UCLA
UCLA Electronic Theses and Dissertations

Title
Investigating Predictive Disease Model Transportability through Cohort Simulation and Causal Analysis

Permalink
https://escholarship.org/uc/item/0jz147kc

Author
Singleton, Kyle Wilson

Publication Date
2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Investigating Predictive Disease Model Transportability through Cohort Simulation and Causal Analysis

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Biomedical Engineering

by

Kyle Wilson Singleton

2016

© Copyright by

Kyle Wilson Singleton

2016

ABSTRACT OF THE DISSERTATION

Investigating Predictive Disease Model Transportability through

Cohort Simulation and Causal Analysis

by

Kyle Wilson Singleton

Doctor of Philosophy in Biomedical Engineering

University of California, Los Angeles, 2016

Professor Alex Anh-Tuan Bui, Chair

A tenet of precision medicine is the ability to predict a clinical response or outcome for a given individual. The creation of predictive disease models as a part of this evidence-based process has increased in recent years as electronic medical data and machine learning techniques have grown in popularity. A number of methods are available for testing internal validity when developing models. However, the performance of developed disease models must also be tested in external populations. This practice of external validation, or transportability, can help determine the extent to which information in a predictive model can be applied generally across population samples. Model performance can suffer when applied in external settings because populations have inherent differences or data was collected with different practices. A number of impediments, particularly in reporting, continue to hinder the widespread application of transportability methods. This dissertation addresses these concerns using a methodology for improving the classification of models and evaluating a proposed technique for partially adjusting problematic models. A set of

simulations are proposed, taking advantage of a minimal set of published values to extend reported findings through bootstrap analysis. Interpretations derived from this method are able to guide the assignment and selection of models by transport levels. Causal transportability analysis, a previously proposed transportability theory in graphical models, is also evaluated for use in partial adjustment scenarios. The resulting process allows for evaluation of model transportability with minimal information and for assessing model adjustments. These investigations serve as useful tools for future transportability analysis. In addition, the results of this work introduce new items that can be added to reporting guidelines and support current trends to establish improved validation frameworks.

The dissertation of Kyle Wilson Singleton is approved.

William Hsu

Ricky Kitotaka Taira

Onyebuchi Aniweta Arah

Denise R. Aberle

Alex Anh-Tuan Bui, Committee Chair

University of California, Los Angeles

2016

Dedicated to my parents: John and Lois Singleton

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my utmost gratitude and respect for my advisor, Dr. Alex Bui, for his guidance and support during my graduate studies. I am thankful that he offered me the opportunity to join the Medical Imaging Informatics (MII) program and the UCLA community. It has been my honor to observe the growth of the MII program under his leadership. His initiative, intellect, and patience inspired the exploration of my informatics interests and helped me navigate the challenging waters of academic research. I am also indebted to Dr. William Hsu for taking the mantle of co-advisor to my work. His persistent drive in pursuit of high-quality research kept me focused when my frustrations mounted. I am thankful for his dedication to crafting exemplary written, visual, and oral presentations of research. This dissertation would not have been possible without the support of these exceptional individuals.

I would also like to thank the other members of my committee: Dr. Denise Aberle, for her tremendous enthusiasm in the medical profession and for infecting me and every student with a desire to work across disciplines. Dr. Ricky Taira, for his unique ability to ask difficult and insightful questions that lead me down new avenues of thought. Dr. Onyebuchi Arah, for generously providing his expertise and pointing me to amazing resources in epidemiology that shaped my view of validation research. This dissertation was pushed to new heights by their guidance, feedback, and expertise.

Thank you to the faculty and staff of MII for providing an environment where researchers can flourish. I would like to thank Professors Corey Arnold, Frank Meng, Jim Sayre, Craig Morioka, and Suzie El-Saden, for providing me with mentorship in the diverse field of informatics. I am especially grateful to Dr. El-Saden for contributing her time and expertise to the evaluation of brain cancer imaging features used in this work. I would like to thank Isabel Rippy for making me

feel welcome the moment I joined the program and for handling the requests of MII graduates with skill and grace. Thank you to Audrey, Shawn, Weixia, Hamid, Noah, Brian, Patrick, Denise, and Carlos for helping with administrative issues and providing entertaining lunchtime discussions. I would like to thank Lew Andrada for being a fount of humor and for our many lunches, book sales, 5K races, and movie viewings. I would like to thank Dr. Bill Speier for aiding me in seemingly limitless capacities (roommate, spotter, colleague, editor), providing support and council that enabled me to power through and achieve results. I would like to thank my fellow students who graduate this year and served as founding members of our candidate's club – Mary, Maurine, Anna, Jean – for their support, insight, and ability to endure my long winded attempts at advice. Thank you to all the other graduate students of MII – Johnny, Shiwen, Nova, Nick, Edgar, Panayiotis, Karthik, Simon, Jiayun, and Tianran – for creating a motivating intellectual environment that inspired me to keep moving forward.

I am forever grateful to my parents, John and Lois Singleton, for supporting all my intellectual endeavors since that first story about a pig in West Virginia brought an IBM computer to our home. Thank you to my brother, Drew, for being there when I needed him and reminding me what patience and endurance look like. I would not be where I am today without their love, support, and belief in my abilities in the face of my own doubts.

Chapter 3 contains material adapted for publication with Nova Smedley, Edgar A. Rios Piedra, Suzie El-Saden, William Hsu, and Alex A.T. Bui submitted with the title "Challenges of Applying and Replicating Previously Reported Predictive Models."

Chapter 5 contains material adapted for publication with Alex A.T. Bui and William Hsu submitted with the title “Categorizing the Transportability of Published Predictive Models using Simulated Cohort Analysis.”

Chapter 6 was based on material published with William Speier, Alex A.T. Bui, and William Hsu in the 2014 American Medical Informatics Association Annual Symposium Proceedings; 2014: pages 1930-1939.

Research reported in this dissertation was supported in part by the National Library of Medicine of the National Institutes of Health under award number T15LM007356 and by the National Cancer Institute of the National Institutes of Health under award number R01CA157553. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

VITA

EDUCATION

2006	B.S., Biomedical Engineering; Minor, Computer Science University of Virginia Charlottesville, VA

PROFESSIONAL EXPERIENCE

2012 - 2016	*Graduate Student Researcher* UCLA Medical Imaging Informatics Los Angeles, CA
2008 - 2012	*National Library of Medicine Fellow* UCLA NLM Medical Imaging Informatics Training Program Los Angeles, CA
2006 – 2008	*Post-baccalaureate Fellow* National Institutes of Health Section on Stroke Diagnostics and Therapeutics National Institute of Neurological Disorders and Stroke Bethesda, MD
2005 - 2006	*Undergraduate Researcher* Computational Systems Biology Lab Charlottesville, VA

PUBLICATIONS AND PRESENTATIONS

Singleton KW, Speier W, Bui AA, Hsu W. Motivating the Additional Use of External Validity: Examining Transportability in a Model of Glioblastoma Multiforme. AMIA Annu Symp Proc. 2014; 2014: 1930-1939.

Singleton KW, Bui AAT, Hsu W. Transfer and transport: incorporating causal methods for improving predictive models. J Am Med Inform Assoc 2014; amiajnl–2014–002968. doi:10.1136/amiajnl-2014-002968

Singleton KW, Hsu W, Bui AAT. Comparing Predictive Models of Glioblastoma Multiforme Built Using Multi-Institutional and Local Data Sources. AMIA Annu Symp Proc 2012; 2012: 1385–1392.

Singleton KW, Lan M, Arnold C, Vahidi M, Arangua L, Gelberg L, Bui AAT. Wireless Data Collection of Self-administered Surveys using Tablet Computers. Annual National Library of Medicine Informatics Fellowship Conference, Washington DC, June 2011.

Singleton KW, Lan M, Arnold C, Vahidi M, Arangua L, Gelberg L, Bui AAT. Wireless Data Collection of Self-administered Surveys using Tablet Computers. AMIA Annu Symp Proc. 2011; 2011:1261-9.

Singleton KW, Garcia-Gathright JI, Burns B, Rocks K, Iglesias JE, Bui AAT, Aberle D. Semi-automated Medical Text and Image Selection for Multimedia Presentation at Tumor Board Reviews. 2009 Radiological Society of North America Conference, Education Exhibit, Invited December 2009, Chicago, Illinois.

CHAPTER 1

INTRODUCTION

A primary goal of the medical research community is to assist with and improve decision making in the clinical setting. Clinicians spend a great deal of time applying their personal experience while also considering the wide body of medical evidence and literature. Their medical knowledge must be used to assess symptoms, choose relevant medical tests and treatments, and reason across a vast space of evidence to obtain the best possible outcome for patients. Researchers strive to develop studies and aggregate medical findings into systems that can help the medical community understand and describe disease processes and the behavior of the human body. Through statistical analysis, findings from research update and fill in gaps of knowledge used to train and inform clinicians.

Models are important tools used by researchers to define a set of assumptions regarding the structure of relationships between features and an outcome. The working knowledge of a disease is represented in a disease model to provide a mechanism for statistically analyzing and predicting patient disease states and outcomes for a disease of interest. Clinicians must consider all the variability of each specific case they encounter and attempt to accurately recall past cases and medical knowledge from their realm of experience. Yet, the growing complexity of our understanding of disease, the continuous publication of research results, and the growing number of patients in need of evaluation threaten to overwhelm physicians struggling to track information on a patient by patient basis. Models can capture and summarize this information in a compact form useful in future decision making. In this way, predictive models supplement clinicians' knowledge and reduce the cognitive load placed on healthcare practitioners.

The development of disease models, however, is also a complex task. A great deal of research has been done to determine best practices for selecting features, designing models, and analyzing model accuracy. Models contain evidence of the contribution of features to given treatment and patient outcomes, but their interpretations are often tied closely to study designs and/or patient cohorts. Models are useful to researchers and clinicians when they can be used broadly for prediction of the patient population outside the original cohort. Testing models for use in other cohorts is a growing research need, particularly so that models can be trusted for decision making. In this dissertation, I explore issues restricting model reuse and evaluate a method for improving assessments of model transportability (also commonly called external validity and generalizability). Assessing transportability is important to determining the settings outside of the original model design where predictions are valid.

1.1 Background and Motivation

Medical evidence is obtained using a number of research assessments. The primary scientific method for studying disease features and treatments are randomized controlled trials (RCTs). In RCTs, researchers establish a particular question (query) of interest about the disease being studied. These questions are tied to a specific experimental goal that tests a hypothesis. Examples include attempting to find a link between disease elements (e.g., "Is there evidence of a link between EGFR expression and cancerous cell growth?"), finding evidence that a treatment helps patients with disease (e.g., "Does treatment with temozolomide reduce tumor size in patients?"), or if testing methods in the hospital are effective for diagnosis (e.g., "Does the use of MRI information improve cancer detection?"). Randomization techniques separate patients into treatment and control study arms in order to manage experimental conditions. In this way, RCTs

seek to find causal connections between elements of disease or between a treatment and effect on a disease by comparing the results of separate testing environments.

As RCTs provide the scientific means to understand disease, their use is integral to evidence-based medicine. However, RCTs are not without their faults. To reach a consensus about a hypothesis, multiple trials should be conducted. Often these trials will arrive at different findings or be difficult to compare because of methodological differences. For this reason, the process of systematic review and meta-analysis has been made paramount to demonstrating the validity of findings across the large body of medical research. Despite the rigors employed to ensure statistical accuracy and understanding, the derived knowledge can only represent a level of certainty in regards to the populations examined during experimentation. It is often unclear how much the original findings were biased by study design or the query of interest. These drawbacks can often impede the determination of how applicable findings are to future patients.

While RCTs and meta-analysis help to realize the overall domain of disease processes and treatments in controlled environments, the medical community must apply this knowledge to the uncontrolled environment of clinical practice. Physicians rely on observational data collected from patients to make informed decisions on treatment and survival. In the same way, researchers utilize the combined findings from experimental studies and collected observational data to create predictive disease models. These models are meant to aid in clinical decision making, but are also highly varied in design, much like RCTs and meta-analysis. There remains a substantial difficulty in applying the predictions of models to new cohorts of patients not involved in the original model design. This difficulty in generalizing models has relegated most disease models to research practice and made them impractical for application in a clinical setting.

The growing success of disease modeling for prediction tasks has encouraged the consideration of their use for every-day clinical decision making. This trend has been supported by the upswing of big data, machine learning, and innovative statistical analysis methods. Currently, models are heavily scrutinized for internal validity of predictions by assessing accuracy with repeated tests of held out sets of data. But, additional validation is required to determine if models apply, or transport, to external settings. Subsequently, tests for transportability (external validity) are a recent and growing topic of interest. Existing validation tests follow a paradigm similar to internal validation by assessing accuracy changes. Most current transportability methods interpret cohort differences qualitatively while considering the final results of discrimination and calibration. Often, this interpretation designates a specific type of difference (e.g., temporal or geographic) or reclassifies the analysis into a different goal (e.g., reproducibility) to explain validation success or failure. However, the degree of differences between a source (original) and a target (new) cohort can be caused by a combination of factors that are still difficult to evaluate. Experts and researchers with knowledge of the data collection process can reveal some contributors to source/target difference, but an inability to observe all factors and confounding influences may mean many differences will remain obscured.

Problematic differences are adjusted for new target locations when possible. These approaches are common because new data is often scarce or difficult to gather in a practical time frame. In addition, general approaches are applied as it remains difficult to understand model performance in the context of (typically smaller) cohorts of external data. As such, updates are often performed using broad assumptions about the applicability of predictions in the population or often by simply generating new model parameters *de novo*. To improve the updating process, it would be useful to assign models to particular transportability states based on the success of validation tests. An

example of this assignment is shown in Figure 1.1 where five theoretical models are assigned to four transportability states. Other proposed external validation frameworks and updating techniques can be mapped to these four classes, helping define what types of techniques are applicable to specific transportability settings. In this way, researchers and clinicians can use more specific transportability assignments in place of a binary usable/unusable classification and potentially target the next steps of analysis and adjustment more easily across many models.

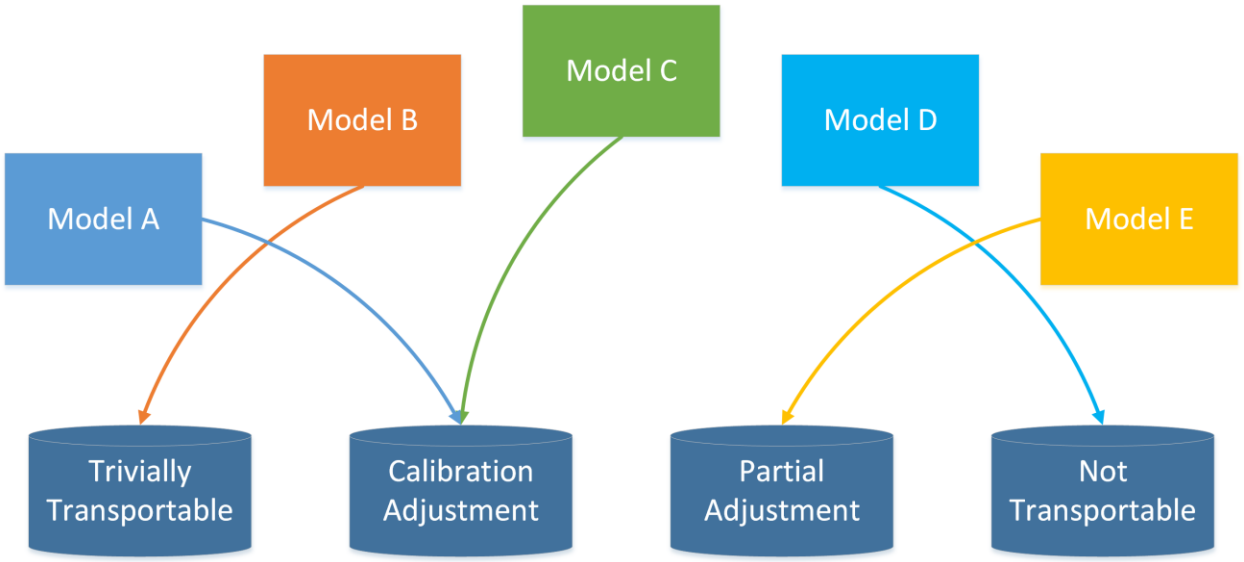


Figure 1.1 Proposed transportability classification levels for more appropriately dividing models considered for future use.

The results of external validity assessments continue to be difficult to interpret. For example, discrimination values such as the c-statistic might show only small variations of one or two percent for a source model applied to multiple target cohorts. Thus, determining if the model is trivially transportable for one target and partial adjustable for another is not a straightforward task. New approaches are needed to provide more detailed tests in order to better determine the ability of models to generalize in one or more cohorts. In addition, these approaches need to be capable of working with limited amounts of source information, as not all values are publicly reported or

available during transportability analysis. Assignment of models to transportability states may be useful in future analysis to help expose classes of source/target difference that could be addressed with new methods. For example, models assigned to the partial transportability state could be candidates for evaluation with a novel method like causal transportability for picking features for re-estimation.

1.2 CONTRIBUTIONS

A great deal of effort is put into the development of models and important training information is supplied in the original effort. Researchers should attempt to take full advantage of the original and complex model building process and reuse information from source environments whenever possible. Transportability of a model is considered by testing the discriminative performance and calibration of models in new settings. Current techniques classify models into binary categories and often require direct access to complete datasets. In addition, ineffective models do not receive additional analysis for potential adjustments to correct issues. In this dissertation I attempt to address these issues through the following two contributions:

1. Improvement of transportability assessments and classifications using a methodology to retrospectively evaluate discrimination variability using a minimal set of reported information to perform simulation.
2. A process for investigating partial adjustments in models with significant performance deficits using causal graphical techniques to extract potential targets in need of updating.

The first contribution takes advantage of previously tested validations and proposed frameworks as a guide in the definition of new transportability classifications. An examination of current

reporting practice in modeling publications and summarization of common issues restricting transportability assessment was reviewed as a first step. Assessment and classification are difficult due in part to the inability to access previous information about the original model. Therefore, a methodology for estimating discrimination performance using a minimal set of published information is proposed and tested. The resulting methodology improves the ability of researchers to define transportability levels for models. A set of examples are used to demonstrate how updated assessments and subsequent model designations can help define additional steps required for model adjustment after initial validation.

The second contribution provides a potential approach to updating models classified with the transportability level of partial adjustment. Few techniques currently exist to define what parts of a model are unaffected during external use. The approach explored in this contribution takes advantage of a novel model description process with causal information to consider what subcomponents of models are transportable between two settings. A set of steps from this method are examined and a process for applying the technique to an example modeling scenario is evaluated. This process takes advantage of additional causal assumptions to make decisions concerning what subsets of model features should be partially adjusted, providing a means for updating problematic models and correcting predictions. Updated models can then be further validated and more frequently used in new settings without requiring full model re-estimations or extensions.

1.3 ORGANIZATION OF THE DISSERTATION

This dissertation is separated into seven chapters, including the introduction and conclusion.

Chapter 2 provides a brief foundation covering previous work related to the areas of evidence based medicine, model validation and updating, and causal transportability analysis. Each area plays a role in the development of robust models and assessing their generalization.

Chapter 3 describes a review of four published models for brain cancer prediction. These models were used to gauge current publication practices and determine the availability of values in reports necessary to perform transportability testing.

Chapter 4 introduces an evaluation of methods for simulating patient data to replace source information that is frequently unavailable as part of the publication process. This evaluation is performed using a large cohort of lung cancer cases collected in the National Lung Cancer Screening Trial (NLST).

Chapter 5 applies the resulting simulation methods to evaluate four published brain cancer models, proposing a novel evaluation method that estimates the variability of an original model's performance based on the size of the external cohort of interest. The proposed evaluation can be generalized to other disease model validations.

Chapter 6 applies causal graphical models as an introduction to novel approaches to partial model adjustment. This process can be used to further expand models classified using the proposed evaluation of Chapter 5.

Chapter 7 summarizes the results of previous chapters, provides suggestions for adjusting current publication and validation processes, and discusses future directions to push transportability evaluation further and approach broader clinical utility for disease models.

CHAPTER 2

BACKGROUND

Many research areas have contributed to the current state of predictive disease modeling including evidence-based medicine, randomized controlled trials, statistical modeling, and validation testing. More recently machine learning and artificial intelligence (e.g., deep learning) have also become popular areas for improving model design and performance. This chapter reviews previous research relevant to medical research from a number of these areas and reviews the current practices for evaluating and adjusting disease models. The first section covers the paradigm of evidence-based medicine research, which has downstream influence on predictive modeling design. The second section reviews current practices for validating models, the metrics used to evaluate model accuracy, and established techniques for updating models. The third section provides a foundational background on graphical and causal models, important tests of their properties, and a causal theory of transportability that holds potential for applying models more broadly when only certain parts of a model are transportable.

2.1 EVIDENCE-BASED MEDICINE

Prognostic modeling research is largely driven by evidence-based medicine (EBM) tasks [1]: randomized controlled trials (RCTs), subsequent systematic reviews of RCTs, and meta-analysis derived from completed systematic reviews play a critical role in establishing accepted clinical practice. Physicians update their understanding of disease, as well as new and existing treatment options, by reviewing these studies. Subsequently, these research efforts in controlled scientific experimentation test important etiological findings and establish focal points for decision making. In essence, clinicians construct their own internal view of relationships between patients and

disease processes from evidence. Predictive disease models derive values for variables and parameters from the same space while also frequently leveraging clinicians' expertise and observational data. The goal of predictive models is then to contribute additional evidence by taking advantage of statistical analysis and computing power to make risk evaluations and create accurate decision-making tools.

2.1.1 Randomized Controlled Clinical Trials

A randomized controlled trial is a scientific study design that takes advantage of the randomization of cases into control and intervention groups in order to reduce bias when testing a hypothesis of an experimental outcome. Different combinations of randomization and case control are applied for general scientific use. The medical community is well-versed in using specific trial designs to perform medical research. For example, clinical trials are most commonly known for their use in testing drug and device effectiveness. The goal of such trials is to evaluate the etiology or causal factors that contribute to a disease or changes in disease (such as after treatment). A wide variety of clinical trial designs are employed each year to generate new evidence and improve medical outcomes and treatments [2].

A randomized controlled clinical trial (RCCT) takes advantage of both randomization and controlled treatments. Two or more study arms are defined with one arm designated as a control that receives no form of treatment/intervention. Enrolled subjects are then randomized to a study arm and their results can be compared against subjects assigned to other intervention or control groups. Randomization helps reduce selection bias, a type of systematic error where subjects are not objectively represented. Selection bias can mask the true effects of the treatment/process under evaluation. For example, if a non-random subject assignment is applied in a trial, patients given a drug therapy might be biased towards certain age ranges, causing an imbalance between treatment

and control groups. The resulting outcome might show the drug treatment as highly effective, but it would be unclear if the effect is a product of the drug itself or the age of the subjects in each group. Control groups are intended to provide an effective null hypothesis for comparison in statistical analysis, but in some cases, control groups may be infeasible or unethical. Therefore, the controlled aspect of RCCT design is not required or universal. In fact, clinical trial designs included complex options such as single and double blinding, paired and crossover assignment, and varying randomization algorithms [3]. A full discussion is beyond the scope of this work, but it is important to understand the fundamental RCCT trial design that drives evidence generation.

Clinical trials are such a ubiquitous part of evidence-based medicine that a specialized guide known as the CONSORT statement has been developed by the clinical research community to guide reporting of results [4]. Despite attempts to design RCCTs to a set of standards, it is often difficult to compare results across trials due to the complexity of their designs, even when two studies examine the same drug's effect on a given condition. While each trial provides results that are validated to their given cohort, it is rarely possible to apply the results externally. Confounding often exists due to differences in sampling size, data collection constraints, the number of patients lost to follow-up, etc. These studies validate the efficacy of an intervention under ideal conditions but do not necessarily address the clinical effectiveness across a real-world population (i.e., the external validity/generalizability of the intervention to routine practice) [5].

While more "practical" or "pragmatic" clinical trials are now promoted [6,7] to relax subject eligibility requirements (thereby broadening the test population), these changes do not definitively overcome the fact that a given investigation encompasses important assumptions about the underlying study group and environment. As such, the majority of current RCCTs could be considered a unique population of patients with constraints that can only be interpreted internally

until evaluated further to determine generalizability. To combat this issue, researchers attempt to evaluate RCCTs and other clinical trials with meta-analysis. Some literature has suggested specific aspects of trials that should be considered for comparison during generalization [5,8].

2.1.2 Meta-analysis

Meta-analyses generate a further statistical evaluation of the reported validity of a set of clinical trials, linking results through a systematic review process. Efforts such as the Cochrane Collaboration recognized the need for maintaining unbiased systematic review and meta-analysis for all of medical research [9]. Meta-analysis publications are now regularly seen from Cochrane and other groups striving to understand the full ramifications of clinical trial results. Providing a grouped evaluation of results can establish which trial findings generalize more widely across the medical community.

Meta-analysis can be seen as aggregating the statistical findings of a group of RCCTs into the semblance of a single, larger study. Yet this analysis is not a combination of all the raw data from the individual populations of each trial, but an examination of the outcome ratios calculated in the results of each trial by performing a weighted average. This revised statistical evaluation of the individual findings is used to make claims about the consensus of medical research on the given topic examined. The conflicting information from individual trials often leads to inconclusive findings concerning a treatment. When consensus cannot be reached, continued research becomes the suggested result.

Significant findings from a meta-analysis are suggestive of strong associations of experimental findings. However, these conclusions are still related to the original trial environments which were internally validated. When the designs and populations across a set of clinical trials are broad,

there is greater opportunity for the contributions from each differing population to support generalization of the meta-analysis findings to a broad set of patients or diseases. But, this same variation can also make it more difficult or impossible to complete the meta-analysis at all. Thus, it remains difficult to apply the knowledge to future cases when the examined populations of each trial in the meta-analysis are varied to enlarge the tested population space. The consensus of research indicates causal connections between variables in the disease studied, but not all circumstances of the many RCCT-defined environments will hold in real-world application for treatment or prediction. Establishing external validity has remained a difficult task despite the growing attempts to provide meta-analysis of RCCTs. Direct application of past studies' results to future analysis and prediction is a continuing goal of many researchers in the meta-analysis area and also extends into the domain of predictive modeling.

2.1.3 Predictive Disease Models

Statistical models provide a means for taking evaluations from many sources (e.g., RCCTs, meta-analysis, and observational data) and making use of them as predictive values for individual cases. In medical research, such models are commonly referred to as disease models, as the outcomes and variables of interest are related directly to disease characteristics. The terms *predictive* and *prognostic* are commonly applied to these models, often depending on their intended goals of predicting a patient state or determining long term prognostic risks. Other titles such as prediction rules, risk scores, and risk calculators are also common [10]. In this dissertation, statistical models will largely be referred to as *predictive disease models* or *predictive models*.

In general, models are built to calculate the probability of an outcome/response variable (dependent variable) as related to a set of candidate predictors (independent variables). A large variety of modeling options are available and model choice is guided, in part, by the outcome of

interest and the predictions targeted. Researchers work diligently to combine disparate sources of knowledge into candidate predictors for the modeling task. Deciding what features to consider in modeling is not straightforward; RCCT findings and expert clinician input provide starting points when choosing potential variables for modeling. Early in the design process, multivariate models start off by considering as many features as possible that have been found in previous research or have a reasonable chance of confounding other features of interest. However, the use of large groups of features can be detrimental if used incorrectly. First, the included features for a disease model must be items that are commonly collected and easily measured in the clinical environment. If clinicians have no access to the features used in a model, it becomes useless to their daily practice. Second, unless the available cohort of cases is orders of magnitude larger than the number of features considered, parameter estimates can become biased. These models have become overfit, as they have been trained to match particular patterns of the original data and not the underlying relationships of the disease in the population as a whole. Feature selection, or pruning, is often used to combat some of these issues by reducing the considered feature space to those that are most useful to prediction.

Following feature selection and model estimation, a model must be evaluated for prediction accuracy on the cohort of interest and should also be tested in outside datasets for reliability. As the main focus of this dissertation is assessments of model validity, different validation stages are considered in more depth in section 2.2. Many other factors, such as variable discretization, sample size, and missing data play important roles in model design and performance, but a full discussion on the breadth of model development is beyond the scope of this dissertation. Chapter 3 includes discussion on why some of these decisions can restrict further validation and replication efforts,

but an assumption is made in this work that sufficient efforts were made by the original model designers to provide robust model training and validation analysis.

Multiple regression models are still dominant in clinical research: including linear regression for continuous outcomes, logistic regression for binary outcomes, and Cox proportional hazard regression for time based measures of outcome (such as time to progression or death). Much of this dissertation works with published models and examples related to logistic and Cox regression. Brief descriptions of these models are provided below with focus given to elements relevant to testing validation in later chapters.

Binary Logistic Regression

In binary logistic regression, the probability of a binary outcome, Y, is estimated as related to a set of candidate predictor variables, $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$. Intercept and model coefficients (parameters) for each of the predictor variables, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ are then estimated using available training data for $\boldsymbol{X}$ and Y. The intercept, β_0, is not always estimated for other model designs (e.g., Cox regression). For each subject, the predictor variables, $\boldsymbol{X}$, are assumed to be known constants and a fixed value of $X_0 = 1$ is added to the vector to account for the intercept value, β_0. By multiplying against the model coefficients, $\boldsymbol{\beta}$, a weighted sum of the contribution of a set of p predictors is generated, commonly called the linear predictor, LP:

$$LP = \boldsymbol{X\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

For the binary logistic regression model, the probability of outcome given predictors, $P(Y|\boldsymbol{X})$ can be calculated based on the linear predictor with the following equation described by Harrell [11]:

$$P(Y|\boldsymbol{X}) = Prob\{Y = 1|\boldsymbol{X}\} = \frac{1}{(1 + e^{-LP})}$$

To test model accuracy, linear predictor values or predicted probabilities can be assessed against the known outcomes of subjects using different thresholds. This process is discussed further in section 2.2.1 covering internal validation. For clinical use, the predicted probability of a newly observed subject provides an indicator of the patients predicted future status.

Cox Proportional Hazards Regression

The proportional hazards model provides the ability to estimate the effects of covariates on subject outcomes over varying amounts of time. The proportional hazards assumption states that the hazards of predictors do not vary over time between subjects which allows for estimation of the regression coefficients, $\boldsymbol{\beta}$. If the proportional hazards assumption does not hold, a different form of hazards analysis using time-varying coefficients or fully parametric models such as the Weibull model should be used [11]. The linear combination of covariates (*LP*) follows the same form as logistic regression but the intercept value is replaced by baseline hazard over time, $h_0(t)$, for all patients. Proportional hazards models are useful for modeling time to event outcomes (e.g., progression or survival) and are able to handle subject censorship as many subjects may not be able to be followed for the entire time period of interest. The Cox model is a semi-parametric form of the proportional hazards model and is most frequently used because it does not require direct estimation or assumption of the baseline hazard function, $h_0(t)$. The Cox proportional hazards model has the following form when stated with the baseline hazard for a given time, *t*:

$$h(t|\boldsymbol{X}) = h_0(t)e^{(LP)}$$

The linear predictor, *LP*, is computed based on feature coefficients and feature values for a set of predictors, *p*, without terms for an intercept value for this model design:

$$LP = \boldsymbol{X\beta} = \beta_1 X_1 + \cdots + \beta_p X_p$$

During model training, regression coefficients necessary for computing the linear predictor are estimated from training data based upon a conditional form of the log partial likelihood. Cox's derivation of this conditional estimator of $\boldsymbol{\beta}$ allows for the baseline hazard function to drop out. This process considers the relative risk between an individual's survival time compared with the range of unique ordered failure times represented by $t_1 < t_2 < \cdots < t_n$, for a set of *n* cases. The set of all individuals with observed survival time, Y_j, greater than or equal to a given time, t_i, comprise the set, R_i, of all individuals at risk at that time. Thus, the estimator for Cox regression takes the following form, as described by Harrell, where the baseline hazard cancels [11]:

$$\frac{h_0(t_i)e^{(X_i\boldsymbol{\beta})}}{\sum_{j\in R_i} h_0(t_i)e^{(X_j\boldsymbol{\beta})}} = \frac{e^{(X_i\boldsymbol{\beta})}}{\sum_{j\in R_i} e^{(X_j\boldsymbol{\beta})}} = \frac{e^{(X_i\boldsymbol{\beta})}}{\sum_{Y_j\geq t_i} e^{(X_j\boldsymbol{\beta})}}$$

This property of the conditional probability allows for estimation of the log partial likelihood including cases that are censored:

$$\log L(\beta) = \sum_{Y_i \ uncensored} \left\{ X_i\boldsymbol{\beta} - log\left(\sum_{Y_j\geq Y_i} e^{(X_j\boldsymbol{\beta})} \right) \right\}$$

While the baseline hazard function cancels out during the model estimation, knowledge of the hazard is necessary for computing survival probabilities. Modern software packages therefore estimate a cumulative baseline hazard function, $\widehat{H}_0(t)$, from training data that is also non-

parametric with respect to the form of the baseline hazard. Applying this cumulative hazard with the estimated coefficients, it is possible to estimate a survival probability at a given time, *t*:

$$S(t|\boldsymbol{X}) = e^{-\hat{H}_0(t)e^{(LP)}}$$

As with logistic regression, the linear predictor and survival probability can be used in model assessment and the predicted survival probability can be used as an estimate to assess the future risk a current patient has given recent feature observations. [11,12]

2.2 Model Validation and Updating

Internal and external validity are properties inherent in the results of both scientific experimentation and predictive models. For example, in clinical trial design a researcher can determine if true statistical differences exist between groups by controlling for important variables and using proper randomization. Claims can be reached concerning the causal connections between variables involved in disease processes for the patients and selection criteria involved in a study. This level of claim falls within the confines of internal validity and physicians can reasonably apply these findings if their patient matches characteristics from the study [5,8,13,14]. Direct reproductions of trials could further support results, but monetary and time costs of such trials make such direct repetition a rare occurrence. Instead, similar trials with variations in populations, protocols, and locations are usually funded to provide further extrapolation of statistical claims to general use. In this way, researchers and physicians are convinced of the external validity, or generalizability, of results across the entire population due to the growing amount of experimental evidence in trials with closely related designs.

Within predictive models, obtaining discrimination between groups of cases is the primary goal. Unlike a clinical trial where subjects are grouped for analysis, a model tries to predict groupings and all the data is mixed for model training. In using all the data to maximize predictive power, the model may be biased and demonstrate overoptimistic performance on the specific set of cases observed (especially if data is used in both training and testing steps). Therefore, internal validation techniques split cohorts and perform repeated testing to check for bias in the model design and training. In situations where overfitting has been avoided, models are considered internally valid for cases drawn from the same cohort, commonly referred to as the source cohort. However, internal results do not guarantee good performance on new cohorts of patient cases, commonly known as target cohorts.

