©Copyright 2017
Tracey Marsh

Distribution-Free Approaches to Assessing the Potential Clinical Impact of Biomarkers

Tracey Marsh

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Marco Carone, Chair

Margaret Pepe

Holly Janes

Program Authorized to Offer Degree:
Public Health - Biostatistics

University of Washington

Abstract

Distribution-Free Approaches to Assessing the Potential Clinical Impact of Biomarkers

Tracey Marsh

Chair of the Supervisory Committee:
Professor Marco Carone
Department of Biostatistics

Recent advances in basic science, combined with new technologies that enable measurement of sophisticated biological processes, present numerous opportunities for advancing the clinical care of patients. A basic tenet of stratified medicine is that utilization of biomarkers can improve identification of which patients may benefit from a particular medical intervention. A precursor to the employment of biomarkers in standard healthcare practices should be a population-level assesment of their impact. Additionally, evaluating biomarkers in accordance with possible clinical applications, at earlier stages of the development pipeline, is important for prioritizing candidates based on the ultimate goal of translating research into improved patient outcomes. In this dissertation, we consider two measures of impact, each relevant for distinct applications of biomarkers to refining medical care. The first measure, net benefit, applies to evaluating biomarkers in clinical decision rules that can guide whether or not a particular clinical intervention is recommended to a patient. The second, a marginal measure of additive interaction, applies to evaluating biomarkers that may be used to define a subgroup of patients for which a treatment may be more, or less, effective than for the whole. The corresponding estimators are either empirical or may be constructed using more general nonparametric approaches. The statistical focus is on efficient inference, an important aspect of evaluating evidence for the adoption of a clinical decision rule in practice or identification of a population for whom an intervention is beneficial.

TABLE OF CONTENTS

LIST OF SYMBOLS

Part I - Clinical Decision Rules

D - disease status $= \begin{cases} 1 & \text{diseased} \\ 0 & \text{non-diseased} \end{cases}$

ρ - outcome probability $= Pr\,(D = 1)$

X - intervention assignment $= \begin{cases} 1 & \text{intervention} \\ 0 & \text{no intervention} \end{cases}$

W - covariates

R - clinical decision rule; $x = R(w)$

r - risk function; $r(w) = Pr(D = 1 \mid W = w)$

r_T - high-risk threshold

$R(\cdot; r, r_T)$ - clinical risk rule; $x = R(w; r, r_T) = I[r(w) > r_T]$

TPR - true-positive rate $= Pr\,(X = 1 \mid D = 1)$

FPR - false-positive rate $= P\,(X = 1 \mid D = 0)$

FNR - false-negative rate $= P\,(X = 0 \mid D = 1)$

TNR - true-positive rate $= P\,(X = 0 \mid D = 0)$

B^{case} - benefit of intervention, over non-intervention, for a case (subject with $D = 1$)

B^{ctrl} - benefit of non-intervention, over intervention, for a control (subject with $D = 0$)

ω - population-level control-to-case benefit ratio $= \frac{\mathrm{B}^{case}}{\mathrm{B}^{ctrl}} \frac{1-\rho}{\rho}$

sNB - standardized net benefit

opt-in formulation; $\text{sNB}^N = \text{TPR} - \omega\text{FPR}$

opt-out formulation; $\text{sNB}^A = \text{TNR} - \frac{1}{\omega}\text{FNR}$

NB - net benefit (unstandardized);

opt-in formulation; $\text{NB}^N = \rho \cdot \text{sNB}^N$

opt-out formulation; $\text{NB}^A = (1-\rho) \cdot \text{sNB}^A$

Part II - Additive Interaction

Y - outcome status $= \begin{cases} 1 & \text{outcome present} \\ 0 & \text{outcome absent} \end{cases}$

X - treatment assignment $= \begin{cases} 1 & \text{treatment} \\ 0 & \text{no treatment} \end{cases}$

B - biomarker status $= \begin{cases} 1 & \text{present/high} \\ 0 & \text{absent/low} \end{cases}$

W - covariates

Δ - phase-II sampling indicator

$Q_{jk}(w)$ - conditional outcome probability $= Pr\left(Y = 1 \mid B = j, X = k, W = w\right)$

$g_B(j \mid k, w)$ - conditional probability of biomarker status $= Pr\left(B = j \mid X = k, W = w\right)$

$g_X(k \mid w)$ - conditional probability of treatment assignment $= Pr\left(X = k \mid W = w\right)$

$\pi(y, x, w)$ - phase-II sampling probability $= Pr\left(\Delta = 1 \mid Y = y, X = x, W = w\right)$

AERI - additive excess risk due to interaction $= Q_{11} - Q_{01} - Q_{10} + Q_{00}$

$Q_{bx} := Pr\left(Y = 1 \mid B = b, X = x\right)$

P-AERI - population adjusted AERI $= \mathbb{E}_W\left[Q_{11}(W) - Q_{01}(W) - Q10(W) + Q_{00}(W)\right]$

$Q_{bx}(w) := Pr\left(Y = 1 \mid B = b, X = x, W = w\right)$

F - annotation for full-data scenario

E - annotation for experimental-data scenario (two-phase sampling scheme)

NP - annotation for unrestricted (non-parametric) estimation

$\perp\!\!\!\perp$ - annotation for estimation reflecting independence

Common Methodological Framework

O - random variable

o - realization of random variable O

RAL - regular and asymptotically linear (estimator)

IF - influence function of a RAL estimator

P - probability distribution

$\mathcal{M}$ - statistical model (set of probability distributions)

0 - annotation for true data-generating mechanism

n - annotation for estimator/estimate constructed from n observations

$N(\mu, \sigma^2)$ - normal distribution with mean μ and variance σ^2

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Stratified Medicine

Consideration of clinically relevant personal factors when making healthcare decisions holds much promise for individual wellness and effective utilization of medical resources. Determining which personal factors are relevant to a given disease or health condition and how they relate to clinical actions is the subject of much research. In its broadest sense, stratified medicine simply describes the organization of patients into subgroups that can be used to guide medical care. Stratified approaches apply to every aspect of care, including preventative, screening, diagnostic, and therapeutic settings.

The quote: "It is far more important to know what person the disease has than what disease the person has.", attributed to Hippocrates, is regularly cited to substantiate the claim that the idea of stratified medicine is not new. Identification of patient blood types and understanding their role in successful treatment by blood transfusion is a simple example that was established in 1901 (Giangrande, 2000). More recently, in 2013 the U.S. Preventative Services Task Force recommended regular screening for lung cancer, using low-dose computed tomography, only for individuals aged 55-80 years who had sufficiently extensive and recent smoking history (Moyer, 2014). The act of cigarette smoking was established as a risk factor for lung cancer in the middle of the 20th century through a combination of basic lab science, pathology, and epidemiology studies (Proctor, 2012). What is fairly new, is a significant expansion in the quantity of personal attributes that can be measured.

Recent advances in basic science, combined with new technologies that enable measurement of sophisticated biological processes, innovations in imaging capability and expanded computational resources, present numerous opportunities for refining patient care. Enthusiasm for the promise of these newfound sources of data is fueled by already realized successes.

One example is the 2012 approval, by the U.S. Food and Drug Administration, of the drug ivacaftor (brand name Kalydeco®) exclusively for the treatment of cystic fibrosis among patients with one of two designated mutations in their CTFR gene. An earlier example is the 1998 approval of trastuzumab (brand name Herceptin®) for the treatment of women with HER2-positive breast cancer. These particular successes are often touted as models for the potential of pharmacogenomics to transform medicine by addressing underlying causes of disease and that the genomics of the tumor is itself variable and can be the source of clinically relevant factors (Simoncelli et al., 2013).

Interest in applying the current confluence of increased technological ability and scientific understanding to medical practice has culminated in a Zeitgeist that extends beyond national or disciplinary boundaries. Summits and conferences have convened throughout the world as a forum of bringing together vested parties, including: industry, academia, clinicians, patients, regulatory agencies, research foundations, and government institutes, to assess the current state, discuss challenges and envision the future of stratified medicine (FORUM, 2015; Uni, 2016; Li, 2016; Moch et al., 2012). Governments are making precision medicine a national priority by funding ambitious programs such as the Precision Medicine Initiatives (PMI) in the U.S. and in China, announced by President Obama during the 2015 State of the Union Address and by the Chinese government during the National People's Congress sessions the following year. Both initiatives include establishing large cohorts of patients who consent to providing their genetic information for research. In Great Britain, the Department of Health created, and both funds and owns, the company Genomics England to oversee the "100,000 Genomes Project" (Genomics England, 2012). This project aims to sequence complete genomes of 100,000 British residents, primarily focused on patients with a rare disease or cancer, for use in research and with the aim of translation into national healthcare practices. As pointed out in the 2013 U.S. Food and Drug Administration's (FDA) report"Paving the Way for Personalized Medicine", stratified medicine is not restricted to genetic-based care and in fact, "a vast variety of medical devices can be used in a personalized approach to improve patient outcomes" (Simoncelli et al., 2013). Of interesting

potential, the cohort of one million American patients set forth in the US PMI intends to allow volunteers to contribute "diverse sources of data - including medical records; profiles of the patients genes, metabolites (chemical makeup), and microorganisms in and on the body; environmental and lifestyle data; patient-generated information; and personal device and sensor data" (Office of the Press Secretary, 2015).

At this point in time, the terms personalized, precision, and stratified medicine are used as synonyms by some and to differentiate subtle nuances by others. We favor the term stratified as it conveys the reality that sound evidence supporting decisions accounting for personal factors is based on an average, e.g., average risk or average benefit, of other patients with similar attributes. Throughout the official White House presentation of the US PMI appear lamentations of the "one-size fits all" approach to healthcare that was designed for the "average patient" (Office of the Press Secretary, 2015). A basic goal of stratified medicine is to improve identification of which patients may benefit from a particular medical intervention. A practical approach is through utilizing new information to create smaller groups of patients with common standards of care. Realizing the promise of stratified medicine will require establishing a "strong evidence base to test the value of stratified approaches" and methodologies that support testing "the efficacy of decision rules and treatments" (FORUM, 2015).

Biomarker Development

A working definition of the term biomarker, a biological marker, is an objective characteristic, of a patient or disease entity, that can be measured reproducibly and can be shown to be an indication of either normal or pathological biological processes (Strimbu and Tavel, 2010). Strictly speaking, most attributes of a patient or their medical condition could be considered a biomarker; familiar examples are human body temperature and insulin level. In practice, the term biomarker is often reserved for novel quantities reliant on sophisticated technologies for measurement. Some examples arise in genomics, proteomics, glycomics, metabolics, and complex imaging features. Under the umbrella of biomarker development

falls momentous research efforts towards identifying relationships between the quantities that are newly accessible and clinically relevant processes. The ultimate goal is to employ established biomarkers to increase treatment options, improve health outcomes and target clinical practice for efficient utilization of resources and improved patient outcomes.

The development process for pharmaceuticals has been organized into a well-established sequence of stages for some time (ICH Guideline E8, 1997). The results of each stage inform research at later stages, foremost in deciding whether to continue or terminate a given research direction. The earliest stages involve discovery research in pre-clinical settings which may involve numerous candidates. Generally, well-defined agents with sufficient supporting evidence proceed to clinical research in human subjects. A candidate therapy that successfully gains licensure will continue to be studied in a post-market safety monitoring stage.

The clinical stage of therapeutic development is further divided into four phases of clinical trials, each having distinct scope and objectives. Phase I trials are pharmacological in nature and are central to studying how characteristics of the therapeutic agent observed in pre-clinical studies transfer to human subjects. Candidates with acceptable profiles continue to Phase II trials, designed as an initial assessment of therapeutic efficacy which is further assessed in confirmatory Phase III trials. Phase IV trials are conducted after licensure, often to refine dosing or gather additional information on potential interactions with other drugs; they are distinct from the stage of routine safety surveillance.

A collaboration between regulatory agencies, industry and academia produced the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH, 1990). Since its formation in 1990, the ICH has continued to release comprehensive and detailed guidelines related to many aspects of drug development, such as quality assurance, efficacy, and safety. Specific guidelines range from topics such as reporting standards for biomarker data (ICH Guideline E16, 2010) to statistical principles for trials (Lewis, 1999). Efforts to bring similar levels of rigor, efficiency, and quality assurance to biomarker development have begun and continue to be proposed. The Tumor Marker Utility Grading System (TMUGS) was put forth as framework for evaluating the

clinical utility of tumor biomarkers (Hayes et al., 1996; Prentice, 1996). An organization of biomarker development into a sequence of distinct phases analogous to those for pharmaceutical development has been proposed for clinical contexts of screening (Pepe et al., 2001). Bossuyt et al. (2003a,b) offered guidelines for establishing rigorous reporting standards in diagnostic contexts. A workshop on co-developing biomarkers and pharmaceuticals, jointly sponsored by the National Cancer Institute (NCI) and the FDA, organized parallel research goals and considerations through a common staged pathway (Taube et al., 2009).

As for pharmaceutical development, distinctions between early, mid, and late stages of biomarker development can be observed in all aspects of research. As the scope of research proceeds from pure discovery, through characterization of performance and clarifications of clinical applications, to evaluation of proposed use in practice, so do all aspects of study design. Early stage research may be conducted on convenient collections of samples, followed by efficient use of mid-sized focused case and control samples, and then a comprehensive prospective study in the full cohort of interest. Ideally, evaluation of a promising candidate culminates in a dedicated clinical trial, though this is not always feasible or deemed necessary. Sample sizes generally increase as more evidence supports the potential utility of a biomarker as does the amount of detail specified prior to each study. Primary analysis measures progress from rough indications of activity to more thorough characterizations of discrimination, and culminate in assessments of clinical utility. Specific to biomarker development, the assay, or means of measuring the biomarker, can be as novel as the clinical relevance of a biomarker. Consequently, assay development itself often occurs in concert with biomarker development and must progress to stable processes with good operating characteristics that can be commercialized, if a biomarker is to successfully translate into improved clinical practices. The importance of assay specifics to defining tumor markers was underscored by TMUGS.

Whereas final evaluation of clinical impact is ideally established in a trial, much prior biomarker research can be conducted *ex vivo* on biological samples collected from humans. These samples are collected relatively non-invasively, such as through nasal swabs, urine sam-

ples, and blood draws, or incidental to standard clinical care, for example tissue collected during a planned surgery. This is in contrast to the many stages of *in vivo* experiments conducted on humans in therapeutic trials. Repositories of biological samples have been, and continue to be, created for the express purpose of biomarker development research. Through the Parkinson's Disease Biomarker Program, the National Institute of Neurological Disorders and Stroke has funded a repository for discovery research (Rosenthal et al., 2015). The Reproductive Medicine Network, funded through the National Institute for Child Health and Human Development, set out to further the impact of completed clinical trials by making remaining biological samples available for secondary research efforts (Casson et al., 2011). The National Cancer Institute oversees the Early Detection Research Network, one component of which is biological reference sets made available to further the mission of accelerating biomarker translation into clinical application (Srivastava and Kramer, 2000). These are to name only a few of the many similar efforts.

Contents and Organization of Dissertation

The work in this dissertation pertains to assessing the potential value of biomarkers for making stratified medical recommendations. Two distinct measures, relevant in different phases of biomarker development, are each the focus in one of the two parts that constitute this thesis.

In Part I we focus on the use of biomarkers in defining clinical decision rules. A clinical decision rule is quite general and can be applied to various medical applications such as diagnosis or prevention of an event that could occur in the future. An example in the diagnostic setting arises in the use of a score derived from clinical symptoms on presentation to the ER for acute wrist injury; the decision is whether or not to conduct an x-ray for definitive diagnosis of fracture when currently most such procedures are retrospectively determined unnecessary. An example in a preventative setting is the use of cholesterol levels, in addition to epidemiological risk factors such as age, sex, and family history, to predict 5-year risk of a cardiac event; the decision could be whether or not to recommend cholesterol-lowering

medications as a preventative measure.

Actions have consequences and employing a rule to group individuals for recommended receipt of an intervention, or not, has both costs and benefits that need to be accounted for. Net benefit is a summary measure that reflects the population-level trade-offs associated with adopting a clinical rule in practice. The primary contribution made in this part is to establish the asymptotic behavior of empirical estimators of the net benefit of a pre-specified clinical decision rule. Analytic descriptions of variability surpass bootstrapped estimates by providing insight and guiding the design of studies intended to evaluate net benefit.

The potential clinical value of a decision rule is one of the final points of evaluation for a promising candidate. Gold-standard evidence for the clinical utility supporting either further assessment or adoption of a rule should be obtained from a large prospective cohort that represents the population of interest. This provides the primary context in which analytic inference for net benefit is established and for extensions to differences of net benefit between competing rules, decision curves, and censored outcomes. Given the importance of case-control samples in biomarker development, especially in mid-to-late phases of research, inference for net benefit from these studies and their design is also addressed.

In Part II our focus shifts to the problem of identifying biomarker-defined subgroups with a differential response to a treatment. In the development of therapeutics, biomarkers that modify treatment effect can give insight into biological mechanisms of either disease etiology or treatment mechanisms. They also have the potential to provide a basis for subgroup-specific FDA approval, as for trastuzumab. Identification of biomarkers that interact with treatment is often conducted during early-to-mid stages of development from case-control samples, often nested within a clinical trial.

We apply a bivariate version of the average exposure-specific outcome to the task of defining a measure of interaction that has a characteristic profile seemingly unique among available methods. Namely, the estimand includes adjustment by covariates, has a marginal and prospective interpretation, and is on the additive scale. The proposed estimator, from a case-control sample nested within a clinical trial, accommodates nonparametric approaches

and can consequently be made robust to model misspecification. In this example, particular interest and effort is directed towards exploiting the known independence between baseline covariates, such as a biomarker measurement, and randomized treatment assignment for efficiency gain.

The author holds the view that public health science, when possible, is best served by statistics that do not rely on presumed parametric structures. The estimands studied herein are distribution-free in that they are well defined over all possible data-generating mechanisms. Further, the corresponding estimators are either empirical or may be constructed using more general nonparametric approaches. The estimators studied in Parts I and II are regular and asymptotically linear (RAL). The theory of RAL estimators and influence function techniques are employed throughout this dissertation. Key results and references for this common methodological framework are provided in Appendix A.

Part I

VARIABILITY IN ESTIMATES OF THE NET BENEFIT OF PRESPECIFIED CLINICAL DECISION RULES

Chapter 1

BACKGROUND

One aspect of stratified medicine is to transform salient information about an individual patient into a decision of whether or not to take a clinical action in regards to a particular potential health outcome. Personal factors could inform decisions by discriminating between patients more likely to benefit from receiving the intervention from those more likely to benefit from avoiding unnecessary care. Standardization measures taken by large health-management organizations, guidelines issued by professional societies and recommendations from task-forces to policy makers create a fertile ground from which the wide-scale adoption of evidence-based rules is possible. In this part of the dissertation we are concerned with assessing the potential clinical value of adopting a clinical decision rule in practice.

In this chapter, we start by introducing a motivating example. After formalizing the notion of a clinical decision rule, with emphasis on rules that are derived from risk prediction models, we introduce various formulations of net benefit. Net benefit, the measure of clinical value studied throughout this work, reflects the trade-offs between the options of intervening and not and we review means of gathering information on cost and benefits. We then introduce the empirical estimator of net benefit and conclude by discussing the importance of analytic formulations of variance.

1.1 Motivating Example: Diagnosis of Incidental Pulmonary Nodules

We set the stage with an example that will be referred to throughout the developments in this part of the dissertation. Suppose a patient received a computerized tomography (CT) scan, conducted as rountine work-up for a shoulder injury, which revealed an indeterminate nodule on a portion of the lung unintentionally included in the image. The patient is then in

a position of determining what action should be taken in response to this incidental finding. The question at hand is whether or not the nodule is cancerous. The definitive diagnosis results from a thoracoscopic surgery with possible resection. Two available actions are to "keep an eye on it" through periodic imaging or to undergo the surgical procedure.

As of 2013, over 135,000 incidental nodules were detected in the U.S.A. each year, and this accounts for over 90% of all solitary pulmonary nodules detected (Furman et al., 2013). The management of such patients presents a challenging clinical problem. Two sets of guidelines, from the American College of Chest Physicians (AACP) (Tan et al., 2003) and from the Fleischner Society (MacMahon et al., 2005, 2017), have been issued to support consistent standards in lung care management, followed by harmonization efforts between the two (Naidich et al., 2013). All recommendations first stratify patients based on the size of the nodule. Nodules less than 8mm are generally considered low-risk and are subject to continued imaging at various intervals. Nodules greater than 8mm are inherently considered higher risk and have distinct management guidelines.

Patients having an incidental nodule with diameter greater than 8mm comprise the population for which improved decision making is sought. The guidelines for these patients recommend estimating the probability of malignancy prior to conducting any other diagnostic testing. Factors contributing to risk assessment are: age, BMI, smoking and cancer history as well as features of the nodule measurable from the original scan. Patients considered at high risk ($> 65\%$ probability) are recommended to undergo surgical diagnosis, in particular a thoracoscopy (Gould et al., 2013).

A recent review of current lung management practices called for better adherence to guidelines by physicians, underscored the importance for improved risk prediction, and proposed the establishment of a national tracking system (Sethi and Parrish, 2016). Current risk assessment, based on epidemiological risk factors and simple characteristics of the nodule, places many patients in an intermediate range ($10 - 65\%$) of risk. While physicians are generally comfortable with continued observation for low-risk patients and surgical diagnosis for high-risk patients, care for intermediate-risk patients is less clear. Biomarkers that can

improve discrimination between patients with incidentally identified pulmonary nodules who do and do not have lung cancer have the potential to reassign currently intermediate-risk patients into either the low or high-risk classes that have more clear treatment decisions.

