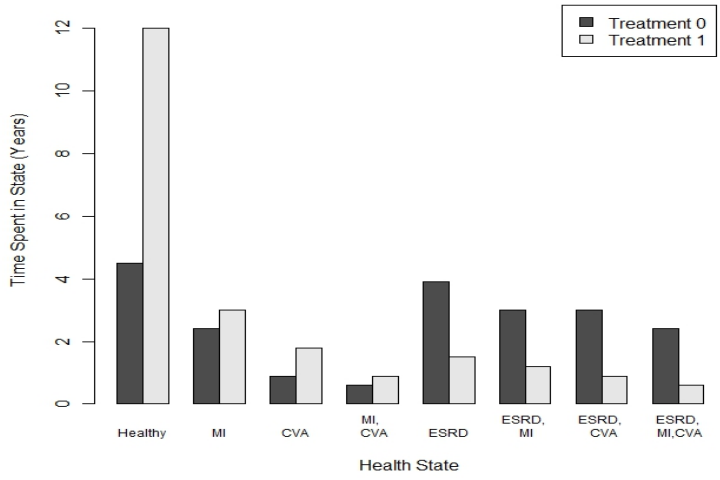

Figure 3.17: Time in Each State.

A difference in disease progression may also indicate a difference in the modes of death. Figure (3.18) shows the percentage of deaths due to each of the 4 possible causes: MI, CVA, ESRD, and competing risk causes (causes not included in the model). If the figure had bars of equal height for treatment 1 and treatment 0, that would indicate that the treatment does not effect the mode of patient death.

However, this is not the case in this example. Figure (3.18) shows that, relative to treatment 0, patients receiving treatment 1 are less likely to die of CVA and ESRD, but are more likely to die of MI or competing risk causes. This suggests that, relative to treatment 0, treatment 1 reduces CVA and ESRD. However, it does not have a fully protective effect against all CVD, since MI death is comparatively more likely under treatment 1.

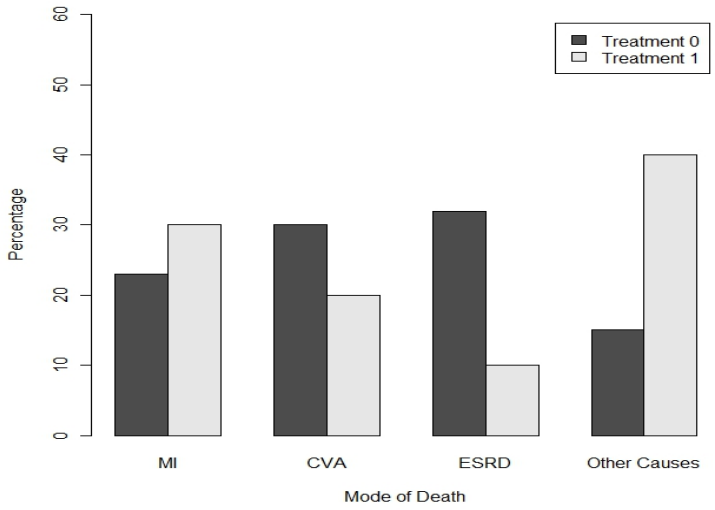

Figure 3.18: Mode of Death.

As we have already seen, stratification of patients by the health state at the start of treatment can provide useful insight into the differences in death mode. Figure (3.19) shows the proportions of subjects in each death mode, conditional upon patient health state at the start of treatment. There is a large difference that depends upon patient's initial health. In the case of patients with a history of MI, CVA, or renal disease, it is very common for the patient to die of whatever complication they have a history of, regardless of treatment. Patients beginning with no history of renal disease or CVD are relatively unlikely to die of CVD or renal disease complications, and more likely to die of competing risk causes.

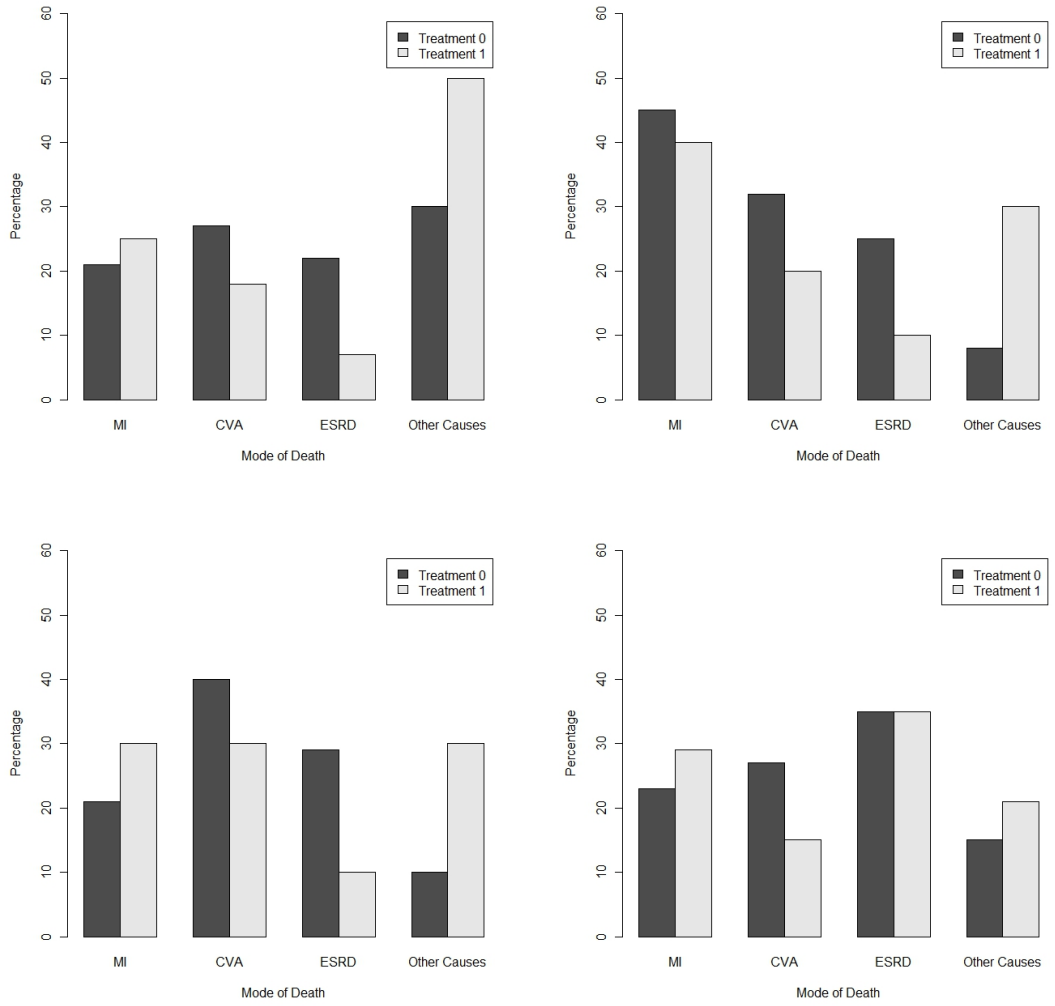

Figure 3.19: *Top Left:* Death Mode for those Starting Healthy. *Top Right*: Death Mode for those Starting with an MI. *Bottom Left:* Death Mode for those Starting with a CVA. *Bottom Right*: Death Mode for those Starting with Renal Disease.

3.4 Conclusion

The paper has demonstrated the usefulness of exploring the data from simulation studies more fully, instead of just relying on simple CEA. We have used hypothetical, illustrative examples in our demonstration, which are by necessity too simple. Nevertheless, we hope that we have been successful in showing the benefits of such an approach.

We have not discussed methods for examining the effects of covariates, such as other indicators of health-status, in this paper. The effect of time-varying covariates, in particular, can be difficult to predict. These can be handled through regression or other types of analyses, suitably modified for the type of data.

CHAPTER IV

Quantifying Treatment Effect in Cost-Effectiveness Analysis

4.1 Introduction

There is growing emphasis in the healthcare area on containing costs while also improving treatment effectiveness (health gained from treatment). The main goal of cost-effectiveness analysis (CEA), which uses metrics based on costs and effectiveness to assess and compare interventions, is to identify effective treatments while also limiting healthcare expenditures. One or more new interventions are compared against the current or baseline intervention. The term intervention is used to characterize a variety of options: screening strategies, pharmaceutical treatments, diagnostic tools, surgical treatments, etc., and they may be implemented via policy- or payer-level requirements or incentives.

While there are several ways to measure effectiveness of a treatment, quality-adjusted-life-years (QALYs) are the most common. Both costs and QALYs are usually discounted at some rate to bring the values at different times to a common reference value. Discounting QALYs is less intuitive than discounting cost, but CEA relies on the assumption that QALYs have a monetary value, and also that there is a constant trade-off between costs and QALYs.

There is a huge literature on the development and use of various metrics for CEA.

Most of them are based on the means of the (random) cost and effectiveness outcomes, where the expected values are taken with respect to some population of interest. We review the commonly used metrics in the next section. There has been some effort to go beyond mean-based analysis in the CEA literature, although this is relatively limited. We discuss the notions of first and second-order stochastic dominance (FSD and SSD) of the distributions and discuss their usefulness in CEA. SD concepts are well known in economics and finance but their use in healthcare decision making has not received much attention (see Laska et al. (1999); Leshno and Levy (2004) for some exceptions). Notions of almost stochastic dominance (ASD) from the literature are also discussed. Even when a new treatment does not stochastically dominate the current one (as will be the case in many situations), it is of interest to examine regions of the outcomes space where the new treatment is better than the standard. Such insights will allow the decision-maker to make more informed decisions.

The concepts of FSD and SSD are popular in economics and finance in part due to their relationships to general classes of utility functions. We also consider comparisons in terms of more specialized, parametric utility functions.

A second, and bigger, focus of the paper is CEA in the presence of statistical uncertainty. In practice, the metrics used for comparison have to be estimated from field data, inducing estimation uncertainty which has to be taken into account in the comparisons. We review selected, important results in the CEA literature for mean-based comparisons: using both large-sample normal approximations and resampling methods. While these methods are applications of well known results in the statistical literature, there appears to be some confusion that we review and clarify. We also propose the use of rank-based methods and discuss analysis of censored and paired comparison data. The new contributions to the CEA literature include SD

comparisons in the presence of statistical uncertainty.

The rest of the paper is organized as follows. In Section 4.2, we describe the common mean-based metrics for CEA. Section 4.3 reviews methods for comparisons based on these metrics in the presence of statistical uncertainty and proposes the use of other methods from the statistics literature to CEA; these include use of rank-based methods and analysis of censored data. In Section 4.4 we expand the comparisons to include utility functions and examining the entire distributions. The role of utility functions is discussed in Section 4.4.1 and the application of first- and second-order stochastic dominance in CEA is the subject of Section 4.4.2. This is followed by the consideration of stochastic dominance comparisons in the presence of uncertainty. Throughout the paper, we demonstrate concepts using illustrative examples.

4.2 Mean-Based Metrics for CEA Analysis

Suppose we are comparing a new intervention (treatment 1) against a current or baseline intervention (treatment 0). Let C_j and E_j, $j = 0, 1$ be the (random) cost and effectiveness outcomes over some appropriate lifetime of a subject. Further, there is some underlying population of subjects that is of interest (such as all patients eligible for Medicare, patients who are at risk of a disease, etc.). These specifics will depend on the particular application, and for the purposes here, we assume that they have already been well formulated.

Let $F_j(t; C)$ and $F_j(t; E)$ be the cumulative distribution functions (CDFs) of C_j and E_j respectively, for $j = 0, 1$. Denote by $\mu_j(C)$ and $\mu_j(E)$ their expected values, i.e., mean cost and mean effectiveness measures. Define $\Delta(C) = \mu_1(C) - \mu_0(C)$ and $\Delta(E) = \mu_1(E) - \mu_0(E)$ be the differences in costs and effectiveness respectively. Most

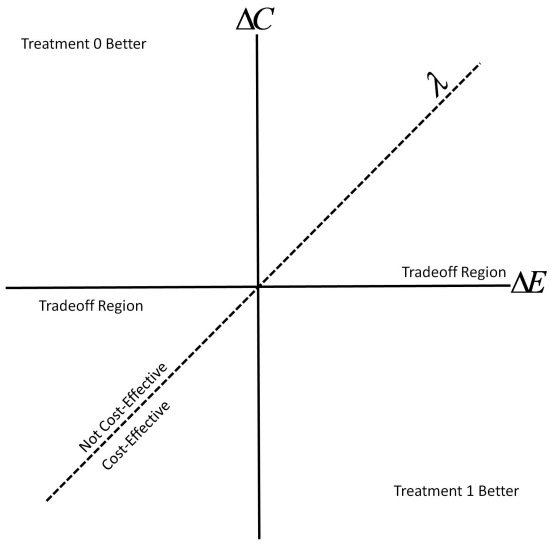

Figure 4.1: Cost-Effectiveness Plane.

of the metrics in the CEA literature are based on these means – more specifically on the differences $\Delta(C)$ and $\Delta(E)$.

Direct two-dimensional analyses have been proposed in the literature (O'Brien et al., 1994). Here, ΔC and ΔE are plotted on the x- and y-axes separately (see Figure (4.1)). If the values fall in the lower right quadrant, the new treatment is more effective and has lower cost than the baseline treatment (termed "dominance" in the CEA literature). If they fall in the upper left quadrant, the baseline treatment has higher effectiveness and lower cost. There is a trade-off between cost and effectiveness if the values fall in the upper right or lower left quadrants. In those quadrants, a intervention is considered cost-effective if it lies below the line through the origin with slope λ, the value of the decision-makers willingness-to-pay.

There are several metrics in the literature for combining the two-dimensional information on costs and effectiveness into a single metric. Perhaps the most common one is the incremental cost-effectiveness ratio (ICER) given by

$$\mu_{ICER} = \frac{\mu_1(C) - \mu_0(C)}{\mu_1(E) - \mu_0(E)} = \frac{\Delta(C)}{\Delta(E)}.$$

μ_{ICER} is compared against a value λ, the amount of money that society is willing to pay for a single unit of effectiveness. We can think of λ as the factor used to convert effectiveness into the same scale as cost. If μ_{ICER} is less than λ (and $\Delta(E) > 0$), then the new treatment is deemed cost-effective. If not, then the current treatment is more cost-effective. The willingness-to-pay value is represented in Figure (4.1) with a dashed line. See O'Brien et al. (1994) for more details.

An assessment of the two treatments in terms of μ_{ICER} can be visualized in the two-dimensional plot in Figure (4.1) as follows. Consider the line with slope λ in the figure. A new intervention is considered cost-effective compared to the baseline, in terms of ICER, if the point $(\Delta(E),\ \Delta(C))$ falls to the right of this line.

Any one-dimensional summary, such as the ICER, will naturally result in a loss of information. For example, a negative ICER may be caused by the numerator or the denominator being negative, and the interpretations are quite different – the new intervention results in higher costs and is less effective or lower costs and greater effectiveness (Stinnett and Mullahy, 1998). The ICER is known to have additional disadvantages. If the difference in effectiveness $\Delta(E)$ is close to zero, μ_{ICER} can be very large. These problems are exacerbated when the ICER metric is estimated from sample data. It can have huge variability if the true value $\Delta(E)$ is close to zero. Moreover, the signs of the estimates can get switched. Some authors have proposed a two-phase approach – first determining if $\Delta(E)$ is positive (treatment 1 is better in terms of effectiveness) and then seeing if the ICER metric is positive and what its value is. This is of course equivalent to a two-dimensional analysis and does not really mitigate the problems.

