

Decision Analysis of Surgical Treatment Indications for Metastatic Epidural Spinal Cord Compression

by

Markian Pahuta

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the Ph.D. degree in
Epidemiology

School of Epidemiology and Public Health
Faculty of Medicine
University of Ottawa

© Markian Pahuta, Ottawa, Canada, 2019

Abstract

Metastatic epidural spinal cord compression (MESCC) occurs when tumour invades the epidural space and compresses the spinal cord. Despite Level 1 evidence that surgery is the most effective treatment for MESCC, there is controversy regarding the role of surgery because of fear that patients who have a short survival will spend a large fraction of their remaining life recovering from surgery and potential complications. This controversy could be resolved by decision-analysis of MESCC treatments using quality-adjusted-life-years (QALYs).

There have been two barriers to conducting decision-analysis of QALYs for MESCC: (a) lack of utility data, and (b) skepticism regarding decision-analysis. The first four research chapters in this thesis address these barriers. The final research chapter reports a decision-analysis of QALYs on the role of surgery in MESCC.

Chapter 1 provides background information on the controversy regarding surgical treatment for MESCC and the rationale for each of the subsequent chapters. Chapter 2 reports a psychometric validation study of a web-based utility valuation module for MESCC. In Chapter 3, application of this module to a general population utility valuation study with a market research panel is described. In Chapter 4, the beneficial properties of Bayesian statistical analysis to minimizing “arbitrariness” in probabilistic sensitivity analysis are described in relation to prognostication for MESCC. Chapter 5 presents a strategy for simplifying and enhancing the transparency of Markov cohort simulation. Finally, the work presented in the research chapters is applied in Chapter 6 to conduct Markov cohort simulation to determine if patients with short survival derive net health-related quality-of-life benefit from surgery.

Pragmatic research around barriers to decision-analysis of QALYs for MESCC was conducted to resolve the controversy regarding the role of surgery in the treatment of MESCC. Under most circumstances, MESCC patients who can ambulate prior to treatment derive net HRQoL benefit from surgery, even if prognosis is poor. Non-ambulatory patients can derive net HRQoL benefit but only if the morbidity of surgery is relatively low. It is my hope that the work used to address barriers to decision-analysis of QALYs will be disseminated and applied in other clinical problems.

Preface/Statement of Academic Achievement

This manuscript-based thesis consists of a series of articles prepared for publication in a peer-reviewed journal.

The first paper, “The Psychometric Properties of a Self-Administered, Open-Source Module for Valuing Metastatic Epidural Spinal Cord Compression Utilities,” was published in *PharmacoEconomics – Open* in 2018.[167] I designed the website described in this paper but retained a web programmer to program and implement the code. The second paper, “A General Population Utility Valuation Study for Metastatic Epidural Spinal Cord Compression Health States,” was accepted for publication by *Spine* in January 2019.[168] The third paper, “Back to Bayesian: a Strategy to Enhance Prognostication of Metastatic Spine Disease,” was accepted for publication by the *International Journal of Clinical Practice* in February 2019.[169] The fourth paper, “A Technique for Estimating Approximate Transition Rates from Published Partitioned Survival Analyses,” was accepted for publication in *Cost Effectiveness and Resource Allocation* in June 2019. I developed the strategy and programmed the models. The fifth paper, “Decision Analysis of Prognostic Contraindications to Surgery for Metastatic Epidural Spinal Cord Compression,” will be submitted to a peer-reviewed journal. I programmed the Markov cohort simulation model using the statistical programming language R.

I am the first and corresponding author on all papers. I wrote all papers with guidance from supervisors and other co-authors. I submitted all manuscripts for publication and addressed all feedback during the peer-review process.

Acknowledgements

First and foremost, I would like to thank my accomplished, intelligent and beautiful wife Dr. Therese Kozminski for her love, unwavering support, patience and prodding during this drawn-out thesis process. I am notorious for multitasking and so embarked on additional “projects” during my time in the PhD program. Therese and I married in my first year in the PhD program. She gave birth to our first son Julian in year two of the PhD. I obtained my Royal College Fellowship in Orthopaedic Surgery in year three during which we also moved from Ottawa to Toronto so that I could start my fellowship in Spine Surgery. That same year, Therese gave birth to our first daughter Kalyna. During year four we emigrated from Canada to the USA where I accepted an academic spine surgeon position. Therese not only tolerated my relative absence from our family while I was completing my PhD, but she sacrificed her career as a Family Physician while encouraging me, motivating me, and making sure I had time to finish my thesis.

Julian and Kalyna, I regret having to frequently hide in the office to work on my thesis. I hope that you will learn from my example and become more efficient and productive than I am. I look forward to spending more time with you and not having to divide my attention.

I would like to express my deepest appreciation to my supervisor Dr. Doug Coyle and co-supervisor Dr. Carl van Walraven. Doug: thank you for your mentorship and willingness to engage with me in esoteric discussions on health economics and decision analysis. Carl: thank you for reviewing esoteric components of this thesis with a fine-tooth comb between the hours of 5 and 6 am so that I could continue working on them the same day. Doug and Carl: thank you for bailing me out when I was kicked-out of the PhD program because I did not check my University of Ottawa e-mail account in a timely manner. Your thoughtful insights and feedback impressed me and challenged me to grow as a researcher. Your engagement has strengthened this thesis.

I am extremely grateful to the members of my thesis committee: Dr. Eugene Wai and Dr. Joel Werier. You have been outstanding clinical mentors. It is because of working with both of you that I pursued spine surgery with a focus in oncology. Your involvement has been instrumental in making this thesis clinically relevant. You were both my champions during residency, and I know you worked behind the scenes to ensure I was well supported during my hiatus from clinical residency training. I value you as colleagues and hope we can collaborate in the future.

Table of Contents

Nomenclature	viii
List of Tables	ix
List of Figures	xi
1 Introduction	1
1.1 History of MESCC Treatment	2
1.2 Surgical Decision Making for MESCC	3
1.3 Barriers to Applying Decision-Analysis of QALYs to MESCC	5
1.4 Overview of Thesis	6
1.4.1 Synthesis of Research Chapters	8
2 SOAP-MESCC Psychometric Evaluation	9
2.1 Introduction	12
2.2 Methods	13
2.3 Results	17
2.4 Discussion	18
3 General Population Utility Valuation Study	21
3.1 Introduction	25
3.2 Material and Methods	25
3.3 Results	28
3.4 Discussion	29
3.5 Conclusion	30
3.6 Figures	31
3.7 Tables	33

4	Bayesian Prognostication	36
4.1	Clinical Scenario	39
4.2	Introduction	39
4.3	Dichotomous Prognostication	40
4.4	Continuous Prognostication	42
4.5	Discussion	44
5	Approximating Transition Rates	47
5.1	Introduction	50
5.2	Transition Rate Approximation	52
5.3	MESCC Case Study	54
5.4	Simulation Study	55
5.5	Discussion	59
6	Decision Analysis	62
6.1	Introduction	65
6.2	Methods	66
6.3	Results	71
6.4	Discussion	71
6.5	Conclusion	76
7	Discussion and Conclusion	77
7.1	Summary of Methodological Findings and Contributions	77
7.2	Summary of Main Clinical Findings	79
7.3	Limitations	80
7.4	Future Work	81
7.5	Conclusion	82
	APPENDICES	83
A	Appendices for Chapter 2	84
A.1	SOAP MESCC Module	84
A.2	SOAP platform	96
A.3	MESCC Health States	99

B	Appendices for Chapter 3	110
B.1	Regression Model for Utilities	110
B.2	Quota Samples Based on 2016 Canadian Census	112
B.3	Regression Modeling	114
C	Appendices for Chapter 4	129
C.1	Confidence Intervals for Median Survival Time	129
C.2	Sensitivity Analysis Through Effective Sample Size Adjustment	131
D	Appendices for Chapter 5	132
D.1	Transition Parameters for an Illness-Death Model	132
D.2	Multistate Estimation of Transition Rates	134
D.3	Worked Example	142
E	Appendices for Chapter 6	145
E.1	Bayesian Estimation	145
E.1.1	Bayesian MCMC Estimation of Event Probabilities for the Treatment-Phase	146
E.1.2	Bayesian MCMC Estimation of Transition Rates	153
E.1.3	Bayesian MCMC Estimation of a Mixed Effects Beta Regression Model for Utilities	161
E.2	Exact Formulas for Markov Cohort Simulation	167
	References	168

Nomenclature

cS+RT laminectomy without stabilization followed by external beam radiotherapy

DOC docetaxel

EORTC European Organisation for Research and Treatment of Cancer

HRQoL health-related quality-of-life

ICC intra-class correlation coefficient

MCID minimally clinically important difference

MESCC metastatic epidural spinal cord compression

mS+RT modern surgery followed by external beam radiotherapy

NSCLC non-small cell lung cancer

OS overall survival

PFS progression-free survival

QALY quality-adjusted life year

RAM ramucirumab

RT-alone external beam radiotherapy alone

SBRT stereotactic body radiation therapy

SOAP Self-directed Online Assessment of Preferences

List of Tables

2.1	MESCC Health State Attribute Phrasing	14
2.2	Participant Characteristics Stratified by Response Pattern	18
2.3	Agreement, Reliability, and Responsiveness Results	18
3.1	Participant Demographics	33
3.2	Utilities for all MESCC Health States	34
3.3	Selection Parameters for all Models	35
3.4	Summary of Beta Regression Analysis for Model 4	35
5.1	Data Abstracted from MESCC RCT PFS and OS Analyses	54
5.2	Comparison of True and Approximate Transition Rates and Hazard Ratios for MESCC RCT	56
5.3	Simulation Results for $(\lambda_{hi} + \lambda_{hd})$	59
5.4	Simulation Results for ρ	60
5.5	Simulation Results for λ_{id}	61
6.1	Selection Criteria used in the MESCC RCT[171]	67
B.1	Quotas Based on the 2016 Canadian Census	112
B.2	Functional Form of Linear Predictor Formulas for Beta Regression	116
B.3	Selection Parameters for all Models	116
B.4	Calibration and Agreement Parameters for all Candidate Models	117
B.5	Mean Utilities by Model	118
B.6	Linear Predictor Formulas with Coefficients	119
B.7	Utilities for all MESCC Health States using Full Dataset	128
C.1	Parameters used for Interpolation of $\text{var}[S(t_m)]$	130

D.2	Comparison of Transition Rates and Hazard Ratios Estimated with Multi-state Techniques under Short and Long Simulation Conditions	143
E.1	Comparison of results from short and long simulation conditions for multi-state estimated transition rates and hazard ratios	148
E.3	Comparison of Transition Rates and Hazard Ratios Estimated with Multi-state Techniques under Short and Long Simulation Conditions	162
E.4	Comparison of results from short and long simulation conditions for beta regression coefficients for utilities	164

List of Figures

3.1	Flow of participants.	31
3.2	Calibration Plot for Model 4.	32
5.1	State-transition diagram for illness-death model	50
5.2	Comparison of state-membership fractions, RT-alone arm	57
5.3	Comparison of state-membership fractions, mS+RT arm	58
6.1	State-transition diagram for decision model	68
6.2	Utilities by number of dysfunctions	70
6.3	Individual simulation Δ QALYs versus number of dysfunctions	72
6.4	Mean Δ QALYs versus number of dysfunctions	73
A.1	MESCC SOAP: page 1	86
A.2	MESCC SOAP: page 2	87
A.3	MESCC SOAP: page 3	88
A.4	MESCC SOAP: perfect health	89
A.5	MESCC SOAP: page 4	90
A.6	MESCC SOAP: probe health state	91
A.7	MESCC SOAP: standard gamble interface	92
A.8	MESCC SOAP: perfect health pop-out	93
A.9	MESCC SOAP: probe health state pop-out	94
A.10	MESCC SOAP: logic check	95
B.1	Variance function check.	120
B.2	Link function check.	121

Chapter 1

Introduction

Over one-third of patients with carcinoma or hematological malignancy will develop spinal metastases.[73, 163] A particularly disabling consequence of these metastases occurs when tumour invades the space between the vertebra and spinal cord (epidural space) and compresses the spinal cord (metastatic epidural spinal cord compression, MESCC).[14, 62] The clinical presentation of MESCC is variable owing to the complex functional anatomy of the spinal cord. Depending on the axial location of compression, patients may experience a motor and/or sensory disturbance. Furthermore, depending on the axial and cranio-caudal location of compression, the neurologic deficit may affect any combination of extremities and the trunk.[216]

A motor deficit can manifest as isolated weakness in a group of muscles, paraplegia, hemiplegia, or quadriplegia. Sensory impairment can present as numbness, pain, or loss of spatial sense in any limb. In concert, a sensory and motor dysfunction can cause loss of hand function, ambulation, urinary and/or faecal continence. These consequences of MESCC can be quite distressing to patients and may lead to clinical depression.[195] Aggravating these functional losses is the loss of independence.

Patients must adapt to these physical and functional limitations. With lower extremity dysfunction, patients may not be able to independently transfer from bed and will require a lift or assistance from others.[158] Furthermore, immobility predisposes patients to skin ulcers and therefore prophylactic skin care must be performed.[34] Sphincter dysfunction is managed by diapers, urinary catheters and bowel regimes.[27, 44] If upper limbs are affected, the manual dexterity to perform basic daily activities such as dressing, toileting and bathing can be lost.[208] Unfortunately, patients with general functional decline from cancer may not have the stamina to learn new adaptive skills.[218] MESCC patients with severe neurologic deficits may be placed in long-term care or hospice institutions if there are insufficient resources to care for them at home. MESCC can have a marked negative impact on health-related quality-of-life (HRQoL).

1.1 History of MESCC Treatment

Surgery for MESCC is a controversial topic. Perceptions on the role of surgery were shaped by high-quality comparative studies published in the late 1970s and early 1980s that found surgery followed by external beam radiotherapy to be inferior to conventional external beam radiotherapy alone (RT-alone) for maintenance of ambulation and survival.[83, 255] Consequently initial surgical indications for MESCC were quite limited:[220, 42]

1. For tissue diagnosis
2. Following deterioration during or relapse after RT-alone
3. Inability to administer RT-alone due to spinal cord volume-dose limit having been reached

It is important to recognize that the surgical technique used in these early studies – laminectomy without stabilization – is not routinely used in the modern surgical era. Herein, classical surgical treatment will refer to laminectomy without stabilization followed by conventional radiotherapy and abbreviated as “cS+RT.”

Laminectomy is often ineffective for MESCC because the majority of spinal metastases originate in the vertebral body (which is in the anterior portion of the spinal column).[83] Therefore, for the majority of patients, a laminectomy will not effectively decompress the spinal cord because laminectomy only directly decompresses the posterior aspect of the spinal cord. Despite this limitation, in cancer patients who are often medically frail, laminectomy was initially preferred over direct decompression of the vertebral body because it avoided the morbidity of formal anterior approaches.[83, 245, 85] It was only after the mid-1980s that less morbid techniques for direct decompression of the vertebral body through an extra-cavitary approach were popularized.[220]

Even if the neural elements are successfully decompressed with a laminectomy, patients who have no stabilization can experience late neurologic deterioration. During a laminectomy, stabilizing ligaments and joints are removed.[184] When the vertebral body is weakened by tumor, the destabilized posteriorly decompressed segment can result in kyphosis and cause neurologic injury.[70, 103, 221] Due to better understanding of spinal stability, and improved instrumentation, modern surgical techniques for MESCC routinely include stabilization.[14, 71, 118, 204]

Modern surgical treatment for MESCC with direct decompression and stabilization followed by conventional external beam radiotherapy addresses the key limitations of cS+RT. Therefore, beliefs on the ineffectiveness of surgical treatment may not be applicable to the modern surgical era. Herein, “mS+RT” will be used to denote the modern surgical treatment approach. mS+RT has been subject to a randomized controlled trial and found to be superior to RT-alone for maintenance of ambulation and survival.[171] Large observational studies on mS+RT have found HRQoL benefit as well.[65]

More recently, treatment paradigms have changed based on advances in radiation therapy.[14] In contrast to conventional external beam radiotherapy, stereotactic body

radiation therapy (SBRT) more precisely delivers higher ablative doses of radiation.[106] Although radiation toxicity is lower with SBRT, high focal doses of radiation can predispose to compression fractures in otherwise stable lesions.[197] With higher focal doses of radiation, direct decompression may not be necessary because SBRT-alone can effectively treat even moderate-grade symptomatic MESCC.[196] However, surgery does still play a role in downgrading compression to facilitate SBRT and optimize local control.[190] Surgery with this intent is termed “separation surgery” and consists of minimal decompression to create a margin of cerebrospinal fluid (CSF) between the spinal cord and tumor. No Level 1 comparative studies of mS+RT to SBRT have been conducted.

1.2 Surgical Decision Making for MESCC

Although mS+RT is effective for MESCC, surgeons make this recommendation cautiously due to prognostic uncertainty. It is generally recommended that appropriate surgical candidates have an estimated survival of 3 – 6 months.[42, 229, 232, 230, 133] Although frequently used in the literature, there is no data to support this prognostic threshold which is based on expert opinion alone.

In a 1986 review article, Sundaresan and colleagues first proposed a 6-month survival threshold.[220] They noted that although neurologic response rates are equivalent with cS+RT and RT-alone, cS+RT conferred a more durable response. Since the majority of patients undergoing RT-alone relapsed within 6 months, they reasoned that patients with predicted survival exceeding 6 months should be offered cS+RT to protect against late neurologic deterioration. This logic may not apply to mS+RT because, in contrast to cS+RT, neurologic response rates are higher with mS+RT than with RT-alone.[255, 171, 189]

Several prognostic indices have been developed and validated for estimating expected survival of patients with metastatic spine disease.[172, 131, 211] The Tokuhashi Score, a widely used survival index for these patients, considers performance status, extra-spinal bone metastases, vertebral metastases, visceral metastases, pathologic diagnosis, and neurologic function.[230] The prognostic value of the Tokuhashi score for predicting 6-month survival has been studied in a recent meta-analysis.[133] The Tokuhashi score had a good discriminative performance (area under the receiver operator curve of 0.748) between patients who die prior to, or survive beyond, 6 months.[64]

Systemic chemotherapy is an important component of treatment for all MESCC patients regardless of treatment offered for spinal cord dysfunction (RT-alone, cS+RT, or mS+RT). Whereas the goal of RT-alone, cS+RT, and mS+RT is to improve physical function; systemic chemotherapy is administered to prolong survival. Due to rapid advances in systemic therapies, some authors have pointed out that existing prognostic scores used to guide surgical decision making may no longer be useful – in particular for non-small cell lung cancer (NSCLC). For example, the Tokuhashi Score was developed in 1990 and revised in 2005.[229, 230] During this time-period, only cytotoxic agents were available for the treatment of NSCLC and, consequently, patients with advanced cancer had a median

overall survival of 4 – 6 months. Currently, patients with advanced NSCLC may be offered targeted therapy such as anti-angiogenic drugs which has been shown to improve overall survival in NSCLC patients.[185] Prognostic uncertainty poses a significant challenge for the surgeon when the complication profile of the two treatment options, mS+RT and RT-alone, is considered.

Patients undergoing RT-alone may experience nausea, dermatitis, and symptoms from inflammation of the gastrointestinal tract (pharyngitis, esophagitis, and gastritis) within a month of receiving spinal external beam radiotherapy. Less than 10% of patients will report these symptoms to be severe.[146, 145, 237] Early toxicity may be even less frequent with the newer technique of stereotactic radiosurgery because radiation is delivered to a smaller field.[107] Infrequently, patients may experience late toxicities months after treatment, including myelopathy and vertebral compression fractures.[144] When compared to those from mS+RT, complications from RT-alone are less common and relatively mild.

Analysis of 30-day National Surgical Quality Improvement Program data found that 14.4% of patients undergoing mS+RT for primary and metastatic spinal tumors experience a major complication with 5.3% undergoing a re-operation and 10.2% being re-admitted.[121] Prospectively collected data from the AOSpine MESCC study reported the most common complications to be urinary tract (25.0%) and wound infections (10.4%).[65] Adverse events were significantly higher in a single-center cohort of patients undergoing emergency surgery for spinal metastases when complications were studied as the primary outcome.[47] In this study, 76.2% of patients experienced at least one complication, and the mortality rate during the admission was 10.9%. Intraoperative adverse events occurred in 31.7% of patients and included: blood loss >2L (16.8%), incidental durotomy (9.9%), mal-positioned implants requiring revision (5.9%), nerve root injury (2%), and cardiac events (1%). Postoperative adverse events occurred in 20.8% of patients and included: delirium (20.8%), transient neurological deterioration (5.9%), pressure sores (4.0%), deep vein thrombosis (4.0%), early construct failure (2.0%), electrolyte imbalance (11.0%), arrhythmia (4.0%), dysphagia (5.0%). The risks of mS+RT are substantial and can negatively affect HRQoL. However, patients experiencing complications can still derive overall net benefit from mS+RT.[47, 3]

The decision to offer mS+RT to a patient with MESCC is difficult because surgical recovery and potential complications must be considered vis-à-vis an uncertain survival time. Some clinicians may take a conservative approach and only offer RT-alone, and not mS+RT, to MESCC patients out of concern for the “worst-case” survival scenario.[196, 136] This is because recovery from surgery and potential complications will impact on a greater fraction of remaining life in patients having short survival, thereby negating net HRQoL benefit. A decision-analysis of quality-adjusted life years (QALYs) could help patients and clinicians jointly assess the trade-offs between survival benefits, HRQoL benefits, recovery and potential complications to reach an optimal treatment decision.[55, 123] This is because QALY concurrently quantify morbidity and mortality within a single parameter. To date, such an analysis has not been conducted for MESCC.

1.3 Barriers to Applying Decision-Analysis of QALYs to MESCC

Barrier 1: Lack of Utility Data

QALYs are calculated using utilities, or health-related HRQoL weights, which can be obtained by direct valuation or from generic health status measures.[249] Direct valuation is the classical approach in which individuals rate hypothetical health state descriptions using the time-trade-off or standard gamble procedures.[76] These procedures can be used to measure utilities for very specific and uncommon health states. However, it can be cumbersome to develop valid health state descriptions for particular diseases. Alternatively, techniques have been developed to convert generic health status measures (e.g. EuroQol-5D, Short Form-6D, or Health Utilities Index Mark 3) to utilities.[55] Conversion of generic health state measures is advantageous because custom health state descriptions are not required. However, utilities can only be obtained for health states actually observed in a cohort of patients completing a generic health survey. For MESCC, few high-quality studies compared mS+RT and RT-alone using generic health status measures, and no study has conducted direct valuation. Therefore, with a paucity of utility data, the ability to apply decision-analysis of QALYs to MESCC has been limited.

Barrier 2: Skepticism Regarding Decision Analysis

Once utility data are available, QALYs can be estimated directly from a study or calculated using simulation-based techniques. Study-based QALY estimates can be prohibitive due to large sample size requirements.[23] Furthermore, study-based decision-analysis is not a viable option when few high-quality comparative studies have collected HRQoL data (as in MESCC). Unfortunately, the alternative, simulation-based QALY estimates, are often viewed with skepticism.[58]

Markov cohort simulation is the most frequently used technique for estimating QALYs through simulation.[22] This technique first involves simulating disease history from a theoretical model of the disease process composed of distinct and clinically-relevant health states. The fraction of simulated patients in each health state over time is simulated using transition probabilities derived from the literature. QALYs are then calculated by weighing the fraction of simulated patients in each health state by the utility of those health states, and then summing the weighed state-membership fractions over the time-horizon of interest.

One criticism of Markov cohort simulation stems from current recommendations for probabilistic rather than deterministic analysis.[21, 36] Deterministic analysis refers to running the simulation model with only one “best” value for each of the model inputs. On the other hand, probabilistic analysis involves randomly selecting many plausible values for each model input and re-running the analysis with each random draw, the goal of which being the examination of the sensitivity of conclusions to the uncertainty in model

inputs. The random draws are taken from probability density functions derived from confidence intervals reported in the literature for each of the model inputs.[22] It is crucial to appreciate that confidence intervals do not quantify the uncertainty in the model inputs.[28] Furthermore, converting a confidence interval to a probability density function can be arbitrary and subjective. Hence, Markov cohort simulation can be criticized because the analysis can be biased by the choice of probability density functions.[21]

Many studies do not report sufficient data to properly compute transition probabilities which are needed for Markov cohort simulation. Many studies will only report progression-free survival (PFS) and overall survival (OS) analyses.[251] PFS is the probability of remaining healthy to time t ; while OS is the probability of being alive (either healthy or ill) at time t . Notably, the PFS and OS do not directly analyze ill patients. Multistate estimation of transition rates is needed to properly calculate transition probabilities.[105, 4, 181] Unfortunately sufficient data for multistate estimation is infrequently reported. Despite this, transition probabilities are often opaquely and improperly derived from PFS and OS analyses.[253]

Not surprisingly, another criticism of Markov cohort simulation is that the technique is complex and poorly reported in the literature.[57, 63, 63, 13, 56, 176, 192, 29] Perhaps due to the resulting lack of transparency in modeling, some authors characterize simulation-base techniques as a “black box.”[29] To be useful for treatment decision making, clinicians and patients must have faith in, and understanding of, these models.[16]

1.4 Overview of Thesis

This thesis consists of seven chapters: an introduction and background chapter; five research chapters; and a discussion chapter. The first four research chapters, Chapters 2, 3, 4, and 5, address the barriers to decision-analysis in MESCC. In the final research chapter, Chapter 6, I present a QALY decision-analysis of mS+RT and RT-alone for MESCC.

Barrier 1: Lack of Utility Data

Given that few high-quality studies have compared mS+RT and RT-alone for MESCC using generic health status measures, direct utility valuation is the most efficient way to collect utility data. Best practices in economic evaluation involve recruiting a sample of healthy individuals from the general population for direct utility valuation (*ex ante* utilities).[248, 198] It is important to appreciate that *ex ante* MESCC utilities from healthy individuals may not be equivalent to *ex post* utilities from patients who have experienced MESCC.[15, 256, 46] Although *ex ante* utilities are theoretically restricted to system policy decisions,[187] *ex ante* utilities have become the *de facto* standard for individual patient decisions. This is because utilities obtained from generic health surveys such as the EuroQol-5D, Short Form-6D, and Health Utilities Index Mark 3 are actually *ex ante* valuations.[55] Therefore, I sought to elicit *ex ante* MESCC utilities to conform with conventions in the literature.

Traditional general population utility valuation are conducted using face-to-face interviews, phone interviews, or postal surveys.[59] These forms of survey administration are time-intensive and costly. As a result, web-based surveys are increasingly used.[60, 61, 180, 101, 143, 201, 239, 84, 109, 240, 6, 117, 159, 247] Web-based studies are typically conducted using proprietary software which limit application to other disease contexts. Furthermore, the psychometric properties of these propriety software programs have not been assessed.[223]

It is important to determine whether web-based utility valuation has acceptable psychometric properties. If this mode of administration has acceptable psychometric properties, it would be beneficial and efficient for other investigators to be able to build disease-specific modules on a common platform that has been used to develop modules with acceptable psychometric properties. To meet this need, in Chapter 2 “The Psychometric Properties of a Self-Administered, Open-Source Module for Valuing Metastatic Epidural Spinal Cord Compression Utilities,” I describe the development of a new open-source (non-proprietary) web-based self-directed utility valuation platform, called the Self-directed Online Assessment of Preferences (SOAP), that can be used on major computer systems (including touch-screen devices). In Chapter 2, I determine the validity, reproducibility, and responsiveness of a SOAP module for MESCC.

Chapter 3 “A General Population Utility Valuation Study for Metastatic Epidural Spinal Cord Compression Health States” presents a general population utility valuation study using the SOAP MESCC module. For this study, a sample of 1138 adult Canadians was recruited using a market research company. I report utility valuations for 31 unique MESCC health states and explore the relative importance of MESCC-related consequences on HRQoL.

Barrier 2: Skepticism Regarding Decision-Analysis

Regarding the concern that probabilistic analysis is arbitrary and prone to manipulation, I propose that Bayesian statistics offer a solution. This branch of statistics was developed in the 18th century and predates the more commonly used Frequentist statistics which were developed in the early 20th century.[68] This type of analysis provides a rigorous and non-arbitrary framework for analyzing data. The output of a Bayesian analysis is a posterior distribution which can be used in place of user defined distributions in economic analysis. The purpose of Chapter 4 “Back to Bayesian: a Strategy to Enhance Prognostication of Metastatic Spine Disease” is to introduce the concepts of Bayesian statistics to clinicians. This chapter also reviews the difficulties in treatment decision making for MESCC.

Credibility of Markov cohort simulation suffers when models are improperly populated with the results of partitioned-survival analysis.[251] However, the correct technique for these situations, partitioned-survival modeling, does not yield as informative of results as state-transition modeling. Chapter 5 “A Technique for Approximating Transition Rates from Published Survival Analyses” presents a solution to this problem. In this chapter, I develop and validate a technique to derive transition probabilities from summary partitioned-survival analysis data.

1.4.1 Synthesis of Research Chapters

The ultimate purpose of this thesis is to determine whether MESCC patients derive net HRQoL benefit from mS+RT compared to RT-alone. This work is motivated by controversy regarding the role of mS+RT despite Level 1 evidence that this treatment is more effective than RT-alone for MESCC.[196, 136]

I will try to resolve the controversy regarding the role of mS+RT by determining if patients with short survival derive net HRQoL with mS+RT relative to RT-alone. If patients benefit from surgery even under the “worst-case” survival scenario, clinicians and patients could make treatment decisions with greater comfort. In Chapter 6 “Decision Analysis of Prognostic Contraindications to Surgery for Metastatic Epidural Spinal Cord Compression,” I determine the conditions under which patients with poor prognosis benefit from surgery using decision-analysis of QALYs.

The first four research chapters set the foundation for Chapter 6. In this final chapter, I conduct Markov cohort simulation using a simple and transparent multistate MESCC disease model developed in Chapter 5. QALYs are calculated using the utilities valued in Chapter 3. In an effort to minimize perceptions of “arbitrariness,” the analysis is run using posterior Bayesian distributions (Chapter 4).

Chapter 2

The Psychometric Properties of a Self-Administered, Open-Source Module for Valuing Metastatic Epidural Spinal Cord Compression Utilities

Markian A. Pahuta¹, Aaron Frombach², Emile Hashem³, Stewart Spence³, Christina Sun³, Eugene K. Wai^{4,5}, Joel Werier⁴, Carl van Walraven^{5,6,7}, Doug Coyle⁵

- 1 – Divisions of Orthopaedic & Neurological Surgery, Henry Ford Health System, Detroit, Michigan, USA
- 2 – Postgraduate Medical Education, The University of Ottawa, Ottawa, Ontario, Canada
- 3 – Undergraduate Medical Education, The University of Ottawa, Ottawa, Ontario, Canada
- 4 – Division of Orthopaedic Surgery, The University of Ottawa, Ottawa, Ontario, Canada
- 5 – School of Epidemiology and Public Health, The University of Ottawa, Ottawa, Ontario, Canada
- 6 – The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
- 7 – The Institute for Clinical Evaluative Sciences, Ottawa, Ontario, Canada

Context

In this chapter, I describe the development of a new open-source (non-proprietary) web-based self-directed utility valuation platform, called the Self-directed Online Assessment of Preferences (SOAP), that can be used on major computer systems (including touch-screen devices). I also determine the validity, reproducibility, and responsiveness of a SOAP module for MESCC. The purpose of this chapter is to present a valid tool that will be used to address the lack of utility data for MESCC.

Abstract

Objectives

Web surveys are often used for utility valuation. Typically, custom utility valuation tools that have not undergone psychometric evaluation are used.

This study aims to determine the psychometric properties of a metastatic epidural spinal cord compression (MESCC) module run on a customizable open-source internet-based self-directed utility valuation platform (Self-directed Online Assessment of Preferences; SOAP).

Methods

Individuals accompanying patients to the emergency department waiting room in Ottawa, Canada were recruited. Participants made SOAP MESCC health state valuations in the waiting room, and 48 hours later at home. Validity, agreement reliability, and responsiveness were measured by logical consistency of responses, Smallest Detectable Change, the Interclass Correlation Coefficient, and Guyatt's Responsiveness Index respectively.

Results

Of 285 participants who completed utility valuations, only 113 (39.6%) completed the re-test. Of these 113 participants, 92 (81.4%) provided valid responses on the first test, and 75 (66.4%) provided valid responses on the test and re-test. Agreement for all groups of health states was adequate since their Smallest Detectable Change was less than the Minimally Clinically Important Difference. The mean Interclass Correlation Coefficient s for all health states were greater than 0.8 indicating at least substantial reliability. Guyatt's Responsiveness Indices all exceeded 0.80, indicating high level of responsiveness.

Conclusions

To our knowledge, this is the first validated open-source, web-based, self-directed utility valuation module. We have demonstrated the SOAP MESCC module to be a valid, reproducible and responsive for obtaining *ex ante* utilities. Considering the successful psychometric validation of the SOAP MESCC module, other investigators can consider developing modules for other diseases where direct utility valuation is needed.

2.1 Introduction

Quality-adjusted-life-years (QALYs) are used to concurrently quantify morbidity and mortality within a single parameter.[55] For this reason, QALYs can facilitate the discussion of risks and benefits during patient counselling regarding treatment options.[123] To help make funding decisions, policy makers may also combine QALYs with cost estimates to calculate the incremental cost effectiveness ratio.[37] QALYs are calculated using “utilities,” or health-related quality of life weights, which are obtained by direct valuation or from generic health status measures.[249]

The choice of utility valuation approach is driven by available data. Direct valuation is the classical approach in which individuals rate hypothetical health state descriptions using the time-trade-off or standard gamble procedures.[76] These procedures can be used to measure utilities for very specific and uncommon health states. However, it can be cumbersome to develop valid health state descriptions for particular diseases. Alternatively, techniques have been developed to covert generic health status measures (e.g. EuroQol-5D, Short Form-6D, or Health Utilities Index Mark 3) to utilities.[55] Conversion of generic health state measures is advantageous because custom health state descriptions are not required. However, utilities can only be obtained for health states actually observed in a cohort of patients involved in the generic health survey.

Unfortunately, generic health scores have not been collected for many diseases, thereby making direct valuation a necessary option for measuring utilities. Best practices in economic evaluation are to recruit a sample of healthy individuals from the general population for utility valuation.[198, 248] Traditionally, general population utility valuation was conducted using face-to-face interviews, phone interviews, or postal surveys.[59] These forms of survey administration are time-intensive and costly, and as a result web-based surveys are increasingly used. [60, 61, 6, 117, 160, 247, 179, 101, 143, 201, 238, 84, 109, 241] Typically these studies are conducted using proprietary software which limits application to other disease contexts. Furthermore, the psychometric properties of these propriety software programs have not been assessed.[223] [84]

It is important to determine whether web-based utility valuation has acceptable psychometric properties. If this mode of administration has acceptable psychometric properties, rather than building custom software for new utility valuation studies, it would be beneficial and efficient for investigators to be able to build disease-specific modules on a common platform that has been used to develop modules with acceptable psychometric properties. To meet this need, we developed a new open-source (non-proprietary) web-based self-directed utility valuation platform usable on major computer systems (including touch-screen devices) called the Self-directed Online Assessment of Preferences (SOAP) (Appendix A.1 and Appendix A.2). SOAP was designed with flexibility in mind and can accept new health state descriptions (modules) with minimal programming.

We decided to first create a SOAP module for metastatic epidural spinal cord compression (MESCC), a condition for which there is limited health-related quality of life data. MESCC can be treated with surgery or radiotherapy. But few high-quality studies compare these interventions using generic health status measures for patients. However,

surgery and/or radiotherapy outcomes could be compared using utilities obtained by direct valuation of hypothetical probe health state descriptions. The European Organisation for Research and Treatment of Cancer (EORTC) MESCC working group has developed a health-related quality of life questionnaire for MESCC.[151] Items from this questionnaire could be used to generate health state descriptions for a SOAP module.

The objective of this study was to determine if the SOAP platform can be used to develop a valid, reproducible, and responsive module for MESCC. For this first application of the SOAP platform, we developed a MESCC module based on the work done by the EORTC and measured psychometric properties in a general population sample.

2.2 Methods

SOAP Platform

Electronic utility valuation protocols are distinguished by the form of health state descriptions, assessment approach, navigation rules, and auxiliary functions.[138] A detailed description of these elements for the SOAP MESCC module are provided in Appendix A.1 and Appendix A.2.

MESCC Module

EORTC Phase 1 development of a MESCC questionnaire in Canada found that patients and health care providers felt that ambulation, urinary continence, pain, and independence were important health-related quality of life issues for MESCC. Since Phase 1 development was restricted to health-related quality of life, and did not specifically consider treatment effects and adverse events, we reviewed prospective studies on MESCC to identify reported outcomes and adverse events.[145, 171, 147, 65] The EORTC items captured all treatment outcomes identified in our review. However, the review identified a large and disparate set of adverse effects. To develop a manageable decision analytic model, all adverse effects were grouped as an “other symptoms” attribute.

A tabular (point-form) presentation of health states was chosen as it is preferred by participants, is believed to decrease cognitive burden compared to the narrative format, and produces similar results to the narrative format.[142, 199] Therefore, we presented health states as a point-form list of five dysfunctional attributes: non-ambulatory (N), incontinent of urine (I), pain (P), dependent (D), and “other symptoms” (S). To reduce the number of potential health states, EORTC items were collapsed to indicate the presence (+) or absence (-) of the dysfunctional attribute, producing 32 discrete health states (Appendix A.3).

When possible, the phrasing for presence or absence of dysfunctional attributes was created using same EORTC items identified in the MESCC module development process (Table 2.1). Items were rephrased to the second person and restructured as declarative sentences. Items describing feelings or worries were not utilized as it was desired to make the

Table 2.1: MESCC Health State Attribute Phrasing

Attribute	Symbol	Item	EORTC Code
Dependence	D-	You do not need help with eating, dressing, washing yourself or using the toilet. You are not dependent on others.	C05, BM49
	D+	You need help with eating, dressing, washing yourself or using the toilet. You are dependent on others.	C05, BM49
Non-ambulation	N-	You walk without assistance.	NA
	N+	You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.	BM48, BN48, C04
Incontinence	I-	You have no trouble controlling your bladder.	BN50
	I+	You need to insert a tube in your bladder or you need to wear a diaper full time, to control your bladder.	BN50, BLM44
Pain	P-	You have back pain relieved by pain medications.	BM31, BM38
	P+	You have back pain not relieved by pain medications.	BM31, BM38
Other Symptoms	S-	You have no other uncomfortable symptoms.	NA
	S+	You have one or more of: nausea, vomiting, shortness of breath, lack of appetite, diarrhea.	C14, C15, C08, C17

The symbol (+) indicates the presence of a dysfunctional attribute, and (-) indicates absence. The European Organisation for Research and Treatment of Cancer (EORTC) codes refer to identifiers in the EORTC Item Bank.

health states descriptions as objective as possible. The rationale for the specific attribute formulation was:

1. **Dependence (D).** The two items identified by the MESCC working group were combined into one attribute to highlight the implications of loss of independence. The qualifiers “do” and “do not” were added to indicate complete function and dysfunction.
2. **Lack of Ambulation (N).** The MESCC working group developed a new item which was used as the functional level. Again, two items were combined to highlight the implications of loss of mobility.
3. **Incontinence of Urine (I).** The item identified by the MESCC working group with a qualifier was used as the functional level. An item from the EORTC bladder cancer module (BLM44) was used to highlight the implications of loss of bladder control.
4. **Pain (P).** As MESCC can only occur in the cervical spine to the thoracolumbar junction, pain was not differentiated by the terms “upper” and “lower” back as was identified by the MESCC working group. As most patients with spine metastasis will have some element of pain, the functional state had patients requiring pain medications. Use of pain medication served as a qualifier, and was taken from the EORTC bone metastasis module (BM38).
5. **Other Symptoms (S).** Again, to maintain efficiency, all adverse effects were characterized by several common adverse symptoms. These items were all taken from the core EORTC questionnaire.

Evaluations were obtained using standard gamble method using a ping-pong search algorithm. Typically in the standard gamble, success is framed as “perfect” health for an undermined period of time. In this context, this can be inferred to be the absence of any dysfunctions. Therefore the fully functional health state (D-,N-,I-,P-,S-) was chosen as the success anchor. To eliminate confusion around life expectancy, all scenarios were framed as having a certain life expectancy of five years; that is, for both the probe health scenario and success health scenario, participants were told their life expectancy would certainly be five years. Five years was the maximum survival reported in a randomized controlled trial on treatments for MESCC.[171] Probe health states were presented in a random order.

The MESCC module was pilot tested in a sample of 40 participants to assess acceptability and ease of use. Participants were asked to rate the SOAP MESCC module using a five-point Likert rating for the statement “[t]his website is easy to use.” 92.5% of participants strongly agreed or agreed with the statement.

Subjects

To be compliant with best practices in economic evaluation, we sought to conduct a direct utility valuation study with a sample of the general population who have not experienced

MESCC using the SOAP MESCC module (*ex ante* valuation).[248] Prior to this general population direct valuation study, psychometric properties of the SOAP MESCC module had to be evaluated. To approximate a general population sample for this psychometric validation study, participants were recruited from the emergency department waiting rooms at The Ottawa Hospital, an academic hospital in Ottawa, Ontario, Canada. Only patients’ family members or friends (i.e. individuals accompanying patients) at least 18 years of age were eligible to participate. Participants were required to be able to read English and have access to the internet outside of the hospital. A minimum sample size of 50 participants has been recommended in published guidelines for reliability and responsiveness evaluations.[223] To ensure robust results, we set the sample size for this study at 75.

Survey Procedures

Participants completed the first survey in the emergency department using a touch screen device. Investigators did not assist participants in navigating or completing the survey. Each participant valued the health state D+N+I+P+S+, one randomly selected singly dysfunctional health state, and another triply dysfunctional health state. Dysfunctional elements were nested to ensure a logical ordering of utilities for the three health states. For example, if the singly dysfunctional health state was D-N-I+P-S-, the triply dysfunctional health state could include incontinence and two of: dependence, lack of ambulation, pain, or other symptoms.

Two days after the initial survey, investigators contacted participants through e-mail and/or phone with information to access the re-test. Participants completed the second survey using their personal device. For the retest, participants were presented with the same probe health states as they completed in the emergency room, but the states were presented in a random order (and not necessarily identical to the initial survey).

Statistical Analysis

Validity refers to whether a tool under investigation measures what it is supposed to measure.[152] Specifically, construct validity concerns whether results obtained using the tool under investigation are consistent with *a priori* hypotheses.[152] We hypothesized that utility valuations should follow the logical ordering of health states with utility valuations following the relationship: singly $\geq$ triply $\geq$ fully dysfunctional. We considered singly = triply = fully a valid response because we could not exclude the possibility of a ceiling effect with one dysfunction. Participant responses were deemed “valid” if their utilities followed this order. The proportion of participants providing valid responses on the test and re-test was computed.

Reproducibility concerns the stability of participants’ responses on repeated testing, and can be characterized by agreement and reliability.[152] Agreement quantifies the absolute differences in participants’ repeated responses. We assessed agreement using the

Smallest Detectable Change.[223] We classified agreement as adequate if the Smallest Detectable Change was less than the Minimally Clinically Important Difference (MCID).[223] By anchoring to Eastern Cancer Oncology Group functional levels, an MCID of 0.05 for cancer utilities obtained by the EuroQol-5D-3L has been proposed.[175] This MCID has also been used for direct utility valuation by the standard gamble and time-trade-off of EuroQol-5D-3L health states.[246] The precision of the standard gamble algorithm used in our study was also 0.05. Therefore, we use an MCID of 0.05 in this study. Systematic differences between the test and re-test sessions were quantified using the Smallest Detectable Change calculation. Reliability concerns the fraction of pooled study variance across the repeated tests attributable to differences between participants (participant variance) and individual test-retest variability (noise).[152] If responses are stable, the ratio of noise to participant variance should be small, and the ratio of participant variance to variance for the pooled results from test and re-test should be high. Reliability accounting for systematic differences between the test and re-test, stratified for the number of dysfunctions in the health state, was quantified using the Intra-Class Correlation Coefficient (ICC) using the following categories: < 0.21 , slight reliability; $0.21-0.40$, fair reliability; $0.41-0.60$, moderate reliability; $0.61-0.80$, substantial reliability; > 0.80 , almost perfect reliability.[129] An $ICC \geq 0.70$ was considered adequate.[223]

Responsiveness reflects the ability of a tool to detect clinically important changes and can be quantified using Guyatt’s Responsiveness Index.[91] The Guyatt’s Responsiveness Index is proportional to the ratio of the MCID to the root mean squared error of the difference between the test and re-test value. If test-retest variability is small relative to the MCID, the tool is deemed responsive because meaningful changes are of greater magnitude than test-retest fluctuation.[90] Values of 0.20, 0.50, and 0.80 were interpreted as small, moderate and large levels of responsiveness, respectively.[108]

Statistical analysis was performed using the statistical programming language R.[183] The distribution of age (Kruskal-Wallis test) and gender (Chi-squared test) was compared between participants providing valid and invalid responses on the test and re-test. We considered logically ordered responses to be valid, that is decreasing utilities assigned to the singly, triply, and fully dysfunctional states. Age was assessed using a one-way ANOVA. Gender was assessed using the chi-squared test. Reproducibility, agreement, reliability, and responsiveness were only measured on participants providing valid responses on both the test and re-test. Since the SOAP tool is intended for measuring average utilities from the general public, average measures (rather than individual measures) of Smallest Detectable Change, ICCs, and Guyatt’s Responsiveness Indices were calculated.[45]

2.3 Results

Of 285 participants who completed utility valuations in the emergency department, only 113 (39.6%) completed the re-test. Of these 113 participants, 92 (81.4%) provided valid responses on the first test, and 75 (66.4%) provided valid responses on the test and re-test (Table 2.2). The response validity pattern was not associated with age ($p = 0.2336$) or sex ($p = 0.971$) (Table 2.2). Only data from the participants providing valid responses

on both the test and re-test were used for reproducibility and responsiveness analysis. Seven respondents skipped at least one scenario during the test and were classified as providing invalid responses. Only one respondent skipped one question during the re-test, and the responses were also classified as invalid. Agreement for all groups of health

Table 2.2: Participant Characteristics Stratified by Response Pattern

	Valid to Valid	Valid to Invalid	Invalid to Valid	Invalid to Invalid	<i>p</i> -value
no. (%)	75 (66.4)	17 (15.0)	17 (15.0)	4 (3.6)	
Age – yr [†]	53.0 (33.5, 61.0)	43.0 (33.0, 48.0)	51.0 (38.0, 62.0)	43.0 (33.25, 52.0)	0.2336
Sex – no. (%)					0.971
Female	47 (62.7)	11 (64.7)	11 (58.8)	3 (75.0)	
Male	28 (37.3)	6 (35.3)	6 (41.2)	1 (25.0)	

[†]median and interquartile range in parentheses.

states was adequate since their Smallest Detectable Change was less than the MCID of 0.05 (Table 2.3). Mean ICCs were all greater than 0.8 indicating substantial reliability, and all ICCs were significantly greater than the pre-specified threshold of 0.7 (Table 2.3). Guyatt’s Responsiveness Indices all exceeded 0.80, indicating large responsiveness for the utility evaluation (Table 2.3).[90]

Table 2.3: Agreement, Reliability, and Responsiveness Results

Number of dysfunctions	Smallest detectable change	ICC (95% CI) [†]	Guyatt’s responsiveness index
Single	0.036	0.879 (0.808, 0.923)	1.45
Triple	0.035	0.898 (0.838, 0.935)	1.47
Full	0.042	0.826 (0.724, 0.890)	1.24

[†]Intra-Class Correlation Coefficient (ICC) and 95% Confidence Interval (95% CI).