Numerous such endeavors are already underway. A diagnostic biomarker development program within the Early Detection and Research Network, funded by the National Cancer Institute, is one such example. The Percepta® genomic signature, analyzed on samples obtained by bronchial brushings, is one test available for improving diagnostic decision making following brochoscopy (Silvestri et al., 2015; Ferguson et al., 2016; Vachani et al., 2016). Less invasively measured biomarkers that could be applied to the reduction of unnecessary surgeries conducted to diagnose an incidental lung nodule include: imaging features on CT scans, signatures of circulating micro RNA obtainable from serum, plasma or sputum, and proteomic signatures analyzed from blood (Atwater et al., 2016).

1.2 Clinical Decision Rules

Formally, a decision rule is a function that maps an observation to an action (Berger, 2013, pp. 9). Using $\mathcal{W}$ to denote the set of possible patient-specific measurements and $\mathcal{X}$ the set of possible actions, a decision rule rule can be written $R : \mathcal{W} \to \mathcal{X}$. We restrict our attention to binary decision rules for which $\mathcal{X} = \{0, 1\}$, where 1 indicates clinical intervention and 0 no intervention. We use the term *clinical* decision rule to emphasize that the decision rule will be both employed and evaluated within a clinical context defined by elements such as the underlying disease, an observable clinical outcome, and a specific intervention. Clinical decision rules can be employed in a variety of medical contexts such as diagnosis or prevention.

Rules Derived From Risk Models

The formal definition of a decision rule is quite general. Decision rules derived from risk models have many desirable properties. First and foremost, a risk model gives both patient and provider meaningful information about the probability of the outcome. A key feature

of employing a risk model for making decisions based on predictions is the risk threshold (Steyerberg et al., 2010, 2012; Van Calster et al., 2013). A single risk threshold defines the boundary between patients considered high-risk and not.

Letting W represent patient-specific variables and using D to denote a binary clinical outcome, where $D = 1$ denotes presence and $D = 0$ denotes absence of the disease or medical condition, a risk model r calculates the probability of the outcome among individuals with similar attributes. In symbols, $r(w) = Pr(D = 1 \mid W = w)$, for each value w of W. For a given risk threshold r_T and risk model r, the decision rule R that assigns intervention ($x = 1$) to high risk individuals and no intervention ($x = 0$) to all others, can be expressed: $R\left(w; r, r_T\right) = I\left[r(w) > r_T\right]$.

For the example of an incidental lung nodule, introduced in Section 1.1, the Mayo risk model r_{Mayo} is a validated model for calculating the probability that the nodule is cancerous ($D = 1$) based on patient variables W including age, smoking and cancer history, and features of the nodule readily readable from the image (Swensen et al., 1997). The ACCP guidelines consider a probability of malignancy greater than 65% high enough risk to warrant prompt surgical diagnosis, whereas lower risks are recommended treatment plans that involve continued monitoring and noninvasive diagnostic procedures. The simplified decision of whether or not to perform definitive surgical diagnosis, can be written $x := R\left(w; r_{\text{Mayo}}, r_T = 0.65\right) = I\left[r_{\text{Mayo}}\left(w\right) > 0.65\right]$.

The results herein pertain to evaluating general clinical decision rules. Our notation will maintain this generality while properties specific to rules derived from risk models will be highlighted. For concreteness we will often adopt the perspective of rules derived from risk models in illustrations and simulations.

Classification

A binary clinical decision rule R places patients into either treat or don't treat classes: $X = 1$ or $X = 0$, where $X := R(W)$. The underlying motivation to develop such a rule is the notion that not only should some patients be treated while others should not, but that there are

patient-specific pieces of information W that can inform this decision. Ideally, only patients with the condition, e.g., their nodule is cancerous, or who would experience the outcome in the absence of treatment, e.g., first coronary event within 10 years, would be intervened on. We refer to such patients as cases, or would-be cases, to emphasize the potential outcome that would have occurred in the absence of treatment. Similarly, we refer to all other patients as controls, or would-be controls. The classification terminology of medical tests (Pepe, 2003, p. 14), summarized in Table 1.1, describes the agreement between case-control status ($D = 1$ or $D = 0$) and the intervention class ($X = 1$ or $X = 0$) resulting from a decision rule. The words true and false describe the concordance or discordance between the decision and case-control status. The words positive or negative describe the decision to intervene or not; when the rule is derived from a risk model, these words also describe the classification as high-risk or not.

		Clinical Outcome	
		$D = 0$	$D = 1$
Decision	$X = 0$	True negative	False negative
	$X = 1$	False positive	True positive

Table 1.1: Classification of intervention decisions by clinical outcome status.

The probabilities of being assigned the intervention class by the decision rule, conditional on clinical outcome are:

$$\text{TPR}_R := Pr\left(X = 1 \mid D = 1\right)$$
$$\text{FPR}_R := Pr\left(X = 1 \mid D = 0\right),$$

where TPR and FPR abbreviate true-positive rate and false-positive rate, respectively, and the subscript is optionally employed to emphasize the dependence on a particular rule R; $X = R(W)$. While it has been noted that these are probabilities, not rates, and are better termed fractions (Pepe, 2003, p. 15), in this dissertation, where there is no potential for

confusion with other true rates, we adopt the common convention. The true and false-negative rates (TNR and FNR) can be defined similarly. We note the two relationships: TNR = 1 - FPR and FNR = 1 - TPR; working with either pair of probabilities is sufficient to summarize the classification accuracies of a decision rule.

Universal and Oracle Rules

We now introduce four simple decision rules that do not rely on a risk model or any available patient-specific information. The first two are universal in that they make the same decision for all individuals. These are the treat-all and treat-none rules, denoted by A or N, respectively. The decision to intervene on no-one is correct for all controls, but incorrect for all cases; this rule has only true-negatives and false-negatives. Conversely, the decision to intervene on everyone is incorrect for all controls, but correct for all cases; this rule has only true-positives and false-positives. The two other rules also achieve extreme classification probabilities as if they had divine knowledge of the unknown case-control status and are thus termed 'oracle' rules; they are not practical rules but are of theoretical interest. The completely incorrect rule I assigns all cases to the no-intervention action and all controls to intervention; this rule has only false-positives and false-negatives. The perfect rule P assigns all cases to the intervention and all controls to no intervention; this rule has only true-positives and true-negatives. The two oracle rules bound the classification performances of any practical decision rule in the sense that: any rule R has both sensitivity and specificity at least as good as that of I and no better than that of P.

1.3 Measures of Net Benefit

Employing a clinical decision rule to guide intervention decisions produces actions, the consequences of which must be included in any assessment of value. Decision analytic measures have been proposed as a means of reflecting both the costs and benefits associated with possible outcome and intervention combinations. In particular, Net Benefit (NB), first introduced by Peirce (1884) and more recently promoted for use in clinical applications by

Elkin and Vickers (2006), has become a popular metric for evaluating the potential value of adopting a clinical decision rule in practice.

Standardized net benefit sNB_R^N is a measure of clinical utility from employing a decision rule R in a population. It is defined as a weighted difference between the true and false-positive rates of the rule:

$$\text{sNB}^N := \text{TPR} - \omega\text{FPR}, \tag{1.1}$$

where the weight, $\omega := \frac{B^{ctrl}}{B^{case}}\frac{1-\rho}{\rho}$, depends on the probability of the clinical outcome $\rho = Pr(D = 1)$ in the population, and a fixed control-to-case benefit trade-off: $\frac{B^{ctrl}}{B^{case}}$; B^{ctrl} is the benefit to a control from avoiding treatment rather than unnecessarily receiving treatment; similarly, B^{case} is the benefit to a case from receiving treatment rather than not receiving treatment. The weight ω reflects the population-level trade-offs between correctly caring for all controls (treatment not recommended) relative to correctly caring for all cases (treatment recommended). These benefits will be explicitly defined in terms of utilities in the next section. The dependence of sNB^N, TPR, and FPR on the rule (e.g., TPR_R) has been suppressed for brevity.

Consider a control-to-case benefit ratio of 1:2; the benefit of a would-be case receiving the intervention is twice as much as the benefit of a would-be control avoiding the intervention. For a disease with 10% prevalence in the population, the weight is $\frac{1}{2}\frac{0.9}{0.1} = 4.5$. At the population-level, in which there are many more controls than cases, the benefit of properly caring for all controls outweighs that of properly caring for all cases by 4.5 times. In this sense, net benefit penalizes the false-positive rate of a clinical decision rule 4.5 times as much as it rewards the true-positive rate. Consequently, a rule must have a true-positive rate more than 4.5 times the false-positive rate in order to achieve $\text{sNB}^N > 0$ and potentially have more clinical utility than the universal rule of treating no-one.

This formulation of sNB is intended for evaluating a clinical rule as an opt-in decision, where a subset of individuals are identified to receive intervention that would not otherwise be administered routinely. In terms of an underlying risk model, high-risk individuals are identified for intervention that is considered less reasonable for low-risk individuals. This

measure implicitly evaluates a rule in terms of its improvement over the universal decision to treat no-one, as indicated by the superscript N, and confirmed by noting $\text{sNB}_N^N = 0$.

There is an analogous measure of standardized net benefit for evaluating a clinical decision rule in an opt-out context, i.e., for the identification of individuals who could reasonably not undergo an intervention that is otherwise routinely administered. Standardized net benefit sNB^A is a measure of clinical usefulness defined:

$$\text{sNB}^A = \text{TNR} - \frac{1}{\omega}\text{FNR}, \tag{1.2}$$

where TNR and FNR denote the true and false-negative rates and ω is the same weight as for sNB^N. This measure implicitly evaluates a clinical rule in terms of its improvement over the universal decision to treat everyone, $\text{sNB}_A^A = 0$, as indicated by the superscript A.

Relative Utility

Standardized net benefit was first introduced by Baker et al. (2009) using the term relative utility. The terminology *standardized net benefit* connects relative utility to net benefit (NB), the quantity employed by Elkin and Vickers (2006). The two versions differ by a factor of the outcome rate: $\text{sNB}^N = \rho \cdot \text{NB}^N$ and $\text{sNB}^A = (1-\rho) \cdot \text{NB}^A$. The desirable features of measuring net benefit on the standardized scale — a consistent range of values (maximum of 1), and better interpretation (Kerr et al., 2016; Pepe and Janes, 2013) — are ultimately due to its underlying derivation in terms of formal decision theory.

Suppose for each patient there is a common utility u_{dx} dependent on their case-control status $D = d$ and assigned intervention class $X = x$. For a fixed decision rule R, D and $X = R(W)$ are random variables with joint probability determined by the clinical population. The expected utility of R is a weighted sum of the utilities, $\mathbb{E}\left[u_{DX}\right] = \sum_{i,j=0}^{1} u_{ij} Pr(X = i \mid D = j) Pr(D = i)$, where the weights can be expressed in terms of the true- and false-positive rates of the rule and the outcome probability. We denote this quantity by $\mathbb{E}U_R$ to emphasize the dependence on the rule R. The difference in expected utility of using a rule R over treating no-one, $\mathbb{E}U_R - \mathbb{E}U_N$, can be expressed in terms two differences in utilities: $B^{case} := u_{11} - u_{10}$,

the benefit of a case receiving treatment over not, and $B^{ctrl} := u_{00} - u_{01}$, the benefit of a control avoiding treatment over unnecessarily being treated. Compared to treating no-one, B^{case} is gained among true-positives and B^{ctrl} is lost among false-positives (the second term in $\text{sNB}^N$is subtracted). Compared to treating everyone, B^{case} is lost among false-negatives (the second term in sNB^A is subtracted) and B^{ctrl} is gained among true-negatives.

The opt-in formulation of standardized net benefit can be written

$$\text{sNB}_R^N = \frac{\mathbb{E}U_R - \mathbb{E}U_N}{\mathbb{E}U_P - \mathbb{E}U_N}$$

which reveals its interpretation as the fraction of the possible increase in expected utility, over treating no-one, achieved by the rule R. The denominator represents the maximum increase possible, which is achieved by the perfect rule; $\text{sNB}_P^N = 1$. An analogous expression and interpretation holds for the opt-out formulation.

Assumptions

The value of net benefit depends directly on the utilities through the control-to-case benefit ratio. Establishing the utilities, even just the single ratio of differences in $\frac{\text{B}^{ctrl}}{\text{B}^{case}}$, is not a simple task. The reliance of net benefit on the trade-off ratio is fairly transparent and can be explored by straight-forward sensitivity analyses that calculate the net benefit, of a given rule, at a few reasonable trade-off ratios.

Somewhat more subtle is the supposition of common utilities, u_{dx}, made in the derivation of sNB through expected utilities. This is a strong assumption that is not always appreciated. For example, it has been pointed out that proposals for adjusting risk rules based on individual cost-benefit trade-offs have not properly reflected this assumption (Kerr et al., 2016). A realistic exception arises when age increases the impact of toxicity associated with the intervention and the utility of the intervention could be less for older patients than younger ones. If age is also a risk factor, then a risk-based decision rule could be assigning more older patients to intervention compared to younger ones. In terms of net benefit, each true-positive gains B^{case}, which may be more credit than is due for treating an older

would-be case patient, and each false-positive loses B^{ctrl}, which may understate the penalty for incorrectly treating an older would-be control patient.

The definition of net benefit in terms of expected utility also reveals a connection between the cost-benefit trade-offs and assumptions on adherence to intervention recommendations determined by a decision rule. The benefit of a true-positive is counted by sNB^N for each case correctly recommended intervention, and analogously for the benefit lost by each control incorrectly recommended intervention. If the benefit ratio is determined from evaluation of the intervention, assuming total adherence to the intervention recommendation, then this assumption is also reflected in the assessment of net benefit. In this case, the calculated net benefit would not account for physicians who make other recommendations or patients that choose other options. Similarly, if the benefit trade-off accounts for some level of non-adherence, as perhaps a ratio of average benefits, across adherent and non-adherent, then the calculated net benefit will reflect the same assumption on adherence.

The formulation of net benefit assumes that the cost of measuring the variables, from which decisions are made, is the same for all rules being considered. Since standardized net benefit implicitly contrasts against the expected utility of the universal decision to treat no-one (or perhaps all), which does not rely on any covariates, the presented formulation does not account for cost of measuring the predictors of the rule R. For simple variables, such as demographic and family history, this seems sensible. It could also be a reasonable simplification for routinely collected and processed laboratory values. However, for a novel biomarker that is financially or logistically expensive to obtain, this may be less reasonable and the estimate of net benefit will be overstated. If the cost of variable acquisition can be summarized in units of B^{case}, or B^{ctrl} for the opt-out formulation, then a measure of net benefit accounting for this reduction can be calculated (Baker et al., 2009). This accounting is also relevant when comparing the net benefit between two rules that have differing costs associated with measurement of the respective predictors. Otherwise, the difference in net benefits that do not explicitly adjust for this expense will be correct when the common costs of variable acquisition cancel.

To emphasize the dependence of net benefit on underlying assumptions, we can recognize it as a measure of the *potential* clinical utility of adopting a clinical decision rule in practice. Nonetheless, net benefit can still provide a useful summary measure of the population performance of a clinical decision rule. It is one metric that can be considered alongside other clinical utility measures, before a biomarker is used in clinical practice to make decisions and the consequent outcomes can be studied.

Identifiability

Implicit in the derivation of standardized net benefit is that D and ρ describe the outcome in the absence of intervention. When the standard of care is not described by the universal rule to treat no-one, this can lead to statistical identifiability issues since the outcome status measured, and outcome probability inferred from the study population, may reflect the standard of care. Numerous applications of rules for making opt-out decisions are motivated by the desire to reduce unnecessary interventions. In a diagnostic setting, it may be reasonable to believe that the outcome status is independent of whether or not the diagnostic intervention is performed. In the example of an incidental lung nodule, not only do we reasonably believe that the cancerous state of the nodule is not impacted by the surgical diagnosis, the fact that these surgeries are somewhat routinely conducted is precisely what provides the definitive case-control status necessary for assessing the potential clinical utility of a rule proposed to reduce unnecessary surgeries. The statistical identifiability of net benefit will need to be considered carefully for each clinical context.

1.4 Costs and Benefits

The utilities required to evaluate the expected utility represent a synthesis of benefits and costs: physical, emotional, financial and so forth on a common scale, and are thus difficult to elicit in practice. Standardized and unstandardized net benefit reduce the need for four distinct utilities to one ratio of differences. The greater feasibility of eliciting $\frac{B^{ctrl}}{B^{case}}$ was appreciated by Elkin and Vickers (2006) and used to promote net benefit.

Consider again a patient who has had a pulmonary nodule incidentally detected and must decide whether or not to undergo surgery to definitively establish or rule out a lung cancer diagnosis. Reasons for hesitation are that lung cancer is rare and the necessary procedure is quite invasive, expensive, and has an extended recovery period. However, lung cancer is a very serious outcome that could have better patient outcomes when detected early. Further, the anxiety associated with not knowing could be substantial. These are a few of the trade-offs that must be considered by patient and physician when deciding which course of action to follow.

Rational Risk Rules

When clinical decision rules are derived from a risk model, the notion of a rational decision connects the benefit trade-offs and the risk threshold. Rational decisions would have individuals elect to take the treatment under which they would expect the benefits to outweigh the costs (Van Calster et al., 2013; Kerr et al., 2016). A choice of risk threshold r_T implicitly reflects the relative costs and benefits of a break-even point at which either option, e.g., treatment or no treatment, is considered equally good and implies $\frac{B^{ctrl}}{B^{case}} = \frac{r_T}{1-r_T}$ (Pauker and Kassirer, 1975). This result can be derived in terms of the utilities u_{dx} introduced in Section 1.3. With ρ representing the probability of being a case, the expected utility of receiving treatment is $\rho \cdot u_{11} + (1-\rho) \cdot u_{10}$ while the expected utility of not receiving treatment is $\rho \cdot u_{10} + (1-\rho) \cdot u_{00}$. The break-even probability r_T satisfies: $r_T \cdot u_{11} + (1 - r_T) \cdot u_{10} = r_T \cdot u_{10} + (1 - r_T) \cdot u_{00}$, which can be algebraically manipulated into the above formula in terms of $B^{case} = u_{11} - u_{10}$ and $B^{ctrl} = u_{00} - u_{01}$. When the probability of being a case is greater than r_T, the expected utility of receiving treatment exceeds that of not receiving treatment and conversely for probabilities less than r_T. For example, employing 65% probability of malignancy as the high-risk threshold for an individual with an incidental lung nodule, used as a basis for recommendation to surgical diagnosis, implies a control-to-case trade-off ratio of 13:7 or $\frac{B^{ctrl}}{B^{case}} = 1.86$.

Conversely, a control-to-case benefit ratio determines the rational threshold according to:

$$r_T = \frac{\frac{B^{ctrl}}{B^{case}}}{1 + \frac{B^{ctrl}}{B^{case}}}.$$

In this sense, the rational risk threshold reflects the trade-off between costs and benefits, information beyond the purely statistical aspects of evaluating a decision rule, expressed on a common scale. The combination of a risk model with the rational threshold defines a rational risk rule. When evaluating the net benefit of a rational risk rule, the weight ω that reflects population level control-to-case benefit trade-offs may be written in terms of the threshold as $\omega = \frac{r_T}{1-r_T}\frac{1-\rho}{\rho}$.

Net benefit, for general clinical decision rules as well as for risk rules, is a population-level summary measure of the potential utility of using a rule in practice. However, a rational risk rule has the additional property of being the rational decision, in terms of the expected utility criterion, on an individual level. This optimality is not enjoyed by general decision rules nor risk rules that are not rational.

Partial Information from Universal Rules

Baker et al. (2009) reason that when the current standard of care is to treat no-one or to treat all, then the case-to-control benefit ratio should be, assuming current practice is rational, greater than or less than the odds of the outcome probability $\frac{\rho}{1-\rho}$, respectively. This follows from thinking in terms of risk. In the absence of a risk model, the presumed risk of each member in the population would be ρ, the overall probability of outcome. If the universal decision is to treat everyone (A) is rational, then everyone in the population should be considered high-risk and consequently the rational threshold must be less than ρ. Conversely, if treating the universal decision to treat no-one (N) is rational, then everyone in the population should be considered low-risk and the rational threshold must be greater than ρ. This also motivates evaluating rules using the opt-in formulation of net benefit, sNB^N in the latter case and the opt-out formulation, sNB^A in the former scenarios.

Eliciting Benefit Tradeoffs

In the absense of an extensive cost-benefit analysis, various other approaches for eliciting acceptable risk thresholds, such as ambivalence, test trade-off, or in terms of regret, have been proposed (Tsalatsanis et al., 2010; Baker et al., 2012). Elicitation of informal risk thresholds, used in decisions by many practicioners, could provide insight on plausible benefit trade-off ratios. In cases when there is no consensus on an appropriate trade-offs, evaluation at a few reasonable control-to-case benefit ratios can be useful (Pepe and Janes, 2013).

1.5 Analytic Formulations of Variance

In a clinical context in which there is currently no intervention, knowing that the net benefit of a proposed rule is positive would support its adoption in practice. Similarly, for two competing rules, a straight-forward comparison of their net benefit would provide information about which rule has more clinical value. However, evaluations of the utility of a clinical decision rule are based on estimates of the unknown net benefit. Estimation introduces uncertainty that should be taken into account when deciding whether to employ a rule in practice, study a candidate further, or abandon it entirely.