Stinnett and Mullahy (1998) proposed two (equivalent) alternatives called Net Health Benefit (NHB) and Net Monetary Benefit (NMB). To define these, define a

random variable called monetary benefit (MB) of treatment j as

$$\text{MB}_j = \lambda E_j - C_j,\ j = 0, 1, \tag{4.1}$$

where, as before, λ is used to convert effectiveness into its monetary equivalent. Then, NMB is defined as

(4.2)

$$\mu_{NMB}(\lambda) = E(MB_1 - MB_0) = \lambda[\mu_1(E) - \mu_0(E)] - [\mu_1(C) - \mu_0(C)] = \lambda\Delta(E) - \Delta(C).$$

If the monetary equivalent of the effectiveness is greater than the cost of the treatment, NMB will be positive and the new treatment is preferred. Otherwise, the current treatment is preferred. Similarly, NHB is defined as

$$\mu_{NHB}(\lambda) = \Delta(E) - \Delta(C)/\lambda. \tag{4.3}$$

Since NMB and NHB are equivalent, we will focus on NMB only in the rest of this paper. Note the one-to-one relationship: $\mu_{ICER} = \lambda$ if and only if $\mu_{NMB}(\lambda) = 0$. We will return to this connection later in the paper.

We note for use later in the paper that NMB is linear in all of its terms, and can be rewritten as

$$\begin{aligned} \mu_{NMB}(\lambda) &= (\lambda\mu_1(E) - \mu_1 C) - (\lambda\mu_0(E) - \mu_0 C) \\ &= \mu_1(MB) - \mu_0(MB), \end{aligned} \tag{4.4}$$

where $\mu_1(MB) = \lambda\mu_1(E) - \mu_1 C$, and as such is the mean monetary benefit of treatment 1 (i.e., the mean weighted combination of costs and effectiveness resulting from treatment 1). $\mu_0(MB)$ has an analogous definition, and is the mean monetary benefit of treatment 0. If $\mu_1(MB) > \mu_0(MB)$, then $\mu_{NMB} > 0$, and this relationship allows us to determine the significance of μ_{NMB} by comparing the means of the two samples.

Since NMB is also a one-dimensional reduction, it loses important information. For example, if it is positive, all we know it that $\lambda\Delta(E) > \Delta(C)$; similarly if it is negative $\lambda\Delta(E) < \Delta(C)$.

Another alternative, proposed by Laska et al. (1997) is the difference in cost-effectiveness ratios (DCER):

$$\mu_{DCER} = \frac{\mu_1(C)}{\mu_1(E)} - \frac{\mu_0(C)}{\mu_0(E)}. \tag{4.5}$$

Note that this metric is not a function of just the differences $\Delta(E)$ and $\Delta(C)$, so it does not suffer from problems of $\Delta(E)$ being close to zero, like the ICER. However, this metric has been criticized by Briggs et al. (2007) who note that economic decisions should be made based on the marginal difference between an option and the next-best comparator, which is accomplished using μ_{ICER}. Therefore, we will not consider it further in this paper.

We make some final comments about the mean-based metrics before moving to estimation uncertainty. Note that, if the underlying distributions are symmetric, the mean coincides with the median, so these metrics can also be viewed as comparisons of differences in medians, There are other cases also where the differences in the means coincide with the differences in medians. Suppose $F_1(t; C) = F_0(t-\eta_C; C)$ and $F_1(t; E) = F_0(t-\eta_E; E)$, i.e., a location (or shift) model holds. Then, the differences in means, medians and other location parameters coincide with the differences in the location parameters η_C and η_E. One consequence is that we can use other estimators besides the sample means to estimate the differences. This will be taken up later in the paper.

In practice, however, the treatment difference will be complex, affecting the dispersion and other aspects of the distribution. The focus on means for CEA inference comes from an economic perspective (Briggs and Gray, 1998; Thompson and Bar-

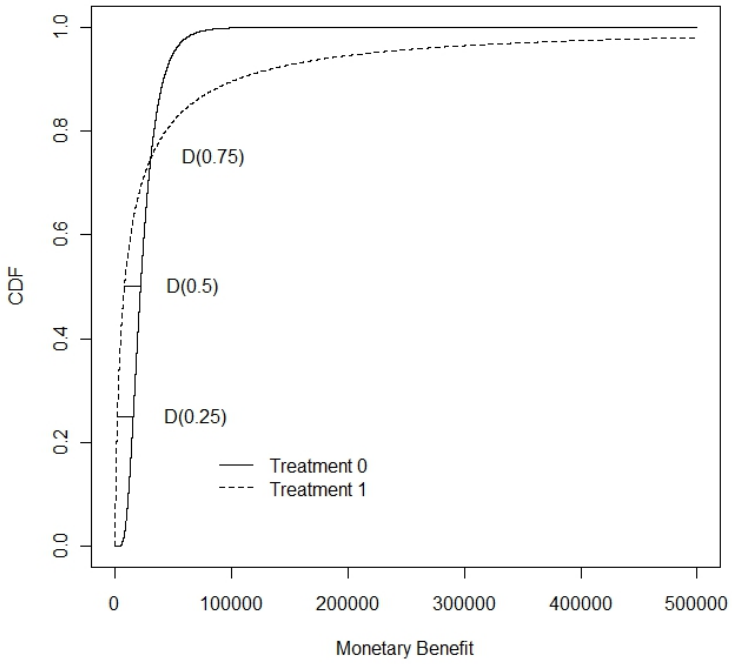

Figure 4.2: Monetary Benefit Distributions.

ber, 2000; O'Hagan and Stevens, 2002). This certainly makes sense for cost as the decision-maker would be very interested in the differences in total (expected) costs of the two treatments. It is less compelling for effectiveness and MB, where there are important ethical considerations. If the distribution of effectiveness or MB outcomes is highly skewed, the median is a more appropriate singe-number summary than the mean. For example, consider the two distributions (F_1 corresponding to treatment 1 and F_0 corresponding to baseline) in Figure (4.2). The expected value of the new treatment is $39,400$ and the standard is $25,000$, so the new treatment is better in terms of mean. But a comparison of the two CDFs in the figure shows that the standard distribution is better than the new treatment even up to the 75-th percentile.

We discuss some alternatives to mean-based comparisons later in the paper. These

include median of the differences in treatments and comparisons in terms of utility functions and stochastic dominance.

4.3 Incorporating Statistical Uncertainty with Mean-Based Metrics

As noted earlier, the actual mean values required for CEA analysis are not known in practice. Rather, they are estimated from data, typically from clinical trials (see, for example Meenan et al. (1998), van Hout et al. (1994), and Kinlay et al. (1996)). The estimation of the parameters induces uncertainty which can be substantial when the sample sizes are small to moderate, so this must be incorporated in the comparisons. This topic has been considered by many authors in the CEA literature. We provide a review and critical assessments of the key methods.

We first consider the situation where the field data are uncensored, the only case that appears to have been considered in the CEA literature. Extensions to censored data will be described at the end of this section.

Suppose we have data on n_0 subjects for the baseline treatment and n_1 subjects for the new treatment. Typically n_0 will be larger than n_1, although this is not an important issue for our purposes here. We assume the data from the different subjects are independent and identically distributed from the corresponding cost and effectiveness distributions. Let $\hat{\mu}_j(C)$ and $\hat{\mu}_j(E)$, $j = 0, 1$, be the sample averages of the observed costs and effectiveness values. Further, let $\widehat{\Delta}(C) = \hat{\mu}_1(C) - \hat{\mu}_0(C)$ and $\widehat{\Delta}(E) = \hat{\mu}_1(E) - \hat{\mu}_0(E)$

Then, we can define the corresponding estimated quantities as

$$\hat{\mu}_{ICER} = \frac{\widehat{\Delta}(C)}{\widehat{\Delta}(E)},$$

and

$$\hat{\mu}_{NMB}(\lambda) = \lambda\widehat{\Delta}(E) - \widehat{\Delta}(C).$$

We also need notation for the variances and their estimators. Let $V_j(C)$ be the variance $\hat{\mu}_j(C)$ and $V_j(E)$ be the variance of $\hat{\mu}_j(E),\ j = 0, 1$ and the corresponding lower cases as their estimated values. Further, let CV_j be the covariance of $\hat{\mu}_j(C)$ and $\hat{\mu}_j(E)$ and cv_j be the corresponding estimated values, $j = 0, 1$. We define the variances of $\widehat{\Delta}C$ and $\widehat{\Delta}E$ to be

$$V_C = V_0(C) + V_1(C) \tag{4.6}$$

and

$$V_E = V_0(E) + V_1(E), \tag{4.7}$$

with their estimated values being denoted v_c and v_e, respectively. Likewise, $Cov(\Delta C, \Delta E)$ is given by

$$V_{C,E} = CV_0 + CV_1, \tag{4.8}$$

with estimated values denoted cv_0 and cv_1, respectively.

The estimated values can be obtained from field data. For example, if $S_0^2(C)$ is the sample variance of the n_0 cost values in the sample for the baseline treatment, $S_0^2(C)/n_0$ is an unbiased estimator of $V_0(C)$. We denote the estimated standard deviation (also called estimated standard error) of $\widehat{\mu}_0(C)$ as $\widehat{SE}(\widehat{\mu}_0(C))$. The standard errors and estimated covariances for the other quantities can be computed similarly.

There are two ways by which the estimation uncertainty can be incorporated into the CEA analysis: hypothesis testing and confidence intervals. For example, hypotheses of interest include the null hypothesis of: a) $H_0 : \Delta(E) \leq 0$ against the alternative $H_1 : \Delta(E) > 0$; b)$H_0 : \Delta(C) \geq 0$ against the alternative $H_1 : \Delta(C) < 0$; c)$H_0 : \mu_{ICER} \leq \lambda$ against the alternative $H_1 : \mu_{ICER} > \lambda$; or d) $H_0 : \mu_{NMB}(\lambda) \leq 0$ against the alternative $H_1 : \mu_{NMB}(\lambda) > 0$. Hypotheses (a) and (b) are often tested

can sequentially – a) and then b) – since an advantage in terms of cost would be attractive only if the treatment is more effective. In all cases, the problems are formulated so that the null hypotheses favor the standard or baseline treatment, and sufficient evidence is needed to reject it in favor of the new treatment. One can also test two-sided hypotheses which test just whether the standard and new treatments have the same effect, although this is not as meaningful in the CEA context.

In practice, hypothesis testing is less useful than estimating the difference in the treatment effects and getting confidence intervals around the point estimates. A confidence interval is more useful because it gives a set of plausible values for the difference in treatment effects. The interval can also be used to test the null hypothesis by determining if the hypothesized null value falls inside the confidence interval or not. One (minor) disadvantage is that one has to specify a confidence level a priori before the interval is computed, so different confidence levels may lead to different conclusions about the hypotheses. On the other hand, one can compute the p-values once for all for the hypothesis tests, and they provide an indication of the evidence for or against the null hypotheses. We will illustrate the issues later in the paper.

There are many papers in the CEA literature that have discussed the problem of statistical uncertainty for mean-based comparisons. These papers discuss examples and provide summary statistics on $v_0(E)$, $v_0(C)$, $v_1(E)$, $v_1(C)$, cv_0, cv_1, which are needed for the analysis (see, for example, Willan (2001), Laska et al. (1997), van Hout et al. (1994), and Chaudhary and Stearns (1996)). However, some of the methods discussed in this paper, including resampling methods and stochastic dominance comparisons, require the entire sample data, not just the summary statistics. Such data are difficult to find in the literature. Thus, we will use artificial data, given in

Appendix 4.6.1, to illustrate the methods and compare them.

4.3.1 Use of Large Sample Approximations for One-Dimensional CEA

Our goal here is to review the main techniques and also clarify some confusion that seems to be present in a few CEA papers. The methods in this section depend on a straightforward application of the central limit theorem, i.e., the distributions of all the sample means are approximately normal when the sample sizes n_0 and n_1 are sufficiently large.

NMB

Since $\widehat{\mu}_{NMB}(\lambda)$ is a linear combination of the sample means, it also has a normal distribution in large samples. Specifically, $[\widehat{\mu}_{NMB}(\lambda) - \mu_{NMB}(\lambda)]/\widehat{SE}(\widehat{\mu}_{NMB}(\lambda))$ is distributed approximately as a standard normal random variable. The standard error of $\widehat{\mu}_{NMB}(\lambda)$ can be calculated as follows. Its variance is

$$\begin{aligned} Var(\widehat{\mu}_{NMB}(\lambda)) &= \lambda^2 Var(\widehat{\Delta}(E)) + Var(\widehat{\Delta}(C)) - 2\lambda Cov(\widehat{\Delta}(E), \widehat{\Delta}(C)) \\ &= \lambda^2[V_1(E) + V_0(E)] + [V_1(C) + V_0(C)] - 2\lambda[CV_1 + CV_0]. \end{aligned} \tag{4.9}$$

The last expression follows from the independence of the data for treatments 0 and 1. By plugging in estimates on the right-hand side and taking the square root, we get the estimated standard error as

$$\widehat{SE}(\widehat{\mu}_{NMB}(\lambda)) = \left(\lambda^2[v_1(E) + v_0(E)] + [v_1(C) + v_0(C)] - 2\lambda[cv_1 + cv_0]\right)^{1/2}. \tag{4.10}$$

One can now construct approximate confidence intervals as well as tests of hypotheses relating to $\mu_{NMB}(\lambda)$ (see Stinnett and Mullahy (1998)). For example, to test if $\mu_B(\lambda) > 0$ at level α, one can use the approximate $(1-\alpha)-$level lower confidence bound (LCB) for $\mu_B(\lambda)$: $[\widehat{\mu}_B(\lambda) + z_\alpha \widehat{SE}(\widehat{\mu}_B(\lambda))$ where z_α is the $\alpha-$th quantile of the standard normal distribution. If this LCB does not include zero, then one can

decide that $\mu_B(\lambda) > 0$ or that the new treatment is better than the current one in terms of NMB (at this value of λ), even after accounting for statistical uncertainty. Otherwise, one may decide to go with the current treatment. Similar tests can be constructed for other hypotheses of interest.

The large-sample procedures will work well when the sample sizes for the field data are sufficiently large. The general rule of thumb in statistics is that both sample sizes should be at least 30. Clearly, the quality of the approximation will improve as the sample sizes get bigger. As an aside, sometimes one sees the use of a t-statistic (Laska et al., 1999) with the value of z_α replaced by a quantile from a t-distribution with some degrees of freedom when the sample sizes are small. The theory for t-distribution is valid when the underlying distributions of both cost and effectiveness random variables are (close to) normal, which is not likely to be the case here. We note that there is no justification for doing this in the present context.

For our example, with $\lambda = 50,000$, $\hat{\mu}_{NMB}(\lambda) = 43,810$. A comparison based on just the point estimate will suggest that the new treatment is cost-effective. Construct a confidence interval, with $\widehat{SE}(\hat{\mu}_{NMB}(\lambda)) = 30,128$, we get a 90% (large-sample) lower confidence bound for $\mu_{NMB}(\lambda)$ as -15240; the corresponding 95% two-sided confidence interval is $(-15240, 102861)$. We see that, after incorporating the statistical uncertainty, the new treatment is not cost-effective.

ICER

We can approximate $[\hat{\mu}_{ICER} - \mu_{ICER}]/\widehat{SE}(\mu_{ICER})$ by a standard normal random variable, and, as before, use this as the basis of large-sample inference. For example, we can decide that μ_{ICER} is greater than λ_0 if the lower end-point of a one-sided lower confidence bound for μ_{ICER} at some specified level, say λ_U is greater than or equal to λ_0.

This large-sample approach has been discussed in the CEA literature (see, for, example, O'Brien et al. (1994); van Hout et al. (1994); Briggs and Fenn (1998)). There is, however, some confusion. Some authors claim that the ratio will be approximately a ratio of Cauchy random variables in large samples and that the distribution will not have finite moments (Wakker and Klaasen, 1995; Siegel et al., 1996; Zethraeus et al., 2003). The point about Cauchy random variables is wrong and appears to arise from mistakenly equating the ratio

$$\left[\frac{\widehat{\Delta}(C)}{\widehat{\Delta}(E)}\right] - \left[\frac{\Delta(C)}{\Delta(E)}\right] \text{ with } \frac{[\widehat{\Delta}(C) - \Delta(C)]}{[\widehat{\Delta}(E) - \Delta(E)]}.$$

The latter does converge to a ratio of (correlated) normal random variables and does not have any moments. But the former expression is the quantity of interest. It can be approximated by a linear combination of random variables, each of which converges to normal (as seen by equation (4.11)), so it has a limiting normal distribution.