2.4 Discussion

Traditional utility valuation studies have been conducted using face-to-face interviews, phone interviews, or postal surveys. These modes of administration have undergone psychometric validation. There is increasing use of web surveys for utility valuation which

are usually conducted using custom and proprietary valuation tools which have not been psychometrically validated. It would be beneficial and efficient for investigators to be able to build disease-specific modules on a common platform that has been used to develop modules with acceptable psychometric properties.

We developed a new platform called the Self-directed Online Assessment of Preferences (SOAP) (Appendix A.1 and Appendix A.2). For the first application of this platform, we developed a module for MESCC health states. The SOAP platform met published benchmarks for reproducibility (both agreement and reliability) and responsiveness for utility measurement. This study demonstrates that the SOAP platform can be used to develop modules with acceptable psychometric properties.

81.4% of participants provided valid responses on the first test, and 66.4% of participants provided valid responses on both the test and re-test. These results should be considered in the context of other *ex ante* valuation studies reported in the literature. We classified a participant's responses valid if their utility valuations decreased with increasing dysfunctional attributes in the health state. For example, if a participant valued the fully dysfunctional health state higher than the single dysfunctional health state, their responses were classified as invalid. This definition of validity is termed "logical consistency" and has been used in traditional general population *ex ante* utility valuation studies of EuroQol-5D-3L health states.

Logical consistency rates for face-to face valuations have been reported for the UK and Netherlands.[51, 128] In the UK study, 12 pairs of health states per participant, could be evaluated for logical consistency. The median rate of logical consistency, per participant, ranged from 83.8% to 91.7%. In the Dutch study, 87.6% of participants provided at least one pair of logically inconsistent valuations. Postal surveys conducted in the US and New Zealand reported at least one logically inconsistent pairing in 88% and 79% of participants respectively.[113, 49] With 81.4% of participants providing a valid response (28.6% providing a logically inconsistent response), the logical consistency rate for the SOAP MESCC module is similar to traditional population studies. Logical consistency and has also been assessed for other self-administered general population *ex ante* utility valuation studies of EuroQol-5D-3L health states over the internet.[6, 12, 7] Each study reported a logical consistency rate below 70%.

Compared to the SOAP MESCC module, the face-to-face, postal and web-based EuroQol-5D-3L utility valuation studies required greater cognitive effort because participants rated a greater number of health states (between 5-10) that were also more complex (5 attributes and 3 levels of dysfunction). Furthermore, these studies did not provide error checking whereas the SOAP MESCC module notified participants of a logical error if they rejected a lottery with 100% of success. Considering these differences, a logical consistency rate of 81.4% on the first test with the SOAP MESCC module is consistent with the literature.

Valuing MESCC health states using the classical standard gamble is problematic for two reasons. The first, the classical standard gamble uses perfect health as a top anchor which is an unrealistic outcome for metastatic cancer. The second, the classical standard gamble considers timeless (ie. perpetual) health states which are incongruent with the metastatic cancer disease process. To make the standard gamble more realistic, we characterised

perfect health as the absence of dysfunctions, and restricted all health states (including the top anchor) to a survival period of five years. These modifications may impact on the interpretation of our results relative to classic utility assessment.

Utilities are typically estimated for specific health states and are used to weight the time in such health states. Consequently, a utility value for a specific state is typically considered “timeless,” that is utilities are usually assumed not to change with time spent in a health state.[10] As a reflection of this, the duration of time spent in a probe health state is not specified in the classical standard gamble.[76] For MESCC health states, we were concerned that the most severe health states would connote poor survival, and therefore confound the measurement of health-related quality of life using the standard gamble with quantity-of-life. To alleviate this difficulty, we explicitly stated a five-year duration for each health state which was the longest survival observed in a randomized controlled trial of treatments for MESCC.[171] This approach has also been used in other utility valuation studies for cancer health states.[214] This modification to health state descriptions should not affect results because the standard gamble (and all other utility elicitation methods) relies on the utility independence assumption.[177] Under this assumption, if a health state has a utility of x , the utility of this health state for 5 years should still be x . Unfortunately, a systematic review concluded that individuals tend not to satisfy the utility independence assumption with no consistent pattern of violation.[234] We are unaware of any algorithm to convert utilities for fixed period of time to “timeless” utilities. Consequently, the utilities measured in this study may not be directly comparable to utilities obtained using the classical standard gamble.

A strength of our study is that we built on the work conducted by the EORTC MESCC working group to ensure that the attributes in the MESCC module were appropriate and representative of the MESCC disease process. A limitation of our study is that we did not assess criterion validity by comparing utilities obtained by SOAP MESCC and a “gold standard.”[152] This could be done by having MESCC patients to value their own health using the SOAP MESCC module and comparing these utility valuations to those derived from a generic health questionnaire. We did not have the resources to conduct such a study. Furthermore, measures of logical validity, reproducibility and responsiveness are more relevant than MESCC criterion validity to investigators considering developing modules for new diseases.

To our knowledge, this is the first validated open-source, web-based, self-directed utility valuation module. For the first application of the SOAP platform, we developed a module for MESCC health states. We have demonstrated the SOAP MESCC module to be a valid, reproducible and responsive for obtaining *ex ante* utilities. Considering the successful psychometric validation of the SOAP MESCC module, other investigators can consider developing modules for other diseases where direct utility valuation is needed.

Chapter 3

A General Population Utility Valuation Study for Metastatic Epidural Spinal Cord Compression Health States

Markian A. Pahuta¹, Eugene K. Wai^{2,3}, Joel Werier², Carl van Walraven^{3,4,5}, Doug Coyle³

- 1 – Divisions of Orthopaedic & Neurological Surgery, Henry Ford Health System, Detroit, Michigan, USA
- 2 – Division of Orthopaedic Surgery, The University of Ottawa, Ottawa, Ontario, Canada
- 3 – School of Epidemiology and Public Health, The University of Ottawa, Ottawa, Ontario, Canada
- 4 – The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
- 5 – The Institute for Clinical Evaluative Sciences, Ottawa, Ontario, Canada

Context

In this chapter, I present a general population utility valuation study using the validated SOAP MESCC module described in Chapter 2. This is the first comprehensive set of *ex ante* utility estimates for MESCC health states. These data will be used to conduct QALY decision analysis in Chapter 6.

Abstract

Study Design

General population utility valuation study.

Objectives

This study obtained utility valuations from a Canadian general population perspective for 31 unique Metastatic Epidural Spinal Cord Compression (MESCC) health states and determined the relative importance of MESCC-related consequences on quality-of-life.

Summary of Background Data

Few prospective studies on the treatment of MESCC have collected quality-adjusted-life-year weights (termed “utilities”). Utilities are an important summative measure which distills health outcomes to a single number that can assist healthcare providers, patients, and policy makers in decision making.

Methods

We recruited a sample of 1138 adult Canadians using a market research company. Quota sampling was used to ensure that the participants were representative of the Canadian population in terms of age, gender, and province of residence. Using the validated MESCC module for the “Self-administered Online Assessment of Preferences” (SOAP) tool, participants were asked to rate 6 of the 31 MESCC health states, each of which presented varying severities of 5 MESCC-related dysfunctions (dependent; non-ambulatory; incontinent; pain; other symptoms).

Results

Participants equally valued all MESCC-related dysfunctions which followed a pattern of diminishing marginal disutility (each additional consequence resulted in a smaller incremental decrease in utility than the previous). These results demonstrate that the general population values physical function equal to other facets of quality-of-life.

Conclusions

We provide a comprehensive set of *ex ante* utility estimates for MESCC health states that can be used to help inform decision making. This is the first study reporting direct utility valuation for a spinal disorder. Our methodology offers a feasible solution for obtaining

quality-of-life data without collecting generic health status questionnaire responses from patients.

3.1 Introduction

The decision to offer a patient with metastatic epidural spinal cord compression (MESCC) surgery followed by radiotherapy (S+RT) or radiotherapy alone (RT) is difficult because survival, physical function, and complications must be considered jointly. Quality-adjusted-life-year (QALY) analysis could allow patients and clinicians to determine the relative weight of these factors and reach an optimal decision. QALYs are calculated using “utilities,” or health-related quality of life weights, which are usually derived from patient responses to generic health status questionnaires (e.g. EuroQol-5D, Short Form-6D, or Health Utilities Index 3).[55] Unfortunately there is a paucity of quality-of-life data for MESCC as few high-quality studies compare interventions using generic health status measures.

When generic health status measures are not available, utilities can be derived by direct valuation. Direct valuation is the classical approach in which individuals rate hypothetical health state descriptions using the time-trade-off or standard gamble procedures.[76] These procedures can be used to measure utilities for very specific and uncommon health states. Best practices in economic evaluation are to recruit a sample of healthy individuals from the general population for direct utility valuation.[248, 198]

Based on the work of the European Organisation for Research and Treatment of Cancer (EORTC),[151] our group has developed a comprehensive set of 31 unique MESCC health state descriptions.[167] We have also developed an online utility valuation module for MESCC using the “Self-administered Online Assessment of Preferences” (SOAP) tool. This module was found to be valid, reproducible and responsive in a sample of individuals who have not experienced MESCC.[167]

The primary objective of this study is to obtain general population utility valuations for the 31 unique MESCC health states. The secondary objective of this study is to determine the relative importance of various aspects to quality-of-life in MESCC.

3.2 Material and Methods

Subjects

Recruiting a general population probability sample is a time-intensive and costly undertaking. Typically this is done by visiting homes, using random-digit-dialing, and selecting random phone numbers. In Canada, with over 80% of Canadians age 16 years and older having access to the internet,[94] web recruitment offers a practical alternative to traditional strategies. As has been done by several other investigators, we recruited participants from a proprietary market research panel (Toluna Group Ltd) for this utility elicitation study for MESCC.[6, 12, 7]

The market research panel was composed of over 80,000 individuals across Canada recruited by the company into a panel of potential survey participants through random-digit-dialing, internet banner advertisements, and partnerships with corporations to become panel members.[231] Panel members agree to be contacted about new surveys. We

did not provide an incentive for participating in our study. However, the market research company managing the panel does award monthly prizes to panel members based on the number and length of surveys completed. Quota sampling was used to ensure that the sample of the market research panel represented the general Canadian population in terms of region of residence, gender, and age based on the 2016 Canadian Census (Table B.1). The market research company sent panel members an e-mail invitation to participate in our study. Interested panel members were redirected to a secure website hosting the utility valuation exercise.

Survey Procedures

Participants were asked to value six health states in the online SOAP MESSC module. The SOAP MESSC module has previously been described.[167] Briefly, the first three pages of the module explained the utility valuation task, and provided an overview of MESSC. Following this, participants completed the standard gamble health state valuation exercises.[167]

Health states were derived from the EORTC item bank and were presented as a point-form description of five dysfunctional attributes: dependent (D), non-ambulatory (N), incontinent of urine (I), pain (P), and “other symptoms” (S). Each attribute was characterized by the presence (+) or absence (-) of the dysfunctional attribute. Dependence was described as “you need help with eating, dressing, washing yourself or using the toilet. You are dependent on others.” Other symptoms were described as “You have one or more of: nausea, vomiting, shortness of breath, lack of appetite, diarrhea.” Valuations were obtained using standard gamble method.[243]

In this study, “perfect” health is represented by the fully functional health state (D-, N-, I-, P-, S-). To eliminate any bias introduced by respondents assuming different life expectancies for each scenario, all scenarios were framed as having a certain life expectancy of five years; that is, for both the probe health scenario, and perfect health, participants were told their life expectancy would certainly be five years.(13) Five years was the maximum survival reported in a randomized controlled trial on treatments for MESSC.[171]

Participants were asked to value six health states. Two health states formed a “test pair.” These pairs shared non-dysfunctional attributes (eg. D-, N+, I-, P-, S- and D-, N+, I-, P-, S+). The other four health states were selected at random.

Statistical Analysis

We deemed participants to have misunderstood the task, or not engaged, if: they provided an illogical valuation for the test pair; or provided the same valuation for more than 4 health states. Such participants were excluded from the analysis. A sensitivity analysis including all participants is provided in Appendix B.3.

Utility values lie between 0 and 1 with 0 representing death and 1 perfect health. Therefore beta regression analysis was chosen to explore the relationship between health

attributes and utility valuation (Appendix B.1).[67] Regression was performed on the mean (μ) of the beta model. The dispersion parameter (σ) was treated as a constant.[67] To account for the presence of six observations per participant (indexed by i), we incorporated a random intercept term for each participant (ϵ_i) in the model for μ . [78]

The SOAP MESCC module is designed for *ex ante* (without experience) utility valuation from a general population sample. These utilities should reflect population preferences for the health states and were used to compute expected (mean) quality-of-life changes in economic analysis.

Within this context, the objective of the analysis was to estimate mean utilities for the population, not predict individual utility valuations. Consequently, goodness-of-fit measures such as the R^2 statistic or root mean squared (RMSE) are not appropriate because they quantify how well the model fits individual observations.[174] To instead quantify the performance of the regression model in estimating the mean utility for the 31 health states relative to direct estimation of mean utilities, we used the absolute agreement intraclass correlation coefficient (ICC). For this ICC calculation, health states were targets and utilities were ratings.[203] A two-way model was used because ratings can only be obtained by direct estimation or regression estimation.

In an effort to strengthen the generalizability of the regression analysis, we implemented internal validation by allocating participants to a test set and validation set in a 1 : 1 ratio.[213] Regression models were fit using the test set, and the optimal model was identified by jointly the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).[125] The optimal regression model was used to compute mean utilities for each health state (regression estimation of means). Mean utilities for each health state were then directly estimated using the validation set (direct estimation of means). The performance of the optimal regression model was quantified using the ICC by comparing mean utilities obtained from the regression model to mean utilities obtained by direct estimation.

Four regression models were considered. In Model 1, each dysfunctional attribute was coded as a categorical variable. In Model 2, the number of dysfunctional attributes was coded as nominal categorical variables. This strategy was used because it avoided assumptions of linear or extra-linear effects. Model 3 incorporated all first-order interactions in Model 1. Model 4 combined Model 1 and Model 2.

Before proceeding with model selection, we graphically checked all four models for misspecification of the variance and link function.[149] We attempted to simplify the optimal model using standard regression model building procedures.[100] These procedures ultimately led us to consider two additional models. In Model 5, the square root of the number dysfunctional attributes (num dys) was included as a continuous predictor. In Model 6, the natural logarithm of the number dysfunctional attributes (num dys) was included as a continuous predictor. The formula for the beta regression linear predictor for each model is outlined in Table B.2.

Sample Size Calculation

The sample size for the regression analysis was based on a commonly cited rule-of-thumb that linear regression requires 10 to 20 subjects per variable.[86] Our regression models considered a total of 21 variables: four numerical variables, five categorical variables, ten interaction terms, and two intercepts. Following this rule of thumb, we would require $15 \times 21 = 315$ subjects. As we planned to split into a test and validation set in a 1 : 1 ratio, we required 630 subjects. However, to ensure at least 2 participants in each quota, 650 participants were required (Table B.1).

3.3 Results

1138 panel members logged into the SOAP MESCC module and provided consent. However, 488 were excluded for not valuing all health states, providing an illogical valuation for the test pair, or providing invariant responses (Figure 3.1). Of the 1029 participants who valued six health states, 379 (36.8%) were excluded for not having understood or engaged in the task.

The socio-demographic characteristics of 650 included and 488 excluded participants are shown in Table 3.1. There were no statistically or qualitatively significant differences between the groups in terms of sex, age, or province of residence.

Mean valuations for the 31 health states tended to decrease with an increasing number of dysfunctions (Table 3.2). For example, the mean utility valuation for 1, 2, 3, 4, and 5 dysfunctions was 0.691, 0.581, 0.471, 0.364, and 0.333 respectively (p -value for trend <0.0001).

The functional form of each model and example mean utility calculations are provided in Appendix B.3. Calibration and agreement parameters for all candidate models are provided in Table B.4 and Figures B.4, B.5, B.6, 3.2, and B.7. The model selection procedure identified the square root of the number of dysfunctions to be the optimal predictor of mean utility based on both the AIC and BIC criteria (Table B.3).

The coefficients for Model 4 are given in Table 3.4. This model was well calibrated, estimating mean utilities in the validation set with a mean absolute error of 0.047 and mean signed difference of -0.02 (both of which are less than the MCID of 0.05 for utilities)(Figure 3.2).[175, 246] There was excellent agreement between mean utilities obtained from this regression model and mean utilities obtained by direct estimation with an ICC of 0.936 (95% CI: 0.86, 0.97).[129]

Mean and median utility valuations were not sensitive to the inclusion of all participants (Table B.7).

Formulas for the linear predictor in the beta regression are provided in Table B.2. Formulas for mean utility valuation for a given MESCC health state using the fitted models are provided in Table B.6. To illustrate the use of these formulas, we will calculate mean utilities for the health state D-, N+, I-, P+, S- using the optimal Model 4. This health

state is coded as: num dys=2. Therefore using Model 4 the mean utility valuation is calculated as:

$$\text{inverse logit} \left(1.71 - 1.11 \times \sqrt{2} \right) = 0.53.$$

3.4 Discussion

In this paper we report directly valued utilities for MESCC health states using the validated SOAP MESCC module.[167] The study sample was representative of the population in all English-speaking Canadian provinces. Our findings are valid with 63.2% of participants having understood and engaged in the task. This rate is superior to validity rates reported in general population direct valuation studies for the EuroQol-5D.[12, 128, 114, 49]

The regression model building exercise revealed that members of the general population value all dysfunctions characterizing MESCC health states equally. Furthermore, dysfunction follows a pattern of diminishing marginal disutility. That is, each additional dysfunction effects a smaller incremental change in utility than the previous dysfunction. These results demonstrate that from the societal perspective, physical function is valued equal to other facets of well-being.

It must be recognized that *ex post* utilities are not equivalent to *ex ante* utilities obtained from patients who have experienced the health states.[156] In part due to adaptation, patients tend to provide higher valuations for health states which predominantly affect physical health than the general population for the same health state. The *ex ante* utilities collected in this study are highly appropriate for facilitating healthcare decision making and can be used to conduct decision analysis and cost-utility analysis for MESCC.[54]

Although it may seem that applying lower *ex ante* utilities may infringe on patient autonomy and deny care, healthcare system decision making impacts patients with various conditions. If the objective of healthcare decision making is to maximize the benefit of all patients, utilities across different disease must be comparable in order to set priorities. Rawles argues that *ex ante* utilities can be used ethically if valued under a “veil of ignorance.”[187] If we assume that the general population providing *ex ante* utility valuations may eventually develop the condition of interest, out of self-interest, they should provide fair valuations. Although *ex ante* utilities are theoretically restricted to system policy decisions, *ex ante* utilities have become the *de facto* standard for individual patient decisions. Utilities obtained from generic health surveys such as the EuroQol-5D, Short Form-6D, and Health Utilities Index 3 are actually *ex ante* valuations.[55] Therefore we have chosen to evaluate *ex ante* SOAP MESCC module to conform with conventions in the literature.

Utility valuations for a single health state were highly variable across participants. This was evidenced by wide Inter-quartile ranges. However, since our objective was to measure general population *ex ante* utilities for health policy decision making, the expected values and the underlying uncertainty represented by the 95% confidence intervals are pertinent.

Health policy decisions concern the allocation of scarce resources to health programs with superior cost-effectiveness ratios at the expense of those with inferior cost-effectiveness

ratios.[55] If cost-effectiveness ratios are uncertain, there is a risk that a wrong policy decision will be made (ie. funding the program with an inferior cost-effectiveness ratio). Even if the risk of a wrong policy decision is large (ie. great uncertainty in cost-effectiveness ratio), from both ethical and economic perspective a decision should still be made.[35] From an ethical perspective, deferring a decision (ie. not funding any health program under consideration) denies care to several groups of patients. From the economic perspective, convincing arguments have been given to defend the practise of expected value decision making – i.e. the irrelevance of inference with respect to public decision making.[5] For these reasons, unbiased estimates of costs and effects are more important than precise estimates.

The results of the regression model building exercise have relevance for clinicians counseling MESCC patients regarding treatment. As dysfunctions follow a pattern of diminishing marginal disutility, each dysfunction is valued equal to the others – it is the total number of dysfunctions that drive quality-of-life. Surgeons should be cognizant that ambulation and continence, which are dysfunctions addressed by surgery, are no more important than other attributes (pain, other symptoms, and level of independence). Furthermore, attempting to reverse a single dysfunction in a patient with high functional status will lead to a greater increase in quality-of-life relative to a patient with low functional status.

To our knowledge, this is the first study reporting direct utility valuation for a spinal disorder. Utilities for spine disease specific instruments such as the NDI, ODI, and SRS-22 have been developed using an indirect “cross-walk” protocol.[193, 31, 32, 19] Patient responses are collected using both the disease specific instrument and a generic health measure (eg. EuroQol-5D, Short Form-6D, and Health Utilities Index 3). Next, a regression model is developed to relate the disease specific score to the generic health measure. Then the regression model estimated generic health measure score can be used to compute a utility. However, utilities obtained from generic health measure scores are actually computed from another regression model relating the generic health measure score to directly valued utilities.[55] Our study demonstrates that disease specific direct utility valuation is feasible and valid. Investigators may consider applying our study protocol and the validated SOAP tool to other disease contexts. This approach may eliminate the potential error of the cross-walk approach introduced by the need for two regression models.

3.5 Conclusion

We provide a comprehensive set of *ex ante* utility estimates for MESCC health states. The utility values derived from this study can be used to help inform population level healthcare decision making, such as allocation of limited resources for specific treatments. The results of this study can also help clinicians counsel MESCC patients regarding treatment.

This is the first study reporting direct utility valuation for a spinal disorder. We demonstrate that direct utility valuation over the internet is a feasible solution for obtaining quality-of-life data when generic health status questionnaire data is lacking. Investigators may consider applying our approach to other disease contexts.

3.6 Figures

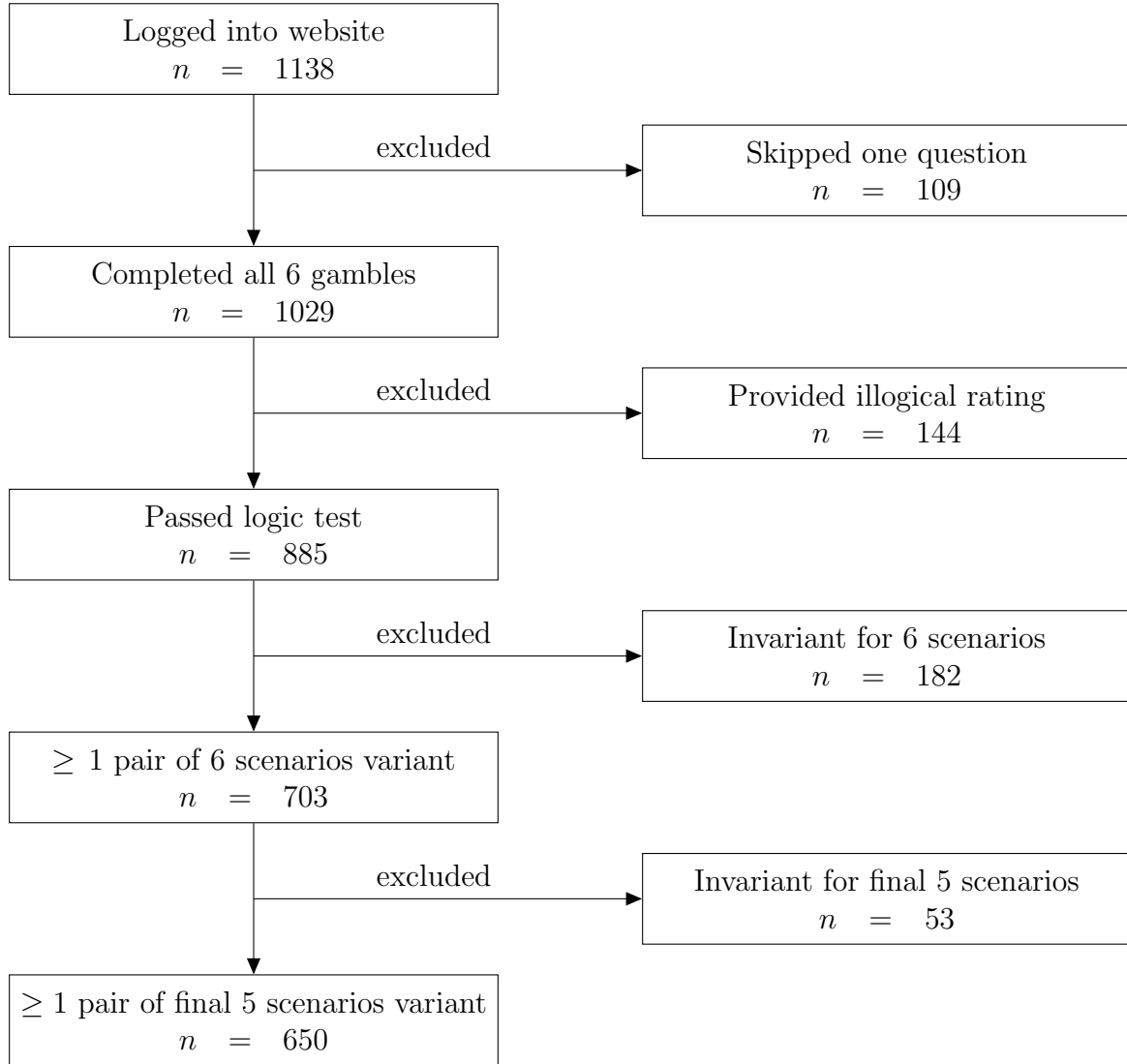
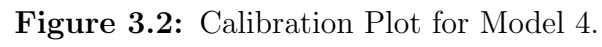


Figure 3.1: Flow of participants.



3.7 Tables

Table 3.1: Participant Demographics

		Included N = 650	Excluded N = 488
Sex – no. (%)			
	Female	331 (51)	234 (48)
	Male	319 (49)	254 (52)
Age – no. (%)			
	18 – 24 yrs	78 (12)	41 (8)
	25 – 34 yrs	113 (17)	76 (16)
	35 – 44 yrs	106 (16)	83 (17)
	45 – 54 yrs	118 (18)	91 (19)
	55 – 64 yrs	107 (17)	79 (16)
	≥ 65 yrs	128 (20)	118 (24)
Province – no. (%)			
	Maritimes	57 (9)	40 (8)
	Ontario	326 (50)	265 (54)
	Manitoba	30 (5)	27 (6)
	Saskatchewan	26 (4)	9 (2)
	Alberta	98 (15)	65 (13)
	British Columbia	113 (17)	82 (17)

Table 3.2: Utilities for all MESCC Health States

Health State	Number of Dysfunctions	N	Mean (95% CI)	Median (IQR)
D+N-I-P-S-	1	181	0.676 (0.628, 0.725)	0.75 (0.45, 0.95)
D-N+I-P-S-	1	185	0.717 (0.671, 0.763)	0.85 (0.55, 0.95)
D-N-I+P-S-	1	187	0.681 (0.632, 0.73)	0.85 (0.45, 0.95)
D-N-I-P+S-	1	182	0.695 (0.647, 0.743)	0.85 (0.512, 0.95)
D-N-I-P-S+	1	202	0.685 (0.639, 0.732)	0.85 (0.462, 0.95)
D+N+I-P-S-	2	88	0.548 (0.474, 0.623)	0.65 (0.225, 0.9)
D+N-I+P-S-	2	108	0.619 (0.554, 0.684)	0.75 (0.4, 0.95)
D+N-I-P+S-	2	106	0.542 (0.475, 0.608)	0.575 (0.25, 0.85)
D+N-I-P-S+	2	112	0.6 (0.534, 0.666)	0.7 (0.35, 0.95)
D-N+I+P-S-	2	88	0.659 (0.589, 0.728)	0.75 (0.45, 0.95)
D-N+I-P+S-	2	107	0.57 (0.505, 0.635)	0.65 (0.325, 0.875)
D-N+I-P-S+	2	105	0.513 (0.44, 0.586)	0.6 (0.1, 0.9)
D-N-I+P+S-	2	100	0.583 (0.514, 0.652)	0.65 (0.35, 0.9)
D-N-I+P-S+	2	89	0.601 (0.525, 0.676)	0.65 (0.3, 0.95)
D-N-I-P+S+	2	113	0.582 (0.514, 0.65)	0.65 (0.25, 0.95)
D+N+I+P-S-	3	81	0.512 (0.428, 0.596)	0.55 (0.1, 0.9)
D+N+I-P+S-	3	76	0.443 (0.363, 0.524)	0.45 (0.1, 0.75)
D+N+I-P-S+	3	80	0.492 (0.413, 0.572)	0.5 (0.075, 0.75)
D+N-I+P+S-	3	88	0.445 (0.367, 0.524)	0.45 (0, 0.75)
D+N-I+P-S+	3	96	0.491 (0.415, 0.566)	0.5 (0.1, 0.85)
D+N-I-P+S+	3	88	0.479 (0.398, 0.56)	0.6 (0, 0.813)
D-N+I+P+S-	3	86	0.478 (0.401, 0.554)	0.525 (0.1, 0.75)
D-N+I+P-S+	3	81	0.407 (0.329, 0.485)	0.45 (0, 0.75)
D-N-I+P+S+	3	91	0.501 (0.426, 0.576)	0.5 (0.15, 0.85)
D-N+I-P+S+	3	81	0.457 (0.374, 0.539)	0.45 (0, 0.8)
D+N+I+P-S+	4	192	0.408 (0.357, 0.46)	0.35 (0, 0.75)
D+N+I+P+S-	4	199	0.359 (0.308, 0.41)	0.25 (0, 0.65)
D+N+I-P+S+	4	191	0.342 (0.292, 0.391)	0.25 (0, 0.65)
D+N-I+P+S+	4	173	0.357 (0.303, 0.411)	0.25 (0, 0.65)
D-N+I+P+S+	4	170	0.351 (0.297, 0.404)	0.25 (0, 0.65)
D+N+I+P+S+	5	174	0.333 (0.281, 0.384)	0.2 (0, 0.638)

95% Confidence Interval (95% CI). Interquartile range (IQR).

Table 3.3: Selection Parameters for all Models

	Description	AIC	BIC
Model 1	Attributes	-2750.24	-899.148
Model 2	Number of dysfunctions (nominal variable)	-2751.78	-906.259
Model 3	Attributes and second-order interactions	-2747.42	-840.572
Model 4	Square root of number of dysfunctions (continuous variable)	-2756.85	-928.061
Model 5	Logarithm of number of dysfunctions (continuous variable)	-2754.9	-926.11

Table 3.4: Summary of Beta Regression Analysis for Model 4

	Variable	Coefficient	SE	<i>t</i> -value	<i>p</i> -value
Mean (μ) coefficients	Intercept	1.7058	0.11594	14.71	<0.0001
	Square root of number of dysfunctions	-1.11247	0.07175	-15.51	<0.0001
Dispersion (σ) coefficient	Intercept	0.74727	0.02139	34.94	<0.0001

Chapter 4

Back to Bayesian: a Strategy to Enhance Prognostication of Metastatic Spine Disease

Markian A. Pahuta¹, Joel Werier², Eugene K. Wai^{2,3}, Carl van Walraven^{3,4,5}, Doug Coyle³

- 1 – Divisions of Orthopaedic & Neurological Surgery, Henry Ford Health System, Detroit, Michigan, USA
- 2 – Division of Orthopaedic Surgery, The University of Ottawa, Ottawa, Ontario, Canada
- 3 – School of Epidemiology and Public Health, The University of Ottawa, Ottawa, Ontario, Canada
- 4 – The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
- 5 – The Institute for Clinical Evaluative Sciences, Ottawa, Ontario, Canada

Context

This chapter addresses skepticism regarding decision-analysis. In this chapter I review Bayesian statistics which I believe can address the concern that probabilistic analysis is arbitrary and prone to manipulation. This is because the output of a Bayesian analysis is a posterior distribution which can be used in place of user defined distributions in economic analysis. Bayesian analysis will be used to conduct QALY decision analysis in Chapter 6.

Abstract

Clinicians must consider prognosis when offering treatment to patients with spine metastases. Although several prognostic indices have been developed and validated for this purpose, they may not be applicable in the current era of targeted systemic therapies.

Even before the introduction of targeted therapies, these prognostic indices should not have been directly used for individual patient decision making without contextualizing with other sources of data. By contextualizing, we mean that prognostic estimates should not be based on these scores alone and should formally incorporate clinically relevant factors not part of prognostic indices. Contextualization requires the use of Bayesian statistics which may be unfamiliar to many readers.

In this paper we show readers how to correctly apply prognostic scores to individual patients using Bayesian statistics. Through Bayesian analysis, we explore the impact of new targeted therapies on prognostic estimates obtained using the Tokuhashi score.

We provide a worked calculation for the probability a patient surviving to 6 months using dichotomous prognostication. We then demonstrate how to calculate a patient's expected survival using continuous prognostication. Sensitivity of the posterior distribution to prior assumptions is illustrated through effective sample size adjustment.

When the predicted prognosis from the Tokuhashi score is contextualized with data on contemporary systemic treatments, patients previously deemed non-surgical candidates may be eligible for surgery.

Bayesian prognostication generates intuitive results and allows multiple data points to be synthesized transparently. These techniques can extend the usefulness of existing prognostic scores in the era of targeted systemic therapies.

4.1 Clinical Scenario

Tom is a 60 year-old man with Stage IV non-small-cell lung cancer (NSCLC). He has metastases to four vertebral bodies with no epidural disease. He is treated with platinum-based chemotherapy cisplatin and radiotherapy. Four months into treatment, the disease at T5 progresses and Tom develops an incomplete spinal cord injury from epidural disease. As a result, Tom experiences a functional decline in his Karnofsky Performance Score to 70 and is admitted to hospital.

The consulting spine surgeon, radiation oncologist, and medical oncologist all agree that direct decompressive surgery and stabilization would be the most effective treatment for neurologic dysfunction.[171] However, Tom’s Tokuhashi Score is 6, which is associated with an expected survival of less than 6 months.[228] Furthermore, neither the spine surgeon nor the radiation oncologist has seen a patient like Tom (with stage IV NSCLC refractory to first-line treatment) survive beyond 3 months. Therefore, both the spine surgeon and radiation oncologist recommend against surgery because Tom’s prognosis is poor.

However, contemporary trials have reported improvement in overall survival for patients with advanced NSCLC refractory to first-line treatment with targeted therapies. Therefore, the medical oncologist disagrees with the spine surgeon and medical oncologist because she/he plans to give Tom targeted therapy and therefore expects Tom to survive longer than 6 months.

The three clinicians plan to meet to resolve their disagreement.

4.2 Introduction

Surgical treatment of metastatic spine disease is effective for maintaining and improving patients’ physical function and health-related quality-of-life.[171, 65] However, benefit from a surgical intervention requires that patients survive longer than the time needed to recover and rehabilitate from surgery. Surgeons typically believe that patients should have an expected survival of at least 3 – 6 months to be surgical candidates.[227]

Several prognostic indices have been developed and validated for estimating expected survival of patients with metastatic spine disease.[172, 132, 211] The Tokuhashi Score, widely used survival index for these patients, considers performance status, extra-spinal bone metastases, vertebral metastases, visceral metastases, pathologic diagnosis, and neurologic function.[228] The Tokuhashi score has good performance in discriminating between patients who die prior to or survive beyond 6-months.[64]

However, some authors have pointed out that existing prognostic scores are not useful in the era of targeted therapies, in particular for non-small cell lung cancer (NSCLC). The Tokuhashi Score, for example, was developed in 1990 and revised in 2005.[229, 228] During this time-period, only cytotoxic agents were available for the treatment of NSCLC, and consequently patients with advanced cancer had a median overall survival was 4 – 6 months. In the contemporary period, patients with advanced NSCLC may be offered

targeted therapy such as antiangiogenic drugs which has been shown to improve overall survival in NSCLC patients.[185]

However, even before the introduction of targeted therapies, the Tokuhashi Score should not have been directly used for individual patient decision making without contextualizing with other sources of data. By contextualizing, we mean that estimates of prognosis should not be based on the Tokuhashi score alone: prognosis should be estimated from multiple data points. This is because the Tokuhashi Score does not consider all clinically important prognostic factors.

Failing to contextualize statistical results, such as the Tokuhashi Score, is a common mistake.[41] This is because contextualization requires Bayesian statistics – a form of statistics that may be unfamiliar to many readers. This form of statistics that actually predates the more commonly used Frequentist statistics (the branch of statistics underpinning confidence intervals and p -values) and is distinguished by the fact that it incorporates multiple sources of information into statistical analysis.[68] Although the term “Bayesian” may be new to many readers, its concepts are likely familiar to many readers as they are also used for diagnostic test interpretation. In essence, the Bayesian statistical approach that integrates new information with previously known information.[68]

In this paper we aim to show readers how to correctly apply the Tokuhashi Score to individual patients using Bayesian statistics. We illustrate the impact of new targeted therapies on prognostic estimates. In Section 4.3, we will determine the probability of Tom surviving to 6 months (dichotomous prognostication). In Section 4.4, we will determine Tom’s expected survival using continuous prognostication. We then conclude the paper by reviewing the debate between Frequentist and Bayesian statistics and its implications on clinical decision making.

4.3 Dichotomous Prognostication

Returning to the clinical scenario, the medical oncologist decides to argue her/his point by estimating the probability of Tom surviving ≥ 6 -months with targeted therapy (docetaxel combined with ramucirumab, DOC+RAM) versus standard therapy (docetaxel alone, DOC). This is an example of dichotomous prognostication because there are only two mutually exclusive outcomes: ≥ 6 -month survival and < 6 -month survival. For Tom, dichotomous prognostication is the process of updating beliefs on the probability of ≥ 6 -month survival with the different treatments (prior probability, PriorPr) with the Tokuhashi Score (converted to a likelihood ratio, LR) to yield a new belief on the probability of ≥ 6 -month survival (posterior probability, PostPr).[2]

A search of the literature reveals that no studies have recruited patients that exactly match Tom’s pathologic diagnosis and disease stage. However, the REVEL investigators performed a planned subgroup analysis in patients like Tom: those with refractory NSCLC within 9-months of starting first-line treatment.[188] Therefore this analysis has the best external validity for the clinical scenario at hand.[186] Six-month survival probabilities for this subgroup are not reported in the main paper; however, we can determine this

probability by digitizing the survival curve shown in supplementary Figure S1a.[92] From the digitized survival curve, we determine that the 6-month survival in the DOC+RAM group was 68%, while in the DOC group 6-month survival was 56

Prognostic performance of the Tokuhashi Score have been reported in several cohort studies and a meta-analysis conducted by Lee and colleagues.[133] We should use the highest level of evidence provide by the meta-analysis reported by Lee *et al.* The authors reported that a Tokuhashi Score of ≥ 8 has a sensitivity and specificity of 0.657 and 0.676, respectively, for ≥ 6 -month survival. Tom has a negative test result because his Tokuhashi Score is ≤ 8 . Therefore, a negative likelihood-ratio (LR)¹ is required:

$$\begin{aligned} \text{LR} &= \frac{1 - \text{Sensitivity}}{\text{Specificity}} \quad [89] \\ &= \frac{1 - 0.676}{0.657} \\ &= 0.493. \end{aligned}$$

The LR quantifies how much more likely a Tokuhashi Score < 8 is found in patients who survive ≥ 6 -months compared to those who survive < 6 -months. We must emphasize that the LR *does not* quantify how much more likely a ≥ 6 -month survival is found in patients with a Tokuhashi Score < 8 compared to those with a Tokuhashi score of ≥ 8 . Therefore the LR provides information in a way that is “backwards” to what is needed for clinical decision making – the spine surgeon and oncologist require the likelihood of ≥ 6 -month survival, not the likelihood of a Tokuhashi Score of < 8 .

The posterior probability can be calculated from the prior probability and *LR* using the formula

$$\text{PostPr} = \frac{\text{PriorPr} \times \text{LR}}{1 - \text{PriorPr} + \text{PriorPr} \times \text{LR}}$$

For DOC only treatment we compute

$$\begin{aligned} \text{PostPr} &= \frac{0.56 \times 0.493}{1 - 0.56 + 0.56 \times 0.493} \\ &= 0.386. \end{aligned}$$

Repeating the same steps for DOC+RAM treatment generates a posterior probability of 0.512.

These calculations indicate that while a Tokuhashi Score of < 8 generally predicts a < 6 -month survival, the score should be integrated with other pieces of information more generalizable to the clinical problem at hand. Given the effectiveness of DOC+RAM, a negative Tokuhashi Score does not translate to a 6-month survival probability $< 50\%$. In contrast, DOC is less effective and a negative Tokuhashi Score results in a survival probability less than 50%.

¹If Tom had a positive test result, we would instead compute the positive likelihood ratio $\text{LR} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$

Based on this analysis, the medical oncologist concludes that if treated with DOC+RAM, Tom is a surgical candidate because he has a greater than 50% probability of surviving ≥ 6 -months. The spine surgeon and radiation oncologist are not entirely convinced and request more analysis.

4.4 Continuous Prognostication

Tom, like many cancer patients, has difficulty interpreting and applying probabilities.[79, 74, 141] Like many cancer patients, Tom has also expressed that he wishes to know the typical survival experience for patients like him, that is, he wants to know the worst and best case scenarios.[93, 215] The medical oncologist would like to present Tom with his expected median survival time because it is a less biased measure than mean survival time.[202] To communicate the best and worst case scenarios, the medical oncologist would like to present range of values which she/he is 95% confident the true median survival lies. The medical oncologist also believes this type of analysis will satisfy the spine surgeon and radiation oncologist.

The medical oncologist now wishes to conduct *continuous prognostication*. Recall that the outcome of interest for dichotomous prognostication could only take on two values (≥ 6 -month survival: yes or no). In contrast the median survival is a continuous variable, that is, it can theoretically take on an infinite number of values between 0 and infinity. We can apply the same studies used for dichotomous prognostication (the REVEL trial and the Lee *et al.* meta-analysis).[188, 133]

To perform continuous prognostication for Tom, beliefs on the median survival time reported by the REVEL trial need to be updated with the median survival time associated with a Tokuhashi Score of ≥ 8 to yield a new belief on the median survival time. A prior probability is needed for each potential median survival time; this collection of prior probabilities is called a *prior distribution*. A likelihood ratio is also needed for each of the potential median survival times; this information is contained in a *likelihood distribution*. The updated beliefs on median survival time are contained in the *posterior distribution*.

To develop a prior distribution from the REVEL subgroup analysis, we require an estimate of the median survival time and standard deviation for each treatment group. From supplementary Figure S1a, we find the median survival time was 7.0 months (95% CI, 6.1 – 8.5 months) for the DOC group; while for the DOC+RAM group, median survival was 9.3 months (95% CI, 8.4 – 10.1 months). These confidence intervals are asymmetric, therefore additional calculations are needed to calculate standard deviation (Appendix C.1).[25] The standard deviation for the DOC group is 0.68, while the standard error for the DOC+RAM group is 0.42.

To develop a likelihood distribution for Tokuhashi Score ≥ 8 , we require an estimate of the median survival time and standard deviation for this group of patients. This analysis was not reported in the publication, but was provided by the authors on request.[133] Using individual patient analysis, Lee and colleagues determined that the median survival and

95% CI for a Tokuhashi Score ≥ 8 was 5.0 ± 0.6 months. Therefore the standard deviation is $0.6 \div 1.96 = 0.33$.

It is important for readers to appreciate that the meaning of confidence intervals from the REVEL trial and Lee *et al.* meta-analysis. These intervals ensure that 95% of similarly computed intervals computed in an infinite number of times would contain the true median survival.[200] The qualifier 95% applies to the technique of computing confidence intervals and not the true expected survival.[148] Confidence intervals do not indicate the range within which we are 95% confident the true median survival lies – such an interval is a credible interval. Much in the same way that the likelihood ratio provides information in a way that is “backwards” to what is needed for clinical decision making, the confidence interval provides information in a “backwards” to a credible interval. Bayesian continuous prognostication is required to compute a credible interval for the median survival.