Bootstrapping is the current approach to evaluating uncertainty in estimates of net benefit (Vickers et al., 2008). While the increasing number of clinical studies including decision analytic approaches in analyses reflects an important philosophical shift in assessing the clinical value of biomarkers, communicating the limits of current evidence is an important step that is often omitted. This reveals that while the importance of clinically relevant measures is appreciated, decision analytic approaches are secondary and studies are not being designed to properly support their evaluation.

Analytic expressions of variance surpass descriptions of uncertainty captured through bootstrapping. They provide insights on sources of variability and support the design of studies for the express purpose of evaluating the utility of a clinical decision rule before it is adopted in practice. To the author's knowledge, only expressions of limited applicability

exist in the literature. In particular, none directly apply to estimates of net benefit calculated from a simple sample of the population of interest. Distribution theory and analytic formulations of the limiting variability of empirical estimators of net benefit constitute the primary contributions of this part of the dissertation.

Chapter 2

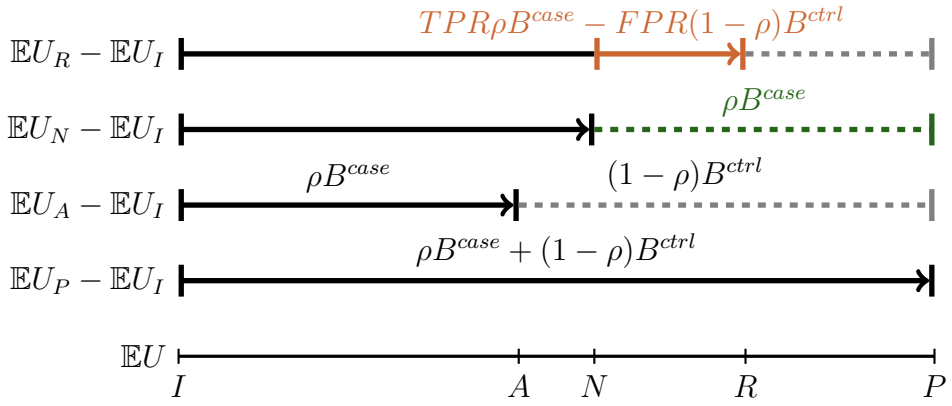
VISUAL REPRESENTATIONS OF STANDARDIZED NET BENEFIT

In this chapter we introduce two visual representations of standardized net benefit. The first relies on ordering rules in terms of expected utility and is intended to illuminate the interpretation of sNB as well as the rationality of using opt-in and opt-out formulations in different clinical contexts. The second approach views sNB as a function of the classification accuracies of rules. This representation is oriented towards evaluating rules within a presumed clinical context and will be employed and extended in later chapters.

2.1 Expected Utility Orders Clinical Decision Rules

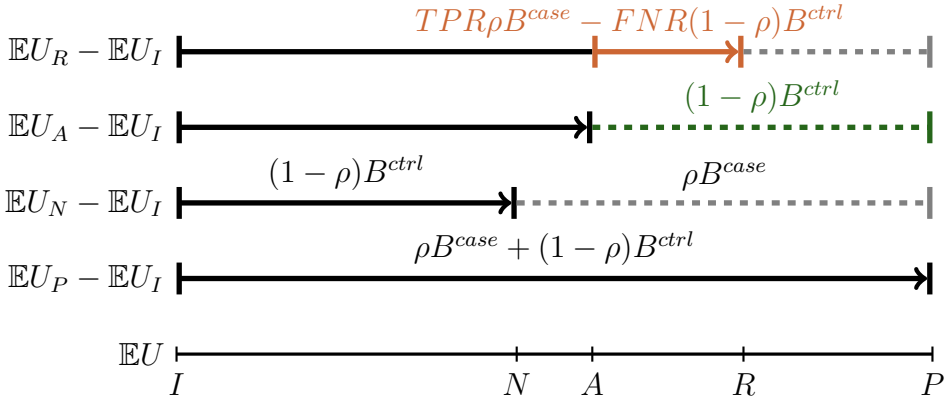
Expected utility provides an ordering of clinical decision rules which ranges from that of a completely incorrect rule, $\mathbb{E}U_I$ to that of a perfect rule, $\mathbb{E}U_P$. Recall that the expected utility of a rule is a weighted sum of the utilities and that the weights can be expressed in terms of the true- and false-positive rates of the rule and the outcome probability. This also applies to the difference in expected utility between two rules, which provides one measure of how far apart, on the expected utility scale, two rules are. For example, $\mathbb{E}U_R - \mathbb{E}U_I$ represents how much more useful, how much further to the right, R is than the totally incorrect rule I. Similarly, $\mathbb{E}U_P - \mathbb{E}U_R$ represents how much less useful, how much further to the left, R is than the perfect rule P. The greatest difference in expected utility is $\mathbb{E}U_P - \mathbb{E}U_I$ which equals $\rho \cdot B^{case} + (1 - \rho) \cdot B^{ctrl}$.

Figure 2.1 provides pictorial interpretations of the opt-in (a) and opt-out (b) formulations of standardized net benefit. In both scenarios, ordering rules by expected utility is illustrated in the bottom line segment; the expected utility of treat-all (A) and that of treat-none



$$sNB_R^N = \frac{\mathbb{E}U_R - \mathbb{E}U_N}{\mathbb{E}U_P - \mathbb{E}U_N} = \frac{TPR\rho B^{case} - FPR(1-\rho)B^{ctrl}}{\rho B^{case}} = TPR - \omega FPR$$

(a) $\omega = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho}{\rho} > 1$



$$sNB_R^A = \frac{\mathbb{E}U_R - \mathbb{E}U_A}{\mathbb{E}U_P - \mathbb{E}U_A} = \frac{TNR(1-\rho)B^{ctrl} - FNR\rho B^{case}}{(1-\rho)B^{ctrl}} = TNR - \frac{1}{\omega}FNR$$

(b) $\omega = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho}{\rho} < 1$

Figure 2.1: Ordering rules by expected utility and pictorial representations of the opt-in (a) and opt-out (b) formulations of standardized net benefit.

(N) lie symmetrically around the midpoint of the range. A general rule R, of even greater expected utility and consequently of positive standardized net benefit, is also included in both diagrams. Above the ordering appear additional segments in which the distance between the completely incorrect rule I and each of the rules P, A, N, and R, on the expected utility scale, is represented by a vector. For all rules other than P, the distance to P is demarcated by a dashed line. The length of various segments, in terms of differences in expected utility, are labeled. The numerator and denominator of net benefit are colored orange and green, respectively.

When the weight ω is greater than one (Figure 2.1a), $\mathbb{E}U_A < \mathbb{E}U_N$, the population level benefit of intervening correctly on controls, $(1-\rho)\cdot \mathrm{B}^{ctrl}$, outweighs that of intervening correctly on cases, $\rho \cdot \mathrm{B}^{case}$. Thus, treating no-one would be universal decision rule with greater expected utility, and is the rational choice in the absence of a superior clinical decision rule. When $\omega > 1$, the benefit trade-offs satisfy $\frac{B^{ctrl}}{B^{case}} > \frac{\rho}{1-\rho}$, which was the previously noted implication of assuming a treat-none decision was rational. In this context, we consider using sNB^N to evaluate a more stratified rule R in terms of its improvement over N. Standardized net benefit, also known as relative utility, can be described as the ratio of two differences in expected utility. In the case of sNB^N_R, this ratio has numerator $\mathbb{E}U_R - \mathbb{E}U_N$ and denominator $\mathbb{E}U_P - \mathbb{E}U_N$. The interpretation of increased expected utility as a fraction of the increase achieved by a perfect rule, both relative to treat-none, can be visualised as the orange segment as a portion of the green segment. When ω is less than one (Figure 2.1b), analogous reasoning leads to using treat-all as a reference which defines sNB^A_R. When ω equals one (not shown), $\mathbb{E}U_A = \mathbb{E}U_N$, $sNB^N_R = sNB^A_R$, and the two formulations are equivalent.

Implications for Selecting Formulations of sNB

When the rule R is a candidate replacement for the current practice of treating no-one, interpreting sNB^N_R as a fraction of the possible increase in expected utility over N, achieved by rule R, is meaningful. When the rule is a candidate for replacing the current practice of treating everyone, the interpretation of sNB^A_R is analogously meaningful. In this latter

case, when the rule is a candidate for replacing the current practice of treating everyone, the interpretation of the opt-in formulation is not as meaningful as the opt-out formulation. Further, we note that

$$\text{sNB}_R^N = \frac{\mathbb{E}U_R - \mathbb{E}U_A + \delta}{\mathbb{E}U_P - \mathbb{E}U_A + \delta},$$

where $\delta = \mathbb{E}U_A - \mathbb{E}U_N$ and has been added to both the numerator and denominator of sNB_R^A. When $\omega < 1$, and treating everyone is a better universal rule than treating no-one, then δ is positive and we conclude that $\text{sNB}_R^N > \text{sNB}_R^A$. That is, the less relevant formulation can give the impression of more benefit. This same behavior applies when using sNB^A to evaluate a rule when N is a better universal decision than A.

Versions of net benefit that implicitly contrast against any decision rule S can be defined analogous to the opt-in and opt-out formulations. In general, we may define:

$$\text{sNB}_R^S = \frac{\mathbb{E}U_R - \mathbb{E}U_S}{\mathbb{E}U_P - \mathbb{E}U_S}.$$

As expected, $\text{sNB}_S^S = 0$. This quantity represents the increase in expected utility of rule R as a fraction of the possible increase achieved by a perfect rule, both over the expected utility of rule S. We point out that this formulation can be achieved as a change of variables from the opt-in formulation according to

$$\text{sNB}_R^S = \frac{\text{sNB}_R^N - \text{sNB}_S^N}{\text{sNB}_P^N - \text{sNB}_S^N}.$$

This observation establishes that sNB^S retains the desirable property of relying on the underlying utilities only through the control-to-case benefit ratio.

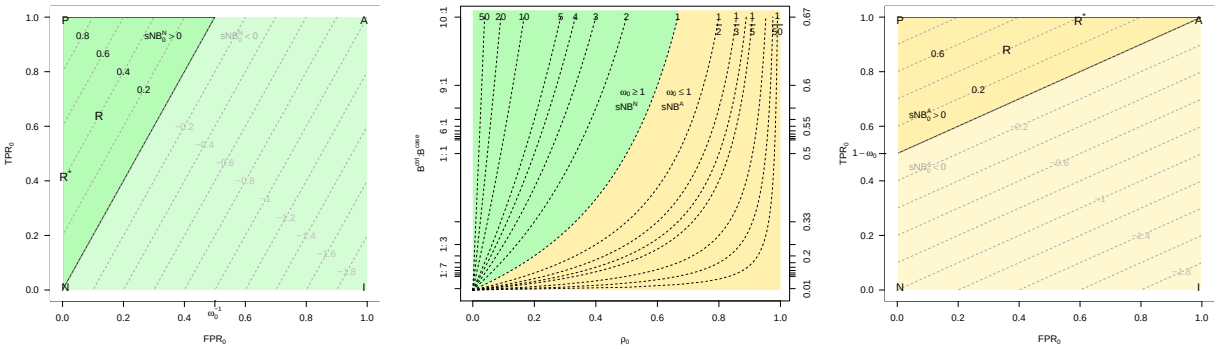
Such a formulation could be of interest when clinical decision rule S represents the current standard of care. For example, if physicians adhered better to using the Mayo risk model when determining whether or not a patient with an incidentally detected pulmonary nodule should undergo surgical diagnosis, then this rule may reasonably represent current practice S. An extended version of the Mayo risk model, reflecting the addition of a novel biomarker, would define an new rule R that could be evaluated in terms of its improvement over current practice using the above measure. However, if the expected utility of S is high,

the denominator of sNB^S could be quite small, which could make this a highly sensitive and difficult to estimate quantity.

2.2 As a Function of Classification Rates, Within a Clinical Context

Before focusing on the variability of estimators of net benefit, we first examine the behavior of the estimands themselves. A clinical decision rule is developed for facilitating the decision of whether or not to recommend a particular intervention pertaining to a specific clinical outcome within a well-defined population. The outcome probability, ρ, and trade-offs between making a correct choice for a control versus a case, B^{ctrl}:B^{case}, determine the features of the clinical context relevant to evaluating the net benefit. We use $C = \left(\rho_0, B^{ctrl}{:}B^{case}\right)$ to refer to these features and we sometimes employ a subscript 0 to emphasize that a quantity is an unknown truth. The true- and false-positive rates are the characteristics of a particular clinical decision rule. Development of risk rules within a clinical context naturally leads to consideration of multiple candidate rules with differing classification characteristics. We are thus led to consider sNB, for a single clinical context, as a function of the possible classification accuracies of a rule. We note that a clinical context determines both the risk threshold for a rational rule and the weight ω that reflects the population level trade-offs between making correct decisions for all controls versus for all cases. Further, as a function of true- and false-positive rates, the behavior of sNB is completely determined by the value of ω_0. If there is debate about the relative benefit trade-offs or a fair amount of uncertainty about the outcome probability, a few plausible contexts could be considered.

Figure 2.2(a) illustrates characteristics of net benefit when the clinical context corresponds to $\omega_0 = 2$. The (FPR,TPR) points corresponding the two universal decisions to treat all (A) and treat none (N) as well as so-called oracle rules that are perfect (P) and totally incorrect (I) are identified on the plot. Sets of classification characteristics that yield the same net benefit for a given clinical context, $\text{sNB}^N = c$, are lines with slope ω_0 and TPR-intercept equal to the common value of net benefit c. For example, the origin represents rule N and $\text{sNB}^N_N = 0$ by construction; all other points on the line with slope ω_0 represent rules yield-



(a) (b) (c)

Figure 2.2: Level sets of sNB^N (a) and sNB^A (c), over possible accuracy characteristics of a rule, for clinical contexts in which $\omega_0 = 2$ (a) or 0.5 (c). In either context, R and R^* represent rules achieving $\text{sNB} = 0.4$. Regions of positive net benefit are outlined and regions of negative net benefit are faded. (b) The relationship between trade-offs and the event rate in determining the weight. Level sets of ω_0 are plotted with dashed lines. Regions corresponding to the rational use of the opt-in and opt-out formulation of net benefit are differentiated by color.

ing increases in TPR that are exactly offset by the weighted penalty on the corresponding increase in FPR. The perfect rule achieves the maximum net benefit of 1, $\text{sNB}_P^N = 1$. The upper left triangle, including the perfect rule and bounded by the $\text{sNB}^N = 0$ level set, has been outlined in black, and the rest of the FPR-TPR square faded to indicate classification performances of rules that do not confer benefit in the given clinical context. The unfaded triangle comprises the classification accuracies of a rule that yield a positive benefit; other regions of minimal sNB^N could be defined similarly. We note that the region of positive net benefit gets smaller as ω_0 increases from 1, which reflects the increased constraints on acceptable false-positive rates as the population-wide consequences of incorrectly intervening on a would-be control become more severe relative to correctly intervening on a would-be case. The largest FPR that corresponds to a test with non-negative benefit is $\frac{1}{\omega_0}$. The totally

incorrect rule achieves the worst standardized net benefit equal to $-\omega_0$.

We plot the performance of a rule R at (FPR=0.1, TPR=0.6) which has standardized net benefit equal to 0.40 in the clinical context of Figure 2.2(a). The rule R*, corresponding to the TPR-intercept of the level set containing R, is the rule that would achieve equivalent sNB^N by correctly recommending intervention to 40% of the cases while not incorrectly recommending it to any controls. In the spirit of relative utility, the length of the TPR axis between N and R* is a geometric representation of sNB$_{\mathrm{R}}^N$ as the increase in expected utility (achieved by R over N) as a fraction of the possible increase in expected utility (that achieved by the perfect rule and represented by the unit-length segment of the TPR-axis between rules N and P).

Different combinations of trade-offs B^{ctrl}:B^{case} and event rate ρ_0 can determine the same weight ω_0. Figure 2.2(b) illustrates this relationship through plotting select level sets of the population level case-to-control benefit ratio ω. The value of the weight ω_0 determines whether, in the absence of a better performing risk rule, the decision to treat no-one ($\omega_0 > 1$) or to treat everyone ($\omega_0 < 0$) is rational. For clinical contexts in each of these regions, the opt-in or the opt-out formulation of net benefit, respectively, would be rationally employed. Notice on Figure 2.2(a), where the weight is greater than 1, the net benefit of the universal decision to treat no-one (sNB$_N^N = 0$) is greater than that of the universal decision to treat everyone (sNB$_A^N = 1 - \omega_0 = -1$). Table 2.1 highlights various combinations of trade-offs and event rates yielding select discounting weights ω_0 and additionally states the rational risk threshold corresponding to the trade-off.

Figure 2.2(c) repeats the illustration of Figure 2.2(a) when the clinical context corresponds to $\omega_0 = 1/2$ for the opt-out formulation of net benefit. Sets of classification characteristics that yield the same net benefit for a given clinical context, sNB$^A = c$, are lines with slope ω_0 that intersect the TPR=1 line at FPR=1-c. For example, the point (FPR=1, TPR=1) represents rule A and sNB$_A^A = 0$ by construction; all other points on the line with slope ω_0 represent rules yielding decreases in FPR that are exactly offset by the weighted penalty on the corresponding decrease in TPR. The perfect rule achieves the maximum net

	$\omega_0 = 2.8$		$\omega_0 = 5$		$\omega_0 = 16.2$	
ρ_0	r_T	$B^{ctrl} : B^{case}$	r_T	$B^{ctrl} : B^{case}$	r_T	$B^{ctrl} : B^{case}$
0.03	0.08	1:11.5	0.13	1:6.5	0.33	1:2
0.05	0.13	1:6.8	0.21	1:3.8	0.46	1:1.2
0.1	0.24	1:3.2	0.36	1:1.8	0.64	1.8:1
0.15	0.33	1:2	0.47	1:1.1	0.74	2.86:1
0.2	0.41	1:1.4	0.56	1.25:1	0.8	4.05:1

Table 2.1: Benefit trade-offs (with corresponding rational risk threshold) and event rates corresponding to the same weight, $\omega_0 = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_0}{\rho_0}$.

benefit of 1, $\text{sNB}_P^A = 1$. The upper left triangle, including the perfect rule and bounded by the $\text{sNB}^A = 0$ level set, has been outlined in black, and the rest of the FPR-TPR square faded to indicate classification performances of rules that are not of benefit in the given clinical context. The unfaded triangle comprises the classification accuracies of a rule that yield a positive benefit; other regions of minimal sNB^A could be defined similarly. We note that the region of positive net benefit gets smaller as ω_0 decreases from 1, which reflects the increased constraints on acceptable true-positive rates as the population-wide consequences of incorrectly not intervening on a would-be case become more severe relative to correctly intervening on a would-be control. The largest TPR that corresponds to a test with non-negative benefit is $1 - \omega_0$. The totally incorrect rule achieves the worst standardized net benefit equal to $-\frac{1}{\omega_0}$.

We plot the performance of a rule R at (FPR=0.34, TPR=0.86) with standardized net benefit equal to 40% in the clinical context of Figure 2.2(c). The rule R^*, corresponding to the intersection of the level set containing R with the TPR=1 line at (FPR=0.6,TPR=1), is the rule that would achieve equivalent sNB^A by correctly recommending no intervention to 40% of the controls while not incorrectly recommending any cases to opt-out. In the spirit of relative utility, the length of the TPR=1 line segment between A and R^* is a geometric

representation of $\text{sNB}_{\text{R}}^{A}$ as the increase in expected utility (achieved by R over A) as a fraction of the possible increase in expected utility (achieved by the perfect rule and represented by the unit-length segment of the TPR=1 line between rules A and P).

The opt-out formulation of sNB is naturally described in terms of true and false-negative rates: $\text{sNB}^{A} = TNR - \frac{1}{\omega}FNR$. Relabeling the axes of Figure 2.2(a) with TNR replacing TPR, FNR replacing FPR and noting that $\frac{1}{\omega} = 2$ when $\omega = 1/2$, reveals that Figure 2.2(a) could also be used to visualize the sNB^{A} presented in Figure 2.2(c).

Chapter 3

VARIABILITY OF EMPIRICAL ESTIMATORS FROM A COHORT

In this chapter, we are concerned with analytic inference for empirical estimators of net benefit defined on a cohort sample. We begin by introducing the importance of cohort samples in the validation setting. After establishing large-sample distribution theory, we apply our results to describe the uncertainty in estimates of net benefit from results of two published studies. The analytic expression of the asymptotic variance is explored for insights in scenarios that may require larger or smaller samples in order to assess net benefit with precision. Example sample size calculations are made for a cohort study designed to assess net benefit. The distribution theory is then extended to the difference in net benefit between two clinical decision rules. Finally, the results of this chapter are illustrated and evaluated using simulated data from the cardiovascular literature. In keeping with the literature, we will develop methodological results for opt-in scenarios and simply state results for opt-out scenarios and unstandardized formulations of net benefit. When there is no risk of ambiguity, we may write sNB for sNB^N and simply refer to net benefit without emphasizing whether the scale is standardized or not.

3.1 Validation Setting

The importance of validating the clinical utility of a proposed risk rule, independent from model development and using prospectively gathered data, is generally agreed upon (Steyerberg et al., 2010, 2012). This is consistent with the objectives of late-stage biomarker development to establish definitive evidence for the utility of proposed rules as they will be used in practice in the intended population. Validation of a rule could be the last stage of

evaluation of a clinical decision rule, prior to a dedicated clinical trial or wide-scale adoption in practice, and hence should be based on a large sample in conjunction with assessment of other utility measures. Validation of rules derived on data from earlier stage studies, perhaps on smaller cohorts or in slightly different populations, can also be a useful step in the accrual of evidence prior to conducting a full-scale final validation.