The issue of whether $\hat{\mu}_{ICER}$ has finite moments or not does not have any thing to do with the large-sample approximation. It is indeed true that $\hat{\mu}_{ICER} = \frac{\widehat{\Delta}(C)}{\widehat{\Delta}(E)}$ may not have a mean in finite samples. This is so when the distribution of $\widehat{\Delta}(E)$ gives positive mass to any region that includes zero, and it can also be true in other cases. However, as the sample size gets larger, the variance of the estimator decreases, so both the numerator and the denominator will get closer and closer to their true values. So, in large samples, the existence of moments is not an issue.

The simplest approach for large-sample inference is to use a direct approximation of the distribution of $\hat{\mu}_{ICER}$ using a first-order Taylor series expansion. This leads to

$$\hat{\mu}_{ICER} - \mu_{ICER} \approx \frac{1}{\Delta^2(E)} \left(\Delta(E)[\widehat{\Delta}(C) - \Delta(C)] - \Delta(C)[\widehat{\Delta}(E) - \Delta(E)]\right). \tag{4.11}$$

Since $\hat{\Delta}(E)$ and $\hat{\Delta}(C)$ have large-sample normal distributions, the large-sample nor-

mality of $\hat{\mu}_{ICER}$ follows. The above linearization also yields the following approximation for the variance:

$$
\begin{aligned}
Var(\hat{\mu}_{ICER}) \approx \frac{1}{\Delta^4(E)} \Big(& \Delta^2(E)\ Var(\widehat{\Delta}(C)) + \Delta^2(C) Var(\widehat{\Delta}(E)) \\
& -2\Delta(E) Cov(\widehat{\Delta}(E), \widehat{\Delta}(C)) \Big) \\
= \frac{1}{\Delta^4(E)} & \left(\Delta^2(E)\ \left[V_0(C) + V_1(C) \right] + \Delta^2(C) \left[V_0(E) + V_1(E) \right] \right. \\
& \left. -2\Delta(E) \left[CV_1 + CV_0 \right] \right).
\end{aligned} \tag{4.12}
$$

The estimated standard error of $\hat{\mu}_{ICER}$ is obtained by substituting the estimates and taking the square root.

For our illustrative example, $\hat{\mu}_{ICER} = 25407$ which is less than the $\lambda = 50000$. $\widehat{SE}(\hat{\mu}_{ICER}) = 300387$, indicating huge variability in the estimate. The large-sample two-sided 95% confidence interval is $(-18249, 69064)$. It contains zero, suggesting that the differences in cost or effectiveness are not very large (most likely the difference in effectiveness). The negative value is difficult to interpret on its own, since this could be due to increased effectiveness and decreased cost or to decreased effectiveness and increased cost, or to estimation uncertainty in either variable. We will address this issue in Section 4.3.3

It is well known in the statistical literature that ratio estimators, such as $\hat{\mu}_{ICER}$, can be unstable and that very large sample sizes are needed before the normal approximation will be adequate. For this reason, alternative methods of constructing confidence intervals and hypothesis tests have been proposed. The most common one is Fieller's method, which has been discussed also in the CEA literature (Chaudhary and Stearns, 1996; Polsky et al., 1997; Briggs and Fenn, 1998). Perhaps the simplest way to motivate this approach is as follows. It is well known in statistics and has also been rediscovered in the CEA literature (see Zethraeus et al. (2000)).

Recall that $\mu_{ICER} = \lambda$ if and only if $\mu_{NMB}(\lambda) = 0$. Suppose we used the large-sample distribution of $\widehat{\mu}_{NMB}(\lambda)$ to determine the set of values of λ for which we will we accept the null hypothesis that $\mu_{NMB}(\lambda) = 0$. Then, from the 1-1 relationship between $\mu_{ICER} = \lambda$ and $\mu_{NMB}(\lambda) = 0$, this set of values of λ will yield a confidence region for μ_{ICER}. Because the distribution of $\widehat{\mu}_{NMB}(\lambda)$ is more stable (it is just a linear combination of the random variables) than that of $\widehat{\mu}_{ICER}$, the resulting confidence region will also be more stable and will have better properties. The issue now is how to construct such a confidence region, whether the region will actually be an interval or disconnected, if it is an interval, whether it is finite. (Fieller, 1954) showed that, under certain conditions, the interval exists and is given by

(4.13)

$$\frac{1}{1-g}\left[\hat{\mu}_{ICER} - \frac{gv_{C,E}}{v}_{E} \mp \frac{z_{\alpha/2}}{\widehat{\Delta}E}\left(v_C - 2\hat{\mu}_{ICER}\, v_{C,E} + \hat{\mu}^2_{ICER} v_E - g\left(v_C - \frac{v^2_{C,E}}{v_E}\right)\right)^{1/2}\right]$$

where $g = z^2_{\alpha/2} v_C/(\widehat{\Delta}^2(E))$. For our illustrative data, Fieller's method gives the following 95% confidence interval: $(-23883, 58059)$. This is shifted to the left and shorter than the interval from the direct large-sample normal approximation, but still includes the value of zero.

The methods discussed thus far can also be used to test various hypotheses related to μ_{ICER} or μ_{NMB}. See, for example, Gardiner et al. (2000).

4.3.2 Resampling Methods for One-Dimensional CEA

We discuss two alternatives to large-sample methods based on resampling the data. The first is the randomization test (also called permutation test) and the second is bootstrap. Although both involve resampling the data, their goals are different. Randomization attempts to get the null distribution of a test statistic; here the sampling is done without replacement. Bootstrap, on the other hand, attempts

to get the actual distribution of an estimator or test statistic; further sampling is done with replacement.

Randomization tests have a long history in statistics and are based on a very simple idea. If the new treatment and the current one are not different, then the two sets of data should have the same distribution. Consider a generic two-sample problem where $N = n_0 + n_1$ subjects were randomly assigned to two treatments. Let $\{Y_{0,1}, ..., Y_{0,n_0}\}$ be the data from one treatment and $\{Y_{1,1}, ..., Y_{1,n_1}\}$ be the data from a second treatment. Let Z be an appropriate test statistic, such as $(\bar{Y}_1 - \bar{Y}_1)/\widehat{SE}(\bar{Y}_1 - \bar{Y}_0)$, the standardized difference of the sample means. Let $\{V_1, ..., V_N\}$ be the combined data. If there is no difference among the two treatments, we can randomly partition the V data into two groups of size n_0 and n_1 and compute a new value of the test statistics, say Z^*, for the new data set. The Z^* should have the same distribution as Z under the null hypothesis. If we do this for all $\binom{N}{n_1}$ possible partitions of $\{V_1, ..., V_N\}$ into two samples of size n_1 and n_0, we get $\binom{N}{n_1}$ possible values, and hence the distribution of Z. Now, if the observed value of Z for our original data set falls in the tails of this distribution, we can interpret it as evidence that the null hypothesis of no treatment effect is not appropriate. See Lehmann et al. (2005) and references therein for more details. If the sample size is large, the permutation distribution can be approximated by a normal distribution (and this will lead to the same results as in the last section).

In the case of CEA, randomization tests can be used on the any of the statistics: differences in sample means of costs and effectiveness, NMB, or ICER. We illustrate it on NMB and test whether $\mu_{NMB} > 0$ or equivalently whether $\mu_1(MB) > \mu_0(MB)$. Define the Y_0-random variable to be MB_0 from the illustrative example, and the Y_1-random variable to be MB_1. In this case, $n_0 = n_1 = 32$ and $N = n_0 + n_1 = 64$.

Since the value of all possible partitions, $\binom{N}{n_1}$, is very large, we randomly sampled $10,000$ permutations, and calculate $10,000$ values of Z^* statistics. The test value for the actual sample is 1.59, which has a p-value of 0.07 based on the bootstrap sample. Thus, the effect is marginal.

The bootstrap represents a different approach to the problem. There are many refinements of the original bootstrap approach suggested by Efron (1979). Some of them have also found their way into the CEA literature. See, for example, Briggs and Fenn (1998) and O'Brien et al. (1994), with examples in Chaudhary and Stearns (1996) and Hunink et al. (1998). The basic idea is to resample data, separately and with replacement, from each of the two samples. Repeatedly sampling and computing the value of the statistic gives us a distribution which can be used to approximate the true distribution. Again, the idea can be used to approximate the distribution of any of the statistics we have discussed thus far. There are several variations of the bootstrap for testing hypotheses and constructing confidence intervals (see Efron and Tibshirani (1993) and Briggs and Fenn (1998) in the CEA literature). Perhaps the most common is the percentile method which used the $100(\alpha/2)$-percentile and $100(1-\alpha/2)$-percentile of the bootstrap distribution as confidence limits for the difference in treatment effect. Other common approaches include the bootstrapped t-statistic (see Jiang and Zhou (2004) in the health economics literature) and the bias-corrected, accelerated bootstrap (see Efron and Tibshirani (1993)).

We illustrate bootstrapping with the simple percentile method and compare it with the results from the large-sample approximations for μ_{ICER}. Using a bootstrap procedure to create $10,000$ samples in the illustrative example of $\hat{\mu}^*_{ICER}$ and the percentile method results in a 95% confidence interval for $\hat{\mu}_{ICER}$ of $(-14702, 61726)$. This interval is similar to, but narrower than, those based on the Fieller and large-

sample normal approximations.

4.3.3 Two-dimensional CEA Based on Mean Metrics

Use of Large Sample Approximations

We can incorporate estimation uncertainty in the two-dimensional setting along the same lines as before. For example, we can use the large-sample approximations to get confidence intervals for $\widehat{\Delta}(C)$ and $\widehat{\Delta}(E)$ separately. Recall the definitions of the variance of $\widehat{\Delta}(C)$ given in Equation (4.6) and for $\widehat{\Delta}(E)$ given in Equation (4.7).

We know that $\widehat{\Delta}(E)$ and $\widehat{\Delta}(C)$ are approximately (and jointly) bivariate normal with mean $\Delta(E)$ and $\Delta(C)$ and variance covariance matrix Σ which can be estimated by

$$\hat{\Sigma} = \begin{pmatrix} v_E & v_{C,E} \\ v_{C,E} & v_C \end{pmatrix}.$$

A conservative, rectangular joint confidence region that does not depend on the correlation between the two random variables can be obtained using the Bonferroni bound (see O'Brien et al. (1994) in the context of CEA). Let the one-dimensional confidence intervals with levels $(1-\alpha_1)$ for $\widehat{\Delta}C$ and $(1-\alpha_2)$ for $\widehat{\Delta}E$ be given by $[\widehat{\Delta}C \pm z_{\alpha_1}/\widehat{SE}(\Delta C)]$ and $[\widehat{\Delta}E \pm z_{\alpha_2}/\widehat{SE}(\Delta E)]$. Taking the intersection of these two regions in the two-dimensional plane yields a rectangular confidence region. From the Bonferroni bounds, the confidence level of this region is greater than $(1-\alpha_1-\alpha_2)$. It is common to take $\alpha_1 = \alpha_2 = \alpha/2$ to get an overall $(1-\alpha)-$level region. Figure (4.3) shows the confidence box region for the illustrative example.

It is well known that this rectangular confidence region is not efficient in such cases because: a) the Bonferroni bound used to combine the two intervals is conservative; and b) an ellipsoidal confidence region that takes into account the correlation structure is more efficient (smaller is area for a given level). The ellipsoidal region is

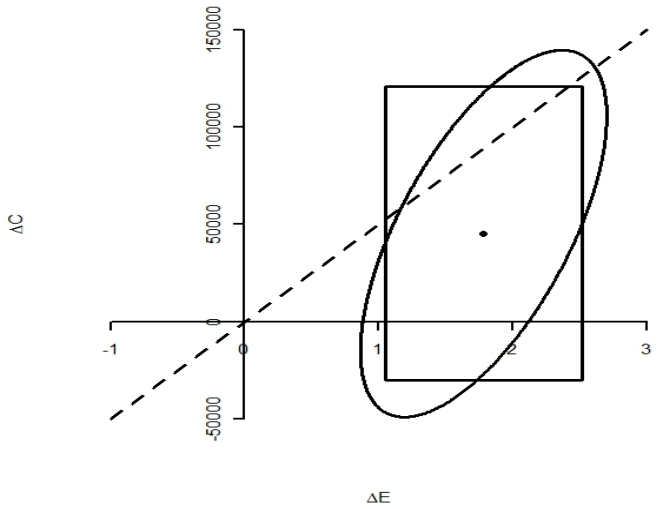

Figure 4.3: Cost-effectiveness plane with 95% confidence regions: Rectangular region based on Bonferroni bound and elliptical region.

constructed as follows (see van Hout et al. (1994) for a discussion of this in the CEA literature). Define the vector

$$X = [\widehat{\Delta}(E) - \Delta(E), \widehat{\Delta}(C) - \Delta(C)]^T.$$

Recall that $X^T\hat{\Sigma}^{-1}X$ is approximately distributed as $\chi^2(2)$, a chi-squared random variable with two degrees of freedom. Let K_α be the $(1-\alpha)-$th upper quantile of the $\chi^2(2)$ distribution. Then, the set of $\{(\Delta(E),\ \Delta(C) : X^T\hat{\Sigma}^{-1}X \leq K_\alpha\}$ yields an approximate $(1-\alpha)$ joint confidence region for $(\Delta(E),\ \Delta(C)$ in the two-dimensional plane. This region is an ellipse as shown in Figure(4.3) for the illustrative example.

Figure (4.3) illustrates the elliptical and rectangular confidence regions for our example CEA; one can discern that the ellipse covers a smaller area than the rectangle. We can also see that the point $(0, 0)$ is outside the ellipse as well as the rectangular region, so the corresponding null hypothesis will be rejected at this confidence level using either method. Both confidence regions include the point $\Delta(C) = 0$, suggesting that there is no statistically significant difference in cost between the two treatments.

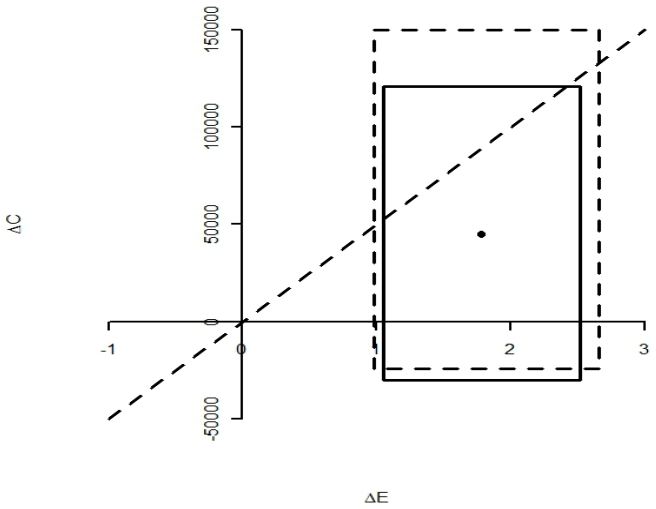

Figure 4.4: Cost-effectiveness plane with 95% rectangular confidence regions: bootstrap rectangle (dashed) with large-sample normal approximation (solid) shown for comparison
.

Resampling Methods

Resampling methods have also been used to get confidence regions in two dimensions. In the CEA context, Briggs and Fenn (1998) discussed the rectangular region using bootstrapping and Bonferroni bounds. We use our data example again to illustrate this approach. The two-dimensional rectangular region, based on $10,000$ replications, using the percentile method is shown in Figure (4.4). The dotted line is the region using the bootstrap, and the large-sample normal approximation (solid line) is also shown for comparison. The region using the bootstrap is somewhat larger and asymmetric (about the estimates), suggesting that the large-sample approximation is not as good, especially for the cost dimension.

One can also use the bootstrap the distribution of $X^T\hat{\Sigma}^{-1}X$ to get the critical value K_α instead of using the χ^2 approximation. The details are very similar and are omitted.