The posterior distribution can be easily derived if we assume the estimates of median survival time follow a normal (or bell-curve) distribution.[209] We denote the median survival time and standard deviation from the REVEL data as m and s respectively. We denote the median survival time and standard deviation from the Lee *et al.* meta-analysis as μ and σ respectively. Through exact mathematical analysis it can be shown that the mean of the posterior distribution for median survival time is

$$\frac{\frac{m}{s^2} + \frac{\mu}{\sigma^2}}{\frac{1}{s^2} + \frac{1}{\sigma^2}}$$

and the standard deviation is

$$\sqrt{\frac{1}{\frac{1}{s^2} + \frac{1}{\sigma^2}}}.$$

Substituting the data for the DOC treatment arm and Tokuhashi Score ≥ 8 yields a posterior distribution with mean

$$\frac{\frac{7.0}{0.68^2} + \frac{5.0}{0.33^2}}{\frac{1}{0.68^2} + \frac{1}{0.33^2}}$$

and standard deviation

$$\sqrt{\frac{1}{\frac{1}{0.68^2} + \frac{1}{0.33^2}}}.$$

We can calculate a 95% credible interval as $5.38 \pm (0.09 \text{ times } 1.96)$ months. In summary, Tom’s estimated median survival with DOC alone is 5.38 months with a 95% credible interval from 5.20 to 5.56 months. By repeating the same calculations for DOC+RAM we determine that Tom’s expected median survival is 6.64 months with a 95% credible interval from 6.50 to 6.78 months

The typical, worst and best case scenarios are all under 6 months if Tom is treated with DOC alone. But if treated with DOC+RAM, Tom’s typical, worst and best case scenarios are all over 6 months. Based on the expectation that recovery from surgery can take over 3 months, the medical oncologist concludes that if treated with DOC+RAM, surgery is a

reasonable proposition because even in the worst-case scenario, expected survival is over 6 months.

However, the spine surgeon and radiation oncologist are not convinced. They have reservations regarding the generalizability of the REVEL trial to Tom’s individual case because the REVEL investigators did not report the number of patients with vertebral metastases in the study sample. The medical oncologist suggests that a sensitivity analysis would demonstrate the robustness of conclusions based on the REVEL trial data, and would address concerns regarding generalizability.

By questioning the the generalizability of the REVEL trial, the spine surgeon and radiation oncologist are implicitly discounting the REVEL trial to the degree that the upper limit of the resulting 95% credible interval is less than and does not include 6 months. To make their discounting explicit, the medical oncologist conducts a “what-if” analysis by re-calculating the posterior distribution for the DOC+RAM treatment multiple times with decreasing effective sizes, but keeping all other parameters constant (Appendix C.2). She/he determines that only if the REVEL subgroup analysis for the DOC+RAM treatment arms consisted of ≤ 64 patients would the 95% credible interval for Tom’s prognosis not include 6 months.

The medical oncologist explains that this exercise formally quantified the degree to which the other two clinicians are implicitly discounting the generalizability of the REVEL trial. In other words, the spine surgeon and radiation oncologist discounted the REVEL trial by 83%. Although the REVEL trial may not be fully generalizable to Tom’s case, the spine surgeon and radiation oncologist do not believe the REVEL trial should be discounted by 83%.² Consequently, all three clinicians agree that Tom should be offered surgery if he receives DOC+RAM treatment.

4.5 Discussion

The motivation for this article was to enhance prognostication for patients with metastatic spine disease in an era of more effective systemic treatments. We introduced Bayesian statistics and sought to demystify this statistical approach. In this article, we focused on the principles and application of Bayesian analysis rather than the mechanics of Bayesian computation. Without knowledge of detailed computational steps, some readers may regard Bayesian statistics as a “black box.” Interested readers are directed to excellent textbooks on the subject.[209, 153, 30, 124] Dichotomous prognostication requires several data-points:(i) a prior survival probability at a time of interest; and (ii) the sensitivity and specificity of the prognostic factor on interest. If a only a survival probability, but not sensitivity and specificity for a prognostic factor of interest are reported, beta-binomial conjugate analysis can be used.[135] Continuous prognostication requires:(i) a prior belief on median survival time and confidence interval; and (ii) median survival time and confidence interval for a prognostic factor of interest.

² $1 - \frac{64}{384} = 0.8333$

We took care to highlight that confidence intervals do not quantify the uncertainty in a value such as median survival time. Although confidence intervals are often used for this purpose, this is because confidence intervals are frequently misinterpreted by the “confidence-level misconception” to be Bayesian credible intervals.[28] Under this misconception, a 95% confidence interval is interpreted to contain the true value with a probability of 95%.[41] Cumming et al. asked 263 researchers in psychology, behavioural neuroscience, or medicine to interpret confidence intervals and found that over 40% suffered from the confidence-level misconception.[41] This misconception was actually anticipated in 1937 by the developers of the confidence interval.[155] As even practicing researchers have difficulty interpreting confidence intervals, the journal *Basic and Applied Social Psychology* recently “banned” the use of confidence intervals in new submissions.[233] The prevalence of the confidence-level misconception suggests that for many individuals the Bayesian credible interval is more intuitive construct for expressing the uncertainty in a reported value. Therefore, we believe Bayesian statistics can enhance clinician-patient communication.

In this article, we focused on the principles and application of Bayesian analysis rather than the mechanics of Bayesian computation. Without knowledge of the computational steps, some readers may regard Bayesian statistics as a “black box.” Interested readers are directed to Appendix C.1 and excellent textbooks on the subject.[209, 153, 30, 124] Without delving into Bayesian computation, we believe readers can gain sufficient understanding of Bayesian analysis through the analogy with diagnostic testing. It is well accepted that diagnosis should be based on the post-test probability of disease rather than the sensitivity and specificity of a test.[110] For similar statistical considerations, credible intervals should be used instead of confidence intervals. Diagnostic test interpretation will be incorrect if the pre-test probability is inappropriate, similarly, credible intervals will be incorrect in the prior probability is inappropriate.

The sensitivity of the posterior distribution to the choice of prior distribution is an important reason for the lack of uptake of Bayesian methods. R.A. Fisher, one of the founders of modern Frequentist statistical methods, opposed Bayesian statistics and characterized the approach as: “...extremely arbitrary... by evolving a vitally important piece of knowledge, that of the exact form of the distribution of [the value of interest], out of an assumption of complete ignorance...”[72] We exploited the sensitivity of the posterior distribution to choice of prior distribution in a positive way for effective sample size sensitivity analysis. However, arbitrary and/or improper selection of prior distributions negatively impacts on Bayesian analysis.

Lack of public trust in science illustrates these issues regarding the prior distributions. Despite scientific evidence on issues such as safety of childhood vaccines, safety of genetically modified foods, and effectiveness of sexual health education these topics remain controversial.[96] Although individual psychology may predispose individuals to embracing controversy, access to “prior distributions” is an important determinant.[139] Controversy is stoked by consumption of selective and biased media.[24] Whereas controversy is lessened when scientific consensus is communicated in a coherent and unified manner.[140] Just as the public is cautioned to consume media critically and identify “fake news,”[150] readers should critically appraise prior distributions before accepting a Bayesian analysis. To mitigate manipulation of the posterior distribution, it has been suggested that

researchers repeat their analysis with several different prior distributions that capture the spectrum of reasonable clinical beliefs.[104] Consequently, readers must assess the sensitivity of results to prior distribution choice, and make a judgment on whether the results are sufficiently robust to make conclusions. Sensitivity analysis was illustrated in Section 4.4 and Appendix C.2 when the effective sample size was adjusted.

We argue that for clinical scenarios with sparse evidence, such as prognostication in metastatic spine disease in the era of targeted therapies, a non-Bayesian approach may lead to suboptimal decision making. A nihilistic course of action based on non-Bayesian analysis would be to ignore recent advances in systemic treatment because their effectiveness for patients with metastatic spine disease has not been formally studied. An overly optimistic course of action based on non-Bayesian analysis would be to assume the results of oncology trials enrolling few patients with metastatic spine disease fully apply to patients with metastatic disease. In this paper we demonstrated how effective sample size sensitivity analysis quantifies the impact of clinicians' assumptions and biases on decision making, and therefore, can make decisions more transparent. In this manner, different – and potentially not fully compatible – sources of evidence can be synthesized for decision making in a rational and transparent way.

It is not known whether Bayesian prognostication for metastatic spine disease is superior to informal clinical decision making.[33] However, it has been shown that informal prognostication for metastatic spine disease is superior to prognostication based on the Tokuhashi Score alone.[126] While for the diagnosis of pulmonary embolism, integrating clinical gestalt with a clinical prediction rule (a Bayesian strategy) is superior to relying solely on a clinical prediction rule.[173] Given the effectiveness of Bayesian clinical decision making in other clinical contexts, and the limitations of existing decision making strategies, Bayesian prognostication for metastatic spine disease should be formally studied.

In this article we present the correct statistical procedures for prognostication. This approach uses Bayesian statistics which is intuitive and can transparently synthesize evidence. Bayesian prognostication can enhance counseling for metastatic spine disease because clinicians' assumptions and biases are formally incorporated into analysis. Future studies should compare Bayesian prognostication with traditional prognostication in a cohort of patients with metastatic spine disease.

Chapter 5

A Technique for Approximating Transition Rates from Published Survival Analyses

Markian A. Pahuta¹, Joel Werier², Eugene K. Wai^{2,3}, Roy A. Patchell⁴, Doug Coyle³

- 1 – Divisions of Orthopaedic & Neurological Surgery, Henry Ford Health System, Detroit, Michigan, USA
- 2 – Division of Orthopaedic Surgery, The University of Ottawa, Ottawa, Ontario, Canada
- 3 – School of Epidemiology and Public Health, The University of Ottawa, Ottawa, Ontario, Canada
- 4 – Department of Neurosurgery, University of Kentucky Medical Center, Lexington, Kentucky, USA

Context

This chapter addresses skepticism regarding decision-analysis. Credibility of Markov cohort simulation suffers when models are improperly populated with the results of partitioned-survival analysis. However, the correct technique for these situations, partitioned-survival modeling, does not yield as informative of results as Markov cohort simulation. In this chapter I present a solution to this problem. I develop and validate a technique to approximate transition rates from published partitioned-survival analysis. Formal multistate techniques, as described in this chapter, will be used to conduct QALY decision analysis in Chapter 6.

Abstract

Background

Quality-adjusted-life-years (QALYs) are used to concurrently quantify morbidity and mortality within a single parameter. For this reason, QALYs can facilitate the discussion of risks and benefits during patient counseling regarding treatment options.

QALYs are often calculated using partitioned-survival modelling. Alternatively, QALYs can be calculated using more flexible and informative state-transition models populated with transition rates estimated using multistate modelling (MSM) techniques. Unfortunately the latter approach is considered not possible when only progression-free survival (PFS) and overall survival (OS) analyses are reported.

Methods and Results

We have developed a method that, under particular conditions, can be used to estimate *approximate transition rates* from published PFS and OS analyses (we will refer to transition rates estimated using full multistate methods as *true transition rates*).

Conclusions

In this article we present the basis for and use of the transition rate approximation method. We then apply the method to a case study and evaluate the method in a simulation study.

5.1 Introduction

Chronic, progressive, and non-communicable diseases (such as cancer, diabetes, cardiovascular disease and chronic respiratory disorders) are now the leading cause of morbidity and mortality around the world. More than 60% of global deaths are attributable to these types of diseases;[224] consequently these diseases now account for up to 50% of the total healthcare budget in some countries.[154] Many of these diseases can be conceptualized as consisting of three health states: healthy (h), ill (i), or dead (d) (Figure 5.1).

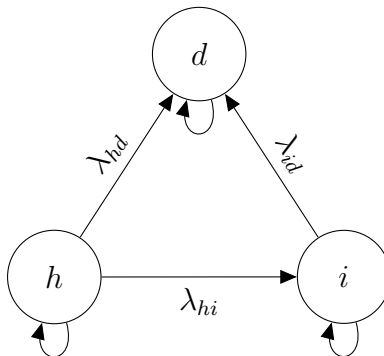


Figure 5.1: State-transition diagram for an illness-death model. The model consists of three health states: healthy (h), ill (i), and dead (d). Variable names adjacent to the solid arrows are transition rates (λ_{hi} , λ_{hd} , and λ_{id}). This model is said to be “progressive” because transitions are irreversible (i.e. unidirectional). The curved arrows indicate that individuals can remain in a particular state over time. See text for more details.

Treatment decisions for chronic, progressive, and non-communicable diseases are difficult because interventions can have distinct, and sometimes opposite, influences on the probability that a patient experiences a given health state. For example, a therapy (e.g. high-risk cancer surgery) may decrease the risk of death (by controlling cancer) but may increase the risk of becoming ill (if a post-operative complication occurs). Quality-adjusted-life-years (QALYs) can be used to concurrently quantify morbidity and mortality within a single parameter.[55] For this reason, QALYs may facilitate the discussion of risks and benefits during patient counseling regarding treatment options.[123] QALY calculation requires knowledge of *state-membership fractions*. These are the proportion of patients from a defined cohort that are in a given health state at a given time t . State-membership fractions can be calculated using *partitioned-survival modelling* or *state-transition modelling*. [226, 253, 251]

Partitioned-survival modelling uses data abstracted from progression-free survival (PFS) curves and overall survival (OS) curves reported in the literature.[251] PFS curves show the fraction of the cohort that is healthy over time t ($PFS(t)$); OS curves show the fraction of the cohort that is alive (either healthy or ill) over time t ($OS(t)$). Since OS curves show the fraction of alive patients, the fraction of dead patients is simply $1 - OS(t)$. The fraction of ill (but alive) patients is the difference between the fraction of alive and healthy

patients $OS(t) - PFS(t)$. We will refer to state-membership fractions calculated in this way as *partitioned-survival fractions*. [226, 253, 251] In contrast, state-transition modelling applies the results of a *multistate analysis*. For the disease shown in Figure 5.1, these techniques would be used to estimate the transition rate (i.e. the instantaneous risk (or hazard) of moving from one state to another) from health to illness ($h \rightarrow i$), from health to death ($h \rightarrow d$), and from illness to death ($i \rightarrow d$). [253] Transition rates can be used to compute transition probability matrices to calculate state-membership fractions (“multi-state fractions”). It is important to recognize that state-transition modelling is based on a set of mutually exclusive health states (health, illness, death), whereas partitioned survival modelling is based on non-mutually exclusive health states (health and illness or death for the PFS curve, and alive and dead for the OS curve). Partitioned-survival modelling is used when sufficient data for state-transition modelling is unavailable.

QALY calculations based on partitioned-survival fractions can suffer from two important limitations that result from the fact that (i) the OS analysis does not consider the survival of ill patients separate from healthy patients, and (ii) the risk of progressing to illness rather than death for healthy patients cannot be determined from PFS analysis. The first limitation of partitioned-survival fractions stems from the difficulty of extrapolating partitioned-survival fractions beyond the study’s observation period. [253] This is a significant deficiency because clinical studies often have a limited observation period that is of insufficient duration to characterize long-term clinical outcomes. [102, 9, 130, 88, 38, 222] The second limitation of partitioned-survival fractions is that computed QALYs are not generalizable to patient cohorts whose baseline fractions of healthy, ill and dead patients differs from those of the study cohort. [253] This is because the OS curve is a weighted average of OS curves for healthy and ill patients; therefore, the shape of the curve will change if the baseline ratio of healthy to ill patients differs. These two limitations restrict the use of partitioned survival fractions for decision analysis. These limitations can be avoided by calculating QALYs using multistate fractions. Because they are based on granular analyses of all transitions, multistate fractions have several advantages over partitioned-survival fractions. First, they can be reliably extrapolated beyond the study observation period. [253] Second, they can be used for decision analysis in cohorts with baseline characteristics that differ from the original study cohort. [4]

Unfortunately, one cannot usually calculate transition rates using data abstracted from PFS and OS analyses. [253, 251] Given the limitations of partitioned-survival fractions and the advantages of multistate fractions, it would be helpful to obtain transition rates and calculate the latter when one only has access to PFS and OS analyses. We have developed a method that, under particular conditions, can be used to estimate *approximate transition rates* from published PFS and OS analyses (we will refer to transition rates estimated using full multistate methods as *true transition rates*).

This article organized as follows. We first present the basis for and use of the transition rate approximation method. A case study is then reported in which we apply transition rate approximation to data from a randomized controlled trial (RCT) of treatments for metastatic epidural spinal cord compression (MESCC). We then report a simulation study evaluating the accuracy of the approximation method. In the last section we summarize and discuss our findings.

5.2 Transition Rate Approximation

The approximation technique is restricted to three-state progressive, time-homogenous Markov disease processes such as the one shown in Figure 5.1.[105] *Progressive* means that transitions are irreversible (i.e. cannot return to health from illness). *Time-homogenous*, means that transition rates do not change over time. *Markov* means that transition rates do not depend on disease history; in other words, the probability that a patient transitions from state x to state y during a particular time period is independent of their previous health state.

The data needed to use the approximation technique can be abstracted from most articles reporting PFS and OS analyses. The number of patients experiencing an event and number of censored patients in both the PFS (N_{pfs}^e and N_{pfs}^c) and OS (N_{os}^e and N_{os}^c) analyses can be determined from the article text or patients-at-risk risk table. To obtain the remaining data points, PFS and OS KM curves need to be digitized. Digitized KM curves can then be used to reconstruct individual patient data using validated algorithms to determine the event times in the PFS and OS analyses.[244] The approximation technique requires that we make note of the maximum observation time (event or censoring) in the PFS and OS analyses (τ_{pfs} and τ_{os} respectively). The area under the PFS and OS curves (AUC_{pfs} and AUC_{os} respectively) are calculated by summing the area under each step of the KM curve.

We denote $h \rightarrow i$, $h \rightarrow d$, and $i \rightarrow d$ transition rates as λ_{hi} , λ_{hd} , and λ_{id} . For the time-homogenous disease processes (i.e. constant transition rates), exit times from the (i) healthy state (i.e. $h \rightarrow i$ or $h \rightarrow d$ transition) and (ii) ill state (i.e. $i \rightarrow d$ transition) are exponentially distributed. Furthermore, once a patient exits health, the probability that they make an $h \rightarrow d$ transition is

$$\rho = \frac{\lambda_{hd}}{\lambda_{hi} + \lambda_{hd}}$$

We will refer to ρ as the risk of death for healthy patients. As there are only two possible transitions out of health, the probability that a transition out of the health state is an $h \rightarrow i$ transition is $1 - \rho$.

The mean time of exit from the healthy state (i.e. mean progression-free survival time) is a biased measure in the presence of right censoring.[43] Instead we calculate the restricted mean progression free-survival time (RMPFST $^{-\tau}$) which is interpreted as the mean progression-free survival time if observation is restricted to a truncation time τ . [134] Since the exit time from health is exponentially distributed, the RMPFST $^{-\tau}$ can be calculated as

$$\text{RMPFST}^{-\tau} = \frac{1 - e^{-(\lambda_{hi} + \lambda_{hd})\tau}}{\lambda_{hi} + \lambda_{hd}}. \quad (5.1)$$

By definition, the area under the PFS curve is equal to RMPFST $^{-\tau}$ when τ is set to the maximum observation time in the PFS analysis, τ_{pfs} . [194, 257] Using Formula 5.1, we can then numerically solve for $\lambda_{hi} + \lambda_{hd}$ using standard algorithmic methods.[18] Simultaneous events in the PFS and OS analyses indicate $h \rightarrow d$ transitions. Therefore, we can

approximate the risk of death for healthy patients as

$$\rho \approx \frac{N_{simul}}{N_{pfs}^e}. \quad (5.2)$$

To approximate λ_{id} we need to use information gathered from the OS analysis. It is more challenging to define an exact formula for the restricted mean overall survival time ($\text{RMOST}^{-\tau}$) than from the RMPFST $^{-\tau}$ because exit from the alive state (i.e. healthy *or* ill) is defined by a mixture of two exponential distributions: exit from health and exit from illness. However, if we know the death times o_i^e and censoring times o_j^c for a cohort of alive patients, N_{os}^e who had an observed event, and N_{os}^c who were right censored, we can approximate $\text{RMOST}^{-\tau}$ truncated to τ_{os} , $\text{RMOST}^{-\tau_{os}}$, using inverse probability weighting

$$\text{RMOST}^{-\tau_{os}} \approx \left(\left(\frac{N_{os}^e + N_{os}^c}{N_{os}^e} \right) \sum_{i=1}^{N_{os}^e} o_i^e + \sum_{j=1}^{N_{os}^c} o_j^c \right) \left(\frac{1}{N_{os}^e + N_{os}^c} \right). [95] \quad (5.3)$$

Next, we determine the total person-time of observation in the OS analysis

$$E_{os} = \sum_{i=1}^{N_{os}^e} o_i^e + \sum_{j=1}^{N_{os}^c} o_j^c. \quad (5.4)$$

If censoring times are not denoted on the OS curve, it is not possible to determine o_j^c . However, we can rearrange Formula 5.3 to yield

$$\sum_{j=1}^{N_{os}^c} o_j^c \approx \text{RMOST}^{-\tau_{os}} (N_{os}^e + N_{os}^c) - \sum_{i=1}^{N_{os}^e} o_i^e \left(\frac{N_{os}^e + N_{os}^c}{N_{os}^e} \right).$$

If we substitute this relationship into Formula 5.4 we obtain

$$\begin{aligned} E_{os} &\approx \sum_{i=1}^{N_{os}^e} o_i^e + \text{RMOST}^{-\tau_{os}} (N_{os}^e + N_{os}^c) - \sum_{i=1}^{N_{os}^e} o_i^e \left(\frac{N_{os}^e + N_{os}^c}{N_{os}^e} \right) \\ &\approx \text{RMOST}^{-\tau_{os}} (N_{os}^e + N_{os}^c) + \sum_{i=1}^{N_{os}^e} o_i^e \left(1 - \frac{N_{os}^e + N_{os}^c}{N_{os}^e} \right) \end{aligned} \quad (5.5)$$

We can repeat the same calculations using the corresponding data from the PFS analysis to approximate total person-time of observation in the OS analysis, E_{pfs} . We then approximate the total person-time of observation in the ill state as

$$E_{ill} \approx E_{os} - E_{pfs}. \quad (5.6)$$

If we make the assumption that the number of $i \rightarrow d$ transitions is

$$N_{id} \approx N_{os}^e - N_{simul}, \quad (5.7)$$

we can compute

$$\lambda_{id} \approx \frac{N_{id}}{E_{ill}}. [250] \quad (5.8)$$

5.3 MESCC Case Study

To evaluate whether the approximation method can generate reasonable results, we compared approximate transitions rates against a gold standard of true transition rates estimated from real study data.

Table 5.1: Data Abstracted from MESCC RCT PFS and OS Analyses

	Description	RT-alone Arm	mS+RT Arm
PFS Analysis			
N_{pfs}^e	total number of PFS events	17	31
N_{pfs}^c	total number of patients censored from PFS analysis	5	6
$\sum_{i=1}^{N_{pfs}^e} p_i^e$	person-time of PFS observation	7.02 years	27.22 years
τ_{pfs}	maximum observation time in the PFS analysis	2.97 years	5.25 years
AUC_{pfs}	Area under PFS curve	0.63	1.16
OS Analysis			
N_{os}^e	total number of OS events	44	45
N_{os}^c	total number of patients censored from OS analysis	1	3
τ_{os}	maximum observation time in the OS analysis	2.99 years	5.25 years
$\sum_{i=1}^{N_{os}^e} o_i^e$	person-time of OS observation	24.17 years	35.89 years
AUC_{os}	Area under OS curve	0.62	0.98
Synthesis of PFS and OS Analyses			
N_{simul}	total number of simultaneous events in PFS and OS curves	10	19

Patchell *et al.* conducted a randomized controlled trial (RCT) comparing modern surgery and radiotherapy (mS+RT) versus radiotherapy alone (RT-alone) for the treatment of metastatic epidural spinal cord compression (MESCC).[171] MESCC occurs when cancer metastasizes to the spine which and can lead to loss of ambulation from paralysis. MESCC can be modelled as in Figure 5.1 if we consider ability to ambulate as the healthy state h and the inability to ambulate due to neurologic dysfunction as the ill state i . True transition rates were estimated using individual patient data provided by the study authors. To eliminate the potential for transcription error and inaccuracy in individual patient data reconstruction, we used actual individual patient data to generate the data listed

in Table 5.1. We estimated true transition rates using the Bayesian modeling language Stan,[212] run through the statistical programming language R (Appendix D).[183] The effect of mS+RT was parametrized as a log-hazard ratio for each RT-alone transition rate.

Prior to comparing true and approximate transition rates, we conducted non-parametric multistate analysis to assess whether our assumed model (progressive, time-homogenous and Markov) was appropriate for MESCC. Non-parametric multistate fractions were estimated from individual patient data from the MESCC RCT using the etm library[1] run through the statistical programming language R.[183] We compared non-parametric multistate fractions and multistate fractions calculated from true transition rates. Goodness-of-fit tests for true multistate analysis of data observed with exact transition times affected by right censoring have not been developed.[226] We therefore used informal graphical methods.

Plots comparing proper non-parametric multistate and proper parametric multistate fractions showed good agreement, and no evidence of systematic deviation (Figures 5.2 and 5.3). Therefore, a progressive time-homogenous three-state Markov model is appropriate for the MESCC RCT data and true transition rates can serve as an appropriate comparator to evaluate approximate transition rates. Calculations for the mS+RT arm are shown in Appendix D.3.

The true transition rates shown in Table 5.2 provides useful insights into the impact of treatment. mS+RT prolongs ambulation with a statistically significant hazard ratio of 0.53 (95% CrI: 0.30, 0.94) on the total transition rate for exit from the ambulatory state ($\lambda_{hi} + \lambda_{hd}$). For patients making a transition out of the ambulatory state, the risk of death was similar with both treatments: relative risk 1.07 (95% CrI, 0.65 – 1.75). mS+RT tended to increase the mortality rate for non-ambulatory patients, hazard ratio for λ_{id} of 1.61 (95% CrI: 0.89, 2.66), but this effect was not statistically significant.

All approximate transition rates lay within the 95% credible intervals for true transition rates. There was no consistent direction of error indicating the approximation method does not consistently under-or over-estimate true transition rates.

5.4 Simulation Study

To assess the validity of the approximation strategy in a wider set of conditions, we conducted a simulation study to assess the impact of censoring on the accuracy of the approximation method for $(\lambda_{hi} + \lambda_{hd})$, ρ , and λ_{id} .

Data were generated randomly for a three-state progressive, time-homogenous Markov disease process with parameters similar to those for the mS+RT arm from the MESCC trial. A simulated cohort of 100 patients, 75 of which were healthy at baseline, was created with $\lambda_{hi} = 0.33$, $\lambda_{hd} = 0.53$, $\lambda_{id} = 3.28$. Events times were independently censored using a uniform distribution to achieve all combinations of 0, 2, 5, and 10 patients censored from the OS and PFS analysis. 100 000 replications were generated for each set of simulation conditions.

Table 5.2: Comparison of True and Approximate Transition Rates and Hazard Ratios for MESCC RCT

	Transition Rates			% Error	Hazard Ratios		
	True (95% CrI)	Approximate	Error		True (95% CrI)	Approximate	Error
RT-alone group							
λ_{hi}	0.66 (0.27, 1.20)	0.65	-0.01	-1.52			
λ_{hd}	1.00 (0.50, 1.67)	0.93	+0.07	+7.53			
λ_{id}	2.12 (1.40, 2.98)	2.24	+0.12	+5.66			
mS+RT group							
λ_{hi}	0.33 (0.17, 0.54)	0.33	0.00	0.00	0.50 (0.21, 1.40)	0.51	+0.01
λ_{hd}	0.53 (0.32, 0.80)	0.52	-0.01	-1.89	0.55 (0.26, 1.16)	0.56	+0.04
λ_{id}	3.28 (2.09, 4.69)	3.63	+0.35	+10.67	1.61 (0.89, 2.66)	1.62	+0.01

CrI, Bayesian credible interval. Error, Approximate - True. % Error, (Approximate - True) $\div$ True.

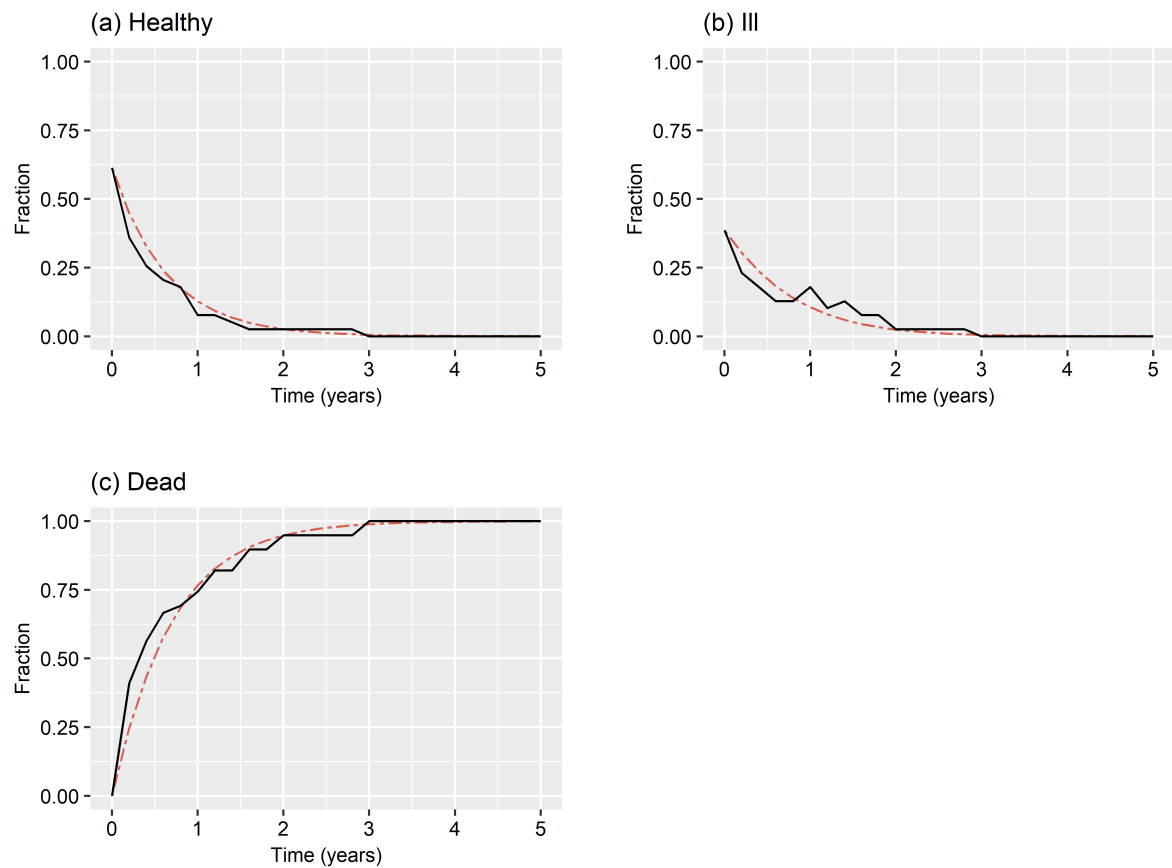


Figure 5.2: State-membership fractions for RT-alone arm. Non-parametric multistate fractions, solid black line. Parametric multistate fractions based on true transition rates, small dashed red line.

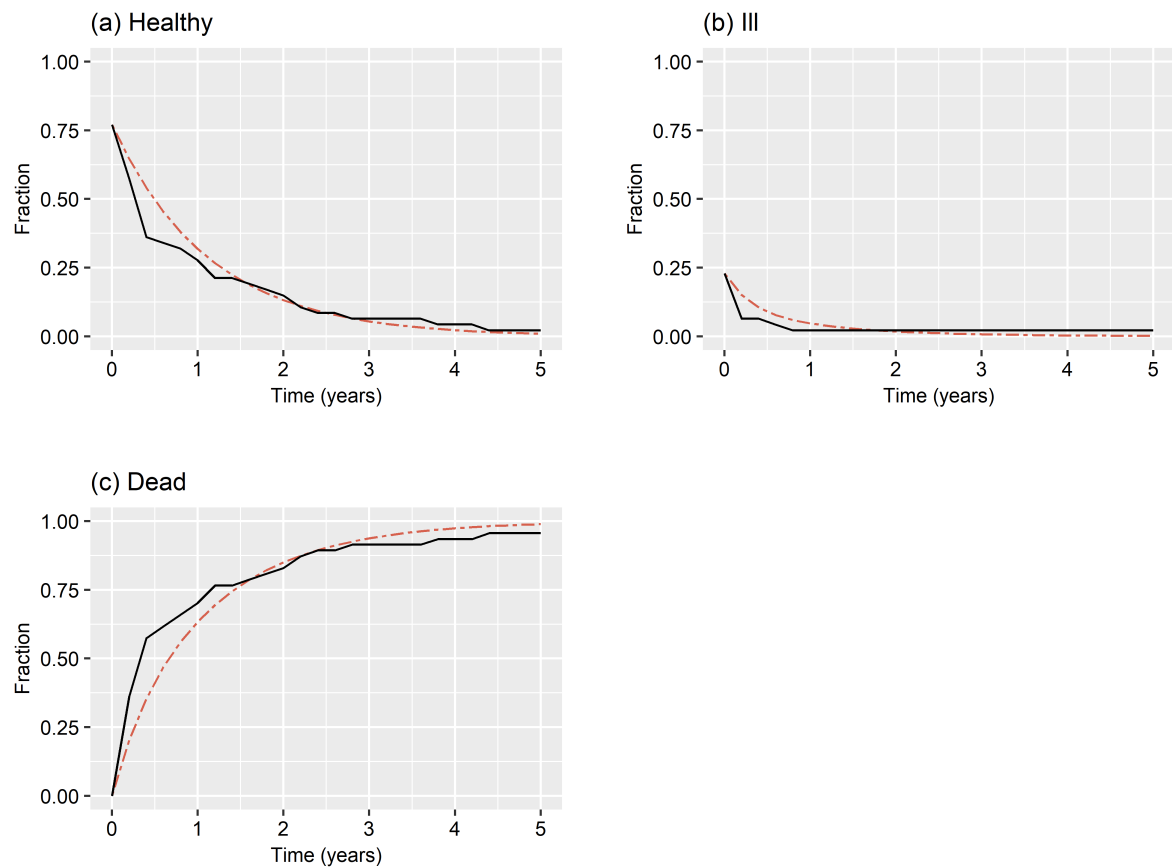


Figure 5.3: State-membership fractions for mS+RT arm. Non-parametric multistate fractions, solid black line. Parametric multistate fractions based on true transition rates, small dashed red line.

We calculated the mean error (ME), mean absolute error (MAE), mean percentage error (%ME), and mean absolute percentage error (%MAE) for each set of simulation conditions (Tables 5.3, 5.4, and 5.5). ME and %ME are a measure of the direction of bias (systematic over- or underestimation). MAE and %MAE are a measure of the magnitude of error, regardless of direction.

The approximation method tended to underestimate $(\lambda_{hi} + \lambda_{hd})$ and ρ as the censoring rate increased, however the bias was small with %ME under 3% in all censoring conditions. Even under no censoring, the approximation method was imprecise with a relatively high MAE and %MAE; increasing censoring did not significantly decrease precision.

The approximation method tended to underestimate λ_{id} as the censoring rate increased, however the bias was small with %ME under 3% in all censoring conditions. Even under no censoring, the approximation method was imprecise with a relatively high MAE and %MAE; increasing censoring did not significantly decrease precision.

The approximation method tended to underestimate λ_{id} as the censoring rate increased, however the bias was small with %ME under 3% in all censoring conditions. Even under no censoring, the approximation method was imprecise with a relatively high MAE and %MAE; increasing censoring did not significantly decrease precision.

Table 5.3: Simulation Results for $(\lambda_{hi} + \lambda_{hd})$

# Censored OS	# Censored PFS	ME	MAE	%ME	% MAE
0	0	0.00	0.08	0.11	9.34
5	8	-0.06	0.10	-7.33	11.19
5	15	-0.10	0.12	-12.18	14.09
5	30	-0.21	0.21	-24.69	24.86
10	15	-0.10	0.12	-11.50	13.70
10	30	-0.20	0.21	-23.69	23.93
20	15	-0.08	0.11	-9.85	12.74
20	30	-0.19	0.19	-22.32	22.65
40	30	-0.17	0.17	-19.30	20.01

5.5 Discussion

Although chronic, progressive, and non-communicable diseases chronic diseases affect both patients' survival and quality-of-life, interventions may impact on these two outcomes differentially. QALYs can simplify decision-making and counselling regarding treatment options.[123] For clinicians and decision makers, QALYs calculated using multistate fractions are useful because they can be used to extrapolate long-term quality-of-life and to

Table 5.4: Simulation Results for ρ

# Censored OS	# Censored PFS	ME	MAE	%ME	% MAE
0	0	-0.00	0.04	-0.01	7.27
5	8	-0.00	0.05	-0.66	7.90
5	15	-0.00	0.05	-0.62	8.24
5	30	-0.00	0.06	-0.58	9.50
10	15	-0.01	0.05	-1.42	8.30
10	30	-0.01	0.06	-1.33	9.52
20	15	-0.02	0.05	-3.75	8.70
20	30	-0.02	0.06	-3.05	9.74
40	30	-0.06	0.08	-9.38	12.21

conduct rich decision analysis. Unfortunately, one cannot usually calculate multistate fractions from PFS and OS curves.[226, 253, 251]

In this paper, we presented a technique for approximating transition rates, which can be used to calculate multistate fractions, from PFS and OS analysis. Our technique requires that three elements be abstracted from each of the PFS and OS analyses: (i) total number of events, (ii) total number of censored patients, and (iii) event times.

Approximate transition rates provide a reasonable estimate of true transition rates estimated using full multistate methods. For the MESCC RCT case study, all approximate transition rates lay within the 95% Bayesian credible intervals for true transition rates. The simulation study indicates that the approximation method is relatively unbiased and precise for estimating the transition rate out of health ($\lambda_{hi} + \lambda_{hd}$) and the risk of death for healthy patients ρ .

It is important to recognize that our techniques only apply to a time-homogenous progressive three-state irreversible disease process. Time-homogeneity is violated if the transition rates change with time (i.e. any parametric model aside from the exponential) or depend on the amount of time spent in the preceding health state (non-Markov phenomenon).[120] Irreversibility is violated if patients can become healthy after being ill.[105] Our approximation approach can be scaled-up to more complex (e.g. reversible transitions, >3 health states) disease models, however, the formulas will become more complex. Furthermore, as was done in this article, it would be necessary to validate the scaled-up approximation approach to evaluate for bias.

In this paper, we have demonstrated that transition rates can be approximated from published PFS and OS analyses. The approximation method is more accurate for estimating the transition rates out of health than the transition rate out of illness. The method tends to under-estimate true transition rates as censoring increases; therefore, approximate transition rates are not a substitute for true transition rates estimated with full multistate

Table 5.5: Simulation Results for λ_{id}

# Censored OS	# Censored PFS	ME	MAE	%ME	% MAE
0	0	0.07	0.41	2.25	12.42
5	8	-0.41	0.56	-12.59	17.19
5	15	-0.55	0.72	-16.81	21.92
5	30	-69.35	71.28	-2114.42	2173.07
10	15	-0.52	0.70	-15.96	21.35
10	30	-0.73	2.36	-22.13	72.08
20	15	-4.54	4.60	-138.32	140.30
20	30	0.79	2.75	24.22	83.88
40	30	-1.89	1.90	-57.67	57.88

methods. However, when proper multistate analysis is not available, approximate transition rates can guide probabilistic modeling and enhance QALY analysis if one considers and accounts for the limitations of the approximation method.

Chapter 6

Decision Analysis of Prognostic Contraindications to Surgery for Metastatic Epidural Spinal Cord Compression

Markian A. Pahuta¹, Joel Werier², Eugene K. Wai^{2,3}, Roy A. Patchell⁴, Carl van Walraven^{3,5,6},
Doug Coyle³

- 1 – Divisions of Orthopaedic & Neurological Surgery, Henry Ford Health System, Detroit, Michigan, USA
- 2 – Division of Orthopaedic Surgery, The University of Ottawa, Ottawa, Ontario, Canada
- 3 – School of Epidemiology and Public Health, The University of Ottawa, Ottawa, Ontario, Canada
- 4 – Department of Neurosurgery, University of Kentucky Medical Center, Lexington, Kentucky, USA
- 5 – The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada
- 6 – The Institute for Clinical Evaluative Sciences, Ottawa, Ontario, Canada

Context

The ultimate purpose of this thesis is to determine whether MESCC patients derive net HRQoL benefit from mS+RT compared to RT-alone. This work is motivated by controversy regarding the role of mS+RT despite Level 1 evidence that this treatment is more effective than RT-alone for MESCC.

In this chapter I will try to resolve the controversy regarding the role of mS+RT by determining if patients with short survival derive net HRQoL with mS+RT relative to RT-alone. If patients benefit from surgery even under the worst-case survival scenario, clinicians and patients could make treatment decisions with greater comfort.

The first four research chapters set the foundation for this final chapter. I conduct Markov cohort simulation using a simple and transparent multistate MESCC disease model developed in Chapter 5. QALYs are calculated using the utilities valued in Chapter 3. In an effort to minimize perceptions of arbitrariness, the analysis is run using posterior Bayesian distributions (Chapter 4).

Abstract

Objectives

To determine whether patients with metastatic epidural spinal cord compression (MESCC) and poor prognosis can derive net health-related quality-of-life (HRQoL) benefit from modern surgery followed by radiotherapy (mS+RT).

Summary of Background Data

To benefit from mS+RT, patients should survive longer than the time needed to recover and rehabilitate. Many clinicians consider an expected survival less than 3-6 months to be a contraindication to surgery and instead offer radiotherapy alone (RT-alone).

Treatment decisions are particularly challenging for MESCC patients who have uncertain prognoses. If survival is short, recovery from mS+RT and possible operative complications will impact on a greater proportion of their remaining life.

Methods

Decision analytic and Markov simulation methods were used to measure changes in incremental quality-of-life-years (Δ QALYs) in patients with a poor prognosis. Poor prognosis was defined as median survival ≤ 3 . The model was populated using outcomes from a randomized controlled trial of mS+RT versus RT-alone for MESCC, and utility values from a general population utility valuation study for MESCC. To model the belief that mS+RT is more morbid than RT-alone, we implemented a “surgical penalty.”

Results

On average, baseline non-ambulatory and ambulatory patients with poor prognosis benefit from mS+RT if the incremental disutility of mS+RT vs RT-alone on an average day (surgical penalty) is less than 4.02 and 4.50 dysfunctions respectively. We are 97.5% confident that mS+RT provides net HRQoL benefit for non-ambulatory patients with poor prognosis when the surgical penalty is at less than 1.78, and for ambulatory patients with poor prognosis when the surgical penalty is less than 3.10.

Conclusions

This is the first study to critically examine prognostic contraindication to surgery for MESCC treatment. We have demonstrated that MESCC patients with poor predicted prognosis can derive HRQoL benefit for S+RT. Given the inaccuracy of prognostication for MESCC, blanket survival indications/contraindications to surgery should be avoided.

6.1 Introduction

Contemporary surgery followed by radiotherapy (direct decompression and stabilization, mS+RT) for symptomatic Metastatic Epidural Spinal Cord Compression (MESCC) provides sustained improvements to patients' physical function and health-related quality-of-life (HRQoL).[171, 65] However, patients should survive longer than the time needed to recover and rehabilitate from surgery to benefit from that surgical intervention. Many surgeons regard an expected survival less than 3 – 6 months to be contraindication to surgery for MESCC.[42, 229, 232, 228, 227] Patients with poor prognosis are offered radiotherapy alone (RT-alone).

Although several prognostic indices have been developed and validated for estimating expected survival of patients with spinal metastases, their use in surgical decision making has been limited.[172, 132, 211] The most commonly used indices, the Tokuhashi and Tomita scores, have only "fair" discriminative capacity for 6-month survival.[133, 99, 98] Furthermore, the Tokuhashi and Tomita scores may not be applicable to patients treated with modern chemotherapy. This is because these scores were developed in the 1990s and 2000s prior to the introduction of targeted therapies.[87, 236] Prognostic uncertainty poses a significant challenge for the surgeon when the complication profile of the two treatment options, mS+RT and RT-alone, is considered.

Patients undergoing RT-alone may experience nausea, dermatitis, and symptoms from inflammation of the gastrointestinal tract (pharyngitis, esophagitis, and gastritis) within a month of receiving spinal external beam radiotherapy. Less than 10% of patients will report these symptoms to be severe.[146, 145, 237] Early toxicity may be even less frequent with the newer technique of stereotactic radiosurgery because radiation is delivered to a smaller field.[107] Infrequently, patients may experience late toxicities months after treatment, including myelopathy and vertebral compression fractures.[144] When compared to those from mS+RT, complications from RT-alone are less common and relatively mild.

Analysis of 30-day National Surgical Quality Improvement Program data found that 14.4% of patients undergoing mS+RT for primary and metastatic spinal tumors experience a major complication with 5.3% undergoing a re-operation and 10.2% being readmitted.[121] Prospectively collected data from the AOSpine MESCC study reported the most common complications to be urinary tract (25.0%) and wound infections (10.4%).[65] Adverse events were significantly higher in a single-center cohort of patients undergoing emergency surgery for spinal metastases when complications were studied as the primary outcome.[47] In this study, 76.2% of patients experienced at least one complication, and the mortality rate during the admission was 10.9%. Intraoperative adverse events occurred in 31.7% of patients and included: blood loss >2L (16.8%), incidental durotomy (9.9%), malpositioned implants requiring revision (5.9%), nerve root injury (2%), and cardiac events (1%). Postoperative adverse events occurred in 20.8% of patients and included: delirium (20.8%), transient neurological deterioration (5.9%), pressure sores (4.0%), deep vein thrombosis (4.0%), early construct failure (2.0%), electrolyte imbalance (11.0%), arrhythmia (4.0%), dysphagia (5.0%). The risks of mS+RT are substantial and can negatively affect HRQoL. However, patients experiencing complications can still derive overall net benefit from mS+RT.[47, 3]

Given prognostic uncertainty and the risks of mS+RT, some clinicians may take a conservative approach and not offer surgery out of concern for the worst-case scenario.[169, 196, 136] This is because recovery from mS+RT and complications will impact on a greater fraction of remaining life in patients having short survival thereby negating net HRQoL benefit. This conservative approach may deny patients with uncertain prognosis the more effective treatment.

To help guide treatment decision making in the context of uncertain prognosis in MESCC, it would be useful to determine whether such patients with poor prognosis can derive net HRQoL benefit from mS+RT. In this paper we study this question using clinical decision analytic methods combined with Markov simulation using data from a randomized controlled trial comparing mS+RT and RT-alone for MESCC[171] and a general population utility evaluation study for MESCC health states[168].