In fact, when applying models in new situations, model accuracy typically suffers because target cohorts have a key systematic difference or because the variability of the true population has not been completely modelled by the source data. Unlike the significant body of work for evaluating internal validity, few techniques have been developed to test external validity. Methods used in assessing internal validity need further study to determine if they can be directly applied to understanding how findings generalize. Epidemiological research has demonstrated the growing exploration of external validity [15–17]. Such research is encouraging a move to focus on testing all data findings with methods relevant to both internal and external validity so that evidence and study conclusions are fully contextualized and applied appropriately. In addition, researchers can begin to study how external performance decreases can identify steps to update models for continued use, rather than having to rebuild and retrain models from scratch.

Validation is therefore one of the most important and also difficult tasks involved in modeling. In this section, I further describe the two major stages of model validation mentioned above: 1)

internal validation, where performance and bias of the model are assessed with data from the original model training environment (the source) and 2) external validation (also referred to as generalizability and transportability), where the model is applied for prediction on new data from cohorts not tied to the original selected population (the target). For each validation stage, I review previously used performance metrics and variability assessments. In addition, I review current suggestions for model updating.

2.2.1 Internal Validity

In order to obtain the best model, the number of cases supplied during training should be as large as possible. Ideally, that means all available source data should be used to determine the selected features and learn the regression coefficients. However, if all data is used in this process, there is no independent data available for testing the accuracy of predictions. While a model can technically be tested on the same data used in training, the results will usually be overoptimistic. Using 100% of a subject cohort for both training and testing is called apparent validation [12]. While some studies report this result, other methods are recommended for splitting or subsampling the available cases to obtain more reasonable assessments of model performance.

Evaluating bias and variability

The simplest internal validation method creates a single split-sample from the source. The source cases are randomly divided into training and testing samples used for model development and model validation respectively. Common split-samples sizes are fifty:fifty or two-thirds:one-third. While simple in practice, there are many drawbacks to this method. First, the validation process may be unstable both in training and validation if the original source cohort is small. For example, there is an increased chance trained coefficients may be unreliable because split sampling can remove important cases from the training dataset. Similarly, if the split is chosen to favor larger

training size, the test set may not be large enough for a robust evaluation. Second, a researcher may be unlucky in the random split and the two samples could end up highly imbalanced in relation to certain features or outcomes, greatly influencing prediction accuracy. Simulation studies demonstrate that large cohorts help mitigate some of these issues and split-sample validation can be used when thousands of subjects are available [18]. Increases in computing power have made it possible to perform other forms of sampling instead of relying on split-sample validation. [11,12,19,20]

Cross-validation is the successor to split-sample validation, extending the technique by completing repeated splits of the data. Each data split is referred to as a fold and by repeating folds across the cohort all subjects contribute to training and testing of the model. For example, in a five-fold cross-validation, the data is divided in five groups and 1/5 of the data is used for testing and the remaining data for training. The test and training sets are changed over five repetitions and prediction results are aggregated. When completed, each case will have served once as a test case. Common cross-validation splits are five-fold, ten-fold, and N-fold cross-validation. The extreme case uses all the cases except one for training and tests the singled test example (from 1…N); this validation is also known as leave-one-out cross-validation (LOOCV) or jack-knife validation. Repeating the analysis across the full cohort in this way provides more stability in the evaluation, but there are tradeoffs in the extent of bias introduced depending on the chosen fold size. Extra stability against the bias of random sample splits can be gained by further repeating the entire cross-validation process M times using different random splits on each iteration. For example, the standard cross-validation procedure could be repeated 50 or 100 times to evaluate the variability caused choosing many different random folds. There is still some debate concerning the need to perform feature selection prior to validation or as a repeated step in the cross-validation process. [11,12,19,20]

Bootstrap validation imitates the process of drawing from an underlying population in order to assess model bias. A bootstrap sample is created by sampling with replacement from the cohort of interest. These resulting samples are the same size as the original data. Each bootstrap is then used as a training source for the model and evaluation is performed on both the bootstrapped sample and the original cohort. The difference between each sample's performance provides an estimate of the model optimism and the final optimism value is subtracted from the apparent validation performance of the original data. Previous research notes that 100-500 bootstraps are usually sufficient for achieving accurate estimates [12]. Bootstrapping has stability advantages because training and testing are always performed at the largest sample size based on the source cohort. In addition, bootstrapping has been shown to be more useful than cross-validation when performing feature selection during the validation process. Computational advances have driven more widespread use of the bootstrapping approach that were not previously possible. However, in some cases discretization and feature selection decisions can make it difficult to take advantage of the bootstrap procedure. Future research will continue to elucidate the impact of these decisions and which form of cross-validation or bootstrapping is most appropriate for successful analysis. In general, the validation procedures above can be applied for testing any metric of interest to a researcher. [11,12,20]

Performance metrics

Discrimination measures test the ability of a model to differentiate between cases with and without the outcome of interest. Cases are classified above and below a chosen decision threshold, establishing a set of correct classifications, true-positives (TP) and true-negatives (TN), and incorrect classifications, false-positives (FP) and false-negatives (FN). The total number of cases with the event of interest is equal the sum of the TP and FN groups, (TP+FN), and the cases without

the event are equal to the sum of the TN and FP groups, (TN+FP). Two important ratios used to consider discrimination accuracy using these counts are sensitivity and specificity. Sensitivity is the ratio of true-positive classifications to the total number of cases with an event, (TP/(TP+FN)), and specificity is the ratio of true-negative classifications to the total number of cases without the event, (TN/(TN+FP)). A receiver operating characteristic (ROC) curve can be created by comparing the performance of sensitivity against 1-specificity across a set of discrimination thresholds for the predicted probability of cases (based on $P(Y|X)$ or LP). ROC curves are commonly plotted against a non-informative model with a random chance of correct prediction. The area under the ROC curve (AUC or AUROC) is derived from the ROC plot and is one of the most common summary measures of discriminative ability. Models with good discrimination between event classes are expected to have AUC values greater than random chance, AUC = 0.5. Models, with perfect discrimination would achieve an AUC = 1.

For binary decision tasks (common for disease models using logistic and Cox regression) the value of the AUC is equivalent to another measure, the concordance statistic (c-statistic). The c-statistic is computed by evaluating all pairs of subjects with and without the outcome being modeled. The final value is calculated by comparing the proportion of pairs where a case with the outcome (i.e., Y=1) has a higher predicted probability (based on $P(Y|X)$ or LP) than the paired subject that did not experience the outcome. Therefore, the c-statistic indicates the proportion of times the model will make a correct decision by assigning an appropriate probability to subjects with events. Interpretation of the c-statistic is the same as the AUC where the c-statistic value is evaluated on a range from 0.5 to 1 (random predictions to perfect predictions). Higher scores indicate that the model can discriminate the outcome classes with higher accuracy. The c-statistic is related to another rank correlation measure, Somers' D_{xy}, through the following identity [11]:

$$D_{xy} = 2(c - 0.5)$$

Somers' D_{xy} is the difference between concordant and discordant probabilities and is interpreted on a range of 0 to 1 rather than starting at a value of 0.5. While the Somers' D_{xy} variation can help interpretation by scaling the probability range, the AUC and c-statistic are the preferred metrics for reporting predictive discrimination in disease modeling literature.

One limitation of current discrimination measures is their summarization across many decision thresholds. In clinical practice, a specific threshold or set of risk cutoffs are likely to be applied. However, significant study is often required to define optimal cutoffs. Summary metrics have become the standard since it is difficult to determine risk levels for many diseases. Another limitation for discrimination metrics include diminishing returns of the magnitude of difference between the internal validity scores of competing models. In particular, minor adjustments to model parameters or the addition of a feature to a model might result in a slight discrimination increase in the c-statistic [21,22]. These incremental changes in discrimination make it difficult to state when one model clearly outperforms another.

As an approach to this issue, a set of reclassification measures were proposed by Pencina to test for finer changes in discrimination. They include the net reclassification index (NRI) and integrated discrimination index (IDI) [21]. Reclassification measures compare two models to determine if adding a feature causes a beneficial shift in the number of true-positive and true-negative classifications. The NRI was originally designed to use risk cutoffs, making the measure susceptible to similar flaws in the c-statistic requiring pre-defined cutoffs. Therefore, a continuous version of the NRI (cNRI or $NRI^{>0}$) was proposed, comparing the relative increase in the predicted probabilities for cases with events and the corresponding decrease for those without the event [23].

The IDI does not require risk assignment as the value is an integration across the valid cutoffs of the predicted sample [21]. These metrics are most commonly discussed in relation to internal validity as they can inform if a model is incomplete and should use additional features. NRI and IDI may have additional utility for external validation by providing a new means for comparing across model designs published in different studies. However, researchers continue to consider the full utility of these recently proposed metrics [24–29] and some controversy still exists about the stability of the NRI score in particular [27,29,30]. Similarly, appropriately defining and evaluating thresholds for clinical use remains a significant hurdle to application of these prediction metrics [31,32].

2.2.2 External Validity (Transportability)

External validation is intended to compare a complete, internally validated model to an external target cohort that is "plausibly similar" to the source. The extent to which a target cohort is similar to a source environment is frequently left open-ended in general discussion. Just as RCCTs can be tweaked into a multitude of designs to target specific bias and confounding, external validation can be assigned a whole variety of terms depending on the known or anticipated differences between source and target cohorts. For example, Justice et al. differentiated between transportability studies with historical, geographical, methodological, spectrum, and follow-up properties [33]. Similarly, when describing the overall evaluation process of models in new settings, authors use a variety of terms such as external validity, generalizability, and transportability interchangeably. In this dissertation the same terminology is widely applied. Preference is given to the term *transportability* due to its common usage in both general and causal modeling publications.

Justice et al., introduced multiple definitions of transportability to help define scenarios of cohort difference where models should be tested. In their discussion, specific definitions of accuracy and generalizability (Table 2.1) were proposed to describe what combinations of transportability could be evaluated in published works. An iterative validation approach was proposed that included a five-level hierarchy of external validation. Each level of Justice's hierarchy accumulates evidence from additional aspects of transportability, ultimately leading to a comprehensive evaluation. The fifth and most comprehensive level, for example, requires evidence from multiple independent validations and varying follow-up time periods to provide a complete assessment of model transportability across Justice's proposed transportability differences [33].

Term	Definition or Criteria
Accuracy	The degree to which predicted outcomes match observed outcomes
Calibration	Predicted probability is neither too high too low (commonly shown with nor calibration curves)
Discrimination	Relative ranking of individual risk is in correct order (observed event rates in those with higher scores are higher); commonly measured with the area under the receiver-operating characteristic curve
Generalizability	Ability of a prognostic system to provide accurate predictions in a new sample of patients
Reproducibility	The system is accurate in patients who were not included in development but who are from an identical population
Transportability	The system is accurate in patients drawn from a different but related population or in data collected by using methods that differ from those used in development
Historical	Accuracy is maintained when the system tested in data from different calendar time
Geographic	Accuracy is maintained when the system is tested in data from different locations
Methodologic	Accuracy is maintained when the system is tested in data collected by using different methods
Spectrum	Accuracy is maintained in a patient sample that is, on average, more or less advanced in disease process or that has a somewhat different disease process or trajectory
Follow-up interval	Accuracy is maintained when the system is tested over a longer or shorter period

Table 2.1 Justice et al.'s definitions of accuracy and generalizability terms. [33]

Historical (or temporal) transportability referred to assessments of source and target cohorts from different time periods, such as data collected 10 years apart where treatment and disease severity

may have changed drastically. Geographic transportability references comparisons of cohorts collected in different locations, such as comparing Los Angeles and San Francisco subjects or European models tested in the United States. When different data collection methods were used for a target cohort, the validation would then assess a model for methodologic transportability. Spectrum transportability requires that a model generalize in both discrimination and calibration to subjects known to have more (or less) advanced cases of the disease. Increased or decreased severity could require a model to predict scenarios that are outside of the original selection criteria used to collect training data. Finally, follow-up period transportability describes the scenario where a model must generalize to previously untested time frames, such as testing a model developed for 5 year survival outcome to predict 2 year survival [33]. These five types of transportability are not mutually exclusive and target cohorts may contain differences from multiple definitions. Justice's groupings serve as useful guides and examples of the types of difference expected between internal and external cohorts. Recently, Steyerberg proposed a more concise variation of these transportability definitions [12]. When known or observed, these differences between cohorts should be clearly reported.

During external analysis, there may be reasons for a researcher to evaluate one type of transportability over another. However, in most cases, a researcher should not perform external validation for the sake of proving a model transports to a particular case of difference (e.g., temporal or geographic). Focusing too heavily on one type of difference can bias findings to specific use cases rather than testing overall generalization. Therefore, model transportability should be tested as widely as possible and greater numbers of successful validations would dictate support for overall use of a given model. For example, the higher a model is placed in Justice's hierarchy of validation, the more confidence there is in the model's generalization. Current

external validation analysis continues to have difficulty broadly evaluating models due to the complexities of reporting models, difficulties formatting target cohorts to source model designs, and the need for new methods for assessing transportability. In the rest of this section, I discuss the current state of external validation methods.

Discrimination and Calibration

To date, metrics used for internal validation are also applied as the standard in external validation. First and foremost, discrimination of the target cohort is tested with either the c-statistic or AUC [19,34]. As the target cohort is independent of the source model, special techniques like cross-validation or bootstrapping are not performed. It is common practice to compare the c-statistics from internal and external evaluations to determine if model accuracy is affected by application to a target cohort. Robust comparison would contrast the external c-statistic against the three values of internal validation: the apparent, optimism corrected, and 95% confidence intervals of discrimination. However, this comparison is rarely made and many discrimination evaluations only consider the apparent c-statistic score. There are no obvious suggestions for testing the significance of difference between internal and external c-statistics. In addition, differences in discrimination scores may be difficult to interpret for iterative changes between similar models when many features are used.

Assessment of model calibration is also warranted for transportability testing. Model calibration is, by definition, perfect for the original model when comparing against the source data. Therefore, calibration is not typically included in internal validation analysis (though there can be uses for this test when performing model selection). Calibration measures explore the model's ability to predict outcomes at rates that match with observed outcomes. For example, if 25 of 100 patients in a cohort have brain cancer, then a well-calibrated model should predict a 25% risk for a newly

observed subject with the same characteristics as the known cancer cases used in model development [12]. For statistical examination of calibration, a re-calibration model equivalent to the regression being analyzed (e.g. logistic regression, Cox regression) should be fit, modeling target outcome versus the linear predictor values calculated using the source model. A perfectly calibrated model will assign probabilities that are equal to the observed outcomes and therefore have a slope of 1 and an intercept of 0. The fit of the calibration model is often visualized graphically and a perfect model falls along a 45-degree angle. Models offset from this optimal line indicate the amount of over/underfit and over/underestimation the model has in the target data. These discrepancies are summarized from the calibration curve by the intercept (overestimation) and slope (overfit) in logistic regression models or a baseline hazard and hazard ratio in Cox survival models. Adjustments can be made to update the linear predictor values using these calibration findings.

Other previously mentioned metrics, such as NRI and IDI, are not commonly discussed for use in external validation. However, there may be ways to utilize these metrics for a deeper understanding of transportability. An additional area being explored due to increased interest in external validity is the concept of clinical utility [22,31,35]. If a given model can pass a majority of transportability assessments, it is likely useful to provide to clinicians for decision making. However, standard discrimination and calibration tests summarize over many thresholds and analysis at specific intervals would be useful to external validation. Vickers introduced a technique called decision curve analysis (DCA) as a way to evaluate performance based on potential clinical decision thresholds [31]. Decision curves are a simple approach for quantifying the clinical usefulness of a prediction model compared to never treating patients, always treating patients, or treating patients using a competing model. The performance of prediction models is calculated at one or more risk

thresholds in order to determine the net benefit of using the predicted probabilities for decisions. However, appropriate risk thresholds can be difficult to define in many domains, limiting the application of decision curve analysis. As DCA is explored further, it would seem appropriate to apply it as part of external validity evaluations since many models might show clinical potential at specific thresholds despite having issues generalizing at other thresholds.

Evaluating external validity

External validation is not a new concept, but it has taken some time for it to become a prominent need and for researchers to formalize a process for evaluation in predictive modeling. Only recently have publications made strides towards more quantitative evaluations, but it is important to understand the foundation these recent methods build upon. One of the earlier and more complete discussions concerning proper external validation steps was provided in 1999 by Justice et al. [33]. Altman and Royston further reviewed the reasoning behind and need for validation [36]. Other useful papers by Bleeker and Konig began to propose formal practices for analysis in ensuing years and Steyerberg went in depth on proper practice in the book, *Clinical Prediction Models* [12,15,16]. Following a four paper series by Moons, Royston, Altman, and Vergouwe in 2009 for The BMJ, publications related to evaluation frameworks and external validity study began to increase. Steyerberg proposed a framework in 2010 that began to include more novel metrics such as NRI and DCA [22]. A paper by Dekkers that same year discussed external validity in clinical trials with interesting similarities in discussion points related to modeling. In 2012, Moons addressed a wider clinical audience with a two part publication in the journal, Heart [37]. Many other papers reviewing external validation or proposing frameworks have been published in the last few years [38–44]

While current practice results in quantitative values for discrimination and calibration and graphical reviews of calibration plots, the discussion and conclusions of these quantitative measures is often quite subjective. Theories are put forward concerning the reasons for the success or failure of the validation. Often the only objective results include statements concerning if the c-statistic is higher or lower in the target sample and if a re-calibration of the model is warranted. Yet, even these results are subjectively interpreted at times; for example, there are not concrete ways to define if a c-statistic that decreased by 0.01 is significantly different from a decrease of 0.05 or more. Similarly, when recalibrated models are presented in publications, there is not a defined method for establishing that the adjusted model is no longer statistically different from the perfect reference. Ultimately, there are many cases where it is unclear if a model should be defined as an external validation success or failure.

One of the most useful current attempts to address some of these issues, by Debray et al., proposes a framework for quantifying the relatedness of development and validation samples (source and target in the parlance of this dissertation) by testing case mix [41]. Depending on the sample difference, external validity evaluations were designated as a test of reproducibility (similar case mix) or transportability (different case mix). Development decisions in the source and target populations can result in varying case mix, where numbers of cases with the outcome or particular feature distribution are different. Such variations could be considered akin to a mix of geographic and spectrum transportability as described by Justice [33]. Thus, Debray's case mix and performance assessment is useful for disambiguating the source and target relatedness in order to better interpret the model contribution to transportibility. Debray used a three step framework to enhance the validation assessment. First, cohort relatedness was assessed with the c-statistic of a relatedness model between source and target cohorts. The mean and standard deviation of linear

predictors in the two datasets were also evaluated. Second, standard external evaluation using discrimination and calibration was performed. Finally, the results of the first two steps were interpreted together to designate a model as “reproducible” or “transportable”.

The “reproducibility” designation was related to the similarity of cohort case mix, and discrimination and calibration were not expected to change a great deal. The “transportability” designation was used for different case-mixes and the expectation was that discrimination and calibration were more likely to change in this scenario. Debray tested three validation datasets and concluded that two of the validation cohorts (with both similar and different case mix) required calibration adjustment while a third validation (different case-mix) did not require adjustment [41]. Therefore, the case-mix assessment was not perfect for disambiguating the affects of of differences between cohorts. Nevertheless, Debray’s framework is an interesting method to consider case differences in parallel to discrimination and calibration tests. Determining the source of differences will likely become an important factor for performing model updating, the next step in the validation when transportability fails.

2.2.3 Updating Techniques

If all transportability tests – such as discrimination, calibration, or others – are insignificant and designate a model as generalizable, then a model can be applied to a target cohort. Given enough positive results on other targets, the model may even be used widely in many situations. Yet in most cases, some level of adjustment will be required before applying a model on outside subjects. Re-calibration is a simple and commonly used technique for model updating. Steyerberg proposed a useful ordering of progressive stages of model updating in relation to logistic regression[12]. These stages are reproduced in Table 2.2 and discussion on their relationship to the model transportability is included below.

Updating Method	Notation
No updating	
Apply original model	β_{source}
Re-calibration	
Update Intercept	$\alpha + \beta_{source}$
Update Slope and Intercept	$\alpha + \delta * \beta_{source}$
Model revision	
Re-calibration + selective re-estimation	$\alpha + \delta * \beta_{source} + \gamma_{target\|p\leq 0.05}$
Re-estimation	$\alpha + \beta_{target}$
Model extension	
Re-calibration + selective re-estimation + selective extension	$\alpha + \delta * \beta_{source} + \gamma_{target\|p\leq 0.05} + \beta_{newfeatures\|p\leq 0.05}$
Re-estimation + selective extension	$\alpha + \beta_{target} + \beta_{newfeatures\|p\leq 0.05}$
Re-estimation + extension	$\alpha + \beta_{target} + \beta_{newfeatures}$

Table 2.2 Steyerberg's proposed methods and notation for updating previously developed logistic regression models for future use [12].

The first case involves no updating of the source model. Findings were deemed transportable during external validation and the coefficients from the source model, β_{source}, are used directly. The next category of updates covers re-calibration procedures. If the model slope or intercept are substantially affected for a target cohort, the intercept or calibration slope can be used to adjust the previous model coefficients. By adding or multiplying coefficient values, the linear predictor values of the model will calibrate more closely with the outcomes observed in the target cohort. Updating in this way is preferable to subsequent updating methods because only two parameters must be estimated and the original regression coefficients do not have to be re-trained. It is important to note, however, that re-calibration has no direct effect on the discrimination of the model in the target cohort and cannot be used to improve model accuracy. If discrimination is substantially affected, more aggressive model updating methods are needed. The next adjustment approach requires that the coefficients of the model be re-trained with target information.

Therefore, at this stage, adjustment becomes more problematic as incorporating target information begins to mix new coefficients into the model that would need to be reviewed with additional internal validity testing. Selective re-estimation attempts to replace specific coefficients in order to make this problem manageable. If a full re-estimation is performed, the original source model is only contributing feature selection information to the modeling process. At this stage, transportability has largely failed as the source coefficients were not able to generalize to the target. Full re-estimation is not preferred because large target datasets would be required to re-train the model and new internal and external validation analysis would be required. A final updating option considers the addition of new features to update a model, creating a model extension. In this case, the source study and cohort missed or removed important features during model design. This finding is problematic because model extensions also require new internal validation analysis that might have been more robust if applied in the source setting where more data is usually available. Therefore, extension methods are indicative of additional situations where models fail to generalize.

Many of Steyerberg's suggested updating methods are used to address failures of external validity. Future research will be required to determine if there are novel methods that can update models in these settings while minimizing repeated evaluation. Methods for adjustment that can maximize the amount of data used from the source model will ensure that time and money are not wasted during research studies in the source setting.

2.3 CAUSAL TRANSPORTABILITY

One future approach to the method of selective re-estimation (or partial re-estimation) is related to the theory of causal transportability published in the domain of artificial intelligence [45]. The

theory states that under certain causal model structures, experimental findings can be deemed transportable between source and target domains. The theory itself has been reviewed for use in observational cases and meta-analysis [45–47]. Generating causal models is another complex process, but the theory provides sound principles for analyzing models that should be adaptable for more general use. In this way, causal analysis could help identify model designs where estimated values from source and target cohorts could be used interchangeably when features are properly controlled for confounding. Chapter 6 of this dissertation presents an exploration into this potential updating process. The rest of this section provides background on graphical disease models, causal models, and the transportability theory developed by Bareinboim and Pearl.

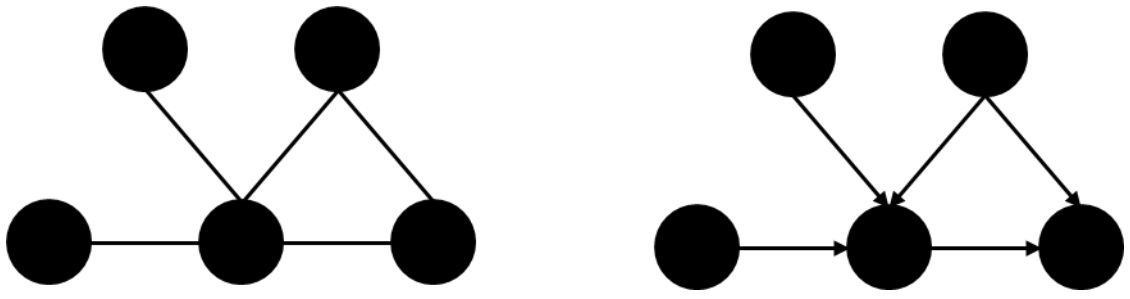
2.3.1 Graphical Disease Models

Graphical models are an increasingly prevalent technique for modeling probabilistic and causal relationships across observation and outcome variables of a disease. Combining probabilistic statistics and graph theory, a graphical model provides the capability to visually interpret and manipulate relationships between variables and move to an algebraic computation of probability distributions in a parsimonious fashion. Graphical representation is particularly advantageous because it provides an intuitive method for examining conditional independence of variables.

Relationships in disease models are descriptive of the inferences derived from experiments, the belief the model developer has about the variables from past experience/observation, and other sources of knowledge. For example, disease models can make use of RCCT and meta-analyses findings to provide guidance to graph structure or prior probability distributions, as they provide the set of presumed belief in the relationships between disease variables.

A graph incorporates well-defined sets of vertices and edges: vertices are representative of the variables chosen for a domain of interest, while edges drawn between vertices describe relationships that exist between a pair of variables. Therefore, a graph G is comprised of a set of edges E, and a set of vertices V: $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$.

Graphs have two important properties that influence their use for prediction: the inclusion or exclusion of directed edges and cycles. The simplest graphical form is a Markov random field, which is a graph with undirected edges (Figure 2.1a). This graphical form denotes correlation relationships between variables but does not designate directionality to the connection. Directed edges impart a specific knowledge concerning the dependence relationship between variables (Figure 2.1b). Modelers can take advantage of important traits of these relationships to reduce computational complexity. Graphs are called cyclic if they include at least one cycle between a set of variables. Their inclusion adds complexity to the system and makes it difficult to understand relationships between variables. Therefore, the most common graphical disease models do not allow for cycles. [3,48]



(a) (b)

Figure 2.1 Examples of Graphical models. a) An un-directed graph and b) a directed acyclic graph

Bayesian Belief Networks

When graphs only contain directional edges and do not provide any cyclical pathway between vertices they are termed directed acyclic graphs (DAGs). A Bayesian belief network (BBN) is a model developed around DAG structure. Directed edges in a DAG characterize the associations between variables, establishing which nodes are parents and children in the graph. When no directed link exists between a pair of variables they are conditionally independent. These independencies allow modelers to take advantage of a property known as the Markov condition.

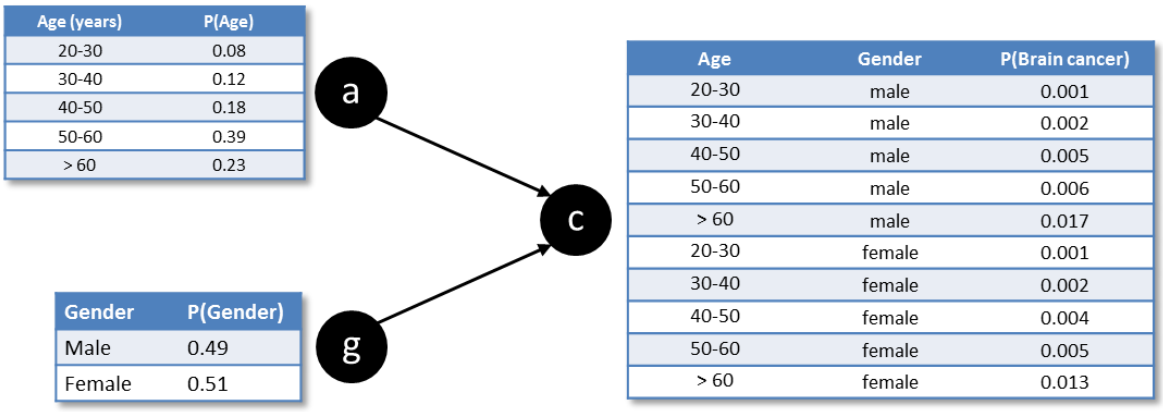
Definition: Markov condition

A variable is conditionally independent of its non-descendants given its parents.

Taking advantage of the Markov condition provides the ability to represent a DAG with a compact factorization to compute the joint probability distribution, the Markov Factorization:

$$P(x) = \prod_{v \in V} P(x_v | x_{pa(v)}) \quad (1)$$

Another way modelers view probabilities is with conditional probability tables (CPTs). For a given node in the graph, the CPT is representative of the node and the parents affecting the node (Figure 2.2). CPTs are an example of the simplification achieved by the Markov condition. The Markov factorization decomposes the joint probability into a product of the individual CPTs. Taking advantage of independence allows for a tractable solution to probabilistic calculations from these models.



Age (years)	P(Age)
20-30	0.08
30-40	0.12
40-50	0.18
50-60	0.39
> 60	0.23

Gender	P(Gender)
Male	0.49
Female	0.51

Age	Gender	P(Brain cancer)
20-30	male	0.001
30-40	male	0.002
40-50	male	0.005
50-60	male	0.006
> 60	male	0.017
20-30	female	0.001
30-40	female	0.002
40-50	female	0.004
50-60	female	0.005
> 60	female	0.013

Figure 2.2 Example of conditional probability tables (CPTs) for parent and children nodes.

D-Separation Criterion and the Markov Blanket

As previously mentioned, independencies are important for understanding the factorization of a graph and the probability distributions a graph can represent. Conditional independencies of variables can be read directly from a given graph using the d-separation criterion. The criterion consists of rules used to determine if a set of variables X are independent of a second set Y, given a third set of known variables Z. The formal definition of d-separation is as follows:

Definition: D-separation

A path p is said to be d-separated by a set of nodes Z if and only if

1. *p contains a chain $i \rightarrow m \rightarrow j$, such that m is in Z*
2. *p contains a fork $i \leftarrow m \rightarrow j$, such that m is in Z*
3. *p contains a collider (or inverted fork) $i \rightarrow m \leftarrow j$, where m is not in Z and no children of m are in Z*

A set Z d-separates X from Y if and only if Z blocks every path from a node in X to a node in Y using the conditions above. [48]

The above rules can be seen graphically in Figure 2.3 below. For rule 1 and 2, inclusion of the node *m* into Z allows for the d-separation of *i* and *j*. However, for rule 3 the exclusion of *m* and its children from Z is necessary so they do not open up a path.

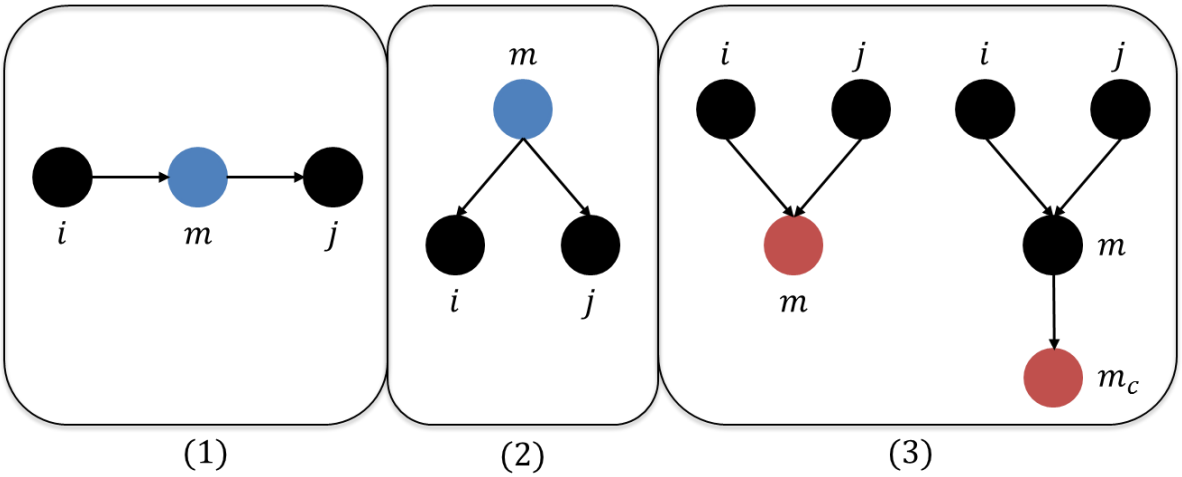


Figure 2.3 Minimal graphs that satisfy the d-separation criteria.

D-separation can be thought of as examining the flow of information between sets of variables. When a path is connected, information can flow between variables and dependence exists. A separated (blocked) path has no information flow between variables due to the inclusion or exclusion of other variables on the path based on the rules above. We use the terms d-connected or d-separated to describe these connections based upon the links between variables and a given separating set.

To provide a bit more context, consider an example based on Figure 2.4 where we consider d-separation as influenced by a set of variables. In the example, the nodes marked in blue, *r* and *v*, are included as members of set Z. The set Z is able to d-separate *x* from *y* since *r* blocks the path to *y*. Though *v* opens a path at the collider *t*, it does not affect the ability of *r* to block the path. However, the set Z is not able to d-separate *s* from *y*. The node *r* has no effect on *s* and the inclusion

of a child of *t* opens the collider path $s \rightarrow t \leftarrow u \rightarrow y$. If *v* is excluded from Z, as described in rule three of the d-separation criteria, then s and y are d-separated.

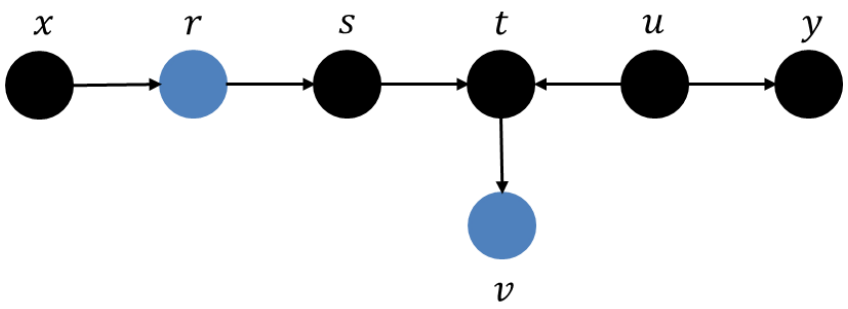


Figure 2.4 A DAG with separating set Z={r,v}

Another special property of DAGs drawn from the rules of d-separation is the Markov blanket. The set of parents, children and spouses of a given node in the graph makes up the Markov blanket for a given node. This collection of neighbors consists of the variables that shield the node from the rest of the network. That is, when values are provided for variables in the blanket or those variables are conditioned on, the target node becomes independent of all other variables outside the blanket.