4.3.4 Use of Rank Statistics

The methods so far deal with mean-based metrics and the statistical uncertainty that arises from estimating them. As noted earlier, the differences in means coincide with differences in other location parameters in some special cases. In this section, we propose a method for assessing the median of the difference in treatment. It is based on ranks of the data rather than the original numerical values. When the median of the differences coincides with the mean, this procedure can be used as a robust alternative to the sample mean-based statistical procedures. The method can be more efficient than sample mean-based methods when the underlying distributions have heavier tails than normal, such as the logistic and double exponential.

There is a huge literature on rank-based methods (see, for example, Lehmann and D'Abrera (1975)). We focus here on just the Wilcoxon test (sometimes called the Mann-Whitney test), which has been widely used. There has been discussion of Wilcoxon rank-sum test in the CEA literature. There is a discussion of sample sizes needed to achieve statistical power for a Wilcoxon test for ICER in Laska et al. (1999). Mark et al. (1995) use the Wilcoxon rank-sum test to identify differences in resource consumption between treated and untreated groups, and uses those differences to estimate cost differences; Mehta et al. (1997) also used the Wilcoxon rank-sum test to identify significant differences in cost categories, and actually forms CE-ratios using medians instead of means.

The Wilcoxon test procedure works as follows. Let $\{Y_{1,i},\ i = 1, \ldots, n_1\}$ be the data for the n_1 patients assigned to Treatment 1 and $\{Y_{0,j},\ j = 1, \ldots, n_0\}$ be the data for the n_0 patients assigned to Treatment 0. Let $N = n_1 + n_0$ and $\{Z_k,\ k = 1, \ldots, N\}$ be the combined Y_0 and Y_1 data. Rank the Z_k's and let $S_1 < ... < S_{n_1}$ be the ranks of the Y_1's among the Z_k's. (We assume there are no ties. The case with ties can be also

handled and we refer readers to Lehmann and D'Abrera (1975)). Let $W = \sum_{i=1}^{n_1} S_i$. If there was no difference between the new and old treatments (null hypothesis), W_s is symmetric with expected value $n_1(N+1)/2$ and variance $n_0 n_1(N+1)/12$. If the new treatment is better, then W_s will tend to be larger. We need to know the null distribution of W_s to develop formal tests.

We can use a randomization framework (same as the one used earlier) to get the null distribution as follows. If there is no difference among the two treatments, all the Z_k values come from a homogeneous group. So we can randomly take a sample of n_1 values from the Z_k's, assign them to treatment 1 (call them Y_1's) and assign the remainder to treatment 0 (call them Y_0's). Doing this will result in a value of W_s^*. There are $\binom{N}{n_1}$ possible ways we can do this. All of these possible values provide the null distribution of W_s. To test the null hypothesis at level α, take the $(1-\alpha)-$th quantile of the null distribution and reject the null hypothesis if the observed value of the data is greater than this quantile. (see Lehmann et al. (2005) for more details).

If n_0 and n_1 are both moderate (say bigger than 30), then the distribution of W_s can be approximated by a normal distribution with mean $n_1(N+1)/2$ and variance $n_0 n_1(N+1)/12$. This can be used to get large-sample test procedures. It turns out that the Wilcoxon test is consistent for testing the general hypothesis $F_1 = F_0$ against $F_1 \leq F_0$ (treatment 1 is more effective than treatment 0).

We illustrate the results on the difference in monetary benefits using our data, so the two distributions of interest are MB_0 and MB_1. The value of the Wilcoxon statistics is $W = 1370$. Using $10,000$ random samples from the full set of possible permutations, the randomization analysis yields a lower 5% critical values as 913, indicating a difference in the two distributions. The use of the large-sample normal approximation, instead of the permutation distribution, gives a critical value of 917,

very close to the randomization result.

A more interesting problem is getting confidence interval for the difference in the treatment effects. Consider the $n_0 n_1$ differences $Y_{1,i} - Y_{0,j}$. Let $D_{(1)} < ... < D_{(n_0 n_1)}$ be the ordered values of these differences. Then, the estimate of the median of the differences $Y_{1,i} - Y_{0,j}$ is estimated (naturally) by the median of the corresponding sample values, i.e., median of the $D_{(k)}$'s. This is referred to as the Hodges-Lehmann estimator in the statistical literature (Lehmann and D'Abrera, 1975). Note that when the medians coincide with the mean, this estimator provides an alternative way to estimate the differences in the means. It tends to be robust to outliers in the data. It can also be more efficient (in a statistical sense) when the underlying distribution has heavier tails than a normal distribution, such as logistic or double exponential. Note also that the ranks are invariant to monotone transformations of the data, provided the same transformation is applied to data from both treatments.

Exact confidence intervals for the Hodges-Lehmann estimator can be based on the Wilcoxon rank-sum test. The critical values $C_{\alpha/2}$ form a two-sided confidence interval when n_0 and n_1 are small have been tabulated in the literature. With sufficiently large samples, one can use large-sample approximation (Lehmann and D'Abrera, 1975) given by

$$C_{\alpha/2} \approx n_0 n_1/2 - Z_{\alpha/2} \left(n_0 n_1 (N+1)/12\right)^{1/2} \tag{4.14}$$

The $C_{\alpha/2}$ should be taken to be the nearest integer. The confidence interval for the median difference is then given by $(D_{(C_{\alpha/2})}, D_{(n_1 n_1 + 1 - C_{\alpha/2})})$.

For our data, the Hodges-Lehmann estimator is 54201, indicating an overall cost-effective treatment effect. The confidence interval using the large-sample approximation for the distribution is $(25250, 81985)$. It is noteworthy that this confidence interval is entirely positive, in contrast to the case for the means. This suggests that

the distributions are not symmetric, so the two quantities are estimating different treatment effects.

The performance of various rank methods have been compared in the statistical literature (Lehmann and D'Abrera, 1975). In particular, it is known that the Wilcoxon procedure does well when the underlying distributions are logistic. Alternatives that work well with other distributions, such as the normal scores, have also been proposed.

4.3.5 Paired Data Analysis

Most of the CEA literature deals with the case where the field data for the standard and new treatment are independent. This is the most common situation – there is likely to be considerable field data on the standard and small studies are then conducted to collect data on the new treatment. It is, however, possible that the two treatments are compared in the same study and that the two treatments are assigned to subjects in such a way that the responses are paired or correlated. In this case, one would just take the difference in the responses (cost and effectiveness) and analyze the differences as one-sample iid data. Both large-sample approximations and resampling methods can be used with the one-sample data.

4.3.6 Censored Data

Incorporating uncertainty when the data are censored does not appear to have been discussed in the CEA literature, so we provide an overview here. Perhaps the most common type of censoring in clinical trials is fixed time censoring: the study is stopped at some fixed (prespecified) time and some subjects are alive at the end of the study. There can be other, more complex forms of censoring, including multiple right censoring where different subjects are right censored at different times.

Consider first the case of fixed time censoring. Specifically, suppose data from both samples are right censored at the same time t_0. Then, we cannot estimate the means of the underlying distributions from the data since we do not have any information on when the subjects that were alive at time t_0 would fail. So, one cannot use the mean-based metrics. However, we can base the comparisons on the conditional means – $E[Y|Y \leq t_0]$ – since this quantity can be estimated from the data. If the data from the two samples are right censored at different times, we have to take the conditional distribution at the minimum of the two censoring times. One could estimate median-based metrics from the sample data if the number of censored observations is smaller than 1/2 for both samples.

The problem is more complex if there is multiple right censoring. However, several of the rank-based methods for comparing two populations have been extended to censored data in the statistical literature. These include generalizations of the Wilcoxon test and the log-rank test (see Lawless (2003)), with the latter being the most common.

4.4 Beyond Mean-Based CEA Analysis

We have focused thus far on CEA based on single-number metrics: means, median, and expected utilities. The remainder of the paper focuses on CEA based on the full distribution of cost and effectiveness outcomes.

To motivate this analysis, consider Figure (4.5) which shows a hypothetical example, the cumulative distribution functions of (say) effectiveness outcomes for a standard and a new treatment, measured in life-years. In this case, the treatment improves the mean length of life from about 4.4 years to 5 years, so mean-centric decision-making would recommend adopting the treatment. However, the median

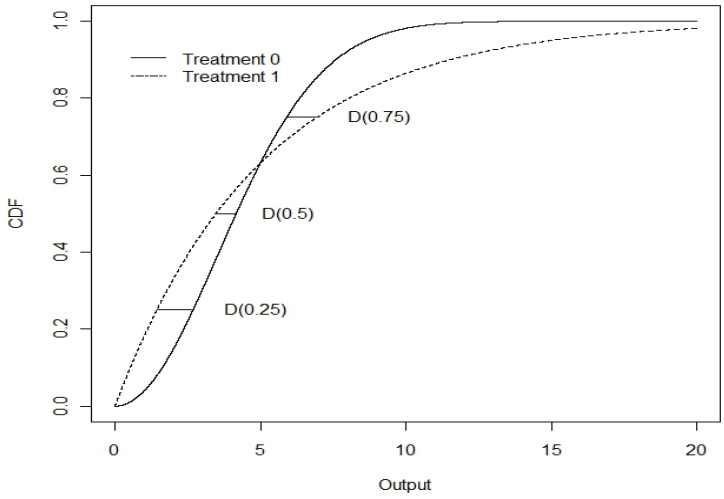

Figure 4.5: Hypothetical Effectiveness Distribution.

length of life under treatment actually decreases from 4.2 years to 3.5 years. We can see from the figure that this occurs due to a long upper tail under the new treatment, which raises the mean for some of the treated population, but has no benefit for most of the population. In fact, the treatment lowers effectiveness for about 60% of the population. The nature of the treatment effects are complex and cannot be characterized by just the differences in means, and different metrics can lead to different decisions. While a health analyst or decision-maker may still prefer to implement the treatment, an analysis of the two distributions will provide much more information about the trade-offs being made and the risks.

4.4.1 Utility Functions

Utility functions provide a natural way to map the cost and effectiveness outcomes into a decision-theoretic framework in the sense of von Neumann and Morgenstern (Von Neumann and Morgenstern, 1947). There is a huge literature on this topic, and we touch on it very briefly.

Using just the means is equivalent to assuming a utility function that is linear

in the outcomes – in other words, being risk neutral (Keeney and Raiffa, 1976). This is inadequate in many situations. For example, a person is not likely to be indifferent between a placebo that has no effect on his/her life expectancy versus a new treatment that offers a 50% chance of doubling life expectancy but carries a 50% risk of instant death, but these two options offer the same expected utility to a risk-neutral decision maker. In fact, it is known that patients will exhibit different behaviors – risk-seeking, risk-averse, and risk-neutral, depending on the situation at hand (Pliskin et al., 1980). It is also generally recognized that policymakers tend to be risk-averse – concerned about new interventions being extremely harmful with a small probability, or concerned about budgetary problems (Koerkamp et al., 2007).

In the rest of this section, we consider comparisons based on expected utilities. If the utility functions are fully specified, then the problem of incorporating statistical uncertainty of estimated parameters using utility functions is the same as with the estimated mean. Consider a particular outcome, say, cost. We can treat the utility outcomes as the new random variables of interest, estimate their means using the sample data and incorporate statistical uncertainty using one of the methods that have already been described.

But the utility functions are rarely known completely. Here we restrict attention to functions of the form $U(x) = (x+b)^c$, with $0 < c < 1$, a class of risk-averse utility functions (Keeney and Raiffa, 1976), and examine the sensitivity of the decisions to the values of b and c. Suppose that the distributions for the MB outcome are, again, the same ones in Figure (4.2): $F_0 = LN(10, 0.5)$ and $F_1 = LN(9, 2)$. Suppose $b = 100$ and $c = 0.3$, and so $u(x) = (x + 100)^{0.3}$. For this case, treatment 1 has an expected utility of 17.9 compared to 20.3 for the standard, so the standard is preferred. (Recall that treatment 1 had a higher expectation than the standard, even

though the standard was better than the new treatment up to the 75th percentile.) However, the parameters b and c are usually unknown, so let us examine how the comparisons change as the values of b and c change. First, consider a sensitivity analysis. The left panel of Figure (4.6) shows how the difference in expected utilities change as b varies with c fixed (at several different values).

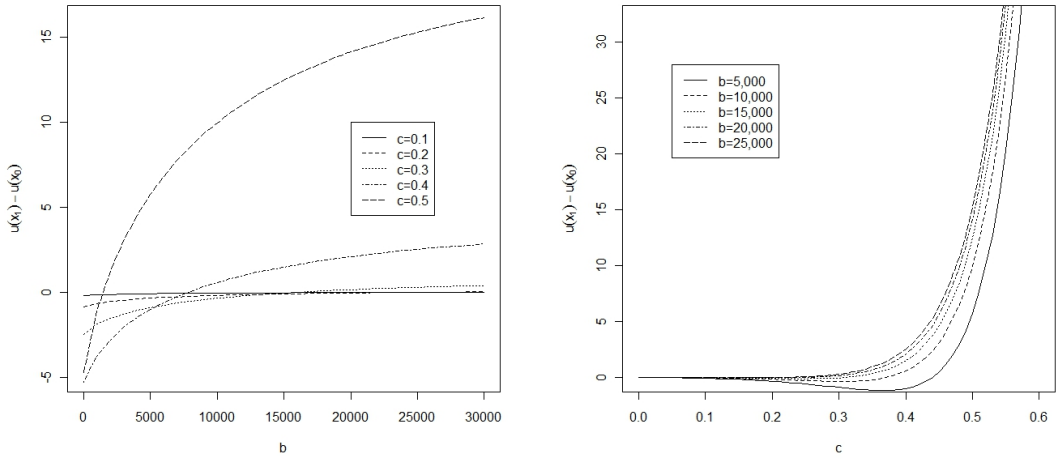

Figure 4.6: An examination of the differences in utilities for the utility function: $u(x) = (x + b)^c$. *Left:* b varies in the x-axis with c fixed at several values. *Right:* c varies in the x-axis with b fixed at several values.

Note that the difference in expected utility is negative over the entire range of b when $c = 0.1$ or $c = 0.2$, indicating that the standard treatment 0 is better than the new treatment 1. When $c = 0.3$, the difference in negative for $b \in (0, 16000)$ and positive thereafter. As the value of c gets larger, treatment 1 has higher expected utility for even relatively small values of b. The right panel shows a different view, with c changing but the value of b fixed at different levels. For smaller values of b, treatment 0 dominates treatment 1 for all values of c up to about 0.5. As b gets larger, the situation reverses, but the behavior in the right panel is more stable than that in the left panel.

We can represent this information simultaneously in the surface plot in Fig-

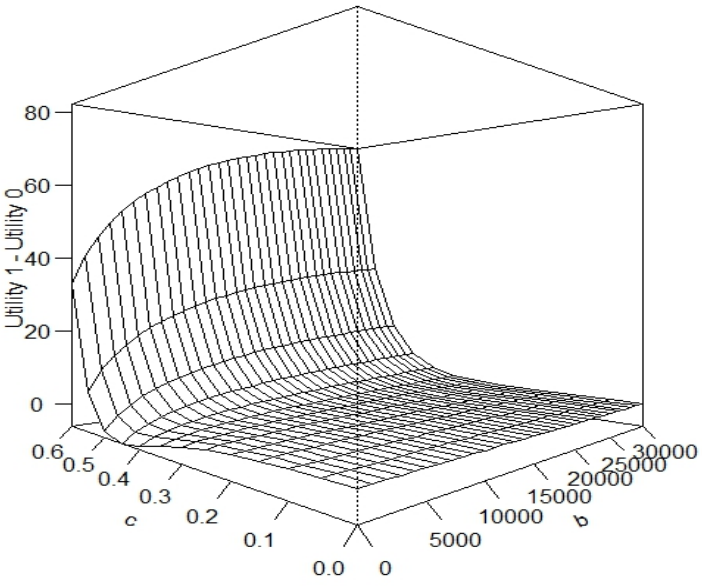

Figure 4.7: Three-dimensional surface plot of the difference in expected utilities as both b and c

ure (4.7). Here, the x- and y-axes show different values of b and c while the $z-$axis shows the difference in expected utilities. Figure (4.8) provides yet another view of the comparison of the two treatments in terms of expected utilities. The plot divides the two-dimensional b and c space into two regions: one where the new treatment 1 is preferred and the other where the standard treatment 0 is better. Such analyses are useful in examining the sensitivity of the parameters in a class of utility functions and provides the decision-maker with additional information.