6.2 Methods

Markov simulation was used to measure changes in the incremental quality-of-life-years (Δ QALYs) for patients with poor prognosis undergoing mS+RT versus those undergoing RT-alone. Poor prognosis was defined as median survival ≤ 3 months because 3 months is the lower limit for which patients are generally considered to be mS+RT candidates.[227] Measurement of QALYs was done from the societal perspective.

Estimation of Effectiveness and Model Structure

Effectiveness parameters were estimated by re-analysis of data from a randomized controlled trial (RCT) by Patchell *et al.*[171] The trial compared radiotherapy alone (30 Gy in ten fractions) to direct decompression and stabilization followed by radiotherapy (30 Gy in ten fractions) for the treatment of symptomatic MESCC (Table 6.1). In this trial, patients' health state (ambulatory, a ; non-ambulatory, n ; or dead, d) was assessed immediately before treatment and within one day of completion of treatment. Time-to-event data was not collected during the treatment phase, while it was collected during the post-treatment phase. Due to differences in data collection, effectiveness parameters were estimated in a different way for the treatment phase and post-treatment phase.

Without time-to-event analysis, the treatment phase data could only be used to compute the event probabilities being in the a , n or d state immediately after treatment. Time-to-event data permitted calculation of transition rates and probabilities for a multistate model. As transitions from the non-ambulatory to ambulatory state were not analyzed in the MESCC RCT, we were unable to compute the rate and probability of this transition.

Event probabilities for the treatment-phase (θ_a , θ_n , and θ_d) were estimated separately for (1) baseline ambulatory patients undergoing RT-alone, (2) baseline non-ambulatory patients undergoing RT-alone, (3) baseline ambulatory patients undergoing mS+RT, and

Table 6.1: Selection Criteria used in the MESCC RCT[171]

Inclusion Criteria	Exclusion Criteria
-Adult patients (≥ 18 years old)	-Multifocal MESCC on MRI
-MESCC	-Primary spinal tumour (sarcoma, chondrosarcoma, chordoma, meningioma)
-Tissue proven diagnosis of cancer	
-Radiologic evidence of spinal cord displacement by tumor	
-At least one neurologic sign attributable to spinal cord compression (including pain)	

(4) baseline non-ambulatory patients undergoing mS+RT. Treatment-phase event probabilities were estimated using Bayesian Markov Chain Monte Carlo (MCMC) methods with non-informative prior distributions (Appendix E.1.1).

We estimated post-treatment transition probabilities (λ_{an} , λ_{ad} , and λ_{nd}) separately for each treatment group. A time-homogenous progressive three-state irreversible multistate model fitted using Bayesian MCMC methods with non-informative prior distributions (Appendix E.1.2).

The state-transition diagram for the Markov simulation model is shown in Figure 6.1. The model consists of three health states: ambulatory (a), non-ambulatory (n), and dead (d). Due to the structure of the effectiveness data reported in the MESCC RCT, the simulation was divided into a treatment-phase and post-treatment phase. As in the MESCC RCT, patients begin the simulation in either the a or n state. During the treatment phase, patients are permitted to transition to any of the three states. After treatment, patients enter a progressive time-homogenous Markov model meaning that transitions between states are irreversible and occur at a constant rate. Furthermore, patients in the n state cannot regain the ability to ambulate.

Population

Immediate post-treatment event probabilities and transition probabilities have not been reported for MESCC patients with poor prognosis. We defined such a cohort by restricted Monte Carlo sampling of the posterior distributions for immediate post-treatment event probabilities. We used samples for which baseline ambulatory patients undergoing RT-alone were expected to have a median survival of ≤ 2.25 months in the post-treatment phase. This threshold was chosen because the treatment phase in the MESCC RCT was 3 weeks, thus the threshold corresponds to ≤ 3 months from the start of treatment.

Valuing Health States

For the MESCC health states shown in Figure 6.1, utilities have been estimated in a general population valuation study and modeled using mixed effects beta regression.[168] This

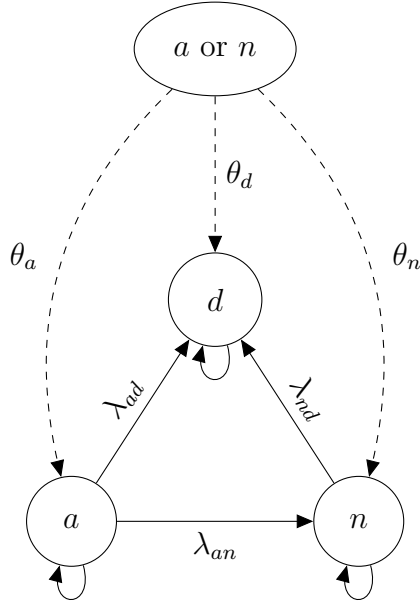


Figure 6.1: State-transition diagram for the Markov simulation. The model consists of three health states: ambulatory (a), non-ambulatory (n), and dead (d). Patients enter the simulation in either state a or n . During the treatment-phase, patients can transition to any of the three states; this transition is denoted by dashed arrows. Variable names adjacent to the dashed arrows are treatment-phase event probabilities (θ_a , θ_n , and θ_d). After treatment, patients enter a progressive time-homogenous Markov model of the MESCC disease process. Variable names adjacent to the solid arrows are post-treatment transition rates (λ_{an} , λ_{ad} , and λ_{nd}). See text for more details.

model was re-fit using Bayesian MCMC methods with non-informative prior distributions (Appendix E.1.3).

However, the health states shown in Figure 6.1 do not account for a longer recovery from mS+RT (compared to RT-alone) and the more severe complications associated with this treatment. Consider two identical patients having the same health state; compared to the patient receiving RT-alone, the patient receiving mS+RT is expected to have a lower utility (at least initially). Therefore we would anticipate an initial *penalty* on utility if a patient undergoes S+RT. The surgical penalty is time-dependent, in that it resolves after (and if) the patient recovers from mS+RT. In some patients, the surgical penalty may persist if a permanent complication occurs. The intensity of the surgical penalty also varies. Unfortunately, the temporal course, frequency, and severity of the surgical penalty of MESCC patients undergoing spine surgery has not been reported in the literature.

Because our analysis uses cohort simulation to compute mean Δ QALYs, granular data on all characteristics of the surgical penalty are not necessary. We apply the surgical penalty as a constant over the entire lifespan to all simulated patients. Therefore, the surgical penalty can be viewed as the average decrease in utility attributable to mS+RT vs RT-alone on an average day for the average patient. Since the surgical penalty is applied as a constant over the entire lifespan, it is not necessary to model the time-course. Since the surgical penalty is applied to all simulated patients, it is not necessary to model the frequency.

As constructed, the impact of the surgical penalty on Δ QALYs also depends on the expected survival of a simulated cohort. For example, a low surgical penalty in a simulation with long survival could result in the same Δ QALYs as a high surgical penalty in a simulation with a short survival. We note that the objective of our analysis is to examine the benefit of mS+RT in patient with poor prognosis in relation to the morbidity of this treatment. Therefore, survival is a nuisance parameter which is addressed through probabilistic Monte Carlo parameter sampling of posterior distributions for immediate post-treatment event probabilities. Thus in our analyses, the confounding effect of expected survival on Δ QALYs is addressed by averaging out over expected survival periods consistent with a cohort with poor prognosis.

The results of our general population MESCC utility valuation study can be used to model the severity of the surgical penalty.[168] In this study, respondents were asked to value a set of dysfunctions that characterize recovery from mS+RT and its potential complications (Table 2.1). Standard gamble utility elicitation revealed that respondents applied equal weight to all dysfunctions and therefore utility was related simply to the number of dysfunctions. Thus we can quantify the average severity of the surgical penalty by specifying the number of dysfunctions listed in Table 2.1.

A uniform distribution on the interval 0 to 4 was placed on the number of dysfunctions in the surgical penalty. Posterior distributions for mean utility valuations and Bayesian 95% credible intervals for the number of dysfunctions are shown in Figure 6.2.

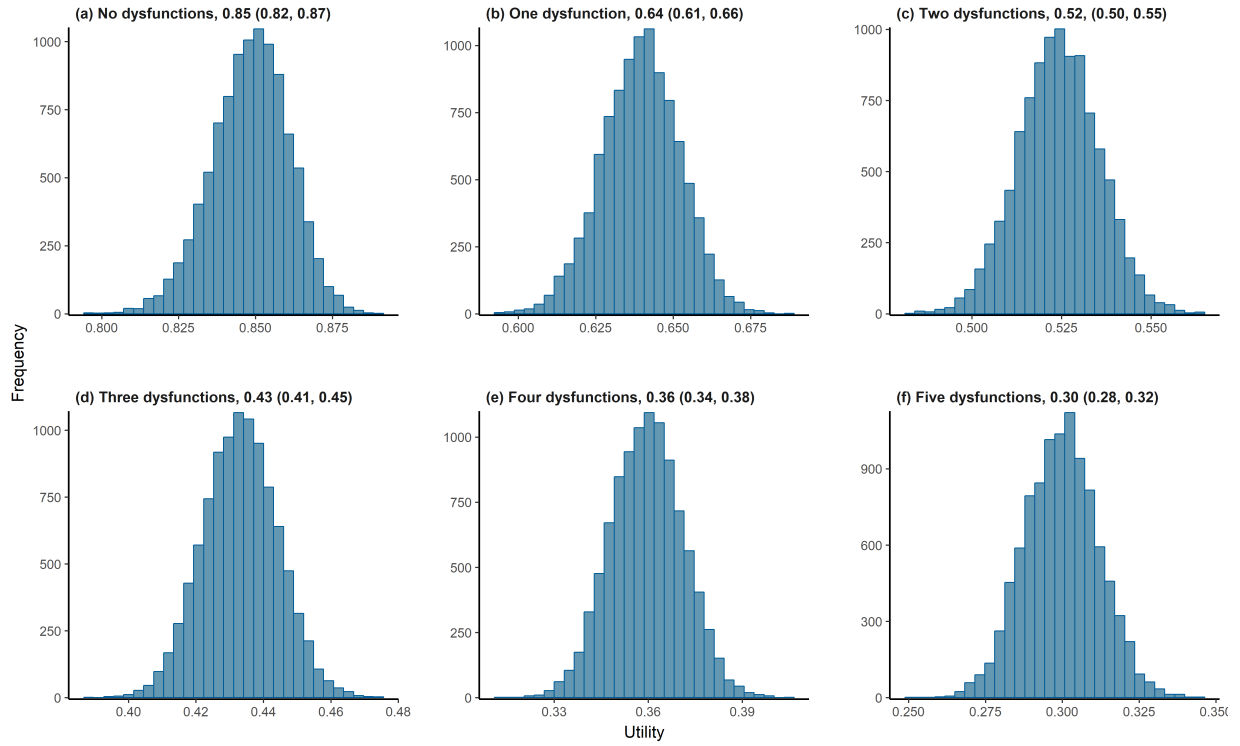


Figure 6.2: Histograms of posterior distributions of MESCC utility values for health states with (a) no dysfunctions, (b) one dysfunction, (c) two dysfunctions, (d) three dysfunctions, (e) four dysfunctions, and (f) five dysfunctions. Mean and 95% credible intervals reported in panel titles.

Analytical Methods, Parameter Uncertainty and Heterogeneity

We derived exact formulas for the results of Markov simulation for a time-homogenous progressive disease process using matrix algebra and integral calculus (Appendix E.2). Since these formulas were derived using exact mathematical analysis, computational considerations such as the cohort size, number of cycles, cycle length, and half-cycle correction are not relevant.[22]

A probabilistic model was implemented through Monte Carlo sampling of the posterior distributions for effectiveness parameters and utilities. Since Bayesian analysis yields probability distributions jointly for all parameters, it was not necessary to directly specify covariance between parameters. The surgical penalty was modeled using a uniform distribution on the number-of-dysfunctions. The model was run using 50 000 replications.

Separate analyses were conducted for baseline ambulatory and non-ambulatory patients. The relationship between mean and surgical penalty was examined using linear regression Δ QALYs. Quantile regression was used to relate the 95% credible interval for Δ QALYs and the surgical penalty.[164]

6.3 Results

When the surgical penalty was between 0 and 1 dysfunctions, the mean Δ QALY (95% credible interval) benefit was 0.35 (0.09, 0.66) and 0.46 (0.21, 0.80) for baseline non-ambulatory and ambulatory patients respectively. For both groups of patients, greater surgical penalty was associated with a greater frequency of simulations with Δ QALYs<0 (Figure 6.3).

For baseline non-ambulatory patients we found that, on average, mS+RT provides net HRQoL benefit if the surgical penalty was less than 4.02 dysfunctions (extrapolated). However, when accounting for uncertainty in the model, we could only be 97.5% confident that mS+RT provides net HRQoL benefit if the number of dysfunctions was less than 1.78 (Figure 6.4a). On average, ambulatory patients also derived HRQoL benefit from mS+RT but with a threshold above 4.50 dysfunctions (extrapolated). We can only be 97.5% confident of HRQoL benefit in non-ambulatory patients to a threshold of 3.10 dysfunctions (Figure 6.4b).

6.4 Discussion

Although mS+RT is the most effective treatment for maintaining neurologic function in MESCC patients, surgeons must make this recommendation cautiously due to prognostic uncertainty.[169] It is generally recommended that appropriate surgical candidates have an estimated survival of 3 – 6 months.[42, 229, 232, 228, 227] Although frequently used in the literature, support and rationale for this prognostic threshold is usually not provided.

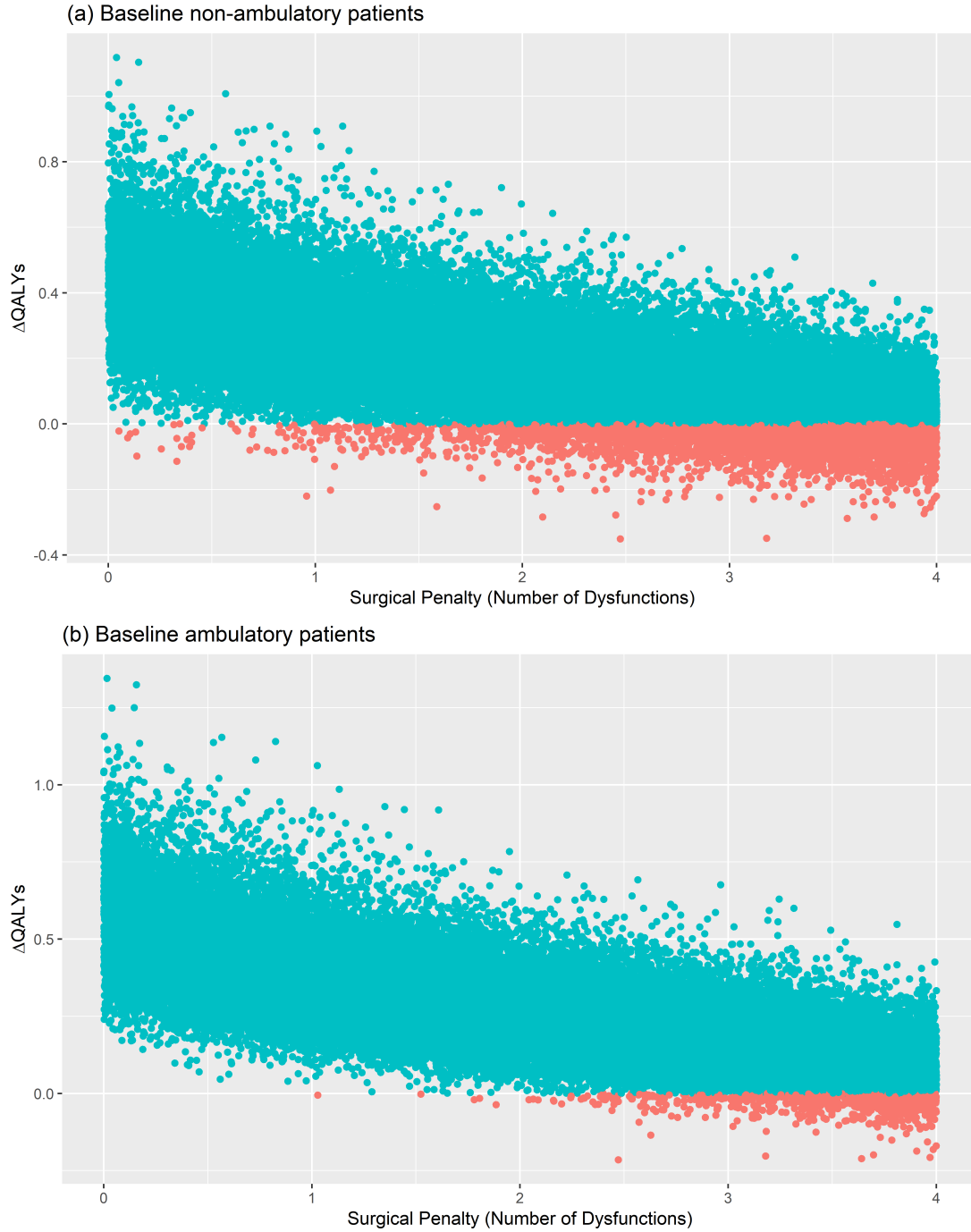


Figure 6.3: Δ QALYs for (a) baseline non-ambulatory patients, and (b) baseline ambulatory patients. Each point represents on replication of the simulation. Blue points represent Δ QALYs > 0 indicating superiority of S+RT. Red points represent Δ QALYs < 0 indicating superiority of RT. Black points represent Δ QALYs $= 0$ indicating equivalence.

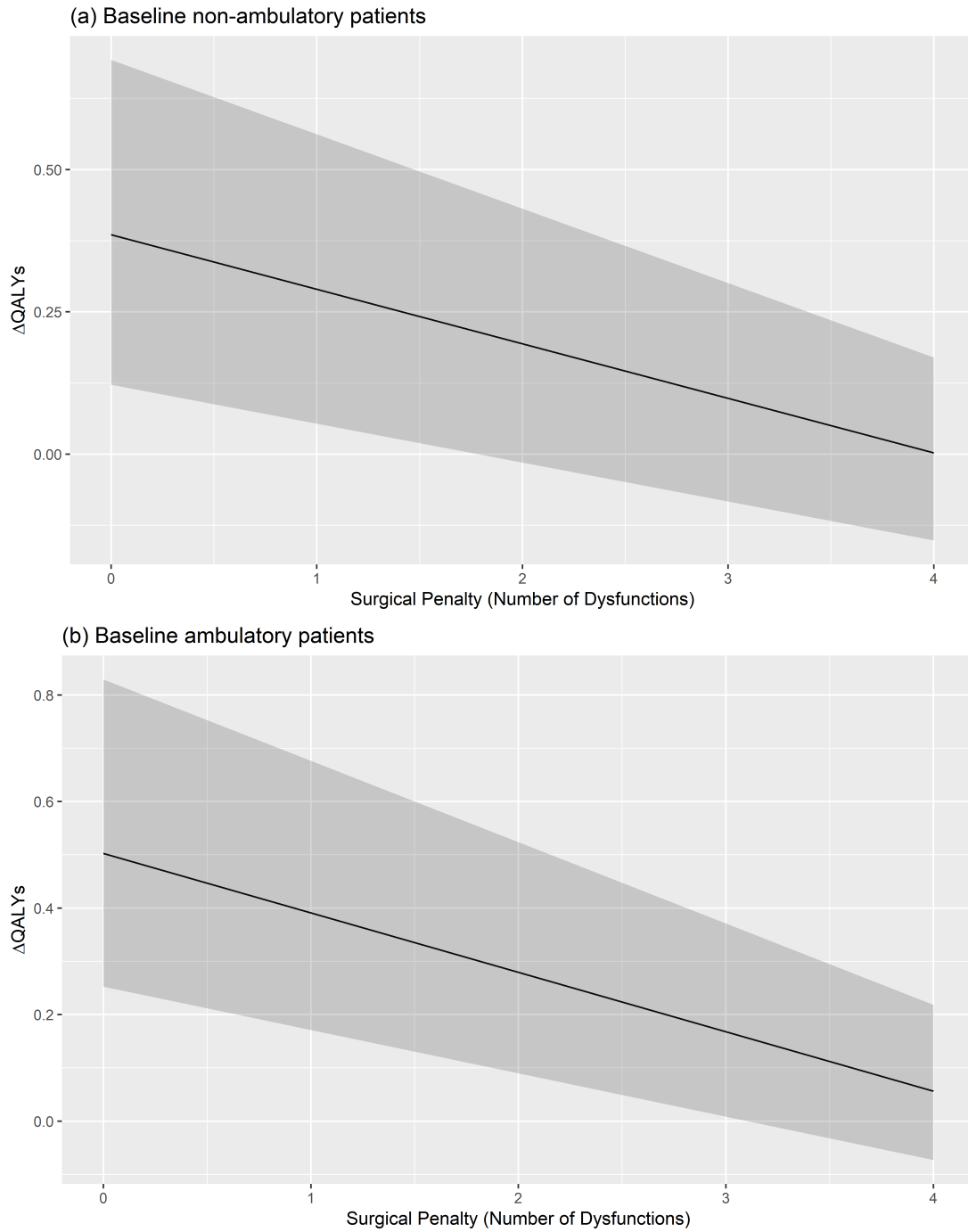


Figure 6.4: Mean $\Delta QALYs$ versus surgical penalty (a) baseline non-ambulatory patients, and (b) baseline ambulatory patients. Shaded area represents 95% credible interval. X-axis intercept is 4.02 and 4.50 (extrapolated) dysfunctions for baseline non-ambulatory and ambulatory patients respectively. The lower limit of the credible interval is 1.78 and 3.10 dysfunctions for baseline non-ambulatory and ambulatory patients respectively.

Sundaresan and colleagues first proposed a 6-month threshold in a review article in 1986.[220] They noted that although neurologic response rates are equivalent with decompressive laminectomy and radiotherapy alone, decompressive laminectomy conferred a more durable response. Since the majority of patients undergoing radiotherapy alone relapsed within 6 months, they reasoned that patients with predicted survival greater than 6 months be offered surgery. This logic may not apply to the modern era because, in contrast to decompressive laminectomy, neurologic response rates are higher with circumferential decompression and stabilization (modern surgery) than with radiotherapy alone.[255, 171]

In this study, we used Markov simulation to evaluate the HRQoL benefit of mS+RT in MESCC patients with poor prognosis (median survival ≤ 3 months). Contrary to widely used guidelines, we found that both ambulatory and non-ambulatory patients with poor prognosis can derive HRQoL benefit from mS+RT but this conclusion depends upon the surgical penalty (measured by number of dysfunctions, Table 2.1) with ambulatory patients able to withstand a larger surgical penalty than non-ambulatory patients.

Our methods permitted a more sophisticated decision analysis compared to previous reports using Patchell *et al.*'s MESCC RCT data.[225, 75] To mitigate the effects of distributional assumptions on results, we employed posterior distributions obtained from Bayesian analysis with non-informative prior distributions.[21] Previous MESCC decision analyses used the results of partitioned survival analysis, whereas we used the results of multistate analysis. Multistate analysis is advantageous because it permits extrapolation of study data beyond the study observation period and application of study data to new patient subgroups.[253] We populated the model using the only reported set of comprehensive of MESCC utilities.[168]

It is important to consider our results in the context of new treatment paradigms based on advances in radiation therapy techniques.[14] In contrast to conventional external beam radiotherapy (cEBRT) used in the MESCC RCT, stereotactic body radiation therapy (SBRT) more precisely delivers higher ablative doses of radiation.[106] Although radiation toxicity is lower with SBRT compared with cEBRT, high focal doses of radiation can predispose to compression fractures in otherwise stable lesions.[197] These competing issues have not been studied and thus the relative treatment disutility of cEBRT versus SBRT is unknown. With higher focal doses of radiation, circumferential decompression used in the MESCC RCT may not be necessary because SBRT-alone can effectively treat even moderate-grade symptomatic MESCC.[196] However, surgery does still play a role in downgrading compression to facilitate SBRT and optimize local control.[190] Surgery with this intent is termed "separation surgery" and consists of minimal decompression to create a margin of CSF between the spinal cord and tumor which has a lower surgical penalty than circumferential decompression used in the MESCC RCT. No Level 1 comparative studies of separation surgery versus SBRT-alone have been conducted. It is expected that our decision analysis of the new treatment paradigm would be influenced by differences in the (i) incremental disutility of separation surgery compared to SBRT-alone and (ii) the impact of improved local control on maintenance of ambulation.

The main limitation of our study relates to the lack of published data on the surgical penalty. We defined the surgical penalty as average decrease in utility attributable to

mS+RT vs RT-alone on an average day for the average patient. We can quantify the average severity of the surgical penalty by specifying the number of dysfunctions listed in Table 2.1. Rather than make a distributional assumption for the surgical penalty, we conducted sensitivity analysis. We are 97.5% confident that mS+RT provides net HRQoL benefit for patients with poor prognosis when the surgical penalty is less than 1.78 and 3.10 dysfunctions for baseline non-ambulatory and ambulatory patients respectively.

We made several assumptions in this study which may impact on results. First, patients could not regain the ability to ambulate after the initial treatment phase. This assumption likely decreases overall QALYs in each treatment arm. This assumption could also potentially impact on the Δ QALYs if neurologic recovery tends to be delayed with a treatment. Conceivably, recovery could be delayed from mS+RT due to need to recover from the surgical insult. Second, we assumed that transition rates are constant, and did verify this assumption by performing diagnostics for multistate models. Third, ambulation was defined as a binary phenomenon. If a more granular definition of ambulation was used, overall QALYs would likely be increased.

It is important to appreciate that the MESCC utilities used in this study were valued by healthy individuals (*ex ante* utilities).[168] *ex ante* MESCC utilities from healthy individuals may not be equivalent to *ex post* utilities from patients who have experienced MESCC. It has been reported that healthy individuals provide lower utility valuations than patients who have experienced living with a colostomy, menopausal symptoms, and hemodialysis.[15, 256, 46] However, the impact of experience may be transient as evidenced by former colostomy patients providing utility valuations equivalent to healthy individuals.[206] Furthermore, in the case of depression, healthy individuals provide higher utility valuations than patients.[182] Currently the understanding of differences between *ex ante* and *ex post* utilities is insufficient to specify general conversion rules.[52]

Rawls argues that *ex ante* utilities can be used ethically if valued under a “veil of ignorance.” If we assume that the general population providing *ex ante* utility valuations may eventually develop the condition of interest, out of self-interest, they should provide fair valuations.[187] Although *ex ante* utilities are theoretically restricted to system policy decisions, *ex ante* utilities have become the *de facto* standard for individual patient decisions. Utilities obtained from generic health surveys such as the EQ-5D, HUI and SF-6D are actually *ex ante* valuations.[55] Therefore our use of *ex ante* utilities conforms with conventions in the literature.

These results can help patients and clinicians make decisions in the face of prognostic uncertainty. By applying the results of this study, there is no need to rely on the worst-case scenario, but instead rely on the most likely scenario identified through simulation methods in this work. Our results indicate that both ambulatory and non-ambulatory patients can gain net HRQoL benefit from mS+RT. However, it is important to only apply these results to patients fitting the recruitment criteria for the MESCC RCT. Notably, patients with multifocal disease and prolonged neurology were not recruited into the study.

6.5 Conclusion

This is the first study to critically examine prognostic contraindications to surgery for MESCC. We have demonstrated that MESCC patients with poor predicted prognosis can derive HRQoL benefit for mS+RT. On average, baseline non-ambulatory and ambulatory patients with poor prognosis benefit from mS+RT if the incremental disutility of mS+RT vs RT-alone on an average day (surgical penalty) is less than 4.02 and 4.50 dysfunctions respectively. We are 97.5% confident that mS+RT provides net HRQoL benefit for non-ambulatory patients with poor prognosis when the surgical penalty is at less than 1.78, and for ambulatory patients with poor prognosis when the surgical penalty is less than 3.10. Given the inaccuracy of prognostication for MESCC, blanket survival indications/contraindications to surgery should be avoided.

Chapter 7

Discussion and Conclusion

Despite Level 1 evidence that surgery is the most effective treatment for metastatic epidural spinal cord compression (MESCC), there is controversy regarding the role of surgery because of fear that patients who have a short survival will spend a large fraction of their remaining life recovering from surgery and potential complications. In this thesis, I sought to resolve this controversy by quality-adjusted life year (QALY) analysis of MESCC treatments.

7.1 Summary of Methodological Findings and Contributions

Prior to addressing the clinical problem, I tackled two methodological barriers: (a) lack of MESCC utility data, and (b) skepticism regarding decision-analysis.

Barrier 1: Lack of Utility Data

QALYs are calculated using utilities, or health-related quality-of-life (HRQoL) weights, which can be obtained by direct valuation or from generic health status measures. For MESCC, few high-quality studies compared modern surgery followed by external beam radiotherapy (mS+RT) and external beam radiotherapy (RT-alone) using generic health status measures, and no study has conducted a direct valuation. Therefore, with a paucity of utility data, the ability to apply decision-analysis of QALYs to MESCC has been limited. Given that few high-quality studies have compared mS+RT and RT-alone for MESCC using generic health status measures, direct utility valuation was the most efficient way to collect utility data. Best practices in economic evaluation involve recruiting a sample of healthy individuals from the general population for direct utility valuation (*ex ante* utilities).[248, 198]

Traditional utility valuation studies have been conducted using face-to-face interviews, phone interviews, or postal surveys. These modes of administration have undergone psychometric validation. However, there is increasing use of web surveys for utility valuation

which are usually conducted using custom and proprietary valuation tools which have not been psychometrically validated. I felt it would be beneficial and efficient for investigators to be able to build disease-specific modules on a common platform that has been used to develop modules with acceptable psychometric properties.

In Chapter 2, I presented a new platform called the Self-directed Online Assessment of Preferences (SOAP). For the first application of this platform, I developed a module for MESCC health states. The SOAP platform met published benchmarks for reproducibility (both agreement and reliability) and responsiveness for utility measurement. To our knowledge, this is the first validated open-source, web-based, self-directed utility valuation module. This study demonstrated that the SOAP platform can be used to develop modules which have acceptable psychometric properties

In Chapter 3, I reported the first general population direct utility valuation study for a spinal disorder in the literature. This is in contrast to existing utilities for spine disease-specific instruments (such as the NDI, the ODI, and the SRS-22) which have been derived using an indirect “cross-walk” protocol.[193, 31, 32, 19] The “cross-walk” protocol involves:

1. collecting responses with a disease specific instrument and a generic health measure (eg. EuroQol-5D, Short Form-6D, or Health Utilities Index Mark 3).
2. developing a regression model to relate the disease-specific score to the generic health measure.
3. estimating a generic health measure score with the regression model.
4. converting generic health measure score to a utility value using another regression model.[55]

The cross-walk protocol is complex as it requires two regression models. Chapter 2 demonstrates that disease-specific direct utility valuation is feasible and valid. Investigators may consider applying our study protocol and the validated SOAP tool to other disease contexts.

Barrier 2: Skepticism Regarding Decision-Analysis

Once utility data are available, QALYs can be estimated directly from a study or calculated using simulation-based techniques. Study-based QALY estimates can be prohibitive due to large sample size requirements.[20] Furthermore, study-based decision analysis of QALYs is not a viable option when few high-quality comparative studies have collected HRQoL data (as in MESCC). Unfortunately, the alternative (simulation-based QALY estimates) are often viewed with skepticism.[58]

Markov cohort simulation is the most frequently used technique for estimating QALYs through simulation.[22] This technique first involves simulating disease history from a theoretical model of the disease process composed of distinct and clinically-relevant health states. The fraction of simulated patients in each health state over time is simulated using

transition probabilities derived from the literature. QALYs are then calculated by weighing the fraction of simulated patients in each health state by the utility of those health states, and then summing the weighed state-membership fractions over the time-horizon of interest.

One criticism of Markov cohort simulation relates to implementation of probabilistic sensitivity analysis.[21, 36] The random draws are taken from probability density functions derived from confidence intervals reported in the literature for each of the model inputs.[22] In Chapter 4, I took care to highlight that confidence intervals do not actually quantify the uncertainty in a value such as median survival time. Although confidence intervals are often used for this purpose, this is done because confidence intervals are frequently misinterpreted by the “confidence-level misconception” to be Bayesian credible intervals.[28] Under this misconception, a 95% confidence interval is interpreted to contain the true value with a probability of 95%.[41] The prevalence of such “confidence-level misconception” suggests that, for many individuals, the Bayesian credible interval is a more intuitive construct for expressing the uncertainty in a reported value. An added benefit is that the posterior distribution, from which a Bayesian credible interval is derived, can be used to directly populate Markov cohort simulation. Therefore posterior distributions obtained from Bayesian analysis with non-informative prior distributions can mitigate the effects of distributional assumptions in Markov cohort simulation.

Another criticism of Markov cohort simulation is that the methods are complex and poorly reported in the literature.[57, 63, 13, 56, 176, 192, 29] Due to the resulting lack of transparency, some authors characterize simulation-base techniques as a “black box.”[29] To be useful for treatment decision making, clinicians and patients must have faith in, and an understanding of, these models.[16] This perception is aggravated by the fact that transition probabilities are often improperly derived from partitioned-survival analysis.

In Chapter 5, I developed and evaluated a technique to approximate transition rates, from PFS and OS analysis. Approximate transition rates reasonably estimate transition rates estimated using full multistate methods. Approximation requires that three elements be abstracted from each of the PFS and OS analyses: (i) total number of events, (ii) total number of censored patients, and (iii) event times. It is my hope that other researchers will use this approximation technique and abandon improper derivation of transition probabilities from partitioned-survival analyses.

7.2 Summary of Main Clinical Findings

In Chapter 3, I reported directly valued utilities for MESCC health states using the validated SOAP MESCC module.[166] The study sample was representative of the population in all English-speaking Canadian provinces. Our findings are valid with 63.2% of participants having understood and engaged in the task. This rate is superior to validity rates reported in general population direct valuation studies for the EuroQol-5D.[12, 128, 113, 49]

The regression model building exercise revealed that members of the general population value all dysfunctions characterizing MESCC health states equally. Furthermore,

dysfunction follows a pattern of diminishing marginal disutility. That is, each additional dysfunction effects a smaller incremental change in utility than the previous dysfunction. These results demonstrate that, from the societal perspective, physical function is valued equal to other facets of well-being. Ambulation and continence (which are dysfunctions that are addressed by surgery) are no more important to overall health than other attributes (including pain, other symptoms, and level of independence).

In Chapter 4, I applied Bayesian analytical methods to prognostication for MESCC in an era of more effective systemic treatments. I explored the impact of new targeted therapies on prognostic estimates obtained using the Tokuhashi score. I provide a worked calculation for the probability of a patient surviving to 6 months using dichotomous prognostication. Continuous prognostication was performed using conjugate analysis. When the predicted prognosis from the Tokuhashi score is contextualized with data on contemporary systemic treatments, patients previously deemed non-surgical candidates may be eligible for surgery.

Some clinicians may be reluctant to adopt Bayesian techniques based on R.A. Fisher’s criticism that the Bayesian approach is “...extremely arbitrary... by evolving a vitally important piece of knowledge, that of the exact form of the distribution of [the value of interest], out of an assumption of complete ignorance ...”[72] This is a fair criticism, but it should be considered in context of the non-Bayesian alternative courses of action.

A nihilistic course of action based on non-Bayesian analysis would be to ignore recent advances in systemic treatment because their effectiveness for patients with metastatic spine disease has not been formally studied. An overly optimistic course of action based on non-Bayesian analysis would be to assume that the results of oncology trials enrolling few patients with metastatic spine disease fully apply to patients with metastatic disease. In Chapter 4, I demonstrated how effective sample size sensitivity analysis quantifies the impact of clinicians’ assumptions and biases on decision making and, therefore, can make decisions more transparent. In this manner, different – and potentially not fully compatible – sources of evidence can be synthesized for decision making in a rational and transparent way. Therefore, Bayesian prognostication can enhance counseling for metastatic spine disease because clinicians’ assumptions and biases are formally incorporated into analysis.

In Chapter 6, I applied the work from the previous four research chapters and used Markov simulation to evaluate the HRQoL benefit of mS+RT in MESCC patients with poor prognosis (median survival ≤ 3 months). Contrary to widely used guidelines, I found that both ambulatory and non-ambulatory patients with poor prognosis can derive HRQoL benefit from mS+RT but this conclusion depends upon the surgical penalty (as measured by number of dysfunctions resulting from surgery, Table 2.1) with non-ambulatory patients able to withstand a smaller surgical penalty than ambulatory patients.

7.3 Limitations

It is important to appreciate that the MESCC utilities used in this study were valued by healthy individuals (*ex ante* utilities).[168] *ex ante* MESCC utilities from healthy in-

dividuals may not be equivalent to *ex post* utilities from patients who have experienced MESCC. It has been reported that healthy individuals provide lower utility valuations than patients who have experienced living with a colostomy, menopausal symptoms, and hemodialysis.[15, 256, 46] However, the impact of experience may be transient as evidenced by former colostomy patients providing utility valuations equivalent to healthy individuals.[206] Furthermore, in the case of depression, healthy individuals provide higher utility valuations than patients.[182] Currently the understanding of differences between *ex ante* and *ex post* utilities is insufficient to specify general conversion rules.[52]

I made several assumptions in Chapter 6 which may impact on results. First, patients could not regain the ability to ambulate after the initial treatment phase. This assumption likely decreases overall QALYs in each treatment arm. This assumption could also potentially impact on the Δ QALYs if neurologic recovery is tends to be delayed with a treatment. Conceivably, recovery could be delayed from mS+RT due to need to recover from the surgical insult. Second, I assumed that transition rates are constant, and did verify this assumption by performing diagnostics for multistate models. Third, ambulation was defined as a binary phenomenon. If a more granular definition of ambulation was used, overall QALYs would likely be increased.

The main limitation of Chapter 6 relates to the lack of published data on the surgical penalty. I defined the surgical penalty as average decrease in utility attributable to mS+RT vs RT-alone on an average day for the average patient. I then quantified the average severity of the surgical penalty by specifying the number of dysfunctions listed in Table 2.1. Rather than make a distributional assumption for the surgical penalty, I conducted sensitivity analysis. I am 97.5% confident that mS+RT provides net HRQoL benefit for patients with poor prognosis when the surgical penalty is less than 1.78 dysfunctions for baseline non-ambulatory patients and 3.10 dysfunctions for ambulatory patients.

7.4 Future Work

The surgical penalty is an abstract concept. Decision-analysis of QALYs could be better conducted through a high-quality comparative study of mS+RT and RT-alone that collects patient-reported HRQoL outcomes. Unfortunately, the utility data obtained through the vignette study described in Chapter 3 has limited application for this purpose.[252] There now exist better tools for this purpose.

The Spine Oncology Study Group (SOSG) has developed a spine oncology-specific outcome questionnaire (SOSGOQ).[217] The original questionnaire consisted of 27 items belonging to five domains: physical function, neurologic function, pain, mental health and social function. The tool has face validity based on expert opinion.[217] SOSGOQ also has demonstrated content validity through high correlation International Classification of Functioning and EQ-5D.[217, 111] A revised version of the tool, SOSGOQ2.0, has recently undergone successful psychometric validation.[242] However, there currently exists no mapping of this tool to utilities. I am collaborating with the AOSpine Knowledge Forum Tumor to create a SOAP module for a shortened version of the SOSGOQ2.0, SOSGOQ-8D with

the aim of conducting a general population valuation study. More recently, discrete choice experiments (DCEs) have been used for online utility valuation because cognitive burden is lessened.[157, 112] We will utilize the EuroQol EQ-VT protocol for DCE, implemented on the SOAP platform, to measure utilities for SOSGOQ-8D.[161, 162]

It is important to appreciate that the MESCC utilities reported in Chapter3 were valued by healthy individuals (*ex ante utilities*). MESCC utilities from healthy individuals may not be equivalent to *ex post* utilities from patients who have experienced MESCC. It has been reported that healthy individuals provide lower utility valuations than patients who have experienced living with a colostomy, menopausal symptoms, and hemodialysis.[15, 256, 46] However, the impact of experience may be transient as evidenced by former colostomy patients providing utility valuations equivalent to healthy individuals.[205] Furthermore, in the case of depression, healthy individuals provide higher utility valuations than patients.[182] Currently the understanding of differences between *ex ante* and *ex post* utilities is insufficient to specify general conversion rules.[52]

Rawls argues that *ex ante* utilities can be used ethically if valued under a “veil of ignorance.” If we assume that the general population providing *ex ante* utility valuations may eventually develop the condition of interest, out of self-interest, they should provide fair valuations.[187] Although *ex ante* utilities are theoretically restricted to system policy decisions, *ex ante* utilities have become the *de facto* standard for individual patient decisions. Utilities obtained from generic health surveys such as the EuroQol-5D, Short Form-6D, and Health Utilities Index Mark 3 are actually *ex ante* valuations.[55] Therefore our use of *ex ante* utilities conforms with conventions in the literature. However, it would be informative to conduct a utility valuation study with MESCC patients and derive *ex post* utilities. These utilities could be used to create a true patient decision aid.

7.5 Conclusion

In summary, my thesis explored prognostic contraindications to surgery for MESCC and determined that existing guidelines based on expert opinion are overly restrictive. As part of this thesis, I presented solutions to two barriers to decision-analysis of QALYs in MESCC: (a) lack of utility data and (b) skepticism regarding decision-analysis. These barriers are not unique to MESCC. It is my hope that this thesis and the methodological contributions will be disseminated to facilitate decision-analysis of QALYs in other disease contexts.

APPENDICES

Appendix A

Appendices for Chapter 2

A.1 SOAP MESCC Module

After logging-in, respondents are first presented with a description of MESCC and its functional consequences (Figure A.1). Next, the basic premise of the standard gamble task is outlined (Figure A.2). The best outcome (perfect health) is then outlined (Figure A.3), and this outcome is encapsulated as “Chris” whose characteristics are easily accessible in the remainder of the survey (Figure A.4). The flow the website is then outlined (Figure A.5).

The standard gamble task begins with a tabular description of the probe health state (Figure A.6). Participants then move to a screen with the probability wheel (Figure A.7). Rejecting a lottery is indicated by “Reject the Pill,” accepting the lottery by “Take the Pill,” and indifference to a lottery by “I Can’t Decide if the Risk is Worth it.” A description of the best outcome is accessible by clicking on “Chris” (Figure A.8). The probe health state is also accessible (Figure A.9).

The search algorithm presents the following sequence of probabilities for the best and worst outcome: 100-0, 10-90, 90-10, 20-80, 80-20,...50-50. The first iteration always proposes a lottery with a 100% probability of the best outcome. If the participant does not accept the lottery, a pop-up window appears which highlights the potential logical error with an explanation (Figure A.10). If the participant still refuses the lottery, the probe health state is assigned a utility of 1.

The search algorithm stops presenting lotteries when a lottery with a probability of the best outcome above 50% is rejected, or a lottery with the probability of the best outcome below 50% is accepted. The algorithm also stops if participants are indifferent to a lottery.

If the participant rejects a lottery above the 50% probability, the utility of the probe health state is imputed to be:

$$\frac{\text{probability best outcome}}{100} + 0.05$$

If the participant accepts a lottery below the 50% probability, the utility of the probe

health state is imputed to be:

$$\frac{\text{probability best outcome}}{100} - 0.05$$

If the participant is indifferent to a lottery, the utility of the probe health is imputed to be the probability of the best outcome.

Introduction

- Patients with spine cancer usually live less than 5 years
- Spine cancer can cause problems:
 1. Inability to eat, dress, wash, and use get to the toilet
 2. Problems walking
 3. Loss of bladder control
 4. Pain
 5. Uncomfortable symptoms
- Patients with spine cancer have to make a difficult decision:
 - A. Reject treatment and put-up with the problems until they die of cancerOR
 - B. Take a risky treatment that will either (1) fix their problems, or (2) hurt them

Next

Figure A.1: Description of impairments caused by MESCC.

Your Job

- You will put yourself in the shoes of someone who is 60 years old with spine cancer
- We will give you THREE examples of problems that spine cancer patients have
- Imagine that the examples describe your life
- Like these patients, you will have to make a difficult decision:
 - A. Reject a Pill, put-up with your problems, and live this way for 5 years
 - OR
 - B. Take a Pill, but you are making a gamble where two things might happen:
 1. If the Pill works, your problems are fixed right away, there are no side effects, and you live for 5 years
 2. If the Pill doesn't work, you die right away

Back

Next

Figure A.2: Explanation of Standard Gamble task.

What happens IF the Pill works?

- If the Pill works, your problems will be fixed and you will live like [Chris](#) for the next 5 years
 - You can always click on Chris button to show the description of Chris' life
 - Please click on the Chris button now
- Please remember!
 - the Pill will NOT lengthen your life, you will still only live 5 years (at which point you will die from cancer)
 - The hard part is deciding whether the example problems are so difficult for you that you would take a Pill that could kill you
 - The purpose of this survey is to measure how much risk you are willing to take

Back

Next

Figure A.3: Caveats to and nuances of best possible outcome.

What happens IF the Pill works?

- If the Pill works, your problems will be fixed and you will live like **Chris** for the next 5 years

- You can always click on Chris button
- Please click on the Chris button now

- Please remember!

- the Pill will NOT lengthen your life, (from cancer)
- The hard part is deciding whether to take a Pill that could kill you
- The purpose of this survey is to measure

Chris is 60 years old and has spine cancer. Chris will only live 5 more years...

Things are going well, Chris:

- doesn't need help with eating, dressing, or washing.
- can walk without assistance
- has no trouble controlling her/his bladder
- has back pain relieved by pain medications
- has no other uncomfortable symptoms

Close

Figure A.4: Description of perfect health.

Now it is your turn to do the survey!

- You will make decisions for THREE examples of problems
 - A pop-up window will tell you when the example changes
 - Read the description, imagine it describes you, and compare your problems to Chris' life
- For each example, you may be asked to make decisions on up to 9 different Pills (Pill A, Pill B, Pill C... up to Pill J)
 - Each Pill has a DIFFERENT chance of killing you
 - For each Pill, decide how much risk of death you are willing to take

Back

Next

Figure A.5: Instructions on navigating webpages.

1 REJECT PILL A

Read Example #1 carefully.

In Example #1 you are 60 years old, you have spine cancer and...