For example, the Amsterdam Pediatric Risk Rule (APRR), discussed further in Section 3.4, was developed on the sub-cohort of study participants enrolled at one of the participating university hospitals and then evaluated on the sub-cohort of study participants enrolled at one of the participating non-university hospitals. In this example, the subjects enrolled at the university hospitals could be different than those seeking care from non-university hospitals. Any confirmatory results on the performance of the rule evaluated on this related but potentially different sub-cohort would not only support that the rule was not an artifact of the particular sample on which is was created, but would also establish a certain amount of robustness to the utility of the rule. A similar team of investigators additionally used the validation cohort of the APRR to evaluate three other clinical decision rules, independently established and published by three other study teams in distinct geographic areas and time periods (Mulders et al., 2017). Any confirmatory results on the performance of these rules in the Amsterdam validation cohort would also validate and establish robustness of the rules. This second validation study has the additional strength of having the data collection and study leadership of the validation work being independent from that of model development.

Assessing performance characteristics of a rule at the time of model development, and from the same single dataset, is not considered validation and is outside of the scope of this dissertation. Estimating the net benefit of a pre-specified clinical decision rule, at one or more benefit trade-off ratios, from a prospectively collected sample representative of the population of interest, is the primary focus of this Chapter.

3.2 Notation

We continue to use the notation introduced in Chapter 1, wherein W is a vector of observable patient-specific variables and the binary clinical outcome D may represent either a current condition or future occurrence (e.g., diagnostic or preventative settings). In practice, the outcome is unknown at the time an intervention decision X will be made. We assume that D is available on all subjects in the study sample. A clinical decision rule R and associated action $x = R(w)$ may be derived from a risk model r and threshold r_T according to $x := I\left[r(w) > r_T\right]$. The classification accuracies of R are referred to as true and false-positive rates and denoted TPR_R and FPR_R, with subscript R optionally used to emphasize the dependence on the rule. The overall probability of the clinical outcome in the population is denoted $\rho := Pr(D = 1)$.

Each member of the cohort contributes an observation to the data set: $O = (X, D)$ which we assume was sampled identically and independently from some unknown population distribution $O_1, \ldots, O_n \sim_{iid} P_0$. The variable X inherits its randomness from W and at times it will be convenient to express the observation as (W, D), $(R(W), D)$, or in the case of a decision rule derived from a risk model, $(r(W), D)$ or $(I[r(W) > r_T], D)$.

Throughout, we will use subscripts n to denote estimators and estimates from a sample of size n and 0 to denote an unknown truth. For example, sNB_n^N represents an estimator of the estimand sNB_0^N constructed from an *i.i.d.* sample of observations $O_i \sim P_0$, where P_0, a probability distribution, is the unknown data-generating mechanism. Random variables will be expressed using upper case letters and a realization or deterministic argument denoted using the corresponding lower case letter. For example, the random variable Y could take the value y.

3.3 Inference for the Empirical Estimator of sNB^N

Empirical Estimation of Net Benefit from a Cohort

Standardized net benefit is a composite measure that relies on three constituents: the overall probability of the clinical outcome and the two classification accuracies of the decision rule: $\text{sNB}_0^N = \text{TPR}_0 - \omega_0 \text{FPR}_0$, where $\omega_0 = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_0}{\rho_0}$.

From a cohort sample, each of the three consitituents has an empirical estimator:

$$\rho_n = \frac{1}{n} \sum_i D_i \tag{3.1}$$

$$\text{TPR}_n = \frac{1}{n\rho_n} \sum_i X_i D_i \tag{3.2}$$

$$\text{FPR}_n = \frac{1}{n(1-\rho_n)} \sum_i X_i (1 - D_i) \tag{3.3}$$

The empirical estimator of net benefit follows naturally by substituting the estimators of each consitutent for the corresponding unknown true values in the expression of sNB:

$$\text{sNB}_n^N = \text{TPR}_n - \omega_n \, \text{FPR}_n, \tag{3.4}$$

where $\omega_n = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_n}{\rho_n}$.

Influence Functions

We will establish the distribution theory for sNB_n^N using influence function techniques. The empirical estimators of the three constituents of sNB_0^N are each simple conditional or marginal averages of binomial variables. They are asymptotically linear estimators with simple influence functions:

$$\begin{aligned} IF_{\rho,0}(o) &= d - \rho_0, \\ IF_{\text{TPR},0}(o) &= \frac{d}{\rho_0} \left\{ x - \text{TPR}_0 \right\}, \quad \text{and} \\ IF_{\text{FPR},0}(o) &= \frac{1-d}{1-\rho_0} \left\{ x - \text{FPR}_0 \right\}. \end{aligned}$$

The composite estimator sNB_n has influence function:

$$\begin{aligned} IF_{\text{sNB}^N,0}(o) &= \frac{d}{\rho_0}\left\{x - \text{TPR}_0\right\} - \omega_0 \frac{1-d}{1-\rho_0}\left\{x - \text{FPR}_0\right\} \\ &\quad + \frac{\omega_0}{\rho_0(1-\rho_0^2)}\text{FPR}_0\left(d - \rho_0\right), \end{aligned} \tag{3.5}$$

which is a combination of the influence functions for ρ_n, TPR_n, and FPR_n and established by application of the delta method for influence functions. An alternate derivation is available in Appendix B.2.

Inference for the Empirical Estimator of sNB

The empirical estimator, sNB_n^N, and its three constituent estimators, are all regular and asymptotically linear estimators. Thus, the asymptotic behavior of sNB_n^N is governed by the Central Limit Theorem:

$$\sqrt{n}\left(\text{sNB}_n^N - \text{sNB}_0^N\right) \xrightarrow{d} N\left(0, \sigma^2_{\text{sNB}_0^N}\right).$$

The limiting variance, $\sigma^2_{\text{sNB}_0} = \mathbb{E}_0\left[IF^2_{\text{sNB},0}(O)\right]$, is the variance of a single observation transformed by the influence function. For n sufficiently large, the variance of the estimator sNB_n will be approximately 1/n times the limiting variance. For this estimator, the limiting variance simplifies to:

$$\sigma^2_{\text{sNB}^N,0} = \frac{1}{\rho_0}\text{TPR}_0\left(1 - \text{TPR}_0\right) + \omega_0^2\frac{1}{1-\rho_0}\text{FPR}_0\left(1 - \text{FPR}_0\right) + \frac{1}{(1-\rho_0)\rho_0}\omega_0^2\text{FPR}_0^2, \tag{3.6}$$

which in part follows from the fact that, pairwise, the product of influence functions for ρ_n, TPR_n and FPR_n are all zero in expectation with respect to P_0, i.e., they are orthogonal functions in $L_0^2(P_0)$. An empirical estimator of the nonparametric variance bound, $\sigma^2_{\text{sNB},n}$, is defined by substituting the empirical estimators of the population parameters into the equation for $\sigma^2_{\text{sNB},0}$:

$$\sigma^2_{\text{sNB}^N,n} = \frac{1}{\rho_n}\text{TPR}_n\left(1 - \text{TPR}_n\right) + \omega_n^2\frac{1}{1-\rho_n}\text{FPR}_n\left(1 - \text{FPR}_n\right) + \frac{1}{(1-\rho_n)\rho_n}\omega_n^2\text{FPR}_n^2. \tag{3.7}$$

Asymptotically correct, approximate, two-sided Wald confidence intervals of level α,

$$\text{sNB}_n^N \mp \sqrt{\frac{\sigma^2_{\text{sNB}^N,n}}{n}} \cdot z_{1-\frac{\alpha}{2}}, \tag{3.8}$$

where $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\alpha/2)$, can be constructed as usual.

Inference for Estimators of Other Formulations of Net Benefit

Other formulations of net benefit, such as sNB^A and NB^N, are similarly composites of the outcome probability and the classification accuracies of a clinical decision rule. The corresponding empirical estimators can also be substituted into the appropriate formulation to define the empirical estimators of each measure net benefit. All estimators are regular and asymptotically linear. The influence functions and limiting variance for each estimator appear in Table 3.1 with calculations in Appendix B.1. Using the estimator-specific limiting variance, Wald confidence intervals can be constructed according to Equation 3.8.

3.4 Examining Variability of Published Results

In this section we present two published studies that exemplify evaluating risk rules for clinical utility. The studies evaluated pre-specified rules on a validation cohort that was independent from the data used for model development. The clinical decision rules being validated were both derived from risk rules using established risk thresholds and were motivated by reducing the number of unnecessary interventions; they are intended for clinical contexts in which the current practice was best described by the "treat all" universal decision. Both studies assessed the classification accuracies of the rule and the second study additionally considered net benefit on the unstandardized scale. Assessment of standardized net benefit, with confidence intervals, would be an improvement to these analyses. Our ability to add confidence intervals here relies on the respective sets of authors following the recommended practice of reporting not just an overall estimate of net benefit, but also each of the constituents (Janes and Pepe, 2013).

Decision	Net Benefit Variant	Influence Function and Limiting Variance
Opt-In	$\text{sNB}^N = \text{TPR} - \omega\text{FPR}$	$IF(o) = \frac{d}{\rho}\{x - \text{TPR}\} - \omega\frac{1-d}{1-\rho}\{x - \text{FPR}\} + \omega\frac{1}{\rho(1-\rho)}\text{FPR}(d-\rho)$ $\sigma^2 = \frac{1}{\rho}\text{TPR}(1-\text{TPR}) + \omega^2\frac{1}{1-\rho}\text{FPR}(1-\text{FPR}) + \omega^2\frac{1}{(1-\rho)\rho}\text{FPR}^2$
	$\text{NB}^N = \rho\,\text{sNB}^N$	$IF(o) = d\{x - \text{TPR}\} - \frac{B^{ctrl}}{B^{case}}(1-d)\{x - \text{FPR}\} + \left(\text{TPR} + \frac{B^{ctrl}}{B^{case}}\text{FPR}\right)(d-\rho)$ $\sigma^2 = \rho\text{TPR}(1-\text{TPR}) + \left(\frac{B^{ctrl}}{B^{case}}\right)^2(1-\rho)\text{FPR}(1-\text{FPR}) + \left(\text{TPR} + \frac{B^{ctrl}}{B^{case}}\text{FPR}\right)^2\rho(1-\rho)$
Opt-Out	$\text{sNB}^A = \text{TNR} - \frac{1}{\omega}\text{FNR}$	$IF(o) = \frac{1-d}{(1-\rho)}\{1 - x - \text{TNR}\} - \frac{1}{\omega}\frac{d}{\rho}\{1 - x - \text{FNR}\} - \frac{1}{\omega}\frac{1}{\rho(1-\rho)}\text{FNR}(d-\rho)$ $\sigma^2 = \frac{1}{1-\rho}\text{TNR}(1-\text{TNR}) + \frac{1}{\omega^2}\frac{1}{\rho}\text{FNR}(1-\text{FNR}) + \frac{1}{\omega^2}\frac{1}{\rho(1-\rho)}\text{FNR}^2$
	$\text{NB}^A = (1-\rho)\,\text{sNB}^A$	$IF(o) = (1-d)\{1 - x - \text{TNR}\} - \frac{B^{case}}{B^{ctrl}}d\{1 - x - \text{FNR}\} - \left(\text{TNR} + \frac{B^{case}}{B^{ctrl}}\text{FNR}\right)(d-\rho)$ $\sigma^2 = (1-\rho)\text{TNR}(1-\text{TNR}) + \left(\frac{B^{case}}{B^{ctrl}}\right)^2\rho\,\text{FNR}(1-\text{FNR}) + \left(\text{TNR} + \frac{B^{case}}{B^{ctrl}}\text{FNR}\right)^2\rho(1-\rho)$

Table 3.1: Asymptotic variance and influence functions for empirical estimators of various formulations of net benefit for opt-in and opt-out clinical decisions (implicitly contrasting against the decision to treat no-one or treat everyone, respectively) on standardized and unstandardized scales. The empirical estimators are conditional on a prespecified clinical decision rule. In all scenarios, the weight $\omega = \frac{B^{ctrl}}{B^{case}}\frac{1-\rho}{\rho}$ and rational decisions derived from a risk model r use threshold $r_T = \frac{B^{ctrl}}{B^{ctrl}+B^{case}}$.

Amsterstam Pediatric Risk Rule

Slaar et al. (2016) consider the utility of a risk rule for reducing the number of unnecessary radiography procedures on children visiting the emergency department (ED) for wrist trauma. In the absence of guidelines, the authors contend that radiography is routinely conducted even though one study estimated that only about half of radiographs performed to evaluate wrist trauma identified a fracture. This practice incurs financial costs, extends waiting time for the family in the ED, and exposes children to harmful radiation, many of whom do not have a fracture. A logistic regression model was developed on the subset of a prospectively gathered multi-site cohort ($N_{dev} = 408$) enrolled at a university hospital. Model predictors included variables such as age, visible deformation, tenderness and swelling of the distal radius. The model, and its derived decision rule, was validated in the distinct subset of the cohort ($N_{valid} = 379$) recruited at visitation to one of three other participating hospitals. The authors considered a 23% predicted risk of wrist fracture high enough to warrant taking a radiograph, citing general consensus in previous literature on thresholds in the range of 20-25%, and evaluated the resulting Amsterdam Pediatric Risk Rule, in terms of its classification properties. The 23% risk threshold defines a rational rule if the control-to-case benefit ratio is roughly 3:10.

Results published in Table 3 of Slaar et al. (2016) indicate that the observed rate of wrist fracture, among children receiving radiography for acute wrist trauma, $\rho_n = 170/379 = 0.449\,(95\%\,\text{CI: } 0.398, 0.499)$, and provide data for the observed true- and false high-risk classification rates of their proposed clinical decision rule: $\text{TPR}_n = 163/170 = 0.959\,(95\%\,\text{CI: } 0.929, 0.989)$ and $\text{FPR}_n = 131/209 = 0.627\,(95\%\,\text{CI: } 0.561, 0.692)$, respectively. We have presented the Wald 95% confidence intervals using the analytic expressions of variance established previously. The authors presented sensitivity and specificity with confidence intervals presumably calculated via bootstrapping, in fair agreement with those above.

Because the standard of care is to perform a radiograph on all children presenting to the ED for acute wrist injury, we assess standardized net benefit using the opt-out for-

mulation. The estimated net benefit of the Amsterdam Pediatric Risk Rule is $\text{sNB}^A = 0.261\,(95\%\,\text{CI: } 0.154, 0.368)$. This benefit results from continuing to perform radiography on 95.9% of kids with a fracture, while reducing unnecessary radiographs to 62.7% of kids with acute write trauma, but no fracture. This benefit can also be described as eliminating 37.3% of radiographs among children without fracture, while not conducting a radiograph on 4.1% of children with a wrist fracture. The proposed rule has the same benefit as a rule that continues to conduct radiographs on all children with wrist fractures, but sparing 26.1% of children without a fracture from being sent for radiograph. In the validation cohort, the Amsterdam Pediatric Risk Rule potentially achieved 26.1% of the benefit that would be gained by using a perfect rule. Based on the analytic asymptotics of the estimator sNB_n^N, the true net benefit could fall between 0.154 and 0.368, with 95% confidence.

Lynch Syndrome in Diagnosed Colorectal Cancer Patients

Lynch Syndrome (LS), a consequence of well-defined germ line genetic mutations and also referred to as hereditary non-polyposis colorectal cancer (HNPCC), is associated with increased risk of cancer, and colorectal cancers (CRCs) in particular. Conversely, CRC patients are a group for whom the implications of having a hereditary variant could be significant and lead to pursuit of more aggressive treatment, extensive surveillance and comprehensive cancer risk management than those without the syndrome. The commercialization of focused panel testing for germline mutations has made genetic evaluation a viable course of action. In guidelines put forth by the US Multi-Society task Force on Colorectal Cancer, a predicted risk of MMR mutation exceeding 5% was sufficient to warrant recommendation of genetic evaluation for Lynch Syndrome among newly diagnosed CRC patients (Giardiello et al., 2014). The clinical usefulness of three established models for predicting individual risk of Lynch Syndrome, all highlighted in the guidelines, were evaluated by Kastrinos et al. (2016) on 11 international cohorts.

The 11 international cohorts comprising the validation set were comprised of 6 clinic-based and 5 population-based cohorts, as determined by the means of recruiting participants.

As might be anticipated, the percentage of patients testing positive for LS was considerably greater in the clinic-based cohorts (23.4%) than in the population-bases cohorts (4.4%). The recommended 5% risk threshold is rational for a control-to-case benefit ratio of 1:19. In the cohort-based populations, this risk is higher than the observed prevalence of LS and the decision to conduct genetic testing on none of these CRC patients is the universal decision of most clinical utility. The opposite holds for the clinic-based cohorts. Other characteristics, such as age, varied between the two cohorts and the risk rules were validated on each cohort type separately. Here we will focus on results for validation of the PREdiction of Mismatch repair gene Mutations (PREMM) model, from a pooled analysis of the clinic-based cohorts.

The previously developed PREMM model ($N_{dev} = 4,539$) was validated in the pooled clinic-based cohorts for a combined $N_{valid} = 2,294$ observations. Results in Table 5 of the article lead us to conclude: $\rho_n = 536/2294 = 0.234\,(95\%\,\text{CI: } 0.216, 0.251)$, and for $r_T = 0.05$, the estimated true-positive rate $\text{TPR}_n = 516/536 = 0.963\,(95\%\,\text{CI: } 0.947, 0.979)$ and false-positive rate $\text{FPR}_n = 1238/1758 = 0.704\,(95\%\,\text{CI: } 0.683, 0.726)$. The authors evaluated PREMM and the decision to treat-all using the opt-in formulation of net benefit and obtained estimates of $\text{NB}^N_{PREMM,n} = 0.197\,(95\%\,\text{CI: } 0.179, 0.214)$ and $\text{NB}^N_{A,n} = 0.193\,(95\%\ 0.175, 0.212)$, to which we have added confidence intervals. We note that no description of variability accompanied the results for any of the above quantities in the publication.

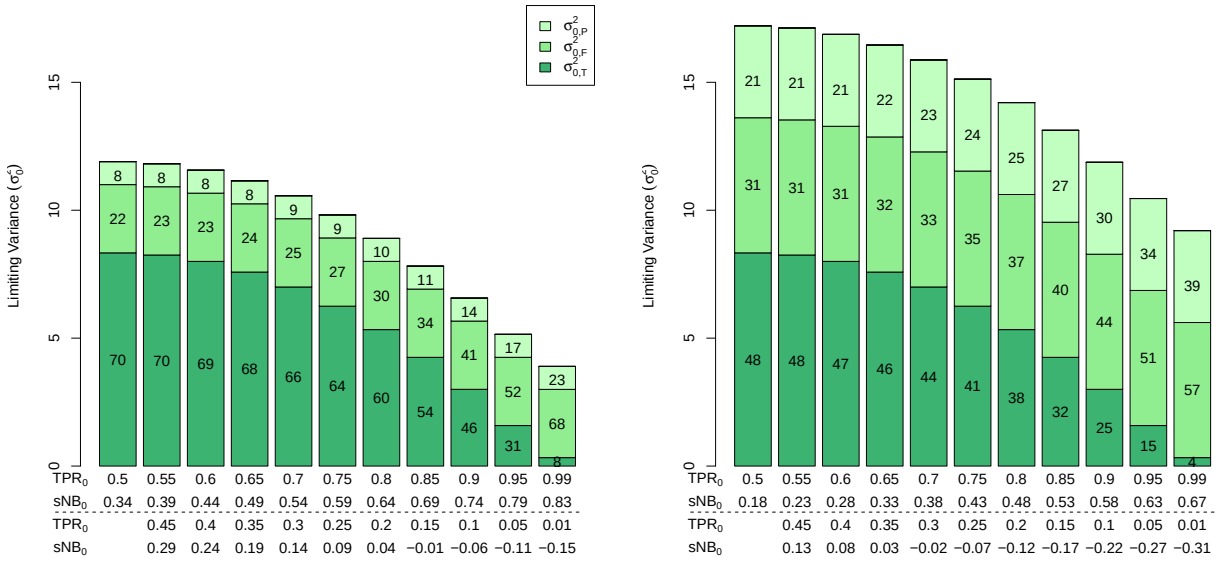
These results indicate that both the PREMM based rule and the universal decision to send all clinic-based CRC patients for genetic testing confer more clinical utility than the universal decision to not test any of them. In this sense, testing everyone is the rational universal decision, consistent with the risk threshold being less than the prevalence observed in this population. The difference, $\text{NB}^N_{PREMM,n} - \text{NB}^N_{A,n} = 0.0032$ provides one measure of the increased utility, in units of B^{case}, and suggests that the PREMM model could offer further clinical utility beyond simply testing everyone. However, an assessment of variability is needed; inference for the difference in net benefit between two rules will be addressed in Section 3.8. For this particular difference, the opt-out formulation of net benefit provides

a natural contrast in units of B^{ctrl}. Using the same constituents as above, we calculate $\text{NB}^A_{PREMM,n} = 0.061$ (95% CI: $-0.04, 0.136$). We note that the point estimate is 19 times $0.0032 = \text{NB}^N_{PREMM,n} - \text{NB}^N_{A,n}$, which accounts for the change of units from B^{case} to B^{ctrl}, $\frac{\text{B}^{ctrl}}{\text{B}^{case}} = 1{:}19$, and that $\text{NB}^A = 0.061$ is roughly 8% of the maximum possible net benefit ($\text{NB}^A_P = 1 - \rho = 0.766$).