Many other assumptions about utility are implicit in standard cost-effectiveness analysis, including the idea of a constant willingness-to-pay that relates cost and QALYs; the particular values chosen for quality weights; and social welfare considerations. Modeling full multiattribute utility functions and also assessing the requisite preference data from decision makers are important tasks which are separate from

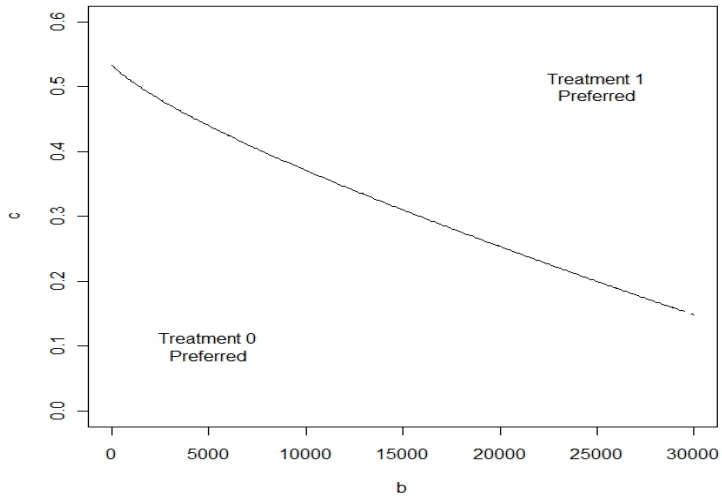

Figure 4.8: Preference Regions Relative to Utility Function Parameters.

the estimation of treatment effects. However, it is useful to consider some of the complexities of utility modeling in order to generate the appropriate results from statistical estimation (most importantly full distributions instead of mean values only).

4.4.2 Stochastic Dominance

When there is consensus about the utility functions or if the decision is robust over a range of functions, it is indeed reasonable to base decisions on expected utility criteria. More often than not, however, this is not the case. Utility functions of various individuals and groups (patients, decision makers) can differ greatly within and across groups.

The finance literature uses mean-variance ordering (Markowitz, 1952) to compare distributions. The mean-variance rule is part of portfolio theory which seeks to maximize a portfolio return for a given amount of risk, or minimize risk for a given amount of return. The rule is used to construct an efficient portfolio of investment opportunities for a risk-averse investor; an investment opportunity is considered in-

efficient (and discarded) if it has a lower mean and higher variance than another (Bawa, 1975). We can also consider the use of a similar rule in CEA to determine an efficient set of interventions. An efficient group of interventions would have relatively high expected NMB and relatively low variance. However, the mean-variance rule depends on the decision-maker having a quadratic utility function (Tobin, 1958; Borch, 1969; Feldstein, 1969).

From a health policy perspective, if a treatment increases the variance of outcomes, it is then important to try and identify the subset of the population for which the treatment helps and the subset for which it is harmful. Of course, if a policymaker is truly indifferent to how medical effects are distributed in the population, the mean-maximizing policy may be fine. In practice, however, implementing policies that overtly help one group of patients while ignoring others would be problematic and possibly unethical.

It is important to recognize at least two different sources of variability – one due to identifiable differences in patient characteristics and a second due to intrinsic variation in a homogeneous group of people. In the case of variability due to patient medical conditions, targeted treatments may be desirable: for instance, treatments that are most effective at preventing heart attacks in high-risk populations, such as those with high blood pressure. Treatments that are more effective or cost-effective along racial or gender lines, however, may be more controversial. Of course, this separation by patient characteristics can often be murky. Further, even if there are obvious patient characteristics that cause the difference, the medical intervention may not be tailored by patient groups for various reasons. Nevertheless, during the analysis stage, it is important to separate patients into homogeneous populations and analyze the populations separately. The results can then be combined at the

decision-making stage.

All of this discussion is intended to make a case for examining the entire distribution of outcomes for the standard and new treatments and determine areas where one is better than the other. There are graphical techniques as well as formal methods for comparing distributions. We focus on formal methods based on the notions of first- and second-order stochastic dominance (SD) here. The SD concepts have been discussed extensively in the finance and economics literature.

SD is a useful, and thus far largely overlooked, tool in CEA. There is brief mention applied to Net Health Benefits (equivalent to NMB) in Stinnett and Mullahy (1998), but with no consideration of the separate outputs of costs and effectiveness or uncertainty. It is also mentioned in Laska et al. (1999) but the focus of the paper is on the estimation of sample size to achieve power in statistical tests in CEA, and there is no analysis of the broader implications of SD on decision-making. There are very few papers in the more general literature about stochastic dominance for medical decisions of which we are aware. The first, Leshno and Levy (2004) deal with the advantages of SD for deciding optimality among a finite set of options. It did not deal with the question of estimation uncertainty. Finally, Sendi et al. (2003) use SD to evaluate a portfolio of health interventions. Their decision-making framework and decision rules are a bit different from those in CEA. Sendi et al. (2003) assume a fixed budget and that health programs are not divisible; in CEA, the existence of any hard upper-bound on budget is usually ignored in favor of discussion about whether an intervention is a good investment, and interventions are implicitly considered to be divisible, so that they may be implemented in whole or in part from most to least cost-effective until the budget is exhausted. As a result of the budget constraint, Sendi et al. (2003) establish that a portfolio is better than a comparison portfolio

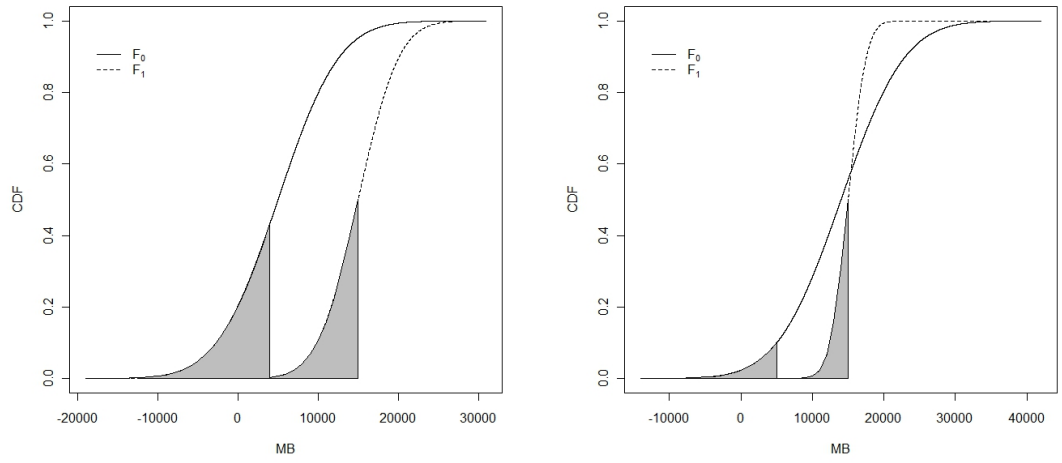

Figure 4.9: Demonstration of Stochastically Smaller Distributions. *Left:* $F_1 \succeq F_0$. *Right:* No FSD.

through dominance in effectiveness while maintaining a pre-specified probability of remaining within the budget. The appropriate decision when a first-order stochastic dominant portfolio does not exist is not discussed.

In the remainder of this section, we review the use of SD methods for CEA (assuming that the distributions are known) and then propose methods for incorporating estimation uncertainty. The latter topic has not been studied in the CEA literature.

First-Order Stochastic Dominance: FSD

Consider a general set up with two CDFs: $F_1(x)$ and $F_0(x)$ and define the difference $D(x) = F_1(x) - F_0(x)$. We say that F_1 is stochastically larger (in the first-order) than F_0 (written as $F_1 \succeq_1 F_0$) if $F_1(x) \leq F_0(x)$ or, equivalently, $D(x) \leq 0$ for all x with strict inequality for some x. A similar statement can be made about F_1 dominating F_0 in first order. In other words, the CDF of the stochastically smaller distribution will always be larger than that of the dominating distribution. Recall that $F_1(x)$ being smaller than $F_0(x)$ means that the outcomes tend to be larger under F_1 than under F_0. Perhaps a more intuitive way to describe this is in terms

survival functions $\bar{F}_j(x) = 1 - F_j(x),\ j = 0, 1$. So F_1 FSD F_0 (denoted $F_1 \succeq_1 F_0$) if $\bar{F}_1(x) \geq \bar{F}_0(x)$ for all x with strict inequality for some x. Clearly, FSD implies $E(X_1) > E(X_0)$.

The left panel of Figure (4.9) demonstrates a situation where $F_1 \succeq_1 F_0$, and the right panel demonstrates a situation where neither distribution dominates the other. In Figure (4.9), regions under the curves in each panel are highlighted; these regions will be discussed in the context of second-order stochastic dominance.

FSD is also known to have a characterization in terms of utility functions. If $F_1 \succeq_1 F_0$, then $E_1(U(X)) > E_0(U(X))$ for all increasing utility functions. (For comparing costs where smaller is better, we can take negative cost values). For this reason, an FSD comparison is the first one to make for a pair of treatments. When such dominance does not exist, then we can compare them in terms of more restricted dominance criteria or examine partial dominance (to be discussed later).

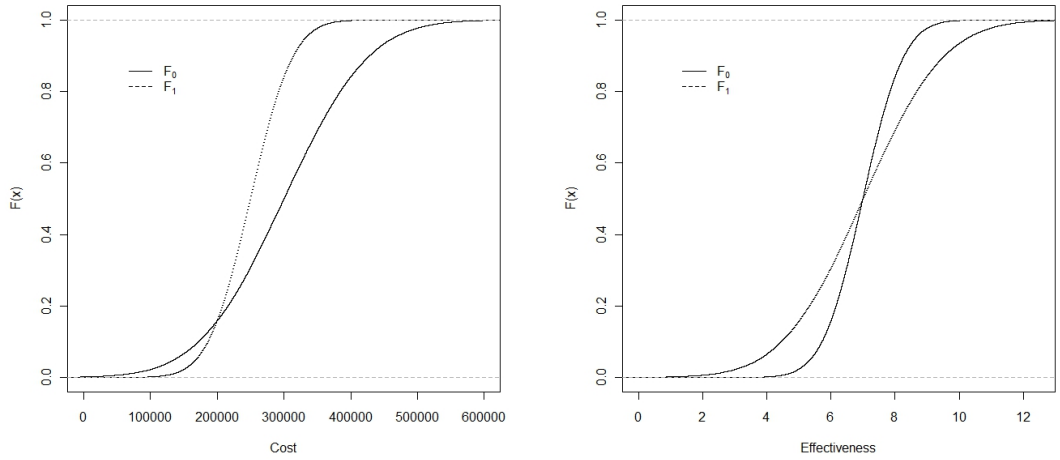

Figure 4.10: Cost and Effectiveness Cumulative Distribution Functions of Treatments with FSD in Monetary Benefit.

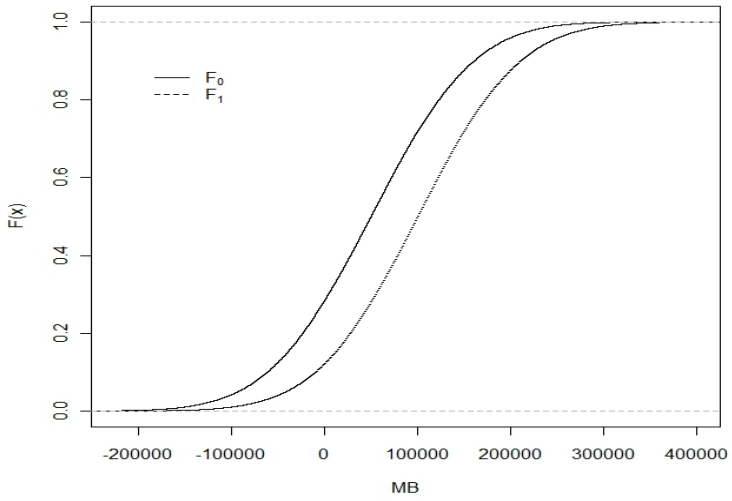

Figure 4.11: Monetary Benefits Cumulative Distribution Functions of Treatments with FSD.

A two-dimensional analysis for FSD involves comparing separately the distributions of costs and effectiveness for the two treatments. A medical decision that is FSD in both cost and effectiveness is unequivocally better and there is no tradeoff between QALYs and cost. However, most new medical treatments typically increase effectiveness and cost. The basis of CEA is that treatments with increased cost are worthwhile as long as there is adequate increases in QALYs.

To understand the trade-off, we can convert effectiveness into monetary units as in the case of NMB. Recall monetary benefit (MB) of treatment j, defined earlier as $\text{MB}_j = \lambda E_j - C_j,\ j = 0, 1$. Letting F_1 and F_0 be the distributions of MB_1 and MB_0 respectively, we can now assess SD in terms of MB. MB can be dominant even if neither costs nor QALYs of that treatment are dominant. Figures (4.10) and (4.11)show an example. The CDFs of cost and QALYs cross, so there is clearly no FSD in these. However, Figure (4.11) shows that there is FSD of treatment 1 over treatment 0 in terms of MB.

If there are multiple treatments, pairs of treatments can be evaluated against

each other, or even portfolios of treatments. In this case, rules based on stochastic dominance are used to discard non-optimal choices. In general, the efficient set of medical choices are those that are not dominated by any other option. If the number of options is large, it may be computationally demanding to do pairwise comparisons for stochastic dominance between all distributions. There exist algorithms for reducing the number of pairwise comparisons and finding the efficient set of options, also referred to as the admissible set (Porter et al., 1973; Bawa, 1975; Bawa et al., 1979). Regardless of the method used to determine the efficient set of options, the result may be a large efficient set of options, as the framework may be unable to rank two risk options (Levy and Hanoch, 1970). We will primarily focus on the pairwise comparison of two interventions in this paper, but simple extensions are available to the construction of an efficient set of interventions.

Second-Order Stochastic Dominance

FSD is a fairly stringent condition, and it may not hold in many situations. So researchers have developed weaker SD criteria called second- and third- order SD that also have interpretations in terms of utility functions. We will focus just on second order SD here.

Define $G_i(x) = \int_{-\infty}^{x} F_i(y)dy$ for $i = 0, 1$. That is, $G_i(x)$ is the area under the curve of F_i up to the point x. Define

$$L(x) = G_1(x) - G_0(x) = \int_{-\infty}^{x} D(y)dy = \int_{-\infty}^{x} (F_1(y) - F_0(y))dy. \tag{4.15}$$

We say that F_1 dominates F_0 in the second order sense ($F_1 \succeq_2 F_0$) if $L(x) \leq 0$ for all x with strict inequality for some x. Figure (4.12) show the areas under the CDF curves for $F_0(x)$ and $F_1(x)$ for two different scenarios. The value of the line representing F_0 in the left panel in Figure (4.12) at NMB= 30000 corresponds to the

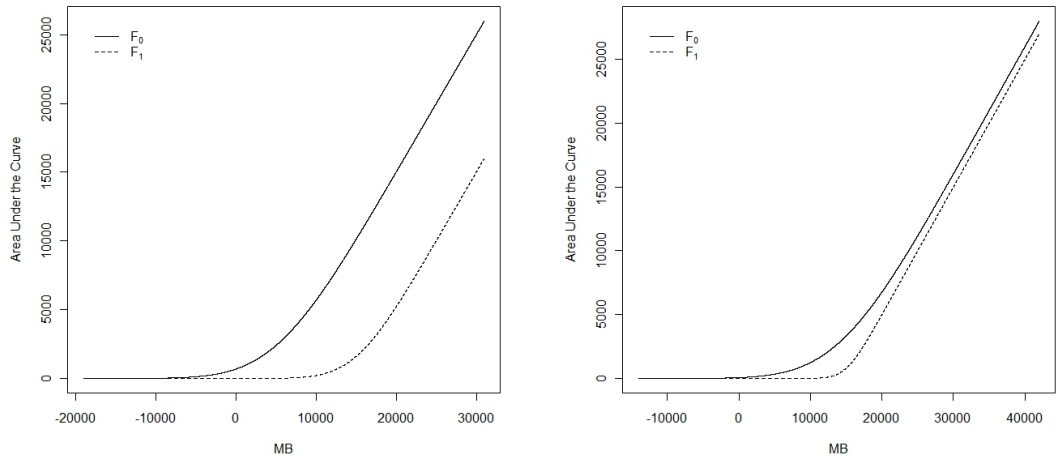

Figure 4.12: Area Under the Curves. *Left and Right:* $F_1 \succeq_2 F_0$

area shaded under the curve of F_0 in the left panel in Figure (4.9). In the left panel, $F_1 \succeq_1 F_0$, so $F_1 \succeq_2 F_0$ by definition. In the right panel $F_1 \succeq_2 F_0$, but Figure (4.9) shows that there is no FSD.