Some things are going well, just like Chris. If you reject the Pill these don't change:

But you also have some problems, which Chris does not have. IF the Pill works it fixes these problems:

- You NEED help with eating, dressing, or washing yourself.
- You have LOST mobility because of weakness in your legs. You NEED TO stay in bed or a chair (or wheelchair) during the day
- You NEED TO insert a tube in your bladder to control your bladder
- You have back pain NOT relieved by pain medications
- Everyday, you HAVE ONE OR MORE of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea

For the following questions, imagine Example #1 describes you

Continue

Figure A.6: Probe health state in tabular form.

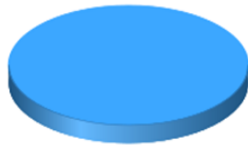
1 REJECT PILL A

If you do not take **Pill A**, you stay in Example #1 until you die of cancer in 5 years

[Click here for a reminder of Example #1...](#)

2 TAKE PILL A

Pill A has a 100% chance of allowing you to live like **Chris** until you die of cancer in 5 years



And there is a 0% chance that **Pill A** will kill you right away

WHICH DO YOU CHOOSE?



(1) REJECT the Pill



(2) TAKE the Pill



I Can't Decide if the Risk is Worth It



I Want to Skip this Example

Figure A.7: Probability wheel.

1 REJECT PILL A

If you do not take **Pill A**, you stay in Example #1 until you die of cancer in 5 years

[Click here for a reminder of Example #1...](#)

2 TAKE PILL A

Pill A has a 100% chance of curing you to live like **Chris** until you die of cancer in 5 years

Chris is 60 years old and has spine cancer. Chris will only live 5 more years...

Things are going well, Chris:

- doesn't need help with eating, dressing, or washing.
- can walk without assistance
- has no trouble controlling her/his bladder
- has back pain relieved by pain medications
- has no other uncomfortable symptoms

chance that
right away

Close



(1) REJECT the Pill



(2) TAKE the Pill



I Can't Decide if the Risk is Worth It



I Want to Skip this Example

Figure A.8: Ability to review definition of perfect health.

1 REJECT PILL A

If you do not take **Pill A**, you stay in Example #1 until you die of cancer in 5 years

[Hide the details of Example #1](#)

In Example #1 you are 60 years old, you have spine cancer and...

Some things are going well, just like Chris. If you reject the Pill these don't change:

But you also have some problems, which Chris does not have. IF the Pill works it fixes these problems:

- You NEED help with eating, dressing, or washing yourself.
- You have LOST mobility because of weakness in your legs. You NEED TO stay in bed or a chair (or wheelchair) during the day
- You NEED TO insert a tube in your bladder to control your bladder
- You have back pain NOT relieved by pain medications
- Everyday, you HAVE ONE OR MORE of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea

2 TAKE PILL A

Pill A has a **100%** chance of allowing you to live like **Chris** until you die of cancer in 5 years



And there is a **0%** chance that **Pill A** will kill you right away

Figure A.9: Ability to review probe health state.

1 REJECT PILL A

If you do not take **Pill A**, you stay in Example #1 until you die of cancer in 5 years

[Click here for a reminder of Example #1](#)

A 100% chance of Pill A working means there is NO risk.

If you take Pill A you will CERTAINLY live like Chris for 5 years.

Chris's life is better than your Example life.

You should take Pill A.

Do you want to change your answer?

Yes

No

☐ I Want to Skip this Example

Figure A.10: Logic check.

A.2 SOAP platform

Assessment Approach

Scaling Method

The time-trade-off and standard gamble can be used to directly elicit utilities. However, it was felt that framing health scenarios in the time-trade-off would be difficult as the life-expectancy of patient with MESCC is less than five years. Therefore, the standard gamble was used for the SOAP MESCC module.

Frame

Typically, in the standard gamble success is framed as “perfect” health for an undermined period of time. This anchor is not appropriate for MESCC because metastatic cancer cannot be cured, and life expectancy is relatively short. Therefore the fully functional health state, D-N-I-P-S- was chosen as the success anchor. To eliminate confusion around life expectancy, all scenarios were framed as having a certain life expectancy of five years; that is for both the probe health scenario, and success health scenario, participants were told their life expectancy would certainly be five years.

Search Procedure

For each health state, it is necessary to vary the probabilities of the gamble until participants are indifferent to the lottery. Three “search procedures” have been described for varying probabilities:

1. **Ping-Pong Variant (PPV).** This is the classic search algorithm proposed by von Neumann and Morgenstern.[243] In contemporary applications, the gambles oscillate between extremes of probability: 100-0, 10-90, 90-10, 20-80, 80-20,... 50-50. The algorithm stops when a gamble with success above 50% is rejected, or a gamble with success below 50% is accepted. The algorithm also stops if participants are indifferent to a gamble. With 10% decrements, the algorithm iterates a maximum of 11 times.[76]
2. **Titration Up Variants.** The Titration Up Variant (TUV) algorithm was first described in a transportation economics paper.[116] Here the probability of success progressively decreases: 100-0, 90-10, 80-20,... 0-100. In the Titration Down Variant (TDV), the probability of success increases: 0-100, 10-90, 20-80,... 100-0.
3. **Bisection Variant (BV).** This algorithm was first applied in U-titer, a computerized utility elicitation tool.[219] The algorithm is applied by offering 100-0, 0-100 and then successively bisecting the difference between the probe gamble and either 100 or 0, depending on whether the gamble was accepted or rejected. For example,

if 50-50 was rejected, the next gamble would be 75-25; while 25-75 would be posed if the gamble was accepted. An arbitrary limit is usually place on the number of iterations. To obtain a precise result, it is necessary to use at least six iterations; consequently, respondent fatigue can be an issue with this procedure.

Unfortunately, there are no guidelines for choosing among the four algorithms. Moreover, three studies have demonstrated that the titration variants and PPB do not produce equivalent results (the bisecting algorithm has not yet been compared to other methods). In a computer utility elicitation study, precision and utility estimates were found to be higher (mean difference between 0.1 to 0.15) with the BUV compared to the PPV.[137] Comparing interviewer-administered PPV to self-administered paper-based TDV, Hammerschmidt found PPV to generate higher utility estimates than TDV.[97] In the same study, no significant difference was found between estimates using the TDV and BUV. Another study compared interviewer led PPV and TDV, and reported utility estimates to be higher with the TDV.[17] These three papers are heterogeneous in methods and results, making it difficult to synthesize a coherent recommendation. Brazier and Dolan, however, offer a sensible conclusion on the differences between the TDV and PPV.[17] They suggest that based on the starting-point, the TDV anchors participants to the upper end of the utility scale. Alternatively, they suggest the PPV may confuse participants and encourage risk taking. However, the alternative conclusion is not supported by their results as internal consistency was equivalent in both the TDV and PPV groups. Anchoring is further supported by Lenert's results as PPV produced higher utility estimates than the BUV. Although Hammerschmidt's results counter the presence of anchoring, they are likely confounded by the mode of administration. In summary, there is weak evidence for anchoring with titration which is plausible given the ordering of risks posed by the search procedures. We selected the PPV because it is the classical search procedure and may protect against anchoring.

Navigation Rules

Sequence Specification

The SOAP tool is flexible by allowing researchers to vary the number and order of health scenarios administered to participants. Each scenario asks participants to rate up to ten different combinations of probabilities.[76]

Error Checking

Participants may go back one iteration to correct an error. The first iteration always proposes a lottery with 100% chance of success. If participants do not choose to take treatment, a pop-up window appears which identifies the logical error, and explains why this choice was not logical.

Auxiliary Functions

Questionnaire Administration

Traditionally, direct valuation administered face-to-face by an interviewer using props to illustrate probabilities.[76] Clearly, this method is time-intensive, and presents logistical difficulties in recruiting a representative general population sample. Consequently, there is a growing body of literature utilizing developed self-directed standard gamble elicitation exercises over the internet.

Data Management

SOAP has been developed to work with Microsoft Internet Explorer, Google Chrome, Mozilla Firefox, Apple Safari, and the Blackberry Browser on Windows, Apple OS, iOS, Linux and Blackberry desktop and touch-screen configurations. The website can be run with the most restrictive security settings, and does not generate any security errors. Participants login with a secure password. The first set of web pages prepare the participant for the task. The second set of pages present the standard gambles. The last section elicits demographic information. SOAP is fully customizable, and the ordering can be easily changed. SOAP was programmed in the open-source PHP and JavaScript language. These languages were chosen as it is unlikely they will be depreciated in the future. Results are stored in an open-source mySQL database, however, most major databases are supported.

Training

The first set of web pages prepare the participant for the task.

A.3 MESCC Health States

D-N-I-P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N-I-P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D-N-I-P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N-I-P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D-N-I+P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N-I+P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D-N-I+P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N-I+P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D-N+I-P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N+I-P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D-N+I-P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N+I-P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D-N+I+P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N+I+P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D-N+I+P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D-N+I+P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You do not need help with eating, dressing, washing yourself, or getting on the toilet
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N-I-P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N-I-P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N-I-P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N-I-P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N-I+P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N-I+P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N-I+P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N-I+P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You walk without assistance.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N+I-P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N+I-P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N+I-P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N+I-P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You have no trouble controlling your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N+I+P-S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N+I+P-S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

D+N+I+P+S-

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- You have no other uncomfortable symptoms.

D+N+I+P+S+

In Scenario X you are 60 years old, you have spine cancer and...

- You need help with eating, dressing, washing yourself, or getting on the toilet.
- You have lost mobility because of weakness in your legs. You need to stay in bed or a chair (or wheelchair) during the day.
- You need to insert a tube in your bladder to control your bladder.
- You have back pain not relieved by pain medications.
- Everyday, you have one or more of: nausea, vomiting, shortness of breath, lack of appetite, or diarrhea.

Appendix B

Appendices for Chapter 3

B.1 Regression Model for Utilities

We wish to relate utility values to a set of predictors using regression. A general linear model (GLM) is inappropriate for two reasons. The first, utilities can only take on values in the closed interval $[0, 1]$ whereas a GLM supports $(-\infty, \infty)$. The second, the GLM assumes homoskedasticity which is not a reasonable assumption for utilities.[50] By virtue of the fact that utilities cannot be greater than 1 or less than 0, dispersion in valuations for health states with mean 0.5 will be greater than dispersion in valuations for health states at the extremes near 0 and 1.

Generalized Linear Models (GzLMs) based on the beta distribution, which supports the open interval $(0, 1)$, offer a potential solution to the limitations of GLMs for utilities. However, adjustments are required as utilities can include the values 0 and 1. One solution is to transform utilities to the closed unit interval:[207]

$$y^* = \frac{y(n-1) + 0.5}{n}.$$

An alternative strategy is to use a beta-inflated model which is a mixture model of three separate regression models: the standard beta model and binomial models for values of 0 and 1.[165] We elected to use the simpler and more readily interpretable beta model with utilities transformed to the closed unit interval.

The Beta Distribution

The beta probability density function (pdf) is an equation which describes the contour of the histogram for a beta distributed random variable Y . [67, 115, 149] The relative frequencies generated by a pdf, $f(y)$, are termed probability densities which for the beta distribution satisfy

$$\int_0^1 f(y)dy = 1$$

Using this definition, the probability that $Y \leq y$ can be computed using the cumulative distribution function (cdf) $F(y)$

$$\Pr(Y \leq y) = F(y) = \int_0^y f(y)dy$$

The beta pdf is a function of two shape parameters: $p \in \mathbb{R}^+$ and $q \in \mathbb{R}^+$. Both parameters influence the mean ($E[Y]$), variance ($\text{Var}[Y]$), skewness, and kurtosis (peakedness and tail weight) of the beta pdf.[11, 170] When $p = q$ the beta pdf is symmetric about $Y = 0.5$, and larger values of the equivalent shape parameters increase peakedness while decreasing tail-weight. When $p < q$, the beta pdf is right-skewed, and left-skewed when $p > q$. When $p \neq q$, p tends to control the mode, and q tends to control the tail weight.

The mean of a beta distributed variable Y is

$$E[Y] = \mu = \frac{p}{p+q}$$

The variance of a beta distributed variable Y is

$$\text{Var}[Y] = \frac{pq}{(p+q)^2(p+q+1)}$$

For exponential family of distributions, such as the beta, the variance can be partitioned as

$$\text{Var}[Y] = V(\mu) \sigma$$

where V is the variance function and σ is the precision parameter. For exponential-family distributions, V is a function of the mean μ , and for the beta distribution is

$$V(\mu) \sigma = \mu(1-\mu).$$

It is through the variance function that heteroscedasticity can be modeled with the beta distribution because variance is smaller at the upper and lower bounds of the distribution. While the precision parameter

$$\sigma = 1 + p + q$$

controls the relative dispersion for a given mean.

Beta Regression

The beta GzLM[67, 149] relates the mean for a set of n covariates x_i to a linear predictor using an inverse link function g^{-1} :

$$E[Y] = \mu = g^{-1} \left(\sum_{i=1}^n \beta_i x_i \right)$$

The logit function is often used for g^{-1} for beta regression. The dispersion parameter σ is often treated as a constant for beta regression.

B.2 Quota Samples Based on 2016 Canadian Census

Table B.1: Quotas Based on the 2016 Canadian Census

Region	Gender	Age	Pop.	% Cnd. Pop.	Quota
Maritimes	Female	18-24	98041	0.444064	3
Maritimes	Female	25-34	138965	0.629423	4
Maritimes	Female	35-44	152810	0.692133	5
Maritimes	Female	45-54	184025	0.833517	5
Maritimes	Female	55-64	186327	0.843943	6
Maritimes	Female	65+	242201	1.097017	7
Maritimes	Male	18-24	104383	0.472789	3
Maritimes	Male	25-34	142078	0.643523	4
Maritimes	Male	35-44	145154	0.657456	4
Maritimes	Male	45-54	178200	0.807133	5
Maritimes	Male	55-64	178502	0.808501	5
Maritimes	Male	65+	203558	0.921989	6
Ontario	Female	18-24	655894	2.970785	19
Ontario	Female	25-34	953313	4.317904	28
Ontario	Female	35-44	923760	4.184048	27
Ontario	Female	45-54	1033308	4.680231	31
Ontario	Female	55-64	930168	4.213072	27
Ontario	Female	65+	1219557	5.523821	36
Ontario	Male	18-24	689380	3.122455	20
Ontario	Male	25-34	921938	4.175795	27
Ontario	Male	35-44	882531	3.997307	26
Ontario	Male	45-54	1023430	4.63549	30
Ontario	Male	55-64	892852	4.044054	26
Ontario	Male	65+	991595	4.491298	29
Manitoba	Female	18-24	62942	0.285087	2
Manitoba	Female	25-34	90091	0.408055	3
Manitoba	Female	35-44	83631	0.378795	2
Manitoba	Female	45-54	86261	0.390708	3
Manitoba	Female	55-64	79745	0.361194	2
Manitoba	Female	65+	105817	0.479284	3
Manitoba	Male	18-24	65881	0.298399	2
Manitoba	Male	25-34	91173	0.412956	3
Manitoba	Male	35-44	80922	0.366525	2
Manitoba	Male	45-54	88596	0.401284	3
Manitoba	Male	55-64	81220	0.367875	2
Manitoba	Male	65+	86100	0.389978	3
Saskatchewan	Female	18-24	51435	0.232968	2
Saskatchewan	Female	25-34	83327	0.377419	2

Continued on next page

Table B.1 – *Continued from previous page*

Region	Gender	Age	Pop.	% Cnd. Pop.	Quota
Saskatchewan	Female	35-44	68416	0.309881	2
Saskatchewan	Female	45-54	71431	0.323537	2
Saskatchewan	Female	55-64	70679	0.320131	2
Saskatchewan	Female	65+	90911	0.411769	3
Saskatchewan	Male	18-24	57438	0.260158	2
Saskatchewan	Male	25-34	87269	0.395273	3
Saskatchewan	Male	35-44	72312	0.327528	2
Saskatchewan	Male	45-54	74792	0.33876	2
Saskatchewan	Male	55-64	72894	0.330164	2
Saskatchewan	Male	65+	74983	0.339626	2
Alberta	Female	18-24	190216	0.861558	6
Alberta	Female	25-34	347060	1.571962	10
Alberta	Female	35-44	297495	1.347464	9
Alberta	Female	45-54	276512	1.252424	8
Alberta	Female	55-64	244383	1.1069	7
Alberta	Female	65+	263549	1.19371	8
Alberta	Male	18-24	203940	0.923719	6
Alberta	Male	25-34	364870	1.65263	11
Alberta	Male	35-44	318403	1.442164	9
Alberta	Male	45-54	296531	1.343098	9
Alberta	Male	55-64	256490	1.161737	8
Alberta	Male	65+	223618	1.012848	7
British Columbia	Female	18-24	212117	0.960756	6
British Columbia	Female	25-34	320180	1.450213	9
British Columbia	Female	35-44	308188	1.395897	9
British Columbia	Female	45-54	347457	1.57376	10
British Columbia	Female	55-64	330475	1.496843	10
British Columbia	Female	65+	434352	1.96734	13
British Columbia	Male	18-24	231772	1.04978	7
British Columbia	Male	25-34	314529	1.424617	9
British Columbia	Male	35-44	302876	1.371837	9
British Columbia	Male	45-54	335889	1.521365	10
British Columbia	Male	55-64	323291	1.464304	10
British Columbia	Male	65+	383712	1.737972	11

B.3 Regression Modeling

Fitting

We considered three base-case regression models for the linear predictor (Model 1, Model 2 and Model 3) (Table B.2). A random intercept for each participant i was incorporated in the models. We considered ϕ a fixed effect and constant in each model.

Prior to comparing models, we checked for misspecification of the variance and/or link function. The variance function was checked by comparing normalized quantile residuals against fitted values (Figure B.1). As utilities lie in the closed interval $[0, 1]$, residuals also lie in an interval which for fitted values of 0 is $(0, 1)$ and for fitted values of 1 is $(-1, 0)$. The smoothed line indicated that the mean residual for all fitted values was 0 and there was no trend. Therefore, we concluded the variance function was properly specified. The link function was checked by comparing the adjusted dependent variable against linear predictor for μ (Figure B.2). Each plot demonstrated a straight line supporting the choice of link function. Of the three base-case regression models, Model 2 was optimal as it had the lowest AIC and BIC (Table B.3).

A plot of these nominal coefficients against the number of dysfunctional attributes revealed a non-linear concave up relationship (Figure B.3(a)). Tukey's and Mosteller's Bulging Rule guided that a square root (Model 4) or logarithmic transformation (Model 5) of the number of dysfunctional attributes could linearize the relationship (Figure B.3(b) and (c)).[235] The beta regression model was re-fitted using Model 4 and Model 5. Model 4 was optimal yielding the lowest AIC and BIC (Table B.3).

Formulas for the linear predictor in the beta regression are provided in Table B.6. Mean utilities for each health state estimated by each regression model are shown in Table B.5. Calibration and agreement parameters for all candidate models are provided in Table B.4 and Figures B.4, B.5, B.6, 3.2, and B.7. We will calculate mean utilities for the health state D-N+I-P+S- to illustrate the use of the formulas listed in Table B.6.

For Model 1, this health states is coded as: D= 0, N= 1, I= 0, P= 1, S= 0. The mean utility valuation is calculated as

$$\text{inverse logit } (0.90 - 0.32(0) - 0.39(1) - 0.26(0) - 0.40(1) - 0.43(0)) = 0.53.$$

For Model 2, this health states is coded as: two dys= 1, three dys= 0, four dys= 0, five dys= 0. The mean utility valuation is calculated as:

$$\text{inverse logit } (0.59 - 0.44(1) - 0.84(0) - 1.12(0) - 1.27(0)) = 0.54.$$

For Model 3, this health states is coded as: D= 0, N= 1, I= 0, P= 1, S= 0, D×N= 0, D×I= 0, D×P= 0, D×S= 0, N×I= 0, N×P= 1, N×S= 0, I×P= 0, I×S= 0, P×S= 0. The mean utility valuation is calculated as:

$$\begin{aligned} \text{inverse logit } (1.18 - 0.58(0) - 0.59(1) - 0.64(0) - 0.45(1) - 0.62(0) \\ + 0.17(0) + 0.16(0) - 0.03(0) + 0.21(0) + 0.32(0) \\ - 0.01(1) - 0.12(0) + 0.05(0) + 0.20(0) + 0.05(0)) = 0.53. \end{aligned}$$

For both Model 4 and Model 5 the health state is coded as: num dys= 2. Using Model 4 the mean utility valuation is calculated as:

$$\text{inverse logit} \left(1.71 - 1.11 \times \sqrt{2} \right) = 0.53.$$

Using Model 5 the mean utility valuation is calculated as:

$$\text{inverse logit} (0.64 - 0.81 \times \ln 2) = 0.52.$$

Table B.2: Functional Form of Linear Predictor Formulas for Beta Regression

	Functional form
Model 1	$\beta_0 + \beta_1 (D) + \beta_2 (N) + \beta_3 (I) + \beta_4 (P) + \beta_5 (S) + \epsilon_i$
Model 2	$\beta_0 + \beta_1 (\text{two dys}) + \beta_2 (\text{three dys}) + \beta_3 (\text{four dys}) + \beta_4 (\text{five dys}) + \epsilon_i$
Model 3	$\beta_0 + \beta_1 (D) + \beta_2 (N) + \beta_3 (I) + \beta_4 (P) + \beta_5 (S) + \beta_6 (D \times N) + \beta_7 (D \times I) + \beta_8 (D \times P) + \beta_9 (D \times S) + \beta_{10} (N \times I) + \beta_{11} (N \times P) + \beta_{12} (N \times S) + \beta_{13} (I \times P) + \beta_{14} (I \times S) + \beta_{15} (P \times S) + \epsilon_i$
Model 4	$\beta_0 + \beta_1 \sqrt{\text{num dys}} + \epsilon_i$
Model 5	$\beta_0 + \beta_1 \ln (\text{num dys}) + \epsilon_i$

Table B.3: Selection Parameters for all Models

	Description	AIC	BIC
Model 1	Attributes	-2750.24	-899.148
Model 2	Number of dysfunctions (nominal variable)	-2751.78	-906.259
Model 3	Attributes and second-order interactions	-2747.42	-840.572
Model 4	Square root of number of dysfunctions (continuous variable)	-2756.85	-928.061
Model 5	Logarithm of number of dysfunctions (continuous variable)	-2754.9	-926.11

Table B.4: Calibration and Agreement Parameters for all Candidate Models

	Mean Absolute Error	Mean Signed Difference	Intraclass Correlation Coefficient (95% CI)
Model 1	0.046036	-0.0161	0.931 (0.857, 0.967)
Model 2	0.046548	-0.01967	0.937 (0.863, 0.97)
Model 3	0.053878	-0.03416	0.917 (0.766, 0.965)
Model 4	0.046582	-0.01934	0.936 (0.86, 0.97)
Model 5	0.049875	-0.02303	0.927 (0.834, 0.966)

Table B.5: Comparison of Model-Based Mean Utilities for all MESCC Health States

Health State	Observed Mean	Model 1	Model 2	Model 3	Model 4	Model 5
D+N-I-P-S-	0.67	0.64	0.644	0.645	0.644	0.654
D-N+I-P-S-	0.73	0.624	0.644	0.644	0.644	0.654
D-N-I+P-S-	0.7	0.655	0.644	0.633	0.644	0.654
D-N-I-P+S-	0.71	0.622	0.644	0.676	0.644	0.654
D-N-I-P-S+	0.71	0.615	0.644	0.636	0.644	0.654
D+N+I-P-S-	0.5	0.546	0.538	0.543	0.533	0.519
D+N-I+P-S-	0.6	0.579	0.538	0.531	0.533	0.519
D+N-I-P+S-	0.56	0.544	0.538	0.529	0.533	0.519
D+N-I-P-S+	0.58	0.537	0.538	0.493	0.533	0.519
D-N+I+P-S-	0.68	0.562	0.538	0.567	0.533	0.519
D-N+I-P+S-	0.54	0.527	0.538	0.533	0.533	0.519
D-N+I-P-S+	0.52	0.519	0.538	0.461	0.533	0.519
D-N-I+P+S-	0.56	0.56	0.538	0.538	0.533	0.519
D-N-I+P-S+	0.61	0.553	0.538	0.531	0.533	0.519
D-N-I-P+S+	0.55	0.518	0.538	0.539	0.533	0.519
D+N+I+P-S-	0.47	0.482	0.438	0.504	0.445	0.437
D+N+I-P+S-	0.41	0.447	0.438	0.421	0.445	0.437
D+N+I-P-S+	0.51	0.439	0.438	0.36	0.445	0.437
D+N-I+P+S-	0.36	0.48	0.438	0.425	0.445	0.437
D+N-I+P-S+	0.51	0.472	0.438	0.426	0.445	0.437
D+N-I-P+S+	0.4	0.437	0.438	0.387	0.445	0.437
D-N+I+P+S-	0.5	0.463	0.438	0.466	0.445	0.437
D-N+I+P-S+	0.39	0.455	0.438	0.431	0.445	0.437
D-N-I+P+S+	0.48	0.453	0.438	0.444	0.445	0.437
D-N+I-P+S+	0.49	0.42	0.438	0.362	0.445	0.437
D+N+I+P-S+	0.37	0.377	0.37	0.37	0.373	0.38
D+N+I+P+S-	0.29	0.384	0.37	0.396	0.373	0.38
D+N+I-P+S+	0.39	0.345	0.37	0.265	0.373	0.38
D+N-I+P+S+	0.35	0.375	0.37	0.337	0.373	0.38
D-N+I+P+S+	0.34	0.359	0.37	0.347	0.373	0.38
D+N+I+P+S+	0.3	0.289	0.337	0.285	0.314	0.339

Table B.6: Linear Predictor Formulas with Coefficients

	Equation
Model 1	$0.90 - 0.32 (D) - 0.39 (N) - 0.26 (I) - 0.40 (P) - 0.43 (S)$
Model 2	$0.59 - 0.44 (\text{two dys}) - 0.84 (\text{three dys}) - 1.12 (\text{four dys}) - 1.27 (\text{five dys})$
Model 3	$1.18 + 0.85 (D) - 0.59 (N) - 0.64 (I) - 0.45 (P) - 0.62 (S) + 0.17 (D \times N) + 0.16 (D \times I) - 0.03 (D \times P) + 0.21 (D \times S) + 0.32 (N \times I) - 0.01 (N \times P) - 0.12 (N \times S) + 0.05 (I \times P) + 0.20 (I \times S) + 0.05 (P \times S)$
Model 4	$1.71 - 1.11\sqrt{\text{num dys}}$
Model 5	$0.64 - 0.81 \ln (\text{num dys})$

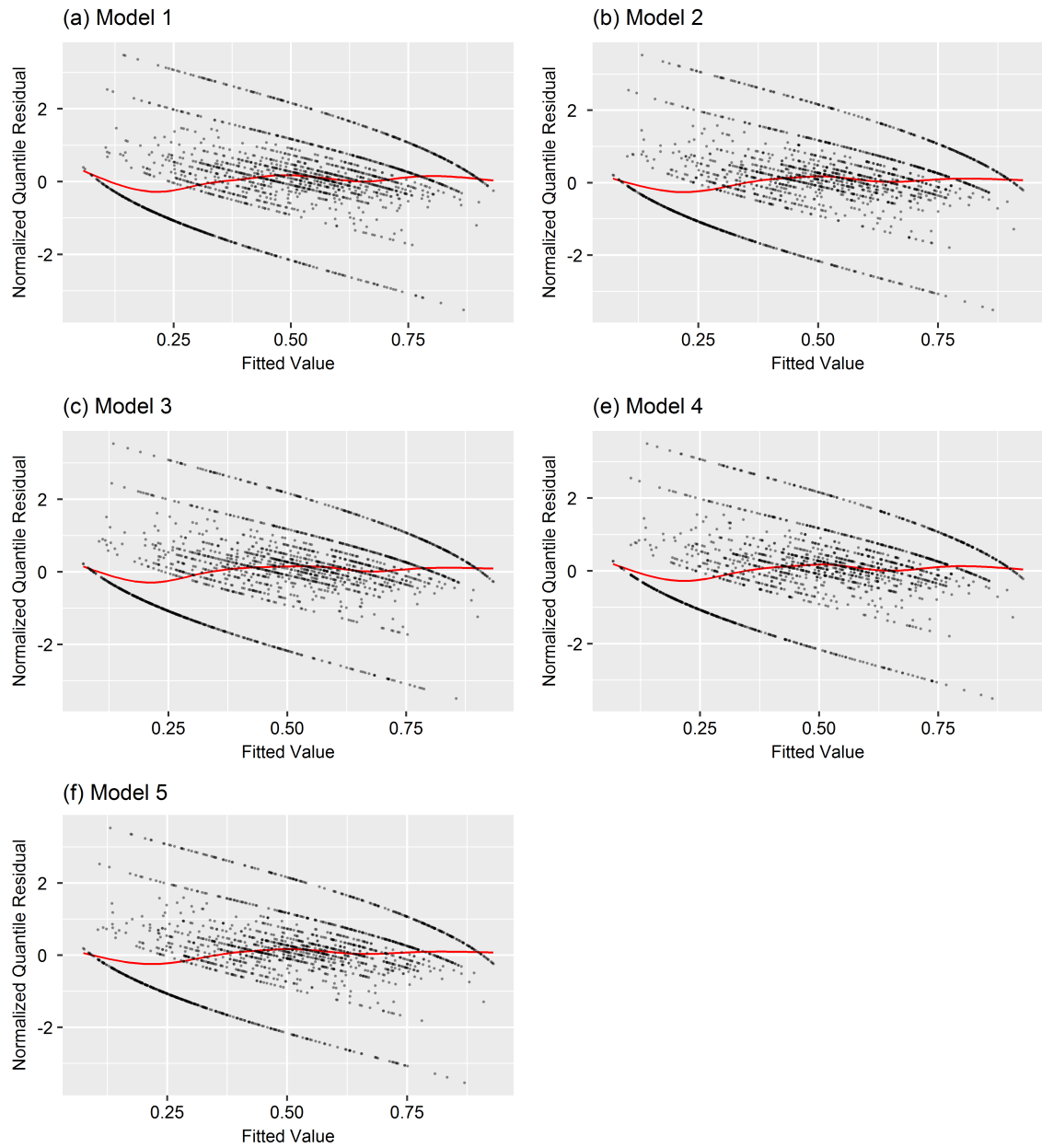


Figure B.1: Variance function check.

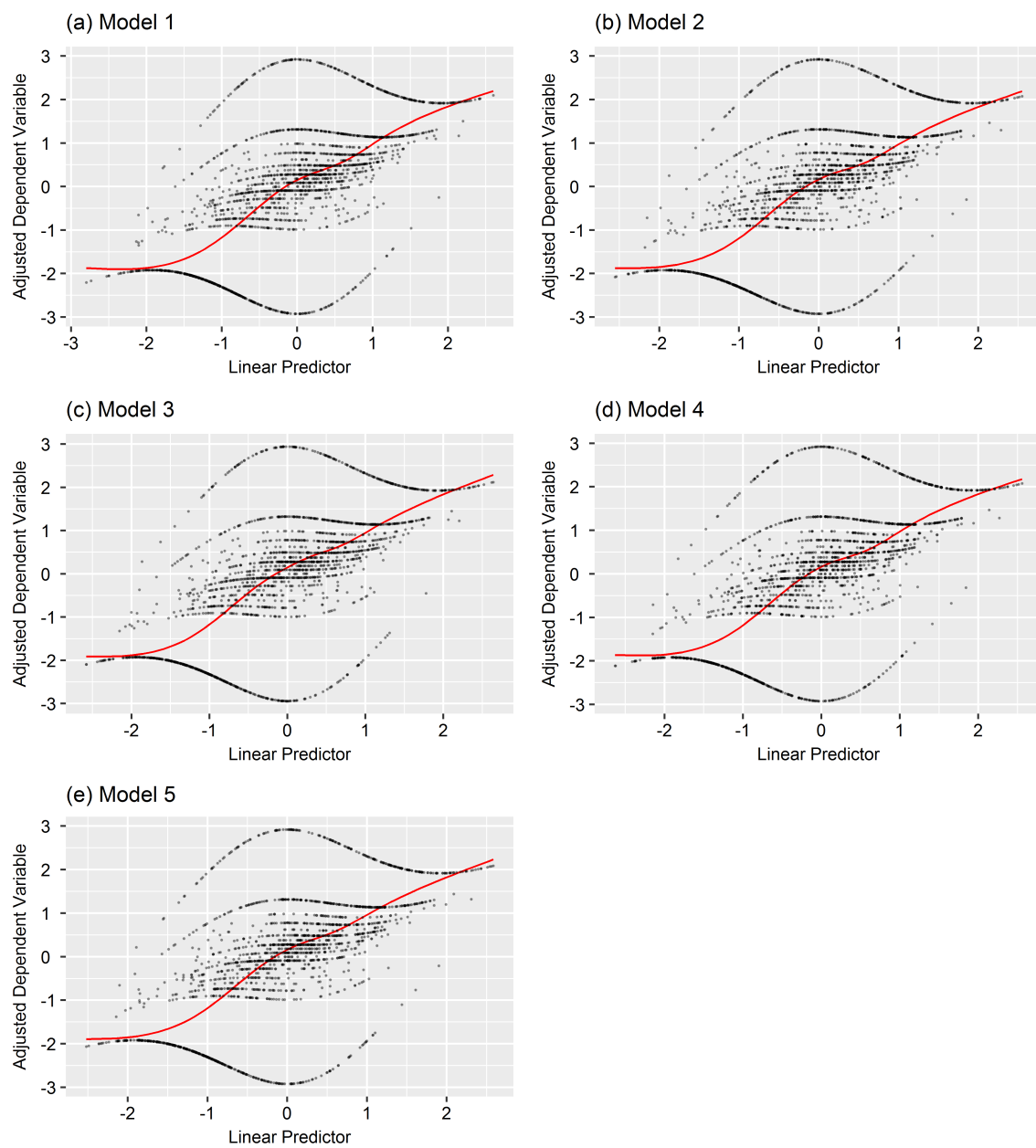


Figure B.2: Link function check.

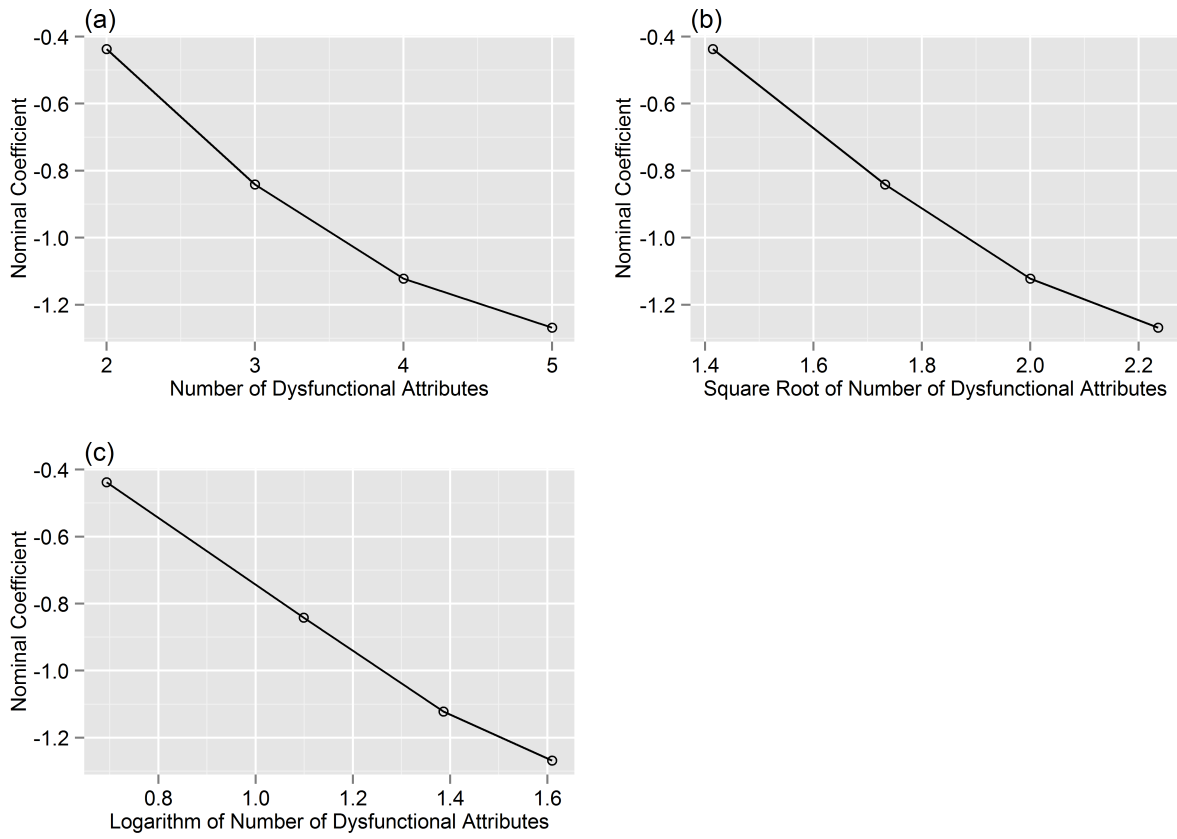
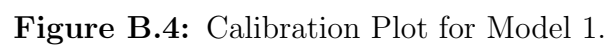
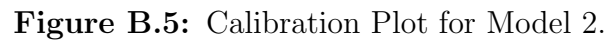
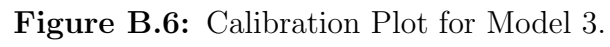
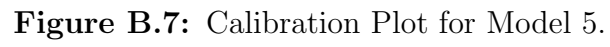


Figure B.3: Plots of nominal coefficients against the number of dysfunctions. No transformation on number of dysfunctions (a), square root (b), and logarithmic (c) transformations.









Sensitivity analysis

Our analysis excluded participants deemed to have not have understood or not have engaged in the utility valuation task. Mean and median utility valuations for each health state including these participants are provided in Supplemental Table B.7. These valuations are quite similar to the results obtained using the more restricted dataset (Table 3.2). Therefore, we deemed our analysis not to be sensitive to the inclusion criteria defined *a priori*.

Table B.7: Utilities for all MESCC Health States using Full Dataset

Health State	Number of Dysfunctions	N	Mean (95% CI)	Median (IQR)
D+N-I-P-S-	1	300	0.597 (0.553, 0.641)	0.75 (0.15, 0.95)
D-N+I-P-S-	1	303	0.641 (0.597, 0.682)	0.8 (0.35, 0.95)
D-N-I+P-S-	1	290	0.596 (0.551, 0.638)	0.75 (0.163, 0.95)
D-N-I-P+S-	1	302	0.605 (0.558, 0.649)	0.75 (0.15, 0.95)
D-N-I-P-S+	1	321	0.603 (0.56, 0.646)	0.75 (0.15, 0.95)
D+N+I-P-S-	2	142	0.539 (0.475, 0.601)	0.65 (0.15, 0.95)
D+N-I+P-S-	2	179	0.569 (0.511, 0.624)	0.65 (0.1, 0.95)
D+N-I-P+S-	2	179	0.478 (0.418, 0.534)	0.45 (0, 0.85)
D+N-I-P-S+	2	179	0.545 (0.487, 0.601)	0.65 (0.1, 0.95)
D-N+I+P-S-	2	134	0.591 (0.523, 0.654)	0.675 (0.15, 0.95)
D-N+I-P+S-	2	169	0.522 (0.463, 0.579)	0.6 (0.1, 0.95)
D-N+I-P-S+	2	169	0.495 (0.433, 0.555)	0.5 (0, 0.95)
D-N-I+P+S-	2	163	0.562 (0.502, 0.618)	0.65 (0.125, 0.95)
D-N-I+P-S+	2	152	0.514 (0.451, 0.577)	0.575 (0, 0.95)
D-N-I-P+S+	2	170	0.561 (0.5, 0.619)	0.65 (0.15, 0.95)
D+N+I+P-S-	3	137	0.473 (0.405, 0.54)	0.45 (0, 0.95)
D+N+I-P+S-	3	132	0.481 (0.413, 0.548)	0.45 (0, 0.95)
D+N+I-P-S+	3	142	0.457 (0.391, 0.525)	0.45 (0, 0.938)
D+N-I+P+S-	3	151	0.494 (0.43, 0.556)	0.55 (0, 0.925)
D+N-I+P-S+	3	153	0.521 (0.459, 0.584)	0.55 (0, 0.95)
D+N-I-P+S+	3	158	0.487 (0.424, 0.55)	0.6 (0, 0.95)
D-N+I+P+S-	3	148	0.467 (0.404, 0.528)	0.45 (0, 0.85)
D-N+I+P-S+	3	135	0.433 (0.367, 0.499)	0.45 (0, 0.825)
D-N+I-P+S+	3	130	0.505 (0.433, 0.572)	0.5 (0, 0.95)
D-N-I+P+S+	3	154	0.496 (0.432, 0.559)	0.5 (0, 0.95)
D+N+I+P+S-	4	325	0.432 (0.388, 0.477)	0.45 (0, 0.85)
D+N+I+P-S+	4	308	0.451 (0.407, 0.495)	0.45 (0, 0.85)
D+N+I-P+S+	4	311	0.415 (0.371, 0.46)	0.35 (0, 0.85)
D+N-I+P+S+	4	280	0.409 (0.364, 0.456)	0.35 (0, 0.75)
D-N+I+P+S+	4	270	0.413 (0.367, 0.46)	0.35 (0, 0.85)
D+N+I+P+S+	5	289	0.402 (0.358, 0.449)	0.35 (0, 0.75)

95% Confidence Interval (95% CI). Interquartile range (IQR).

Appendix C

Appendices for Chapter 4

C.1 Confidence Intervals for Median Survival Time

Bayesian analysis of median survival time requires knowledge of the mean and standard deviation of the median survival time (t_m) for each group of the REVEL study and for each stratum in the Lee *et al.* meta-analysis.

For the Lee *et al.* meta-analysis this can be easily done because they report symmetric asymptotic confidence intervals for t_m . These confidence intervals are based on the assumption that t_m follows a normal distribution. Symmetric asymptotic confidence intervals for t_m are based on the relationship:[191]

$$\text{var}[t_m] = \frac{\text{var}[S(t_m)]}{f(t_m)^2}. \quad (\text{C.1})$$

In the original report, t_m was not stratified by Tokuhashi score. We requested this analysis from the study authors:

- a Tokuhashi Score < 8 was associated with a median survival of 5 months (95% CI, 4.356 – 5.644 months); thus $\mu=5$, and $\sigma=(5.644-5)\div 1.96=0.332$.
- a Tokuhashi Score ≥ 8 was associated with a median survival of 14 months (95% CI, 12.666 – 15.334 months); thus $\mu=14$, and $\sigma=(15.334-14)\div 1.96=0.681$.

Many studies, such as the REVEL trial, report asymmetric confidence intervals for t_m . A commonly used technique for computing asymmetric confidence intervals is that described by Brookmeyer and Crowley and was used to analyze the REVEL trial.[25] Brookmeyer and Crowley confidence intervals contain the times t in an α -level confidence interval if they satisfy the relationship:

$$c_\alpha \text{var}[S(t)] \geq (S(t) - 0.5)^2.$$

Table C.1: Parameters used for Interpolation of $\text{var}[S(t_m)]$

	Lower Confidence Limit for t_m		Upper Confidence Limit for t_m	
	DOC	DOC+RAM	DOC	DOC+RAM
t	6.1	8.4	8.5	10.1
$S(t)$	0.54	0.55	0.44	0.45
$\text{var}[S(t)]^\dagger$	0.00042	0.00065	0.00094	0.00065

[†]calculated using Formula C.2.

Where $S(t)$ is the survival fraction at time t ; $\text{var}[S(t)]$ is the Greenwood variance; and c_α is the critical value of the $\chi^2(1)$ distribution with a right tail area of α . Within the Brookmeyer and Crowley interval,

$$\text{var}[S(t)] \geq \frac{(S(t) - 0.5)^2}{c_\alpha}.$$

Thus outside the interval,

$$\text{var}[S(t)] < \frac{(S(t) - 0.5)^2}{c_\alpha}.$$

We therefore approximate the variance of the survival fraction at the confidence limits by,

$$\text{var}[S(t)] \approx \frac{(S(t) - 0.5)^2}{c_\alpha}. \quad (\text{C.2})$$

Greenwood variances computed from the limits of the 95% confidence intervals for t_m reported in the REVEL study are shown in Table C.1. Datapoints for the DOC and DOC+RAM arms were obtained from Figure S1a of the REVEL publication.[188] We use trilinear interpolation of t , $S(t)$ and $\text{var}[S(t)]$ between the confidence limits to estimate $\text{var}[S(t_m)]$ which was 0.00074 and 0.00065 for the DOC and DOC+RAM groups respectively.

We must still compute the probability density function $f(t_m)$ in order to compute $\text{var}[t_m]$ based on Formula C.1. A variety of strategies have been proposed to compute $f(t_m)$. [254] We calculated $f(t_m)$ using the upper and lower confidence limits for t_m , t_m^l and t_m^u , respectively

$$f(t_m) = \frac{S(t_m^l) - S(t_m^u)}{t_m^u - t_m^l}$$

$f(t_m)$ was -0.04 and -0.06 for the DOC and DOC+RAM groups respectively.

Substituting $\text{var}[S(t_m)]$ and $f(t_m)$ into Formula C.1 yields a $\text{var}[t_m]$ of 0.46 and 0.18 for the DOC and DOC+RAM groups respectively.

C.2 Sensitivity Analysis Through Effective Sample Size Adjustment

It is helpful to consider the effective sample instead of the variance. Larger studies will have a smaller variance and larger precision. The Cutler-Ederer effective sample size (n') is a measure of the number of individuals who survived and contributed to $S(t)$, and is related to the Greenwood variance ($\text{var}[S(t)]$) through:[53]

$$n' = \frac{S(t)(1 - S(t))}{\text{var}[S(t)]} \quad (\text{C.3})$$

Since $S(t_m) = 0.5$, the effective sample size at t_m is

$$n'_{t_m} = \frac{0.25}{\text{var}[S(t_m)]}.$$

Therefore

$$\text{var}[S(t_m)] = \frac{0.25}{n'_{t_m}}. \quad (\text{C.4})$$

Substituting Formula C.4 into Formula C.1 yields an expression for the asymptotic variance of median survival time

$$\text{var}[t_m] = \frac{0.25}{n'_{t_m} (f(t_m))^2}. \quad (\text{C.5})$$

For the REVEL trial, we can calculate a baseline effective sample size for each arm because we calculated $f(t_m)$ and $\text{var}[t_m]$ (Appendix C.1). The baseline effective sample sizes for the DOC and DOC+RAM arms are 337 and 384 respectively. By keeping $f(t_m)$ and $\text{var}[t_m]$ constant while changing n'_{t_m} , we can explore the impact of the effective sample size on the posterior distribution.