On the standardized scale, the estimate of $\text{sNB}^A_{PREMM,0}$ is 0.08 (95% CI: $-0.018, 0.177$). The PREMM based rule achieves 8% of the improvement over genetically testing all CRC patients, that would be achieved by using a perfect rule. We can also say that this rule has the same net benefit, over testing everyone, as a rule that eliminates testing for 8% of the CRC patients that do not have LS while continuing to test all CRC patients with the syndrome.

Decision theoretic perspectives advocate adoption of a new rule if its estimated improvement were positive. However, taking into account variability, we cannot rule out a decrease in improvement resulting from using the PREMM model with a 5% threshold, over the practice of testing all clinic-based patients. The asymmetry of the 95% confidence interval for the estimated sNB^A suggests that a larger sample may have conferred greater confidence in the observed improvement. Statistical significance aside, the PREMM rule appears to achieve only a small portion of the possible increase in clinical utility. Sending all clinic-based CRC patients for genetic evaluation may be a simpler standard of care to adopt until a rule offering a more impactful improvement is available.

The primary difference between using NB^N and sNB^A to compare treating all and treating no-one are the interpretations. As we see, the same qualitative conclusions about the clinical utility of PREMM vs testing everyone are drawn from either approach. As discussed in Section 2.1, using sNB^N when sNB^A is rational and more relevant to current practice can be misleading quantitatively ($\text{sNB}^N_{PREMM} = 0.825$).



(a) (b)

Figure 3.1: Decomposition of the limiting variance of sNB_n^N, σ_0^2, for a clinical context defined by $\rho_0 = 0.03$, B^{ctrl}:B^{case} ($\omega_0 = 16.2$), when the rule has a 1% (a) or 2% (b) false-positive rate, across true-positive rates. The variance is symmetric in TPR_0 and hence each bar represents two scenarios and two values of net benefit. Components represent variability in estimation of ρ_0 (top), TPR_0 (middle), and FPR_0 (bottom). Components are labeled with the corresponding percentage of the total.

3.5 Structure of the Limiting Variance

Limiting Variability Decomposition

The limiting variance for empirical estimators from a cohort is expressed as a sum of three terms, each of which describes the variability in sNB_n due to uncertainty in estimating the true-positive rate, the false-positive rate, and the outcome probability, respectively. To highlight these dependencies, we can decompose the limiting variability $\sigma^{2,cohort}$ into the sum of $\sigma_T^2 = \frac{1}{\rho}\text{TPR}\,(1-\text{TPR})$, $\sigma_F^2 = \frac{1}{1-\rho}\,\omega^2\,\text{FPR}\,(1-\text{FPR})$ and $\sigma_P^2 = \omega^2\text{FPR}^2\frac{1}{\rho(1-\rho)}$, which we

distinguish from the limiting variance for empirical estimators of TPR, equal to $\sigma^{2,cohort}_{TPR}$, and so forth. The factors of ρ or $1-\rho$ in the respective denominators of σ^2_T and σ^2_F account for the portion of the cohort sample contributing to each estimator.

Figure 3.1 represents the total limiting variances as a bar broken into its three components. Given a particular clinical context, each plot is oriented as if moving across ROC curves of increasing discrimination among cases, TPR increases from left to right, for a fixed false-positive rate. Values of the corresponding net benefit of employing a rule with the indicated performance characteristics in the given clinical context are labeled below the true-positive rate on the axis. Since the variance is symmetric in TPR, but the measure of net benefit is not, there are two sets of values.

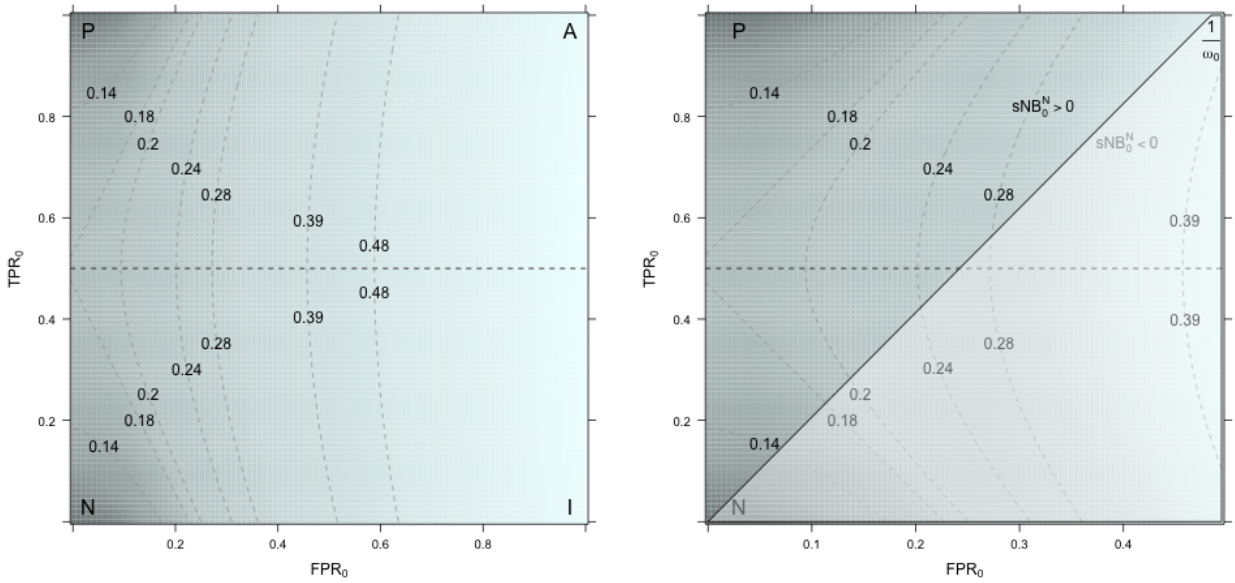
Standardized net benefit penalizes the mistakes introduced by using the rule, compared to treating no-one, more than it rewards an improvement. The weight ω represents this penalty and reflects both the relative frequency of and the relative benefits between controls and cases. Consequently, the variability of net benefit estimators is quite sensitive to the error rate (false-positive rate), increasingly so when the penalty ω_0 is large. Figure 3.1 pertains to a clinical context for which the event is rare, $\rho_0 = 0.03$, and an intervention with a 1:2 cost-benefit trade-off, conferring modestly more potential benefit to intervening on cases than to not intervening on controls, and rules with false-positive rates of 1% (a) and 2% (b). The noticeable difference in the total limiting variance is solely due to absolute increases in the contributions from estimation of FPR_0 and ρ_0 when FPR is 2% compared to 1%. The contribution due to estimation of the true-positive rate does not rely on the corresponding false-positive rate of the rule. The absolute contribution by σ^2_T to the total σ^2 is the same in the two figures, but the proportion of the total is less in (b) than in (a). In particular, the proportion due to estimation of the outcome probability is non-negligible (8-23%) when the false-positive rate is 1% and it almost doubles (21-40%) when FPR is 2% rather than 1%.

Changes to the clinical context correspond to fairly intuitive changes to the variability and its decomposition. For example, when considering a greater relative benefit for a case,

(1:11.5 control-to-case benefit ratio instead of 1:2) the penalty for mistakes is $\omega_0 = 2.8$ rather than 16.2 and the variability contributions from estimation of FPR_0 or ρ_0, both scaled by ω_0^2, are smaller than those in the figure. In this scenario, the absolute contribution due to estimation of the true positive-rate is the same as in Figure 3.1 and the relative contributions due to the outcome probability and false-positive rate are generally only a few percent of the total. If instead we consider the 1:2 cost-benefit trade-off in the context of a more common event, $\rho_0 = 15\%$ instead of 3%, the penalty for mistakes is also $\omega_0 = 2.8$, but a much larger portion of the sample contributes to the estimation of the true-positive rate and the reduction in $\sigma^2_{0,T}$ is significant. In this clinical context $C = \big(\rho_0 = 0.15, B^{ctrl}{:}B^{case} = 1{:}2\big)$, the relative picture is quite similar to that of Figure 3.1, albeit at higher false-positive rates, 5% and 10% rather than 1% and 2%. Overall, the magnitude of the limiting variance of sNB_n is quite sensitive to the rareness of the event, and all the more so with greater false-positive rates. Accompanying figures appear in Appendix B.3.

Level Sets Within a Clinical Context

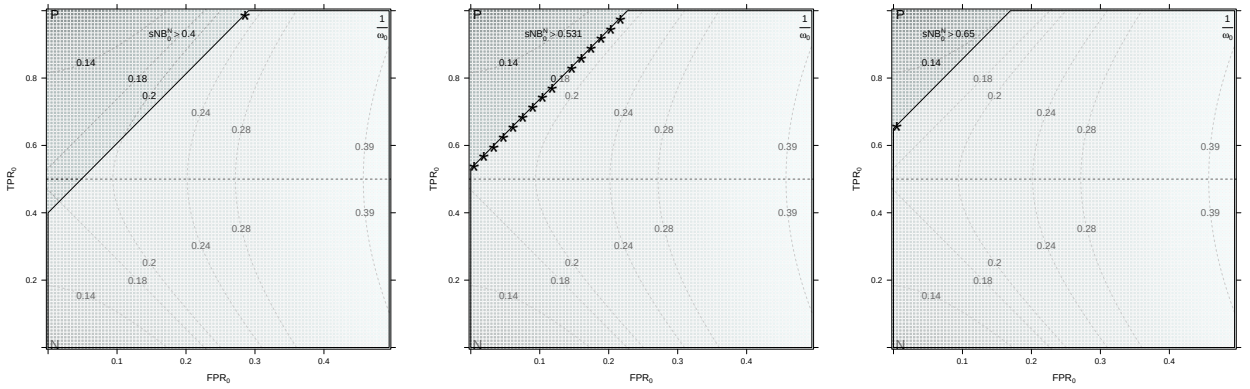
While it is tempting to examine the behavior of σ^2_{sNB} as only one parameter varies, excepting TPR, which only appears in one isolated term, this is quite difficult and not obviously meaningful. In practice, the relevant clinical context will determine both the outcome probability as well as the relative benefit trade-offs of the intervention. As we did in Section 2.2 for values of sNB^N, we now consider the sampling variability of sNB^N_n, in a fixed clinical context, as a function of the possible classification accuracies of a clinical decision rule. We demonstrate this in Figure 3.2(a) for a clinical context defined by $C = \big(\rho_0 = 0.03, B^{ctrl}{:}B^{case} = 1{:}17\big)$ which implies that $\omega_0 = 2$. Contours of variance are indicated with grey dashed lines; to provide a meaningful scale, we quantify variance by the approximate 95%-CI half-width $(1.96\sqrt{\sigma_0^2/N})$ for estimation from a cohort of size N=1,000. Over the entire set of possible true and false-possible rates, variability can be quite extreme. Fortunately, the full set of possible classification accuracies is not of practical interest. Figure 3.2(b) shows the same information as in (a), but focused on the region that yields positive net benefit.



(a) (b)

Figure 3.2: (a) The asymptotic variance σ_0^2 of empirical estimators of sNB^N, over possible classification characteristics of a rule, for a clinical contexts defined by $\rho_0 = 0.03$ and B^{ctrl}:B^{case} = 1:17. Grey dashed lines mark variance contours and have been labeled with the approximate 95%-CI half-width ($1.96\sqrt{\sigma_0^2/N}$) for estimation from a cohort of size 1,000. (b) Focused on classification accuracies yielding positive net benefit.

Algebraic manipulation of the limiting variance formula (3.6) leads to the observation that contours of equal limiting variance lie on hyperbolas centered at ($\text{FPR}_0 = -\frac{\rho_0}{2(1-\rho_0)}$, $\text{TPR}_0 = 0.5$) with asymptotes going through this point with slopes $\pm\omega_0$. Over the asymptote with positive slope, sNB^N is constant and equals $\frac{1}{2}+\frac{\rho_0}{2(1-\rho_0)}\omega_0$; the limiting variance is also constant and equals $\frac{1}{4\rho_0} - \frac{\omega_0^2\rho_0}{4(1-\rho_0)^2}$. Whenever these values are outside meaningful ranges (between 0 and 1 or positive, respectively), they simply don't correspond to points in the unit FPR-TPR square and are mathematically but not practically meaningful. Level sets for variances less than on the asymptotes occur in horizontal pairs (above and below TPR=50%) and for greater variances in vertical singletons (the matching pair is over negative FPR-values). Restricting to rules for which the true-positive rate is greater than 50%, we note that the



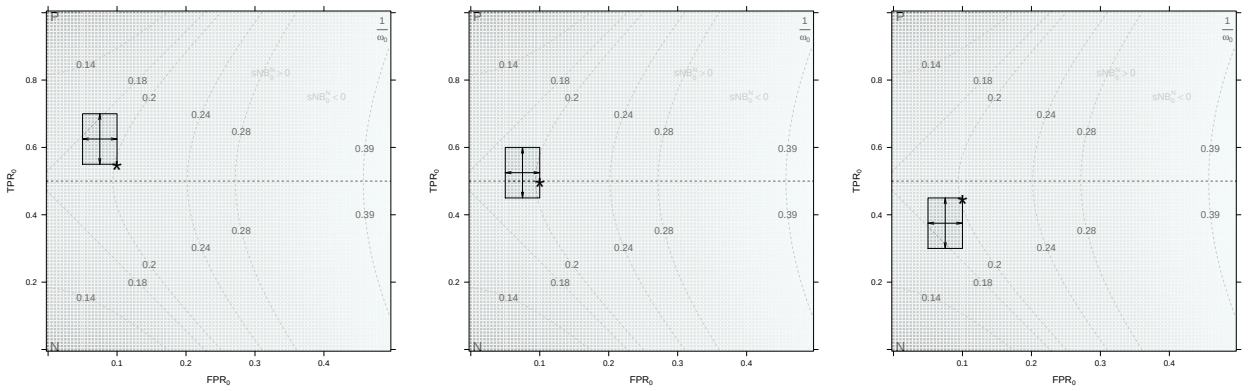
(a) (b) (c)

Figure 3.3: Over regions of minimal net benefit the classification accuracies of a rule for which sNB_n is most variable are denoted with an asterix '*'. Three qualitative scenarios are demonstrated, in which the minimal net benefit is less than (a), equal to (b), or greater than (c) $\text{sNB}^N = \frac{1}{2} + \frac{\rho_0}{2(1-\rho_0)}\omega_0 = 0.53$. All other details are the same as for Figure 3.2(b).

greater the net benefit of the rule, the easier it is to estimate; the variance generally decreases as rules become closer to perfect. Calculations are in Appendix B.4.

Variance Bound Over a Region of Minimal Net Benefit

It can be shown that, within a given clinical context, the limiting variance, over any triangular region corresponding to rules of minimal net benefit, $\left\{R : \text{sNB}_R^N \geq d\right\}$, for d less than, equal to, or greater than $\frac{1}{2} + \frac{\rho_0}{2(1-\rho_0)}\omega_0$, i.e., the level set is below, equal to, or above the asymptote with slope ω_0, is bounded by the variance of the rule with accuracies corresponding to the upper right corner, any point on the boundary, or to the lower left corner of the set, respectively. These three scenarios are illustrated in Figure 3.3, where an asterix '*' is used to denote the classification accuracies of a rule for which sNB_n will be most variable. Plugging the characteristics of these extremal points into the formula for the limiting variance yields a quick way to gauge how difficult it will be to evaluate clinical utility of a minimal performance level within a given clinical context.



(a) (b) (c)

Figure 3.4: Over rectangular regions, the classification accuracies of a rule for which sNB_n is most variable is denoted with an asterix '*'. Three qualitative scenarios, in which the rectangle is above (a), crosses (b), or below (c) the TPR=0.5 line, are demonstrated. All other details are the same as for Figure 3.2(b).

Over the region of positive net benefit, the variance bound simplifies to $\frac{1}{\rho_0}\left(1+\frac{B^{ctrl}}{B^{case}}\right)$ for rational risk rules. In this sense, standardized net benefit is more difficult to evaluate in situations of rare outcomes and all the more so the greater the relative benefit of correctly not intervening on a control. Both the region of clinical usefulness and the difficulty of evaluating that usefulness due to variability are driven by the weight $\omega_0 > 1$ which reflects the relative dominance of the population level benefit of getting all cases right over getting all controls right and consequently penalizes introduced mistakes (false-positives), relative to treating no-one, more than rewarding corrections (true-positives). Calculations are in Appendix B.4.

Variance Bound Over a Rectangular Region

Within a given clinical context, the limiting variance, over any rectangular region is bounded by the variance of the rule with true-positive rate closest to 50% and largest false-positive rate. This follows immediately from previous discussion regarding the level sets of asymptotic variance within a clinical context. These three scenarios are illustrated in Figure 3.4, where

an asterix '*' is used to denote the classification accuracies of a rule for which sNB_n will be most variable.

The availability of $(1-\alpha)$ confidence intervals for the true- and false-positive rates, determined from pre-validation performance of a candidate rule R, motivates interest in rectangular regions with joint confidence level equal to $(1-\alpha)^2$. Plugging the characteristics of these extremal points into the formula for the limiting variance yields a quick way to gauge how difficult it could be to validate the net benefit of R in a known or assumed clinical context. Application of this property to sample size calculations is demonstrated in Section 3.7.

3.6 Efficiency Gain from Known Event Rate

Suppose that the outcome probability ρ_0 is known and that the required outcome and risk model predictors have been measured on a cohort of size n. In this scenario, it is natural to use ρ_0 in the estimation of sNB_0 simply by substituting it for ρ_n in the empirical estimator sNB_n which gives $\text{sNB}_n^\rho = \text{TPR}_n - \omega_0 \text{FPR}_n$, where $\omega_0 = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_0}{\rho_0}$ and $\text{TPR}_n, \text{FPR}_n$ are the usual empirical estimators of the true- and false-positive rates of the rule. Adding ρ to the superscript distinguishes this estimator from the fully empirical cohort estimator.

Knowing the event rate limits the statistical model to distributions for which $P[D=1] = \rho_0$ denoted $\mathcal{M}^\rho$. An influence function of sNB_n^ρ, for estimation over $\mathcal{M}^\rho$, can be derived analogously to that of sNB_n, but with ρ a fixed constant instead of a variable. Conceptually, as a representative of a directional derivative over $\mathcal{M}^\rho$, where there can be no sensitivity to the event rate of the underlying data-generating mechanism, the influence function can have no component corresponding to ρ. Simply, $\sigma_0^{2,\rho}$ equals σ_0^2 less $\sigma_{0,P}^2$, the contribution to the total limiting variance from estimating ρ_0. Using sNB_n^ρ and $\sigma_0^{2,\rho} = \sigma_{0,T}^2 + \sigma_{0,F}^2$ in Equation 3.8, Wald confidence intervals can be constructed for empirical estimators of sNB when the outcome probability is known.

Equality in

$$\sigma_0^{2,\rho} = \sigma_{0,T}^2 + \sigma_{0,F}^2 \leq \sigma_{0,T}^2 + \sigma_{0,F}^2 + \sigma_{0,P}^2 = \sigma_0^2 \tag{3.9}$$

only occurs when ρ_0 is either 0 or 1, which are special cases of ρ_0 known. While the empirical estimator sNB_n^N remains an asymptotically linear estimator of sNB_0 when ρ_0 is known, it is no longer efficient. The estimator sNB_n^ρ is efficient for estimation over $\mathcal{M}^\rho$, but is not a viable estimator of net benefit when ρ_0 is unknown.

From this relationship, one can read the efficiency gain associated with knowing ρ_0 directly from Figure 3.1. For example, in a clinical context $C = \left(\rho_0 = 0.03, B^{ctrl}{:}B^{case} = 1{:}2\right)$, with a rule having true- and false-positive rates of 85% and 1% respectively, there would be an 11% efficiency gain associated with using sNB_n^ρ to estimate $\text{sNB}_0 = 0.69$, relative to using sNB_n^ρ which includes estimating the event rate. The relative gain in efficiency jumps to 27% when the false-positive rate is 2% instead of 1% and $\text{sNB}_0 = 0.53$.

See Appendix B.5 for additional details on the efficiency of sNB_n and sNB_n^ρ.

3.7 Cohort Study Design

The limiting variance formula (3.6) and the resulting confidence intervals directly support sample size calculations for a cohort on which to validate a specified rule R. We demonstrate this use with a few simple examples. We suppose that the decision rule proposed for validation is a rational risk rule, defined from a risk model r with high-risk threshold $r_T = 0.2$. The rational threshold is determined by a 1:4 control-to-case benefit ratio for the intervention in a clinical context for which previous development studies observed an outcome probability of roughly 11%. We also suppose that true- and false-positive rates equal to 80 and 10%, respectively, have been previously observed for the rule $R = (r, r_T)$. In all examples, the goal is to determine a minimum cohort size for assessing the clinical utility of a decision rule.

Plan to Achieve Minimum Precision with Assumed Constituent Values

One analysis criterion for the study could be to estimate the net benefit with 95% confidence within a fixed precision. Assuming that the validation population is similar to those represented by development data, the simplest calculation would plug the preliminary estimates of decision rule accuracy and outcome probability into the variance formula and Wald

confidence interval formula and then solve for the sample size. In this example, a standardized net benefit of roughly 0.6 is anticipated and the assumed limiting variance would be $\sigma^2_{\text{sNB}^N} = 2.29$. For precision within ± 0.05, the computation for a 95% CI half-width of length 0.05:

$$z_{0.975} * \sqrt{\frac{2.29}{n}} = 0.05$$

becomes an equation in n, the sample size, with solution equal to $3,513$. Enrolling a minimum of 3,513 subjects would allow estimation within $\pm$ 0.05 under the assumed design parameters. Even in this fairly successful context, fair discrimination and not especially rare outcome, the variability is considerable and a rather large sample is required to meet a fairly modest precision requirement. If the lower precision of ± 0.10 were deemed reasonable, a similar calculation would determine 878 participants to be sufficient.