It is known that if $F_1 \succeq_2 F_0$, then $E_1(U(X)) > E_0(U(X))$ for decision makers who are risk-averse (which is to say they have increasing and concave utility functions). More specifically, for any utility function that is differentiable, increasing and concave, $E_1(U(X)) > E_0(U(X))$. Thus, the SSD criterion orders distributions within a more restricted class of utility functions than FSD. It can be shown that $L(\infty) = \mu_1 - \mu_0$, so SSD also implies $\mu_1 > \mu_0$.

We can again apply the SSD comparisons separately to the distributions of cost and effectiveness or do a reduced comparison in terms of MB_1 and MB_0 as discussed in the case of FSD. A reduced comparison would suffice for a risk-averse decision maker who cares only about cost-effectiveness, whereas the separate comparisons are more informative for decision-makers with specific effectiveness goals or cost constraints.

Almost Stochastic Dominance

It is quite likely that there is no strict dominance of the one treatment over another, in terms of cost and effectiveness, or even the combined measure of MB. In this case, there may be ethical issues concerned with the choice of treatment due to its differential effect in different regions and whether it helps some segment of the population and harms others. The question of costs and ethics in healthcare is more complicated due to the many different methods of insurance and payment in the United States healthcare system.

Suppose there is no FSD and let

$$S_1(F_1, F_0) = \{x \in (0, \infty) : F_0(x) < F_1(x)\}, \tag{4.16}$$

the region where $F_0(x) < F_1(x)$. Recall that no FSD implies that the CDFs cross at one or more points or that $D(x) = 0$ for some values of $x \in (\infty, \infty)$. If the crossings occur at only one point, say x_0, then the region will be an interval of the form $(-\infty, x_0)$ or (x_0, ∞); otherwise it will be a union of multiple intervals. Note that the probability of the area S_1 is the same under F_0 and F_1 since they cross at the end-points of the intervals that form the regions S_1. For example, consider Figure (4.13). $D(x) = 0$ at $x = a_1$, $x = a_2$, and $x = a_3$, so $P(a_1 < X_0 < a_2) = F_0(a_2) - F_0(a_1)$, equals $P(a_1 < X_1 < a_2) = F_1(a_2) - F_1(a_1)$.

In any case, S_1 provides information about the segments of the population where F_0 is better than F_1. Suppose $F_1 \succeq_1 F_0$ for the most part, and the probability of the regions where $F_0(x) < F_1(x)$ is small. Then, one possible notion of almost or $\epsilon-$FSD is that the total probability of the regions where FSD is violated is smaller than some threshold ϵ.

One problem with this notion is the difference $F_0(x) - F_1(x)$ can be big in this

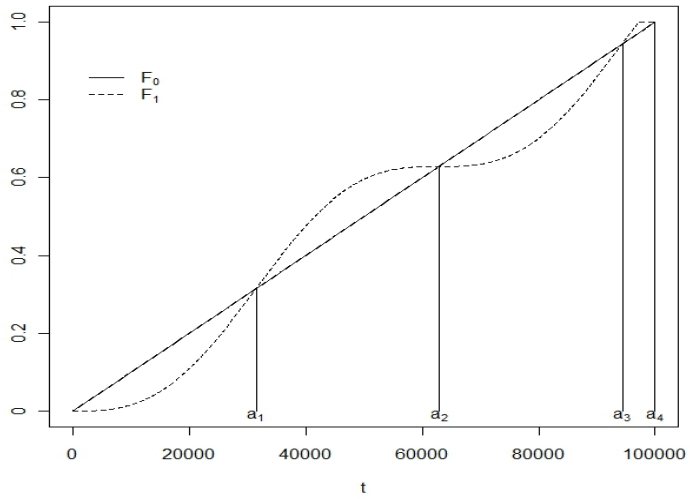

Figure 4.13: Conditional Dominance.

region even though the probability of the region is small. To get around this problem, Leshno and Levy (2002) use an alternative notion of almost stochastic dominance that depends on the area under the curve. Denote the integrated absolute difference between F_1 and F_0 as $||F_1 - F_0|| = \int_{-\infty}^{\infty} |F_1(x) - F_0(x)|dx$. This quantity exists if both F_1 and F_0 have finite absolute first moments. It can be viewed as one measure of the difference between the two distributions. Then, Leshno and Levy (2002) say that F_1 almost dominates F_0 in FSD (AFSD) if

$$\int_{S_1} [F_1(x) - F_0(x)]dx \leq \varepsilon ||F_1 - F_0||, \tag{4.17}$$

with $\varepsilon \in (0, 0.5)$. In other words, the difference in the area under the curve where $F_0(x) < F_1(x)$ is relatively small compared to the total area. Leshno and Levy (2002) show that, if F_1 dominates F_0 in terms of AFSD, then F_1 is preferred over F_0 for all utility functions that are increasing but have a bounded derivative, where the bound depends on ε. As $\epsilon \to 0$, this class approaches the class of all increasing utility functions.

Leshno and Levy (2002) also define ASSD similarly. Let $S_2(F_1, F_0) = \{x \in$

Table 4.1: Almost Stochastic Dominance Example Probability Distribution.

Rate of Return: z	5%	7%	9%	12%
$P(X = z)$	0.1	0.0	0.0	0.9
$P(Y = z)$	0.0	0.4	0.6	0.0

$S_1(F_1, F_0) : \int_{-\infty}^{x} F_0(y)dy < \int_{-\infty}^{x} F_1(y)dy\}$, the region where the SSD dominance of F_1 over F_0 fails. Then, Leshno and Levy (2002) define almost SSD as follows. F_1 almost dominates F_0 in SSD (ASSD) if

$$\int_{S_2} [F_1(x) - F_0(x)]dx \leq \varepsilon ||F_1 - F_0|| \tag{4.18}$$

and $\mu_1 \geq \mu_0$. As before, $\varepsilon \in (0, 0.5)$. Leshno and Levy (2002) show that, if F_1 dominates F_0 in terms of ASSD, then F_1 is preferred over F_0 for all utility functions that are increasing and concave with bounded second derivative, where the bound depends on ε. Again, as $\epsilon \to 0$, this class approaches the class of all increasing and concave utility functions.

As shown in Leshno and Levy (2002), there are many situations where one would prefer F_1 over F_2 in terms of AFSD or ASSD (and for most reasonable utility functions in their respective classes) but there is no (absolute) FSD or SSD. One example in Leshno and Levy (2002) is a comparison of the utility of stocks (X) and bonds (Y). The distribution of the one-year rate of return is given in Table 4.1. Consider $X^{(n)}$ to be the distribution of X after n years, and $Y^{(n)}$ to be the distribution of Y after n years. The CDF of $X^{(n)}$ starts to the left of the CDF of $Y^{(n)}$, so FSD of $X^{(n)}$ over $Y^{(n)}$ does not formally exist, no matter how large n is. However, Leshno and Levy (2002) show that as n increases the CDF of X shifts right faster than the CDF of Y, and that nearly all utility functions will prefer $X^{(n)}$ to $Y^{(n)}$ for large n. The concepts of AFSD and ASSD are definitely applicable to CEA.

Assessing First-Order Stochastic Dominance with Statistical Uncertainty

As before, we use a generic two-sample set up to describe the procedures. Let $\{X_1, ..., X_{n_0}\}$ be the data from one treatment and $\{Y_1, ..., Y_{n_1}\}$ be the data from a second treatment, and $N = n_0 + n_1$. Let $\hat{F}_j(x)$ be the empirical CDF based on the data, the well-known estimator of the unknown CDF $F_j(x)$, $j = 0, 1$. Let $\hat{D}(x) = \hat{F}_1(x) - \hat{F}_0(x)$. With the data being estimated, methods are needed to infer from the sample whether FSD exists in the population.

One possibility is using methods to construct a confidence region for $D(x) = F_1(x) - F_0(x)$ and using this to assess FSD. Since $D(x)$ is a function, the confidence region has to be valid for all (or most) values of x, i.e., a simultaneous confidence region. Perhaps the simplest method to construct such a confidence region is through the use of the two-sample Kolmogorov-Smirnov (KS) statistic or its weighted versions. We will consider first the unweighted KS statistic. Let

$$K = \sup_{-\infty<x<\infty} |\hat{D}(x) - D(x)| = \sup_{-\infty<x<\infty} |(\hat{F}_1(x) - F_1(x)) - (\hat{F}_0(x) - F_0(x))|. \tag{4.19}$$

It is well known that the distribution of K under the null hypothesis that $F_1(x) = F_0(x)$ (and the distribution is continuous) is distribution-free, i.e., it does not depend on the underlying distribution. This is true even in finite samples. The null distribution of K has been well studied and is available in most software packages. Let $K_\alpha = K_\alpha(n_0, n_1)$ be the $(1-\alpha)-$th quantile of the distribution of K with sample sizes n_0 and n_1. From this and equation 4.19 above, we have

$$P\left(\sup_{-\infty<x<\infty} |\hat{D}(x) - D(x)| < K_\alpha\right) \geq 1 - \alpha. \tag{4.20}$$

In other words,

$$\hat{D}(x) \pm K_\alpha \tag{4.21}$$

is a two-sided confidence region for $D(x)$ for all x – i.e., it is a simultaneous confidence region. We can now use this band to assess FSD as follows. If the upper bound of this region is below zero for all x, we can conclude that $D(x) \leq 0$ for all x even after taking the statistical uncertainty into account. This implies that $F_1 \succeq_1 F_0$. Conversely, if the lower bound is strictly above zero, it implies that $F_0 \succeq_1 F_1$.

The two-sided confidence region treats F_1 and F_0 symmetrically. In CEA, treatment 1 is new and we may be interested only in assessing if $F_1 \succeq_1 F_0$ or not. In that case, we can just construct a one-sided upper bound for the confidence region. If the upper bound is below zero for all x, we conclude that $F_1 \succeq_1 F_0$.

One disadvantage of the $K-$band above is that the width of the band $\pm K_\alpha$ is constant for all x. This is not desirable as the variance of $\hat{D}(x)$ tends to zero in both the lower and upper tails and is highest in the middle regions. To account for this, we can use a weighted version of the $K-$ statistic of the form

$$W = \sup_{\{x:a<x<b\}} |\hat{D}(x) - D(x)|/\widehat{SE}(\hat{D}(x)). \tag{4.22}$$

Recall that

$$SE(\hat{D}(x)) = \sqrt{F_1(x)(1-F_1(x))/n_1 + F_0(x)(1-F_0(x))/n_0},$$

which can be estimated by plugging in the corresponding empirical CDF's. The range of the statistic W has to be restricted to $\{a < x < b\}$ as the variance tends to zero in the lower and upper tails and dividing by the standard deviation can blow up the value of the statistic in the tails. In practice, one can take a and b to the lower and upper percentiles of the data, corresponding to some level such as .01 or .05. With this choice, the distribution of the statistic is also distribution-free under the null-hypothesis. Using this test statistic, we get the following band, sometimes

called the W-band (Doksum and Sievers, 1976):

$$\hat{D}(x) \pm W_\alpha \sqrt{\hat{F}_1(x)(1-\hat{F}_1(x))/n_1 + \hat{F}_0(x)(1-\hat{F}_0(x))/n_0}. \tag{4.23}$$

Note that the width of this band decreases as one moves to the tails, and it will be smaller in the lower and upper tails compared to the K-band.

We use a new example to demonstrate the use of these bands for FSD, and continue this example into the discussion of SSD also. The data are given in Appendix 4.6.2; as with the previous example, we suppose that the data has been observed at 32 points for each of the treatments. Figure (4.14) shows the empirical distribution functions of the monetary benefit for treatments 1 and 0 in our illustrative example; the distributions do not cross, so we cannot rule out FSD. Figure (4.15) shows the two-sided confidence band for the monetary benefit, using both the weighted and unweighted statistics. It shows that the K-band has constant width while the W-band is narrower in the tails. In both cases, the upper band is not completely below zero, so we cannot conclude that $F_1 \succeq_1 F_0$. For the true distributions, shown in Appendix 4.6.2, we do have FSD but the statistical uncertainty from estimating the distributions masks this. If the dominance is stronger or if the sample size is much larger, we would have been able to determine the truth.

Assessing Second-Order Stochastic Dominance with Statistical Uncertainty

There has been less attention in the statistical literature on assessing SSD from sample data, but some work has been done in economics (see McFadden (1989); Barrett and Donald (2003)). Also, the problem is more complex as the corresponding K and W statistics are no longer distribution-free; i.e., the distributions of the test statistics depend on the underlying distribution even if $F_1 = F_0$. One would have to compute the distribution for each case separately, but the distributions are unknown,

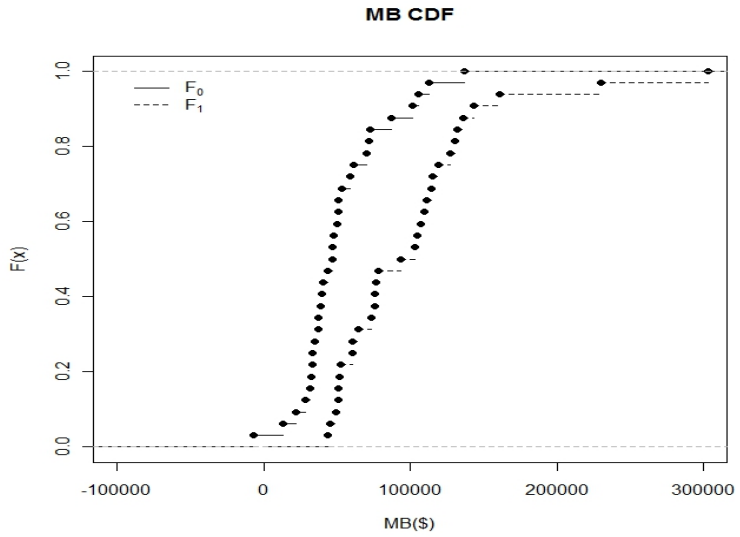

Figure 4.14: Empirical Cumulative Distribution Function.

making the problem difficult analytically. However, resampling methods can be used to solve this problems.

As in the FSD case, we use a generic two-sample set up to describe the procedures. Let $\{X_1, ..., X_{n_0}\}$ be the data from one treatment and $\{Y_1, ..., Y_{n_1}\}$ be the data from a second treatment, and $N = n_0 + n_1$. Let $\hat{G}_j(x)$ be the empirical estimator of the unknown $G_j(x) = \int_{-\infty}^{x}(F_1(y) - F_0(y))dy$, $j = 0, 1$. Let $\hat{L}(x) = \hat{G}_1(x) - \hat{G}_0(x)$.