Appendix D

Appendices for Chapter 5

D.1 Transition Parameters for an Illness-Death Model

The transition rates between states can be organized in a transition rate matrix $\mathbf{R}$ where rows correspond to the “from” state, and the columns to the “to” state. States in the illness-death model are healthy (h), ill (i), and dead (d). Disallowed transitions are given a value of 0, and all rows must sum to 0. We model the MESCC disease process using time-homogenous transition rates: that is transition rates do not change over time and thus $\mathbf{R}$ is constant over time. Disallowed transitions are i to h , and d to either h or i . The transition rate matrix is

$$\begin{bmatrix} -(\lambda_{hi} + \lambda_{hd}) & \lambda_{hi} & \lambda_{hd} \\ 0 & -\lambda_{id} & \lambda_{id} \\ 0 & 0 & 0 \end{bmatrix}. [40]$$

The transition rate matrix can be converted to a transition probability matrix $\mathbf{P}$. This matrix records the probability of making transitions in the time interval from $t = 0$ to the time t . Using the transition rate matrix $\mathbf{R}$, the transition probability matrix is computed as

$$\begin{aligned} \mathbf{P}(t) &= e^{\mathbf{R}t} \\ &= \begin{bmatrix} p_{hh}(t) & p_{hi}(t) & p_{hd}(t) \\ 0 & p_{id}(t) & p_{id}(t) \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{-(\lambda_{hi} + \lambda_{hd})t} & -\frac{\lambda_{hi}e^{-\lambda_{id}t}(-1 + e^{-(\lambda_{hi} + \lambda_{hd} - \lambda_{id})t})}{\lambda_{hi} + \lambda_{hd} - \lambda_{id}} & 1 - p_{hh}(t) - p_{hi}(t) \\ 0 & e^{-\lambda_{id}t} & 1 - p_{ii}(t) \\ 0 & 0 & 0 \end{bmatrix}. [40] \end{aligned}$$

The fraction of patients in each state at time t is recorded in the state membership vector $\mathbf{m}(t)$. The initial state membership fractions are $\mathbf{m}(0) = [\theta_h, \theta_i, \theta_d]$. $\mathbf{m}(t)$ is computed as the product of $\mathbf{m}(0)$ and the transition probability matrix $\mathbf{P}(t)$,

$$\mathbf{m}(t) = \mathbf{P}(0) \cdot \mathbf{P}(t).$$

The first element of $\mathbf{m}(t)$ is $H(t)$, the fraction of healthy patients

$$H(t) = \theta_h e^{-(\lambda_{hi} + \lambda_{hd})t}.$$

The second element of $\mathbf{m}(t)$ is $I(t)$, the fraction of ill patients

$$I(t) = \theta_i e^{-\lambda_{id}t} - \frac{\theta_h \lambda_{hi} e^{-\lambda_{id}t} (-1 + e^{-(\lambda_{hi} + \lambda_{hd} - \lambda_{id})t})}{\lambda_{hi} + \lambda_{hd} - \lambda_{id}}.$$

The third element of $\mathbf{m}(t)$ is $D(t)$, the fraction of dead patients

$$D(t) = 1 - \theta_h e^{-(\lambda_{hi} + \lambda_{hd})t} - \theta_i e^{-\lambda_{id}t} + \frac{\theta_h \lambda_{hi} e^{-\lambda_{id}t} (-1 + e^{-(\lambda_{hi} + \lambda_{hd} - \lambda_{id})t})}{\lambda_{hi} + \lambda_{hd} - \lambda_{id}}.$$

D.2 Multistate Estimation of Transition Rates

Methods

Likelihood

From Patchell *et al.*'s trial data, patients' health state trajectories can be categorized into seven different sequences (Table E.2). For some patients, the state was unknown at some time intervals. The time spent in the h , i and unknown state are denoted t_h , t_i , and t_u respectively. The transition rate and transition probability matrices are used to calculate the likelihood contribution for each trajectory.[122, 119, 66]

Table D.1: Health state trajectories and corresponding likelihood contributions

Sequence of states	Likelihood contribution
$h \rightarrow i \rightarrow d$	$\lambda_{hi}p_{hh}(t_h) \times \lambda_{id}p_{ii}(t_i)$
$h \rightarrow d$	$\lambda_{hd}p_{hh}(t_h)$
$i \rightarrow d$	$\lambda_{id}p_{ii}(t_i)$
$h \rightarrow \text{censored}$	$p_{hh}(t_h)$
$i \rightarrow \text{censored}$	$p_{ii}(t_i)$
$h \rightarrow \text{unknown} \rightarrow d$	$p_{hh}(t_h) \times (\lambda_{hd}p_{hh}(t_u) + \lambda_{id}p_{ii}(t_u) + \lambda_{id}p_{hi}(t_u))$
$\text{unknown} \rightarrow d$	$\lambda_{hd}p_{hh}(t_u) + \lambda_{id}p_{ii}(t_u) + \lambda_{id}p_{hi}(t_u)$

Prior Distributions

Non-informative prior distributions were used for multistate model estimation

$$\begin{aligned}\log \lambda &\sim \text{Normal}(\text{mean} = 0, \text{variance} = 10), \\ \beta &\sim \text{Normal}(\text{mean} = 0, \text{variance} = 10).[178]\end{aligned}$$

An upper truncation limit of 10 was applied to all prior distributions.[178]

Implementation

The model was estimated using the Bayesian modelling language Stan[212] run through the statistical programming language R.[183] Four Markov chains were implemented with different random initial values for each chain.

Initially all chains were run for 5000 iterations without thinning; 2500 burn-in iterations were discarded. Post-burn-in traceplots for each parameter were examined to ensure that

all four chains reached a similar mean and sampled similar regions of the distribution.[80] Iterations were increased until (1) the potential scale reduction factor, $\hat{R}$, was < 1.1 ; and (2) histograms of the posterior distribution were smooth and without gaps.[39, 26, 48] To mitigate autocorrelation, thinning was increased so that effective sample size, N_{eff} , was similar for all parameters, and the effective sample size as $\geq 20\%$ of the number of sampling iterations.[80] Post-burn in iterations were increased until Monte Carlo (MC) error for each parameter was $< 5\%$ of the sample standard deviation.[210] The final simulation conditions were termed the short conditions.

To ensure that global convergence was achieved (rather than local convergence), models was re-run with iterations doubled.[48] Simulation conditions were adjusted to ensure convergence and mitigate autocorrelation. These simulation conditions were termed the long conditions. Percent relative deviation for the estimates from the long conditions to the short conditions was computed. Global convergence was indicated if the percent relative deviation was less than $\pm 2\%$ for each parameter.[48]

STAN Code

```

functions{
  real MarkovKnownlpdf(real t, real [] lam, real censind, int state, int trans){
    //log-likelihood function when state known
    real lamstay [3]; real LL;
    lamstay[1]=-(lam[1]+lam[2]);
    lamstay[2]=-lam[3];
    LL = (lamstay[state])*t + log(lam[trans])*censind;
    return LL;
  }
  real MarkovUnknownlpdf(real t, real [] lam){
    //log-likelihood function when state unknown
    real LL; real LikeA; real LikeB; real LikeC;
    LikeA=exp((-lam[1] - lam[2])*t)*lam[2];
    LikeB=(lam[1]/(lam[1]+lam[2]-lam[3]))*(exp(-lam[3]*t)-exp((-lam[1] - lam[2])*t))*lam[3];
    LikeC=exp(-lam[3]*t)*lam[3];
    LL = log(LikeA + LikeB +LikeC);
    return LL;
  }
}
data {
  int <lower=0> Nknown; int <lower=0> Nunknown; vector [Nknown] tknown; vector [Nunknown] tunknown;
  int <lower=0> stateknown [Nknown]; int <lower=0> transknown [Nknown]; vector [Nknown] censindknown;
  vector [Nknown] grpknown; vector [Nunknown] grpunknown;
}
parameters {
  real llam [3]; real bet [3];
}
transformed parameters {
  real lam [3]; real slam [3];
  for (k in 1:3){
    //transition rates for RT-alone group
    lam[k]=exp(llam[k]);
    //transition rates for mS+RT group
    slam[k]=lam[k]*exp(bet[k]);
  }
}
model {
  real tempknown [3]; real tempunknown [3];
  //likelihood for transitions where from and to state known
  for (i in 1:Nknown){
    tempknown[1]=lam[1]*exp(bet[1]*grpknown[i]);
    tempknown[2]=lam[2]*exp(bet[2]*grpknown[i]);
    tempknown[3]=lam[3]*exp(bet[3]*grpknown[i]);
    target += MarkovKnownlpdf(tknown[i], tempknown, censindknown[i], stateknown[i], transknown[i]);
  }
  //likelihood for transitions where either from or to state unknown
  for (j in 1:Nunknown){
    tempunknown[1]=lam[1]*exp(bet[1]*grpunknown[j]);
    tempunknown[2]=lam[2]*exp(bet[2]*grpunknown[j]);
    tempunknown[3]=lam[3]*exp(bet[3]*grpunknown[j]);
    target += MarkovUnknownlpdf(tunknown[j], tempunknown);
  }
  //Prior Distributions
  llam ~ normal(0, 10)T[,10];
  bet ~ normal(0, 10)T[,10];
}

```

Results

To meet pre-specified benchmarks for autocorrelation, thinning was increased to 10 to ensure that effective sample size as $\geq 20\%$ of the number of sampling iterations (Table D.2).

Traceplots for λ , and β showed a stable mean and variance under both short (Figure D.1) and long (Figure D.2) conditions. $\hat{R}$ was < 1.1 for all parameters under all conditions; and percent relative deviation was $< \pm 2\%$ for all parameters (Table D.2). We conclude that global convergence was reached.

Histograms of the posterior distribution under both short (Figure D.3) and long (Figure D.4) simulation conditions show smooth histograms without empty bins. Under both short and long simulation conditions, Monte Carlo error was $< 5\%$ of the sample standard deviation for each parameter (Table D.2). We conclude that both the short and long simulation conditions were sufficient to characterize the shape of the posterior distribution.

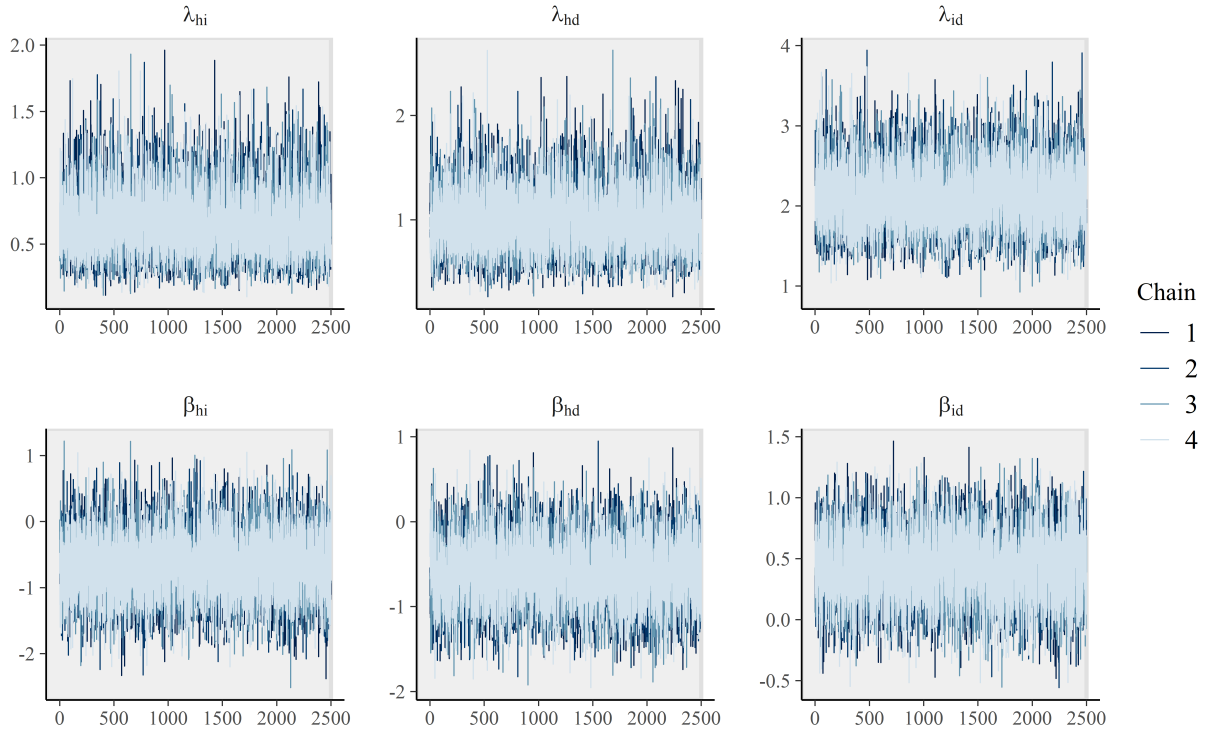


Figure D.1: Post-burn-in traceplots for transition rates and hazard ratios under short simulation conditions.

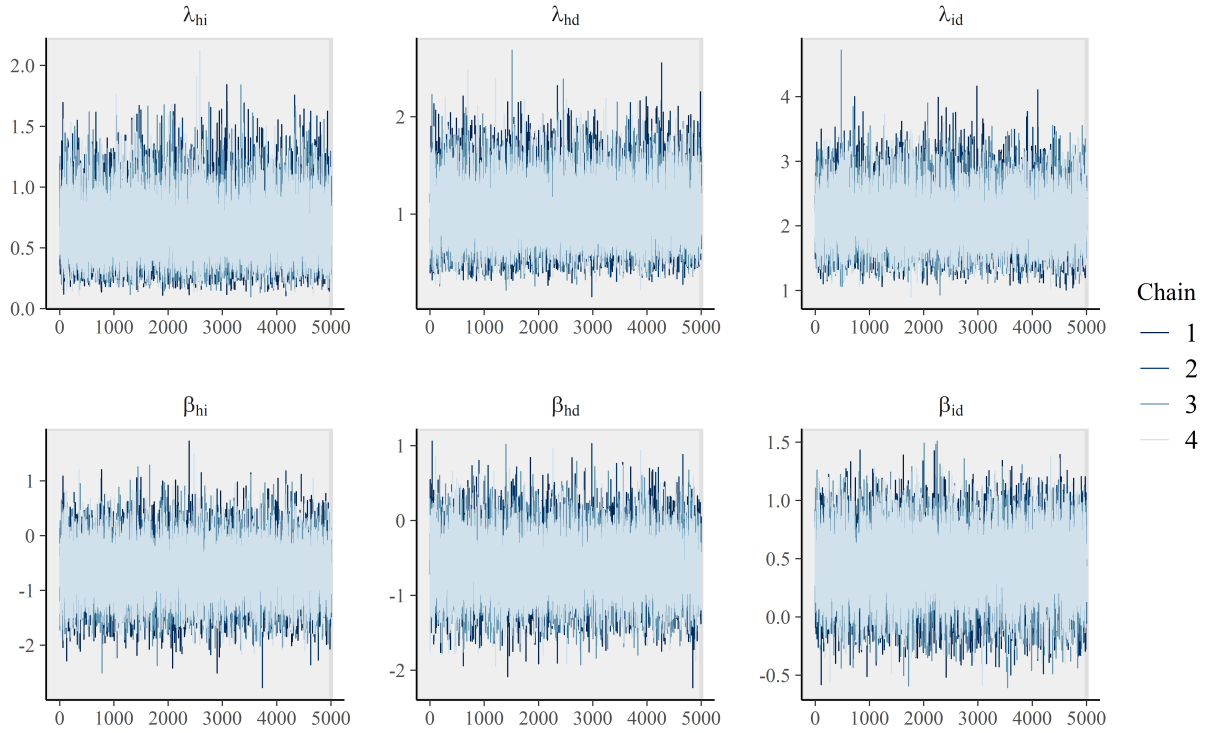


Figure D.2: Post-burn-in traceplots for transition rates and hazard ratios under long simulation conditions.

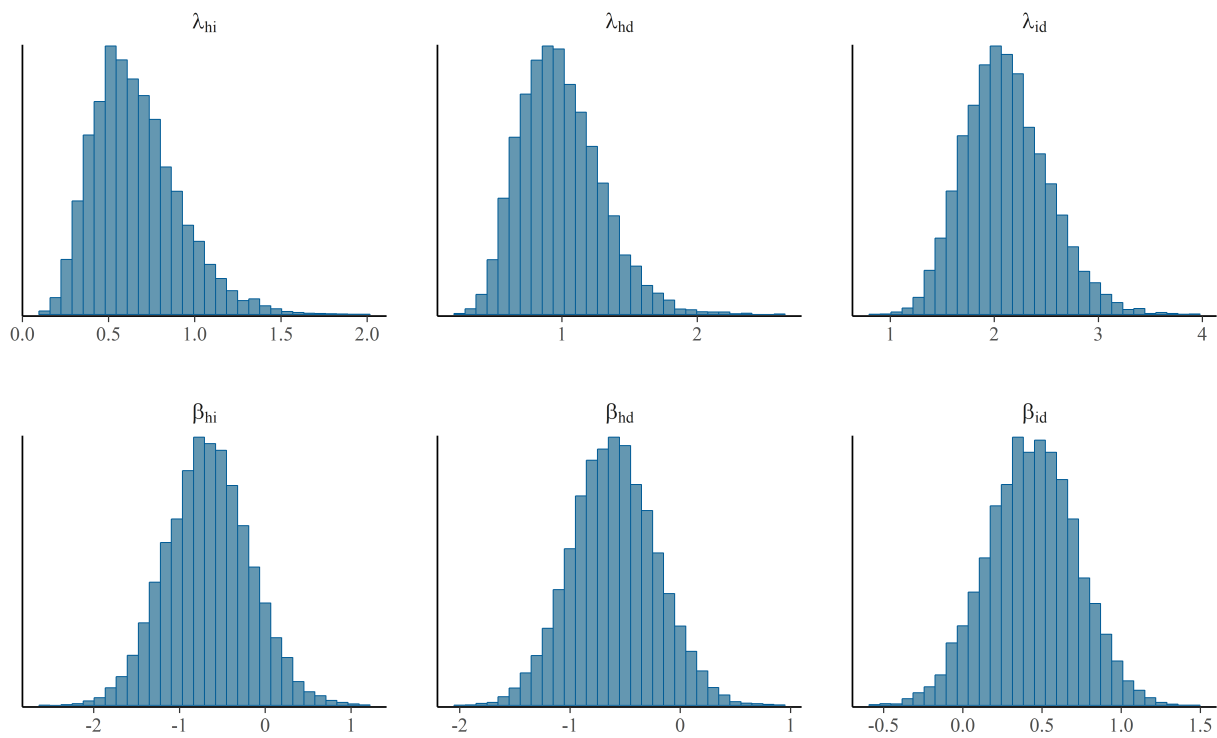


Figure D.3: Histograms of posterior distributions for transition rates and hazard ratios under short simulation conditions.

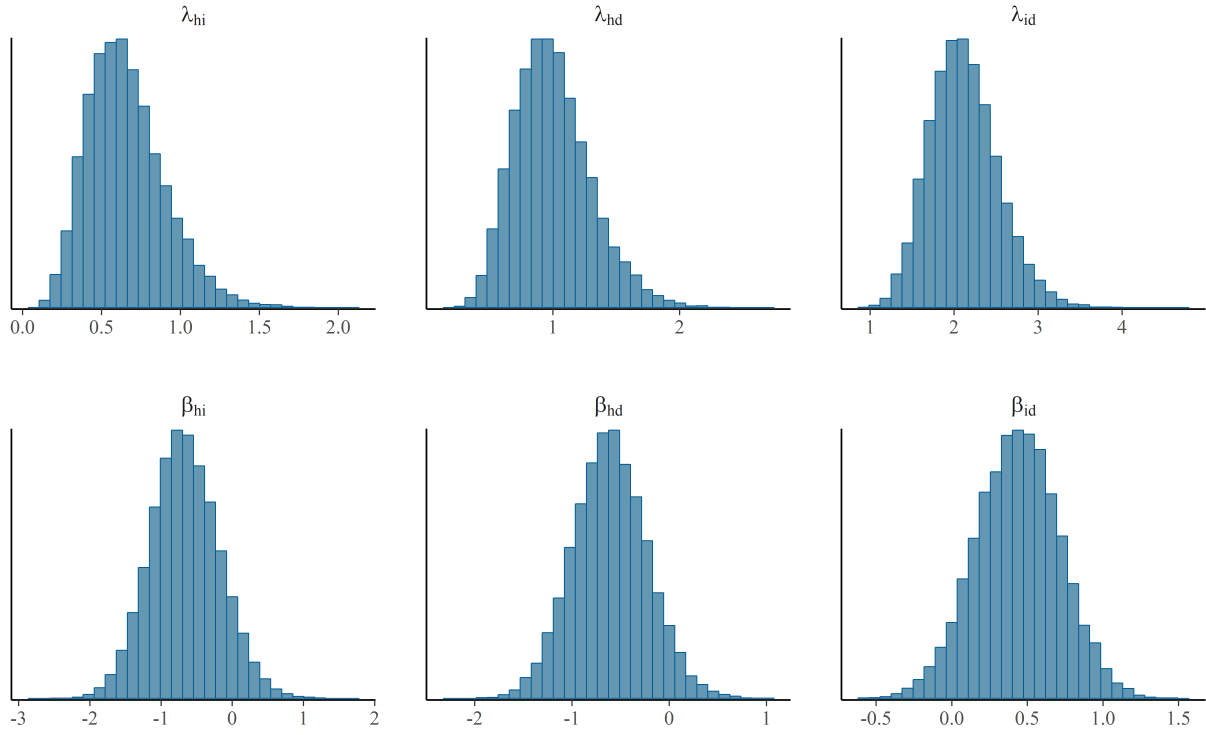


Figure D.4: Histograms of posterior distributions for transition rates and hazard ratios under short simulation conditions.

D.3 Worked Example

In this appendix we perform transition rate approximation for the mS+RT arm in the MESCC RCT using the formulas developed in Chapter 5.

Table 5.1 is used to populate Formula 5.1 with $AUC_{pfs} = \text{RMPFST}^{-\tau_{pfs}}$,

$$\begin{aligned}\text{RMPFST}^{-\tau_{pfs}} &= \frac{1 - e^{-(\lambda_{hi} + \lambda_{hd})\tau_{pfs}}}{\lambda_{hi} + \lambda_{hd}} \\ 1.159 &= \frac{1 - e^{-(\lambda_{hi} + \lambda_{hd})5.25}}{\lambda_{hi} + \lambda_{hd}}.\end{aligned}$$

This equation can be solved using `uniroot` function in the statistical programming language `R` with `lambdaS` as the variable for $(\lambda_{hi} + \lambda_{hd})$

```
lambdaS_solve<-function(lambdaS, tau_pfs, RMPFSTtau_pfs)(1-exp(-lambdaS*tau_pfs))/lambdaS-RMPFSTtau_pfs
uniroot(lambdaS_solve,c(0.00001,10000),5.25,1.16)
```

Using this code we obtain $\lambda_{hi} + \lambda_{hd} = 0.85$.

Table 5.1 is also used to populate Formula 5.2,

$$\begin{aligned}\rho &\approx \frac{N_{simul}}{N_{pfs}^e} \\ &\approx \frac{19}{31} \\ &= 0.61.\end{aligned}$$

We can now compute

$$\begin{aligned}\lambda_{hi} &= \rho (\lambda_{hi} + \lambda_{hd}) \\ &= 0.61 \times 0.85 \\ &\approx 0.52,\end{aligned}$$

and

$$\begin{aligned}\lambda_{hd} &= (\lambda_{hi} + \lambda_{hd}) - \lambda_{hi} \\ &\approx 0.85 - 0.52 \\ &\approx 0.33.\end{aligned}$$

We populate Formula 5.5 with values from Table 5.1 to yield

$$\begin{aligned}E_{os} &\approx \text{RMOST}^{-\tau_{os}} (N_{os}^e + N_{os}^c) + \sum_{i=1}^{N_{os}^e} o_i^e \left(1 - \frac{N_{os}^e + N_{os}^c}{N_{os}^e} \right) \\ &\approx 0.98 + 35.89 \left(1 - \frac{45 + 3}{45} \right) \\ &\approx 45.27.\end{aligned}$$

Table D.2: Comparison of Transition Rates and Hazard Ratios Estimated with Multistate Techniques under Short and Long Simulation Conditions

	Short Conditions					Long Conditions					% Rel	
	Mean	SDev	MCSE	$\hat{R}$	N_{eff}	Mean	SDev	MCSE	$\hat{R}$	N_{eff}	Dev	
λ_{hi}	0.66	0.25	0.00	10094	1.00	0.66	0.24	0.00	19751	1.00	0.80	
λ_{hd}	0.99	0.30	0.00	9956	1.00	1.00	0.30	0.00	19084	1.00	-0.50	
λ_{id}	2.12	0.40	0.00	10039	1.00	2.12	0.40	0.00	19627	1.00	0.00	
β_{hi}	-0.66	0.49	0.00	10303	1.00	-0.66	0.48	0.00	19906	1.00	0.30	
β_{hd}	-0.60	0.39	0.00	9665	1.00	-0.61	0.38	0.00	19387	1.00	-1.00	
β_{id}	0.44	0.28	0.00	9805	1.00	0.44	0.28	0.00	19750	1.00	0.00	

SDev, standard deviation. MCSE, Monte Carlo standard error. % Rel. Dev., percent relative deviation

We adapt Formula 5.5 for the PFS analysis and substitute relevant values from Table 5.1 to yield

$$\begin{aligned}
E_{pfs} &\approx \text{RMPFT}^{-\tau_{pfs}} (N_{pfs}^e + N_{pfs}^c) + \sum_{i=1}^{N_{pfs}^e} p_i^e \left(1 - \frac{N_{pfs}^e + N_{pfs}^c}{N_{pfs}^e} \right) \\
&\approx 1.16 + 27.22 \left(1 - \frac{31 + 6}{31} \right) \\
&\approx 37.65.
\end{aligned}$$

We can then use E_{os} and E_{pfs} in Formula 5.6

$$\begin{aligned}
E_{ill} &\approx E_{os} - E_{pfs} \\
&\approx 45.27 - 35.94 \\
&\approx 7.00.
\end{aligned}$$

We then substitute values from Table 5.1 in Formula 5.7

$$\begin{aligned}
N_{id} &\approx N_{os}^e - N_{simul} \\
&\approx 45 - 19 \\
&\approx 26
\end{aligned}$$

N_{id} and E_{ill} are used in Formula 5.8 to obtain

$$\begin{aligned}
\lambda_{id} &\approx \frac{N_{id}}{E_{ill}} \\
&\approx \frac{26}{7.00} \\
&\approx 3.71.
\end{aligned}$$

Appendix E

Appendices for Chapter 6

E.1 Bayesian Estimation

This section describes the methods for and results of Bayesian MCMC estimation for

- treatment-phase event probabilities
- a mixed effects beta regression model for utilities

Details on the likelihood function and prior distribution for each application of Bayesian MCMC estimation are provided in the following subsections. However, a common analytic approach was used for each application.

Each model was estimated using the Bayesian modelling language Stan[212] run through the statistical programming language R.[183] Four Markov chains were implemented with different random initial values for each chain.

Initially all chains were run for 10 000 iterations without thinning; 5000 burn-in iterations were discarded. Post-burn-in traceplots for each parameter were examined to ensure that all four chains reached a similar mean and sampled similar regions of the distribution.[80] Iterations were increased until (1) the potential scale reduction factor, $\hat{R}$, was < 1.1 ; and (2) histograms of the posterior distribution were smooth and without gaps.[39, 26, 48] To mitigate autocorrelation, thinning was increased so that effective sample size, N_{eff} , was similar for all parameters, and the effective sample size as $\geq 0.1\%$ of the number of sampling iterations.[80] The final simulation conditions were termed the short conditions.

To ensure that global convergence was achieved (rather than local convergence), models was re-run with iterations doubled.[48] Simulation conditions were adjusted to ensure convergence and mitigate autocorrelation. These simulation conditions were termed the long conditions. Percent relative deviation for the estimates from the long conditions to the short conditions was computed. Global convergence was indicated if the percent relative deviation was less than $\pm 1\%$ for each parameter.[48] Results for long conditions were used to run the decision model (Section 6.3).

E.1.1 Bayesian MCMC Estimation of Event Probabilities for the Treatment-Phase

Likelihood

In Patchell *et al.*'s RCT, patients were evaluated at $t = 0$ and immediately after treatment. At $t = 0$ patients could be ambulatory (a) or non-ambulatory (n). After treatment patients could be a , n or dead (d). If we stratify by treatment group and health state at $t = 0$, the health state immediately after treatment can be modeled using a multinomial distribution.[81]

The multinomial distribution is a generalization of the binomial distribution. The multinomial distribution models the probability of observing a vector of counts after repeated trials. This distribution is parametrized by $\theta_1 \dots \theta_k$ event probabilities with $\sum_{i=1}^k \theta_i = 1$. [170] Concretely for the Patchell *et al.*'s RCT, the vector of counts is the number of ambulatory, non-ambulatory and dead patients, $[N_a, N_n, N_d]$, for a given treatment arm stratified by baseline ambulatory status at is distributed as:

$$[N_a, N_n, N_d] \sim \text{Multinomial}(\theta_a, \theta_n, \theta_d)$$

Prior Distributions

We utilized non-informative prior distributions to form a hierarchical multinomial-Dirichlet model.[8] Event probabilities were organized in the vector $\Theta = [\theta_a, \theta_n, \theta_d]$. θ was stratified for treatment arm and baseline ambulatory status. Stratum was listed in the subscript on $\theta_{x,y}$ with x denoting the treatment arm, and y denoting baseline ambulatory status. Consequently four prior distributions were specified:

$$\begin{aligned}\Theta_{rt,a} &\sim \text{Dirichlet}(\tau_a A_{rt}) \\ \Theta_{rt,n} &\sim \text{Dirichlet}(\tau_n A_{rt}) \\ \Theta_{s+rt,a} &\sim \text{Dirichlet}(\tau_a A_{s+rt}) \\ \Theta_{s+rt,n} &\sim \text{Dirichlet}(\tau_n A_{s+rt}).\end{aligned}$$

Note that treatment arms share a common concentration parameter vector A (subscripted for treatment arm). In this way, Θ vectors for the same arm are dependent. A non-informative hyper-prior distribution was placed on A

$$A \sim \text{Dirichlet}(1, 1, 1).$$

The informativeness of each prior distribution was controlled by an equivalent sample size parameter τ . To make Θ vectors for baseline ambulatory status strata dependent, τ was common for baseline ambulatory status. A non-informative hyper-prior distribution was placed on τ

$$\tau \sim \text{Uniform}(0, 100).$$

Results

To meet pre-specified benchmarks for autocorrelation, thinning was increased to 10 to ensure that effective sample size as $\geq 20\%$ of the number of sampling iterations (Table E.1).

Traceplots for $\Theta_{rt,a}$, $\Theta_{rt,n}$, $\Theta_{s+rt,a}$, and $\Theta_{s+rt,n}$ showed a stable mean and variance under both short (Figure E.1) and long (Figure E.2) conditions. $\hat{R}$ was < 1.1 for all parameters under all conditions; and percent relative deviation was $< \pm 2\%$ for all parameters (Table E.1). We conclude that global convergence was reached.

Histograms of the posterior distribution under both short (Figure E.3) and long (Figure E.4) simulation conditions show smooth histograms without empty bins. Under both short and long simulation conditions, Monte Carlo error was $< 5\%$ of the sample standard deviation for each parameter (Table E.1). We conclude that both the short and long simulation conditions were sufficient to characterize the shape of the posterior distribution.

Table E.1: Comparison of results from short and long simulation conditions for multistate estimated transition rates and hazard ratios

	Short Conditions					Long Conditions					% Rel	
	Mean	SDev	MCSE	$\hat{R}$	N_{eff}	Mean	SDev	MCSE	$\hat{R}$	N_{eff}		Dev
$\theta_{a,rt,a}$	0.54	0.08	0.00	9703	1.00	0.54	0.07	0.00	20000	1.00		-0.40
$\theta_{n,rt,a}$	0.32	0.07	0.00	9695	1.00	0.32	0.07	0.00	19790	1.00		0.00
$\theta_{d,rt,a}$	0.14	0.05	0.00	10000	1.00	0.14	0.05	0.00	20000	1.00		0.00
$\theta_{a,rt,n}$	0.41	0.10	0.00	9723	1.00	0.42	0.10	0.00	20000	1.00		-0.50
$\theta_{n,rt,n}$	0.41	0.09	0.00	9961	1.00	0.41	0.09	0.00	20000	1.00		0.20
$\theta_{d,rt,n}$	0.17	0.07	0.00	9964	1.00	0.17	0.07	0.00	20000	1.00		0.60
$\theta_{a,s+rt,a}$	0.77	0.07	0.00	9703	1.00	0.77	0.07	0.00	20000	1.00		0.00
$\theta_{n,s+rt,a}$	0.18	0.06	0.00	9695	1.00	0.18	0.06	0.00	20000	1.00		-0.60
$\theta_{d,s+rt,a}$	0.05	0.03	0.00	10000	1.00	0.05	0.03	0.00	19538	1.00		0.00
$\theta_{a,s+rt,n}$	0.63	0.09	0.00	9723	1.00	0.63	0.09	0.00	19823	1.00		0.20
$\theta_{n,s+rt,n}$	0.30	0.08	0.00	9961	1.00	0.30	0.09	0.00	20000	1.00		-0.70
$\theta_{d,s+rt,n}$	0.07	0.05	0.00	9964	1.00	0.07	0.05	0.00	18866	1.00		0.00

SDev, standard deviation. MCSE, Monte Carlo standard error. % Rel. Dev., percent relative deviation

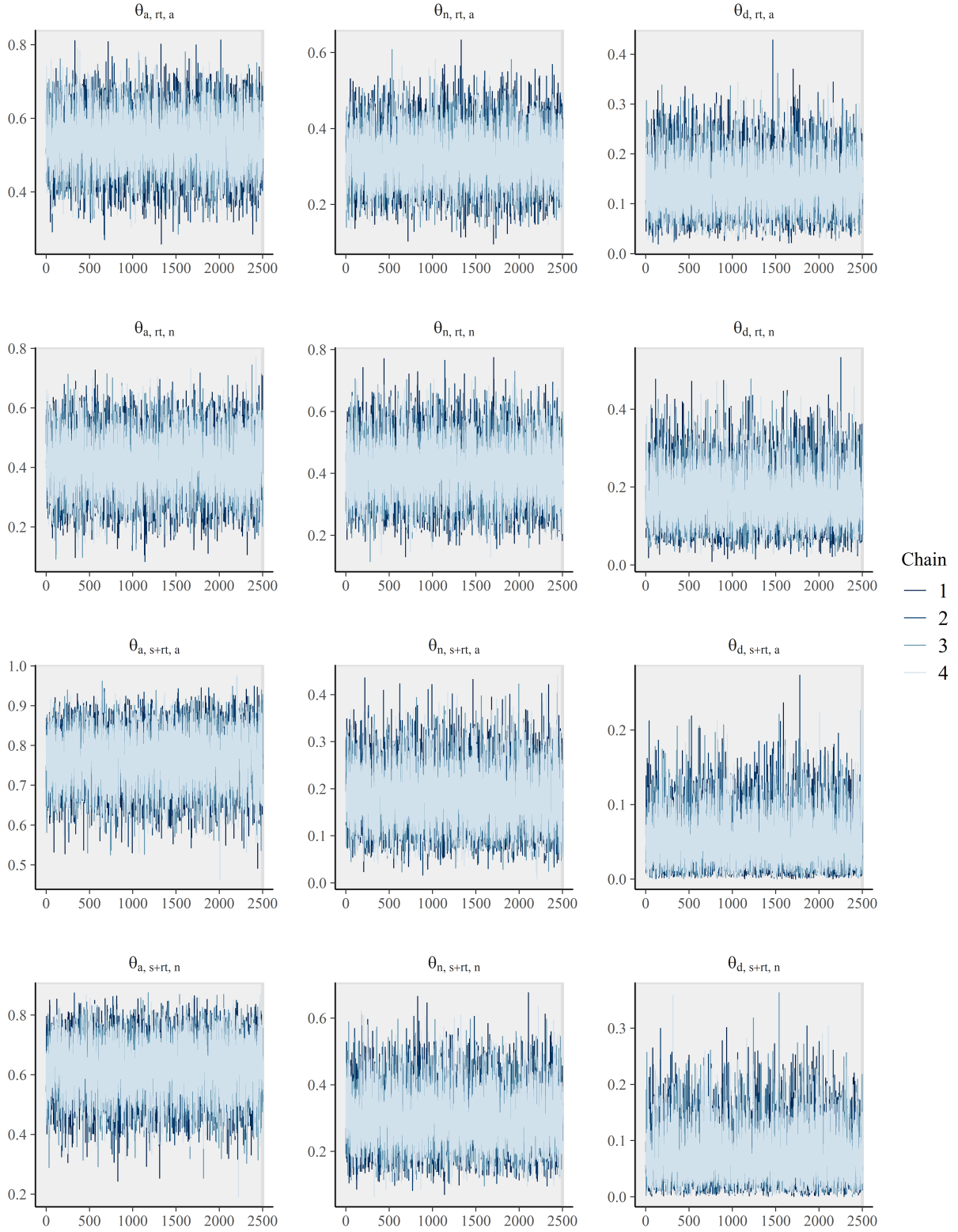


Figure E.1: Post-burn-in traceplots for event probabilities under short simulation conditions.

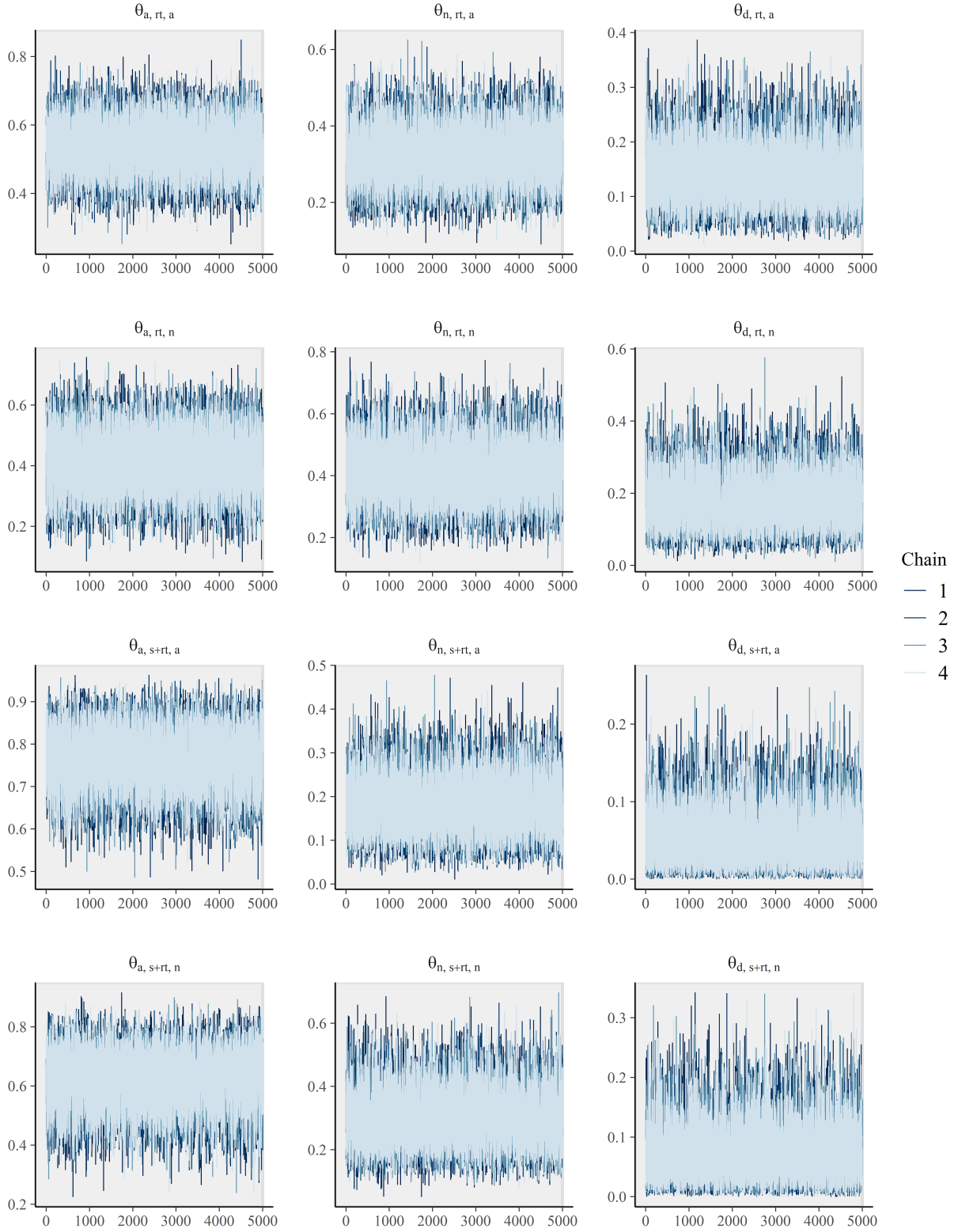


Figure E.2: Post-burn-in traceplots for event probabilities under long simulation conditions.

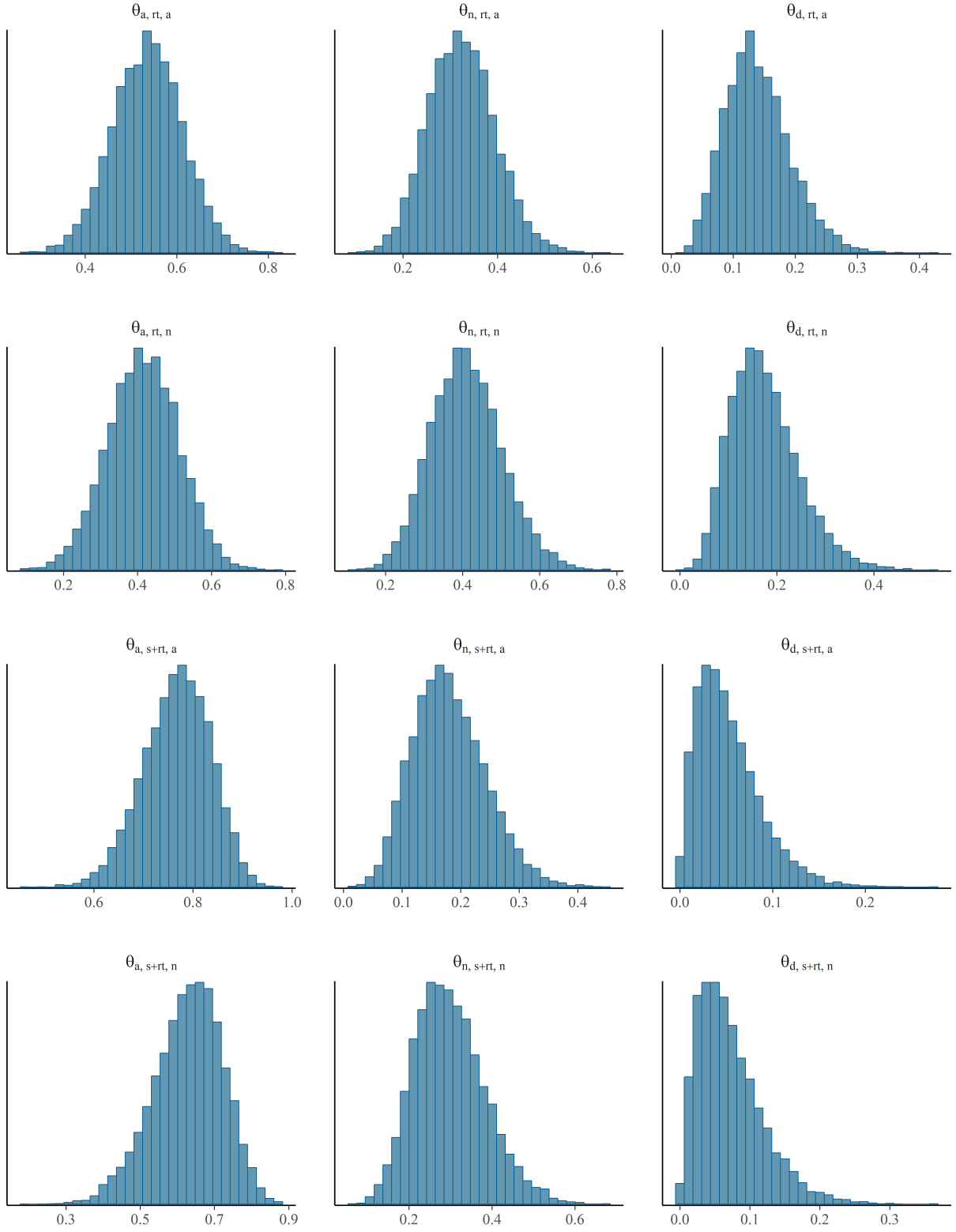


Figure E.3: Histograms of posterior distributions for event probabilities under short simulation conditions.