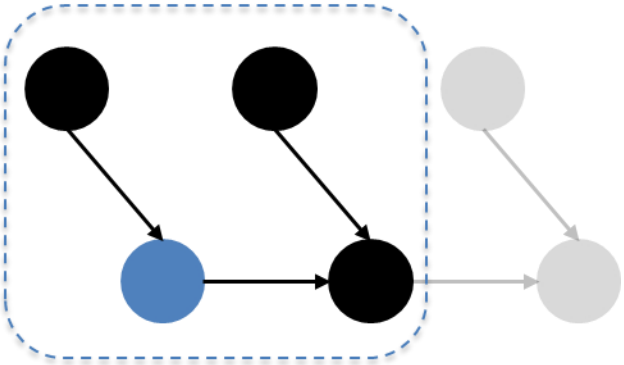


Figure 2.5 An example of the Markov blanket for a selected node (blue)

Bayesian Belief Networks in Biomedical Informatics

BBNs have become a widely used tool for prognostic and prediction models in medicine. Bayesian techniques have also become popular for modeling in many different domains including econometrics [49], artificial intelligence [48,50], and epidemiology [51]. BBNs provide particular advantages over other methods due to their parsimonious representation of probabilities using independence and a capability to compute probabilities with partial data. Early diagnostic and prognostic efforts for classification were reviewed by Shipster [3] include HEPAR, MUNIN, and Pathfinder [52–54]. Each of these efforts tackled models for different medical domains: HEPAR for liver and biliary tract, MUNIN for muscle and nerves, and Pathfinder for lymph-node pathology. These varying targets exemplify the capability of BBNs for use across the medical field as graphical models are flexible in handling a variety of variables and causal considerations.

One long running example of Bayesian network use in medicine can be found in predicting breast cancer risk. MammoNet was one of the first networks for mammographic screening based around patient history, current physical data, and radiologic features [55]. Information from experts and literature were used to establish conditional probabilities and 77 cases were evaluated for benign or malignant lesion status. Later, another Bayesian network was developed around well studied breast cancer imaging features (BI-RADS features) and was able to increase predictive accuracy over the previous model [56]. Further refinement of the model for predicting lesion malignancy was able to demonstrate prediction accuracy similar to an expert radiologist [57,58]. In addition, the results of ROC analysis indicated that combining the model and expert information would increase the accuracy over using either technique alone [58].

Bayesian techniques have also led to comparison work against other predictive modeling options. Lung cancer predictions using Bayesian networks and support vector machines (SVMs) were

compared by Jayasura et al. [59]; both techniques performed with similar accuracy, but Bayesian techniques retained higher accuracy when missing data became a factor. Bioinformatics researchers also utilize BBNs to search genomic and proteomic data for predictive cell markers to expand the features available for clinical models. Guha et al. provide an example where a Bayesian network was used to target lung cancer by analyzing proteins related to two highly studied cancer genes, EGFR and KRAS [60]. Cancer is a popular target for predictive modeling and publications continue to point to increased use of BBNs for targeting different forms [55,56,61]. BBNs are also increasingly seen for other predictive tasks in medicine such as head-injury [62], trauma [63], and venous thrombosis [64]. In addition to predictive and prognostic tasks, BBNs can be constructed for use in diagnostic reasoning or treatment selection; a more in depth discussion of these methods can be found in [65].

Utilizing Bayesian models for prediction in glioblastoma multiforme (GBM) is relatively unexplored. Efforts are still largely focused on determining appropriate prognostic variables. Recent studies have considered standard clinical and treatment variables [66,67], imaging markers [68,69], and genetic factors [70,71] for use in modeling GBM evolution. Many of these efforts utilize logistic regression models or Cox proportional hazard models for testing appropriate predictive variables, but few have considered transitioning to Bayesian techniques.

2.3.2 Causal Models

The special constraints of a DAG are useful for the purposes of probabilistic prediction modeling. But, a distinction should be made concerning the differences between causal and probabilistic models (e.g., BBNs). A causal model provides a directed edge from node X to node Y if the value assumed by X is used in the function that determines the value of Y. These connections in a causal

model are representative of an “X causes Y” relationship, conveying an inherent sequential ordering of events and representing a direct functional relationship.

Within the probabilistic (Bayesian) context, the edges between nodes are commonly interpreted in a similar fashion as causal connections because modelers derive their assumptions from the intuition of experts or experimental findings. However, BBNs are not causal models of the disease at hand simply because modelers have an intention of using directed edges as assumptions of causality. The relationships between variables are encoded by means of a conditional probability table of variables given their parents. These probabilistic dependencies can be described by a number of equivalent graphs. For example, Figure 2.6 presents a network with three variables where the directionality of the edges is presented in two ways. A single probability distribution can be described by either of these directed models because they contain the same set of conditional independencies. In contrast, arrows in a causal model not only represent probabilistic dependence but also direct causation.

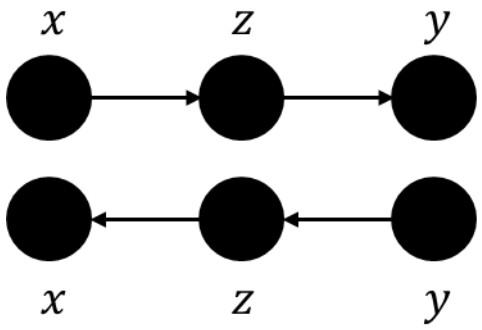


Figure 2.6 Two graphs with different directed edges but the same joint probability.

Thus, a causal graph should be utilized to construct a causal model (network), embedding the strong causal claims that move beyond probabilistic descriptions. Examining the model with these properties provides a modeler with the ability to consider interventions on the system.

External intervention on a model can be described as forcing a variable to take a set value rather than allowing it to follow the underlying function it naturally contributes to the model. For example, for intervention on a single variable X, the value of X in the graph is forced to the value $\boldsymbol{x_i}$ rather than following the natural distribution. This procedure is commonly described in equation form using a do() operator. The previous example is the intervention $do(x)$ on the variable X. Interventions are used for determining the causal effect of one variable on another. This can be easily equated to the experimental process in RCCTs, where randomization steps are used to force a value of interest to a specific value to study its causal effect on another variable.

Intervention on a set of variables has implications on the causal graph. As intervention is a forced action, a variable under the influence of intervention is no longer dependent on its parents. Graphically, this is represented by a break in the directed link between the parent and child nodes. The updated graph under intervention is commonly noted as a mutilated graph. Using different sets of mutilated graphs, inference rules can be used to convert probabilistic sentences involving interventions to sentences that involve only observations. These inference rules are known as the do-calculus and are described further below.

In addition to the added capability to explore interventions using a causal model, it is also possible to consider arcs in the graph representing uncontrolled confounders. Represented by bi-directed edges, confounding links describe correlations caused by unknown or unmeasurable variables related to the system. Uncontrolled confounders are very problematic for the d-separation characteristics of a graph as there is no available data or distribution to condition upon. This results in connections that obscure the direct causal connections between the variables, ultimately disrupting the ability to identify effects and perform accurate predictive inference.

Figure 2.7a contains an example of a possible graphical causal model for prediction of GBM survival from treatment and genetic variables. Additionally, the graph can also incorporate confounding links as shown in Figure 2.7b by dashed bi-directed arcs. The square nodes of Figure 2.7b are an additional item used for the purpose of transportability theory and are discussed in section 2.3.3.

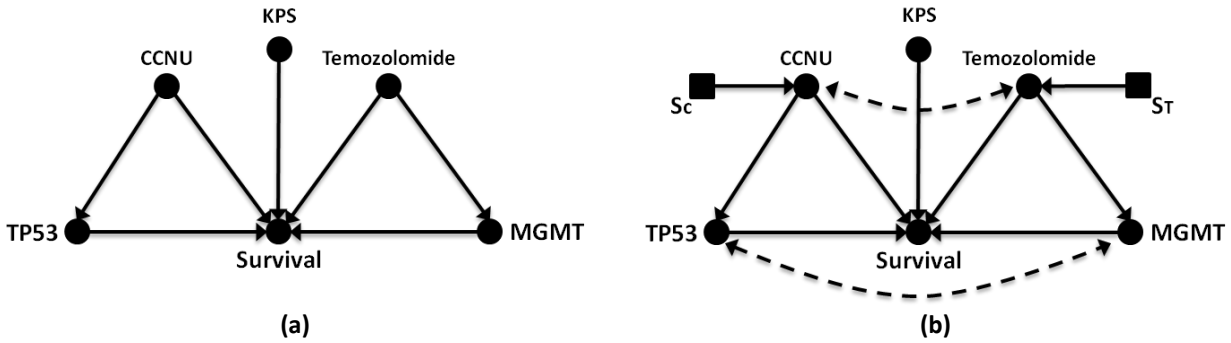


Figure 2.7 Example causal diagram for (a) GBM survival prediction and (b) the same causal diagram of GBM with links and nodes represented expected confounding information and population differences for variables. In the diagram, solid circular nodes represent observed variables; while square nodes indicate selection nodes controlling for population differences. Causal links are represented with solid lines with directional arrows. Bi-directional dashed lines indicate a variables linked by confounders. The selected observational variables are Tumor Protein 53 (TP53); O6-methylguanine-DNA-methyltransferase (MGMT); Temozolomide (Temodar); CCNU (Lomustine); and Karnofsky Performance Score (KPS). Unique selection nodes for CCNU and temozolomide are shown as SC and ST

Identification

The end goal of a disease modeler is typically to provide a prediction concerning a patient or population utilizing passive observations. Yet, causal models are built around assumptions and interventions that explain events based upon experimentation. To utilize causal models appropriately for probabilistic inference, we must determine when a model is *identifiable* from observational findings.

Following from Pearl's definition of causal effect identifiability, we can state that when $P(y|do(X = x))$ is identifiable, we can infer the effect of the $do(X = x)$ intervention from observational data and the causal graph, *G*. That is, identifiability allows for non-experimental data

to be used with our incomplete causal knowledge in order to estimate values when large samples exist for estimating the probability distribution of a variable, $P(v)$. Two graphical tests are particularly useful when dealing with identification of effects with existing covariates: the back-door and front-door criteria.

The back-door criterion indicates if a set of variables is able to block specific paths that point into a node X. Therefore, the set Z is able to block edges with arrows that enter X through the "back-door". The back-door criterion is defined as follows:

Definition: Back-door criterion

A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X, Y) in a DAG G if:

i. No node in Z is a descendent of X; and

ii. Z blocks every path between X and Y that contains an arrow into X. [48]

The criterion is easy to examine graphically as seen in the example in Figure 2.8. Particular sets of variables from $X_1 \ldots X_6$ can be a part of Z and satisfy the criterion. In the example, the sets Z=$\{X_3, X_4\}$ and Z=$\{X_4, X_5\}$ both satisfy the back-door criterion. The set Z=$\{X_4\}$ does not satisfy the needed criteria, because a path remains that can cross the collider at $X_1 \rightarrow X_4 \leftarrow X_2$.

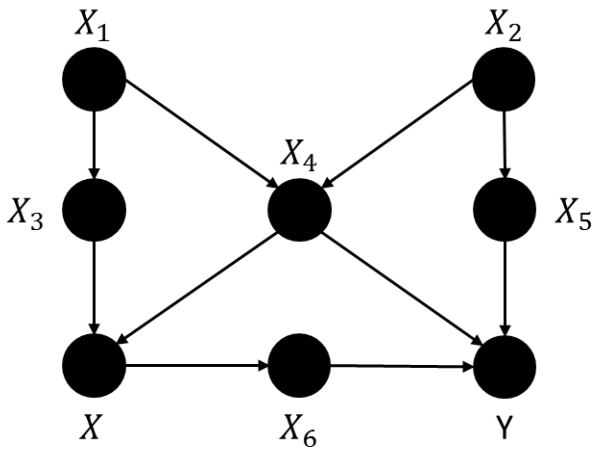


Figure 2.8 Example DAG for Back-door criterion.

The Front-door criterion examines the case where covariates affected by the variable X can be used to identify the causal effect unlike in the back-door criterion above. The front-door criterion is defined as follows:

Definition: Front-door criterion

A set of variables Z satisfies the font-door criterion relative to an ordered pair of variables (X,Y) if:

i. Z intercepts all directed paths from X to Y;

ii. There is no unblocked back-door path from X to Z; and

iii. All back-door paths from Z to Y are blocked by X. [48]

Just as with the back-door criterion above, we can examine this test graphically. Figure 2.9 is drawn from Figure 2.8 where $X_1 \dots X_5$ are now unmeasured variables. As $X_1 \dots X_5$ can no longer serve to block back-door paths we need another mechanism for identification. With the front-door criterion, we see that X_6 intercepts the path from X to Y as required in (*i*) while X itself is able to block the only back-door path across the unknown variables, U.

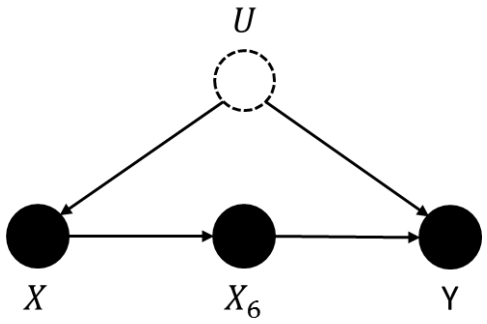


Figure 2.9 Example DAG for Front-door criterion.

Pearl notes that the front-door criterion is overly restrictive and that some paths denied by conditions (*ii*) and (*iii*) can be loosened depending on other covariates of the graph. Due to this issue, Pearl moved forward with developing a formal set of inference rules for working with identification in causal graphs called the *do-calculus*. [48]

Do-Calculus

The do-calculus is a set of inference rules for transforming probabilistic sentences with interventions and observations into new sentences. The rules follow the use of the do() operator mentioned previously for describing interventions in causal graphs. The application of the do-calculus rules can be used iteratively to reduce an expression for P(y|$do(X = x)$) into an expression with no do() terms. When this reduction is possible, the causal effect of X on Y is identifiable for the graph, G. An equivalent notation for $do(X = x)$ is $\hat{x}$ and is used to simplify the do-calculus notation.

The do-calculus rules follow from the interpretation of the do() operator causing a sub-model of the original model that is represented by a mutilated graph. A mutilated graph, $G_{\overline{X}}$, represents a graph where all incoming arrows to node X have been removed. A graph $G_{\underline{X}}$ represents the graph where outbound arrows from node X have been removed. The first rule of do-calculus, therefore,

relates to the conditional independencies after the graph changes from a do intervention, $\hat{x}$. Removing variables from the set Z do not introduce new dependencies to the graph. Rule two observes that an intervention on Z has the same effect as simply observing Z=z. This rule is related to the back-door criterion as Z is blocking those paths. Rule three deals with interventions on Z that have no effect on Y at all. The independence of Y and Z allows for the intervention to be removed completely. Z should not include ancestors of W, however, as they are important for blocking back-door paths involved in the front-door criterion. The rules of the do-calculus are as follows:

Definition: Rules of do-calculus

1. *(Insertion/deletion of observations):*

   $P(y|\hat{x}, z, w) = P(y|\hat{x}, w)$ if $(Y \perp Z|X, W)_{G_{\overline{X}}}$

2. *(Action/observation exchange):*

   $P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w)$ if $(Y \perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$

3. *(Insertion/deletion of actions):*

   $P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w)$ if $(Y \perp Z|X, W)_{G_{\overline{X},\overline{Z(W)}}}$

   Where Z(W) is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$ [48]

The do-calculus has been proven complete [72,73], meaning that it is sufficient for deriving all identifiable causal effects. Repeated application of the do-calculus rules can render a given interventional sentence to be "hat-free", where the sentence is written in terms of only observational values. When such a statement cannot be completely reduced, then the causal graph is unable to be used to identify the effect of X on Y. [48]

2.3.3 Transportability Theory

Transportability theory provides a basis for describing the ability to "transport", or move, data between populations based on the relations expressed through a causal model. For example, a physician in a rural setting might wish to apply the results of an RCCT conducted at a large research hospital to decision making for patients locally under his/her care. The RCCT findings can be understood in the context of a causal graph, per Figure 2.7, as a treatment (A) with effect on a patient outcome (B), with additional measured factors such as clinical history, imaging, or genetics (C, D, E). Transportability allows a researcher to identify potential confounding evidence between variables (represented by dotted lines) and population differences (indicated by the square nodes, S_C and S_T) that influence if the findings from the RCCT can be applied to the rural patients in a principled way. For instance, the physician may not have genetic information for his population; applying transportability can help ascertain whether the genetic information collected in the RCCT can be reused (i.e., transported) with the local group (and if not, under what additional circumstances such data transport is valid). Likewise, differences between the hospital and local populations (e.g., demographics, disease prevalence within the region) can be accommodated via transportability. Thus, if no substantial differences can be proven to exist, or existing differences can be accounted for utilizing properties such as the do-calculus, then the external validation can be used to confirm the model is transportable to the new population. A suitably constructed causal graph can be used to ascribe the set of variables which are not portable and must contribute probabilities to an inference task for a given population. The remaining variables in the graph are those capable of reuse to transport information from a study population to a new population.

To properly describe the full set of causal connections in the graph and make it useful for transportability testing, additional information not commonly captured in a BBN must be explicitly

represented. First, unmeasured confounding information expected to exist between any two nodes needs to be marked accordingly. These confounders are represented by bi-directional dashed edges and cover the counterfactual circumstances of variables that may be impossible to observe or measure. An example of this situation could be the potential interaction of a non-prescription pain-killer and treatments prescribed by the physician (Figure 2.7, the dashed arc between A and B). Patients may not report their non-prescription use and there may be unmeasurable interactions even if the physician knows both drugs are being taken. Second, when population differences are suspected or known to exist for a particular variable, a selection node is added that embeds the need to control for variation. A selection node serves this purpose by explicitly identifying population differences (e.g., disparities in demographics, socioeconomic status) that are responsible for assigning a value to that variable. By way of illustration, if age differences were significant between two populations, a selection node could be used to define a patient selection that maintains age-matching. I discuss these points further in the context of the simplistic Bayesian model of GBM in Figure 2.7.

In the example GBM model in Figure 2.7a, we have a set of six variables and their causal connections. The example network comprises no connective links other than the direct causal connections derived from the literature. With no additional links to consider, we would find ourselves in a state where the findings for one population are (in theory) transportable to another. However, causal assumptions have been made in constructing this graph, and we must consider the differences that likely exist between our nodes and the target populations upon which the model could be applied. Rather, most networks will more reasonably be in a state where there are specific confounders and selection nodes to manipulate. A graph with a number of these issues, such as in Figure 2.7b, is a case where data may not be transportable unless certain constraints can be met

either by transportability rules and the use of do-calculus. Otherwise, modelers should have strong evidence or support that removal of the connections can be made without affecting the outcomes. The goal is to map between the real-world graph in Figure 2.7b and the ideal causal graph in Figure 2.7a to enable the transport of information.

Having properly described the links assigned to the graph, the application of Pearl's work with transportability is now possible. The algebraic rules of do-calculus [45,48] enable a formal mathematical statement to be derived that determines what elements of information are transportable with the given variables, relationships, confounders, and selection nodes. Further graphical analysis via d-separation and back-/front-door criteria can help determine which variables of the model are identifiable. In this way, the full spectrum of techniques related to causal graphical models can be used to determine identification. When a causal graph is not identifiable, its findings are not transportable. Further examples drawn for these situations are reviewed in more detail in the available work from Pearl and Barenboim [45,46].

2.4 SUMMARY

As reviewed in this chapter, randomized controlled clinical trials and meta-analysis provide a wealth of knowledge to clinicians and researchers, and disease models have the potential to provide decision support through statistical prediction. Validation is required to assess predictive model accuracy and to evaluate the capability of models to generalize to the population at large. In this chapter, I established the primary types of validation and prior work that has laid the foundation of validation analysis. This dissertation proposes new methods for calculating internal validation statistics with limited information and methods for improving the classification of external validation (transportability) decisions. Partial model updating is an important area of study for

providing effective tools for addressing transportability failures. Thus, graphical and causal model concepts were reviewed in order to provide a background in causal transportability theory. Transportability theory is considered for updating models evaluated in this dissertation. As the number of disease modeling papers continues to grow each year, understanding and improving model validation and updating techniques have become more important. This dissertation builds from the previously discussed works in order to improve transportability assessments and work towards the wide use of models in clinical decision making.

CHAPTER 3

GATHERING AND ANALYZING PUBLISHED MODELS

Predictive disease models are becoming more widespread in medical literature. A PubMed search for "prediction model" shows that the number of papers making references to models has been increasing steadily (2005: 1622 publications, 2010: 2741 publications, 2015: 5135 publications) [74]. Models are useful to medical research and clinical decision making because of their ability to combine multivariate relationships from varied data sources for more detailed disease analysis. The transition to electronic health records and new analytic techniques in imaging and genetics have caused an explosion of medical data contained in clinical trial reports, treatment guidelines, laboratory results, drug interaction databases, imaging results, and genetic profiles. The increased wealth of electronically stored information along with more powerful computing resources have driven data analysis and machine learning techniques that have brought modeling to the forefront as a tool for prediction in many fields including medicine.

Taking advantage of the knowledge produced across many modeling investigations is challenging. Models are built to answer many types of clinical questions and predictions are meant to provide support for important medical decisions. However, determining if a specific model or set of models can reliability support decision-making tasks is an important area under investigation. Insufficient reporting of methods and findings is one of the primary problems when considering models for application in decision making. A number of important steps in data collection, cleaning, discretization and imputation occur before models are trained. Discretization and imputation decisions can be particularly important when applying a model to future cases. Feature and model selection decisions are also important and provide context for what medical data is important to

collect and what outcomes a model is attempting to predict. Finally, the inclusion of interval validation analysis indicates the performance and potential bias of the model to the original data. Reporting standards for these decisions and tests are not fully developed and review papers in different disease domains have described the inconsistency of published values [5,75–78]. One set of recommendations based on literature review, surveying, and discussions between statisticians, epidemiologists, methodologists, health care professionals, and journal editors was recently put forward in the TRIPOD statement (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) [79,80]. Increased exposure of reporting recommendations should improve the quality and consistency of model publications and enable improved comparisons of model results and validity.

Another difficulty of applying models in new environments relates to the underlying differences that exist between the source cohort (i.e., subjects collected and analyzed during model development) and the target cohort (i.e., the subjects in need of future prediction and decision making). Internal validation provides an understanding of the bias in a model. Biased models are frequently overfit, leading to strong source performance that will not carry forward to different cases. Nevertheless, future model performance is not guaranteed even when bias is minimized. Target environments will frequently include patient cases the original model did not cover. Insights into potential differences could be used to control data collection and model factors and improve results. Clear reporting can provide awareness of potential differences. New techniques tied to external validation could also increase the number of models considered after failures by providing model adjustment or updating options.

The development of predictive models in cancer is driven by current medical knowledge and treatments. Future therapies and disease prognosis methods build from previous care and research.

However, it takes time to reach consensus in clinical trial research and predictive models have seen similar trends in determining a unifying set of biomarkers. For example, studies of glioblastoma multiforme (GBM), a type of primary brain cancer, have been unable to reach a definite consensus on the most effective predictive variables for GBM patients [66,71,81–86]. Researchers continue to work towards integrative models of disease that better utilize the influx of experimental data. Important to this task is the creation of models that are able to generalize between patient populations. External validation provides a means for determining the success or failure of a model to transport to target cohorts, but these analyses are not yet able to pinpoint differences and their causes. Previous epidemiological experiments have demonstrated the difficulty of carrying source results forward. Bleeker et al. examined model parameters using derivation (source) and validation (target) test sets. Coefficients derived for both sets of data were compared and the derivation model was unable to predict validation cases with sufficient accuracy [15]. A similar review of stroke models found some generalization was possible with temporally separated cases (i.e., temporal transportability), but all examined models had issues when applied to external data with other forms of difference [16].

In this chapter, a set of papers using predictive models for the analysis of GBM survival were selected. After review of a large set of queried papers, four modeling papers were reviewed in detail. Each model was then applied to a target cohort of UCLA patients and tested for model discrimination, the most common test for deciding model transportability. The difficulties of gathering relevant details concerning data processing, model design, and validation results are discussed, highlighting the obstacles clinicians and researchers face when considering the application of models to their data. These key issues were considered in relation to widely published reporting issues and influenced research presented in other chapters of this dissertation.

3.1 METHODS

To better understand the issues of gathering published data to support external validation, published literature was reviewed for recent models of GBM disease survival. A set of relevant papers were obtained from PubMed, following the common search practices of clinicians searching for publications applicable to their patients. PubMed results were reviewed and a set of models were selected for external application. The external dataset was created by mining a UCLA database of GBM patients treated after 2005. Parameters of the selected models were then used to determine performance and assess transportability by predicting UCLA outcomes. Figure 3.1 provides and overview of the analysis stages which are expanded on below.

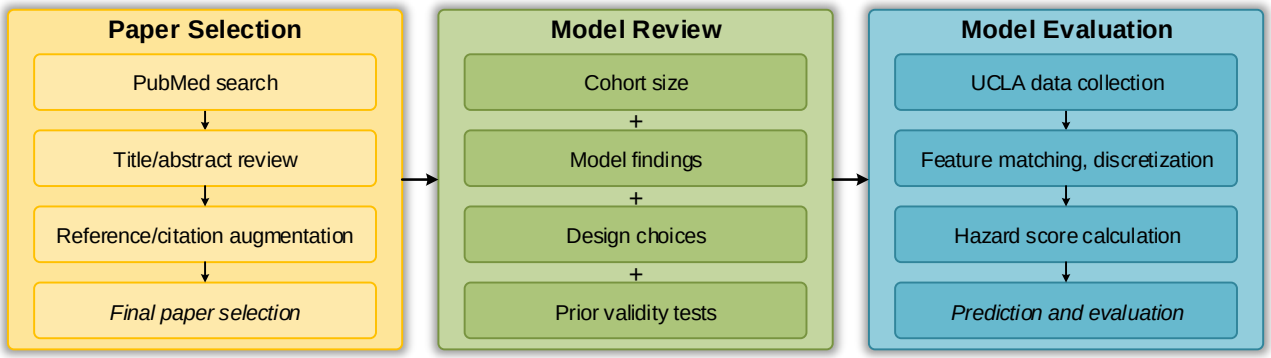


Figure 3.1 Selection, review, and evaluation process for this chapter. Paper selection combined search engine and manual review methods to choose a final paper set. Model review gathered and analyzed relevant elements from papers. Evaluation applied gathered knowledge to test prediction accuracy using a local dataset.

3.1.1 Paper selection

Papers were selected through a combination of PubMed database querying and manual curation. First, a PubMed search string was created using input from the PubMed online query builder. The search targeted papers containing references to glioblastoma multiforme and predictive modeling. The final search string combined free text phrases with expanded MESH terms to account for different wording choices in publications.

("glioblastoma"[MeSH Terms] OR "glioblastoma"[All Fields] OR ("glioblastoma"[All Fields] AND "multiforme"[All Fields]) OR "glioblastoma multiforme"[All Fields]) AND (prognostic[All Fields] OR predictive[All Fields] OR cox[All Fields] OR ("risk"[MeSH Terms] OR "risk"[All Fields]) OR prediction[All Fields] OR statistical[All Fields] OR hazard[All Fields]) AND ("Survival "[All Fields]) AND model[All Fields]

A total of 220 related papers were returned in the PubMed query in January 2014. Titles and abstracts of returned entries were manually reviewed to constrain the set further. Paper selection criteria used in this step included: studied survival outcome, published in 2010 or later, and included a detailed model description. This filtering reduced the PubMed set to 18 papers. Many of these remaining papers examined specific predictive targets in univariate analysis rather than multivariate prediction models. These univariate models were excluded and the references and citations of three remaining multivariate papers were used to expand the search for multivariate modeling papers [69,87,88]. Four papers [87–90] were chosen as the final modeling set. They included descriptions of the following model elements: cohort size, cohort demographics, model variables, significant variables chosen in multivariate Cox regression, and coefficients of the Cox model. One multivariate modeling paper by Mazurowksi et al., for example, did not include description of model coefficients and was removed from consideration [69]. Cox proportional hazard models were explored because they were the most studied model design in GBM papers during the review process. Table 3.1 summarizes details of the selected papers.

Title	Author	Published	Study Period	Cases Considered	Cases Used	Median Survival Time (Months)
Overall survival, prognostic factors, and repeated surgery in a consecutive series of 516 patients with glioblastoma multiforme [89]	Helseth	2010	2003-2008	516	516	9.9
Clinical variables serve as prognostic factors in a model for survival from glioblastoma multiforme: an observational study of a cohort of consecutive non-selected patients from a single institution [87]	Michaelsen	2013	2005-2010	225	225	14.6
MR Imaging Predictors of Molecular Profile and Survival: Multi-institutional Study of the TCGA Glioblastoma Data Set [90]	Gutman	2013	2006-2008+	75	68	13.3*
Evaluation of outcome and prognostic factors in patients of glioblastoma multiforme: A single institution experience [88]	Kumar	2013	2002-2009	439	360	7.67**
Comparison UCLA Dataset	Singleton	-	2005-2015	482	125	16.9

*Table 3.1 Summary of selected papers for model comparison. +End date approximated from first publication on the Cancer Genome Atlas (TCGA) data in Nature. *Value derived from data, not reported in literature. **Higher survival (7.97 months) reported for Group II (KPS>70).*

3.1.2 Model review

The four selected GBM modeling papers were inspected for elements important to understanding the creation of the patient cohort and the purpose of the original model. Some important items considered include the source cohort size, model hazards for calculating feature coefficients, descriptions of design choices (i.e., discretization, feature selection), and discussion points concerning data complexities or limitations that might imply population differences likely to

influence external validity. Each publication is identified by the first author of the work when referenced in the remainder of this dissertation: Helseth, Michaelsen, Gutman, and Kumar. Details of the source study size and survival times are summarized in Table 3.1 and compared to the local target dataset. The hazard ratios of significant multivariate features used in discrimination testing are reported in Table 3.2 for each model. Discretization and imputation findings are discussed in the next section as they influenced the target cohort creation. Other relevant details are discussed in the results when considering their implications in relation to validation tests.

	Helseth	Michaelsen	Gutman	Kumar
Sex	1.44			
Age (a, b)	1.02	1.31		
ECOG (c)	2.13	1.22		
		2.06		
KPS			0.972	
Tumor Location	2.31			1.52
Tumor Site				2.34
Surgery	2.72			
Corticosteroids		2.06		
Radiation Dose				2.03
Chemotherapy				0.44
Proportion of Contrast Enhancement			7.745	
Tumor Major Axis			1.016	

Table 3.2 Cox Regression Hazard Ratio Summary: a – Helseth Age continuous, b – Michaelsen Age discretized by 10, c – Michaelsen ECOG split into two dummy variables.

Authors for each of the papers were contacted in an attempt to gather training data, as the use of such data would provide the ability to replicate the original models and estimate additional metrics for transportability analysis. Unfortunately, data was unavailable either because authors did not respond or because data protections restricted access. The Gutman dataset was an exception; the source cohort was built form a public database and was therefore available for follow-up analysis. Without previous training data, evaluation is limited to more qualitative comparisons based on reported findings and discussion. Additional quantitative assessment is difficult beyond direct comparison of discrimination metrics. For instance, depending on the level of internal validation

reporting, an assessment of the calibration and overfit of hazard values may not be clear. Replication with the original data would provide an opportunity to review the internal validity of the model estimation. Similarly, design details were typically interpreted from published text summaries that data access could clear up. Ambiguities in language could lead to misinterpretations in application of the model that would affect the transportability. Thus, it can be difficult to pinpoint sources of error if the model performs poorly on new cases.

3.1.3 Transportability analysis

UCLA data collection and preprocessing

A retrospective cohort of UCLA patients was obtained by manually reviewing the records of patients seen for GBM assessment from 2005 to 2014. Manual chart review was performed and cases were removed from consideration when baseline clinical reports, pre-surgical imaging, or long-term follow-up information were missing, as this made it impossible to gather information relevant to the features in the selected published models. Patient and feature information was recorded to a local database as part of the chart review. Figure 3.2 summarizes the selection process. The final dataset used for evaluation included 125 cases. A primary reason for missing reports or imaging in the manual review was patient referral; initial evaluations from other hospitals were often not shared when subjects were referred to UCLA.

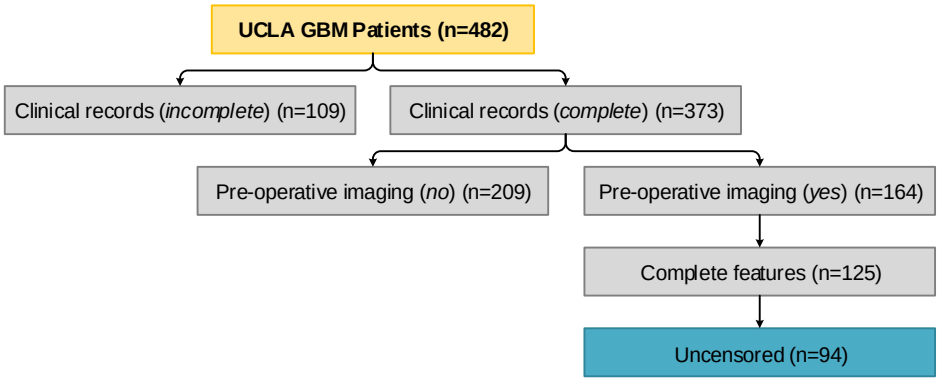


Figure 3.2 Patient selection process for building the UCLA test cohort.

Manual review attempted to gather complete case data for all clinical and imaging features seen in the Cox models. Collected values were recorded to a database of UCLA GBM subjects. The system was queried following collection to further determine the completeness of requisite clinical and imaging information. Baseline reports were important for determining age, sex, baseline diagnosis information, and initial Karnofsky performance score (KPS), a cognitive assessment variable. Follow-up reports were necessary for treatment variables such as surgery, chemotherapy, radiation therapy, and corticosteroid treatment (used to control cerebral edema and inflammation, thereby altering imaging appearance). Information for KPS and corticosteroid treatment proved to be the most difficult to obtain from reports. Imaging features were obtained from pre-surgical magnetic resonance (MR) imaging. A subset of the VASARI imaging feature set was recorded from these images by an expert neuro-radiologist (22 years of experience).

A total of 482 cases diagnosed with GBM were examined, yielding a final cohort of 125 cases with clinical and imaging data matching the feature sets of the previous models. Given that most of the papers did not report an imputation process, selection was conservative to avoid missing data and match the absence of imputation in previous works. Missing data was allowed only for

the KPS feature (six cases) as it was the only feature with a described imputation approach. Missing KPS values were imputed to a score of 80 in the Gutman paper.

Target cohort prediction

The risk rates of features in Cox models are usually reported as hazard ratios as seen in Table 3.2. Hazard ratio values can be easier to interpret than model coefficients as they indicate the increased or decreased hazard rate of an outcome (often death) for a patient with the presence of a feature. For example, a hazard ratio of 2 for a binary feature like Sex (e.g., Sex: Male = 1, Female=0) describes that that hazard rate for a male patient is double the rate for females.