A minimum precision requirement could also be expressed in terms of a percentage of the estimand. For $sNB = 0.6$, estimation within 10% of the true value requires estimation within ± 0.06. The above calculation, for a half-width equal to 0.06, determines a minimum enrollment requirement of 2,458.

Additionally Plan for Minimum Precision for Constituent Values

The constituents of net benefit are as important as the composite measure for assessing the performance of a rule. Consequently, designing a study solely for net benefit is not sufficient. The same calculations can be conducted for each of the constituents simply by using the assumed values of TPR, FPR, and ρ, and the implied values for $\sigma^2_{\text{TPR}} = \frac{\text{TPR}(1-\text{TPR})}{\rho}$, $\sigma^2_{\text{FPR}} = \frac{\text{FPR}(1-\text{FPR})}{1-\rho}$, and $\sigma^2_\rho = \rho(1-\rho)$, and solving the corresponding equation. The minimum cohort size to establish all four quantities at a minimum fixed precision level is the maximum size needed to estimate any one of the estimands at the given level. The samples sizes required for estimation within ± 0.05 and ± 0.1 fixed precision levels for the composite measure, each constituent, and all four quantities jointly, are presented in Table 3.2.

The decomposition of σ^2_{sNB}, introduced in Section 3.5, into the sum $\sigma^2_T + \sigma^2_F + \sigma^2_P$, es-

	Assumed	More Precise		Less Precise	
Estimand	Value	Within	n^{cohort}	Within	n^{cohort}
sNB^N	0.60	±0.05	3,513	±0.10	878
TPR	80%	±0.05	2,235	±0.10	559
FPR	10%	±0.05	155	±0.10	39
ρ	0.11	±0.05	150	±0.10	38
All	-	±0.05	3,513	±0.10	878

Table 3.2: Sample sizes for achieving absolute precision levels, 5% and 10%, designing for the composite and consituent estimands individually, or jointly (All).

tablishes trivially that σ^2_{sNB} is greater than any one of the three components. As the first component is exactly the limiting variance of estimators TPR_n, and the second component equals the limiting variance of estimators FPR_n scaled by a factor larger than 1, it is immediate that the sample size designed for estimating sNB is larger than that for either the true or false-positive rate, within a common absolute precision. As seen in this example, it can be by quite a margin. The relationship is less obvious for estimating the outcome probability.

In the scenario examined, the variability in estimating net benefit is considerably greater than that of any constituent. This will often, but not always, be the case. Counterexamples arise for a frequent outcome and highly accurate rule in a context with fairly moderate benefit ratio. For example, when the true-positive and true-negative rates are both 99%, $\rho = 0.25$ and $\omega = 3$ ($\frac{B^{ctrl}}{B^{case}} = 1$), $\sigma^2_\rho = 0.1875$ while $\sigma^2_{\text{sNB}} = 0.1632$; the estimator ρ_n is asymptotically more variable than sNB_n and larger cohorts would be required to estimate ρ_0 at the same absolute level of precision as sNB_0. As was discussed in Section 3.5, where the limiting variance of sNB^N_n was examined within a given clinical context, for rules with classifications near that of the perfect rule, or the treat none rule, the variability can be arbitrarily low. Hence, for any given ρ, there are some rules R for which $\sigma^2_{\text{sNB}^N_R}$ is less than the limiting variability in ρ_n, and if this is important, it should be confirmed explicitly.

Establishing sample sizes for estimation of all constituents at a fixed relative precision level, for example within 5% of the estimand, is more involved. Each quantity is being estimated at a different absolute level of precision and sample size requirements can vary wildly. Sample size calculations for the same example and for a fixed relative precision level is presented in Table 3.3. Similar tables can be constructed for precision requirements that vary by constituent.

	Assumed	More Precise		Less Precise	
Estimand	Value	Within	n^{cohort}	Within	n^{cohort}
sNB^N	0.60	±0.03	9,832	±0.06	2,458
TPR	80%	±0.04	3,492	±0.08	873
FPR	10%	±0.005	15,538	±0.01	3,884
ρ	0.11	±0.006	12,432	±0.011	3,108
All	-	-	15,538	-	3,884

Table 3.3: Sample sizes for achieving relative precision levels, 5% and 10% of the true value, designing for the composite and constituent estimands individually, or jointly (All).

Plan Conservatively from Range of Rule Accuracies

Another sample size calculation could be conducted making use of anticipated performance characteristics of a risk model, within a given clinical context. Over a square region of plausible performances, the variance formula establishes that the limiting variance corresponding to the largest false-positive rate and the true-positive rate closest to 50% bounds that of all other points considered plausible. Use of this variance would provide a more conservative approach to calculating sample size.

For example, suppose that the studies used to develop and refine the decision rule R, intended for a clinical context with outcome probability roughly 11% and intervention decision with a 1:4 control-to-case benefit ratio, had been designed to estimate the true- and

false-positive rates within ± 0.05, and with 95% confidence, the investigators believe the true-positive rate is within $[0.75, 0.85]$ and the false-positive rate lies within $[0.05, 0.15]$. Using $TPR = 75\%$ and $FPR = 0.15\%$ would produce a more conservative approach to ensuring the desired precision level is achieved. For an absolute precision level of 10%, the conservative design would require a cohort of 1,241 subjects in order to estimate sNB_R^N over the assumed accuracy range within ± 0.10. At $(TPR = 75\%, FPR = 15\%)$, the net benefit would be 0.47. Estimation of the true-positive rate to within ± 0.1 would require a cohort with 654 participants, whereas a cohort of only 54 would be required for similar estimation of the false-positive rate. Sample size requirements for estimation of the outcome probability have not changed with this approach.

3.8 Contrasting the Net Benefit of Clinical Decision Rules

A cohort validation study provides a valuable opportunity to make direct comparisons between competing decision rules. The difference in net benefit, $\Delta\text{sNB} := \text{sNB}_2 - \text{sNB}_1$, between a pair of candidate decision rules, R_1 and R_2, provides one approach. When one rule relies on the same set of patient-specific information as the other, plus an additional novel biomarker, ΔsNB gives information on the incremental improvement attributable to the biomarker.

For example, R_1 could be a decision rule defined on routine clinical and imaging variables W_1, used to determine whether or not a patient with an indeterminate lung nodule found on an incidental CT scan should undergo an invasive procedure to establish or rule-out cancer, and R_2 a rule that additionally uses a novel proteomic biomaker obtained through a minimally invasive nasal brushing, W_2; in terms of the resulting decisions, $x_1 := R_1(w_1)$ and $x_2 := R_2(w_1, w_2)$. When the two rules are rational risk rules, the two risk models will differ, but the risk threshold will be necessarily be the same as it is determined by the common control-to-case benefit ratio.

An empirical estimator of ΔsNB_0 is naturally given by $\Delta\text{sNB}_n := \text{sNB}_{2,n} - \text{sNB}_{1,n}$. If the two estimators were constructed on independent samples, then the variance of ΔsNB_n is

simply the sum of the two variances as previously established. When two decision rules are evaluated on the same data set, correlation between the two estimators, $\text{sNB}_{1,n}$ and $\text{sNB}_{2,n}$, must be appropriately accounted for when making inferential statements about ΔsNB_0 using ΔsNB_n.

By linearity, the influence function of a difference is the difference of the respective influence functions. For ΔsNB_n, this is: $IF_{\Delta\text{sNB}} := IF_{\text{sNB}_2} - IF_{\text{sNB}_1}$. The limiting variance $\sigma^2_{\Delta_{\text{sNB}},0} = \mathbb{E}_0\left[IF^2_{\Delta_{\text{sNB}},0}(O)\right]$ simplifies for the opt-in formulation to:

$$\begin{aligned}\sigma^2_{\Delta\text{sNB}^N} \quad := \quad & \frac{1}{\rho}\left\{\text{TPR}_2(1-\text{TPR}_2) + \text{TPR}_1(1-\text{TPR}_1) - 2\text{Cov}\left(\text{TPR}_1, \text{TPR}_2\right)\right\} \\ & + \omega^2 \frac{1}{1-\rho}\left\{\text{FPR}_2(1-\text{FPR}_2) + \text{FPR}_1(1-\text{FPR}_1) - 2\text{Cov}\left(\text{FPR}_1, \text{FPR}_2\right)\right\} \\ & + \omega^2 \frac{1}{\rho(1-\rho)}\left(\text{FPR}_2 - \text{FPR}_1\right)^2\end{aligned}$$

where we have dropped explicit reference to the true data-generating mechanism (0 subscripts), have continued to use subscripts i as shorthand for dependence on the decision rule R_i, and, in a slight abuse of notation, we use

$$\text{Cov}_0\left(\text{TPR}_1, \text{TPR}_2\right) := \mathbb{E}_0\left\{\frac{D}{\rho_0} X_1 X_2\right\} - \text{TPR}_{1,0}\text{TPR}_{2,0}$$

and $\text{Cov}\left(\text{FPR}_1, \text{FPR}_2\right)$ defined analogously. Each of these quantities has its own empirical estimator, given by:

$$\text{Cov}_n\left(\text{FPR}_1, \text{FPR}_2\right) := \frac{1}{n}\sum_i \frac{1-D_i}{(1-\rho_n)} X_{1,i} X_{2,i} - \text{FPR}_{1,n}\text{FPR}_{2,n}$$

and similarly for $\text{Cov}_n\left(\text{TPR}_1, \text{TPR}_2\right)$. The first term in each of these expressions estimates the proportion of cases or controls respectively, for which both decision rules assign intervention. For rules that are derived from risk models, this is the proportion that both models predict to be high-risk.

We can plug these estimators, along with the constituent estimators of sNB_n, into the above formula to get an estimator of the limiting variance of ΔsNB_n and construct asymptotically correct Wald confidence intervals as usual. In cases where one rule is an extension

of the other, one might anticipate that the covariances are both positive, and the variability is less than that for the difference of two uncorrelated rules. This implies that estimating the difference in net benefit between two rules from one same data set could be more efficient than using the difference in net benefit estimated from two different data sets; in addition to ruling out any question of difference between the samples.

3.9 Illustration

We now illustrate the entire process of risk model development and subsequent evaluation of the risk rule on independent data sampled from the same population used in our simulations (Section 3.10). The measurements of interest are three continuous covariates $W = (W_1, W_2, W_3)$ and we also record in D an indication of whether or not a particular outcome was observed. We suppose that a careful weighting of costs and benefits has led to using 20% probability of the undesirable clinical event as a threshold for defining a high-risk group whom is targeted for an intervention that would not reasonably be given to the general population (e.g., prescription of cholesterol-lowering medication). This threshold implies a 1:4 control-to-case benefit ratio.

This simulated population was employed by Pepe (2011) for discussion of reclassification methods applied to risk categories modeled after those used in the ATP-III cholesterol treatment guidelines for the management of CHD. For example, we suppose in this illustration that we are interested in predicting the risk of a first coronary event within 10 years using two established covariate (sex and age) and possibly a biomarker (ratio of total cholesterol to high-density lipoprotein levels). The goal is to identify high-risk patients for targeted prophylactic prescription of cholesterol-lowering medication.

Development

A development cohort of 800 subjects, 77 (9.6%) of whom were observed to experience the outcome, was used to fit predictive risk models using logistic regression. We call the model using the two established biomarkers PredMod; for a subject with biomarker values (W_1, W_2)

its predicted risks are calculated by taking the inverse logit of $-4.17 + 1.68W_1 + 1.06W_2$. An extended model, called $\text{PredMod}_{\text{ext}}$, additionally includes the novel biomarker,W_3, and its linear predictor equals $-6.49 + 1.47W_1 + 1.23W_2 + 2.33W_3$.

Estimates of net benefit and classification characteristics, evaluated on the development data, for the base model equal ($\text{sNB}^N_{\text{PredMod},n} = 0.523, \text{TPR}_n = 70.1\%, \text{FPR}_n = 7.6\%$), for the extended model equal ($\text{sNB}^N_{\text{PredMod}_{\text{ext}},n} = 0.779, \text{TPR}_n = 88.3\%, \text{FPR}_n = 4.4\%$), and for the difference equal ($\text{sNB}^N_{\Delta,n} = 0.256, \Delta\text{TPR}_n = 18.2\%, \Delta\text{FPR}_n = -3.2\%$), where we have chosen not to provide confidence intervals. Methods that account for the finite sample bias from evaluating the performance on the same data to which the risk model was fit and that provide valid inference in this setting are beyond the scope of this dissertation, which is focused on evaluating pre-specified decision rules. We note that in practice, cross-validation or other attempts at accounting for the potential optimism in performance may be preferred for making sample size calculations.

Design

These results observed during the development stage are deemed promising enough to warrant a validation study. For simplicity, we assume the previously observed performances of PredMod and $\text{PredMod}_{\text{ext}}$, and use these values to calculate an assumed limiting variability for the net benefit estimators of each model. By the approach demonstrated in Section 3.7, the sample sizes for estimating $\text{sNB}^N_{\text{PredMod}}$ and $\text{sNB}^N_{\text{PredMod}_{\text{ext}}}$ within ± 0.1 are 1,142 and 558 respectively. Similarly, the estimates made from development data can be used to calculate an assumed limiting variance for the difference in net benefit between the two rules, and analogous calculations lead to a minimum of 794 enrollees to estimate sNB^N_{Δ} within ± 0.1. In each of these scenarios, the sample size required for estimation of net benefit at the given precision will guarantee estimation of each constituent within the same absolute precision level.

The extended rule using the novel biomarker appeared to provide noticeably better accuracy and clinical utility than the base rule which relied only on standard clinical variables.

As a compromise between the large sample size required to estimate all parameters of interest within ± 0.01, which is driven by variability in $\text{sNB}^N_{\text{PredMod},n}$, and the much smaller enrollment requirement for a study only assessing the performance of the extended rule, we decide to enroll 800 participants in a validation study. Assuming similar performance as in development, estimation of the net benefit and classification accuracies for the extended rule and its incremental performance increases over the base rule should be achieved with the desired accuracy level; performance of the base rule slightly less so.

Validation

The same biomarker values have been collected on an independent validation cohort comprised of another 800 individuals. The outcome probability of $D = 1$, experienced by 79 participants, is observed to be $0.099\,(\,95\%\,\text{CI: } 0.078, 0.119)$ which is similar to that seen in the development cohort. The risks, and the corresponding high-risk classification and intervention decisions, are calculated from the two pre-specified models, PredMod and $\text{PredMod}_{\text{ext}}$, using the biomarker values for each subject.

The rule based on established biomarkers, $R_1 = (\text{PredMod}, r_T = 0.2)$, has modest true-$0.759\,(95\%\,\text{CI: } 0.665, 0.854)$ and false-positive $0.069\,(95\%\,\text{CI: } 0.051, 0.088)$ rates. Combining this with the observed outcome probability, yields a standardized net benefit estimate of $0.601\ (95\%\,\text{CI: } 0.492, 0.711)$, which is equivalent to that of a rule that identifies as high-risk 60.1% of would-be cases and none of the would-be controls (i.e., the equivalent control-perfect rule). The extended rule, $R_2 = (\text{PredMod}_{\text{ext}}, r_T = 0.2)$, does indeed have better discrimination with estimated true and false positive rates of $0.861\,(95\%\,\text{CI: } 0.784, 0.937)$ and $0.039\,(95\%\,\text{CI: } 0.025, 0.053)$, respectively. The estimated standardized net benefit of the extended rule equals $0.772\ (95\%\,\text{CI: } 0.687, 0.858)$, which is greater than that of the base rule.

The inclusion of the additional biomarker in the model leads to capture of roughly 17.1% more would-be cases, comparing equivalent rules that do not identify any would-be controls as high-risk; this is an interpretation the incremental difference in standardized net benefit, $\Delta_{\text{sNB}} = 0.171\,(95\%\,\text{CI: } 0.059, 0.283)$, of using $(\text{PredMod}_{\text{ext}}, r_T = 0.2)$ over

(PredMod, $r_T = 0.2$) for targeting high risk individuals for intervention. While the rule using the novel biomarker appears to offer a fair bit more clinical utility, over treating no-one, than using only the established predictors, the true incremental increase in net benefit could reasonably, with 95% confidence, be anywhere between 6 and 28% of the benefit gain achievable by a perfect rule. Estimation from the validation set provided accuracy of ± 0.11, which is quite close to the design target of a ± 0.10 precision level.

3.10 Simulated Performance of sNB_n^N

All simulations are based on the same super-population used in the example data analysis. Briefly, three covariates $W = (W_1, W_2, W_3)$ are distributed according to a mixture of two trivariate normal distributions with identity variance matrices and means $\mu_0 = 0_3$ and $\mu_1 = (1.7, 1, 2)$, with mixture probabilities of 90 and 10%, respectively. The outcome D is binomial with probability conditional on Y following a logistic relationship and overall probability of 10%. Using these features, a super-population consisting of 1 million observations was generated. For full details of this illustration population, we refer the reader to the supplemental materials of Pepe (2011).

We first evaluate the performance of a pre-specified baseline risk rule, PredMod, on a validation cohort. A logistic model for D, as predicted by two of the covariates W_1 and W_2, was fit on a development set of 600 randomly selected observations. Validation samples of sizes N_{valid} equal to 400, 800, and 2000 were sampled from the remaining super-population. The net benefit for four possible clinical decision rules, combining the pre-specified risk model with risk thresholds of 10, 20, 30 or 40% (corresponding to control-to-case ratios of 1:9, 1:4, 1:2.5 and 2:3), is evaluated in the context of identifying high-risk individuals to target for an intervention that is not routinely administered.

To study the frequentist properties of the analytic variance and confidence intervals proposed for validating a clinical decision rule in terms of net benefit, selection of the validation samples and evaluation of standardized net benefit were conducted 5,000 times for each model developed. The true net benefit, sNB_0, of a pre-specified clinical decision rule, R=(r, r_T), is

determined by using the predicted values for the entire super-population.

			Std. Dev.		Coverage	
	sNB_0	% Bias	MC	Wald	BS	Wald
r_1	0.531	-0.40	0.060	0.060	94.7	94.6
r_2	0.526	-0.41	0.059	0.060	95.4	95.2
r_3	0.517	-0.68	0.062	0.062	94.9	94.6
r_4	0.530	-0.55	0.060	0.061	95.2	94.9
r_5	0.532	-0.06	0.061	0.060	95.1	94.4

Table 3.4: Estimation of sNB(rT=0.2) for 5 prespecified models (PredMod) developed from 5 independent samples of N=400) over 5000 replicates of external validation sets (N=800).

Table 3.4 summarizes estimates of the net benefit for each of 5 different pre-specified models r_i , from validation samples of size 800, for the five corresponding decision rules using a 20% high-risk threshold: R_i=$(r_i, r_T = 20\%)$. Each independently developed model has a different true, model-dependent, value of sNB; this is stated in the column labeled sNB_0. Overall, the bias, as a percent of true sNB, is fairly small and analytic estimates of variance match the observed monte carlo variance of estimated performance quite well. Coverage of 95% confidence intervals are similarly on target with slightly greater coverage for bootstrap constructed compared to analytic Wald confidence intervals, but there is no clear trend for which is superior.

Table 3.5 examines in more depth the properties of validation for one pre-specified model (model r_1 in Table 3.4). Overall, the point estimates and estimates of variability behave quite well in terms of both bias and average estimates of variability matching observed Monte Carlo variability across simulations. The usual improvements with increasing sample size are apparent.

The coverage rates of empirical estimators of the true- and false-positive rates display the well-known performance drop associated with using Wald intervals for estimates of proba-

		sNB				TPR				FPR				
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
							True Value							
		0.671	0.531	0.431	0.353	0.840	0.732	0.643	0.563	0.168	0.090	0.055	0.035	0.100
							% Bias							
400		-0.48	-1.05	-1.53	-1.95	0.14	-0.03	-0.09	-0.17	-0.01	0.05	0.24	0.18	-0.02
800		-0.32	-0.40	-0.56	-0.86	0.04	0.02	-0.03	-0.15	-0.08	-0.38	-0.46	-0.44	-0.18
2000		0.01	0.01	0.01	0.05	0.02	0.02	0.03	0.06	-0.17	-0.17	-0.13	-0.13	0.26
							Standard Deviation							
400	observed	0.069	0.086	0.098	0.106	0.058	0.071	0.078	0.079	0.020	0.015	0.012	0.010	0.015
	analytic	0.068	0.086	0.098	0.106	0.058	0.070	0.076	0.078	0.020	0.015	0.012	0.010	0.015
800	observed	0.048	0.060	0.068	0.074	0.041	0.050	0.054	0.056	0.014	0.011	0.008	0.007	0.011
	analytic	0.048	0.060	0.068	0.074	0.041	0.049	0.054	0.055	0.014	0.011	0.008	0.007	0.011
2000	observed	0.030	0.038	0.043	0.046	0.026	0.032	0.034	0.035	0.009	0.007	0.005	0.004	0.007
	analytic	0.030	0.038	0.043	0.046	0.026	0.031	0.034	0.035	0.009	0.007	0.005	0.004	0.007
							95% Coverage							
400	bootstrap	94.6	95.3	95.1	94.8	93.2	94.5	94.3	94.8	94.8	94.7	94.7	94.5	94.3
	analytic	93.6	94.7	94.7	94.7	91.8	93.0	93.6	93.9	94.8	94.2	94.4	93.4	94.7
800	bootstrap	95.1	94.7	94.6	94.8	94.5	94.4	94.9	94.7	95.7	94.6	94.6	94.8	94.5
	analytic	94.4	94.6	94.8	94.6	93.6	93.8	94.4	94.2	95.6	94.4	94.4	94.0	95.0
2000	bootstrap	95.0	94.8	95.2	95.2	94.6	94.4	94.5	94.8	94.5	94.8	94.9	94.1	94.5
	analytic	94.8	94.7	95.1	95.1	94.3	94.1	94.3	94.5	94.5	94.9	94.7	93.9	94.7

Table 3.5: Simulation of validating a prespecified model (PredMod) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

bilities in smaller sample sizes and when the true values are extreme (close to 0 or 1). This is most pronounced in estimators of TPR because their effective sample size (cases only) averages only 10% of the full cohort. When the validation sample contains 400 observations, only 40, on average, contribute to estimation of $\text{TPR}_0 = 84\%$ ($r_T = 0.1$ scenario) and the observed coverage of the Wald constructed confidence intervals is 91.8% while that of bootstrapped CI is 93.2%. Similarly, 360 observations, on average, contribute to estimation of $\text{FPR}_0 = 3.5\%$ ($r_T = 0.4$ scenario); though the contributing sample size is many fold greater than that for the true-positive rates, the more extreme estimand works to yield analytic Wald confidence intervals with slightly low coverage of 93.4% which is lower than the 94.5% coverage exhibited by bootstrapped CI .