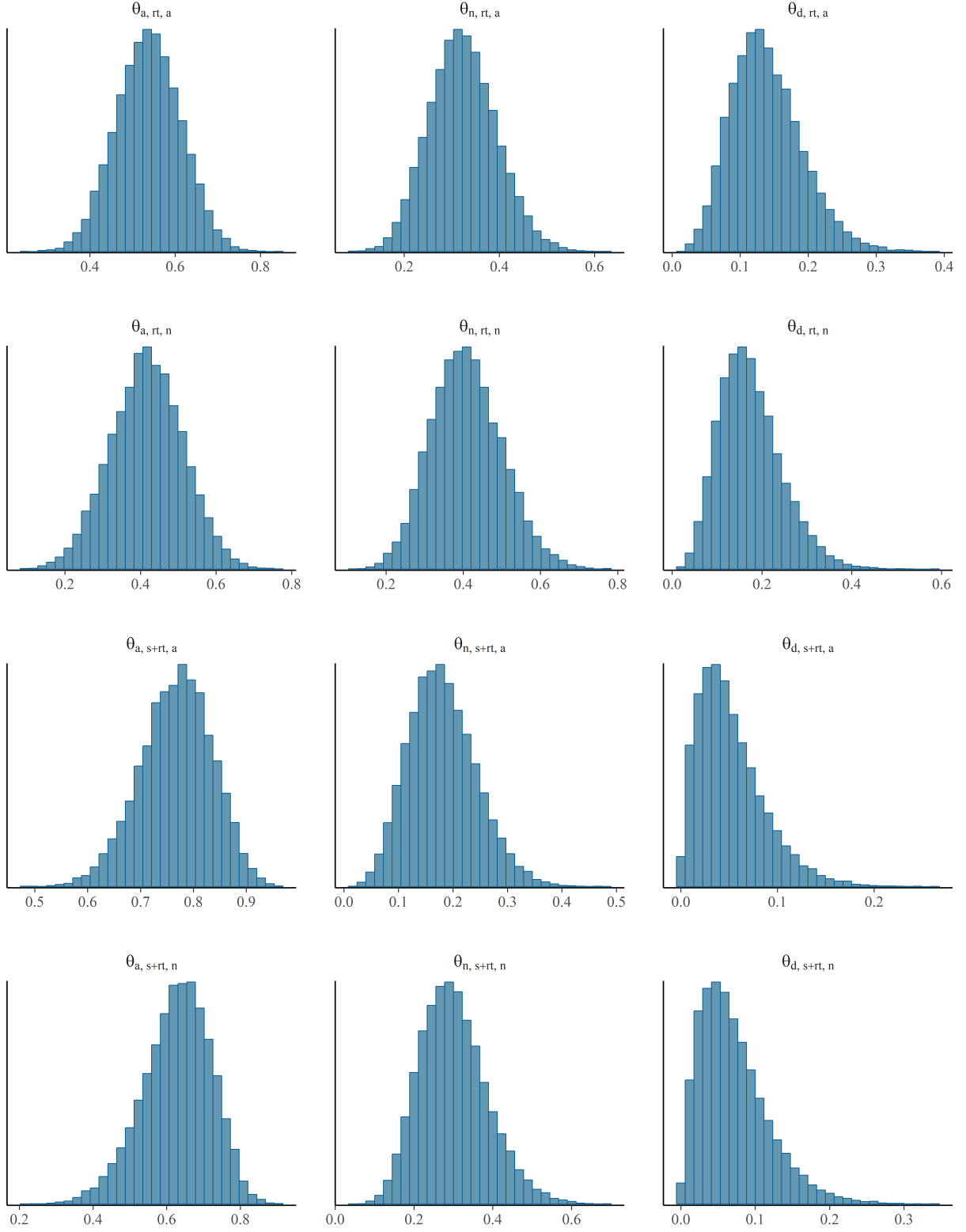


Figure E.4: Histograms of posterior distributions for event probabilities under long simulation conditions.

E.1.2 Bayesian MCMC Estimation of Transition Rates

Likelihood

From Patchell *et al.*'s trial data, patients' health state trajectories can be categorized into seven different sequences (Table E.2). For some patients, the state was unknown at some time intervals. The time spent in the a , n and unknown state are denoted t_a , t_n , and t_u respectively. The transition rate and transition probability matrices are used to calculate the likelihood contribution for each trajectory.[122, 119, 66]

Table E.2: Health state trajectories and corresponding likelihood contributions

Sequence of states	Likelihood contribution
$a \rightarrow n \rightarrow d$	$\lambda_{an}p_{aa}(t_a) \times \lambda_{nd}p_{nn}(t_n)$
$a \rightarrow d$	$\lambda_{ad}p_{aa}(t_a)$
$n \rightarrow d$	$\lambda_{nd}p_{nn}(t_n)$
$a \rightarrow \text{censored}$	$p_{aa}(t_a)$
$n \rightarrow \text{censored}$	$p_{nn}(t_n)$
$a \rightarrow \text{unknown} \rightarrow d$	$p_{aa}(t_a) \times (\lambda_{ad}p_{aa}(t_u) + \lambda_{nd}p_{nn}(t_u) + \lambda_{nd}p_{an}(t_u))$
$\text{unknown} \rightarrow d$	$\lambda_{ad}p_{aa}(t_u) + \lambda_{nd}p_{nn}(t_u) + \lambda_{nd}p_{an}(t_u)$

Prior Distributions

Non-informative prior distributions were used for multistate model estimation

$$\begin{aligned}\log \lambda &\sim \text{Normal}(\text{mean} = 0, \text{variance} = 10), \\ \beta &\sim \text{Normal}(\text{mean} = 0, \text{variance} = 10).[178]\end{aligned}$$

An upper truncation limit of 10 was applied to all prior distributions.[178]

Implementation

The model was estimated using the Bayesian modelling language Stan[212] run through the statistical programming language R.[183] Four Markov chains were implemented with different random initial values for each chain.

Initially all chains were run for 5000 iterations without thinning; 2500 burn-in iterations were discarded. Post-burn-in traceplots for each parameter were examined to ensure that all four chains reached a similar mean and sampled similar regions of the distribution.[80] Iterations were increased until (1) the potential scale reduction factor, $\hat{R}$, was < 1.1 ; and (2) histograms of the posterior distribution were smooth and without gaps.[39, 26, 48]

To mitigate autocorrelation, thinning was increased so that effective sample size, N_{eff} , was similar for all parameters, and the effective sample size as $\geq 20\%$ of the number of sampling iterations.[80] Post-burn in iterations were increased until Monte Carlo (MC) error for each parameter was $< 5\%$ of the sample standard deviation.[210] The final simulation conditions were termed the short conditions.

To ensure that global convergence was achieved (rather than local convergence), models was re-run with iterations doubled.[48] Simulation conditions were adjusted to ensure convergence and mitigate autocorrelation. These simulation conditions were termed the long conditions. Percent relative deviation for the estimates from the long conditions to the short conditions was computed. Global convergence was indicated if the percent relative deviation was less than $\pm 2\%$ for each parameter.[48]

STAN Code

```

functions{
  real MarkovKnownlpdf(real t, real [] lam, real censind, int state, int trans){
    //log-likelihood function when state known
    real lamstay [3]; real LL;
    lamstay[1]=-(lam[1]+lam[2]);
    lamstay[2]=-lam[3];
    LL = (lamstay[state])*t + log(lam[trans])*censind;
    return LL;
  }
  real MarkovUnknownlpdf(real t, real [] lam){
    //log-likelihood function when state unknown
    real LL; real LikeA; real LikeB; real LikeC;
    LikeA=exp((-lam[1] - lam[2])*t)*lam[2];
    LikeB=(lam[1]/(lam[1]+lam[2]-lam[3]))*(exp(-lam[3]*t)-exp((-lam[1] - lam[2])*t))*lam[3];
    LikeC=exp(-lam[3]*t)*lam[3];
    LL = log(LikeA + LikeB +LikeC);
    return LL;
  }
}
data {
  int <lower=0> Nknown; int <lower=0> Nunknown; vector [Nknown] tknown; vector [Nunknown] tunknown;
  int <lower=0> stateknown [Nknown]; int <lower=0> transknown [Nknown]; vector [Nknown] censindknown;
  vector [Nknown] grpknown; vector [Nunknown] grpunknown;
}
parameters {
  real llam [3]; real bet [3];
}
transformed parameters {
  real lam [3]; real slam [3];
  for (k in 1:3){
    //transition rates for RT-alone group
    lam[k]=exp(llam[k]);
    //transition rates for mS+RT group
    slam[k]=lam[k]*exp(bet[k]);
  }
}
model {
  real tempknown [3]; real tempunknown [3];
  //likelihood for transitions where from and to state known
  for (i in 1:Nknown){
    tempknown[1]=lam[1]*exp(bet[1]*grpknown[i]);
    tempknown[2]=lam[2]*exp(bet[2]*grpknown[i]);
    tempknown[3]=lam[3]*exp(bet[3]*grpknown[i]);
    target += MarkovKnownlpdf(tknown[i], tempknown, censindknown[i], stateknown[i], transknown[i]);
  }
  //likelihood for transitions where either from or to state unknown
  for (j in 1:Nunknown){
    tempunknown[1]=lam[1]*exp(bet[1]*grpunknown[j]);
    tempunknown[2]=lam[2]*exp(bet[2]*grpunknown[j]);
    tempunknown[3]=lam[3]*exp(bet[3]*grpunknown[j]);
    target += MarkovUnknownlpdf(tunknown[j], tempunknown);
  }
  //Prior Distributions
  llam ~ normal(0, 10)T[,10];
  bet ~ normal(0, 10)T[,10];
}

```

Results

To meet pre-specified benchmarks for autocorrelation, thinning was increased to 10 to ensure that effective sample size as $\geq 20\%$ of the number of sampling iterations (Table E.3).

Traceplots for λ , and β showed a stable mean and variance under both short (Figure E.5) and long (Figure E.6) conditions. $\hat{R}$ was < 1.1 for all parameters under all conditions; and percent relative deviation was $< \pm 2\%$ for all parameters (Table E.3). We conclude that global convergence was reached.

Histograms of the posterior distribution under both short (Figure E.7) and long (Figure E.8) simulation conditions show smooth histograms without empty bins. Under both short and long simulation conditions, Monte Carlo error was $< 5\%$ of the sample standard deviation for each parameter (Table E.3). We conclude that both the short and long simulation conditions were sufficient to characterize the shape of the posterior distribution.

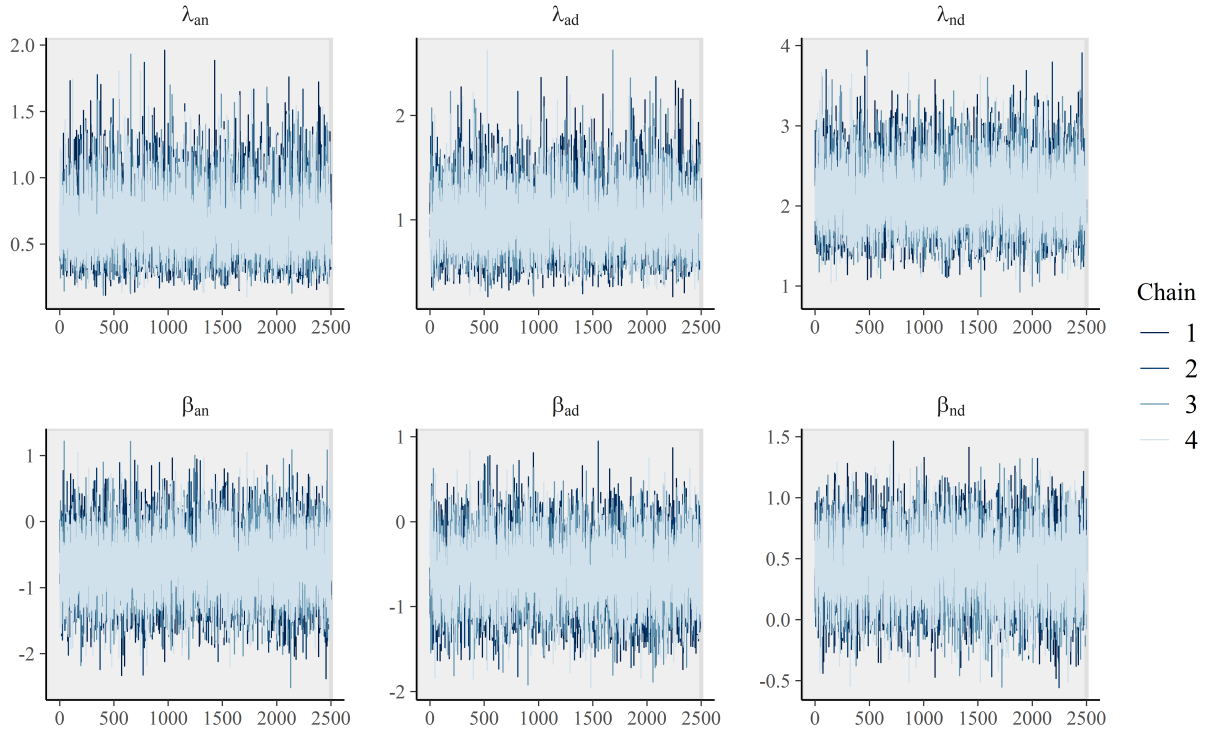


Figure E.5: Post-burn-in traceplots for transition rates and hazard ratios under short simulation conditions.

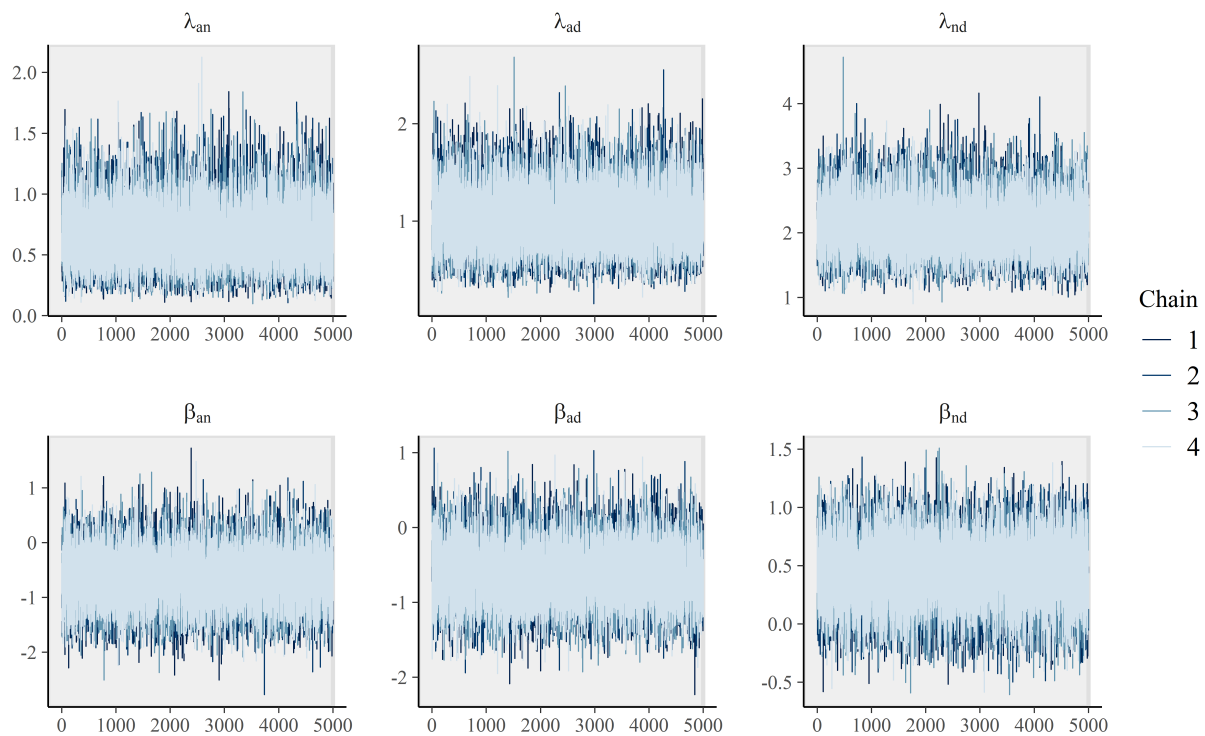


Figure E.6: Post-burn-in traceplots for transition rates and hazard ratios under long simulation conditions.

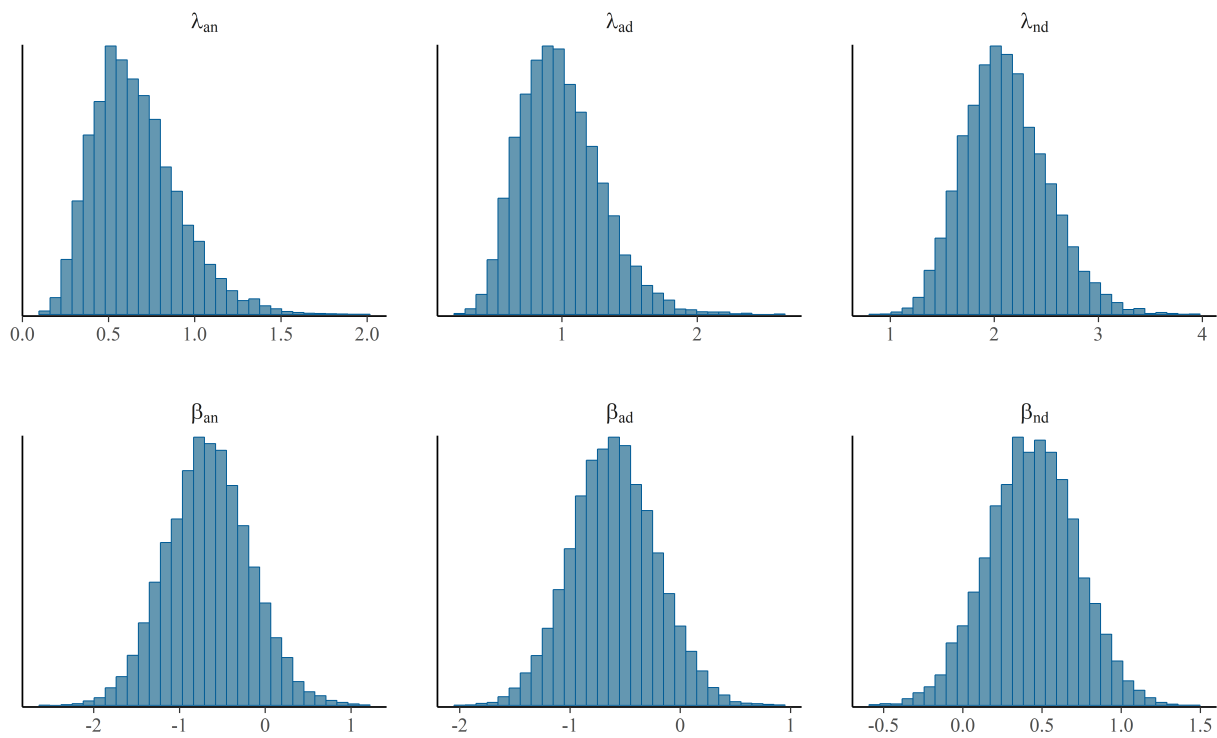


Figure E.7: Histograms of posterior distributions for transition rates and hazard ratios under short simulation conditions.

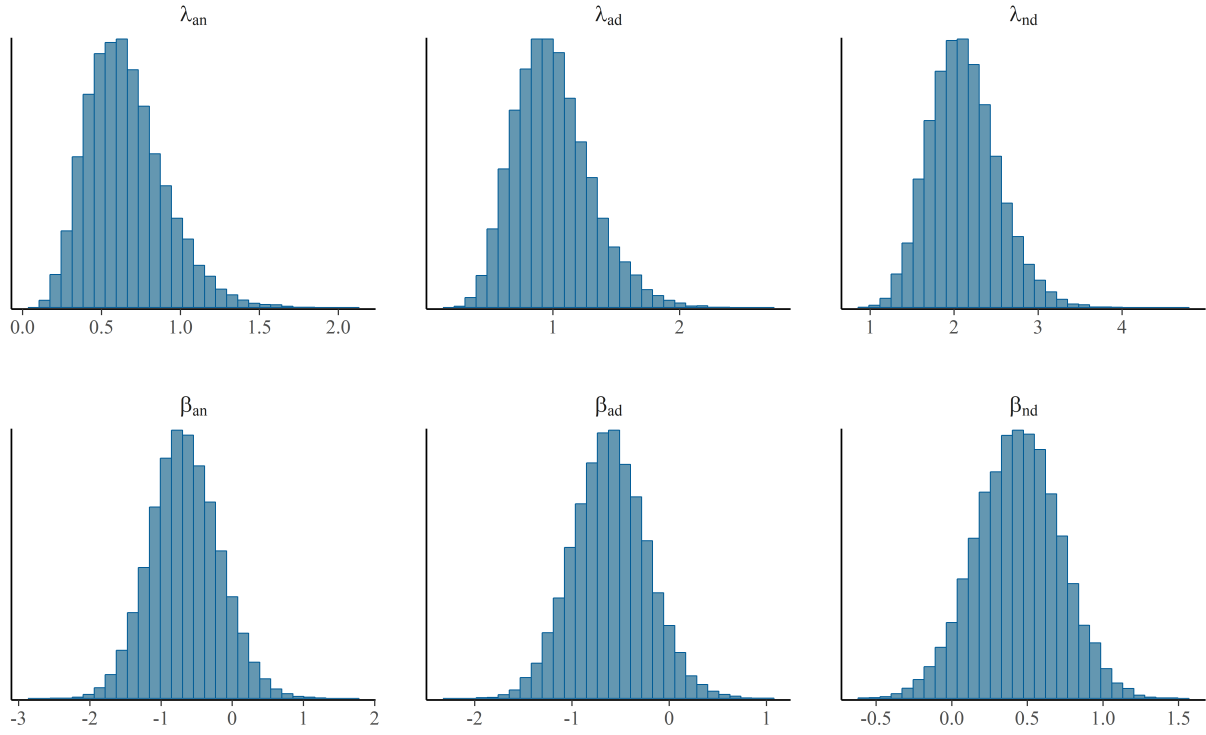


Figure E.8: Histograms of posterior distributions for transition rates and hazard ratios under short simulation conditions.

E.1.3 Bayesian MCMC Estimation of a Mixed Effects Beta Regression Model for Utilities

Model

In previous work, a beta regression model with the square root of the number of dysfunctional attributes in a health state was found to be well calibrated for estimating mean utility.[168] For the purposes of decision analysis, we re-fit this model using Bayesian methods.

Beta regression was used because utilities are restricted to the interval $[0, 1]$. The beta distribution is typically parametrized using two shape parameters: $p \in \mathbb{R}^+$ and $q \in \mathbb{R}^+$. [115] For regression estimation, it is more useful to parametrize with the mean (μ) and precision (ϕ). [67] This alternative parametrization of the beta distribution is related to the typical parametrization by $p = \mu\phi$ and $q = (1 - \mu)\phi$. [67]

Mixed beta regression model for the mean, μ , was implemented as described by Figueroa-Zúñiga and colleagues.[69] It was assumed that participant i 's j -th utility valuation, Y_{ij} , is drawn from a beta distribution

$$Y_{ij} \sim \text{Beta}(\mu_{ij}\phi, (1 - \mu_{ij})\phi).$$

The inverse logit function was used to map linear predictors to the $[0, 1]$ interval.[149] Based on rigorous model selection, the covariate for participant i 's j -th utility valuation was the square root of the number of dysfunctions. β_0 and β_1 are fixed effect regression coefficients. To account for six responses per participant a random intercept term ϵ_i was added for each participant i . [78] The precision parameter was common to all utility valuations.[67]

The expectation of participant i 's j -th utility valuation is the mean of then beta distribution (μ_{ij}) served as the dependent variable of the regression model:

$$\text{E}[Y_{ij}] = \mu_{ij} = \text{inverse logit}(\beta_0 + x_{ij}\beta_1 + \epsilon_i).$$

Prior Distributions

Non-informative prior distributions were used

$$\begin{aligned}\beta_0 &\sim \text{Normal}(\text{mean} = 0, \text{standard deviation} = 10), \\ \beta_1 &\sim \text{Normal}(\text{mean} = 0, \text{standard deviation} = 10), \\ \epsilon_i &\sim \text{Normal}(\text{mean} = 0, \text{standard deviation} = 1), \\ \phi &\sim \text{Gamma}(0.1, 0.01). [127, 82, 77]\end{aligned}$$

Results

To meet pre-specified benchmarks for autocorrelation, thinning was increased to 3 to ensure that effective sample size as $\geq 20\%$ of the number of sampling iterations (Table E.4).

Table E.3: Comparison of Transition Rates and Hazard Ratios Estimated with Multistate Techniques under Short and Long Simulation Conditions

	Short Conditions					Long Conditions					% Rel	
	Mean	SDev	MCSE	$\hat{R}$	N_{eff}	Mean	SDev	MCSE	$\hat{R}$	N_{eff}	Dev	Dev
λ_{an}	0.66	0.25	0.00	10094	1.00	0.66	0.24	0.00	19751	1.00	0.80	0.80
λ_{ad}	0.99	0.30	0.00	9956	1.00	1.00	0.30	0.00	19084	1.00	-0.50	-0.50
λ_{nd}	2.12	0.40	0.00	10039	1.00	2.12	0.40	0.00	19627	1.00	0.00	0.00
β_{an}	-0.66	0.49	0.00	10303	1.00	-0.66	0.48	0.00	19906	1.00	0.30	0.30
β_{ad}	-0.60	0.39	0.00	9665	1.00	-0.61	0.38	0.00	19387	1.00	-1.00	-1.00
β_{nd}	0.44	0.28	0.00	9805	1.00	0.44	0.28	0.00	19750	1.00	0.00	0.00

SDev, standard deviation. MCSE, Monte Carlo standard error. % Rel. Dev., percent relative deviation

Traceplots for β_0 , β_1 and ϕ showed a stable mean and variance under both short (Figure E.10) and long (Figure E.10) conditions. $\hat{R}$ was < 1.1 for all parameters under all conditions; and percent relative deviation was $< \pm 2\%$ for all parameters (Table E.4). We conclude that global convergence was reached.

Histograms of the posterior distribution under both short (Figure E.11) and long (Figure E.12) simulation conditions show smooth histograms without empty bins. Under both short and long simulation conditions, Monte Carlo error was $< 5\%$ of the sample standard deviation for each parameter (Table E.4). We conclude that both the short and long simulation conditions were sufficient to characterize the shape of the posterior distribution.

Table E.4: Comparison of results from short and long simulation conditions for beta regression coefficients for utilities

	Short Conditions					Long Conditions					% Rel	
	Mean	SDev	MCSE	$\hat{R}$	N_{eff}	Mean	SDev	MCSE	$\hat{R}$	N_{eff}	Dev	
β_0	1.73	0.10	0.00	9655	1.00	1.72	0.10	0.00	19892	1.00	0.10	
β_1	-1.15	0.06	0.00	10378	1.00	-1.15	0.05	0.00	20695	1.00	0.00	
ϕ	0.96	0.02	0.00	10081	1.00	0.96	0.02	0.00	21283	1.00	0.10	
SDev, standard deviation. MCSE, Monte Carlo standard error. % Rel. Dev., percent relative deviation												

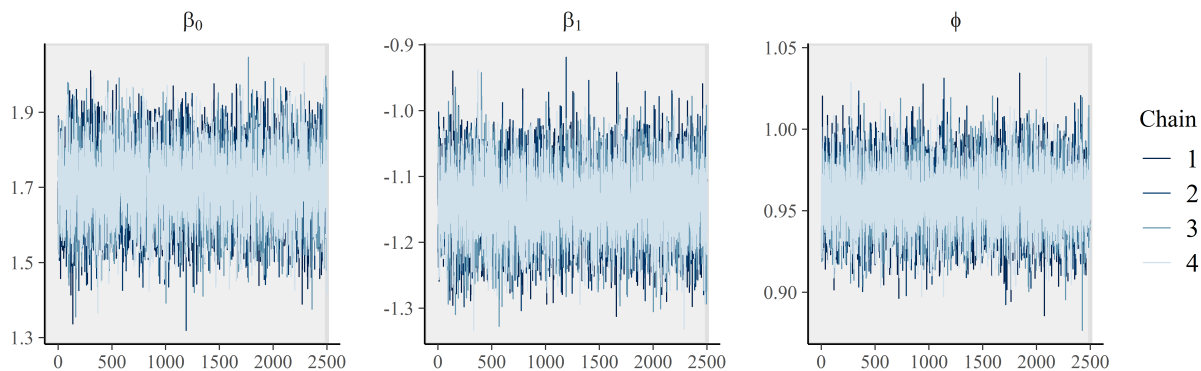


Figure E.9: Post-burn-in traceplots for regression coefficients under short simulation conditions.

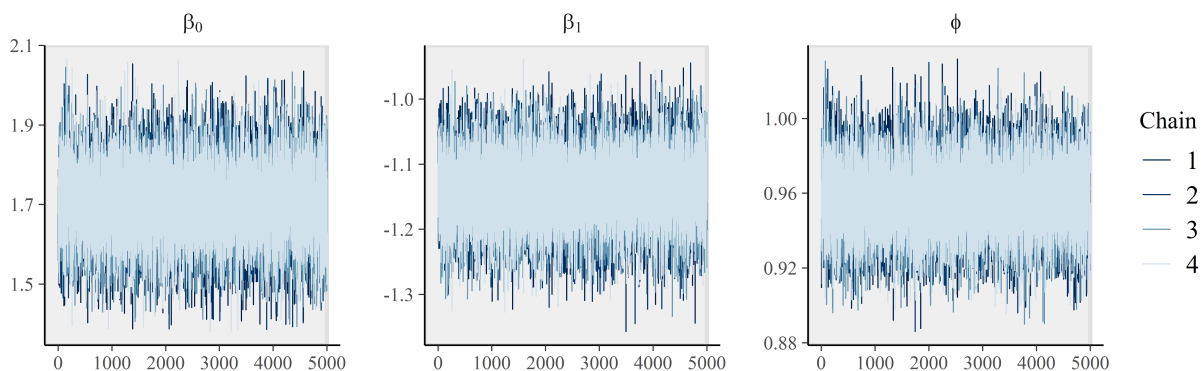


Figure E.10: Post-burn-in traceplots for regression coefficients under long simulation conditions.

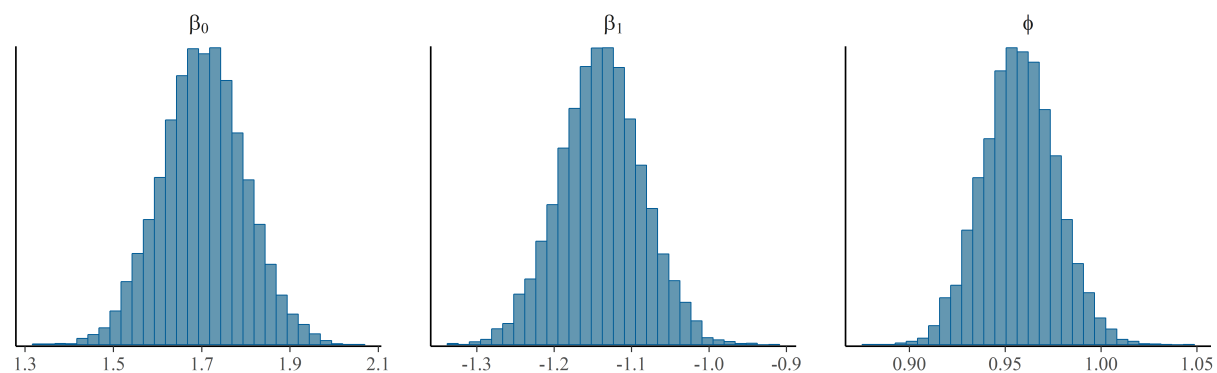


Figure E.11: Histograms of posterior distributions for regression coefficients under short simulation conditions.

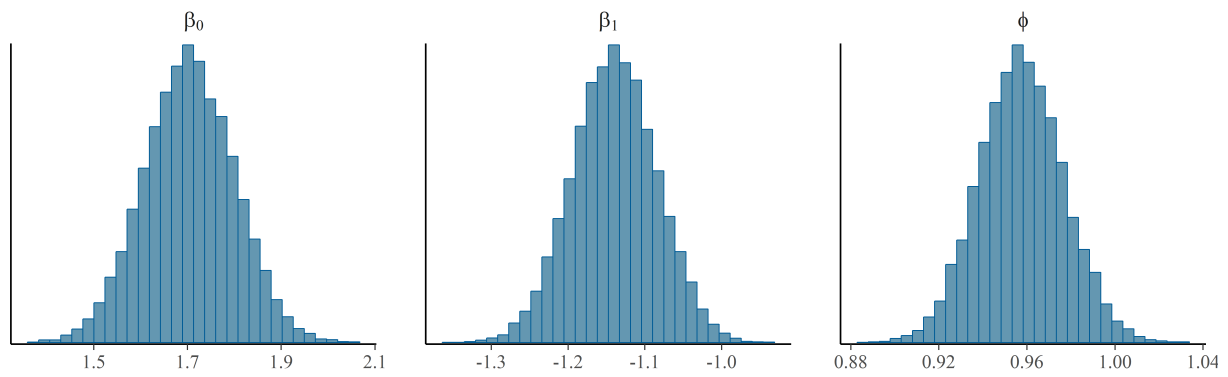


Figure E.12: Histograms of posterior distributions for regression coefficients under short simulation conditions.

E.2 Exact Formulas for Markov Cohort Simulation

Markov cohort simulation is used to determine the incremental quality-of-life-years (ΔQALYs) gained/lost by patients undergoing treatment y versus treatment x . Markov cohort simulation involves three steps for each treatment arm:[22]

1. simulate state membership fractions over a time-horizon of interest
2. weigh state membership fractions by their utility
3. sum the weighed state membership fractions over the time-horizon of interest

The formulas presented in the remainder of this section apply to either the RT-alone or mS+RT treatment arm. θ will be used to denote event probabilities for any treatment arm and baseline ambulatory status stratum (Appendix E.1.1). θ should be replaced by the appropriate strata-specific parameter in the formulas. λ will be used to denote transition rates. For the mS+RT arm, λ should be replaced by $\lambda\beta$.

The simulation model (Figure 6.1) consists of three health states: ambulatory (a), non-ambulatory (n), and dead (d). The fraction of simulated patients in each health state at time t is recorded in the state membership vector $\mathbf{m}(t)$. The state membership fractions at the start of simulation are the event probabilities for the appropriate treatment arm and baseline ambulatory status stratum (Section E.1.1); $\mathbf{m}(0) = [\theta_a, \theta_n, \theta_d]$. $\mathbf{m}(t)$ is computed as the product of $\mathbf{m}(0)$ and the transition probability matrix $\mathbf{P}(t)$. [40]

The first element of $\mathbf{m}(t)$ is $A(t)$, the fraction of ambulatory patients

$$A(t) = \theta_a e^{-(\lambda_{an} + \lambda_{ad})t}.$$

The second element of $\mathbf{m}(t)$ is $N(t)$, the fraction of non-ambulatory patients

$$N(t) = \theta_n e^{-\lambda_{nd}t} - \frac{\theta_a \lambda_{an} e^{-\lambda_{nd}t} (-1 + e^{-(\lambda_{an} + \lambda_{ad} - \lambda_{nd})t})}{\lambda_{an} + \lambda_{ad} - \lambda_{nd}}.$$

If the utility of the ambulatory state is W_a ; the utility of the non-ambulatory state is W_n ; and the utility of the dead state is 0, the total QALYs gained/lost by a simulated cohort can be expressed as a sum of integrals,

$$\begin{aligned} \text{QALYs} &= W_a \int_0^\infty A(t) dt + W_n \int_0^\infty N(t) dt \\ &= \frac{\theta_a (W_a \lambda_{nd} + W_n \lambda_{an}) + \theta_n W_n (\lambda_{an} + \lambda_{ad})}{(\lambda_{an} + \lambda_{ad}) \lambda_{nd}}. \end{aligned}$$

References

- [1] A. Allignol, M. Schumacher, and J. Beyersmann. Empirical Transition Matrix of Multi-State Models: The etm Package. *J. Stat. Softw.*, 38(4), 2011.
- [2] D. G. Altman and J. M. Bland. Diagnostic tests 2: Predictive values. *BMJ*, 309(6947):102, jul 1994.
- [3] A. Amelot, L. Balabaud, D. Choi, Z. Fox, H. A. Crockard, T. Albert, C. M. Arts, J. M. Buchowski, C. Bunger, C. K. Chung, M. H. Coppes, B. Depreitere, M. G. Fehlings, J. Harrop, N. Kawahara, E. S. Kim, C. S. Lee, Y. Leung, Z. J. Liu, J. A. Martin-Benlloch, E. M. Massicotte, B. Meyer, F. C. Oner, W. Peul, N. Quraishi, Y. Tokuhashi, K. Tomita, C. Ulbricht, J. J. Verlaan, M. Wang, and C. Mazel. Surgery for metastatic spine tumors in the elderly. Advanced age is not a contraindication to surgery! *Spine J.*, 17(6):759–767, jul 2017.
- [4] P. K. Andersen, S. Z. Abildstrom, and S. Rosthøj. Competing risks as a multi-state model. *Stat. Methods Med. Res.*, 11(2):203–15, apr 2002.
- [5] K. J. Arrow and R. C. Lind. Uncertainty and the Evaluation Decisions of Public Investment. *Am. Econ. Rev.*, 60(3):364–378, 1970.
- [6] L. A. Augestad, K. Stavem, I. S. Kristiansen, C. H. Samuelsen, and K. Rand-Hendriksen. Influenced from the start: anchoring bias in time trade-off valuations. *Qual. Life Res.*, 25(9):2179–2191, 2016.
- [7] F. Augustovski, L. Rey-Ares, V. Irazola, M. Oppe, and N. J. Devlin. Lead versus lag-time trade-off variants: does it make any difference?. *Eur. J. Health Econ.*, 14 Suppl 1:S25–31, 2013.
- [8] L. Azzimonti, G. Corani, and M. Zaffalon. Hierarchical Multinomial-Dirichlet model for the estimation of conditional probability tables, aug 2017.
- [9] A. Bagust and S. Beale. Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach. *Med. Decis. Making*, 34(3):343–51, 2014.
- [10] M. V. Bala, L. L. Wood, G. A. Zarkin, E. C. Norton, A. Gafni, and B. J. O’Brien. Are health states ”timeless”? The case of the standard gamble method. *J. Clin. Epidemiol.*, 52(11):1047–53, nov 1999.

- [11] K. P. Balanda and H. L. Macgillivray. Kurtosis : A Critical Review. *Am. Stat.*, 42(2):111–119, 1988.
- [12] N. Bansback, A. Tsuchiya, J. Brazier, and A. Anis. Canadian Valuation of EQ-5D Health States: Preliminary Value Set and Considerations for Future Valuation Studies. *PLoS One*, 7(2):e31115, feb 2012.
- [13] C. M. Bell, D. R. Urbach, J. G. Ray, A. Bayoumi, A. B. Rosen, D. Greenberg, and P. J. Neumann. Bias in published cost effectiveness studies: systematic review. *BMJ*, 332(7543):699–703, mar 2006.
- [14] M. H. Bilsky, I. Laufer, and S. Burch. Shifting Paradigms in the Treatment of Metastatic Spine Disease. *Spine (Phila. Pa. 1976)*., 34(Supplement):S101–S107, oct 2009.
- [15] N. F. Boyd, H. J. Sutherland, K. Z. Heasman, D. L. Tritchler, and B. J. Cummings. Whose utilities for decision analysis? *Med. Decis. Making*, 10(1):58–67, 1990.
- [16] C. A. Brauer, P. J. Neumann, and A. B. Rosen. Trends in Cost Effectiveness Analyses in Orthopedic Surgery. *Clin. Orthop. Relat. Res.*, PAP, jan 2007.
- [17] J. Brazier and P. Dolan. Evidence of preference construction in a comparison of variants of the standard gamble method. 2005.
- [18] R. P. Brent. *Algorithms for minimization without derivatives*. Prentice-Hall, Englewood Cliffs, N.J., 1973.
- [19] K. H. Bridwell, W. Cats-Baril, J. Harrast, S. Berven, S. Glassman, J.-P. Farcy, W. C. Horton, L. G. Lenke, C. Baldus, and T. Radake. The validity of the SRS-22 instrument in an adult spinal deformity population compared with the Oswestry and SF-12: a study of response distribution, concurrent validity, internal consistency, and reliability. *Spine (Phila. Pa. 1976)*., 30(4):455–61, feb 2005.
- [20] A. Briggs. Economic evaluation and clinical trials: size matters. *BMJ*, 321(7273):1362–3, 2000.
- [21] A. Briggs. Probabilistic analysis of cost-effectiveness models: statistical representation of parameter uncertainty.[comment]. *Value Heal.*, 8(1):1–2, 2005.
- [22] A. Briggs, K. Claxton, and M. Sculpher. *Decision Modelling for Health Economic Evaluation (Handbooks in Health Economic Evaluation)*. Oxford University Press, USA, 2006.
- [23] A. H. Briggs. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics*, 17(5):479–500, may 2000.
- [24] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *Am. J. Public Health*, 108(10):1378–1384, oct 2018.

- [25] R. Brookmeyer and J. Crowley. A Confidence Interval for the Median Survival Time. *Biometrics*, 38(1):29, mar 1982.
- [26] S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.*, 7(4):434–455, dec 2013.
- [27] A. S. Burns, D. A. Rivas, and J. F. Ditunno. The management of neurogenic bladder and sexual dysfunction after spinal cord injury. *Spine (Phila. Pa. 1976)*., 26(24 Suppl):S129–36, dec 2001.
- [28] P. R. Burton. Helping doctors to draw appropriate inferences from the analysis of medical studies. *Stat. Med.*, 13(17):1699–1713, 1994.
- [29] M. J. Buxton, M. F. Drummond, B. A. Van Hout, R. L. Prince, T. A. Sheldon, T. Szucs, and M. Vray. Modelling in economic evaluation: An unavoidable fact of life. *Health Econ.*, 6(3):217–227, 1997.
- [30] B. P. Carlin and T. A. Louis. *Bayesian methods for data analysis*. CRC Press, Boca Raton, 3rd edition, 2009.
- [31] L. Y. Carreon, P. a. Anderson, C. M. McDonough, M. Djurasovic, and S. D. Glassman. Predicting SF-6D utility scores from the neck disability index and numeric rating scales for neck and arm pain. *Spine (Phila. Pa. 1976)*., 36(6):490–494, 2011.
- [32] L. Y. Carreon, S. D. Glassman, C. M. McDonough, R. Rampersaud, S. Berven, and M. Shainline. Predicting SF-6D utility scores from the Oswestry disability index and numeric rating scales for back and leg pain. *Spine (Phila. Pa. 1976)*., 34(19):2085–9, sep 2009.
- [33] G. Cervellin, L. Borghi, and G. Lippi. Do clinicians decide relying primarily on Bayesians principles or on Gestalt perception? Some pearls and pitfalls of Gestalt perception in medicine. *Intern. Emerg. Med.*, 9(5):513–519, 2014.
- [34] Y. Chen, M. J. Devivo, and A. B. Jackson. Pressure ulcer prevalence in people with spinal cord injury: age-period-duration effects. *Arch. Phys. Med. Rehabil.*, 86(6):1208–13, jun 2005.
- [35] K. Claxton. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J. Health Econ.*, 18(3):341–364, jun 1999.
- [36] K. Claxton, M. Sculpher, C. McCabe, A. Briggs, R. Akehurst, M. Buxton, J. Brazier, and T. O’Hagan. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health Econ.*, 14(4):339–347, apr 2005.
- [37] I. Cleemput, M. Neyt, N. Thiry, C. De Laet, and M. Leys. Using threshold values for cost per quality-adjusted life-year gained in healthcare decisions. *Int. J. Technol. Assess. Health Care*, 27(1):71–6, jan 2011.

- [38] M. Connock, C. Hyde, and D. Moore. Cautions regarding the fitting and interpretation of survival curves: examples from NICE single technology appraisals of drugs for cancer. *Pharmacoeconomics*, 29(10):827–37, oct 2011.
- [39] M. K. Cowles and B. P. Carlin. Markov Chain Monte Carlo Convergence Diagnostics : A Comparative Review. *J. Am. Stat. Assoc.*, 91(434):883–904, 1996.
- [40] D. R. Cox and H. D. Miller. *The theory of stochastic processes*. Wiley, New York, 1965.
- [41] G. Cumming, J. Williams, and F. Fidler. Replication and Researchers’ Understanding of Confidence Intervals and Standard Error Bars. *Underst. Stat.*, 3(4):299–311, 2004.
- [42] G. R. Cybulski. Methods of surgical stabilization for metastatic disease of the spine. *Neurosurgery*, 25(2):240–52, aug 1989.
- [43] S. Datta. Estimating the mean life time using right censored data. *Stat. Methodol.*, 2(1):65–69, mar 2005.
- [44] D. De Looze, M. Van Laere, M. De Muynck, R. Beke, and A. Elewaut. Constipation and other chronic gastrointestinal problems in spinal cord injury patients. *Spinal Cord*, 36(1):63–6, jan 1998.
- [45] H. C. de Vet, L. M. Bouter, P. D. Bezemer, and a. J. Beurskens. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. *Int. J. Technol. Assess. Health Care*, 17(4):479–487, 2001.
- [46] G. A. De Wit, J. J. V. Busschbach, and F. T. De Charro. Sensitivity and perspective in the valuation of health status: Whose values count? *Health Econ.*, 9(2):109–126, 2000.
- [47] N. Dea, A. Versteeg, C. Fisher, A. Kelly, D. Hartig, M. Boyd, S. Paquette, B. K. Kwon, M. Dvorak, and J. Street. Adverse events in emergency oncological spine surgery: a prospective analysis. *J. Neurosurg. Spine*, 21(5):698–703, 2014.
- [48] S. Depaoli and R. van de Schoot. Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychol. Methods*, 22(2):240–261, 2017.
- [49] N. J. Devlin, P. Hansen, P. Kind, and A. Williams. Logical inconsistencies in survey respondents’ health state valuations - A methodological challenge for estimating social tariffs. *Health Econ.*, 12(7):529–544, 2003.
- [50] P. Dolan, C. Gudex, P. Kind, and A. Williams. A social tariff for EuroQol: results from a UK general population survey. 1995.
- [51] P. Dolan, C. Gudex, P. Kind, and A. Williams. Valuing health states: a comparison of methods. *J. Health Econ.*, 15(2):209–31, apr 1996.