Regression coefficients, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_i)$, can be obtained by taking the $\log_{10}$ of the hazard ratio values. Regression coefficients can be multiplied against the respective feature states of a subject, $\boldsymbol{x} = (x_1, \dots, x_i)$, and summed to obtain a linear predictor value used to predict outcome.

$$LP = \boldsymbol{x\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

Discrimination is commonly measured with either the area under the curve (AUC) of a receiver operating characteristic (ROC) curve or with the concordance statistic (c-statistic). Both measures are summary statistics of the overall predictive capability of a model. In this work we use the concordance measure, a rank statistic that compares predicted values for pairs of subjects with the outcome (Y=1) and without the outcome (Y=0). The proportion of pairs where a subject with positive survival outcome at a given time (i.e., Y=1|t) has a lower linear predictor, LP, than a paired subject that dies indicates the proportion of times the model will make a correct decision between these outcomes.

Using this procedure to convert reported hazards (Table 3.2), linear predictor values were calculated for target UCLA cases from the Helseth, Michaelsen, Gutman, and Kumar papers [87–90]. The c-statistic was calculated using these linear predictor values, outcome status, and survival times. Statistical analysis was programmed in R (v 3.1.3) and run in RStudio [91,92]. Linear predictors were calculated manually and adjusted based on feature means derived from the published summary statistics. Mean feature values were not published for the Gutman model and were instead calculated from shared data. The c-statistic was calculated using the *rcorr.cens* function of the 'rms' package [93]. The resulting target c-statistics for discrimination of each model were compared against published values of the source cohort c-statistic where available.

3.1.4 Replication analysis

In addition to applying the four models, source data from Gutman et. al was used to perform a replication analysis. Patient selection and discretization were attempted on the source cohort using the Gutman paper's design description. A set of 68 cases were selected, removing cases with missing outcome information as closely as possible to the process described in the original work. Missing KPS values were imputed to a value of 80. This 68 case cohort was used to train a Cox proportional hazards model in with the *cph* function in the 'rms' package [93]. The hazards from this Cox model were compared directly with the published values to determine if a replication of the original process was successful. In addition, the resulting model was used to compute linear predictors and c-statistic for the target UCLA cohort.

As the original Gutman work did not report an assessment of internal validation, the c-statistic from apparent validation of this replication model was used for comparing performance c-statistics found when applying to the target cohort.

3.2 RESULTS

3.2.1 Selected Papers

Four papers [87–90] were selected as final targets for analysis based on the overall strength of their reporting. Descriptions of selection criteria, population statistics, and modeling methods were more robust than other modeling papers that were considered. However, each paper had data collection and modeling decisions that were difficult to decode or apply. For example, Helseth made a number of references to significant factors related to patient ages greater than 60 and 70 years (likely from a stratified analysis that was not completely included). However, the reported hazard ratio for age was based on a continuous measurement during modeling. This added confusion, initially causing an incorrect analysis to be run with a binary discretization based on age greater than 60 before the misinterpretation was detected. In Michaelsen, corticosteroid therapy was determined at the start of treatment following surgery. When collecting the status of this treatment for UCLA patients, initial treatment time was not static. Some patients received steroids immediately after surgery; others received steroids weeks later, following radio-chemotherapy. Subsequently, decisions concerning corticosteroid treatment were sometimes unclear in UCLA cases, making it difficult to code values based on the descriptions given by Michaelsen. In Gutman, a standardized feature set for imaging (VASARI) was evaluated using a team of readers. However, VASARI features have yet to be fully validated using radiologists who were not trained as part of the original study [69]. Classification of percentages of enhancing tumor in the UCLA dataset were lower than values determined from cases in the Cancer Genome Atlas (TCGA) dataset. As only one expert reader was available for this analysis, additional tests could not be performed to determine whether this difference was due to radiologist interpretation, MRI protocols, or disease traits. Finally, given patients' limited access to facilities in India, the Kumar

study was partly focused on complex analysis of radiation treatment schedules. UCLA follows a more standardized treatment pattern. To make these difference more comparable, radiation treatment groups were simplified into treatment and non-treatment classes rather than being judged at a specific dosage. Each of the described issues impart the need to clearly define model constraints. Even with the detailed reporting in this set of papers, some constraints had to be adjusted due to the availability of information and difference of treatment at UCLA. When reports were unclear, application became a game of assumptions. Errors based on assumptions are more likely, such as incorrect discretization or improper feature collection, which might affect model performance.

Additional threats exist to the external validation of models beyond interpretations of modeling decisions. For the Helseth model, the inclusion of data from before and after 2005 might be problematic. In 2005, Stupp's report [94] became a milestone publication revealing that tumor resection and temozolomide therapy result in the longest GBM survivals. Including cases prior to this milestone might threaten temporal validity since survival chances increased when treatment protocols were replaced by the Stupp protocol. Patient selection in Michaelsen's evaluation excluded cases where ECOG scores were high (i.e., individuals with poorer outcomes). Michaelsen's population, therefore, is likely skewed to predicting cases with longer survival. This difference in survival rates could influence the generalizability of the model to other populations with higher ECOG scores and worse prognosis. In Gutman, the size of the original dataset is of concern. The primary goal of the work was to assess predictive ability of imaging features, but prediction was not directly evaluated with an internal validation. Follow-up replication showed a large amount of variability in the predictions made with this feature set, implying that the model coefficients might not be sufficiently trained and more cases could be added to the cohort. Like

Michaelsen, Kumar stratified patients prior to analysis. Based on KPS values, two groups of patients were created and all reported multivariate results are related to Group II with KPS>70. Patients with low KPS values received a different treatment protocol and did not contribute to model training. Future cases with low KPS might not be accurately predicted by the model as a result. Finally, the Helseth, Michaelsen, and Kumar models include populations of patients from Europe and India. An Indian cohort with known differences in race, overall health, and socioeconomic status from patients in our California dataset might strongly influence the transportability of a model. European cohorts might have less difficulty transporting to other locations like those in the United States, but population differences could still play a role in model performance depending on the nature of the disease being studied.

3.2.2 Transportability evaluation

Decreased performance was seen in all models when comparing against available source c-statistic values (Table 3.3). Helseth and Michaelsen c-statistics indicated the best overall external performance, approaching a value of 0.7. Michaelsen, had the largest total decrease in performance on target data, dropping by about 12%. The original performance for Michaelsen was quite strong at 0.82. This large decrease is disappointing as Michaelsen's external performance, while still best overall, is now much closer the Helseth model. Gutman's performance decreased the least, but the source model had the worst original performance. Kumar did not report internal validity findings and values could not be obtained from the authors after multiple follow-up inquiries. Therefore, no direct interpretation can be made about the change in discrimination performance between the source and target. Similarly, it is unclear how to interpret the external performance compared to other models without some idea of the original c-statistic.

Model	Internal c-statistic	External c-statistic
Helseth	0.72	0.679
Michaelsen	0.82[a]	0.696
Gutman	0.632[b]	0.601
Kumar	-	0.633

Table 3.3 Previously published or derived internal c-statistics and external validation c-statistics when applied to a target UCLA cohort. a) Five-fold cross-validation c-statistics were reported as >0.80; b) Value derived from shared data, reported value was unavailable.

Based on current external validity assessments, the implication is that none of the models are transportable due to their performance decreases. Gutman might be considered the most generalizable to the target setting given that it had the smallest performance decrease between source and target. However, the overall utility of the model is lower than the other models because the c-statistic is closest to the random decision threshold of 0.5. If overall discrimination performance is preferred, Michaelsen would be chosen due its high discrimination performance (c-statistic = 0.696). However, the large difference in performance compared to internal validation indicates that the Michaelsen model does not generalize as well to the target as other models. Therefore, direct comparisons of internal and external c-statistics are not helpful for deciding whether each of these models are transportable to the target. Instead, confidence intervals should be obtained in order to perform significance testing of the differences between c-statistics. Further investigation of these models with CI's and additional statistics could provide a more accurate assessment of which models are transportable.

3.2.3 Replication evaluation

The replication attempt resulted in hazard coefficients that did not exactly match the previously reported values (Table 3.4). The hazard of KPS values remained the same as previously reported with a similar p-value. Smaller hazard coefficients were seen for the imaging features of major axis length (a measure of tumor size) and proportion of contrast-enhanced tumor (pCET). The decrease in hazard was not substantial for tumor length, but pCET hazard dropped from 7.7 to 5.1.

In addition, the significance of the imaging features was decreased in the replicated model attempt. Both imaging features were previously reported as significant predictors, but these features appeared more borderline (Length p=0.051) or insignificant (pCET p=0.066) in the replication attempt. However, the overall model was still able to significantly predict cases compared to a null model ($\chi^2 = 16.6$, p=0.0009). The apparent performance measured with the c-statistic was low (c-statistic = 0.632) compared to the other Cox regression models considered (c-statistic > 0.7).

	Reported		Replicated	
Variable	**Hazard Ratio**	**P-Value**	**Hazard Ratio**	**P-Value**
KPS	0.972	0.006	0.972	0.004
Major Axis Length	1.016	0.030	1.013	0.051
Proportion contrast-enhanced tumor	7.745	0.037	5.861	0.066

Table 3.4 Reported and replicated Cox proportional hazard values for the Gutman paper. Likelihood ratio tests with three degrees of freedom – Reported: $\chi^2 = 17.4$ (P=0.00059), Replicated: $\chi^2 = 16.6$ (P=0.0009).

Much of this instability in the model parameters is related to the small sample size of the Gutman cohort. When removing cases from the original 75 Gutman cases, there was some ambiguity in how to select survival outcome. Multiple survival features existed based on the reports allowed in TCGA. This made it unclear what 7 cases were removed in the original analysis, as multiple combinations seemed possible depending on the outcome features considered. Therefore, the changes seen are likely due to a different selection of the final 68 cases used in analysis. In discussions with Dr. Gutman, it was determined that the original selection method could not be repeated because the contributing statistician from that paper had moved to another institution.

3.3 DISCUSSION

Discovering relevant papers is one impediment to model application due to the varied goals of predictive disease models. In this work, an attempt to create a focused search string returned over two hundred papers with references to predictive models and GBM. However, only a handful

studied the target outcome of interest (overall survival) using multivariate prediction. Instead, univariate analysis was most prevalent and a number of loosely related papers were included in the results. Researchers and clinicians must have a clear understanding of the clinical question they wish to answer in order to select papers with designs that are useful. For example, other outcome targets such as time to progression or survival after progression might have more study with multivariate models. However, these modeling goals would require different decisions during data collection and feature specification, resulting in a separate analysis from the models in this work.

Even when the best papers can be found in literature searches, there are deficiencies in reporting when describing model choices and assumptions. During selection of GBM papers, similar reporting issues were observed to previous investigations in other disease domains [5,75–78]. Publications typically include population summaries and model coefficients, but the details of the processing used to obtain the case data and models are often lacking. Even when discussion appears complete, the ambiguity of free text description and inclusions of stratification analysis can obscure the most important decisions made during model construction. For example, in Helseth et al., discussion involving the effects of stratification at certain age cutoffs was included. However, these stratifications had no bearing on the final model design. In both Helseth and Michaelsen's papers, feature discretization used in modeling did not match with the population splits presented in summary tables. Without reading closely, a researcher may collect feature values for the target cohort incorrectly and influence the transportability comparison. The ability to combine relevant findings from the growing publication base is concerning if the quality of reporting cannot be improved so that inherent differences between studies can be clarified and compared.

Variance of populations and modeling decisions across papers is often the largest impediment to model application. Some of these differences can be detected with robust reporting. Within our example publications, data was drawn from four different countries. Gutman drew on public data from a large multi-institutional collaboration, which might have some extra strength when applied across domains, but might also suffer if similar protocols could not be applied to a target cohort. The other papers used data from local institutions, which can more easily follow a specific set of protocols and standardization, but might also be overfit to predicting patients from those institutions. Helseth and Michaelsen used ECOG score, more common to European clinics, to measure cognitive deficits while KPS measures were used in the UCLA target cases. This required feature conversion to use their models which adds a potential source of error [95]. These examples illustrate how quickly data collection and model design differences can accumulate across studies.

When the chosen source models were applied to the UCLA cohort, performance decreased in all cases. Such decreases are common in external analysis [5,75–78]. The many factors already discussed and additional population difference that may not be identifiable from descriptions of the source population are the primary contributors to this decrease. Current assessments of external validity using only discrimination cannot provide enough detail to gauge the extent of differences. Debray et al. discussed how performance decrease is often tied to the homogeneity, or case-mix, of a target sample [41]. Therefore, additional metrics can further distinguish between source and target differences and their relationships to transportability outcomes. Models will continue to be published exploring potential clinical markers in disease. However, without a means to apply these models more generally, a number of opportunities to translate research to the clinical setting are being wasted.

Standardized sets of expected items in reporting should continue to be pushed to aid the transition to increased transportability assessment. Twelve journals have publicly supported the TRIPOD statement and additional journals should consider TRIPOD or alternative guidelines. Regular use of appendices can provide extra space for detailed model information and programming code. Breaking down barriers to data sharing should also be explored further. Additional data may have been available for this analysis, but patient data access was restricted by legal protections. The Observational Health Data Sciences and Informatics (OHDSI) network could serve as an example system for applying standards in data sharing. OHDSI uses the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to create datasets from multiple institutions. Datasets shared with the CDM can be easily run with multiple open source tools with fewer restrictions as researchers do not store data locally [96–99]. Finally, formalized model standards could provide computer readable model descriptions. Encapsulating all the design choices and steps into a standard, annotated representation removes any ambiguity when reapplying modeling steps to process future data. The Predictive Model Markup Language (PMML) is an example of an XML standard for model description used within the data mining community [100]. Reporting, data sharing, and computerized model standards could be combined to open the door to easier identification and transport analysis of models. In particular, these standards can provide easier access that will facilitate new analytical tests that can enhance definitions of source and target difference.

3.3.1 Limitations

The paper selection process targeted modeling papers with reference to GBM. The original search string did not make a distinction between univariate and multivariate models, increasing the number of returned results and likely allowing for additional false positives. In addition, the

targeted outcome of overall survival was important and many returned papers had other outcome targets. For this analysis, a reference and citation review was used to expand the search for more relevant works to partly address this limitation. In depth review of the most relevant works could be performed in future work to better define a new query that could more accurately return relevant papers.

Deriving all of the relevant modeling information from publications was limited to available free-text. This process mirrors how a clinician reviews literature to apply a published model. However, some items concerning distribution, discretization, and variable selection techniques were not always clear. Internal validity tests of discrimination were also not available for the Gutman and Kumar papers. Authors of each paper were contacted by email to attempt to address this limitation, requesting access to data or reports of missing c-statistics in addition to inquiring about interpretations of the text. Gutman and Michaelsen were responsive, but extra input was usually limited due to time and sharing constraints. In the end, Gutman data was shared and used for additional analysis. No other data or updated values were able to be obtained for the other three publications.

The examination of four separate model designs limited the overlap of relevant features, as many potential predictive features are still under investigation in GBM research. This may have biased the cases selected during construction of the UCLA cohort. When selecting patients and gathering data for relevant features, more effort was necessary to gather the full set of considered features. This limited the total number of patients that were reviewed since data entry required manual review of patient reports. In addition, during final patient selection, patients who only had data relevant to one or two of the model designs were often excluded in favor of a more complete case selection. A database of GBM cases including these cases and future additions will be available to

better track these findings and may help with future imputation attempts rather than dropping many cases with missing values.

3.3.2 Conclusion

Inconsistencies in reporting were seen in a review of predictive modeling papers on GBM patient survival. These reporting issues are a substantial hurdle to transportability analysis, as demonstrated in an analysis of four previously published models. Overall performance dropped in all model applications to target UCLA data and in one case a comparison to the original results was not possible due to reporting issues. Without further understanding of the significance of these performance changes, researchers would consider these previous modeling attempts non-transportable. Reporting guidelines have been proposed and should continue to be pushed by journals. In the meantime, researchers should attempt to improve transportability assessments under the restriction of limited information and encourage collaborations to share data. Such These efforts can help emphasize the importance of transportability findings to the research community which will further encourage the desire to provide structured, sharable model designs.

CHAPTER 4

SIMULATING SOURCE DATA FROM PUBLISHED SUMMARY DATA

Data sharing can be challenging in medical research. Patient data contains a great deal of protected personal information that must be adjusted or removed before communicating values between researchers. For example, sharing of protected health information (PHI) in the United States is regulated based on a set of 18 identifiers as part of the Health Insurance Portability and Accountability Act (HIPAA). Similar laws exist in the European Union and sharing between countries requires additional paperwork. The protection of PHI often means data cannot be shared as part of the publication process and that extra work is required to share data retrospectively with other researchers.

As a consequence, summary statistics (such as mean, standard deviation, and group counts) of the source population are typically provided in publications. These values are meant to put the current population under study in context with other publications. Yet, inconsistencies in reporting often hinder the ability to compare the internal validation results of many models due to author-specific choices concerning what values and discretizations are used in tables and figures. In addition, these varied decisions make it difficult to construct an external cohort that can be evaluated against multiple models. Thus, many transportability assessments are unworkable because of the time required to process patient data to match previous models.

Unclear values could be clarified by requesting feedback from authors. Unreported values could also be sought retrospectively or access could be given to the original data to run analysis directly. Unfortunately, both of these options are largely impractical. Some authors may be unresponsive to inquiries concerning their results. In other cases, the communicating author may be a clinician

not completely familiar with the statistical process. The staff or students that performed statistical analysis may no longer be available to clarify values or reprocess data for new statistics. In extreme circumstances, the original data may have been mislaid and the steps to recreate the cohort may not be repeatable. Given these major difficulties in obtaining previous source data, two procedures for simulating a set of cases similar to the study population using only summary data were evaluated. I hypothesize that these simulated cohorts contain enough information to substitute for unavailable source information when retrospectively evaluating the variability of the c-statistic reported in published models.

4.1 METHODS

4.1.1 Evaluation dataset

To assess this application of cohort simulation, an evaluation was performed using previously collected data from the National Lung Screening Trial (NLST) [101,102]. The NLST was a large multicenter trial that enrolled 53,454 subjects with high lung cancer risk between 2002 and 2004. The trial was designed to randomly assign participants to screenings to evaluate the effectiveness of using either low-dose CT or standard radiography. Subjects were followed until 2009, during which data was collected on tumor development and deaths from lung cancer. The study was conclusive in demonstrating that low-dose CT was useful in reducing mortality in lung cancer compared with radiography. The NLST cohort is useful for secondary evaluation due to its previous application in the external validation of a lung cancer risk prediction model built using data from the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO). The primary PLCO risk model is a logistic regression model constructed in 2011 that was updated for comparison to NLST data in 2013 [103,104]. An additional Cox regression model is included in

PLCO publications. In addition to its use in PLCO evaluations, the large size of the NLST cohort helps to simplify the simulation evaluations.

Prior to simulation, the NLST cohort was randomly split into a training and testing group (N_{train}=40,000 and N_{test}=10,000) to mirror a split-sample internal validation. Split-sample evaluation is possible in this case because the original dataset is large. However, split-sample analysis is primarily used to simplify the interpretation of comparisons between the test data and simulations rather than having multiple results to consider in a cross-validation. The evaluation process could be adapted to run with cross-validation or bootstrapping if desired in future analysis.

The PLCO model was used to derive the features of interest for model building. Training cases were used to develop a logistic regression and Cox proportional hazards model of cancer risk. Training cases are only used to determine the source models to be evaluated using the test data and simulated data. In order to simulate cohorts based on the NLST test split, test data was also used to train logistic and Cox models. However, these models were only used to supply regression coefficients to calculate risk probabilities necessary for simulating outcomes. The application of these coefficients from the test cohort for event assignment is described in more detail in the section 4.1.2.

Test cases were also used to compute additional values employed as inputs of the simulation process that are commonly reported in the literature. Namely, the c-statistic as a test of internal validation, summary statistics of features in the population, and for Cox model examples a Kaplan Meier curve of overall survival. In addition to these common values, a covariance matrix of the test cohort was calculated for use in analyzing the importance of including feature correlations in simulation.

4.1.2 Simulation designs

The simulation approach in this work takes advantage of the information in the previous modeling publications to inform the designs used to sample values and create simulated cases. Creating each case in the simulation requires two steps: 1) sampling of values for each feature described in the model and 2) assignment of the event status of the case based on the simulated features.

Feature sampling is performed by making assumptions on the shape of a feature's distribution of values and assigning a sampling distribution to draw simulated values. Commonly assigned distributions are: binomial distributions for binary features, multinomial distributions for categorical features, and normal distributions for continuous features. The inputs for the sampling distributions are drawn from reported summary statistics. In addition, feature correlations can be considered using covariance information. In this chapter, simulations are designated as naïve or covariance simulations depending on the inclusion or exclusion of correlation information. Two types of sampling are performed depending on the simulation design used.

Naïve simulation with independent sampling

Independent sampling is a simple and straightforward way to generate feature and outcome values. Each feature is independently sampled from an assigned distribution, then transformed into a probability for sampling from an outcome distribution. A full pass of this sampling process creates a single simulated case. Cases are drawn until the simulated cohort reaches the desired size. As independent sampling does not take feature correlation into account, simulated cases may have significant differences from the source data. Yet, the independent approach may be preferable if these differences do not affect later applications, because it only requires summary statistics commonly reported in literature (mean, standard deviation, and probabilities derived from feature

counts). Therefore, it requires a minimal set of information and is less cumbersome to collect and report when simulating cases.

Covariance simulation with Gibbs sampling

Although feature correlation information is not readily available in current reporting, the influence of correlations may have significant effects on the similarity and variability of a simulated cohort. For this evaluation, a second sampling method was performing incorporating covariance information to determine the importance of correlation to the final performance of simulated cohorts. This method was performed by computing the covariance matrix of the source NLST test cohort. A covariance matrix is a generalization of the pairwise comparison of features into a multidimensional space, providing feature correlation information in the form of covariance. Use of the covariance matrix is attractive because it is easy to compute and is a summary of the original dataset that is more easily shared between researchers when data access is restricted. Features compared against themselves during pairwise tests define the variance of that predictor. Therefore, a covariance matrix provides data necessary to sample correlated features from a joint multivariate distribution of Gaussian models (e.g., Table 4.1).

	age	bmi	smokeday	smokeyr	smokequittime
age	25.18079	-1.68555	0.060203	19.44629	3.542062
bmi	-1.68555	24.89754	-0.08511	-4.40085	3.772599
smokeday	0.060203	-0.08511	0.015884	0.198613	-0.16917
smokeyr	19.44629	-4.40085	0.198613	54.30047	-18.051
smokequittime	3.542062	3.772599	-0.16917	-18.051	23.988

Table 4.1 Example covariance matrix of continuous features from the NLST test cohort. A full multivariate covariance matrix is included in Table A.1.

In the case of a simulation with only continuous features, summary means and a covariance matrix make it possible to easily draw values from a joint multivariate normal distribution. In many models, however, there are binary and categorical features. To efficiently sample from a

multivariate distribution with a mix of normal, binomial, and multinomial distributions, Gibbs sampling was performed.

Gibbs sampling is a Markov chain Monte Carlo algorithm for obtaining a sequence of values from a multivariate probability distribution [105]. Sampling is performed by taking draws while performing a "random walk" through the state space of the features. The Gibbs sampling process iteratively updates one feature at a time, while holding remaining features constant. The chosen feature is updated to a new value based on the assumed sampling probability distribution, which uses an updated probability conditioned on the states of other constant features [106]. After completing a full pass by updating each feature, a new random case is completed.

The state space for this Gibbs sampling application was defined by the summary statistics and covariance matrix of the NLST test cohort. For features following a normal distribution, these values are defined by the following matrices for the j^{th} sampled feature and the other remaining features from the set K, where $K = \{1, \dots, j-1, j+1, \dots, n\}$. The sampled feature vector, x, is partitioned for the j^{th} feature being estimated by:

$$x = \begin{bmatrix} x_j \\ x_K \end{bmatrix}$$

The mean, μ, and covariance, Σ, matrices are similarly split for the jth feature as follows:

$$\mu = \begin{bmatrix} \mu_j \\ \mu_K \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{jj} & \Sigma_{jK} \\ \Sigma_{Kj} & \Sigma_{KK} \end{bmatrix}$$

When sampling normal features, updated inputs for the mean and covariance matrix are obtained using conditional distributions computed with the *j* and *K* subsets of the mean and covariance matrices [107]. These updated values are used as input for sampling from the multivariate normal distribution. The multivariate normal distribution used for sampling a new value for feature x_j, given that the remaining features x_K are equal to *a*, is defined with the following updated conditional mean and variance distributions:

$$\left(x_j \middle| x_K = a\right) \sim N(\bar{\mu}, \bar{\Sigma})$$

$$\bar{\mu} = \mu_j + \Sigma_{jK}{\Sigma_{KK}}^{-1}(a - \mu_K)$$

$$\bar{\Sigma} = \Sigma_{jj} - \Sigma_{jK}{\Sigma_{KK}}^{-1}\Sigma_{Kj}$$

In order to sample values from binary and multinomial distributions, the conditional probabilities of possible choices must be determined. The equation for determining the conditional probabilities for a binary example is:

$$P\left(x_j = 1 \middle| x_K\right) = f\left(\begin{bmatrix} x_j = 1 \\ x_K \end{bmatrix}\right) \Big/ \sum_{b=\{0,1\}} f\left(\begin{bmatrix} b \\ x_K \end{bmatrix}\right)$$

where the probability of a state is determined from the density function of the multivariate normal distribution. Input probabilities for the previous equation were calculated using the *dmvn* function from the 'mvnfast' package in R. The mvnfast package supplies efficient C++ implementations of multivariate normal functions [108]. For example, the probability of a binary feature equal to 1 given other feature states, $P\left(x_j = 1 \middle| x_K\right)$, was drawn in R by setting values for the feature being drawn and calling the *dmvn* function as follows:

$$f = dmvn(\begin{bmatrix} x_j = 1 \\ x_K \end{bmatrix}, \mu, \Sigma)$$

These steps are easily generalized to multinomial features by calculating the conditional probability of each potential choice.

The assumed sampling distributions used in the covariance simulation are the same as those employed in naïve simulation. Gibbs sampling requires an initialization state before starting the sampling process. A single randomly generated case was drawn using the naïve simulation to serve as the initialization point. It is standard when using Gibbs sampling to perform an initial set of random samples using an assigned burn-in period. Burn-in is a specified set of initializing runs that allow for the sampling to reach a stable state space that is not correlated with the initialization state. Burn-in samples are discarded and are not used for the final cohort. Complete samples are then drawn to act as cases of the cohort using a thinning technique. Selecting within successive samples is not recommended for creating a simulated sample because neighboring samples will have some degree of correlation. Instead, a thinning parameter is set to select every i^{th} Gibbs iteration to create the simulated cohort. For this simulation, burn-in was set at 10,000 iterations and thinning was set to select every 10^{th} iteration as a simulated case.

Event assignment

Assignment of an event or outcome state occurs for each sampled case selected by the thinning parameter. Information from a trained model is required to transform the features into a risk probability that can be used for sampling binary outcome from a binomial distribution. The form of this probability distribution function depends on the model chosen to provide simulation inputs. Binary logistic regression and Cox proportional hazards models are relevant in this evaluation.

Binary Logistic regression

In the case of binary logistic regression, the probability of an event given a set of features, *P(Y|X)*, can be computed with a set of estimated feature coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_i)$, and features, $\boldsymbol{x} = (x_0 = 1, x_1, \ldots, x_i)$, with the following probability distribution function:

$$P(Y|X) = Prob\{Y = 1|X\} = \frac{e^{\boldsymbol{\beta x}}}{1 + e^{\boldsymbol{\beta x}}}$$

Therefore, sampling an outcome requires the reported coefficient values of a model and the features of a simulated case created with naïve or covariance sampling to compute outcome risk and make an assignment for logistic model examples.

Cox Proportional Hazards Regression

The corresponding estimation for Cox proportional hazards is more complicated. Outcome is sampled based on an estimate of risk, much like logistic regression. However, determining survival risk requires further sampling of an associated time to event which follows a hazard function. The probability of survival at a given time based on features, *S(t|X)*, is estimated using a cumulative baseline hazard function, $\widehat{H}_0(t)$, and the linear coefficients, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_i)$, and features, $\boldsymbol{x} = (x_1, \ldots, x_i)$, defined by the following survival probability function:

$$S(t|X) = e^{-\widehat{H}_0(t)e^{(\boldsymbol{\beta x})}}$$

The cumulative baseline hazard function is additional information that is not directly reported as part of Cox regression analysis. It can, however, be estimated based on characteristics of the cases in a dataset. One common approximation of the cumulative hazard is based on the Kaplan-Meier estimator, $\hat{S}(t)$, a non-parametric statistic that estimates the survival function of a cohort. When a

model is trained, Kaplan-Meier estimates are often derived as part of the Cox regression modeling. These values are not commonly published as a table of raw values, but instead are summarized using a survival curve plot. Using the survival curve of the NLST test cohort, graph digitizer software [109] was used to extract values of survival time and survival probability (Figure 4.1). This process mimics steps necessary to obtain these survival values from a published graph.

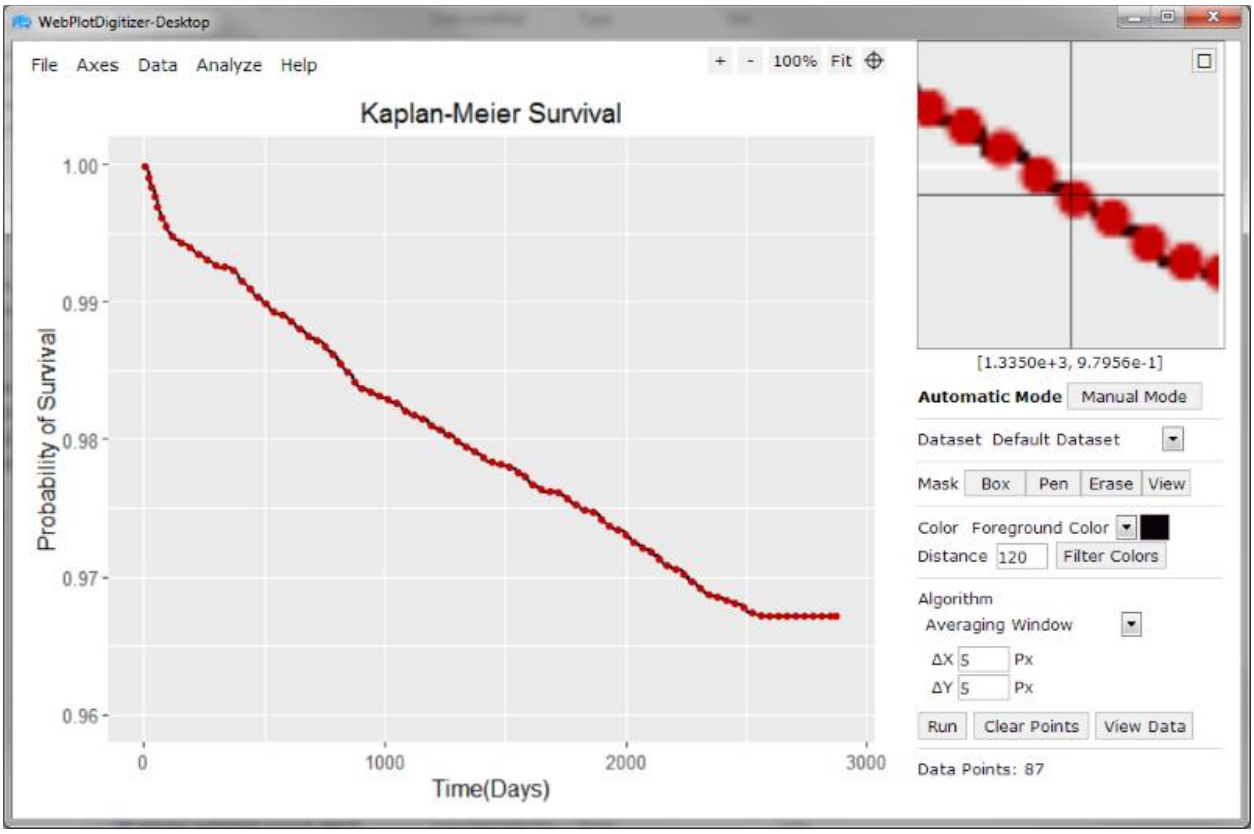


Figure 4.1 Kaplan-Meier survival curve for the NLST 10k test cohort (black) and extracted survival data points (red) using graph digitizer software.

The extracted probabilities of the Kaplan-Meier survival function were then log transformed to provide an estimate of the cumulative baseline hazard function:

$$\hat{H}_0(t) = -\log\left(\hat{S}(t)\right)$$

Survival times, $\boldsymbol{t} = \left(t_{s_1}, \ldots, t_{s_N}\right)$, were simulated using a method described by Bender et. al. that uses a set of transformations to the inverse of the cumulative baseline hazard function, $\widehat{H}_0{}^{-1}$ [110]. The resulting expression allows for sampling a survival time, t_s, by first sampling from a uniform random distribution on the interval of zero to one, *U~U[0,1]*:

$$T = \widehat{H}_0{}^{-1}[-\log(U)\mathrm{e}^{-\boldsymbol{\beta x}}]$$

The sampled time and previously sampled features are then used to estimate a survival probability, $p_s(t_s|\boldsymbol{x})$, that serves as input for sampling outcome.

$$p_s(t_s|\boldsymbol{x}) = e^{-\widehat{H}_0(t_s)e^{(\boldsymbol{\beta x})}}$$

A binary outcome was simulated using a binomial distribution for each of these individual survival probabilities to complete the simulated case data.

4.1.3 Evaluations

Simulated cohorts were created for logistic and Cox regression analysis using both the naïve and covariance simulation designs described above. Each cohort consisted of 10,000 simulated cases, matching the size of the source NLST test cohort used as a gold standard in comparisons.

Ability of simulated values to model source

The test statistic of interest in this examination is the c-statistic, a measurement of the performance of a model on a cohort of cases. As simulation is a random process, it is important to test whether the c-statistic of the original data source is significantly different from c-statistics computed from repeated simulations. Necessary values were first obtained from the NLST test cohort to perform simulations. Summary statistics, variances, logistic coefficients, and Cox coefficients are reported

in Table 4.2. The source NLST survival function is shown in Figure 4.1a. A simplified covariance matrix example for continuous features is shown in Table 4.1 and the full covariance matrix is provided in Table A.1 in the appendix. Next, 1000 simulated cohorts were created for both the naïve and covariance simulation methods in the logistic regression modeling setting (2000 simulations total). Gibbs sampling was performed using a burn-in of 10,000 and thinning of 10. The c-statistic was computed for each cohort and used to determine the distribution of c-statistics created with each simulation design. The c-statistic of the NLST test set was compared for significant difference against the simulated c-statistic distribution using a two-sided t-test. The same process was performed in the Cox regression setting for the naïve simulation design.