These effects on the coverage of the true- and false-positive rates translate, somewhat tempered, to estimators of net benefit. Coverage of sNB_0^N is somewhat low, 93.6% in the scenario most challenging for estimating TPR_0. Otherwise, the coverages observed generally fall between 94.5 and 95.1% for the analytic intervals and agree well with the coverage demonstrated by bootstrapped confidence intervals. Wald confidence intervals rely on a Normal approximation which holds asymptotically, but for extreme probabilities requires larger samples before being reasonably good. This is an advantage of confidence intervals defined by percentiles of a bootstrap distribution. Wald confidence intervals constructed using a bootstrapped estimate of standard error exhibit similar points of weakness in terms of coverage as the Wald intervals using the analytic estimates of variability.

The same results were considered in the evaluation of the clinical utility associated with using predicted risk from the extended model PredMod_{ext}. The same general trends are observed, but the improved discrimination of the extended model require estimation of true- and false-positive rates that are more extreme than those achieved by counterparts for PredMod. In the most extreme scenario ($N_{valid} = 400$ and TPR=91.8%), coverage as low as 88.2% by TPR_n intervals and 91.6% by sNB_n intervals are observed. Rules with very good performance characteristics, especially true-positive rates in a low-event rate scenario, will require larger sample sizes to achieve better analytic coverage. The analytic estimators of

variability perform well in these settings, though the Normal distribution approximation assumed by the Wald approach is the weakness.

			Std. Dev.		Coverage	
	ΔsNB_0	% Bias	MC	Wald	BS	Wald
r_1	0.230	0.13	0.056	0.057	95.3	95.2
r_2	0.233	0.25	0.055	0.056	94.8	94.8
r_3	0.226	0.38	0.051	0.052	95.0	94.9
r_4	0.228	0.80	0.052	0.052	94.9	94.5
r_5	0.213	0.34	0.050	0.050	94.4	94.5

Table 3.6: Estimation of ΔsNB(rT=0.2) for 5 prespecified models developed from 5 independent samples of N=400) over 5000 replicates of external validation sets (N=800).

Tables 3.6 and 3.7 present the same results for the contrast between the net benefits of the two risk models. On average, the contrasts between the true-positive rates are all positive (at each risk threshold, the model extended to include the novel biomarker has more high-risk classification among controls) and the contrasts between the false-positive rates are all negative (the extended model has less high-risk classifications among controls). The same general trends in performance are observed for the contrast as for each individual measure of net benefit. Overall, bias is small, analytic estimates of variability match the Monte Carlo variability observed across simulations. However, coverage is quite close to the target 95% and on par with that of bootstrap percentile intervals. The observed improvements in coverage are explained by estimators for differences in proportions generally achieving normality faster, approximations are more reasonable at smaller samples, than for single proportions.

Extended results tables, including bootstrapped standard errors and corresponding Wald CI, and results for the extended model are in Appendix B.6.

		ΔsNB				ΔTPR				ΔFPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
						True Value							
		0.164	0.230	0.268	0.289	0.079	0.144	0.196	0.241	-0.086	-0.038	-0.019	-0.008
						% Bias							
400		-0.49	0.23	0.22	0.07	-3.45	-0.74	-0.15	0.02	-0.38	-0.75	-1.32	-2.40
800		0.25	0.13	0.34	0.45	-0.96	-0.19	0.39	0.57	-0.22	-0.91	-1.32	-1.41
2000		-0.06	-0.01	-0.01	-0.06	-0.02	0.16	0.04	0.07	-0.31	-0.50	-0.35	-0.96
						Standard Deviation							
400	observed	0.065	0.081	0.093	0.105	0.059	0.072	0.077	0.081	0.020	0.016	0.013	0.011
	analytic	0.064	0.080	0.093	0.106	0.059	0.070	0.076	0.079	0.020	0.016	0.013	0.011
800	observed	0.045	0.056	0.065	0.073	0.042	0.049	0.054	0.056	0.014	0.011	0.009	0.008
	analytic	0.045	0.057	0.065	0.074	0.042	0.049	0.053	0.056	0.014	0.011	0.009	0.008
2000	observed	0.029	0.036	0.041	0.046	0.026	0.032	0.034	0.035	0.009	0.007	0.006	0.005
	analytic	0.028	0.036	0.041	0.046	0.026	0.031	0.034	0.035	0.009	0.007	0.006	0.005
						95% Coverage							
400	bootstrap	94.0	94.4	94.5	94.9	92.8	93.1	94.3	94.3	94.6	94.4	94.1	94.6
	analytic	93.5	94.1	94.5	95.0	92.4	92.7	93.8	93.8	94.6	94.7	94.5	95.0
800	bootstrap	94.9	95.3	95.1	94.9	94.6	94.6	94.3	94.8	95.4	95.2	95.2	94.5
	analytic	95.1	95.2	95.0	95.1	94.6	94.7	94.1	94.6	95.5	95.3	95.2	95.1
2000	bootstrap	94.8	94.4	94.3	94.8	94.7	94.0	94.5	94.3	94.5	94.9	95.4	95.2
	analytic	94.8	94.3	94.4	95.0	94.9	93.9	94.3	94.3	94.5	95.1	95.5	95.4

Table 3.7: Simulation of validating a prespecified model (PredMod - PredMod$_{ext}$) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

3.11 Summary

In this chapter, we established analytic inference for empirical estimators of net benefit in the context of a validation cohort. We presented formulas for the influence functions and limiting variance of empirical estimators of net benefit on either the standardized or unstandardized scale and with respect to both the opt-in and the opt-out formulations of net benefit. The usefulness of net benefit as a summary measure of clinical utility and the importance of assessing variability were demonstrated by augmenting the results from published studies. We used the analytic formulas to explore the structure of the variability through the decomposition of the asymptotic variance into components associated with estimation of each constituent of net benefit and by viewing the values of the asymptotic variance as a function of the classification accuracies of possible rules, within a given clinical context. Insights from the former approach revealed efficiency gains available to estimation scenarios in which the outcome probability were known. Insights from the latter approach were employed in example sample size calculations. Distribution theory for empirical estimators of the net benefit of a single clinical decision rule was extended to support conducting inference for the difference in net benefit between two rules. After the full process of risk rule development, validation study design, and assessment were illustrated, the frequentist properties of proposed estimators were evaluated through a simulation study.

Chapter 4

ADDING CONFIDENCE BANDS TO RELATIVE UTILITY CURVES

In this chapter we are concerned with conducting inference on estimates of relative utility curves. After introducing the notions of relative utility and decision curves, we propose confidence bands as a means of describing the variability inherent to estimating relative utility curve that is more appropriate than current approaches. We present a numerical algorithm for constructing approximate confidence bands and provide the formulas necessary to apply the algorithm in the context of relative utility curves. The use of confidence bands is demonstrated in an example and their performance evaluated by simulation.

4.1 Relative Utility and Decision Curves

The concept of a decision curve is generally restricted to clinical decision rules that are derived from a risk model. By varying the high-risk threshold r_T over the entire [0,1] range, a risk model r gives rise to a family of classifiers that could be summarized by a receiver operating characteristic (ROC) curve. Each choice of threshold defines a clinical decision rule $R = (r, r_T)$. If we restrict consideration of these rules to contexts in which the control-to-case trade-off ratio, of a potential intervention, makes the rule rational, then the net benefit could be calculated for each rule using $\omega = \frac{r_T}{1-r_T}\frac{1-\rho}{\rho}$. Both the collection of rules and the corresponding values of net benefit are indexed by the risk threshold.

A decision curve is a plot of the unstandardized net benefit as a function of the risk threshold r_T, or equivalently as a function of the benefit tradeoff ratio for which the rule is rational $\frac{r_T}{1-r_T} = \frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}$ (Elkin and Vickers, 2006). A version on the standardized scale was introduced as a relative utility curve (Baker et al., 2009; Baker, 2009). Elkin and Vickers

(2006) proposed decision curves as a tool for understanding regions of relative benefit over which decision making could be improved from utilizing a risk model. A decision curve, or relative utility curve, provides an assessment of a risk model, or the biomarker predictors, that could be useful at earlier stages of development, when the clinical contexts and decisions for which the candidate biomarker may prove useful are still being identified. Decision curves also provide a means of conducting sensitivity analysis on the assumed control-to-case benefit ratio. This could be useful when expert opinions fall in a range rather than consensus on a single value.

While a decision curve can be defined over the entire range of possible thresholds between 0 and 1, this is not wholly rational or practical. As previously noted, for control-to-case benefit ratios that correspond to a rational risk threshold below the outcome probability, the rational universal decision is to treat everyone and the opt-out formulation sNB^A would provide the relevant contrast of expected utility. Similarly, for benefit ratios above the outcome probability, sNB^N would be the relevant contrast of expected utility. In many applications, one or the other of these regions, $(0, \rho)$ or $(\rho, 1)$, will clearly be of interest. Further, thresholds at either extreme, very close to 0 or to 1, correspond to benefit ratios that are either extremely close to zero (a correctly identified control confers negligible benefit relative to a case) or get arbitrarily large (a correctly identified case confers negligible benefit relative to a control). These ratios are either unrealistic or pertain to situations in which risk rules are likely not good candidates for guiding decisions.

4.2 Confidence Bands

When the estimand of interest is a function, such as for a decision curve, a generalized notion of confidence interval is available. Confidence bands are often used in survival analysis when the survival curve is of interest, not just the survival probability at one time point. They are equally applicable when the decision curve is of interest, not just the net benefit of a single rule defined from the risk model at one risk threshold. A confidence band is a pair of functions, with the same domain as the estimand, that provide upper and lower bounds

for the estimated function, analogously to how a confidence interval provides an upper and lower bound for a point estimate. The frequentist concept of coverage for a 95% confidence band is that, when constructed around estimates from many repeated samples, 95% of the bands would contain the estimand, the true curve, at every point of its domain. This coverage property is sometimes emphasized by referring to confidence bands as simultaneous confidence bands.

A confidence interval can be calculated for a estimator of a function at a single point. When this is done for every point, or a fine grid of points, in the domain, a set of upper and lower bounds for the estimated function is created. We refer to these apparent 'bands' as pointwise confidence intervals in order to distinguish them from true confidence bands. Publicly available software supporting decision curve analysis, on either the standardized or unstandardized scale (DecisionCurve package for R; www.decisioncurveanalysis.org), currently supports adding pointwise confidence intervals estimated by bootstrapping. The results of Chapter 3 could be used to add analytically calculated pointwise confidence intervals.

Confidence bands, unlike pointwise confidence intervals, account for the multiple estimates of net benefit, corresponding to rules defined by different thresholds, and their correlation. Confidence bands are more appropriate for assessing the variability of a function estimand, and in particular for assessing the variability in an estimated decision curve. Because of the more stringent coverage criteria incorporated into the definition of a confidence band, they are generally anticipated to be wider than their pointwise counterparts.

4.3 Theoretical Construction of Confidence Bands

Let $\psi_0(t)$ be a functional estimand defined on $\mathcal{T}$, e.g., sNB_0 as a function of the risk threshold, and $\psi_n(t)$ an unbiased estimator of $\psi_0(t)$. Analogous to the construction of approximate Wald confidence intervals for asymptotically normal point estimators, we consider $1-\alpha$ confidence bands of the form $\psi_n(t) \pm u_n(t)$, $t \in \mathcal{T}$, where

$$u_n(t) = c\sqrt{\frac{\sigma_n^2(t)}{n}}$$

for a critical value c that satisfies

$$Pr\left(\sup_{t\in\mathcal{T}}\left|\frac{\psi_n(t)-\psi_0(t)}{\sqrt{\frac{\sigma_n^2}{n}}}\right|\leq c\right)\approx 1-\alpha.$$

The coverage property for the confidence band applies to all $t \in \mathcal{T}$ simultaneously, as captured by the supremum within the probability. The centered and scaled functional estimator is an empirical process. Calculation of the constant c requires weak convergence of the empirical process and knowledge of the limit process.

In the case of a cumulative distribution function $\mathbb{F}$, the empirical process $\sqrt{n}\,(\mathbb{F}_n - \mathbb{F}_0)$, not normalized to unit variance, converges weakly to a Brownian Bridge process composed with the cumulative distribution. Further, the supremum of a Brownian Bridge process composed with a cumulative distribution function is independent of $\mathbb{F}$ and follows the Kolmogorov distribution (Kolmogorov, 1933). The true- and false-positive rates, as a function of the risk threshold, are both complements of case or control-specific cumulative distribution functions. Confidence bands for these two processes can be constructed by looking up the appropriate constant from the known Kolmogorov distribution. Such an approach was proposed by Campbell (1994) as a means of constructing a confidence band for an empirical ROC curve.

For more general processes, such as the net benefit process over varying risk thresholds, establishing the limiting distribution and the corresponding supremum is a difficult task. Due to the many possible covariance or correlation processes, this is true even when the empirical process is known to converge weakly to a Gaussian process. Consequently, numerical approaches have been embraced as a practical means to constructing approximate confidence bands. (Degras, 2011)

The multiplier bootstrap provides one approach to constructing approximate confidence bands for a convergent empirical process when the the influence function is known and estimable at all $t \in \mathcal{T}$. This method scales the estimated influence functions of the process by samples from a standard normal distribution and then calculates their supremum over $t \in \mathcal{T}$. Repeating this many times produces an approximation of the limiting distribution of

the supremum of the process, from which a critical value can be calculated as the appropriate percentile. This method relies on the fact that the process of the product of a standard normal random variable and the influence functions of a functional estimator has the same first and second moments as the original empirical process and is justified by the multiplier bootstrap theorems in Section 2.9 of Van Der Vaart and Wellner (1996) .

Another quite similar strategy is to approximate the distribution of by sampling directly from an approximation of the limiting process. For the empirical estimators of net benefit studied within this dissertation, the limiting variance and covariances of the process can be calculated explicitly and estimated directly from estimators of the true- and false-positive rate constituent processes. We demonstrate this approach in the following section.

4.4 Numerical Construction of Confidence Bands

We now present a numerical algorithm, that we attribute to Dudoit et al. (2004), for constructing approximate confidence bands, applied to relative utility curves, over a specified interval of thresholds $[t_a, t_b]$. Dependence on the underlying pre-specified risk model is suppressed in the notation that follows — for example, we write $\text{sNB}_n(t)$ to indicate the estimated standardized net benefit of the rule $R = (r, t)$.

i Fix a sufficiently fine set of thresholds $\{t_1, \ldots, t_M\}$, evenly spaced over $[t_a = t_1, t_b = t_M]$.

ii Estimate an MxM correlation matrix Σ with $\Sigma_{ii} = 1$ and for $i \neq j$
$\Sigma_{ij} = \frac{\text{Cov}_n(t_i, t_j)}{\sqrt{\sigma_n^2(t_i)\sigma_n^2(t_j)}}$, where $\sigma_n^2(t_i)$ is the plug-in estimate of the asymptotic variance of $\text{sNB}_n(t_i)$ and $\text{Cov}_n(t_i, t_j)$ is the estimate of the asymptotic covariance given in Equation 4.1.

iii Draw B samples from $N(0, \Sigma)$, and for each sample $Z_i = (z_{1,i}, \ldots, z_{M,i})$, let
$m_i := max_j |z_{j,i}|$

iv Let $Q_{1-\alpha}$ be the $(1-\alpha)^{\text{th}}$ percentile of the set $\{m_1, \ldots, m_B\}$.

v Values of the $(1-\alpha)^{\text{th}}$ confidence band at the set thresholds are given by $CB(t_i) := \text{sNB}_n(t_i) \pm Q_{1-\alpha}\sqrt{\frac{\sigma^2_{\text{sNB},n}(t_i)}{n}}$.

The normal random variables Z play the role of deviations from the estimand; taking their maximum in step (iii) keeps track, on a uniform scale, of how wide a band centered around the estimated curve would need to be in order to cover the set of deviations. The quantile of the deviations is calculated from absolute values, and hence the $(1-\alpha)^{\text{th}}$ percentile in step (iii), after being rescaled in step (v) to reflect the observed variability, produces the desired level of coverage for the two-sided band.

The algorithm above could be applied to any estimand that is a function of a single argument; we have referred to relative utility curves for concreteness. For example, using the point estimator ΔsNB in place of sNB, and the limiting variance σ^2 and covariance Cov specific to Δ_{sNB} in place of those specific to sNB, the process above could be used to approximate a confidence band for the difference in two relative utility curves. The formula for the covariance term specific to Δ_{sNB} is given in Equation 4.2.

Correlation Matrix for a Single Clinical Decision Rule

When assessing the net benefit of a single clinical decision rule, we define the limiting covariance function of the sNB process as $\text{Cov}\,(t_i, t_j) = \mathbb{E}\left[IF_{\text{sNB}(t_i)}(O) IF_{\text{sNB}(t_j)}(O)\right]$. For each pair (t_i, t_j) of thresholds, this is the limiting covariance between the estimators of performance for two risk rules, defined by one risk model at two different risk thresholds. An empirical estimator of the limiting covariance can be defined in terms of the constituents of $\text{sNB}_n(t_i)$ and $\text{sNB}_n(t_j)$. To highlight the connection to a Brownian Bridge process, we express it as follows:

$$\begin{aligned}
\text{Cov}_n\,(t_i, t_j) \;=\; & \frac{1}{\rho_n}\left\{\text{TPR}_n(t_i) \wedge \text{TPR}_n(t_j) - \text{TPR}_n(t_i)\text{TPR}_n(t_j)\right\} \\
& + \omega_{i,n}\omega_{j,n}\frac{1}{1-\rho_n}\left\{\text{FPR}_n(t_i) \wedge \text{FPR}_n(t_j) - \text{FPR}_n(t_i)\text{FPR}_n(t_j)\right\} \\
& + \omega_{i,n}\omega_{j,n}\frac{1}{(1-\rho_n)\rho_n}\text{FPR}_n(t_i)\text{FPR}_n(t_j),
\end{aligned} \tag{4.1}$$

where $\omega_{k,n}$ denotes the usual weight $\frac{t_k}{1-t_k}\frac{1-\rho_n}{\rho_n}$, and the $\wedge$ symbol indicates the minimum of the two arguments. By monotonicity of the false-positive rates, $\text{FPR}_n(t_i) \wedge \text{FPR}_n(t_j)$ equals the false-positive rate at the maximum threshold $(t_i \vee t_j)$: $\text{FPR}_n(t_i \vee t_j)$. The weight corresponding to the rule defined by threshold t_k is defined in terms of the trade-off ratio: $\frac{\text{B}^{ctrl}}{\text{B}^{case}} = \frac{t_k}{1-t_k}$ for which the rule $R_k = (r, t_k)$ would be rational.

The entries in the correlation matrix, used in step (ii) of the confidence band construction, have the form $\Sigma_{ij} = \frac{\text{Cov}_n(t_i,t_j)}{\sigma_n(t_i)\sigma_n(t_j)}$, where Cov_n is defined in Equation 4.1 and σ_n can be calculated using Formula 3.7.