- [52] M. G. T. Dolders, M. P. A. Zeegers, W. Groot, and A. Ament. A meta-analysis demonstrates no significant differences between patient and population preferences. *J. Clin. Epidemiol.*, 59(7):653–64, jul 2006.
- [53] F. J. Dorey and E. L. Korn. Effective sample sizes for confidence intervals for survival probabilities. *Stat. Med.*, 6(6):679–687, 1987.
- [54] M. Drummond, D. Brixner, M. Gold, P. Kind, A. McGuire, and E. Nord. Toward a Consensus on the QALY. *Value Heal.*, 12:S31–S35, mar 2009.
- [55] M. F. Drummond, M. J. Sculpher, G. W. Torrance, B. J. O’Brien, and G. L. Stoddart. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford University Press, USA, 2005.
- [56] D. M. Eddy. Accuracy versus transparency in pharmacoeconomic modelling: finding the right balance. *Pharmacoeconomics*, 24(9):837–44, 2006.
- [57] D. M. Eddy, W. Hollingworth, J. J. Caro, J. Tsevat, K. M. McDonald, and J. B. Wong. Model transparency and validation: A Report of the ISPOR-SMDM Good Research Practices Task Force -7. *Value Heal.*, 15:843–850, 2012.
- [58] N. C. Edwards, A. C. Skelly, J. E. Ziewacz, K. Cahill, and M. J. McGirt. The role of decision analytic modeling in the health economic assessment of spinal intervention. *Spine (Phila. Pa. 1976)*., 39(22 Suppl 1):S16–42, 2014.
- [59] L. Engel, N. Bansback, S. Bryan, M. M. Doyle-Waters, and D. G. Whitehurst. Exclusion Criteria in National Health State Valuation Studies: A Systematic Review. *Med. Decis. Mak.*, 36(7):798–810, 2016.
- [60] M. Evans, H. H. Jensen, M. Bogelund, J. Gundgaard, B. Chubb, and K. Khunti. Flexible insulin dosing improves health-related quality-of-life (HRQoL): a time trade-off survey. *J. Med. Econ.*, 16(11):1357–1365, 2013.
- [61] M. Evans, K. Khunti, M. Mamdani, C. B. Galbo-Jorgensen, J. Gundgaard, M. Bogelund, and S. Harris. Health-related quality of life associated with daytime and nocturnal hypoglycaemic events: a time trade-off survey in five countries. *Health Qual. Life Outcomes*, 11:90, 2013.
- [62] Expert Panel on Radiation Oncology-Bone Metastases, S. S.-M. Lo, S. Ryu, E. L. Chang, N. Galanopoulos, J. Jones, E. Y. Kim, C. D. Kubicky, C. P. Lee, P. S. Rose, A. Sahgal, A. E. Sloan, B. S. Teh, B. J. Traughber, C. Van Poznak, and A. D. Vassil. ACR Appropriateness Criteria® Metastatic Epidural Spinal Cord Compression and Recurrent Spinal Metastasis. *J. Palliat. Med.*, 18(7):573–84, jul 2015.
- [63] K. Fairman and F. Curtiss. It’s Only a Pharmacoeconomic Model–Believe It or Not. *J Manag Care Pharm*, 14(1):83–85, 2008.

- [64] J. Fan, S. Upadhye, and A. Worster. Understanding receiver operating characteristic (ROC) curves. *CJEM Can. J. Emerg. Med. care = JCMU J. Can. soins médicaux d'urgence*, 8(1):19–20, 2006.
- [65] M. G. Fehlings, A. Nater, L. Tetreault, B. Kopjar, P. Arnold, M. Dekutoski, J. Finkelstein, C. Fisher, J. France, Z. Gokaslan, E. Massicotte, L. Rhines, P. Rose, A. Sahgal, J. Schuster, and A. Vaccaro. Survival and Clinical Outcomes in Surgically Treated Patients With Metastatic Epidural Spinal Cord Compression: Results of the Prospective Multicenter AOSpine Study. *J. Clin. Oncol.*, 34(3):268–276, nov 2015.
- [66] P. D. Feigin. Maximum likelihood estimation for continuous-time stochastic processes. *Adv. Appl. Probab.*, 8(04):712–736, dec 1976.
- [67] S. Ferrari and F. Cribari-Neto. Beta Regression for Modelling Rates and Proportions. *J. Appl. Stat.*, 31(7):799–815, aug 2004.
- [68] S. E. Fienberg. When did Bayesian inference become "Bayesian"? *Bayesian Anal.*, 1(1):1–40, mar 2006.
- [69] J. I. Figueroa-Zúñiga, R. B. Arellano-Valle, and S. L. Ferrari. Mixed beta regression: A Bayesian perspective. *Comput. Stat. Data Anal.*, 61:137–147, may 2013.
- [70] G. F. Findlay. The role of vertebral body collapse in the management of malignant spinal cord compression. *J. Neurol. Neurosurg. Psychiatry*, 50(2):151–4, feb 1987.
- [71] L. H. Fischer, J. M. Abzug, A. L. Osterman, P. J. Stern, and J. Chang. Complications of common hand and wrist surgery procedures: flexor and extensor tendon surgery. *Instr. Course Lect.*, 63:97–103, jan 2014.
- [72] R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 222(594-604):309–368, jan 1922.
- [73] V. L. Fornasier and J. G. Horne. Metastases to the vertebral column. *Cancer*, 36(2):590–4, aug 1975.
- [74] R. Fuller, N. Dudley, and J. Blacktop. How informed is consent? Understanding of pictorial and verbal probability information by medical inpatients. *Postgrad. Med. J.*, 78(923):543–4, sep 2002.
- [75] J. C. Furlan, K. K.-W. Chan, G. a. Sandoval, K. C. K. Lam, C. a. Klinger, R. A. Patchell, A. Laporte, and M. G. Fehlings. The combined use of surgery and radiotherapy to treat patients with epidural cord compression due to metastatic disease: a cost-utility analysis. *Neuro. Oncol.*, 14(5):631–40, may 2012.
- [76] W. Furlong, D. Feeny, G. W. Torrance, R. Barr, and J. Horsman. Guide to design and development of health-state utility instrumentation. 1990.
- [77] D. M. Galvis, D. Bandyopadhyay, and V. H. Lachos. Augmented mixed beta regression models for periodontal proportion data. *Stat. Med.*, 33(21):3759–3771, 2014.

- [78] J. C. Gardiner, Z. Luo, and L. A. Roman. Fixed effects, random effects and GEE: what are the differences? *Stat. Med.*, 28(2):221–39, jan 2009.
- [79] M. Gattellari, P. N. Butow, M. H. N. Tattersall, S. M. Dunn, and C. A. MacLeod. Misunderstanding in cancer patients: Why shoot the messenger? *Ann. Oncol.*, 10(1):39–46, 1999.
- [80] A. Gelman, B. P. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, second edition, 2003.
- [81] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. R. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, third edition, 2014.
- [82] M. Gheorghe, S. Picavet, M. Verschuren, W. B. Brouwer, and P. H. van Baal. Health losses at the end of life: a Bayesian mixed beta regression approach. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 180(3):723–749, 2017.
- [83] R. W. Gilbert, J. H. Kim, and J. B. Posner. Epidural spinal cord compression from metastatic tumor: diagnosis and treatment. *Ann. Neurol.*, 3(1):40–51, jan 1978.
- [84] E. Goodwin, C. Green, and A. Spencer. Estimating a Preference-Based Index for an Eight-Dimensional Health State Classification System for Multiple Sclerosis. *Value Health*, 18(8):1025–1036, 2015.
- [85] E. J. Graham, L. G. Lenke, T. G. Lowe, R. R. Betz, K. H. Bridwell, Y. Kong, and K. Blanke. Prospective pulmonary function evaluation following open thoracotomy for anterior spinal fusion in adolescent idiopathic scoliosis. *Spine (Phila. Pa. 1976)*., 25(18):2319–25, sep 2000.
- [86] S. B. Green. How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behav. Res.*, 26(3):499–510, 1991.
- [87] T. M. Gregory, R. Coriat, and O. Mir. Prognostic scoring systems for spinal metastases in the era of anti-VEGF therapies. *Spine (Phila. Pa. 1976)*., 38(11):965–966, 2013.
- [88] R. Grieve, N. Hawkins, and M. Pennington. Extrapolation of survival data in cost-effectiveness analyses: improving the current state of play. *Med. Decis. Making*, 33(6):740–2, 2013.
- [89] D. a. Grimes and K. F. Schulz. Refining clinical diagnosis with likelihood ratios. *Lancet*, 365(panel 1):1500–1505, 2005.
- [90] G. Guyatt, S. Walter, and G. Norman. Measuring Change Over Time- Aseessing the Usefulness of Evaluative Instruments. *J. Chronic Dis.*, 40(2):171–178, 1987.
- [91] G. H. Guyatt, R. A. Deyo, M. Charlson, M. N. Levine, and A. Mitchell. Responsiveness and validity in health status measurement: A clarification. *J. Clin. Epidemiol.*, 42(5):403–408, jan 1989.

- [92] P. Guyot, A. E. Ades, M. J. N. M. Ouwers, and N. J. Welton. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med. Res. Methodol.*, 12(1):9, feb 2012.
- [93] R. G. Hagerty, P. N. Butow, P. A. Ellis, E. A. Lobb, S. Pendlebury, N. Leighl, D. Goldstein, S. K. Lo, and M. H. N. Tattersall. Cancer patient preferences for communication of prognosis in the metastatic setting. *J. Clin. Oncol.*, 22(9):1721–1730, 2004.
- [94] M. Haight, A. Quan-Haase, and B. a. Corbett. Revisiting the digital divide in Canada: The impact of demographic factors on access to the Internet, level of online activity, and social networking site usage. *Information, Commun. Soc.*, 17(4):503–519, 2014.
- [95] E. F. Halpern. Behind the numbers: inverse probability weighting. *Radiology*, 271(3):625–8, jun 2014.
- [96] L. C. Hamilton, J. Hartter, and K. Saito. Trust in Scientists on Climate Change and Vaccines. *SAGE Open*, 5(3), 2015.
- [97] T. Hammerschmidt, H.-P. Zeitler, M. Gulich, and R. Leidl. A comparison of different strategies to collect standard gamble utilities. *Med. Decis. Making*, 24(5):493–503, 2004.
- [98] J. A. Hanley. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit. Rev. Diagn. Imaging*, 29(3):307–35, 1989.
- [99] J. a. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, apr 1982.
- [100] F. E. Harrell. *Regression Modeling Strategies*. Springer Series in Statistics. Springer International Publishing, Cham, 2 edition, 2015.
- [101] S. Harris, M. Mamdani, C. B. Galbo-Jorgensen, M. Bogelund, J. Gundgaard, and D. Groleau. The effect of hypoglycemia on health-related quality of life: Canadian results from a multinational time trade-off survey. *Can. J. diabetes*, 38(1):45–52, 2014.
- [102] N. Hawkins and R. Grieve. Extrapolation of Survival Data in Cost-effectiveness Analyses: The Need for Causal Clarity. *Med. Decis. Making*, 37(4):337–339, 2017.
- [103] J. M. Herman and V. K. Sonntag. Cervical corpectomy and plate fixation for post-laminectomy kyphosis. *J. Neurosurg.*, 80(6):963–70, jun 1994.
- [104] J. P. T. Higgins and D. J. Spiegelhalter. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int. J. Epidemiol.*, 31(1):96–104, 2002.
- [105] P. Hougaard. Multi-state Models: A Review. *Lifetime Data Anal.*, 5(3), 1999.

- [106] M. Huo, A. Sahgal, D. Pryor, K. Redmond, S. Lo, and M. Foote. Stereotactic spine radiosurgery: Review of safety and efficacy with respect to dose and fractionation. *Surg. Neurol. Int.*, 8(1):30, 2017.
- [107] M. Huo, A. Sahgal, D. Pryor, K. Redmond, S. Lo, and M. Foote. Stereotactic spine radiosurgery: Review of safety and efficacy with respect to dose and fractionation. *Surg. Neurol. Int.*, 8:30, 2017.
- [108] J. A. Husted, R. J. Cook, V. T. Farewell, and D. D. Gladman. Methods for assessing responsiveness. *J. Clin. Epidemiol.*, 53(5):459–468, 2000.
- [109] R. Hutchins, A. J. Viera, S. L. Sheridan, and M. P. Pignone. Quantifying the utility of taking pills for cardiovascular prevention. *Circ. Cardiovasc. Qual. Outcomes*, 8(2):155–163, 2015.
- [110] R. Jaeschke, G. H. Guyatt, and D. L. Sackett. Users’ guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA*, 271(9):703–7, mar 1994.
- [111] S. J. Janssen, T. Teunis, E. van Dijk, M. L. Ferrone, J. H. Shin, F. Hornicek, and J. H. Schwab. Validation of the Spine Oncology Study Group—Outcomes Questionnaire to assess quality of life in patients with metastatic spine disease. *Spine J.*, 17(6):768–776, 2017.
- [112] F. R. Johnson, E. Lancsar, D. Marshall, V. Kilambi, A. Mühlbacher, D. A. Regier, B. W. Bresnahan, B. Kanninen, and J. F. Bridges. Constructing experimental designs for discrete-choice experiments: Report of the ISPOR conjoint analysis experimental design good research practices task force. *Value Heal.*, 16(1):3–13, 2013.
- [113] J. A. Johnson, N. Luo, J. W. Shaw, P. Kind, and S. J. Coons. Valuations of EQ-5D health states: are the United States and United Kingdom different? *Med Care*, 43(3):221–228, 2005.
- [114] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. Wiley, third edition, 2005.
- [115] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Volume 2*. Wiley, second edition, 1994.
- [116] M. Jones-Lee, G. Loomes, D. O’Reilly, and P. Philips. The Value of Preventing Non Fatal Road Injuries: Findings of a Willingness to Pay National Sample Survey. Technical report, Transport Research Laboratory, Wokingham, United Kingdom, 1993.
- [117] T. R. Jorgensen, C. Emborg, K. Dahlen, M. Bogelund, and A. Carlborg. The effect of the medicine administration route on health-related quality of life: Results from a time trade-off survey in patients with bipolar disorder or schizophrenia in 2 Nordic countries. *BMC Psychiatry*, 16:244, 2016.

- [118] M. B. Kabins and J. N. Weinstein. The History of Vertebral Screw and Pedicle Screw Fixation. *Iowa Orthop. J.*, 11:127–136, 1991.
- [119] J. D. Kalbfleisch and J. F. Lawless. The Analysis of Panel Data under a Markov Assumption. *J. Am. Stat. Assoc.*, 80(392):863–871, dec 1985.
- [120] J. D. Kalbfleisch and R. L. Pretince. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, Hoboken, NJ, 2nd ed. edition, 2002.
- [121] A. Karhade, V. Vasudeva, H. Dasenbrock, Y. Lu, W. Gormley, and W. Groff. Thirty-day readmission and reoperation after surgery for spinal tumors: a National Surgical Quality Improvement Program analysis. *Neurosurg Focus*, 41(2):E5, 2016.
- [122] R. Kay. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42(4):855–65, dec 1986.
- [123] P. Kind, J. E. Lafata, K. Matuszewski, and D. Raisch. The use of QALYs in clinical and patient decision-making: issues and prospects. *Value Health*, 12 Suppl 1:S27–30, mar 2009.
- [124] J. R. Kruschke. *Doing bayesian data analysis: a tutorial with r and bugs*. Academic Press Inc., Oxford, 2011.
- [125] J. Kuha. AIC and BIC: Comparisons of Assumptions and Performance. *Sociol. Methods Res.*, 33(2):188–229, nov 2004.
- [126] K. Y. H. Kwan, T. C. Lam, H. C. W. Choi, H. Y. Koh, and K. M. C. Cheung. Prediction of survival in patients with symptomatic spinal metastases: Comparison between the Tokuhashi score and expert oncologists. *Surg. Oncol.*, 27(1):7–10, 2018.
- [127] B. M. Lagos-Álvarez, R. Fustos-Toribio, J. Figueroa-Zúñiga, and J. Mateu. Geostatistical mixed beta regression: a Bayesian approach. *Stoch. Environ. Res. Risk Assess.*, 31(2):571–584, 2017.
- [128] L. M. Lamers, P. F. M. Stalmeier, P. F. M. Krabbe, and J. J. V. Busschbach. Inconsistencies in TTO and VAS values for EQ-5D health states. *Med. Decis. Mak.*, 26(2):173–181, 2006.
- [129] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- [130] N. R. Latimer. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Med. Decis. Making*, 33(6):743–54, aug 2013.
- [131] I. Laufer, J. B. Iorgulescu, T. Chapman, E. Lis, W. Shi, Z. Zhang, B. W. Cox, Y. Yamada, and M. H. Bilsky. Local disease control for spinal metastases following ”separation surgery” and adjuvant hypofractionated or high-dose single-fraction stereotactic radiosurgery: outcome analysis in 186 patients. *J. Neurosurg. Spine*, 18(3):207–14, mar 2013.

- [132] I. Laufer, D. G. Rubin, E. Lis, B. W. Cox, M. D. Stubblefield, Y. Yamada, and M. H. Bilsky. The NOMS framework: approach to the treatment of spinal metastatic tumors. *Oncologist*, 18(6):744–51, jun 2013.
- [133] C.-H. Lee, C. K. Chung, T.-A. Jahng, K.-j. Kim, C. H. Kim, S.-J. Hyun, H.-J. Kim, S. R. Jeon, U.-K. Chang, S.-H. Lee, S.-H. Moon, H. Majeed, D. Zhang, G. Gravis, C. Wibmer, N. Kumar, K. Y. Moon, J. H. Park, E. Tabouret, and S. Fuentes. Which one is a valuable surrogate for predicting survival between Tomita and Tokuhashi scores in patients with spinal metastases? A meta-analysis for diagnostic test accuracy and individual participant data analysis. *J. Neurooncol.*, 123(2):267–75, jun 2015.
- [134] C. H. Lee, J. Ning, and Y. Shen. Analysis of restricted mean survival time for length-biased data. *Biometrics*, 74(2):575–583, 2018.
- [135] J. C. Lee and D. J. Sabavala. Bayesian Estimation and Prediction for the Beta-Binomial Model. *J. Bus. Econ. Stat.*, 5(3):357–367, jul 1987.
- [136] J. S. Lee, K. T. Suh, J. I. Kim, and T. S. Goh. Analysis of Sagittal Balance of Ankylosing Spondylitis Using Spinopelvic Parameters. *J. Spinal Disord. Tech.*, 27(3):E94–E98, 2014.
- [137] L. A. Lenert, D. J. Cher, M. K. Goldstein, M. R. Bergen, and A. Garber. The effect of search procedures on utility elicitations. *Med. Decis. Making*, 18(1):76–83, 1998.
- [138] L. A. Lenert and A. E. Sturley. Use of the internet to study the utility values of the public. *Proc. AMIA Symp.*, pages 440–444, 2002.
- [139] S. Lewandowsky, G. E. Gignac, and K. Oberauer. The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science. *PLoS One*, 8(10):e75637, oct 2013.
- [140] S. Lewandowsky, G. E. Gignac, and S. Vaughan. The pivotal role of perceived scientific consensus in acceptance of science. *Nat. Clim. Chang.*, 3(4):399–404, apr 2013.
- [141] I. M. Lipkus. Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Med. Decis. Mak.*, 27(5):696–713, 2007.
- [142] H. Llewellyn-Thomas, H. J. Sutherland, R. Tibshirani, A. Ciampi, J. E. Till, and N. F. Boyd. Describing health states. Methodologic issues in obtaining values for health states. *Med. Care*, 22(6):543–52, jun 1984.
- [143] J. C. Lloyd, T. Yen, R. Pietrobon, J. S. Wiener, S. S. Ross, P. J. Kokorowski, C. P. Nelson, and J. C. Routh. Estimating utility values for vesicoureteral reflux in the general public using an online tool. *J. Pediatr. Urol.*, 10(6):1026–1031, 2014.

- [144] S. Lutz, T. Balboni, J. Jones, S. Lo, J. Petit, S. E. Rich, R. Wong, and C. Hahn. Palliative radiation therapy for bone metastases: Update of an ASTRO Evidence-Based Guideline. *Pract. Radiat. Oncol.*, 7(1):4–12, 2017.
- [145] E. Maranzano, R. Bellavita, R. Rossi, V. De Angelis, A. Frattegiani, R. Bagnoli, M. Mignogna, S. Beneventi, M. Lupattelli, P. Ponticelli, G. P. Biti, and P. Latini. Short-course versus split-course radiotherapy in metastatic spinal cord compression: results of a phase III, randomized, multicenter trial. *J. Clin. Oncol.*, 23(15):3358–65, may 2005.
- [146] E. Maranzano, P. Latini, E. Perrucci, S. Beneventi, M. Lupattelli, and E. Corgna. Short-course radiotherapy (8 Gy x 2) in metastatic spinal cord compression: an effective and feasible treatment. *Int. J. Radiat. Oncol. Biol. Phys.*, 38(5):1037–44, jul 1997.
- [147] E. Maranzano, F. Trippa, M. Casale, S. Costantini, M. Lupattelli, R. Bellavita, L. Marafioti, S. Pergolizzi, A. Santacaterina, M. Mignogna, G. Silvano, and V. Fusco. 8Gy single-dose radiotherapy is effective in metastatic spinal cord compression: results of a phase III randomized multicentre Italian trial. *Radiother. Oncol.*, 93(2):174–9, nov 2009.
- [148] D. Mayo. In Defense of the Neyman-Pearson Theory of Confidence Intervals. *Philos. Sci.*, 48(2):269–280, 1981.
- [149] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, London, 2 edition, 1989.
- [150] R. M. Merchant and D. A. Asch. Protecting the Value of Medical Science in the Age of Social Media and "Fake News". *JAMA*, nov 2018.
- [151] G. Mitera, A. Loblaw, A. Sahgal, and B. Danielson. Canadian-led International Development of a European Organization for Research and Treatment of Cancer Quality of Life Module for Malignant Spinal Cord Compression: Results of Phase I. *Int. J. Radiat. Oncol.*, 78(3):S604, nov 2010.
- [152] L. B. Mokkink, C. B. Terwee, D. L. Patrick, J. Alonso, P. W. Stratford, D. L. Knol, L. M. Bouter, and H. C. W. de Vet. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J. Clin. Epidemiol.*, 63(7):737–745, 2010.
- [153] L. A. Moye. *Elementary bayesian biostatistics*. Chapman & Hall/CRC, Boca Raton, 2008.
- [154] T. Muka, D. Imo, L. Jaspers, V. Colpani, L. Chaker, S. J. van der Lee, S. Mendis, R. Chowdhury, W. M. Bramer, A. Falla, R. Pazoki, and O. H. Franco. The global impact of non-communicable diseases on healthcare spending and national income: a systematic review. *Eur. J. Epidemiol.*, 30(4):251–277, 2015.

- [155] J. Neyman. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philos. Trans. R. Soc. London. Ser. A, Math. Phys. Sci.*, 236(767):333–380, 1937.
- [156] E. Nord, N. Daniels, and M. Kamlet. QALYs: Some Challenges. *Value Heal.*, 12(10):S10–S15, mar 2009.
- [157] R. Norman, R. Viney, N. K. Aaronson, J. E. Brazier, D. Cella, D. S. J. Costa, P. M. Fayers, G. Kemmler, S. Peacock, A. S. Pickard, D. Rowen, D. J. Street, G. Velikova, T. A. Young, and M. T. King. Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format. *Qual. Life Res.*, 25(3):637–49, mar 2016.
- [158] J. Nyland, P. Quigley, C. Huang, J. Lloyd, J. Harrow, and A. Nelson. Preserving transfer independence among individuals with spinal cord injury. *Spinal Cord*, 38(11):649–57, nov 2000.
- [159] S. Olofsson, H. Norrlid, and U. Persson. Preferences for improvements in attributes associated with basal insulin: a time trade-off and willingness-to-pay survey of a diabetic and non-diabetic population in Sweden. *J. Med. Econ.*, 19(10):945–958, 2016.
- [160] S. Olofsson, H. Norrlid, and U. Persson. Preferences for improvements in attributes associated with basal insulin: a time trade-off and willingness-to-pay survey of a diabetic and non-diabetic population in Sweden. *J. Med. Econ.*, 19(10):945–58, oct 2016.
- [161] M. Oppe, N. J. Devlin, B. van Hout, P. F. M. Krabbe, and F. de Charro. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*, 17(4):445–53, jun 2014.
- [162] M. Oppe, K. Rand-Hendriksen, K. Shah, J. M. Ramos-Goñi, and N. Luo. EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. *Pharmacoeconomics*, 34(10):993–1004, 2016.
- [163] J. A. Ortiz Gómez. The incidence of vertebral body metastases. *Int. Orthop.*, 19(5):309–11, 1995.
- [164] M. Osborne. An effective method for computing regression quantiles. *IMA J. Numer. Anal.*, 12(2):151–166, 1992.
- [165] R. Ospina and S. L. P. Ferrari. Inflated beta distributions. *Stat. Pap.*, 51(1):111–126, jan 2010.
- [166] M. Pahuta, A. Formbach, G. Mitera, D. Coyle, J. Werier, and E. Wai. Validation of the self-administered online assessment of preferences (SOAP) utility elicitation tool. *Can. J. Surg.*, 59(3 Suppl 2):S40, 2016.

- [167] M. Pahuta, A. Frombach, E. Hashem, S. Spence, C. Sun, E. K. Wai, J. Werier, C. van Walraven, and D. Coyle. The Psychometric Properties of a Self-Administered, Open-Source Module for Valuing Metastatic Epidural Spinal Cord Compression Utilities. *PharmacoEconomics - open*, 3(2):197–204, jun 2019.
- [168] M. A. Pahuta, E. K. Wai, J. Werier, C. van Walraven, and D. Coyle. A General Population Utility Valuation Study For Metastatic Epidural Spinal Cord Compression Health States. *Spine (Phila. Pa. 1976)*., 44(13):943–950, 2019.
- [169] M. A. Pahuta, J. Werier, E. K. Wai, C. van Walraven, and D. Coyle. Back to Bayesian: A strategy to enhance prognostication of metastatic spine disease. *Int. J. Clin. Pract.*, 73(4):e13322, apr 2019.
- [170] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Inc., New York, third edition, 1991.
- [171] R. a. Patchell, P. a. Tibbs, W. F. Regine, R. Payne, S. Saris, R. J. Kryscio, M. Mo-hiuddin, and B. Young. Direct decompressive surgical resection in the treatment of spinal cord compression caused by metastatic cancer: a randomised trial. *Lancet (London, England)*, 366(9486):643–8, aug 2005.
- [172] G. R. Paton, E. Frangou, and D. R. Fourney. Contemporary treatment strategy for spinal metastasis: the "LMNOP" system. *Can. J. Neurol. Sci.*, 38(3):396–403, may 2011.
- [173] A. Penaloza, F. Verschuren, G. Meyer, S. Quentin-Georget, C. Soulie, F. Thys, and P.-M. Roy. Comparison of the unstructured clinician gestalt, the wells score, and the revised Geneva score to estimate pretest probability for suspected pulmonary embolism. *Ann. Emerg. Med.*, 62(2):117–124.e2, aug 2013.
- [174] R. R. Picard and R. D. Cook. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.*, 79(387):575–583, sep 1984.
- [175] A. S. Pickard, M. P. Neary, and D. Cella. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health Qual. Life Outcomes*, 5(1):70, nov 2007.
- [176] M. Pignone, S. Saha, T. Hoerger, K. N. Lohr, S. Teutsch, and J. Mandelblatt. Challenges in systematic reviews of economic analyses. *Ann. Intern. Med.*, 142(12 Pt 2):1073–9, jun 2005.
- [177] J. Pliskin, D. Shepard, and M. Weinstein. Utility Functions for Life Years and Health Status. *Oper. Res.*, 28(1):206–224, 1980.
- [178] M. J. Price, N. J. Welton, and A. E. Ades. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *Stat. Med.*, 30(2):140–51, jan 2011.

- [179] G. H. Prosser, P. J. Yates, D. J. Wood, S. E. Graves, R. N. de Steiger, and L. N. Miller. Outcome of primary resurfacing hip replacement: evaluation of risk factors for early revision. *Acta Orthop.*, 81(1):66–71, 2010.
- [180] L. A. Prosser, K. Payne, D. Rusinak, P. Shi, and M. Messonnier. Using a discrete choice experiment to elicit time trade-off and willingness-to-pay amounts for influenza health-related quality of life at different ages. *Pharmacoeconomics*, 31(4):305–315, 2013.
- [181] H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Stat. Med.*, 26(11):2389–430, may 2007.
- [182] J. M. Pyne, J. C. Fortney, S. Tripathi, D. Feeny, P. Ubel, and J. Brazier. How Bad Is Depression? Preference Score Estimates from Depressed Patients and the General Population. *Health Serv. Res.*, 44(4):1406–1423, aug 2009.
- [183] R Core Team. R: A Language and Environment for Statistical Computing, 2018.
- [184] W. Rahman, C. J. Richards, and C. P. Duncan. Surgical nuances to minimize muscle damage during the anterolateral intermuscular approach in minimally invasive hip replacement. *Instr Course Lect*, 57:243–247, 2008.
- [185] J. Raphael, K. Chan, S. Karim, R. Kerbel, H. Lam, K. delos Santos, R. Saluja, and S. Verma. Antiangiogenic Therapy in Advanced Non-small-cell Lung Cancer: A Meta-analysis of Phase III Randomized Trials. *Clin. Lung Cancer*, 18(4):345–353.e5, 2017.
- [186] M. Rawlins. De Testimonio: On the evidence for decisions about the use of therapeutic interventions. *Clin. Med. J. R. Coll. Physicians London*, 8(6):579–588, 2008.
- [187] J. Rawls. *A Theory of Justice: Revised Edition*. Belknap Press, Cambridge, Massachusetts, United States, 2 edition, 1999.
- [188] M. Reck, L. Paz-Ares, P. Bidoli, F. Cappuzzo, S. Dakhil, D. Moro-Sibilot, H. Borghaei, M. Johnson, R. Jotte, N. A. Pennell, F. A. Shepherd, A. Tsao, M. Thomas, G. C. Carter, F. Chan-Diehl, E. Alexandris, P. Lee, A. Zimmermann, A. Sashegyi, and M. Pérol. Outcomes in patients with aggressive or refractory disease from REVEL: A randomized phase III study of docetaxel with ramucirumab or placebo for second-line treatment of stage IV non-small-cell lung cancer. *Lung Cancer*, 112(March 2017):181–187, 2017.
- [189] K. J. Redmond, S. S. Lo, C. Fisher, and A. Sahgal. Postoperative Stereotactic Body Radiation Therapy (SBRT) for Spine Metastases: A Critical Review to Guide Practice. *Int. J. Radiat. Oncol. Biol. Phys.*, 95(5):1414–28, 2016.
- [190] K. J. Redmond, S. S. Lo, C. Fisher, and A. Sahgal. Postoperative Stereotactic Body Radiation Therapy (SBRT) for Spine Metastases: A Critical Review to Guide Practice. *Int. J. Radiat. Oncol. Biol. Phys.*, 95(5):1414–1428, 2016.

- [191] N. Reid. Estimating the Median Survival Time. *Biometrika*, 68(3):601, dec 1981.
- [192] D. Rennie and H. S. Luft. Pharmacoeconomic analyses: making them transparent, making them credible. *JAMA*, 283(16):2158–60, apr 2000.
- [193] S. S. Richardson and S. Berven. The development of a model for translation of the Neck Disability Index to utility scores for cost-utility analysis in cervical disorders. *Spine J.*, 12(1):55–62, jan 2012.
- [194] P. Royston and M. K. Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med. Res. Methodol.*, 13(1):152, dec 2013.
- [195] R. L. Ruff, V. W. Adamson, S. S. Ruff, and X. Wang. Directed rehabilitation reduces pain and depression while increasing independence and satisfaction with life for patients with paraplegia due to epidural metastatic spinal cord compression. *J. Rehabil. Res. Dev.*, 44(1):1–10, 2007.
- [196] S. Ryu, J. Rock, R. Jain, M. Lu, J. Anderson, J. Y. Jin, M. Rosenblum, B. Movsas, and J. H. Kim. Radiosurgical decompression of metastatic epidural compression. *Cancer*, 116(9):2250–2257, 2010.
- [197] A. Sahgal, C. M. Whyne, L. Ma, D. A. Larson, and M. G. Fehlings. Vertebral compression fracture after stereotactic body radiotherapy for spinal metastases. *Lancet Oncol.*, 14(8):e310–e320, jul 2013.
- [198] G. D. Sanders, P. J. Neumann, A. Basu, D. W. Brock, D. Feeny, M. Krahn, K. M. Kuntz, D. O. Meltzer, D. K. Owens, L. A. Prosser, J. A. Salomon, M. J. Sculpher, T. A. Trikalinos, L. B. Russell, J. E. Siegel, and T. G. Ganiats. Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second panel on cost-effectiveness in health and medicine. *JAMA - J. Am. Med. Assoc.*, 316(10):1093–1103, 2016.
- [199] H. J. Schünemann, E. Ståhl, P. Austin, E. Akl, D. Armstrong, and G. H. Guyatt. A comparison of narrative and table formats for presenting hypothetical health states to patients with gastrointestinal or pulmonary disease. *Med. Decis. Making*, 24(1):53–60.
- [200] Schuyler W Huck. *Statistical Misconceptions*. Taylor and Francis Group, New York, 2009.
- [201] E. D. Sears, R. Shin, L. A. Prosser, and K. C. Chung. Economic analysis of revision amputation and replantation treatment of finger amputation injuries. *Plast. Reconstr. Surg.*, 133(4):827–840, 2014.
- [202] K. Seppä and T. Hakulinen. Mean and median survival times of cancer patients should be corrected for informative censoring. *J. Clin. Epidemiol.*, 62(10):1095–1102, 2009.

- [203] P. E. Shrout and J. L. Fleiss. Intraclass correlations: Uses in assessing rater reliability., 1979.
- [204] R. M. Slone, M. MacMillan, W. J. Montgomery, and M. Heare. Spinal fixation. Part 2. Fixation techniques and hardware for the thoracic and lumbosacral spine. *Radiographics*, 13(3):521–43, may 1993.
- [205] D. M. Smith, R. L. Sherriff, L. Damschroder, G. Loewenstein, and P. A. Ubel. Misremembering colostomies? Former patients give lower utility ratings than do current patients. *Health Psychol.*, 25(6):688–695, 2006.
- [206] W. R. Smith, Y.-J. Kim, B. H. Ziran, R. Ganz, J. O. Anglen, S. B. Murphy, P. F. Stahel, R. L. Buly, and M. B. Millis. Hip joint-preserving surgery: beyond the classic osteotomy. *Instr. Course Lect.*, 55:145–158, jan 2006.
- [207] M. Smithson and J. Verkuilen. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods*, 11(1):54–71, mar 2006.
- [208] G. J. Snoek, M. J. IJzerman, H. J. Hermens, D. Maxwell, and F. Biering-Sorensen. Survey of the needs of patients with spinal cord injury: impact and priority for improvement in hand function in tetraplegics. *Spinal Cord*, 42(9):526–32, sep 2004.
- [209] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, Chischester, 2004.
- [210] D. J. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. WinBUGS User Manual: version 1.4. Technical report, MRC Biostatistics Unit, Cambridge, UK, 2003.
- [211] D. E. Spratt, W. H. Beeler, F. Y. de Moraes, L. D. Rhines, J. J. Gemmete, N. Chaudhary, D. B. Shultz, S. R. Smith, A. Berlin, M. Dahele, B. J. Slotman, K. C. Younge, M. Bilsky, P. Park, and N. J. Szerlip. An integrated multidisciplinary algorithm for the management of spinal metastases: an International Spine Oncology Consortium report. *Lancet Oncol.*, 18(12):e720–e730, 2017.
- [212] Stan Development Team. The Stan Core Library, Version 2.17.0, 2017.
- [213] E. W. Steyerberg, M. J. Eijkemans, F. E. Harrell Jr., and J. D. Habbema. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med.Decis.Making*, 21(0272-989X (Print)):45–56, 2001.
- [214] A. M. Stiggelbout, G. M. Kiebert, J. Kievit, J. W. Leer, G. Stoter, and J. C. de Haes. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Med. Decis. Making*, 14(1):82–90, 1994.
- [215] M. R. Stockler, M. H. N. Tattersall, M. J. Boyer, S. J. Clarke, P. J. Beale, and R. J. Simes. Disarming the guarded prognosis: predicting survival in newly referred patients with incurable cancer. *Br. J. Cancer*, 94(2):208–12, jan 2006.

- [216] J. Street, B. Lenehan, S. Berven, and C. Fisher. Introducing a new health-related quality of life outcome tool for metastatic disease of the spine: Content validation using the international classification of functioning, disability, and health; On behalf of the spine oncology study group. *Spine (Phila. Pa. 1976)*., 35(14):1377–1386, 2010.
- [217] J. Street, B. Lenehan, S. Berven, and C. Fisher. Introducing a new health-related quality of life outcome tool for metastatic disease of the spine: content validation using the International Classification of Functioning, Disability, and Health; on behalf of the Spine Oncology Study Group. *Spine (Phila. Pa. 1976)*., 35(14):1377–86, jun 2010.
- [218] M. Stubblefield. Barriers to Rehabilitation of the Malignant Spine. *Top. Spinal Cord Inj. Rehabil.*, 14(2):19–30, oct 2008.
- [219] W. Sumner, R. Nease, and B. Littenberg. U-titer: a utility assessment tool. *Proc. Annu. Symp. Comput. Appl. Med. Care*, pages 701–5, jan 1991.
- [220] N. Sundaresan, G. V. Digiacinto, and J. E. Hughes. Surgical treatment of spinal metastases. *Clin. Neurosurg.*, 33:503–22, 1986.
- [221] N. Sundaresan, A. A. Steinberger, F. Moore, V. P. Sachdev, G. Krol, L. Hough, and K. Kelliher. Indications and results of combined anterior—posterior approaches for spine tumor surgery. *J. Neurosurg.*, 85(3):438–446, 1996.
- [222] P. Tappenden, J. Chilcott, S. Ward, S. Eggington, D. Hind, and S. Hummel. Methodological issues in the economic analysis of cancer treatments. *Eur. J. Cancer*, 42(17):2867–75, nov 2006.
- [223] C. B. Terwee, S. D. M. Bot, M. R. de Boer, D. van der Windt, D. L. Knol, J. Dekker, L. M. Bouter, and H. C. W. de Vet. Quality criteria were proposed for measurement properties of health status questionnaires. *J. Clin. Epidemiol.*, 60(1):34–42, 2007.
- [224] The World Health Organization. *Global status report on noncommunicable diseases 2010*. Geneva, 2010.
- [225] K. C. Thomas, B. Nosyk, C. G. Fisher, M. Dvorak, R. A. Patchell, W. F. Regine, A. Loblaw, N. Bansback, D. Guh, H. Sun, and A. Anis. Cost-effectiveness of surgery plus radiotherapy versus radiotherapy alone for metastatic epidural spinal cord compression. *Int. J. Radiat. Oncol. Biol. Phys.*, 66(4):1212–8, nov 2006.
- [226] A. C. Titman and L. D. Sharples. Model diagnostics for multi-state models. *Stat. Methods Med. Res.*, 19(6):621–51, dec 2010.
- [227] Y. Tokuhashi, Y. Ajiro, and M. Oshima. Algorithms and planning in metastatic spine tumors. *Orthop. Clin. North Am.*, 40(1):37–46, v–vi, jan 2009.
- [228] Y. Tokuhashi, H. Matsuzaki, H. Oda, M. Oshima, and J. Ryu. A revised scoring system for preoperative evaluation of metastatic spine tumor prognosis. *Spine (Phila. Pa. 1976)*., 30(19):2186–91, oct 2005.

- [229] Y. Tokuhashi, H. Matsuzaki, S. Toriyama, H. Kawano, and S. Ohsaka. Scoring system for the preoperative evaluation of metastatic spine tumor prognosis. *Spine (Phila. Pa. 1976)*., 15(11):1110–3, nov 1990.
- [230] Y. Tokuhashi, Y. Tokuhashi, H. Matsuzaki, H. Matsuzaki, H. Oda, H. Oda, M. Oshima, M. Oshima, J. Ryu, and J. Ryu. A revised scoring system for preoperative evaluation of metastatic spine tumor prognosis. *Spine (Phila. Pa. 1976)*., 30(19):2186–91, oct 2005.
- [231] Toluna Ltd. Esomar 26: 26 questions to help research buyers of online samples. Technical report, 2013.
- [232] K. Tomita, N. Kawahara, T. Kobayashi, A. Yoshida, H. Murakami, and T. Akamaru. Surgical strategy for spinal metastases. *Spine (Phila. Pa. 1976)*., 26(3):298–306, feb 2001.
- [233] D. Trafimow and M. Marks. Editorial. *Basic Appl. Soc. Psych.*, 37(1):1–2, 2015.
- [234] A. Tsuchiya and P. Dolan. The QALY model and individual preferences for health states and health profiles over time: a systematic review of the literature. *Med. Decis. Making*, 25(4):460–7, 2005.
- [235] J. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, PA, 1977.
- [236] H. Uei, Y. Tokuhashi, and M. Maseda. Treatment Outcome of Metastatic Spine Tumor in Lung Cancer Patients. *Spine (Phila. Pa. 1976)*., 42(24):E1446–E1451, 2017.
- [237] Y. M. van der Linden, S. P. D. S. Dijkstra, E. J. A. Vonk, C. A. M. Marijnen, J. W. H. Leer, and Dutch Bone Metastasis Study Group. Prediction of survival in patients with metastases in the spinal column: results based on a randomized trial of radiotherapy. *Cancer*, 103(2):320–8, jan 2005.
- [238] F. E. van Nooten, X. Koolman, and W. B. F. Brouwer. The influence of subjective life expectancy on health state valuations using a 10 year TTO. *Health Econ.*, 18(5):549–558, 2009.
- [239] F. E. van Nooten, X. Koolman, J. J. V. Busschbach, and W. B. F. Brouwer. Thirty down, only ten to go?! Awareness and influence of a 10-year time frame in TTO. *Qual. Life Res.*, 23(2):377–384, 2014.
- [240] F. E. van Nooten, N. J. A. van Exel, X. Koolman, and W. B. F. Brouwer. ”Married with children” the influence of significant others in TTO exercises. *Health Qual. Life Outcomes*, 13:94, 2015.
- [241] F. E. van Nooten, N. J. A. van Exel, X. Koolman, and W. B. F. Brouwer. ”Married with children” the influence of significant others in TTO exercises. *Health Qual. Life Outcomes*, 13:94, jul 2015.

- [242] A. L. Versteeg, A. Sahgal, L. D. Rhines, D. M. Sciubba, J. M. Schuster, M. H. Weber, P. P. Varga, S. Boriani, C. Bettgowda, M. G. Fehlings, M. J. Clarke, P. M. Arnold, Z. L. Gokaslan, and C. G. Fisher. Psychometric evaluation and adaptation of the Spine Oncology Study Group Outcomes Questionnaire to evaluate health-related quality of life in patients with spinal metastases. *Cancer*, 124(8):1828–1838, 2018.
- [243] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, Princeton, 1953.
- [244] X. Wan, L. Peng, and Y. Li. A Review and Comparison of Methods for Recreating Individual Patient Data from Published Kaplan-Meier Survival Curves for Economic Evaluations: A Simulation Study. *PLoS One*, 10(3):e0121353, mar 2015.
- [245] R. G. Watkins III and R. G. Watkins IV. *Surgical Approaches to the Spine*. Springer New York, New York, NY, 2015.
- [246] H.-L. Wee, S.-C. Li, F. Xie, X.-H. Zhang, N. Luo, D. Feeny, Y.-B. Cheung, D. Machin, K.-Y. Fong, and J. Thumboo. Validity, feasibility and acceptability of time trade-off and standard gamble assessments in health valuation studies: a study in a multiethnic Asian population in Singapore. *Value Health*, 11 Suppl 1:S3–10, 2008.
- [247] M. G. M. Weernink, C. G. M. Groothuis-Oudshoorn, M. J. IJzerman, and J. A. van Til. Valuing Treatments for Parkinson Disease Incorporating Process Utility: Performance of Best-Worst Scaling, Time Trade-Off, and Visual Analogue Scales. *Value Health*, 19(2):226–232, 2016.
- [248] M. C. Weinstein, J. E. Siegel, M. R. Gold, M. S. Kamlet, and L. B. Russell. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA*, 276(15):1253–8, oct 1996.
- [249] M. C. Weinstein, G. Torrance, and A. McGuire. QALYs: the basics. *Value Health*, 12 Suppl 1:S5–9, mar 2009.
- [250] N. J. Welton and A. E. Ades. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Med. Decis. Making*, 25(6):633–45, 2005.
- [251] C. Williams, J. D. Lewsey, D. F. Mackay, and A. H. Briggs. Estimation of Survival Probabilities for Use in Cost-effectiveness Analyses: A Comparison of a Multi-state Modeling Survival Analysis Approach with Partitioned Survival and Markov Decision-Analytic Modeling. *Med. Decis. Mak.*, 37(4):427–439, 2017.
- [252] S. E. Wolowacz, A. Briggs, V. Belozeroff, P. Clarke, L. Doward, R. Goeree, A. Lloyd, and R. Norman. Estimating Health-State Utility for Economic Models in Clinical Studies: An ISPOR Good Research Practices Task Force Report. *Value Heal.*, 19(6):704–719, 2016.

- [253] B. Woods, E. Sideris, S. Palmer, N. Latimer, and M. Soares. Partitioned survival analysis for decision modelling in health care: a critical review. Technical Report June, NICE Decision Support Unit, Sheffield, 2017.
- [254] G. Wu and L. Peng. Pharmacokinetics of propofol in patients undergoing total hip replacement : effect of acute hypervolemic hemodilution. *Anaesthetist*, 60(9):835–840, 2011.
- [255] R. F. Young, E. M. Post, and G. A. King. Treatment of spinal epidural metastases. Randomized prospective comparison of laminectomy and radiotherapy. *J. Neurosurg.*, 53(6):741–8, 1980.
- [256] N. Zethraeus and M. Johannesson. A comparison of patient and social tariff values derived from the time trade-off method. *Health Econ.*, 8(6):541–5, sep 1999.
- [257] L. Zhao, B. Claggett, L. Tian, H. Uno, M. A. Pfeffer, S. D. Solomon, L. Trippa, and L. J. Wei. On the restricted mean survival time curve in survival analysis. *Biometrics*, 72(1):215–221, 2016.