	Means	StdDev	Logistic Coefficients	Cox Coefficients
age	61.3974	25.1808	0.0530	0.0525
race=1	reference class	reference class	reference class	reference class
race=2	0.0438	0.0419	0.1898	0.2308
race=3	0.0193	0.0189	-0.0789	-0.0811
race=4	0.0031	0.0031	1.2064	1.1092
race=5	0.0044	0.0044	-6.8030	-3.0059
hispanic	0.0124	0.0122	0.0400	0.0552
educat	3.6466	2.3273	-0.0855	-0.0870
bmi	27.8757	24.8975	-0.0317	-0.0311
diagcopd	0.0516	0.0489	0.4315	0.4192
histcancer	0.0418	0.0401	0.7368	0.7032
famhistcancer	0.2212	0.1723	0.1855	0.1777
cigsmok	0.4856	0.2498	-0.1244	-0.1175
smokeday_nl	0.3986	0.0159	-2.0911	-2.0655
smokeyr	39.8328	54.3005	0.0472	0.0459
smokequittime	3.6414	23.9880	-0.0329	-0.0327
Intercept	-	-	-6.4666	-

Table 4.2 Means, standard deviations, and logistic and Cox coefficients for the NLST test cohort used for simulation.

Comparing bootstrapped variances

The simplicity of the current simulation process makes it likely that c-statistics of simulations would be different from the c-statistic of actual NLST cases. However, the current goal of these

proposed simulations is not to perfectly match the original c-statistic, but to provide a simulated cohort with similar variability to the original data when bootstrapping values. When assessing transportability, a target cohort's performance is compared to the reported c-statistic, but the confidence interval of this value is rarely provided. If the variation of bootstrapped values in simulated data match with real data, then the simulation can be used to compute retrospective confidence intervals to aid in external validity evaluations.

Assessing the ability of a simulated cohort to match bootstrapped variance required a comparison of the performance of bootstrapped c-statistics from naïve simulation, covariance simulation, and the NLST test cohort. In addition, since c-statistics of simulated cohorts were anticipated to be different, the comparison of these methods was further stratified by taking two simulations from the previously created 1000 cohorts. One cohort with c-statistic closest to the original NLST c-statistic and a randomly selected cohort were chosen. Bootstrap analysis was performed on these chosen simulated cohorts and the NLST test cohort, taking 1000 bootstrap samples and calculating the variability over the resulting 1000 c-statistics.

In most real world applications, the target dataset used to evaluate external validity is smaller than the number of cases in the source. With available data, it would be possible to bootstrap values at smaller cohort sizes to determine variability at the level of the target's size. Therefore, another comparison of the four simulation cohorts was used to evaluate the stability of bootstrapping subsamples of the original cohort size. If bootstrapped variation is similar between simulation and source data in the previous test, then subsample bootstrapping is expected to also be stable. The bootstrap subsamples used for evaluation ranged in size from 200-1000 in increments of 100 and from 1000-10000 in increments of 1000. The resulting bootstrapped simulation distributions at varying sample sizes were compared to the NLST bootstrap distribution using the Kolmogorov-

Smirnov test (KS-test). The means of the randomly selected examples were centered to the NLST c-statistic before the KS-test was performed. Multiple comparisons correction was performed using a Bonferroni adjustment to address the increased number of tests calculated across these many sample sizes.

4.2 RESULTS

As anticipated, the c-statistic of the NLST source cohort was significantly different from simulated cohorts created with naïve simulation (p=0.0017). Covariance simulation incorporating correlation between features showed a distinct improvement in similarity, but the NLST c-statistic was still significantly different from simulated examples (p=0.0304). The relationships between the source c-statistic and 1000 simulated cohorts for logistic regression is shown in Figure 4.2.

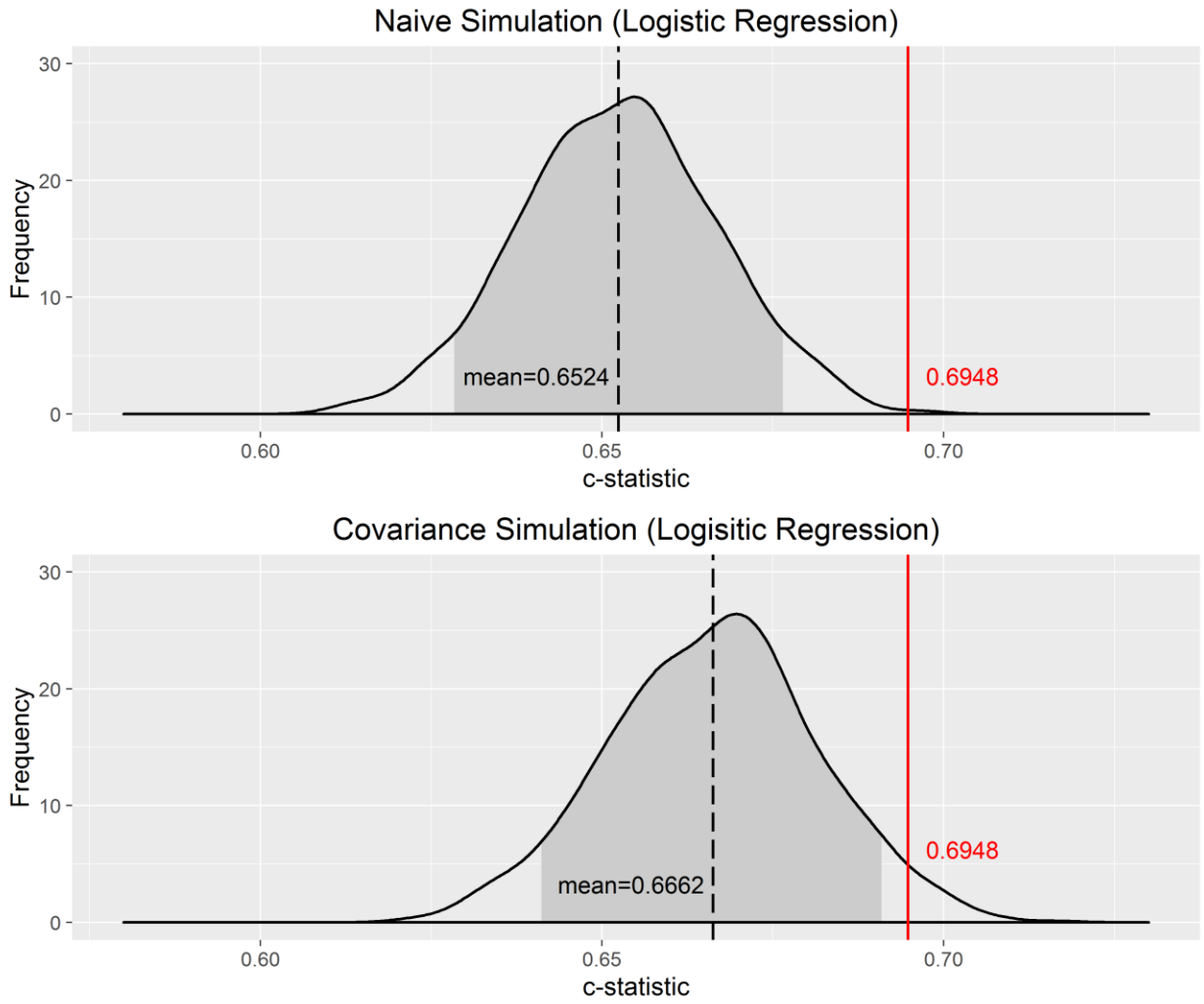


Figure 4.2 Comparison c-statistic distributions from naive and covariance simulation (mean c-statistic: black dashed line, confidence interval: grey) to source NLST c-statistic (red line) predicted with logistic regression.

Performance of naïve simulations for Cox proportional hazards regression were also significantly different from the source NLST c-statistic (p=0.007). The distribution of simulated c-statistics was similar to naïve performance in logistic regression (Figure 4.3). Therefore, the proposed simulation designs were not able to generate a cohort that is statistically indistinguishable from the original NLST data based upon calculation of the c-statistic.

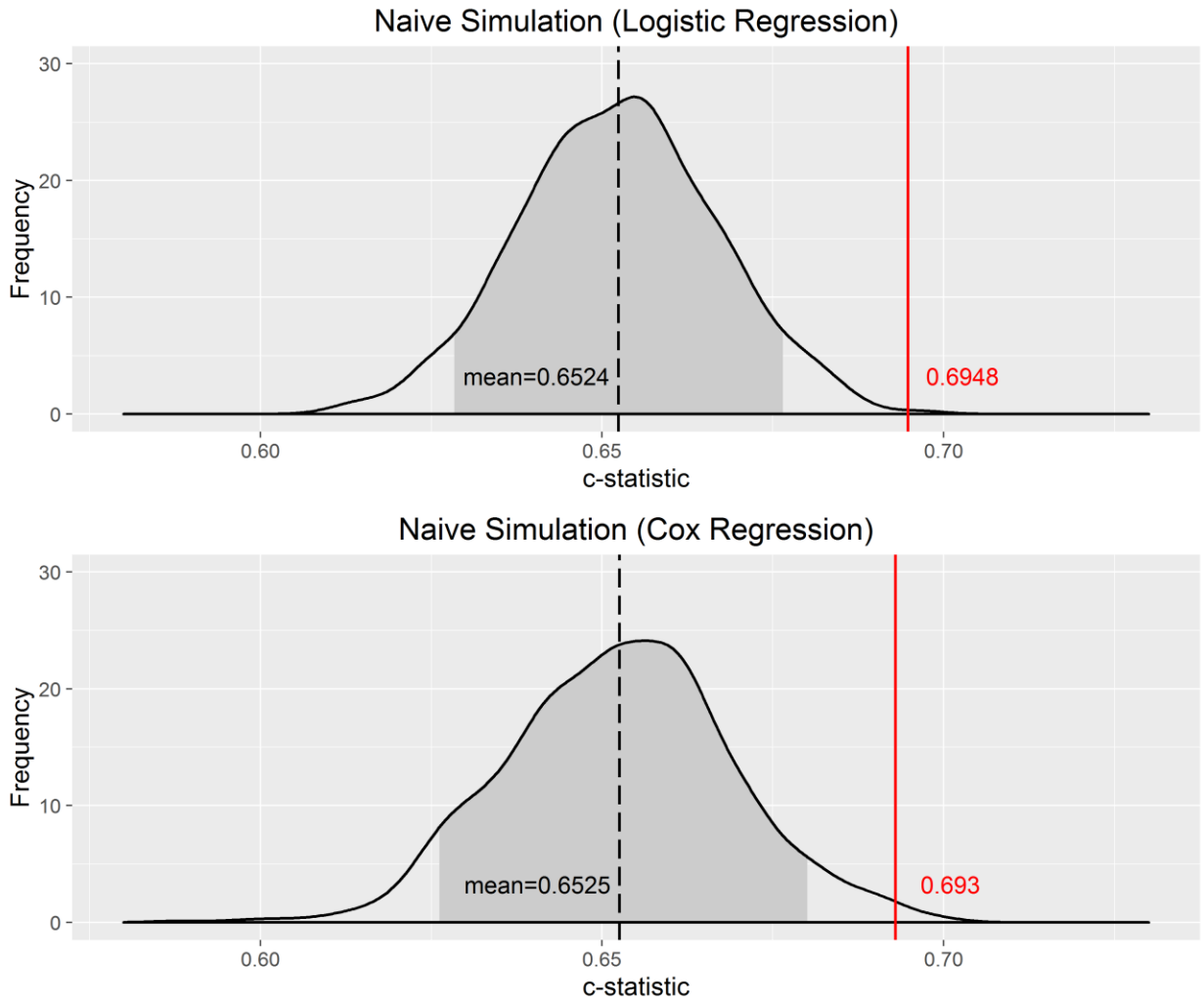


Figure 4.3 Comparison of naïve simulations for logistic regression and Cox proportional hazards regression. Simulated values (mean c-statistic: black dashed line, confidence interval: grey), NLST c-statistic (red line).

Results comparing bootstrap analysis of the four selected simulation cohorts to the source NLST dataset were more promising. Differences between the bootstrapped variance of simulated cohorts at sample size 10,000 was minimal (Figure 4.4). None of the four simulation cohorts were significantly difference from the source NLST bootstrap using the KS-test with Bonferroni correction for multiple comparisons (Table 4.3).

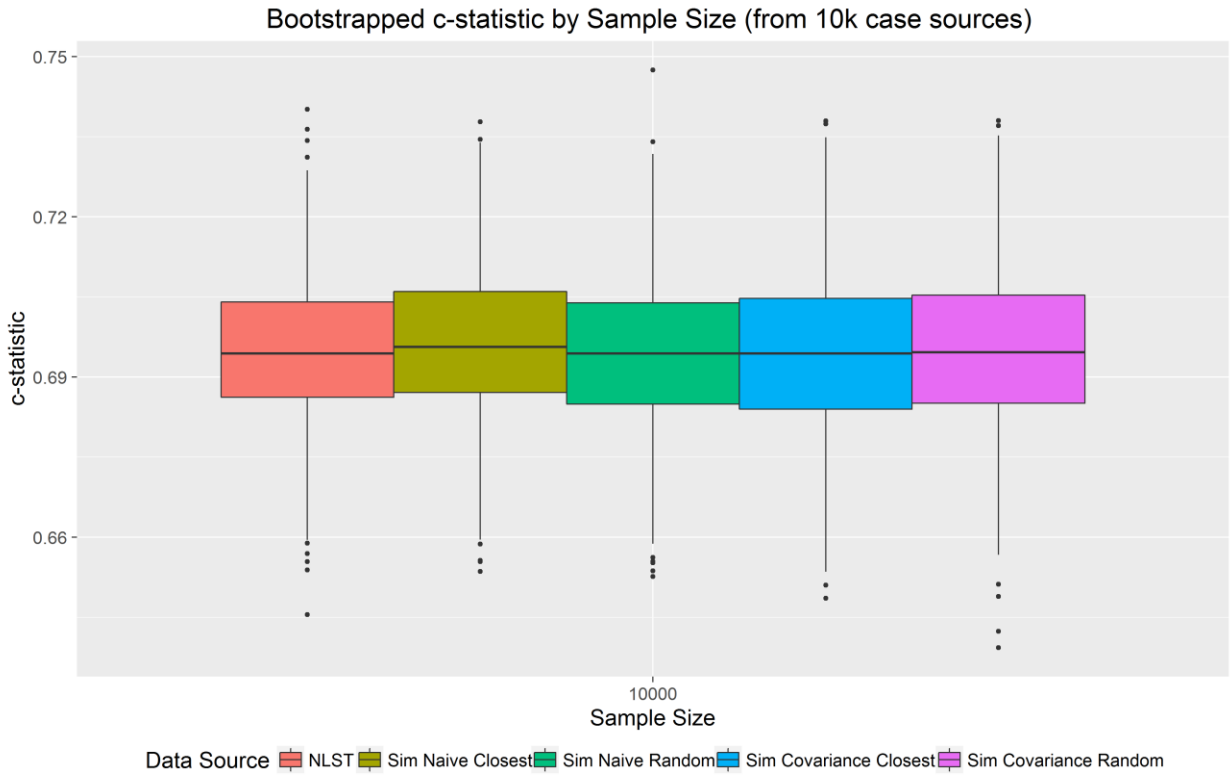


Figure 4.4 Bootstrapped variance of the c-statistic in source NLST and four simulated cohorts. (Random cohorts were normalized to the NLST c-statistic mean for graphical comparison)

Bootstrapping at smaller sizes showed similar results. C-statistic variances increased as smaller bootstrap samples were drawn, with larger increases seen as sample size decreased. An increase in the number of outliers and the disagreement in variance between source and simulation were seen at samples below 1000, but none of the bootstrapped differences were statistically significant at the subsample sizes (Figure 4.5 and Table 4.3). These bootstrap results are based on simulation cohorts for the logistic regression model design.

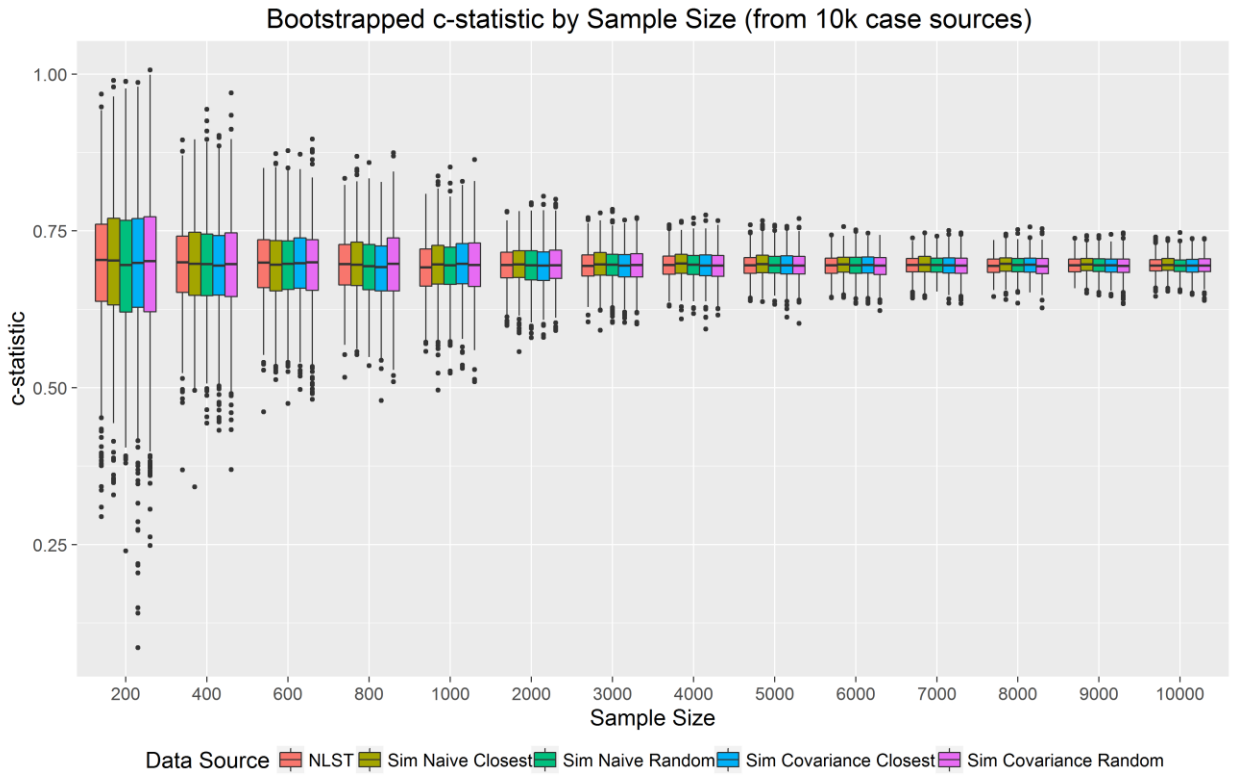


Figure 4.5 Comparison of bootstrap sampling at varying sample sizes.

Sample Size	Naive Closest	Naive Random	Covariance Closest	Covariance Random
200	0.370	0.069	0.263	0.055
400	0.648	0.573	0.263	0.164
600	0.536	0.828	0.859	0.723
800	0.181	0.341	0.062	0.006
1000	0.148	0.241	0.087	0.005
2000	0.288	0.219	0.097	0.466
3000	0.087	0.859	0.723	0.466
4000	0.033	0.500	0.181	0.121
5000	0.020	0.794	0.219	0.400
6000	0.370	0.573	0.723	0.241
7000	0.002	0.048	0.241	0.148
8000	0.001	0.723	0.011	0.164
9000	0.055	0.794	0.859	0.536
10000	0.017	0.241	0.008	0.134

Table 4.3 Kolmogorov-Smirnov test p-values for significant difference between NLST bootstraps and simulated bootstraps at various sample sizes. Bonferroni corrected significance value = 0.00089.

4.3 DISCUSSION

Simulation is a complex process comprising a set of assumptions when assigning distributions, determining sampling procedures, and collecting important values for setting parameters and initializing values. Many different items can be adjusted in an attempt to simulate values that are statistically similar to a source dataset. However, a great deal of time can be spent to maximize this similarity. The simulations in this chapter, by contrast, attempt to make use of a minimal set of information from the source environment. This approach is used, in part, because current reporting practice indicates that many values will be unavailable and are unlikely to be supplied even when requests are extended to authors.

The cost of this approach is that the resulting simulations are typically different from the source cohort they attempt to match. The c-statistic was used to judge difference in this work and over 1000 simulations, a significant difference was seen between the source data and the simulated statistics. The difficulty of significantly matching the original cohort with naïve simulation was an expected issue. Covariance simulation was included in an attempt to include enough information on the original cohort to closely match source. The addition of covariance to the simulation process was helpful in moving the resulting simulated c-statistics closer to the NLST source. This improvement suggests it may be useful to add a covariance matrix of the source data in publication results. This addition would be much less cumbersome than requiring researchers to spend time de-identifying data or completing paperwork before sharing information. However, the covariance simulation was still significantly different from the source cohort, indicating that additional adjustments would be necessary to most accurately represent the source.

Nevertheless, the intended use of the simulated cohorts in this work was to retrospectively calculate the variance of the source cohort c-statistic through bootstrapping. It was anticipated that

very closely simulating the source cohort was not a requirement for this capability. Comparison of two simulated cohorts, one with very similar performance and another selected at random, demonstrated that the simulated cases were able to accurately represent the variance of the source population even when those cases did not have a c-statistic performance comparable to the original data. In addition, there was no significant difference between simulating values using the naïve and covariance simulations for this purpose. Since naïve simulation is sufficient for computing the variance of the c-statistic retrospectively, this method provides a simple way for researchers to obtain extra information about discrimination performance while only requiring a minimal amount of reported information.

A set of comparisons for bootstrapping subsamples of a cohort (i.e., smaller sample sizes) indicate that simulated cohorts can be of additional use for computing variance assessments at many sample sizes. This capability allows for variance statistics to be generated flexibly to match the size of a target cohort being considered in transportability assessments. Such an adjustment to a confidence interval can provide useful context when comparing small target validation datasets to source statistics calculated with large cohorts. In this analysis, the calculated variance remained stable down to sizes 10%-20% of the size of the original cohorts. Values for sample sizes between 200-1000 had much higher variability, causing the confidence interval to be more susceptible to the random selection of cases. It is unclear if the larger differences between source and simulation examples for sample sizes below 1000 is related to sample size relative to the original source cohort (i.e., less than 10% of source) or if this variability is common to all samples below 1000. The community should be wary of the potential for higher error rates, but in this analysis none of the differences were statistically significant. Additional bootstrap samples might be able to stabilize the measurements in future analysis.

4.3.1 Limitations

A large, highly curated dataset from lung cancer research was used to provide inputs for this analysis. Few published models are built on such substantial collections of subjects. Consequently, some of the bootstrap findings may not be as stable when applied to the simulation of smaller source cohorts. Performing this same analysis with example datasets of varying size would help to clarify if the simulation process can be applied widely to model validation.

Logistic regression and Cox proportional hazards models were considered in this evaluation due to their common use on the NLST dataset and also in medical prediction models. Many additional model options are available and appropriate derivations of probability distributions are required as part of simulating values for other model designs. The use of simulation for calculating the performance using the c-statistic can be generally applied to models of many types, but there may be additional difficulty employing this technique on complex models used in machine learning without additional development and research.

The ability of the simulations to significantly match with the source data was shown to be inconsequential for this approach where the primary goal was performing bootstrap analysis. However, other metrics may be more heavily influenced by the ability of a simulation to create a cohort that matches closely to the original data. Future research could test this approach in other validation metrics such as the calibration intercept, calibration slope, and integrated discrimination index.

4.3.2 Conclusion

Cohorts simulated with naïve and covariance methods can be used to retrospectively asses the bootstrap variability of the c-statistic. In the current implementation, however, they should not be

used for analyzing the central tendency of the c-statistic which should instead be inherited from reported values. Simulated cohorts also appear to be useful for calculating variance at sample sizes smaller than the original data, but future work is needed to determine how accurate these tests are in smaller cohorts and at small sample sizes.

CHAPTER 5

CATEGORIZING MODEL TRANSPORTABILITY WITH SIMULATED COHORTS

Disease models are an important step in exploring the relationships between multiple predictive features. Each individual model can provide some insight into the predictive power of distinct feature combinations in different situations. Thus, disease models have become frequently used to help solidify understanding of the measurements and observations that describe disease and outcome relationships. Some predictive disease models are currently trusted to provide risk estimates as seen in risk calculators for cardiovascular disease [111–113] and lung cancer [104,114,115]. Predicted risks can be informative to both patients and clinicians when making lifestyle and treatment decisions. Models must be properly validated in cases outside of the development environment before being widely accepted for these risk calculation, prediction, and decision-making tasks.

Most models are not evaluated for transportability to other settings. In previous chapters, a number of limitations were discussed concerning why evaluations are difficult to accomplish. Most assessments also take an all or nothing approach, where failure to validate implies that the model only has use in the original environment. Under this view, even if all models were validated, a majority would be set aside as unusable when an external validation failed. It would be useful to categorize models further into levels of transportability in order to consider what problematic models might updated for continued use.

In Chapter 3, four published Cox regression models for brain cancer were reviewed and applied to an external UCLA cohort. The predictive performance on the target UCLA data, measured with the c-statistic, was compared to the reported performance of the original studies. C-statistic

performance was lower in the target data for all of the applied models, indicating that the models were generally non-transportable. Building from this initial analysis, the four brain cancer survival models were evaluated with an extended simulation analysis used to interpret the level of transportability achieved by each model.

5.1 METHODS

Given the decreased performance of the brain cancer models in Chapter 3, it might appear that none of the models are appropriate for additional use. However, model performance is known to vary across different test datasets and the external c-statistic values might be lower by chance. Determining if a c-statistic decrease is significant is not possible without information concerning the confidence interval on the original internal validation. In addition, the cohort sizes of the source and target environments are not identical. At a smaller sample size, the variability of the c-statistic will be larger, so larger differences in c-statistic performance should be expected to occur by chance.

To get a better perspective on the variation in the source cohort, the naïve cohort simulation method described in Chapter 4 was used to determine the bootstrapped variability of the source c-statistic. Simulations were built for each published model and 1000 simulated cohorts were generated matching the sample size of the unavailable source cohorts (Helseth=516 cases, Michaelsen=225 cases, Gutman=68 cases, Kumar=312 cases). Based on the results in Chapter 4, a single cohort with minimum difference in the c-statistic from the reported value was selected from the 1000 simulations. These selected simulated cohorts were used for bootstrap analysis. Determining the significance of performance changes is a first step in determining if a model is trivially transportable or should be assigned to a more restrictive transportability level.

Calibration assessments were also performed for each model. Calibration tests a model's ability to predict outcomes at rates that match with observed outcomes. The observed rates in a target cohort may be different enough from the source that the predictions generated for new cases are systematically over or underestimated. Detecting such bias in combination with the previous c-statistic examination can help determine when validated models should be assigned to the trivial or calibration adjustment transportability levels.

All simulations and statistical analysis were coded in R (v3.1.3) and run with RStudio (v0.98) [91,92].

5.1.1 Simulation assumptions

Simulated cohorts were constructed for each model based on published values, following the process for naïve simulation of a Cox regression cohort presented in Chapter 4. Summary statistics of patient characteristics and regression coefficients of the Cox regression model were obtained from published text or tables in three of the four models. Regression coefficients were available for the Gutman model, but summary statistics were not. Gutman et al. were able to provide access to the source dataset, but the original case selection could not be perfectly duplicated (see Chapter 3.2.3). Therefore, summary statistics were drawn from a subset of 68 cases that were similar by not identical to the original training set. These collected coefficients and summary statistics are provided in Table 5.1 with the sampling distribution assumptions used to sample each feature.

Helseth				Michaelsen			
Features	**Statistics**	**Hazard Ratio**	**Distribution**	**Features**	**Statistics**	**Hazard Ratio**	**Distribution**
Sex	M: 0.411 F: 0.589	1.44	Binomial	**ECOG**	0: 0.608 1: 0.304 2: 0.088	Reference: 0 1: 1.22 2: 2.06	Multinomial
Age	63.7 (13.26)	1.02	Normal	**Corticosteroids**	Yes: 0.733 No: 0.267	2.06	Binomial
ECOG Score	Low: 0.812 High: 0.188	2.13	Binomial	**Age (by 10)**	59.2 (12.40)	1.31	Normal
Tumor Location	Unilateral: 0.853 Bilateral: 0.147	2.31	Binomial				
Primary surgery	Biopsy: 0.089 Resection 0.911	2.72	Binomial				
Gutman				**Kumar**			
Features	**Statistics**	**Hazard Ratio**	**Distribution**	**Features**	**Statistics**	**Hazard Ratio**	**Distribution**
KPS	40: 0.015 60: 0.176 80: 0.721 100: 0.132	0.972	Multinomial	**Tumor Site**	Central: 0.045 Other: 0.955	2.336	Binomial
Tumor Major Axis Length	78.95 (19.62)	1.016	Normal	**Tumor Location**	Parietial: 0.564 Other: 0.446	1.516	Binomial
pCET	0.025 0.195 0.505 0.815	7.745	Multinomial	**Radiation Dose**	Yes: 0.776 No: 0.224	2.026	Binomial
				Chemotherapy	Yes: 0.715 No: 0.285	0.435	Binomial

Table 5.1 Extracted values and assumptions for simulation process. Summary statistics and hazard ratios were derived from published text and tables. Distribution choices were based on published feature discretizations.

Survival times and probabilities were extracted from Kaplan-Meier overall survival figures using offline graph digitizer software [109]. A screenshot was taken from PDF copies of the Helseth and Michaelsen papers. The survival figure for the Kumar model was downloaded from the full text version of the online paper. For these images, points were generated along the curve using "Automatic Mode" using the averaging window algorithm set to draw points at steps of 5 pixels in X and Y directions. Prior to running the automatic process, the foreground color of the curve was set and a masking box excluding the figure axes and text was drawn. For Michaelsen and Kumar, points were automatically generated along secondary curves in the figure. These extra points were manually removed before extracting values.

An overall survival figure was not available from the Gutman paper, which only included Kaplan-Meier plots stratified by tumor enhancement categories. Therefore, Kaplan-Meier estimates were derived using the 68 cases from the original cohort. Extracted values for each cohort are plotted in Figure 5.1. The raw values for survival time and survival probability are included in Table A.3 of the appendix.

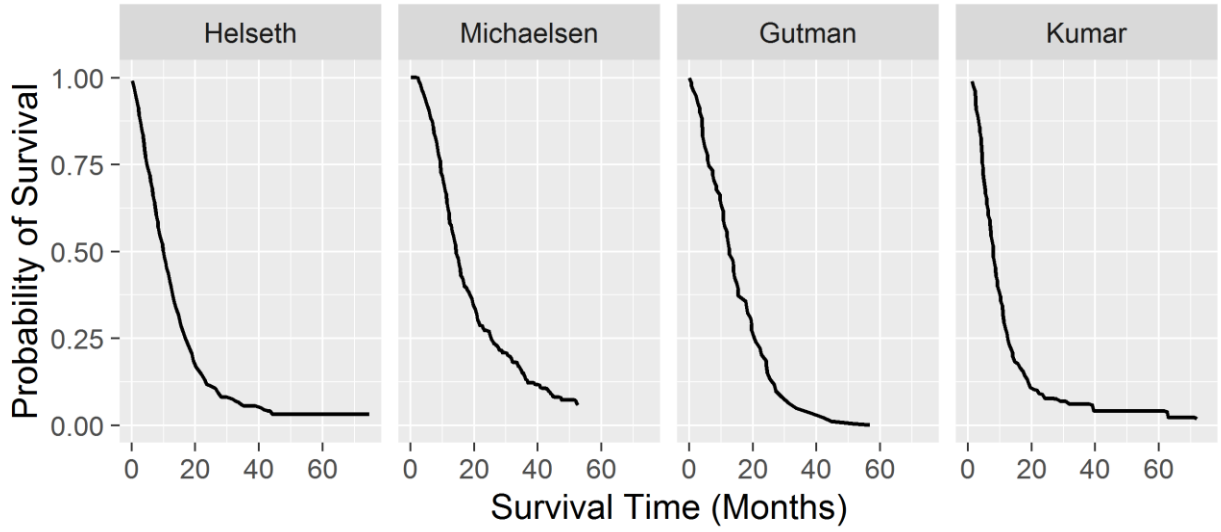


Figure 5.1 Extracted survival curve values from published figures (Helseth, Michaelsen, and Kumar) and replicated data (Gutman). Values were extracted from published figures using WebPlotDigitizer software (examples shown in Figure 4.1 and Appendix A).

5.1.2 Bootstrapped c-statistic assessment

One simulated cohort was drawn from a set of 1000 simulations for the Helseth, Michaelsen and Gutman cohorts based on the mean c-statistic closest to the reported value. Kumar et al. did not report a c-statistic value for their model. Since this ground truth was unavailable, two simulated cohorts were drawn to examine the interval ranges at extremes of model performance. A low and high c-statistic value, ± two standard deviations from the mean c-statistic of the 1000 simulation set, were used to select two simulated cohorts. Bootstrap analysis was performed to create 1000 bootstrap samples for each of the selected set of five simulations. C-statistics were calculated for

each bootstrap sample using the *survConcordance* function of the 'survival' package [116] to determine an associated confidence interval. External c-statistics of the target cohort were then compared against these intervals.

Next, 1000 bootstrap samples were drawn at size N=125, equal the sample size of the target UCLA cohort. The bootstrap evaluation at N=125 for the Gutman cohort is actually oversampled using this technique (Gutman cohort N=68). This oversampling was included as a comparison point against the other models, but may be a biased representation. C-statistics were computed for each bootstrap using the same steps described above to determine confidence intervals. External c-statistics were compared to this second confidence interval to determine if different conclusions would be drawn with subsample bootstraps.

5.1.3 Calibration assessment

Calibration was evaluated for each model by fitting a Cox proportional hazards model regressing survival time and outcome against the computed linear predictor values of the target cohort. This regression results in a calibration coefficient, β_c, that summarizes the ability of the linear predictors of each model to match with target outcomes. A calibrated model will predict target cases accurately and have a calibration coefficient equal or close to one. Values above and below one are indicative of the overall under and overestimation of the linear predictor values obtained with a source model. A simple re-calibration can be achieved by using the calibration coefficient as a scaling factor for the linear predictor. Cox calibration models were fit using the *cph* function from the 'rms' R package [93]. Calibration was also assessed by reviewing calibration curves created using the *val.surv* function from the 'rms' package.

5.2 RESULTS

Based upon the bootstrap analysis of the simulated cohorts, the decreased performances of the Helseth and Gutman models on target UCLA cases were not significant when compared against the source and target sized samples. Therefore, these models are reasonable candidates for transportability as they are able to predict UCLA cases at rates not significantly different than predictions in the source environment. However, the Gutman model is built using a dataset smaller than the target cohort and the overall variability of the c-statistic from bootstrap analysis is high. The confidence interval of the N=68 bootstrap includes a prediction rate of 0.5, indicating that the model does not perform better than random chance in some cases. When bootstrap oversampling was performed at N=125, the confidence interval decreased and was close to excluding the 0.5 rate. Though the oversampling test can be biased, this demonstrates that a larger sample would be more effective at indicating the predictive capabilities of the Gutman model. These comparisons are shown in Figure 5.2.