We start with a discussion of methods for constructing a simultaneous confidence region for $L(x) = G_1(x) - G_0(x)$, As with the FSD case, we use weighted and unweighted statistics. Let

$$M = \sup_{-\infty<x<\infty} |\hat{L}(x)| = \sup_{-\infty<x<\infty} |(\hat{G}_1(x) - \hat{G}_0(x))|. \tag{4.24}$$

As we have noted, the distribution of M depends on the common (but unknown) distribution under the null hypothesis that $F_1(x) = F_0(x)$. As discussed in Barrett and Donald (2003), we will use bootstrap techniques. Specifically, a large number of bootstrap replications of $\hat{F}_1(x)$ and $\hat{F}_0(x)$ are created. Denoting them as $\hat{F}_1^*(x)$ and $\hat{F}_0^*(x)$, we compute $\hat{G}_1^*(x)$ and $\hat{G}_0^*(x)$ and $M^* = \sup_{-\infty<x<\infty} |(\hat{G}_1^*(x) - \hat{G}_0^*(x))|$.

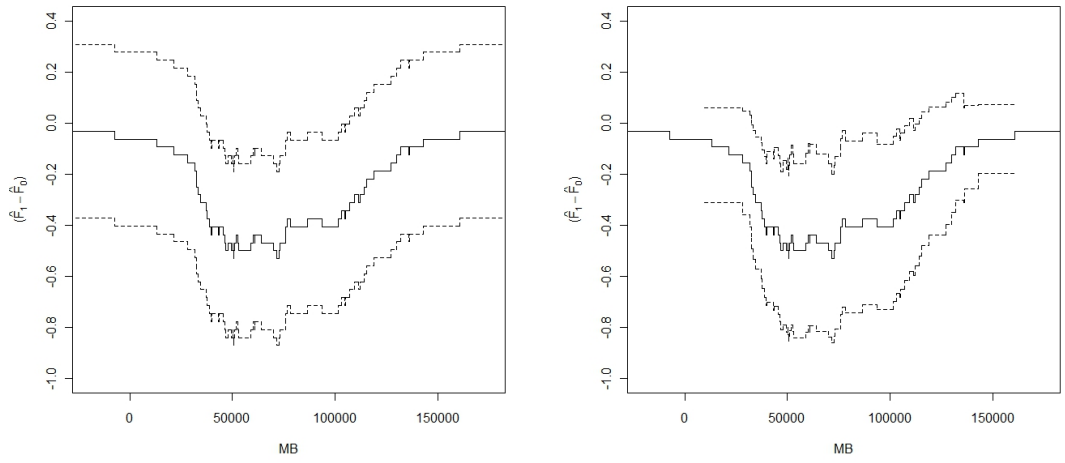

Figure 4.15: Left: $\hat{D}(x)$ and $K-$ band. Right: $\hat{D}(x)$ and $W-$ band.

Let M^*_α be the $(1-\alpha)-$th quantile of the distribution of M^*. Using this we get the unweighted two-sided band as

$$\hat{L}(x) \pm M^*_\alpha. \tag{4.25}$$

We also consider an analogous weighted statistic, as in the FSD case:

$$U = \sup_{\{x:a<x<b\}} |\hat{L}(x) - L(x)|/\widehat{SE}(\hat{L}(x)), \tag{4.26}$$

which leads to the weighted confidence band:

$$\hat{L}(x) \pm U_\alpha \widehat{SE}(\hat{L}(x)). \tag{4.27}$$

Again the critical value will U_α will have to be obtained via bootstrap. But there is an additional complication here in that $\widehat{SE}(\hat{L}(x))$ can be computed only after all the bootstrap samples are available, while it is needed for each bootstrap computation. Similar problems have been handled by the double-bootstrap in the statistical literature Efron and Tibshirani (1993). This involves running another bootstrap within each bootstrap computation to estimate $\widehat{SE}(\hat{L}(x))$ and then computing the weighted statistic.

The left and right panels of Figure (4.16) show the $M-$ or $U-$ bands. Neither one provides support for SSD of F_1 over F_0. The bands are rather wide, indicating the large variability of the estimated $\hat{G}_j$'s. The main conclusion to be made here is that establishing SSD in the presence of statistical uncertainty is difficult and requires very large sample size.

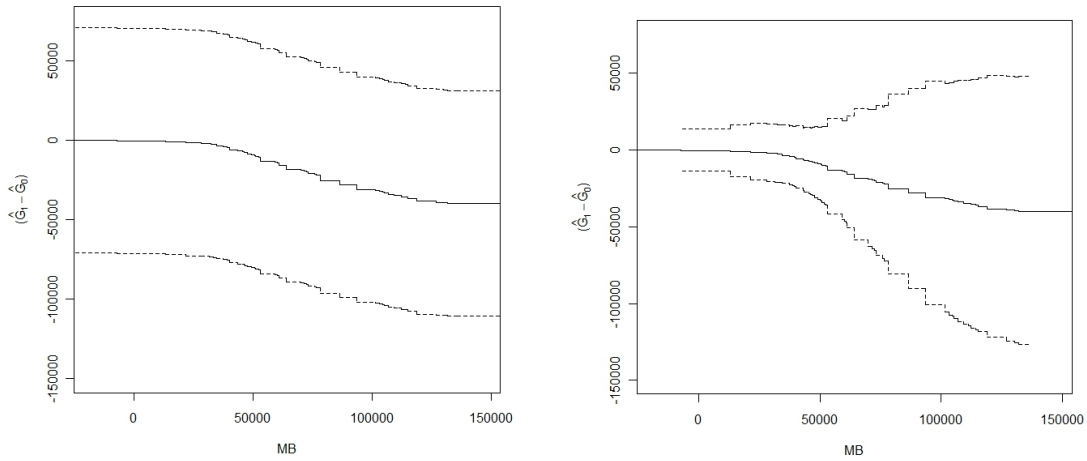

Figure 4.16: *Left:* $M-$ band for Assessing SSD. *Right*: $U-$ band for Assessing SSD.

4.5 Conclusion

We have reviewed metrics for quantifying treatment effect in CEA and methods for making comparisons in the presence of statistical uncertainty. We have tried to make a case for going beyond mean-based analysis to examine the entire distributions and using concepts of stochastic dominance. We also discussed comparisons based on utility functions. Because the utility functions of decision-makers are largely unknown, we propose sensitivity analysis about the parameters of the utility function. The procedures we examined for determining SSD in the presence of estimation uncertainty have very low power, and additional research on alternative methods is needed.

4.6 Appendix

4.6.1 Illustrative Data

This appendix contains the data used for the examples about statistical inference (except for stochastic dominance) in this paper. Assume the true underlying distribution of $(E_0, C_0) \sim LN\left(\begin{pmatrix} 0.3 \\ 7.2 \end{pmatrix}, \begin{pmatrix} 0.4, 1 \\ 1, 7.3 \end{pmatrix}\right)$. Likewise, assume the true underlying distribution of $(E_1, C_1) \sim LN\left(\begin{pmatrix} 0.8 \\ 7.2 \end{pmatrix}, \begin{pmatrix} 0.25, 1 \\ 1, 7 \end{pmatrix}\right)$. Only $n_0 = n_1 = 32$ samples are available from each population. The summary statistics for these samples are given in Table 4.2.

Table 4.2: Illustrative Example Summary Statistics

Treatment	$\hat{\mu}(E)$	$\hat{\mu}(C)$	$v(E)$	$v(C)$	cv
0	1.30	17,916	0.017	182,592,614	1400
1	3.09	3,180	0.124	1,301,030,431	7879

The full sample data is in Table 4.3.

Table 4.3: Illustrative Example Data

Sample	$\mathbf{E_0}$	$\mathbf{C_0}$	$\mathbf{E_1}$	$\mathbf{C_1}$
1	0.92	163	3.05	41628
2	1.03	496	0.92	10
3	1.47	24715	4.93	19870
4	1.22	8164	2.21	6387
5	1.22	197	2.76	6028
6	0.83	84	1.59	3001
7	0.89	24	1.17	108
8	0.54	25	3.17	13130
9	1.65	12264	1.17	21
10	1.76	9049	1.52	3499
11	0.85	222	0.83	666
12	1.07	1088	1.06	11
13	1.13	222	1.52	92
14	1.14	8748	7.82	344820
15	0.63	1	2.59	256
16	1.46	6774	3.75	39273
17	2.44	29236	1.57	395
18	1.06	400	1.29	38
19	1.96	1415	2.43	6013
20	1.48	7731	3.06	22258
21	0.59	96	7.32	1103670
22	1.33	1013	3.93	22566
23	1.91	535	3.20	110
24	4.37	434992	7.83	271106
25	2.04	13612	1.14	215
26	1.34	4718	5.01	29924
27	0.68	142	1.34	164
28	0.74	165	3.13	25471
29	1.45	4878	2.84	2170
30	0.83	1197	5.20	14646
31	0.75	475	4.36	13537
32	0.94	474	5.00	30678

4.6.2 Stochastic Dominance Section Illustrative Data

This appendix contains the data used to demonstrate the use of confidence bands for FSD and SSD. Assume the true underlying distribution of $(E_0, C_0) \sim LN\left(\begin{pmatrix} 0.3 \\ 7.2 \end{pmatrix}, \begin{pmatrix} 0.25, 1 \\ 1, 7. \end{pmatrix}\right)$. Likewise, assume the true underlying distribution of $(E_1, C_1) \sim LN\left(\begin{pmatrix} 0.8 \\ 7.2 \end{pmatrix}, \begin{pmatrix} 0.25, 1 \\ 1, 7 \end{pmatrix}\right)$. Only $n_0 = n_1 = 32$ samples are available from each population. The true distribution functions of these distributions are shown in Figure 4.17. It is clear from this that there is FSD at a population level, but the work in this paper shows that the uncertainty is too high to conclude there is FSD in the case of the 32-patient samples.

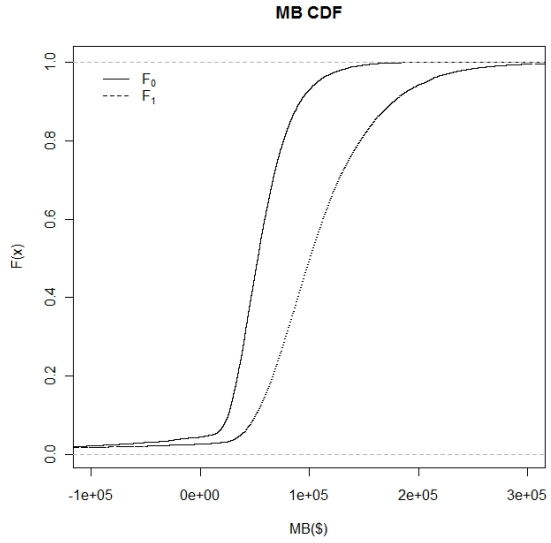

Figure 4.17: True Distribution of Stochastic Dominance Illustrative Data.

Table 4.4 contains the sample data used for the stochastic dominance examples.

Table 4.4: SSD Illustrative Example Data

Sample	$\mathbf{E_0}$	$\mathbf{C_0}$	$\mathbf{E_1}$	$\mathbf{C_1}$
1	1.01	38	1.29	498
2	1.19	8780	0.99	15
3	2.69	22391	1.55	727
4	0.67	114	2.45	3681
5	0.78	7	2.49	63999
6	3.14	20697	2.08	53816
7	0.69	53	1.56	2219
8	1.36	28034	1.56	118
9	2.27	12041	2.61	340
10	0.65	51	2.19	127
11	1.46	3280	2.39	4480
12	1.06	10	6.42	18387
13	0.63	13	3.45	12200
14	0.77	893	4.76	101872
15	0.96	272	1.52	111
16	2.88	71579	2.43	7133
17	1.32	6786	4.67	3982
18	2.11	286	1.20	57
19	0.56	26	2.88	1131
20	1.06	3009	1.05	19
21	0.79	158	1.87	139
22	0.26	1	1.02	3
23	2.68	112409	2.07	245
24	1.40	77431	2.55	317
25	1.33	5469	0.91	13
26	0.75	256	2.65	815
27	0.93	349	2.28	2604
28	0.94	191	1.49	1269
29	1.49	2730	2.21	3748
30	0.87	111	0.87	105
31	0.66	141	1.04	24
32	1.85	5711	2.10	510

CHAPTER V

Conclusion and Future Work

5.1 Summary and Contributions

Cost-effectiveness analysis will increase in importance as healthcare becomes more expensive and resources become more constrained. CEA is interdisciplinary, drawing on knowledge from medical doctors and epidemiologists, and increasingly from statisticians and operations research professionals. This thesis contributes to the CEA literature by developing methods of evaluation and analysis of cost-effectiveness data. The following are key components of the contribution.

1. Expressions were developed to evaluate stationary, progressive multi-state models of disease used for CEA. With the assumption of independence between transition and sojourn distributions, as is made in the case of Markov and semi-Markov models common in the literature, the expressions can be easily evaluated analytically. In the case of non-stationary processes, which have a dependence between transitions, the expressions can still be evaluated through a straightforward simulation. This approach has the advantage over traditional DES of evaluating and weighting all possible patient paths, similar to stratified sampling in population studies.

2. The stochastic process underlying the analysis of disease progression was made

explicit in Chapter III. We demonstrated the usefulness of using graphical methods to analyze the simulation data more extensively and to develop key insights into the performance of treatments.

3. A review of CEA metrics and methods for incorporating statistical uncertainty was presented, including a clarification of confusion in the CEA literature regarding the use of large-sample theory in determining confidence intervals. We also proposed the use of rank-based methods in CEA, and discussed the determination of uncertainty in the case of censoring. We expanded the discussion of CEA to comparisons of entire distributions of rewards. We proposed two methods for evaluating these distributions: explicit analysis using utility functions and first- and second-order stochastic dominance, for the cases when the decision-maker utility function is unknown. New contributions to the CEA literature in this chapter include stochastic dominance comparisons in the presence of estimation uncertainty, use of rank methods, and analysis with censored data.

5.2 Limitations and Future Research

There are several directions for future research related to this dissertation. The main limitation of this research is that the examples used in all three papers are illustrative, and are not based on actual clinical data. This suits the purposes of methodology development well, but the ideas here should be applied and proved using actual data, and the availability of such data can be limited. In particular, the methods Chapter IV should be applied alongside the design and execution of a clinical or observational trial, and an original simulation study should be conducted using the methods in Chapter II and III.

Additionally, here are several comments, specific to each of the papers, about

opportunities for future research.

The expressions in Chapter II are limited to the case of progressive models of disease. While in general, many CEA models are progressive or can be designed that way, many existing models are not progressive, and analysts will continue to create non-progressive models. The development of methods to handle non-progressive models is a challenging problem; existing expressions for mean rewards in the rewards literature are recursive (Howard, 1971; Janssen and Manca, 2006), due to the infinite number of paths through a non-progressive multi-state model. These recursive calculations are tedious in many cases, and not adopted to CEA. The development of algorithms and non-DES simulation methods specific to CEA of more general (non-progressive) evaluation of random and expected rewards would be of use to the CEA field, and novel in the CEA literature.

The analysis in Chapter III, while useful, is more time-consuming and requires more knowledge of statistical software than many of the end-users of the information, and even the analysts who create simulation models, may have. The development of software to conduct the analysis, including plotting densities and quantiles, and stratifying on covariates, would be an advancement in the use of this analysis. We created prototype software as a proof-of-concept, but significantly more information about end-user requirements and computer science knowledge is needed to create such software. It would also be interesting to use existing simulation data and apply the graphical techniques in the paper to see which prove the most useful across many studies, and from that develop a recommended order of analysis, or even a prescriptive decision-tree directing analysis for those less familiar with analyzing data.

For the research presented in Chapter IV, further work is needed in assessing

stochastic dominance under uncertainty. This dissertation used only KS-type statistics, and these often have limited power, which was obvious in the very wide confidence limits, particulary for second-order stochastic dominance. Particularly challenging is the fact that it would be desirable to develop statistical tests for dominance in which the null hypothesis is non-dominance and the alternative is dominance; this makes it difficult to use many of the L_2 tests standard in the literature, where L_2 tests are based on measures of the squared integrated distance between the empirical CDFs (for instance, Cramer-von Mises and Anderson-Darling tests).

Bibliography

Aalen, O. O. (1995). Phase type distributions in survival analysis. *Scandinavian Journal of Statistics*, 22(4):447–463.

Barrett, G. F. and Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, 71(1):71–104.