Michaelsen's performance in the UCLA cohort was significantly decreased. The decrease was also significant in the subsample bootstrap evaluation. These significant decreases indicate that this model is unable to transport directly to the target cases. The UCLA and Michaelsen cohorts should be reviewed in more detail to determine if there is an underlying difference that could be adjusted. When comparing results in Kumar, no final determination can be made because the c-statistic of the source data is not known. As shown in Chapter 4, the naïve simulation design is unable to represent the source cohort with enough accuracy to derive this value retrospectively. In Figure 5.2, c-statistics at the 95% confidence extremes of 1000 simulated examples were used to draw cohorts as an alternative approach to reviewing the performance significance. In the lower c-statistic case, Kumar performance changes are not significant. However, the high c-statistic

simulation was significantly different and a final determination of transportability is not clear. This finding highlights the importance of providing the internal validation analysis including the AUC or c-statistic during publication in order to enable future validation reviews.

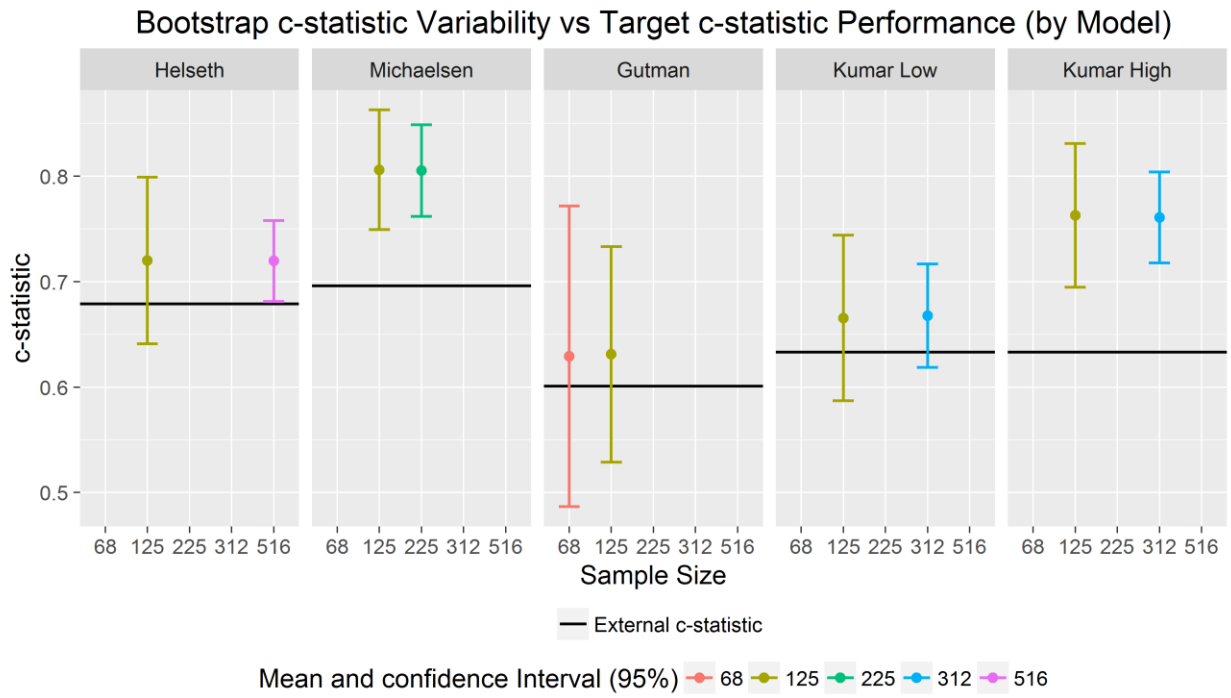


Figure 5.2 Bootstrap results of simulated cohort data at source and target sample sizes, displayed by source model. Two simulations (low and high c-statistic performance) were considered for Kumar as a c-statistic value for the source data was not reported.

Calibration analysis demonstrated that all of the models had some degree of miscalibration. The Helseth model showed a minor overestimation of predicted probabilities compared to observed rates in the UCLA cohort (β_c = 0.922). A calibration plot of prediction at median survival time (506 days) makes this apparent (Figure 5.3), as the calibration curve is below the perfect calibration reference line. Since target performance was not significant in the discrimination analysis and re-calibration looks reasonable, the Helseth model can be assigned for calibration adjustment.

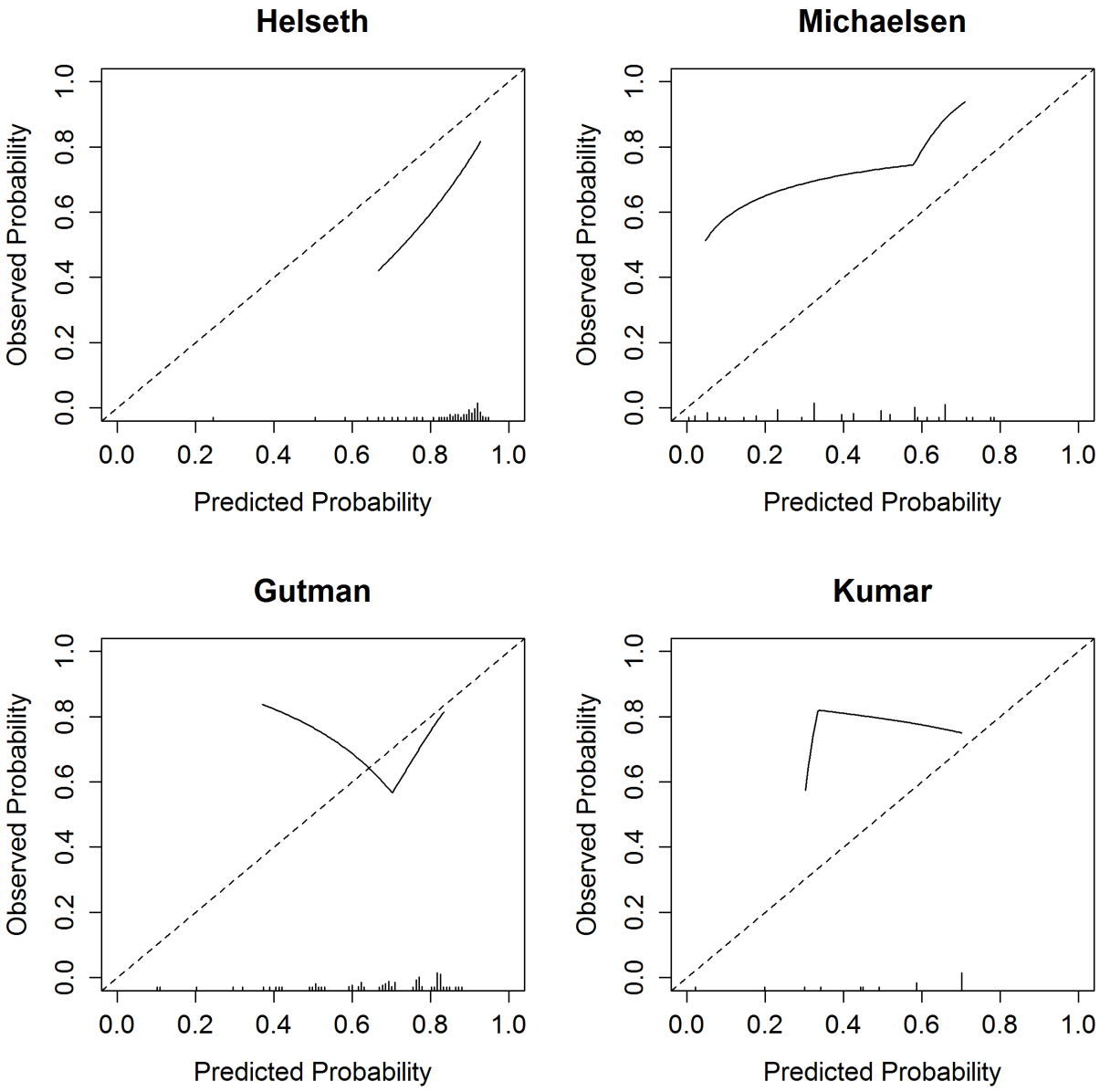


Figure 5.3 Calibration plots for each model at median UCLA survival time (506 days)

Michaelsen showed a strong and consistent underestimation (β_c = 0.608). The calibration plot shows that this underestimation was worse for cases with low probabilities of survival. A re-calibration would be useful to put observed and predicted rates in line in the target setting, but the significant performance decrease means that adjusting the calibration alone would not be enough

to make the Michaelsen model transportable. Therefore, the Michaelsen model cannot be assigned for calibration adjustment. Instead, it should be considered for partial adjustment. If further investigation into the source and target cohorts does not yield potential adjustment targets, Michaelsen might ultimately be assigned as a non-transportable model.

The calibration plot of Gutman crosses the perfect calibration line. Some cases with higher observed outcomes appear to be assigned very low predicted probabilities as seen by a bowing in the calibration plot. The calibration coefficient is also low (β_c = 0.576) and it is unclear if a re-calibration would be effective with many highly observed outcomes being assigned low probabilities. The small sample size of the training data for the Gutman model may be one explanation for these poor predictions. As the performance was similar between the simulated source and target cohorts in bootstrap analysis, we can consider the model for calibration adjustment. However, more data should be used to reconstruct the model. Therefore, the model should be revisited in case the external validation changes with more data and the model needs reclassification to a partial or non-transportable level in the future.

Kumar's overall calibration (β_c = 0.84) was better than Michaelsen and Gutman. However, the calibration plot shows a bowing point where many high probability cases were assigned low probability predictions, similar to the calibration plot for Gutman. This apparent mixing of predictions would suggest that a simple scaling with the calibration coefficient will not provide an effective re-calibration. Recalibrating in this fashion does not change the order of assigned predictions and will not improve discrimination performance. Without a reported c-statistic for the source cohort, it is difficult to judge the transportability of the Kumar model. If the target c-statistic was determined to be high, then the model would likely need partial adjustments given the significant c-statistic difference and mixture of calibration values. If the difference was found to

be insignificant, then the model might be able to be calibration adjusted with further review of the calibration issues

	Helseth	Michaelsen	Gutman	Kumar
C-statistic Performance	Not significant	Significant	Not Significant	Inconclusive+
Calibration Coefficient	0.922	0.608	0.576	0.84
Calibration Curve Analysis	Overestimation	Underestimation	Mixed	Mixed
Transportability Determination	Calibration Adjustment	Partial Adjustment	Calibration Adjustment*	Inconclusive+

*Table 5.2 Summary of Transportability Evaluation Findings. * Larger sample size could update transportability determination. + Missing source c-statistic makes assessment inconclusive.*

5.3 DISCUSSION

Predictive model transportability is too strictly defined into usable and non-usable groupings using current methods that compare source and target c-statistics. Incorporating the variability of the c-statistic from internal validation can more completely describe the significance of performance changes when applying source models to target cohorts. Calibration testing is also suggested for understanding source and target predictions with better context in the external validation process [11,35,37]. By combining information from both types of evaluation, a more detailed interpretation can be made of a given model's transportability. Most models may not be directly applicable to target locations, but many could be strong candidates after appropriate adjustment.

In the four models above, c-statistic variability was calculated retrospectively by performing bootstrap analysis using simulated cohorts created with a naïve simulation method. All four models were able to be assessed with this method, though final interpretation was not obvious for the Gutman and Kumar models. Gutman's interpretation suffered due to the small sample size of the source cohort. Kumar could not be fully assessed because the c-statistic performance in the source data was not available, even after repeated requests were extended to the author. However,

simulated cohort analysis appears useful for determining the significance of performance changes in target data when these obstructions can be avoided.

The calibration analysis in this chapter was based on current practice [12]. A calibration model was fit to the linear predictor to test the overall difference in predicted probabilities and observed outcomes. This calibration fit can supply a slope and intercept value in logistic regression, but in Cox regression only a calibration coefficient is estimated. In either modeling case, a calibration plot should also be reviewed to visual assess if target locations have different outcome rates. Yet, interpretation of calibration plots appears to be a rather subjective, often reducing to a general claim of calibration/miscalibration. This issue is apparent in this analysis as some degree of miscalibration occurs in each validation, but interpretation of the effects on re-calibration adjustment is unclear. Being unable to make a strong assessment, the classification of models between the calibration adjustment and partial adjustment levels of transportability remains difficult as seen in the uncertain classifications of the Gutman and Kumar models. Calibration adjustment cannot correct for poor performance in a target cohort and, in general, significant c-statistic decreases point towards partial or non-transportable situations. Nevertheless, future improvements to calibration assessments might prove useful to clearly define the line between the calibration and partial adjustment categories. A bootstrap analysis akin to the c-statistic comparison might be used to attain this goal in future work.

Overall, the addition of the simulation cohort analysis appears to provide a more definitive decision in regards to discrimination between source and target (Table 5.2). For example, the small sample size of the Gutman cohort could have been used to screen out this model before testing transportability. But, simulated cohort evaluation made this fact clear; variability was too large at the Gutman sample size to use the model for prediction. In the case of the Helseth and Michaelsen

models, the ability to compare against intervals at two sample sizes made the appropriateness of each model more obvious. These additional comparisons are therefore helpful to further define transportability groups beyond simple transportable and non-transportable divisions. Detailed separation of models will help guide future frameworks searching to test specific models for adjustment or immediate application. Reporting bootstrapped c-statistic variability by estimating directly from source data rather than a simulated cohort would be ideal and is highly recommended for future publications.

5.3.1 Limitations

The models chosen for this analysis were selected based upon the inclusion important modeling details (i.e., clear descriptions of summary statistics, feature selection and discretization, and multivariate Cox regression coefficients). However, the initial development of disease models may be performed to explore the predictive impact of features of interest rather than targeting direct prediction. For example, Kumar et al. were interested not only in multivariate predictors, but also in the influence of specific features such as treatment regimens [88]. Some of the current hurdles to obtaining internal validity values from papers may be explained by these alternative goals of modeling. Therefore, additional analysis of this methodology to interpret transportability groups may be warranted in other disease modeling domains where values are missing more/less frequently from publications.

Many cohorts used for analysis are limited in size. In previous analysis of the simulation process using lung cancer data, confidence intervals appeared less stable at smaller sample sizes. However, the potential influence of this contribution could not be assessed with brain cancer data as source cohorts were not readily available. Brain cancer cohorts are rarely larger than 500 cases as seen by the cohort range in this analysis (68-516 cases) and publicly available data from the Cancer

Genome Atlas (529 cases). Additional analysis should be performed to confirm if sample size is a significant issue in the simulation and bootstrapping task.

The comparisons used in this evaluation were based on currently suggested validation practices [10–12,37]. The field has settled on specific practices such as the analysis of discrimination and calibration. However, incorporating additional tests to consider case-mix [41], discrimination indexes [21], and decision curve analysis [31] may further inform the interpretations of transportability assignments. Other existing simulation approaches for survival data might prove useful for investigating the influence of specific population changes on these statistics. In addition, methods targeting partial model adjustments are currently limited. New evaluation methods such as causal graphical analysis (Chapter 6) might help pinpoint the exact features degrading transportability and allow targeting model updates. The methodology presented here is one piece of a large set of potential improvements to transportability assessment and should be assessed for use when these other items are included in analysis.

5.3.2 Conclusion

Using variability measurements of the c-statistic from simulated cohorts and calibration tests of the target cohort, model transportability was more specifically classified for two brain cancer models. In the remaining models, sample size and reporting issues made classification less clear. Overall, when adequate information is available from previous work, assessments of transportability seem to benefit from the inclusion of discrimination variability.

CHAPTER 6

EXPLORING CAUSAL TRANSPORTABILITY

Substantial amounts of time, money, and effort are exerted to run scientifically sound experiments to determine the efficacy of medical therapies and natural progression of disease. The primary focus of such studies is to determine the ability of internal validated findings to impart knowledge about clinical questions. Over many studies, findings can be combined to assess the generalization of this knowledge to the population as a whole. The combined findings from many experimental studies and collected observational data can also be used to approach generalized knowledge about broad disease classes. Scientific findings are often used to inform the development of multivariate models for predicting disease treatments and progression. Accurate predictions of risk and outcomes have potential to aid clinical decision making. However, examination of models beyond internal applications in the original population has yet to become common practice. External validation is a necessary test to explore an individual model's ability to transport to new environments.

External validation (transportability) assessments are becoming more prevalent, but a majority of validation results show decreased performance in target cohorts. As seen in the previous chapters, classification of model transportability can be extended beyond a transportable and non-transportable paradigm. Defining more specific classification levels can help to determine what adjustments would be appropriate for correcting differences between source and target locations. For example, scaling of a model's linear predictor values is a straightforward approach for models in need of calibration adjustments. These models have similar performance in source and target locations, but consistently over/underestimate the correct probabilities for patients. When models

have substantial discrimination decreases, calibration alone cannot adjust the model in a meaningful way for target application. Such models need to consider the potential differences in the cohorts in order to define if partial adjustments to specific features can adjust predictions into more accurate ranges.

Partial adjustments are an important area of study in transportability because it is preferential to reuse as much information as possible from previous studies. This is particularly true for models trained with cohorts larger in size than a target domain; these models have reduced variability due to access to more data. Consequently, models trained with more data can often avoid overfit compared to models trained with a target cohort of smaller size. Another example where partial adjustment can be beneficial is in the collection of expensive or difficult features. High costs or limited resources might make it impossible for some features to be collected in all locations. In contrast, understanding of population differences can also indicate features that are required for proper model predictions. Few approaches currently exist to consider which partial adjustments are appropriate before classifying a model as non-transportable. Transportability theory, introduced by Pearl and Bareinboim, is a novel method from the artificial intelligence community that has potential for evaluating when partial model adjustments are appropriate [45].

Transportability theory is based on graphical analysis and assessment of causality assumptions of a model to determine if particular features meet the proper constraints to be transported from the source population in combination with specific target information. For example, a physician in a rural setting might wish to apply the results of a clinical trial conducted at a large research hospital to decision making for patients under their care. The trial findings can be understood in the context of a causal graph, per Figure 6.1, as a treatment (A) with effect on a patient outcome (B), with additional measured factors such as clinical history, imaging, or genetics (C, D, E).

Transportability theory allows a researcher to identify potential confounding evidence between variables (represented by dotted lines) and population differences (indicated by square nodes, S1 and S2) that are known or believed to exist between two cohorts. The influence of these constraints can be used to determine what data from the trial environment can be applied to the rural patients in a principled way. For instance, the physician may not have enough genetic information for the population to build a model on their own; applying transportability can help ascertain whether the genetic information collected in the trial can be transported (i.e., reused) to the local group (and if not, under what other graphical circumstances such data transport would be valid). Similarly, differences between the hospital and local populations (e.g., demographics) can be accommodated via transportability. In general, if all existing differences can be accounted for, then model variables can be considered transportable to the new population through a partial transport utilizing source values in combination with updates from target information. If substantial differences cannot be addressed, the reviewed model would likely be classified as non-transportable and other models should be considered instead. In depth discussion concerning the properties of graphical and causal models that contribute to transportability theory are included in background section 2.3.

Transportability theory is investigated in this chapter by 1) developing a limited Bayesian belief network (BBN) disease model of glioblastoma multiforme (GBM) targeting overall survival; 2) reviewing the ability of transportability theory to encode information concerning the potential issues of transporting source findings; and 3) testing the transport of partial data between different source and target cohorts under the assumption of an evaluated causal graph. Publicly available data from The Cancer Genome Atlas (TCGA)[117] initiative of the National Cancer Institute (NCI) were used to create source and target cohorts based on the multi-institutional nature of the TCGA dataset. These source and target examples were applied to train and test results of the model

under different partial transportability settings. The results from applied TCGA data help to demonstrate the use of transportability for partial model adjustments, as well as helping to describe the complexities involved in this practice. The graphical model used in this paper is simplified to provide an opportunity to examine the characteristics of a disease model and transportability theory without the complications that a large set of predictive variables may add to the process.

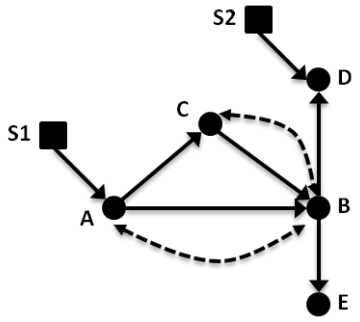


Figure 6.1 Example causal diagram for lung cancer treatment. Variables: (A) treatment; (B) tumor progression; (C) tumor biopsy gene expression; (D) clinical history; and (E) CT imaging findings. In this causal diagram, solid circles represent standard variables (such as in a Bayesian belief network), and solid arrows between these nodes represent causal relationships. Dashed arrows/arcs indicate confounding influences between two variables that may exist when considering other populations. Selection nodes, shown as squares, provide a means to sub-select or filter a given variable so that the evidence is comparable between two groups.

6.1 METHODS

6.1.1 Evaluation dataset

A number of multi-institutional efforts now exist to establish observational databases, supplementing experimental datasets. Two efforts focused on building databases for GBM research are The Cancer Genome Atlas (TCGA) and the Repository for Molecular Brain Neoplasia Data (REMBRANDT)[117,118]. TCGA is a public database containing primarily clinical and genomic (copy number, DNA methylation, gene expression, single nucleotide polymorphisms) information for 20 different types of cancer. In addition, TCGA is a part of ongoing efforts to make radiological and pathological images more readily available in cancer research. The

REMBRANDT database is focused specifically on data obtained for brain gliomas (astrocytoma, mixed glioma, oligodendroglioma, GBM) with a limited number of unmatched non-tumor controls. When considering the design of a simplified graphical model, the TCGA and REMBRANDT datasets were examined to consider the full scope of features of interest in GBM prediction. A large number of clinical, imaging, and genetic features are now collected for research (Table 6.1), but current predictive models have not revealed a particular feature set best adapted to outcome predictions. The final graphical model for this evaluation uses a set of features that overlap with those seen in these multi-institutional databases.

Variable	
Demographics	Total radiation dosage
Presenting age	Other drug name
Family & social history	Other drug Frequency
Environmental exposure	Other drug Dosage
Tumor location	Steroid drug name
Tumor size	Steroid frequency
Tumor grade	Steroid dosage
VEGF	Karnofsky performance score
EGFR VIII	Other performance score
PTEN	Tumor volume (on imaging)
TP53	Necrosis imaging finding
MGMT	Contrast enhancement imaging finding
DNA methylation	Non-contrast enhancing region
Chemotherapy drug name	Tumor multi-focality
Chemotherapy frequency/dosage	Edema volume (on imaging)
Number of chemotherapy cycles	Mass effect
Type of surgical resection procedure	Satellites
Extent of resection	ADC map (imaging)
Type of radiation therapy	Time to progression (TTP)
Radiation therapy fractionation	Time to survival (TTS; death)

Table 6.1 Partial list of collected variables from among two multi-institutional data sources, TCGA and REMBRANDT.

Data for this analysis was obtained from the TCGA public data repository. A total of 579 cases are available in the TCGA database with clinical information. TCGA cases with available clinical and genomic data were evaluated with variable selection and preprocessing. Cases were first removed

if there was prior evidence of a glioma. Overall survival models require new cases of brain cancer for studying the length of survival. Cases with prior evidence of tumor are unlikely to contain necessary baseline survival time data. This selection reduced the number of available cases to 544. Due to the small number of variables in the Bayesian network, complete case analysis was considered to reduce the amount of missing data and imputation. KPS values were blank or unavailable for 143 cases. For this Bayesian analysis, censorship could be problematic and some censored cases were removed to target only cases observed until death or past a median survival cutoff. The final count of selected cases for analysis was 346.

The selected TCGA cohort was discretized prior to evaluation using the categories in Table 6.2. The dataset was divided into three source and target subsets based on contributing hospital locations in the TCGA population. Source cohorts included multiple institutions and were used in model training; target cohorts included cases from a single institution and served as patients from a target location. Three target splits were made (large, medium, and small sets of cases) to examine the effects of prediction when varying amounts of target data were available. The final TCGA locations chosen to serve as targets for analysis were Hospitals 2, 6, and 19. These target cohorts contained 84, 65, and 18 cases respectively. All remaining cases were combined to create source cohorts (262, 281, and 328 cases respectively).

Variable	Range/Categorical values
Age	0 (<40); 1 (40<65) ; 2 (65<80) ;3 (>80)
Karnofsky Performance Score	20,40,60,70,80,90,100 - 7 Category Assignment
Metagene Score	0 (Low, Score <= 0), 1 (High, Score > 0)
Survival past median	0 (No); 1 (Yes)

Table 6.2 Selected model variables and discretization choices.

6.1.2 Bayesian model design

The base Bayesian model used for transportability theory review (Figure 6.2a) contains four variables; 1) a demographic variable, age; 2) a cognitive assessment variable, Karnofsky performance score (KPS); 3) a genetic variable, a 9-gene metagene score derived from Colman, et al. [119]; and 4) an outcome variable, overall survival (survival past median of 12 months). The choice of features in the Bayesian network was based primarily upon common use in past models and the ability to create a simple, yet plausible network. Age and KPS features, for example, were chosen due to their predictive significance in previous GBM models [69,89,90]. Similarly, overall survival was the most common outcome variable used in previous GBM prediction studies.

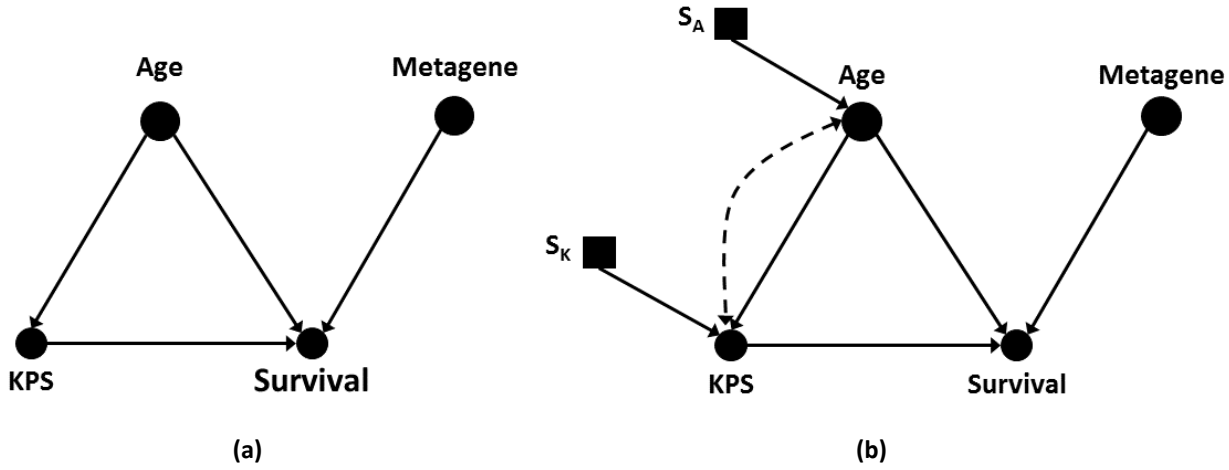


Figure 6.2 Example causal diagram for (a) GBM survival prediction and (b) the same causal diagram of GBM with links and nodes representing expected confounding information and population differences for variables. In the diagram, solid circular nodes represent observed variables; while square nodes indicate selection nodes controlling for population differences. Causal links are represented with solid lines with directional arrows. Bi-directional dashed lines indicate variables linked by confounders. The selected observational variables are Patient Age (Age); Karnofsky Performance Score (KPS); 9-gene metagene score (Metagene); and Patient Survival at Population Median (Survival). Unique selection nodes for Age and KPS are shown as S_A and S_K.

A genetic variable was included based upon the growing interest in genetic prediction variables for cancer. Genetic testing is not available in all clinical locations and large samples are difficult to collect in current clinical practice. The TCGA and REMBRANDT studies were conducted, in part, for obtaining these values for general review and future application to external studies.

Therefore, a genetic variable is a suitable example of an item from a model that would benefit from transport between source and target locations. A number of recent papers discuss the predictive potential of genetic markers such as O6-methylguanine-DNA-methyltransferase (MGMT) methylation and tumor protein 53 (TP53) gene expression. For example, significant up-regulation of MGMT expression when treated with O6-alkylating agents such as temozolomide (Temodar) demonstrate potential benefits for patient survival [70,120–124]. Similarly, up/down regulation of TP53 factors into cell apoptosis; reduced rates of apoptosis are characteristic of many types of cancer and can contribute to large growth rates of cancerous cells [125].

For this evaluation, the genetic feature used was a metagene score derived from nine gene expression values measured in the TCGA dataset. The selection of these significant genes and the metagene scoring technique originated from previous work by Colman, et al. [119]. Colman's score is calculated by summation of the weighted expression levels of the nine genes of interest. Meta-gene scores for each patient were discretized into high and low score classes before model training (Table 6.2).

6.1.3 Transportability theory discussion

In the example GBM model in Figure 6.2a, a set of four variables and their causal connections are shown. The example network comprises no connective links other than the direct causal connections derived from literature review. Given the set of causal assumptions made in the proposed graph, differences that may exist between source and target populations should be considered. Expected differences are encoded as additional nodes or bidirectional links between existing variable nodes. Additional nodes relate to assumed or known population differences. Bidirectional links indicate expected sources of confounding. Some form of confounding and population difference are likely in most graphical networks and must be dealt with when

problematic to the transportability task. An example graph with a considered set of these issues is provided in in Figure 6.2b. A model may not be transportable unless certain constraints can be met either by transportability rules (Chapter 2.3: do-calculus/d-separation) or evidence that the added differences in model connections can be removed or adjusted without affecting the outcomes. Thus, the goal of the transportability theory can be seen as an attempt to map between the messy real-world graph with lots of disruptions in Figure 6.2b and the ideal causal graph in Figure 6.2a to enable the transport of information.

Unobserved and confounding variables.

Consider the dotted connections in the example GBM network (Figure 6.2b). Bidirectional dotted links in the causal graph represent latent confounding variables. A confounding connection represents interactions mediated by unmeasured variables (i.e., data that could not be observed). In this example, the added link between age and Karnofsky performance score might denote the belief that unmeasured complex biological interactions can explain the interaction between KPS and age, masking direct effects of the stated causal assumption. For example, KPS is derived from an examination of a patient's current mental and physical status. This status derives from a combination of the current symptomatic state of disease in the patient and some mix of other prior disease. The patient's symptoms might be tied to a damaged hip from osteoporosis, causing a decreased score due to lost mobility, or symptoms could be tied to a past neurological event such as a stroke, causing a decreased score due to stroke related aphasia. These kinds of effects would mask the attempt to measure age's causal effect on KPS values caused exclusively by GBM involvement. If proof exists in the literature that such an interaction is common between features, additional variables might be required for the model to correct the confounding before a proper transportability assessment can be made. The impact of the particular confounding example used

above would be minimal, as physicians are trained to account for prior disease during KPS evaluation. Therefore, the confounding link could be removed as there is a reasonable belief that clinicians are considering past injuries when quantifying the KPS value.

Population differences.

In addition to confounders, consideration must be given to the population differences that exist in the collected data. For example, chemotherapy treatment may vary between two locations depending on physician treatment preferences/experience, hospital practices, and availability of drugs. The number of patients treated might influence the predicative capabilities of the model depending on how the features were included and their causal links to other features. Selection nodes in Figure 6.2b represent potential cohort differences in age and KPS scoring. Age often varies depending on the type and the location of hospital where data is collected. Available training in performance scoring, overall experience of evaluators with patients in the domain, and standard variability of measurements taken by different examiners can have effects on KPS scoring if these differences are systematic. Adding new selection nodes can change how information flows between variables as the network is examined with transportability theory rules. Unless variance between the populations can be explained with evidence or well supported assumptions, stratification or re-estimation of variables connected to selection nodes may be warranted.

Theory application

Application of Pearl and Bareinboim's work with transportability theory is possible for a given graph following the consideration of difference with relevant links and nodes. A set of algebraic rules called the do-calculus [45,48] (Chapter 2.3) enables a formal mathematical statement to be derived for the causal graphical structure with considerations for known controlled information. Analysis of the function can determine what elements of information are transportable based upon

fixed variables and the relationships, confounders, and selection nodes in the graph. This process is performed with do-calculus by breaking defined causal links between nodes based upon forced experimental constraints. Graphical analysis of the separated nodes via d-separation, and the front- and back-door criteria, can help determine which variables of the model are identifiable. Identification entails the evaluation of the graph edges that remain when observational data is used to set a variable to a specific state, and then determining when the network is not directly affecting the transportation of findings. Thus, when a causal graph is not identifiable, its findings are non-transportable. For example, KPS can serve as a surrogate measure for imaging findings of brain tumor growth, which are influenced by population differences. As long as the KPS variable can be established as conditionally independent of population differences, information from the feature can serve as a replacement for the imaging information using the front door criterion of d-separation. This process can unblock a situation where imaging findings may not be available and their causal relationship is obstructing successful model application. More information on do-calculus and d-separation can be found in Chapter 2.3 or Pearl and Bareinboim's work [45,48]. Similar examples for situations involving back-door paths and bidirectional counterfactual edges can be drawn for the examples above; a number of examples are provided with more detail in the Pearl and Barenboim's discussions [46–48].

6.1.4 Network evaluation

An example of the utility gained by controlling for disruptive factors in the causal graph is provided using a Bayesian belief network evaluation. Cohorts built from TCGA data were applied to the proposed Bayesian network to demonstrate the use of a partial adjustment. After reviewing the obstructed causal graph, the transportability of specific features can be run based on how the

assigned differences are cleared from the graph in causal analysis (such as in Figure 6.2a with no remaining obstructions).

To perform this evaluation, the Bayesian belief network predictive model for GBM was tested using custom code in MATLAB (version 7.10.0, MathWorks, Inc, Natick, MA). Source and target cohorts were built by splitting the TCGA dataset by contributing location; one TCGA participating location was held out as a target cohort while all remaining data formed the source set. A breakdown of the number of cases in each source and target cohort are provided in Table 6.3. Breaking cases into these groups allowed the source cohort to act as a previous evaluation site for model construction and training. Target cohorts served as external locations with new patients in need of prediction. For this analysis, three splits were performed targeting contributing hospitals in the TCGA dataset with large (Hospital 2), medium (Hospital 6), and small (Hospital 19) sample sizes.

TCGA Hospital ID	Target Size	Source Size
2	84	262
6	65	281
19	18	328

Table 6.3 Sample sizes of source and target cohorts created by splitting TCGA data.

Four model combinations were created by varying the training and test cohorts from the three source-target splits. The four combinations were: Source versus Source (SS), Target versus Target (TT), Source versus Target (ST), and Transported Source versus Target (TrST). Each consideration describes the Training-Test setup used for modeling as summarized in Table 6.4. Leave-one-out cross-validation was performed on test cases to determine the prediction rate of the models. Mann Whitney U-tests were used to test for significant difference between prediction classes of the model.

Model	Training Data	Test Data
SS	All Source	Source
TT	All Target	Target
ST	All Source	Target
TrST	Age: Source, KPS: Target, Metagene: Source, Survival: Source	Target

Table 6.4 Description of training and test data used in the model considerations. Test data is cross validated using the leave-one-out cross-validation method.

As a gold standard evaluation, SS and TT examinations represent the construction of a model using data from source and target locations with no knowledge of the other cohort. This emulates the current state of practice where research locations frequently build models for local use and perform internal validation. In this case, past models are not taken into account and transportability is not assessed between locations. In the ST comparison, the external validity of the source model is highlighted by predicting new cases from the target cohort. The ST examination applies information directly from source to target and in the case where all source and target variables are similar, model performance would be similar indicating a case of trivially transportability. All model probabilities are obtained from the source cohort with no training input; target cases provide no model training and are simply tested with the source model. This examination stage parallels with previous transportability investigations of this work. The chance of no differences existing between the source model and target are rare, as seen in previous results in other chapters, and performance was expected to suffer. Finally, the TrST split examines a partial transport adjustment where the probabilities of the KPS variable were retrained by the target data instead of the source. The assumption from the causal analysis was that the source KPS data differed greatly from the observations of target patients. Therefore, retraining this feature with target information can adjust for difference but does not influence the effects of the other features in the model. The trained information of the remaining variables was transported from the source model. The expectation was that the joint use of information from the source and target cohorts will outperform the ST

method because proper partial adjustments have been targeted using the causal transportability considerations. Table 6.4 provides a full breakdown of how source and target data were assigned for training the model in each of the four model combinations.

6.2 RESULTS

6.2.1 Network evaluation

A total of 346 TCGA cases were available for analysis and were split into three source-target cohorts. Each source-target cohort was then used for model training, followed by testing using leave-one-out cross-validation (LOOCV) across the four described training variations (SS, TT, ST, TrST). LOOCV was chosen in order to maximize the number of cases available for the training steps, as the available sample sizes for the target splits were all small. Final results of the analysis of the trained and tested Bayesian networks are presented in Table 6.5. Overall model performance was modest (SS and TT). Directly applied source models showed reduced discrimination in two of the three comparisons (ST) and all showed reduced performance compared to the complete source constructed model (TT). When partially adjusted, two models showed similar and improved performance, while one model continued to have reduced discrimination (TrST). An in depth discussion of these findings for each source-target combination are discussed below.

	Model							
	SS		TT		ST		TrST	
Hospital 2 (262,84)	0.69	(2.7E-08)	0.76	(1.5E-05)	0.74	(1.1E-04)	0.76	(1.7E-05)
Hospital 6 (281,65)	0.72	(1.6E-11)	0.68	(0.007)	0.63	(0.056)	0.63	(0.059)
Hospital 19 (328,18)	0.71	(4.7E-12)	0.94	(0.004)	0.68	(0.248)	0.94	(0.004)

Table 6.5 Leave-one-out validation results of transportability analysis. Values represented are Area under the curve (AUC) and Mann-Whitney U p-value for significant difference between survival prediction classes. Three hospitals in the TCGA dataset are compared to demonstrate the effects of target cohort size. Karnofsky performance score (KPS) was held out as missing/unmeasured data in this model.

6.2.1.1 Prediction using source training data (SS and TT)

Internal validation using LOOCV showed moderate performance for the models trained with source (SS) and target (TT) data. Performance measures for this analysis were calculated using area under the ROC curve (AUC). Discrimination ranged from 0.69-0.72 (LOOCV) for SS models and TT models ranged from 0.68-0.94. While target trained models managed to outperform source models, these performance gains were related to the small sample sizes in the TT combination. The models were more likely to be overfit than their SS counterparts, as noted by the exceptional performance in the target of Hospital 19 (AUC = 0.94). The larger p-values from Mann-Whitney testing also indicate that the SS model performances were more likely to be appropriate. The SS and TT model performances served as gold standards of discrimination performance from internal validation when comparing to subsequent model applications, ST and TrST.

Decreased model discrimination when applying to the target implies the need for partial adjustment. These adjustments were attempted in the TrST case and can be compared to both the SS and TT values to understand the final improvement of a partial adjustment made in light of transportability theory.

6.2.1.2 Prediction using outside training data (ST)

A comparison of the SS with ST performance is equivalent to the analysis process described in Chapter 3. The trained source model is applied for predicting target location values. Measurement of the applied AUC discrimination with LOOCV analysis provides the ST comparison results. Lower AUC values indicated that the training data from the source cohort did not create a model that was transportable for predicting cases in the case of Hospital 6 and 19 splits. Results would be similar and externally valid when populations are similar and variable differences are minimal. Only in the large target cohort split, Hospital 2, was performance the same or improved. A

significant differentiation (p=1.1E-04) between prediction classes was detected for this application by the Mann-Whitney U-test supporting the external validity of this source-target example. For the other splits, transportability did not hold for the source model on outside data as seen by the decreased AUCs and p-values that do not reach significance (0.056 and 0.248).

When comparing to the target trained models, a decrease in performance was seen for all cohort combinations. The TT score served as a theoretical performance maximum where no source information was supplied. AUC values dropped by 2.7%, 5.3%, and 26% respectively for the three hospital splits compared to TT models. Therefore, by comparing ST discrimination performance to the target trained model performance, it can be seen that more accurate discrimination is possible (though potentially overfit).

6.2.1.3 Prediction using transported probabilities (TrST)

Adjusting models with decreased performance by taking appropriate information from both the source and target locations is intended to improve the original model by replacing variable and distribution information to improve overall predictive performance in the target. In the TrST model, KPS values were assumed to vary between source and target populations. Therefore, KPS probabilities were trained using target patient data while other variables use transported data from the source. In two applications of the TrST model, an improvement of AUC over the application of the source trained model, ST, was observed. Performance for Hospital 2 and Hospital 19 improved, matching the prediction accuracies seen in the TT model validations. These values meet or exceed the performance indicated by the SS validation results as well. In the case of Hospital 19, the smallest target cohort, the Mann-Whitney U test statistic changed from being insignificant (ST p=0.248) to significant (TrST p=0.004). These improvements suggest that data from the KPS variable in the target was able to better model local cases. Hospital 6 showed no improvement in

accuracy between ST and TrST attempts. This result suggests that there was actually no significant underlying difference between the source and target KPS data for this split as was assumed for analysis. When these values are similar between populations, no new information is added by retraining information using a target cohort. In the Hospital 6 case, partial adjustment was not warranted for the KPS feature. Additional confounding links and selection nodes in the causal transportability analysis should be considered to search for other partial adjustment targets. However, if no other differences are easy to define or test in a causal graph, then a given model should be classified as a non-transportable case. Detecting and classifying non-transportable cases is important to using transportability theory effectively for model adjustment and improving model accuracies.

6.3 DISCUSSION

Given the general decrease in performance observed when applying models to external populations, it is important to consider mechanisms for adjusting models before classifying them as non-transportable. Examination of the specific differences between source and target cohorts can be informative for combining source and target information into a more effective model. These types of partial adjustments may also be informative for future model building if many complete source and target cohorts can be combined for retraining a model with a larger set of predictive features. Transportability theory is a novel method for considering the influence of feature and cohort differences on the ability to transport partial information from a source to a target.

The transport of probabilities for prediction from a source model to a target cohort imparted an increase in the predictive power of the model over an original source model for two of the three sites in this evaluation. These results demonstrate at a basic level the potential power of

transportability theory to assess a model and determine appropriate variables for transport. Consideration of possible confounding and population difference in new cases is important when determining whether external validity applies to a model.

In this chapter, core concepts of transportability were reviewed in the context of a GBM model to investigate the potential contributions of the method to partial model adjustment. In the examined scenario, feature information for a metagene biomarker based on gene expression was usable in a model applied between two cohorts. Target information was used to retrain another problematic feature, KPS, and performance improvements were observed in two of three cohort scenarios. This result indicates that the retrained target information can work in concert with transported source information to improve the predictive capabilities of a model with previous performance decreases. Performance was even equal to rates observed when completely retraining a full model from scratch. Transporting data in this way can also be used to extend models when data is unmeasurable or unavailable. In addition, the causal transportability analysis has potential to be reviewed prior to data collection. Subsequently, it can prove beneficial by reducing the number of variables that must be measured at the target location, saving time and money. For example, in the Bayesian network evaluation (Figure 6.2) age, metagene, and survival information were transported from the source model in the final comparison (TrST). In doing so, the target was not required to provide information to estimate probabilities for these variables, only KPS data was collected from the target.

Application of transportability theory to increasingly complex model designs will be important to expand the utility of this approach. For instance, expanding a model with the addition of a new variable can cause a number of complications in the causal network not fully discussed in this chapter. Examples of additional considerations that might need review when inserting new features

into the network include: how the variable is measured, what other variables it is causally connected to, how the addition affects the previous assumptions of the links between features (i.e., changes to independence), and if measurement of the variable introduces new differences between populations. As model complexity increases, it appears that the number of considerations may become difficult to appreciate. Yet, inclusion of model information and graph designs to modeling experiments may be a way for investigators to clarify the model's experimental target, assumptions, and design decisions. Published graphs might then act as a template for outside researchers to test the model's transportability against target datasets.

While the effectiveness of transportability theory is demonstrated by this examination, future work is needed to allow transportability theory to be accessible to many researchers. An honest evaluation of confounding arcs and selection nodes is required to faithfully consider the causal nature of the relationships described in a graph. In this way, a researcher can use transportability theory to demonstrate that conflicts have been considered and that assumptions made when attempting to claim findings are externally valid. However, the method to identify this information is not well defined and would currently require additional input from experts or literature review.

Causal transportability assessment is one potential method for improving models in need of partial adjustment. Other options should be explored further to define what adjustment path is simplest for researchers. The transportability theory evaluation in this work makes it clear that partial adjustments are possible and will be useful to applying models more freely rather than discarding the important feature information contained in source models.

6.3.1 Limitations

One limitation of this discussion and evaluation is the overall simplification of the problem. A probabilistic model of GBM should include a number of variables covering clinical, treatment, imaging, and genetic factors. The presented model only examines two such facets (clinical and genetic) and minimized the feature set to four specific nodes for the prediction task. This simplification is necessary for an introductory discussion of transportability theory, but is unrealistic when compared to model designs with larger feature sets seen in regression analysis. Models with 10-20 features will likely complicate the ability to determine and analyze the causal relationships of the model. Future analysis must examine the computational sophistication of larger disease models in order to find tractable solutions to transportability questions. Further efforts must be made to provide descriptions of the theory that are accessible to a broader audience with an interest in testing external validity. This will require communication between computer scientists, statisticians, and informaticians to balance the descriptive language used and ensure that papers related to transportability can be published more widely.

Low prediction rates for the current models are potentially tied to the simplistic model representation chosen to facilitate discussion. Other statistical models have reported higher rates of discrimination in GBM survival prediction (AUC 0.81-0.82) [69,87], though some models perform at a similar level to the proposed Bayesian model. In addition, a limited number of confounding and population differences were considered. Examination of location differences in the TCGA dataset might elicit other variables that influence partial adjustment, allowing for improvement for a split like Hospital 6, where KPS updating did not improve discrimination. Providing a robust examination of factors that can disturb the external validity of data is necessary

to support claims made when completing causal transportability evaluations. Overlooked or ignored confounders could have influenced the capability of the model to perform in this work.

Finally, the use of a 9-gene metagene score to summarize gene expression values allowed for simplification of the Bayesian network, but may not be the best feature design for prediction. The metagene score was used in this work based upon Colman's evaluation that found the grouped feature was more predictive, but this finding itself has not been externally validated in a meaningful way [119]. An examination that treats each gene as a variable of the model might yield improved results in some cases. Many other gene expression rates are also measured for these populations and more statistical examination of the predictive involvement of these genes may suggest changes in the included genetic factors.

6.3.2 Conclusion

The core concepts of transportability theory were discussed in relation to a simple Bayesian model for GBM overall survival. The evaluation and discussion in this simplified context provides an understandable introduction to transportability and the examination of how poor assumptions and population differences may discourage outright dismissal of models from applied use. Increases in AUC when testing a partially adjusted model with information from both source and target cohorts (TrST) improved two of three examined test cases. These improvements are indicative of the utility of transporting partial information from a source model. Additional work in the area of transportability theory can act as a useful tool for defining partial adjustment targets. The complexities of causal analysis require future work, however, before determining that this process is able to outperform other methods for identifying where source findings are valid for transport and use with target populations.

CHAPTER 7

CONCLUSION

The increasing ubiquity of electronic health records, along with improved computational methods in statistical analysis and machine learning, have driven an explosion of predictive modeling for medical decision making. Evidence-based practice would benefit from well-constructed and validated models that provide probabilities of risk or suggest treatment paths for patients. It is well understood that internal validation should be tested when presenting potential models. However, external validation to other related environments is rarely attempted. Subsequently, few models are validated for decision making, despite an expanding supply of modeling publications.

7.1 SUMMARY OF RESULTS

In this dissertation, an investigation of the current state of transportability (external validation) analysis noted many of the current obstructions to providing effective evaluations. Clear reporting is particularly important to allow future researchers to understand the design choices of a given predictive model and apply it appropriately in other settings. Guidelines have already been suggested by consortiums of researchers in hopes of improving these practices. Internal validation steps have been sufficiently established, but many authors do not report discrimination or calibration values in relation to their published models. These validation tests are the most important analytic step for assessing transportability without access to the original source dataset. The results of this work support the need for discrimination values and suggest that the reporting of internal validations assessments should be further emphasized.

The inclusion of retrospectively evaluated variability in discrimination metrics, such as the concordance statistic, was demonstrated as a new method for improving the evaluation and

understanding of model transportability. Simulated information was able to classify models into more distinct transportability groups. As target cohorts are rarely the same size as source cohorts, observed performance differences should also be checked for significance at multiple data sizes. Increases and decreases in discrimination may be related to the random selection of cases in the target rather than a true underlying population difference. Calculating the variability of applied metrics during internal validation is important for understand the bias of the source evaluation and enables comparisons for significance with the target. The method described in this work is particularly useful in present applications as it can make use of a minimal set of values that authors commonly report. Decreases in performance for the examined brain cancer models imply that none of the models were able to generalize, but do not distinguish between the need for calibration or partial adjustments. Final interpretations of transportability were more concrete when appropriate internal values were available, making it possible to separate between calibration and partial adjustment cases in particular. For example, the Michaelsen brain cancer model was able to be classified specifically for partial adjustment based on the significant decrease seen by adding discrimination variability to the transportability assessment. This capability can make it easier to determine what models should be approached for adjustment with specialized methods in the future.

Minimal investigation has been performed to determine how to approach partial adjustments to models. Evaluation of a Bayesian belief network in this work helped demonstrate that the combined use of source and target information can provide performance improvements. Transportability theory was shown as a useful tool for stating assumptions of causality and population differences. By using graphical rules from the theory, a formal process can be followed to define what features should be targeted for adjustment. Performance in the transportability

theory evaluations were able to recover to levels comparable with a complete re-estimation process when appropriate features are targeted. Currently, this process is still difficult to apply as more features and disruptive assumptions are applied to model graph. However, the application holds promise as a targeted approach to updating problematic models in need of partial adjustment.

7.2 FUTURE DIRECTIONS

This dissertation touches on two areas of external validation in need of improvement. However, the design and study of external validation is a relatively young area of research. Many additional areas can be targeted in future research and the currently proposed methods can also be extended further. Furthermore, the results of this dissertation are informative for suggesting updates to recently proposed guidelines and the validation process as a whole.

7.2.1 Additional use of simulations and metrics

Simulated cohorts built from limited published information were useful for making new determinations about models in this work. This process was used to focus on estimating variability of discrimination in the c-statistic, but may be generalizable to many different performance metrics. For example, calibration serves as another primary metric for analyzing external validation, but is not incorporated in current interval validation assessments. However, during cross-validation or bootstrapping, the calibration of the trained model to test cases will vary over folds/bootstraps in the same way that the discrimination metrics vary. Incorporating calibration scores and calibration variance with confidence intervals can help better describe the stability of the model in the internal setting. These values should be computed as a new addition to standard validation analysis of the source data. In addition, the simulation process from this work can be applied to consider these calibration values retrospectively if necessary.

Deeper analysis of model calibration may also prove useful for studying transportability of findings over time. For example, as updated evidence informs new treatments options, patient care may be significantly revised. Observed event rates would change and previously validated models would become miscalibrated. Future research should evaluate how models can be re-calibrated for these types of medical updates and calculate at what point a previously validated model would begin to require partial adjustment or complete retraining. Thus, assessment of calibration over time would inform researchers of what features are most susceptible to change and whether models are broadly applicable over long time periods. Providing a form of continuous model review could generate new forms of transportability adjustment and model designs that robustly handle rapid updating.

Other novel metrics for understanding transportability should be considered moving forward. Some publications make use of additional performance measures such as explained variation (R^2) and the Brier score in reporting [126,127]. Future research should consider what metrics might be more appropriate in certain circumstances and how effectively a simulation process can estimate their values. Two additional discrimination measures, the net reclassification index (NRI) and integrated discrimination index (IDI), help quantify model performance when adding/changing predictors in a model and may provide insight into model updating [21]. However, the application of NRI for validation has been controversial and requires further study to become reliable a measure [27,128]. Finally, model research needs to assess clinical utility in more detail. Clinical utility requires a choice between decision-making thresholds that are difficult to define across many different models. Discrimination measures such as the c-statistic generalize over many thresholds to summarize overall performance. Decision curve analysis (DCA) is a metric for the calculating the benefit of different models when comparing false-positive and false-negative rates

of predictions [32]. DCA is limited by a need for expert input to define threshold comparisons and future research should attempt to learn appropriate thresholds through validation procedures. While additional metrics are unlikely to supplant discrimination or calibration for validation assessment, they may ultimately prove useful for further defining transportability classifications. Thus, future work should explore new metrics broadly to determine how they can provide accurate separation of the most useful models for evidence-based decision making.

Lastly, the proposed simulation approaches in this work are relatively simple, striving to provide ease of implementation while requiring minimal data. The proposed naïve approach, for example, requires only two or three sets of published inputs and makes general assumptions of sampling distributions. As a result, the naïve approach was unable to consistently create a cohort that had a discrimination c-statistic significantly similar to the compared source cohort during NLST investigation. The covariance approach adds complexity to the simulation process, attempting to model feature correlations. Simulated covariance cohorts showed improved similarity in c-statistics compared to the naïve approach, although they continued to be significantly different. In order to more accurately generate simulated values, future work should explore what information is required to reach significant similarity with simulations. One direct improvement would include further constraining the simulation process to follow selection criteria from the source paper. For example, patients are frequently restricted to specific ranges based on certain features (e.g., ages between 50-70 years old). The current naïve and covariance approaches do not take such restrictions into account, generating a set of cases that can potentially fall outside the bounds of the source cohort criteria. Maximizing the similarity of feature bounds to source examples is important for creating a cohort with significantly similar discrimination performance. Another approach to maximizing similarity would include further optimization with more advanced

Markov chain Monte Carlo sampling. This method attempts to optimize performance by choosing samples that move the simulated performance closer to the published c-statistic. The difficulty of this method, however, is in defining the stopping condition for the sampling process. Adding this functionality could complicate the simulation process, making it less accessible to researchers. Nevertheless, simulated cohorts that are insignificantly different from the source would be able to estimate the central tendency of internal validation metrics and help extend the utility of the simulation process. Future adjustments to the simulation approach must balance robust specification while remaining as simple as possible to allow for widespread adoption.

7.2.2 Publication guidelines

Ultimately, retrospective reviews using simulation might not be necessary if researchers follow a proper set of guidelines for reporting predictive model design and validation. The recently released TRIPOD statement [79] is the most comprehensive attempt to provide a checklist of reportable information in modeling papers. It covers what information should be reported in all sections of a paper from abstract to conclusions. The results of evaluations in this dissertation help stress the importance of items such as c-statistic variability and calibration details. These items should be strongly emphasized in any reporting guidelines.

TRIPOD includes both confidence intervals and calibration in the current checklist as part of results sections covering “Model performance” and “Model-updating” [79]. The findings of this work support the descriptions provided by the TRIPOD statement. As seen when interpreting model transportability in this work, discrimination values are crucial and the lack of AUC or c-statistic values make it impossible to fully evaluate external validity. This work found calibration metrics were important for distinguishing between the trivial and calibration adjustment levels of model transportability as well. Other recent publications on proper validation procedures stress

calibration testing but do not always include suggestions for measurements of confidence intervals during internal validation [35,37–39,129]. Some researchers may assume inclusion of items like confidence intervals is implied as part of standard analysis. However, given the general absence of confidence intervals in current reporting, all items relevant to internal and external validation should be stressed explicitly. Therefore, this work emphasizes the need for discrimination, calibration, and confidence interval assessment to support transportability analysis. Repeated evaluations in other domains can further underscore how following TRIPOD guidelines will create well described publications that lead to increased model evaluation and reuse.

Based on the findings in previous chapters, a source covariance matrix could be suggested as an additional reporting requirement for revised TRIPOD guidelines. The covariance matrix contains important feature correlation information and can be shared more easily than the source data in most circumstances. As demonstrated in this dissertation, the addition of information from the covariance matrix can be used to simulate cohorts that are more closely related to the source cases. Consequently, the addition of this matrix is a relatively simple step that can allow for more accurate external analysis. Adding the covariance matrix as a requirement to supplementary materials in publications would allow for more widespread simulation and validation analysis while reducing concerns of exposing protected health information.

The current state of external validation research is moving quickly, often with multiple newly proposed frameworks or revised evaluation processes being published each year. Future analysis of techniques like those presented in this work may support other additions or subtractions from the current guidelines. Therefore, steering committees must act quickly to update guidelines. Similarly, researchers interested in predictive modeling must stay appraised of current suggestions provided by guideline efforts. Finally, the modeling community should broaden the reporting

guideline focus beyond publications and into databases of model findings. The important context of data collection, cleaning, feature selection, and model design can be standardized as seen by efforts such as PMML and ODHSI [96,100]. Researchers should be required to summarize their findings in programmatic forms that allow for easy review, replication, and validation. These computer readable descriptions of model designs and outcomes are less ambiguous than free text descriptions. Creating a database of models would help the community compare and contrast model findings more quickly, particularly when searching for the most generalizable models to apply to specific predictive tasks.

7.2.3 Transportability validations and classifications

There are a variety of proposed options for approaching external validity analysis. Most overlap with each other and the methods in this work also inherit from them closely. Nevertheless, continued evaluation of these approaches can help define the most effective techniques. In this work, heavy focus was given to model discrimination, in line with the majority of other proposed frameworks. Discrimination is certainly the most important analytic measure for initial assessment of a model's use in decision making. Decision classes must be separated into clear groups to achieve accurate conclusions. However, new methods are needed for exploring the intricate differences between cohorts. The primary method of this work takes advantage of c-statistic variability and confidence intervals to provide additional insight into discrimination. Debray et al.'s approach inspects case-mix, linear predictor difference, and standard deviation ratios, providing another novel approach that might change the way researchers analyze transportability [41]. Additional work is needed to push the boundaries of analysis with improved metrics. Calibration, in particular, can still be difficult to interpret. For example, direct comparison of calibration intercepts and slopes does not yield clear interpretations of which models are most

appropriate for adjustment. Novel approaches can more clearly define these interpretations and consider validation issues from the clinical perspective of decision making [31]. However, their use should be clearly grounded in the differentiation of model external validity. Transportability assessments clearly require further extension to reach this goal and provide distinct classifications that can give clinicians confidence that predictive models can become a part of their regular practice.

One contribution of this work includes a proposed set of transportability classifications that attempt to divide models more evenly between levels of utility. The primary goal of this grouping is to provide a clear differentiation between the trivial, calibration, and partial transportability categories. Most research to date focuses on models that meet trivial and calibration adjustment constraints. However, these models are relatively easy cases to evaluate as most of the information is transportable with minor adjustments. More focus is needed to target models with significant predictive issues (i.e., models in the partial adjustment category).

As more assessments of model transportability are completed, the proposed categories of this work can be updated. For example, many frameworks attempt to separate transportability evaluations based on the known difference in cohorts, such as temporal or geographic transportability. New categorizations combining this predefined knowledge with the more general categories of transportability levels might be explored in future work. Subcategorizations for the calibration and partial adjustment categories could also be defined as new processes are developed for adjusting models in these states. Causal transportability evaluation, for example, is just one technique for partial model adjustment. Additional research might define that this method or others are only applicable to certain classes of difference within the partial adjustment category.

7.2.4 Causal analysis and partial model adjustments

Causal graphical analysis with transportability theory stands as an interesting tool for partially adjusting models. However, graphical modeling is not a widely used technique for disease modeling. Future work must consider ways to define graphical assumptions more easily from existing models. For example, a process for converting regression models to a graphical structure by combining information from regression coefficients and a source covariance matrix might allow for a guided definition of the nodes and edges of a Bayesian belief network (BBN). These resulting BBNs would make it easier to define model assumptions and add causal constraints so that transportability theory could be tested more broadly. In addition, researchers might also include their own assumptions of feature and outcome relationships when designing regression models. With widespread adoption and reporting, these expert defined graphs would serve as templates of graphical structure and could inform broad structure designs. Finally, current explorations of transportability theory are often limited in the scope of variables explored. Increasing the size of networks in future analysis will be important for understand the limitations of applying this method in complex model designs.

Overall, cohort differences and partial model adjustments are in need of much deeper examination. Big data and machine learning attempt to surpass issues of difference by including larger and larger training datasets. Yet, in medical practice, the number of patients and data points are still relatively limited due to the rarity of certain diseases and the need for controlled experimentation. Modeling efforts in healthcare must balance between these two approaches and make use of partial adjustments as a tool for correcting specific models. Trained models can then be made generally useful when the collection of multiple, large datasets is not possible. While this work demonstrated

one method for the identification of model difference and partial adjustment, this task remains an open area of research in need of other novel solutions.

7.3 CONCLUSION

Machine learning and predictive analysis can play an important role in improving medical decision making. Before clinicians and patients are comfortable trusting predictive models developed with these techniques, they must undergo robust validation analysis in many related, but slightly different settings. Evaluations of the transportability of models are still obstructed by current reporting practices and the relative novelty of the need for this secondary evaluation. Many researchers from computer science, epidemiology, informatics, and other medical fields have begun driving reliable and standard approaches for this field.

This dissertation investigated improvements to the assessment and categorization of models. In addition, a novel process for evaluating models in need of partial adjustment was explored. These applications show that additional improvements can be made for assessments of external validity by obtaining information on the performance of a model that may be unstudied or unreported in previously published findings.

The methods presented in this dissertation can help solve certain problem cases in validation analysis, but overall can be considered on step in future advances in this field. Some future directions are discussed as part of this chapter, but advancements may elucidate broader approaches than those suggested. The results of this dissertation help demonstrate how transportability can be given a more nuanced analysis so that model reuse can be clearly defined. External validation is now at the forefront of predictive modeling research and through continued

research efforts will begin to provide confidence for applying models to aid evidence-based decision making.

APPENDIX A

SIMULATION CONSTRAINTS: COVARIANCE AND SURVIVAL VALUES

	age	bmi	sday	syr	squit	race2	race3	race4	race5	hisp	educ	copd	hcancer	fhcancer	cigsmok
age	25.181	-1.686	0.060	19.446	3.542	-0.035	0.015	-0.007	-0.002	-0.004	-0.540	0.069	0.064	-0.037	-0.270
bmi	-1.686	24.898	-0.085	-4.401	3.773	0.041	-0.032	0.004	0.009	0.001	-0.321	0.013	-0.015	0.050	-0.496
sday	0.060	-0.085	0.016	0.199	-0.169	0.003	0.000	0.000	0.000	0.000	0.004	-0.001	0.000	-0.001	0.015
syr	19.446	-4.401	0.199	54.300	-18.051	0.043	0.005	-0.003	0.003	0.002	-1.960	0.120	0.076	-0.053	1.394
squit	3.542	3.773	-0.169	-18.051	23.988	-0.073	0.008	-0.002	-0.003	-0.004	0.457	-0.019	-0.007	0.030	-1.762
race2	-0.035	0.041	0.003	0.043	-0.073	0.042	-0.001	0.000	0.000	0.000	-0.024	0.000	0.000	0.000	0.008
race3	0.015	-0.032	0.000	0.005	0.008	-0.001	0.019	0.000	0.000	0.000	0.003	0.000	-0.001	-0.002	-0.001
race4	-0.007	0.004	0.000	-0.003	-0.002	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
race5	-0.002	0.009	0.000	0.003	-0.003	0.000	0.000	0.000	0.004	0.000	-0.003	0.000	0.000	0.000	0.000
hisp	-0.004	0.001	0.000	0.002	-0.004	0.000	0.000	0.000	0.000	0.012	-0.005	0.000	0.000	-0.001	0.000
educ	-0.540	-0.321	0.004	-1.960	0.457	-0.024	0.003	0.000	-0.003	-0.005	2.327	-0.001	-0.004	-0.034	-0.046
copd	0.069	0.013	-0.001	0.120	-0.019	0.000	0.000	0.000	0.000	0.000	-0.001	0.049	0.002	0.002	-0.002
hcancer	0.064	-0.015	0.000	0.076	-0.007	0.000	-0.001	0.000	0.000	0.000	-0.004	0.002	0.040	0.000	0.001
fhcancer	-0.037	0.050	-0.001	-0.053	0.030	0.000	-0.002	0.000	0.000	-0.001	-0.034	0.002	0.000	0.172	-0.003
cigsmok	-0.270	-0.496	0.015	1.394	-1.762	0.008	-0.001	0.000	0.000	0.000	-0.046	-0.002	0.001	-0.003	0.250

Table A.1 Complete covariance matrix of the NLST test cohort. Compact variable names used for spacing, see Table A.2 for full NLST feature names.

NLST Coding	Short Coding
age	age
bmi	bmi
smokday	sday
smokyear	syr
smokquittime	squit
race2	race2
race3	race3
race4	race4
race5	race5
hispanic	hisp
educat	educ
diagcopd	copd
histcancer	hcancer
famhistcancer	fhcancer
cigsmok	cigsmok

Table A.2 NLST and short feature name reference.

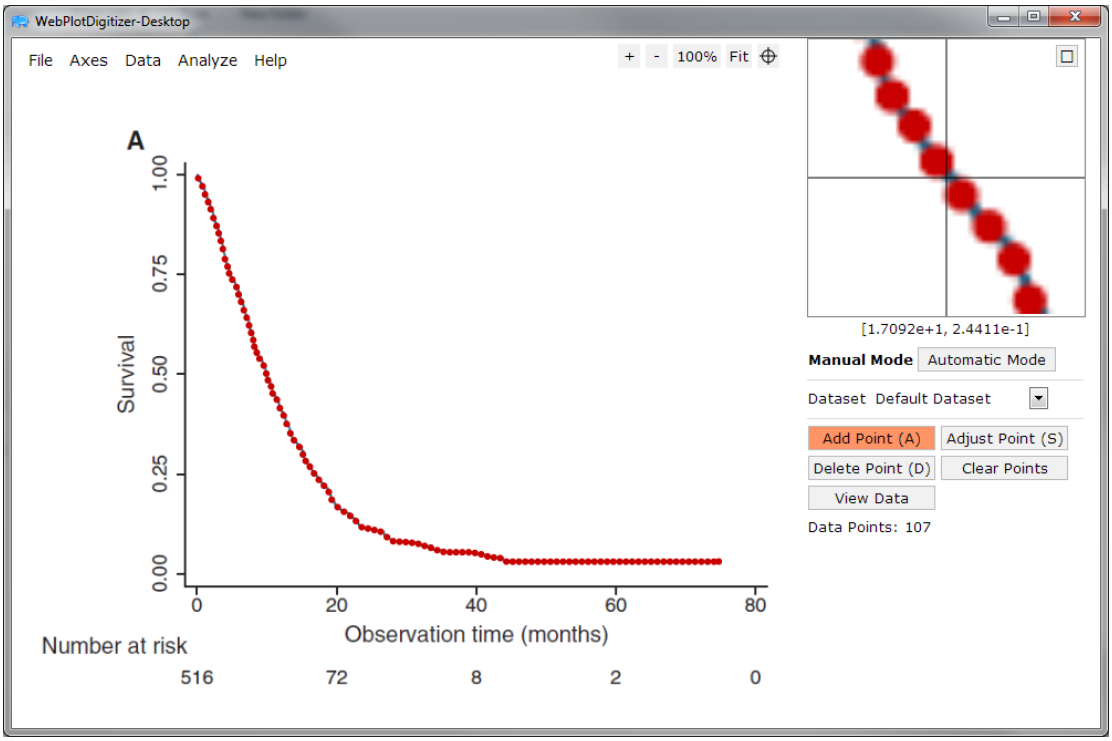


Figure A.1 Data extraction for Helseth overall survival curve.

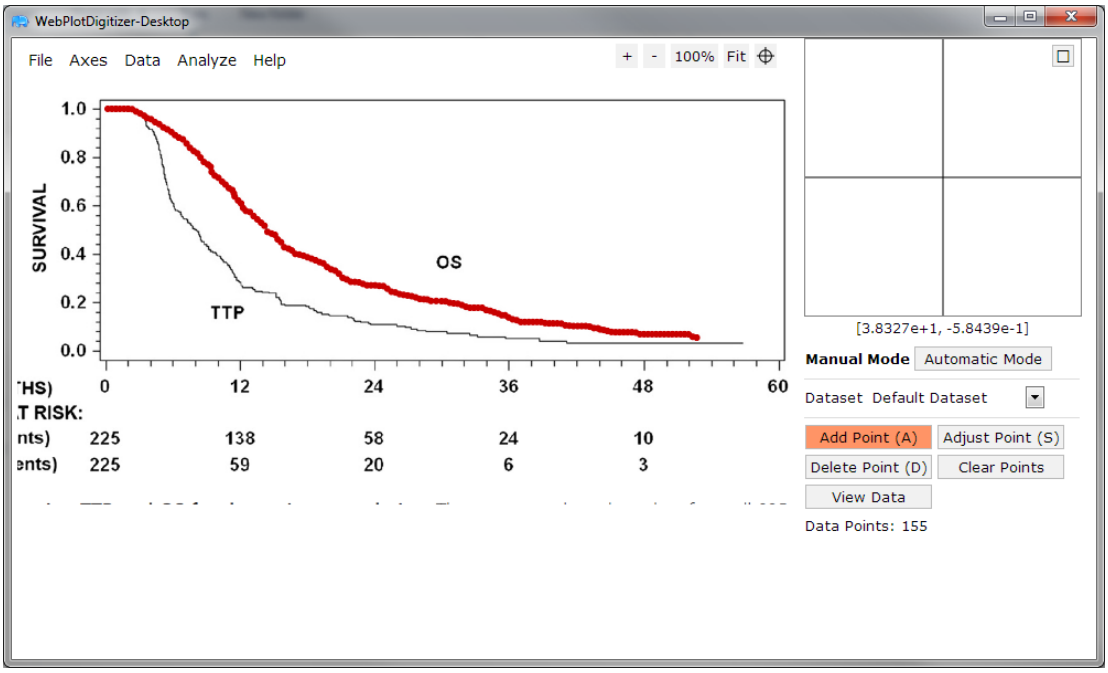


Figure A.2 Data extraction for Michaelsen overall survival (OS) curve. Points generated on the time to progression (TTP) survival curve were removed manually.