Basu, A. (2009). Individualization at the heart of comparative effectiveness research: The time for i-cer has come. *Med Decis Making*, 29(6):NP9–11.

Bawa, V. S. (1975). Optimal rules for ordering uncertain prospects. *Journal of Financial Economics*, 2(1):95–121.

Bawa, V. S., Lindenberg, E. B., and Rafsky, L. C. (1979). An efficient algorithm to determine stochastic dominance admissible sets. *Management Science*, 25(7):609–622.

Beck, J. R. and Pauker, S. G. (1983). The Markov process in medical prognosis. *Med Decis Making*, 3(4):419–458.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. Pearson Prentice Hall, Upper Saddle River, N.J.

Birch, S. and Gafni, A. (1992). Cost effectiveness/utility analyses : Do current

decision rules lead us to where we want to be? *Journal of Health Economics*, 11(3):279–296.

Birch, S. and Gafni, A. (2004). The 'NICE' approach to technology assessment: An economics perspective. *Health Care Management Science*, 7(1):35.

Borch, K. (1969). A note on uncertainty and indifference curves. *The Review of Economic Studies*, 36(1):1–4.

Briggs, A. and Fenn, P. (1998). Confidence intervals or surfaces? uncertainty on the cost-effectiveness plane. *Health Economics*, 7(8):723–740.

Briggs, A. and Gray, A. (1998). The distribution of health care costs and their statistical analysis for economic evaluation. *Journal of Health Services Research and Policy*, 3(4):233.

Briggs, A., Mihaylova, B., Sculpher, M., Hall, A., Wolstenholme, J., Simoons, M., Deckers, J., Ferrari, R., Remme, W. J., Bertrand, M., and Fox, K. (2007). Cost effectiveness of perindopril in reducing cardiovascular events in patients with stable coronary artery disease using data from the europa study.(report). *Heart*, 93(9):1081(6).

Castelli, C., Combescure, C., Foucher, Y., and Duares, J.-P. (2007). Cost-effectiveness analysis in colorectal cancer using a semi-Markov model. *Statistics in Medicine*, 26(30):5557–5571.

Chambers, J. M. (1983). *Graphical Methods for Data Analysis*. Wadsworth statistics/probability series. Wadsworth International Group ; Duxbury Press, Belmont, Calif. : Boston.

Chaudhary, M. A. and Stearns, S. C. (1996). Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in Medicine*, 15(13):1447–1458.

Cleveland, W. S. (1993). *Visualizing Data*. Hobart Press, Murray Hill, N.J. : Summit, N.J.

DeFauw, M., Nair, V. N., and Yang, Y. (2011). Cost-effectiveness analysis in health policy decision making: Direct methods for multi-state processes. *Technical Report; University of Michigan Department of Statistics*, 524.

Doksum, K. A. and Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63(3):421–434.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Feldstein, M. S. (1969). Mean-variance analysis in the theory of liquidity preference and portfolio selection. *The Review of Economic Studies*, 36(1):5–12.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):175–185.

Garber, A. M. and Phelps, C. E. (1997). Economic foundations of cost-effectiveness analysis. *Journal of Health Economics*, 16(1):1–31.

Gardiner, J. C., Huebner, M., Jetton, J., and Bradley, C. J. (2000). Power and sample assessments for tests of hypotheses on cost-effectiveness ratios. *Health Economics*, 9(3):227–234.

Gardiner, J. C., Luo, Z., Bradley, C. J., Sirbu, C. M., and Given, C. W. (2006). A dynamic model for estimating changes in health status and costs. *Statistics in Medicine*, 25(21):3648–3667.

Gerstein, H. C., Yusuf, S., Mann, J. F. E., Hoogwerf, B., and et al. (2000). Effects of ramipril on cardiovascular and microvascular outcomes in people with diabetes mellitus: Results of the HOPE study and MICRO-HOPE substudy. *The Lancet*, 355(9200):253.

Givens, G. H. and Hoeting, J. A. (2005). *Computational Statistics*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J.

Gold, M. R., Siegel, J. E., Russell, L. B., and Weinstein, M. C. (1996). *Cost-effectiveness in health and medicine: report to the U.S. Public Health Service*. U.S. Dept. of Health and Human Services Office of Public Health and Science Office of Disease Prevention and Health Promotion : U.S. G.P.O., Washington, D.C.

Halpern, E. F., Weinstein, M. C., Hunink, M. G. M., and Gazelle, G. S. (2000). Representing both first- and second-order uncertainties by Monte Carlo simulation for groups of patients. *Med Decis Making*, 20(3):314–322.

Hayward, R. A., Krumholz, H. M., Zulman, D. M., Timbie, J. W., and Vijan, S. (2010). Optimizing statin treatment for primary prevention of coronary artery disease. *Annals of internal medicine*, 152(2):69–77.

Hill, S. R., Mitchell, A. S., and Henry, D. A. (2000). Problems with the interpretation of pharmacoeconomic analyses: A review of submissions to the Australian pharmaceutical benefits scheme. *JAMA*, 283(16):2116–2121.

Hjelmgren, J., Berggren, F., and Andersson, F. (2001). Health economic guidelines; similarities, differences and some implications. *Value in Health*, 4(3):225–250.

Howard, R. A. (1971). *Dynamic probabilistic systems*. Series in decision and control. Wiley, New York.

Hunink, M. G. M., Bult, J. R., de Vries, J., and Weinstein, M. C. (1998). Uncertainty in decision models analyzing cost-effectiveness. *Medical Decision Making*, 18(3):337–346.

Jacobson, G. A. (2007). Comparative clinical effectiveness and cost-effectiveness research: Background, history, and overview. *Congressional Research Service*.

Janssen, J. and Manca, R. (2006). *Applied semi-Markov Processes*. Springer Science+Business Media, New York, NY.

Janssen, J. and Manca, R. (2007). *Semi-Markov Risk Models for Finance, Insurance and Reliability*. Springer Science+Business Media, LLC, Boston, MA.

Jiang, H. and Zhou, X.-H. (2004). Bootstrap confidence intervals for medical costs with censored observations. *Statistics in Medicine*, 23(21):3365–3376.

Keeler, E. B. and Cretin, S. (1983). Discounting of life-saving and other nonmonetary effects. *Management Science*, 29(3):300–306.

Keeney, R. L. and Raiffa, H. (1976). *Decisions with Multiple Objectives : Preferences and Value Tradeoffs*. Wiley series in probability and mathematical statistics. Wiley, New York.

Kennedy, W. J. and Gentle, J. E. (1980). *Statistical Computing*. Statistics, textbooks and monographs.v.33. M. Dekker, New York.

Kent, D. M., Hayward, R. A., Griffith, J. L., Vijan, S., Beshansky, J. R., Califf, R. M., and Selker, H. P. (2002). An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. *The American Journal of Medicine*, 113(2):104–111.

Kinlay, S., Leitch, J. W., Neil, A., Chapman, B. L., Hardy, D. B., and Fletcher, P. J. (1996). Cardiac event recorders yield more diagnoses and are more cost-effective than 48-hour holter monitoring in patients with palpitations: A controlled clinical trial. *Annals of Internal Medicine*, 124(1 Part 1):16–20.

Koerkamp, B. G., Hunink, M. G. M., Stijnen, T., Hammitt, J. K., Kuntz, K. M., and Weinstein, M. C. (2007). Limitations of acceptability curves for presenting uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 27(2):101(11).

Laska, E. M., Meisner, M., and Siegel, C. (1997). Statistical inference for cost-effectiveness ratios. *Health Economics*, 6(3):229–242.

Laska, E. M., Meisner, M., and Siegel, C. (1999). Power and sample size in cost-effectiveness analysis. *Medical Decision Making*, 19(3):339–343.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J.

Lehmann, E. L. and D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, volume 351. Holden-Day San Francisco.

Lehmann, E. L., Romano, J. P., and SpringerLink (2005). *Testing Statistical Hypotheses*. Springer texts in statistics. Springer, New York.

Leshno, M. and Levy, H. (2002). Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance. *Management Science*, 48(8):1074–1085.

Leshno, M. and Levy, H. (2004). Stochastic dominance and medical decision making. *Health Care Management Science*, 7(3):207.

Levy, H. and Hanoch, G. (1970). Relative effectiveness of efficiency criteria for portfolio selection. *The Journal of Financial and Quantitative Analysis*, 5(1):63–76.

Liew, D. (2006). A model for undertaking effectiveness and cost-effectiveness analyses of primary preventive strategies in cardiovascular disease. *European Journal of Cardiovascular Prevention and Rehabilitation*, 13(4):515–522.

Long, E. F., Brandeau, M. L., Galvin, C. M., Vinichenko, T., Tole, S. P., Schwartz, A., Sanders, G. D., and Owens, D. K. (2006). Effectiveness and cost-effectiveness of strategies to expand antiretroviral therapy in st. petersburg, russia. *AIDS*, 20(17):2207–2215.

Mark, D. B., Hlatky, M. A., Califf, R. M., Naylor, C. D., Lee, K. L., Armstrong, P. W., Barbash, G., White, H., Simoons, M. L., and Nelson, C. L. (1995). Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *New England Journal of Medicine*, 332(21):1418–1424.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, pages 77–91.

Matchar, D. B., Samsa, G. P., and Liu, S. (2005). Cost-effectiveness of antiplatelet agents in secondary stroke prevention: The limits of certainty. *Value in Health*, 8(5):572–580.

Matchar, D. B., Samsa, G. P., Matthews, J. R., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Wolf, P. A., D'Agostino, R. B., and Lipscomb, J. (1997). The stroke prevention policy model: Linking evidence and clinical decisions. *Ann Intern Med*, 127(8 Part 2):704–711.

McFadden, D. (1989). Testing for stochastic dominance. In Fomby, T. B. and Seo, T. K., editors, *Studies in the Economics of Uncertainty*, pages 113–134. Springer, New York, Berlin, London, and Tokyo.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley series in probability and statistics. Applied probability and statistics section. Wiley, New York.

McMahon, L. F. J., Haywarad, R., Saint, S., Chernew, M., and Fendrick, A. M. (2005). Univariate solutions in a multivariate word: Can we afford to practice as in the "good old days"? *The American Journal of Managed Care*, 11(8):473–474.

Meenan, R. T., Stevens, V. J., Hornbrook, M. C., Chance, P.-A. L., Glasgow, R. E., Hollis, J. F., Lichtenstein, E., and Vogt, T. M. (1998). Cost-effectiveness of a hospital-based smoking cessation intervention. *Medical Care*, 36(5):670–678.

Mehta, M., Noyes, W., Craig, B., Lamond, J., Auchter, R., French, M., Johnson, M., Levin, A., Badie, B., Robbins, I., and Kinsella, T. (1997). A cost-effectiveness and cost-utility analysis of radiosurgery vs. resection for single-brain metastases. *International Journal of Radiation Oncology*Biology*Physics*, 39(2):445–454.

Murray, C. J., Evans, D. B., Acharya, A., and Baltussen, R. M. (2000). Development of who guidelines on generalized cost-effectiveness analysis. *Health Economics*, 9(3):235–251.

O'Brien, B. J., Drummond, M. F., Labelle, R. J., and Willan, A. (1994). In search of power and significance: Issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care*, 32(2):150–163.

O'Hagan, A. and Stevens, J. (2002). The probability of cost-effectiveness. *BMC Medical Research Methodology*, 2(1):5.

Paltiel, A. D., Fuhlbrigge, A. L., Kitch, B. T., Liljas, B., Weiss, S. T., Neumann, P. J., and Kuntz, K. M. (2001). Cost-effectiveness of inhaled corticosteroids in adults with mild-to-moderate asthma: Results from the asthma policy model. *Journal of Allergy and Clinical Immunology*, 108(1):39–46.

Pauker, S. G. and Wong, J. B. (2005). The influence of influence diagrams in medicine. *Decision Analysis*, 2(4):238–244.

Pliskin, J. S., Shepard, D. S., and Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations Research*, 28(1):206–224.

Polsky, D., Glick, H. A., Willke, R., and Schulman, K. (1997). Confidence intervals for cost-effectiveness ratios: A comparison of four methods. *Health Economics*, 6(3):243–252.

Porter, R. B., Wart, J. R., and Ferguson, D. L. (1973). Efficient algorithms for conducting stochastic dominance tests on large numbers of portfolios. *The Journal of Financial and Quantitative Analysis*, 8(1):71–81.

Rosen, A. B., Hamel, M. B., Weinstein, M. C., Cutler, D. M., Fendrick, A. M., and Vijan, S. (2005). Cost-effectiveness of full Medicare coverage of angiotensin-converting enzyme inhibitors for beneficiaries with diabetes. *Annals of Internal Medicine*, 143(2):89.

Ross, S. M. (2003). *Introduction to Probability Models*. Academic Press, 8 edition.

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

Sculpher, M. and Gafni, A. (2001). Recognizing diversity in public preferences: The use of preference sub-groups in cost-effectiveness analysis. *Health Economics*, 10(4):317–324.

Sendi, P., Al, M. J., Gafni, A., and Birch, S. (2003). Optimizing a portfolio of health care programs in the presence of uncertainty and constrained resources. *Social Science & Medicine*, 57(11):2207–2215.

Siegel, C., Laska, E., and Meisner, M. (1996). Statistical methods for cost-effectiveness analyses. *Controlled Clinical Trials*, 17(5):387–406.

Sonnenberg, F. and Beck, R. (1993). Markov models in medical decision making: A practical guide. *Medical Decision Making*, 13(4):322–338.

Stevens, W. and Normand, C. (2004). Optimisation versus certainty: understanding the issue of heterogeneity in economic evaluation. *Social Science & Medicine*, 58(2):315–320.

Stinnett, A. A. and Mullahy, J. (1998). Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making*, 18(2):S68.

Thompson, S. G. and Barber, J. A. (2000). How should cost data in pragmatic randomised trials be analysed? *BMJ*, 320(7243):1197–1200.

Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies*, 25(2):65–86.

Valenstein, M., Vijan, S., Zeber, J. E., Boehm, K., and Buttar, A. (2001). The cost-utility of screening for depression in primary care. *Annals of Internal Medicine*, 134(5):345–360.

van Hout, B. A., Al, M. J., Gordon, G. S., and Rutten, F. F. (1994). Costs, effects and C/E-ratios alongside a clinical trial. *Health Economics*, 3(5):309–19.

Vijan, S., Hofer, T. P., and Hayward, R. A. (1997). Estimated benefits of glycemic control in microvascular complications in type 2 diabetes. *Annals of Internal Medicine*, 127(9):788–795.

Von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton Univ. Press, Princeton.

Wakker, P. and Klaasen, M. P. (1995). Confidence intervals for cost/effectiveness ratios. *Health Economics*, 4:373–381.

Weinstein, M. C. and Stason, W. B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine*, 296(13):716–721.

Weinstein, M. C., Toy, E. L., Sandberg, E. A., Neumann, P. J., Evans, J. S., Kuntz, K. M., Graham, J. D., and Hammitt, J. K. (2001). Modeling for health care and other policy decisions: Uses, roles, and validity. *Value in Health*, 4(5):348–361.

Willan, A. (2001). On the probability of cost-effectiveness using data from randomized clinical trials. *BMC Medical Research Methodology*, 1(1):8.

Yang, Y. and Nair, V. N. (2011). Inference for time-to-failure with multistate data: A comparison of traditional and process-based approaches. *Canadian Journal of Statistics*, (In press).

Zaric, G. S. (2003). The impact of ignoring population heterogeneity when Markov models are used in cost-effectiveness analysis. *Med Decis Making*, 23(5):379–386.

Zethraeus, N., Johannesson, M., Jonsson, B., Lothgren, M., and Tambour, M. (2003). Advantages of using the net-benefit approach for analysing uncertainty in economic evaluation studies. *PharmacoEconomics*, 21:39–48.

Zethraeus, N., Lthgren, M., and Sdertlje, A. Z. R. D. (2000). On the equivalence of the net benefit and the fieller's methods for statistical inference in cost-effectiveness analysis. *Working Paper Series in Economics and Finance*.