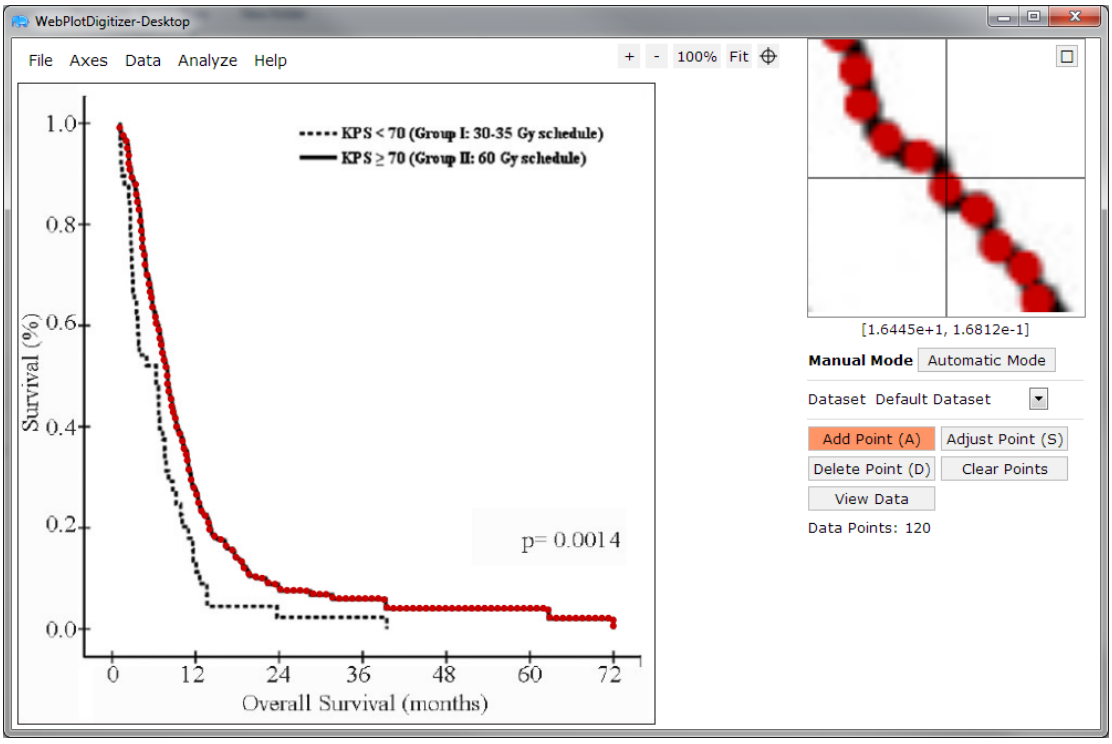


Figure A.3 Data extraction for the Kumar overall survival curve for Group II cases used for Cox regression (KPS >= 70 and 60 Gy chemotherapy). Points generated on the Group I survival curve were removed manually.

Table A.3 Extracted survival times and probabilities from published (Helseth, Michaelsen, Kumar) or derived (Gutman) Kaplan-Meier survival curves. Extracted values were used for Cox hazards cohort simulation.

Helseth		Michaelsen		Gutman		Kumar	
Time (months)	Survival	Time (months)	Survival	Time (months)	Survival	Time (months)	Survival
0.238	0.992	0.140	1.001	0.000	1.000	1.234	0.990
0.838	0.972	0.502	1.001	0.526	0.988	1.750	0.975
1.213	0.951	0.863	1.001	0.723	0.975	2.165	0.964
1.663	0.932	1.225	1.001	1.282	0.963	2.319	0.949
2.023	0.914	1.587	1.001	2.104	0.950	2.517	0.919
2.413	0.892	1.949	1.001	2.696	0.925	2.535	0.935
2.863	0.872	2.311	0.999	3.353	0.912	2.671	0.907
3.163	0.854	2.673	0.990	3.419	0.899	2.980	0.892
3.464	0.835	3.035	0.982	3.945	0.885	3.524	0.878
3.764	0.815	3.385	0.973	4.044	0.872	3.599	0.858
4.064	0.789	3.639	0.963	4.142	0.858	3.752	0.843
4.439	0.771	4.000	0.958	4.175	0.845	3.984	0.828
4.664	0.754	4.362	0.946	4.241	0.831	4.181	0.806
5.114	0.738	4.724	0.938	4.471	0.818	4.321	0.786
5.714	0.719	5.086	0.924	4.800	0.804	4.443	0.770
6.014	0.701	5.448	0.916	5.753	0.776	4.472	0.754
6.374	0.682	5.809	0.906	5.852	0.762	4.724	0.739
6.764	0.661	6.141	0.893	6.148	0.748	4.825	0.720
7.176	0.642	6.473	0.882	7.233	0.734	5.123	0.699
7.514	0.623	6.835	0.875	7.397	0.719	5.396	0.681
7.814	0.605	7.167	0.858	7.726	0.705	5.550	0.666
8.114	0.587	7.468	0.840	8.416	0.691	5.751	0.654
8.264	0.570	7.800	0.827	8.581	0.676	5.904	0.635
8.614	0.555	8.132	0.817	9.764	0.662	6.313	0.616
9.014	0.540	8.403	0.800	9.797	0.647	6.406	0.604
9.614	0.523	8.705	0.782	10.093	0.633	6.786	0.591
9.974	0.503	9.067	0.772	10.652	0.618	6.934	0.574
10.214	0.486	9.308	0.762	10.718	0.603	7.140	0.561
10.664	0.470	9.398	0.741	10.816	0.588	7.309	0.545
10.889	0.453	9.670	0.726	11.079	0.574	7.525	0.532
11.489	0.437	10.001	0.717	11.737	0.558	7.809	0.517
11.924	0.416	10.333	0.701	11.901	0.542	7.988	0.500
12.464	0.398	10.635	0.689	12.164	0.525	8.017	0.484
12.889	0.376	10.936	0.674	12.559	0.509	8.202	0.470
13.424	0.353	11.238	0.666	12.625	0.492	8.527	0.455
13.934	0.336	11.382	0.653	13.644	0.475	8.634	0.440
14.714	0.319	11.449	0.640	13.907	0.458	8.835	0.428
15.214	0.300	11.751	0.623	13.973	0.442	9.144	0.416
15.614	0.283	12.022	0.611	14.301	0.425	9.309	0.400
16.214	0.269	12.203	0.592	14.926	0.408	9.763	0.386
16.814	0.252	12.444	0.580	15.353	0.390	10.151	0.371
17.489	0.236	12.806	0.576	15.386	0.373	10.371	0.356
18.239	0.222	13.138	0.559	17.786	0.356	10.692	0.345
18.914	0.206	13.470	0.545	17.951	0.340	10.846	0.333
19.339	0.186	13.801	0.532	18.378	0.323	11.055	0.315
20.189	0.168	14.103	0.519	19.332	0.307	11.386	0.295

21.089	0.156	14.374	0.494	19.595	0.290	11.773	0.279
21.989	0.146	14.736	0.487	19.660	0.273	12.161	0.266
22.814	0.133	15.038	0.481	20.252	0.257	12.444	0.250
23.640	0.118	15.279	0.463	20.877	0.240	12.857	0.233
24.540	0.114	15.581	0.451	22.290	0.222	13.401	0.224
25.440	0.111	15.861	0.430	22.685	0.204	13.866	0.210
26.340	0.107	16.244	0.424	24.164	0.186	14.020	0.197
27.240	0.093	16.546	0.417	24.230	0.168	14.720	0.183
28.140	0.083	16.847	0.403	24.559	0.150	15.653	0.177
29.040	0.081	17.209	0.400	25.348	0.134	16.391	0.164
29.940	0.080	17.571	0.396	26.696	0.117	17.286	0.156
30.840	0.079	17.933	0.389	27.222	0.100	17.818	0.142
31.740	0.076	18.295	0.383	29.195	0.081	18.607	0.134
32.640	0.071	18.657	0.377	31.101	0.065	18.956	0.121
33.540	0.066	19.019	0.369	33.666	0.049	19.668	0.109
34.440	0.060	19.350	0.364	37.578	0.037	20.628	0.103
35.340	0.056	19.682	0.349	42.148	0.023	21.562	0.101
36.240	0.055	20.044	0.339	44.581	0.012	22.418	0.092
37.140	0.055	20.406	0.334	51.321	0.006	23.430	0.089
38.040	0.055	20.738	0.321	56.877	0.001	24.174	0.079
38.940	0.055	21.069	0.304			25.142	0.077
39.840	0.053	21.431	0.298			26.076	0.077
40.740	0.050	21.793	0.288			27.011	0.077
41.640	0.044	22.155	0.287			27.946	0.075
42.541	0.042	22.517	0.286			28.880	0.070
43.441	0.040	22.879	0.279			29.814	0.069
44.341	0.032	23.241	0.273			30.749	0.069
45.241	0.031	23.603	0.273			31.657	0.062
46.141	0.031	23.965	0.273			32.617	0.061
47.041	0.031	24.327	0.271			33.552	0.061
47.941	0.031	24.688	0.270			34.486	0.061
48.841	0.031	25.050	0.260			35.421	0.061
49.741	0.031	25.412	0.247			36.356	0.061
50.641	0.031	25.774	0.244			37.290	0.061
51.541	0.031	26.136	0.237			38.225	0.061
52.441	0.031	26.498	0.234			39.082	0.058
53.341	0.031	26.860	0.232			39.469	0.043
54.241	0.031	27.222	0.229			40.403	0.041
55.141	0.031	27.584	0.224			41.338	0.041
56.041	0.031	27.946	0.217			42.273	0.041
56.941	0.031	28.307	0.216			43.207	0.041
57.841	0.031	28.609	0.215			44.142	0.041
58.741	0.031	28.942	0.209			45.077	0.041
59.641	0.031	29.408	0.209			46.011	0.041
60.541	0.031	29.876	0.208			46.946	0.041
61.441	0.031	30.238	0.208			47.881	0.041
62.342	0.031	30.599	0.202			48.815	0.041
63.242	0.031	30.961	0.199			49.750	0.041
64.142	0.031	31.323	0.198			50.685	0.041
65.042	0.031	31.685	0.192			51.620	0.041
65.942	0.031	32.047	0.185			52.554	0.041
66.842	0.031	32.409	0.181			53.489	0.041
67.742	0.031	32.771	0.181			54.424	0.041

68.642	0.031
69.542	0.031
70.442	0.031
71.342	0.031
72.242	0.031
73.142	0.031
74.042	0.031
74.717	0.032

33.133	0.181
33.495	0.181
33.857	0.172
34.218	0.169
34.580	0.162
34.942	0.157
35.247	0.150
35.543	0.150
35.847	0.141
36.209	0.134
36.571	0.130
36.933	0.123
37.295	0.123
37.656	0.123
38.018	0.123
38.380	0.123
38.742	0.123
39.104	0.118
39.466	0.117
39.828	0.117
40.190	0.117
40.552	0.116
40.914	0.109
41.275	0.107
41.637	0.106
41.999	0.106
42.361	0.106
42.723	0.106
43.085	0.104
43.447	0.098
43.809	0.097
44.020	0.092
44.344	0.090
44.653	0.086
45.015	0.082
45.377	0.081
45.739	0.081
46.101	0.081
46.463	0.081
46.825	0.081
47.186	0.080
47.548	0.073
47.910	0.073
48.272	0.073
48.634	0.073
48.996	0.073
49.358	0.073
49.720	0.073
50.082	0.073
50.444	0.073
50.805	0.073
51.167	0.073
51.529	0.073

55.358	0.041
56.293	0.041
57.228	0.041
58.162	0.041
59.097	0.041
60.032	0.041
60.966	0.041
61.901	0.041
62.602	0.038
62.833	0.023
63.768	0.022
64.702	0.022
65.637	0.022
66.572	0.022
67.506	0.022
68.441	0.022
69.376	0.022
70.311	0.022
71.245	0.022
71.944	0.006
71.946	0.018

	51.891	0.072		
	52.253	0.062		
	52.615	0.059		

REFERENCES

1. Sackett DL, Rosenberg WM, Gray JA, *et al.* Evidence based medicine: what it is and what it isn't. *BMJ* 1996;**312**:71–2.

2. ClinicalTrials.gov. https://clinicaltrials.gov/

3. Bui AAT, Taira RK. *Medical Imaging Informatics*. Springer Science & Business Media 2009.

4. CONSORT Website. http://www.consort-statement.org/

5. Rothwell PM. External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet* 2005;**365**:82–93. doi:10.1016/S0140-6736(04)17670-8

6. Godwin M, Ruhland L, Casson I, *et al.* Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Med Res Methodol* 2003;**3**:28. doi:10.1186/1471-2288-3-28

7. Tunis SR SD. Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003;**290**:1624–32. doi:10.1001/jama.290.12.1624

8. Dekkers OM, Elm E von, Algra A, *et al.* How to assess the external validity of therapeutic trials: a conceptual approach. *Int J Epidemiol* 2010;**39**:89–94. doi:10.1093/ije/dyp174

9. Bero L RD. The cochrane collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995;**274**:1935–8. doi:10.1001/jama.1995.03530240045039

10. Moons KGM, Royston P, Vergouwe Y, *et al.* Prognosis and prognostic research: what, why, and how? *BMJ* 2009;**338**:b375. doi:10.1136/bmj.b375

11. Harrell F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer 2015.

12. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media 2008.

13. Begg C, Cho M, Eastwood S, *et al.* Improving the quality of reporting of randomized controlled trials: The consort statement. *JAMA* 1996;**276**:637–9. doi:10.1001/jama.1996.03540080059030

14. Kleinberg S, Hripcsak G. A review of causal inference for biomedical informatics. *J Biomed Inform* 2011;**44**:1102–12. doi:10.1016/j.jbi.2011.07.001

15. Bleeker S., Moll H., Steyerberg E., *et al.* External validation is necessary in prediction research:: A clinical example. *J Clin Epidemiol* 2003;**56**:826–32. doi:10.1016/S0895-4356(03)00207-5

16. König IR, Malley JD, Weimar C, *et al.* Practical experiences on the necessity of external validation. *Stat Med* 2007;**26**:5499–5511. doi:10.1002/sim.3069

17. Petersen ML. Compound Treatments, Transportability, and the Structural Causal Model. *Epidemiology* 2011;**22**:378–81. doi:10.1097/EDE.0b013e3182126127

18. Steyerberg EW, Harrell Jr FE, Borsboom GJJM, *et al.* Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;**54**:774–81. doi:10.1016/S0895-4356(01)00341-9

19. Altman DG, Vergouwe Y, Royston P, *et al.* Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605–b605. doi:10.1136/bmj.b605

20. Katz MH. Multivariable Analysis: A Primer for Readers of Medical Research. *Ann Intern Med* 2003;**138**:644–50. doi:10.7326/0003-4819-138-8-200304150-00012

21. Pencina MJ, D' Agostino RB, D' Agostino RB, *et al.* Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27**:157–72. doi:10.1002/sim.2929

22. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;:ehu207. doi:10.1093/eurheartj/ehu207

23. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;**30**:11–21. doi:10.1002/sim.4085

24. Pencina MJ, D'Agostino RB, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med* 2012;**31**:101–13. doi:10.1002/sim.4348

25. Pencina MJ, D'Agostino RB, Pencina KM, *et al.* Interpreting Incremental Value of Markers Added to Risk Prediction Models. *Am J Epidemiol* 2012;:kws207. doi:10.1093/aje/kws207

26. Kerr KF, Wang Z, Janes H, *et al.* Net Reclassification Indices for Evaluating Risk Prediction Instruments: A Critical Review. *Epidemiology* 2014;**25**:114–21. doi:10.1097/EDE.0000000000000018

27. Leening MJG, Vedder MM, Witteman JCM, *et al.* Net Reclassification Improvement: Computation, Interpretation, and ControversiesA Literature Review and Clinician's Guide. *Ann Intern Med* 2014;**160**:122–31. doi:10.7326/M13-1522

28. Cook NR. Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve. *Clin Chem* 2008;**54**:17–23. doi:10.1373/clinchem.2007.096529

29. Cook NR. Clinically Relevant Measures of Fit? A Note of Caution. *Am J Epidemiol* 2012;:kws208. doi:10.1093/aje/kws208

30. Pepe MS, Fan J, Feng Z, *et al.* The Net Reclassification Index (NRI): A Misleading Measure of Prediction Improvement Even with Independent Test Data Sets. *Stat Biosci* 2014;**7**:282–95. doi:10.1007/s12561-014-9118-0

31. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med Decis Making* 2006;**26**:565–74. doi:10.1177/0272989X06295361

32. Vickers AJ, Cronin AM, Elkin EB, *et al.* Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak* 2008;**8**:53. doi:10.1186/1472-6947-8-53

33. Justice AC, Covinsky KE, Berlin JA. Assessing the Generalizability of Prognostic Information. *Ann Intern Med* 1999;**130**:515–24. doi:10.7326/0003-4819-130-6-199903160-00016

34. Collins GS, Moons KGM. Comparing risk prediction models. *BMJ* 2012;**344**:e3186–e3186. doi:10.1136/bmj.e3186

35. Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiol Camb Mass* 2010;**21**:128–38. doi:10.1097/EDE.0b013e3181c30fb2

36. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–473. doi:10.1002/(SICI)1097-0258(20000229)19:4<453: :AID-SIM350>3.0.CO;2-5

37. Moons KGM, Kengne AP, Grobbee DE, *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;:heartjnl-2011-301247. doi:10.1136/heartjnl-2011-301247

38. Debray TPA, Moons KGM, Ahmed I, *et al.* A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* Published Online First: 2013. doi:10.1002/sim.5732

39. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* 2013;**13**:33. doi:10.1186/1471-2288-13-33

40. Collins GS, Groot JA de, Dutton S, *et al.* External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;**14**:40. doi:10.1186/1471-2288-14-40

41. Debray TPA, Vergouwe Y, Koffijberg H, *et al.* A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;**68**:279–89. doi:10.1016/j.jclinepi.2014.06.018

42. Brooks JC, Shavelle RM, Strauss DJ, *et al.* Long-Term Survival After Traumatic Brain Injury Part I: External Validity of Prognostic Models. *Arch Phys Med Rehabil* 2015;**96**:994–999.e2. doi:10.1016/j.apmr.2015.02.003

43. Siontis GCM, Tzoulaki I, Castaldi PJ, *et al.* External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;**68**:25–34. doi:10.1016/j.jclinepi.2014.09.007

44. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *J Clin Epidemiol* 2016;**69**:245–7. doi:10.1016/j.jclinepi.2015.04.005

45. Pearl J, Bareinboim E. Transportability of Causal and Statistical Relations: A Formal Approach. In: *2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW)*. IEEE 2011. 540–7. doi:10.1109/ICDMW.2011.169

46. Bareinboim E, Pearl J. Transportability of Causal Effects: Completeness Results. In: *Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012. http://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5188

47. Bareinboim E, Pearl J. Meta-Transportability of Causal Effects: A Formal Approach. In: *Proceedings of The Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*. JMLR (31) 2013. 135–143.

48. Pearl J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press 2000.

49. Koop G, Poirier DJ, Tobias JL. *Bayesian econometric methods*. Cambridge University Press 2007.

50. Darwiche A. *Modeling and reasoning with Bayesian networks*. Cambridge University Press Cambridge 2009.

51. Greenland S, Pearl J, Robins JM. Causal Diagrams for Epidemiologic Research. *Epidemiology* 1999;**10**:37–48. doi:10.2307/3702180

52. Lucas PJF, Segaar RW, Janssens AR. HEPAR: an expert system for the diagnosis of disorders of the liver and biliary tract. *Liver* 1989;**9**:266–275. doi:10.1111/j.1600-0676.1989.tb00410.x

53. Andreassen S, Woldbye M, Falck B, *et al.* MUNIN: a causal probabilistic network for interpretation of electromyographic findings. In: *Proceedings of the 10th international joint conference on Artificial intelligence - Volume 1*. San Francisco, CA, USA: : Morgan Kaufmann Publishers Inc. 1987. 366–372.

54. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems: Part I. The Pathfinder project. *Methods Inf Med* 1992;**31**:90–105.

55. Kahn Jr CE, Roberts LM, Shaffer KA, *et al.* Construction of a Bayesian network for mammographic diagnosis of breast cancer. *Comput Biol Med* 1997;**27**:19–29. doi:10.1016/S0010-4825(96)00039-X

56. Burnside E, Rubin D, Shachter R. A Bayesian network for mammography. *Proc AMIA Symp* 2000;:106–10.

57. Burnside ES, Rubin DL, Shachter RD, *et al.* A probabilistic expert system that provides automated mammographic–histologic correlation: initial experience. *Am J Roentgenol* 2004;**182**:481–488.

58. Burnside ES, Rubin DL, Fine JP, *et al.* Bayesian Network to Predict Breast Cancer Risk of Mammographic Microcalcifications and Reduce Number of Benign Biopsy Results: Initial Experience1. *Radiology* 2006;**240**:666–73. doi:10.1148/radiol.2403051096

59. Jayasurya K, Fung G, Yu S, *et al.* Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. *Med Phys* 2010;**37**:1401.

60. Guha U, Chaerkady R, Marimuthu A, *et al.* Comparisons of tyrosine phosphorylated proteins in cells expressing lung cancer-specific alleles of EGFR and KRAS. *Proc Natl Acad Sci* 2008;**105**:14112–7. doi:10.1073/pnas.0806158105

61. Galán SF, Aguado F, Díez FJ, *et al.* NasoNet, modeling the spread of nasopharyngeal cancer with networks of probabilistic events in discrete time. *Artif Intell Med* 2002;**25**:247–64.

62. Nikiforidis GC, Sakellaropoulos GC. Expert system support using Bayesian belief networks in the prognosis of head-injured patients of the ICU. *Med Inform Médecine Inform* 1998;**23**:1–18.

63. Ogunyemi OI, Clarke JR, Ash N, *et al.* Combining Geometric and Probabilistic Reasoning for Computer-based Penetrating-Trauma Assessment. *J Am Med Inform Assoc* 2002;**9**:273–82. doi:10.1197/jamia.M0979

64. Kline JA, Novobilski AJ, Kabrhel C, *et al.* Derivation and Validation of a Bayesian Network to Predict Pretest Probability of Venous Thromboembolism. *Ann Emerg Med* 2005;**45**:282–90. doi:10.1016/j.annemergmed.2004.08.036

65. Lucas PJF, van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health-care. *Artif Intell Med* 2004;**30**:201–14. doi:10.1016/j.artmed.2003.11.001

66. Lacroix M, Abi-Said D, Fourney DR, *et al.* A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J Neurosurg* 2001;**95**:190–8. doi:10.3171/jns.2001.95.2.0190

67. Muacevic A, Kreth FW. Quality-adjusted survival after tumor resection and/or radiation therapy for elderly patients with glioblastoma multiforme. *J Neurol* 2003;**250**:561–8. doi:10.1007/s00415-003-1036-x

68. Pope WB, Sayre J, Perlina A, *et al.* MR Imaging Correlates of Survival in Patients with High-Grade Gliomas. *Am J Neuroradiol* 2005;**26**:2466–74.

69. Mazurowski MA, Desjardins A, Malof JM. Imaging descriptors improve the predictive power of survival models for glioblastoma patients. *Neuro-Oncol* Published Online First: 7 February 2013. doi:10.1093/neuonc/nos335

70. Esteller M, Garcia-Foncillas J, Andion E, *et al.* Inactivation of the DNA-Repair Gene MGMT and the Clinical Response of Gliomas to Alkylating Agents. *N Engl J Med* 2000;**343**:1350–4. doi:10.1056/NEJM200011093431901

71. Zinn PO, Majadan B, Sathyan P, *et al.* Radiogenomic Mapping of Edema/Cellular Invasion MRI-Phenotypes in Glioblastoma Multiforme. *PLoS ONE* 2011;**6**:e25451. doi:10.1371/journal.pone.0025451

72. Huang Y, Valtorta M. Identifiability in Causal Bayesian Networks: A Sound and Complete Algorithm. In: *Proceedings of the 21st National Conference on Artificial Intelligence*. Boston, Massachusetts: : AAAI Press 2006. 1149–1154.

73. Shpitser I, Pearl J. Identification of Conditional Interventional Distributions. In: *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*. Arlington, Virginia: : AUAI Press 2006. 437–444.

74. Publication search. "prediction model" NCBI PubMed Database. http://www.ncbi.nlm.nih.gov/pubmed/?term=prediction+model

75. Bleeker S., Moll H., Steyerberg E., *et al.* External validation is necessary in prediction research: *J Clin Epidemiol* 2003;**56**:826–32. doi:10.1016/S0895-4356(03)00207-5

76. Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. *Stat Med* 2004;**23**:907–926. doi:10.1002/sim.1691

77. Mushkudiani NA, Hukkelhoven CWPM, Hernández AV, *et al.* A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *J Clin Epidemiol* 2008;**61**:331–43. doi:10.1016/j.jclinepi.2007.06.011

78. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: Current issues and guidelines. *Int J Med Inf* 2008;**77**:81–97. doi:10.1016/j.ijmedinf.2006.11.006

79. Collins GS, Reitsma JB, Altman DG, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD

StatementThe TRIPOD Statement. *Ann Intern Med* 2015;**162**:55–63. doi:10.7326/M14-0697

80. Moons KGM, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration. *Ann Intern Med* 2015;**162**:W1–73. doi:10.7326/M14-0698

81. Pope WB, Sayre J, Perlina A, *et al.* MR imaging correlates of survival in patients with high-grade gliomas. *AJNR Am J Neuroradiol* 2005;**26**:2466–74.

82. Chaichana K, Parker S, Olivi A, *et al.* A proposed classification system that projects outcomes based on preoperative variables for adult patients with glioblastoma multiforme. *J Neurosurg* 2010;**112**:997–1004. doi:10.3171/2009.9.JNS09805

83. Huse JT, Holland E, DeAngelis LM. Glioblastoma: Molecular Analysis and Clinical Implications. *Annu Rev Med* 2013;**64**:59–70. doi:10.1146/annurev-med-100711-143028

84. Nakamura H, Murakami R, Hirai T, *et al.* Can MRI-derived factors predict the survival in glioblastoma patients treated with postoperative chemoradiation therapy? *Acta Radiol* 2013;**54**:214–20. doi:10.1258/ar.2012.120525

85. Chambless LB, Kistka HM, Parker SL, *et al.* The relative value of postoperative versus preoperative Karnofsky Performance Scale scores as a predictor of survival after surgical resection of glioblastoma multiforme. *J Neurooncol* 2014;**121**:359–64. doi:10.1007/s11060-014-1640-x

86. Chinot OL, Wick W, Mason W, *et al.* Bevacizumab plus Radiotherapy–Temozolomide for Newly Diagnosed Glioblastoma. *N Engl J Med* 2014;**370**:709–22. doi:10.1056/NEJMoa1308345

87. Michaelsen SR, Christensen IJ, Grunnet K, *et al.* Clinical variables serve as prognostic factors in a model for survival from glioblastoma multiforme: an observational study of a cohort of consecutive non-selected patients from a single institution. *BMC Cancer* 2013;**13**:402. doi:10.1186/1471-2407-13-402

88. Kumar N, Kumar P, Angurana S, *et al.* Evaluation of outcome and prognostic factors in patients of glioblastoma multiforme: A single institution experience. *J Neurosci Rural Pract* 2013;**4**:46. doi:10.4103/0976-3147.116455

89. Helseth R, Helseth E, Johannesen TB, *et al.* Overall survival, prognostic factors, and repeated surgery in a consecutive series of 516 patients with glioblastoma multiforme. *Acta Neurol Scand* 2010;**122**:159–167. doi:10.1111/j.1600-0404.2010.01350.x

90. Gutman DA, Cooper LAD, Hwang SN, *et al.* MR Imaging Predictors of Molecular Profile and Survival: Multi-institutional Study of the TCGA Glioblastoma Data Set. *Radiology* 2013;**267**:560–9. doi:10.1148/radiol.13120118

91. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: : R Foundation for Statistical Computing 2015. http://www.R-project.org/

92. RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA: : RStudio, Inc. 2015. http://www.rstudio.com/

93. Harrell FE. *rms: Regression Modeling Strategies*. 2015. http://CRAN.R-project.org/package=rms

94. Stupp R, Mason WP, van den Bent MJ, *et al.* Radiotherapy plus Concomitant and Adjuvant Temozolomide for Glioblastoma. *N Engl J Med* 2005;**352**:987–96. doi:10.1056/NEJMoa043330

95. Oken MM, Creech RH, Tormey DC, *et al.* Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol* 1982;**5**:649–55.

96. Observational Health Data Sciences and Informatics (OHDSI) Website. http://www.ohdsi.org/

97. Observational Medical Outcomes Partnership (OMOP) Website. http://omop.org/

98. OMOP Common Data Model (CDM). Observational Medical Outcomes Partnership Website. http://omop.org/CDM

99. Voss EA, Makadia R, Matcho A, *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc* Published Online First: 9 February 2015. doi:10.1093/jamia/ocu023

100. Data Mining Group - PMML version 4.2. http://www.dmg.org/pmml-v4-2.html

101. The National Lung Screening Trial Research eam. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N Engl J Med* 2011;**365**:395–409. doi:10.1056/NEJMoa1102873

102. Aberle DR, DeMello S, Berg CD, *et al.* Results of the Two Incidence Screenings in the National Lung Screening Trial. *N Engl J Med* 2013;**369**:920–31. doi:10.1056/NEJMoa1208962

103. Tammemagi CM, Pinsky PF, Caporaso NE, *et al.* Lung Cancer Risk Prediction: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Models and Validation. *J Natl Cancer Inst* 2011;**103**:1058–68. doi:10.1093/jnci/djr173

104. Tammemägi MC, Katki HA, Hocking WG, *et al.* Selection Criteria for Lung-Cancer Screening. *N Engl J Med* 2013;**368**:728–36. doi:10.1056/NEJMoa1211776

105. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans Pattern Anal Mach Intell* 1984;**PAMI-6**:721–41. doi:10.1109/TPAMI.1984.4767596

106. Andrieu C, Freitas N de, Doucet A, *et al.* An Introduction to MCMC for Machine Learning. *Mach Learn* 2003;**50**:5–43. doi:10.1023/A:1020281327116

107. Eaton ML. *Multivariate statistics: a vector space approach*. Beachwood, Ohio: : Institute of Mathematical Statistics 2007.

108. Fasiolo M. *An introduction to mvnfast. R package version 0.1.4.* 2014. http://cran.r-project.org/web/packages/mvnfast/vignettes/mvnfast.html

109. Rohatgi A. *WebPlotDigitizer*. 2015. http://arohatgi.info/WebPlotDigitizer

110. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med* 2005;**24**:1713–23. doi:10.1002/sim.2059

111. Wilson PWF, D'Agostino RB, Levy D, *et al.* Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* 1998;**97**:1837–47. doi:10.1161/01.CIR.97.18.1837

112. Hippisley-Cox J, Coupland C, Vinogradova Y, *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007;**335**:136. doi:10.1136/bmj.39261.471806.55

113. Goff J David C, Lloyd-Jones DM, Bennett G, *et al.* 2013 ACC/AHA Guideline on the Assessment of Cardiovascular RiskA Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol* 2014;**63**. doi:10.1016/j.jacc.2013.11.005

114. Tammemägi MC, Church TR, Hocking WG, *et al.* Evaluation of the Lung Cancer Risks at Which to Screen Ever- and Never-Smokers: Screening Rules Applied to the PLCO and NLST Cohorts. *PLoS Med* 2014;**11**:e1001764. doi:10.1371/journal.pmed.1001764

115. Lung Cancer Risk Calculators. https://brocku.ca/lung-cancer-risk-calculator

116. Therneau TM. *A Package for Survival Analysis in S*. 2015. http://CRAN.R-project.org/package=survival

117. McLendon R, Friedman A, Bigner D, *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**:1061–8. doi:10.1038/nature07385

118. Madhavan S, Zenklusen J-C, Kotliarov Y, *et al.* Rembrandt: helping personalized medicine become a reality through integrative translational research. *Mol Cancer Res MCR* 2009;**7**:157–67. doi:10.1158/1541-7786.MCR-08-0435

119. Colman H, Zhang L, Sulman EP, *et al.* A multigene predictor of outcome in glioblastoma. *Neuro-Oncol* 2010;**12**:49–57. doi:10.1093/neuonc/nop007

120. Hegi ME, Diserens A-C, Gorlia T, *et al.* MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *N Engl J Med* 2005;**352**:997–1003. doi:10.1056/NEJMoa043331

121. Karayan-Tapon L, Quillien V, Guilhot J, *et al.* Prognostic value of O6-methylguanine-DNA methyltransferase status in glioblastoma patients, assessed by five different methods. *J Neurooncol* 2010;**97**:311–22. doi:10.1007/s11060-009-0031-1

122. Wiewrodt D, Nagel G, Dreimüller N, *et al.* MGMT in primary and recurrent human glioblastomas after radiation and chemotherapy and comparison with p53 status and clinical outcome. *Int J Cancer* 2008;**122**:1391–1399. doi:10.1002/ijc.23219

123. Chinot OL, Barrié M, Fuentes S, *et al.* Correlation Between O6-Methylguanine-DNA Methyltransferase and Survival in Inoperable Newly Diagnosed Glioblastoma Patients Treated With Neoadjuvant Temozolomide. *J Clin Oncol* 2007;**25**:1470–5. doi:10.1200/JCO.2006.07.4807

124. Rivera AL, Pelloski CE, Gilbert MR, *et al.* MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. *Neuro-Oncol* 2010;**12**:116–21. doi:10.1093/neuonc/nop020

125. Kang H-C, Kim C-Y, Han J, *et al.* Pseudoprogression in patients with malignant gliomas treated with concurrent temozolomide and radiotherapy: potential role of p53. *J Neurooncol* 2011;**102**:157–62. doi:10.1007/s11060-010-0305-7

126. Korn EL, Simon R. Measures of explained variation for survival data. *Stat Med* 1990;**9**:487–503. doi:10.1002/sim.4780090503

127. Royston P. Explained variation for survival models. *Stata J* 2006;**6**:83–96.

128. Kerr KF, Wang Z, Janes H, *et al.* Net Reclassification Indices for Evaluating Risk-Prediction Instruments: A Critical Review. *Epidemiol Camb Mass* 2014;**25**:114–21. doi:10.1097/EDE.0000000000000018

129. Steyerberg EW, Moons KGM, van der Windt DA, *et al.* Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Med* 2013;**10**. doi:10.1371/journal.pmed.1001381