Correlation Matrix for a Difference Between Clinical Decision Rules

When assessing the difference in net benefit between two clinical decision rules, we define the limiting covariance function of the ΔsNB process $\text{Cov}\,(t_1, t_2) = \mathbb{E}\left[IF_{\Delta_{\text{sNB}(t_1)}(O)} IF_{\Delta_{\text{sNB}(t_2)}(O)}\right]$. For each pair of thresholds (t_i, t_j), this is the covariance between estimators of the difference in net benefit, between two rules derived from different risk models, at two different risk thresholds. Because we now have two risk models and two risk thresholds to work with, we extend previous notation. Let $R_{kl} = (r_k, t_l)$ denote the decision rule derived from risk model r_k using t_l as the high-risk threshold and $X_{kl} = I\left[r_k(W_k) > t_l\right]$ the corresponding decision, where W_k are the predictors of model r_k. Further, we define $Cov_n^d(R_{kl}, R_{mn}) := \mathbb{E}_n\left[X_{kl}X_{mn} \mid D = d\right] - d\text{TPR}_n(R_{kl})\text{TPR}_n(R_{mn}) - (1-d)\text{FPR}_n(R_{kl})\text{FPR}_n(R_{mn})$. An empirical estimator of the limiting covariance, for the difference between rules defined by models 1 and 2 at thresholds t_i, t_j, can be written as follows:

$$\begin{aligned}
\text{Cov}_n\,(t_i, t_j) &= \frac{1}{\rho_n}\left\{Cov_n^1(R_{2i}, R_{2j}) - Cov_n^1(R_{2i}, R_{1j}) - Cov_n^1(R_{1i}, R_{2j}) + Cov({}_n^1 R_{1i}, R_{1j})\right\} \quad (4.2)\\
&+ \frac{w_i w_j}{1-\rho_n}\left\{Cov_n^0(R_{2i}, R_{2j}) - Cov_n^0(R_{2i}, R_{1j}) - Cov_n^0(R_{1i}, R_{2j}) + Cov_n^0(R_{1i}, R_{1j})\right\}\\
&+ \frac{w_i w_j}{\rho_n(i-\rho_n)}(\text{FPR}_n(R_{2i}) - \text{FPR}_n(R_{1i}))(\text{FPR}_n(R_{2j}) - \text{FPR}_n(R_{1j})),
\end{aligned}$$

where $\omega_{k,n}$ denotes the weight $\frac{t_k}{1-t_k}\frac{1-\rho_n}{\rho_n}$.

The entries in the correlation matrix, used in step (ii) of the confidence band construction, have the form $\Sigma_{ij} = \frac{\text{Cov}_n(t_i,t_j)}{\sigma_n(t_i)\sigma_n(t_j)}$, where Cov_n is defined in Equation 4.2 and σ_n can be

calculated using Formula 3.10.

Assumptions

Pointwise, for each specific rule or difference, the centered and scaled empirical estimators of sNB^N or ΔsNB^N are asymptotically Normal. The asymptotics of an empirical estimator of a decision curve is governed by an empirical process. The above construction relies on the weak convergence of $\text{sNB}\,(r_T)$ to a Gaussian process. The convergence is with respect to asymptotics of the estimator, sample size $n \to \infty$, the number of sampled deviations $B \to \infty$, as well as the maximum distance between adjacent thresholds in the grid $\min_i |t_i - t_{i+1}| \to 0$. The thresholds do not need to be evenly spaced, but this simplification ensures that the fineness of the set increases with the number of thresholds considered; i.e., as $M \to \infty$.

Weak convergence of sNB as a process should be anticipated by the fact that it is a weighted difference of the complement of the cumulative distribution functions of the risk distributions among cases and among controls. Since the weight involves the odds of the risk threshold, which is unbounded as $r_T \to 1$, we must either trim the allowable domain to have upper limit less than 1, or make assumptions about the risk distribution among controls to ensure uniform control. Neither poses a practical limitation. Additional details are available in Appendix B.7.

4.5 Relation Between Confidence Bands and Hypothesis Tests

A confidence band can be viewed as an inverted hypothesis test. In the case of the decision curve $\text{sNB}_n^N(t)$ for a single rule, the null hypothesis of the corresponding test is that the rule provides zero net benefit across all thresholds in the domain $t \in \mathcal{T}$ considered. For a contrast between two rules, $\Delta\text{sNB}_n^N(t)$, the null hypothesis of the corresponding test is that there is no difference in net benefit between the rules at any threshold in the range considered. In terms of the confidence band, the null hypothesis should be rejected if the band excludes zero, at one or more points of the domain. The coverage property employed in the construction of the confidence band directly ensures that the type-I error is α. This

can be confirmed directly.

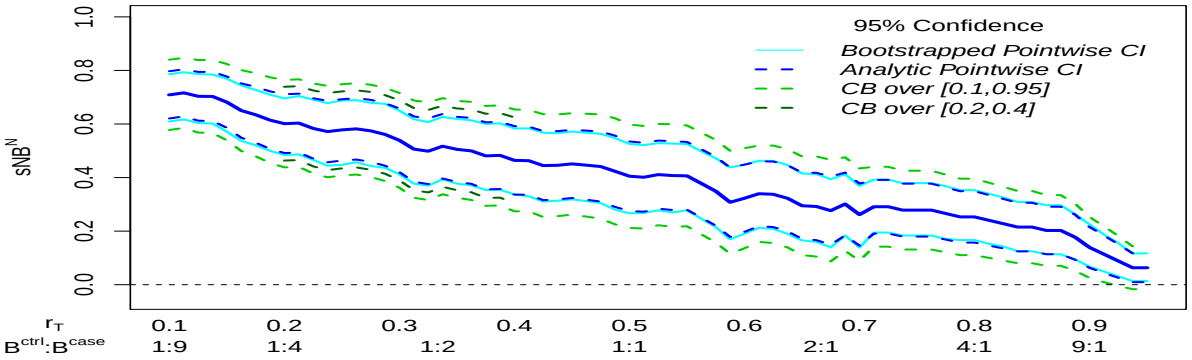
Under the null, $\text{sNB}_0(t) = 0\ \forall t \in \mathcal{T}$. If the null is rejected, then there exists a $t_c \in \mathcal{T}$ for which 0 is not in the confidence band. This implies that $\sqrt{n}\frac{\psi_n(t_c)}{\sigma^2(t_c)}$ does not lie in $[-c, c]$ which in turn implies $\sqrt{n}\left|\frac{\psi_n(t_c)}{\sigma^2(t_c)}\right|$ and hence $\sup_{t\in\mathcal{T}}\sqrt{n}\left|\frac{\psi_n(t)}{\sigma^2(t)}\right|$ are both greater than c. Under the null hypothesis, this last event occurs with probability $\approx 1-(1-\alpha) = \alpha$, by construction.

The power of a statistical test at a particular alternative is often of interest. Computation of the confidence bands and corresponding test statistic rely on uniform consistency of both σ_n^2 and Cov_n as estimators of the respective features of the limit process in order to ensure the designed coverage or error. On the other hand, calculation of power at an alternative requires using the known covariance of the assumed alternative process. The composite nature of net benefit implies that for any alternative, stated in terms of a value of net benefit, there are many possible combinations of constituents $(\rho, \text{TPR}, \text{FPR})$ that give rise to a given alternative but with differing variance and covariance processes. Hence, this task can become quite complex in general. However, the presented formulas provide a basis for exploring power across scenarios plausible for a given application.

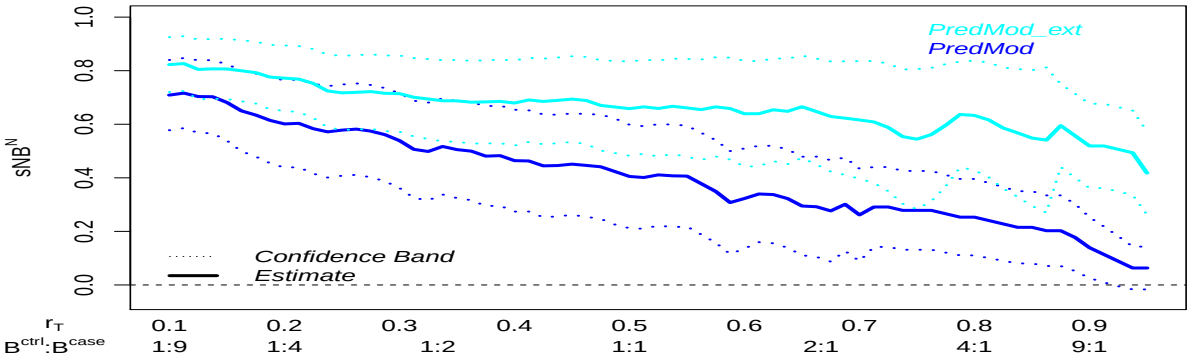
4.6 Illustration

We now continue the illustration from Section 3.9 from a more exploratory perspective that would be common at earlier stages in biomarker development. For brevity we will limit our analysis to thresholds that are at least 10%, the observed outcome rate, which yields rules rational for an opt-in decision. We also decide to limit our analysis to thresholds no greater than 95%, which caps considered control-to-case benefit ratios at 19:1.

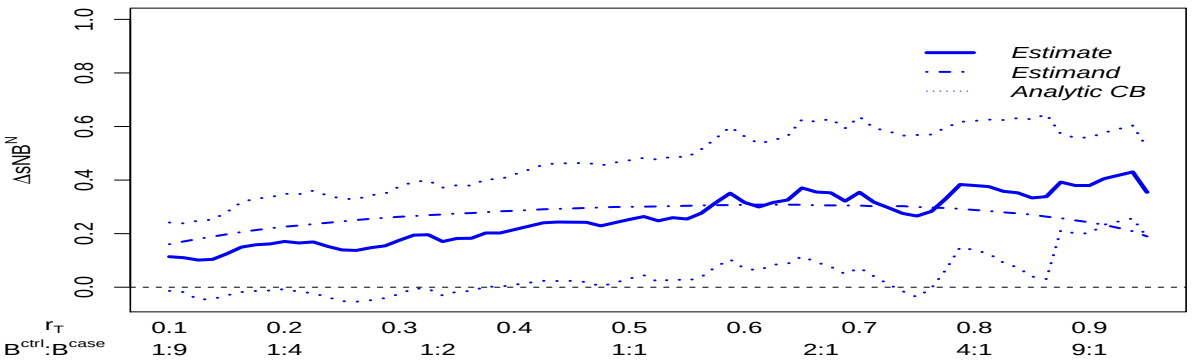
Recall that the PredMod model, based on two routine clinical variables W_1, W_2, was fit on a development cohort of $N = 800$. We now assess its potential clinical utility on the independent validation cohort of the same size by plotting the estimated relative utility curve in Figure 4.1(a). To the estimates of $\text{sNB}^N_{\text{PredMod}}(t)$ for the rules $R = (\text{PredMod}, t)$ at each threshold $t \in [0.1, 0.95]$, we have added descriptors of variability using: pointwise and analytic confidence intervals, and a confidence band, all for a 95% level of confidence.



(a)



(b)



(c)

Figure 4.1: (a) Descriptions of variability for a decision curve. (b) Decision curves for competing risk rules. (c) Confidence band for the difference between the competitors.

The pointwise confidence intervals, based on percentiles of a bootstrap distribution (1,000 resamples) and analytic intervals, using asymptotically correct Wald formulas, are in good agreement. The confidence band is wider than the pointwise intervals, as expected; it captures the variability of the entire curve and simultaneously provides coverage for multiple estimands. There is considerable variability inherent in the estimates. This aspect of the assessment of the risk model is unappreciated when estimates are not accompanied by descriptions of variability, and are underestimated when only pointwise intervals are provided.

We note that the range of thresholds considered, $[0.1, 0.95]$, corresponds to a fairly wide range of control-to-case benefit trade-offs: $\frac{B^{ctrl}}{B^{case}}$ ranges from 1:9 to 19:1. If this evaluation were being conducted at a later stage of development, one would expect a well defined clinical decision to be under consideration and a lack of consensus on benefit trade-offs to be limited to a smaller range. In a more focused relative utility curve analysis, we would still expect the confidence bands to be wider than pointwise confidence intervals, but to a lesser extent. This is demonstrated by the confidence band calculated as if the range of thresholds considered was $[0.2, 0.4]$, which lies between the full range confidence band and the pointwise confidence intervals.

We have also shown the relative utility curves for the base and extended PredMod models in Figure 4.1(b). Comparison of just the decision curves would lead to conclusions that the expected net benefit of rules based on the extended model clearly dominates, across all thresholds, that of the model using only the two established predictors. Even in this setting, where we are evaluating the addition of a useful biomarker with a fair sized validation sample, consideration of 95% confidence bands suggests that caution is warranted prior to making conclusions of dominance. Statements about the dominance of one curve over another should be accompanied by confidence bands calculated for the contrast between the net benefits of the two models under comparison.

The difference between the relative utility curves for the extended model and the base model is shown in Figure 4.1(c) accompanied by the 95% confidence band appropriate for the curve $\Delta\text{sNB}^N(t)$. Again, even in this case of a strong biomarker, and assessment of

net benefit from a reasonably sized sample $N_{valid} = 800$, the claim that $\text{PredMod}_{\text{ext}}$ clearly dominates PredMod over the entire range [0.1,0.95] of thresholds cannot be substantiated with statistical significance at a 95% confidence level. The dashed curve in the figure indicates the true difference in relative utility curves, the estimand, which is not fully covered by the estimated confidence band in this example.

4.7 Practical and Numerical Considerations

When the number of thresholds used in the calculation of the relative utility curve is many and the number of observations in the dataset is relatively small, the number of individuals who change from high-risk to low-risk between t_i and t_{i+1} can be quite small, this is even more so when considering only cases.

Instability in the Tail

When the size of the validation cohort is fairly small, the relative utility curve can look fairly jumpy, especially at higher risk thresholds. This is primarily driven by the small number of cases who, if the risk model is any good, have high risks. A fairly small number of cases who go from being classified as high-risk to low-risk, as the threshold passes from t_i to t_{i+1}, can correspond to a fair drop in estimated true-positive rate. Depending on how many controls similarly underwent a change in risk classification, and the weight ω in the net benefit measure, the relative utility curve could demonstrate a seemingly discrete increase or decrease. In the illustration, from a validation cohort of 79 cases and 721 controls, the confidence bands computed for thresholds between 10 and 95% at increments of 1.25% exhibit this phenomenon. It is an honest portrayal of the data.

However, this does indicate that in the region where corresponding benefit ratios are not generally realistic, the estimation of net benefit from sample sizes can be quite unstable. This is the basic difficulty of estimating a rare event. Further, variance estimates, whether analytic or bootstrapped, are not likely to perform well in these scenarios. This can be seen by considering the extreme scenario when the maximum risk score of participants is below

the highest risk threshold considered. In such a case, TPR, FPR, and sNB reach zero at some t_i and stay zero at all higher thresholds. All estimates of variance of any of these estimators, whether analytic or bootstrap based, will be zero. This prohibits good coverage properties of confidence bands that extend into extreme risk threshold regions. As was seen in the illustration, the calculated confidence did not cover the estimand, primarily because of the estimates for the rules using a thresholds greater than 90%.

Perfect Correlation

In regions where the empirical risk distribution has no mass, the estimated net benefit of the corresponding rules does not change and hence the estimated correlation between the performances is 1. This cause a numerical issue with sampling from a multivariate normal distribution with correlation matrix that is not positive-definite. This can be accounted for with careful coding, basically by sampling from lower dimension multivariate normal distribution and making the sampled deviation for the perfectly correlated rules equal. This issue can be alleviated somewhat by choosing the number of risk thresholds in the context of the range of interest and the sample size available.

4.8 Simulated Performance of CB_n

The most important frequentist property of a confidence band is that it achieves the designed level of coverage. Using the same super-population and approach as for our evaluation of empirical point estimators from independent cohort data, we evaluate the proposed confidence band construction for the relative utility curves of the two pre-specified models PredMod and $\text{PredMod}_{\text{ext}}$, as well as their difference. Relative utility curves, for thresholds ranging between 20% and 80%, were estimated and 95% confidence bands were calculated on 5,000 validation cohort samples that varied from 800 to 9,600 patients. The true relative utility curve was determined by calculating the relative utility curve of each pre-specified risk model on the entire super-population and by taking the difference between the two.

The observed coverage rates are presented in Figure 4.2, which displays coverages between

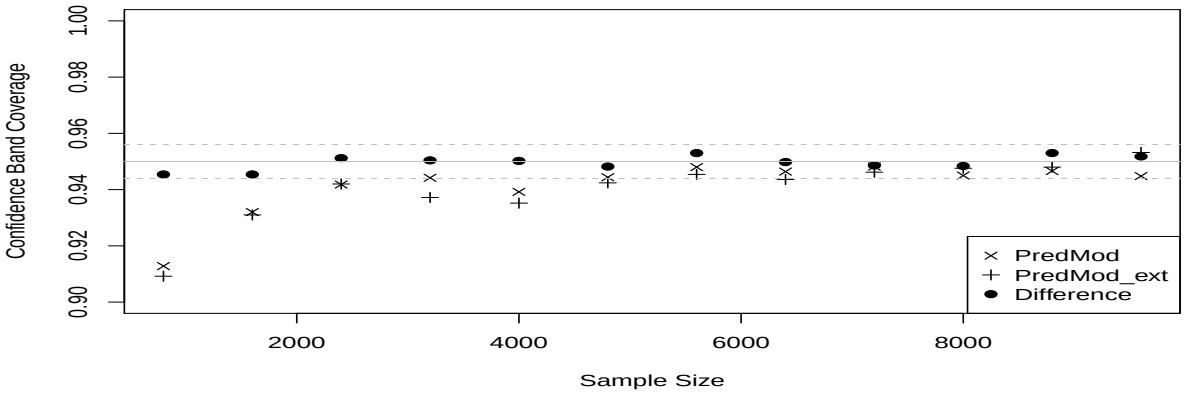


Figure 4.2: Coverage of confidence for $\text{sNB}_n(t)$ with $t \in [0.2, 0.8]$.

90 and 100%. The desired 95% confidence level is indicated with a line. Dashed lines indicate upper and lower limits for the central 95% of coverage estimates, based on 5,000 replicates, when the true coverage is 95%; we will refer to this range as the Monte Carlo range centered at 95%. Asymptotically the coverage of confidence bands around the relative utility curve for either model and their difference are on target. Overall, the coverage is quite good and falls within the range centered at 95% coverage. Coverage of the difference curve is within this range for all sample sizes considered. The confidence bands for individual relative utility curves demonstrate lowest coverage, 91% to 93%, at the smaller sample sizes of 800 and 1,600, but are well within the Monte Carlo range centered at 95% once samples contain at least 4,800 subjects. In the smaller samples sizes considered, coverage of the relative utility curve corresponding to the extended risk model is generally less than for the base model.

All of these trends are analogous to behavior exhibited by asymptotically correct Wald confidence intervals for previously examined point estimates. Good coverage for a proportion, such as a true-positive rate, requires larger sample sizes the more extreme the proportion and generally requires larger sample sizes than good coverage of a more complex estimand, such as net benefit. Similarly, estimation of net benefit requires larger sample sizes for good

coverage than a difference of net benefits. This is primarily accounted for by the Normal approximation becoming reasonable at different sizes depending on the relative skewness of small sample estimate distributions. The confidence band relies on a similar Gaussian approximation and exhibits the same characteristics.

4.9 Summary

In this chapter we introduced decision curves and relative utility curves and provided a means for constructing approximate confidence bands. We provide the formulas necessary to apply the algorithm in the context of single relative utility curves as well as the difference between two relative utility curves. We demonstrated the value of confidence bands in an example and assessed the coverage properties through a simulation study.

Chapter 5

ESTIMATION FROM UNMATCHED CASE-CONTROL STUDIES

In this chapter we extend previous results for conducting inference on empirical estimators of net benefit from cohort studies to empirical estimators from unmatched case-control studies. We first establish the empirical estimators and their distribution theory by calculating influence functions as adaptations of influence functions or estimators from cohort samples. We do this for three variants of unmatched case-control sampling schemes: a stand-alone case-control study, a stand-alone case-control study augmented with an external estimate of the outcome probability, and an outcome-dependent secondary sample of a two-phase study conducted on a cohort. The stand-alone case-control study, comprised of a sample of cases and a sample of controls, requires a known outcome probability and we examine the implications of this assumption. Design for case-control studies intended to assess the potential clinical utility of a decision rule is considered in two ways. First, we present the optimal control-to-case ratio for a fixed sample-size and the efficiency loss associated with non-optimal ratios. Second, the efficiency gain of additional controls in a sample consisting of all cases and a flexible number of controls is discussed. Performance of the proposed estimators is evaluated with a simulation study. Throughout this Chapter, we continue to assume that the clinical risk rule is pre-specified.

5.1 Biomarker Development Setting

Estimation from a cohort representing the population of interest was the focus of previous chapters and constitutes the gold-standard for validation of a rule and assessment of its potential clinical utility. However, a large-scale independent cohort study would only be

justified if results accrued from earlier stages of development consistently support the possible clinical utility of a biomarker. Some earlier studies could also have been cohort studies, to which previous results would support estimation and study design, but many may have been conducted on case-control samples.

Case-control studies are of central importance in early and mid-phase biomarker development research (Gu and Pepe, 2009; Baker et al., 2002; Pepe et al., 2001). As in epidemiological settings, case-control sampling provide statistical efficiency that enables use of smaller sample sizes and less expensive studies. Specific to biomarker development, case-control samples can potentially accelerate the iterations inherent in earlier stages of development. Case-control samples can also provide a level of control for characterizing the discriminatory properties of a biomarker (Pepe, 2003, p. 26). For example, discrimination of migration signature markers TFPI and TNC between pancreatic ductal adenocarcinoma cases and healthy controls can be studied alongside discrimination from chronic pancreatitis controls and from acute biliary obstruction controls (Balasenthil et al., 2017).

Repositories of biological samples are being created as a resource for biomarker development (see, for example, Rosenthal et al., 2015). Some examples arise as efforts to maximize the impact of clinical trials by making remaining biological samples available for secondary research efforts, as in Casson et al. (2011). In other examples, sample collection is conducted expressly for collaborative research projects (see, for example, Srivastava and Kramer, 2000). For mid-stages of biomarker research development, gold standard studies follow a PRoBE design, which entails prospective-specimen-collection and retrospective-blinded-evaluation (Pepe et al., 2008).

While stages of biomarker development prior to final validation may have primary objectives other than the direct assessment of clinical utility, best practices for biomarker research advocate for reflection of potential clinical application throughout all aspect of earlier-phase study design, including sampling population and performance measures (Pepe et al., 2016). Indeed, it is plausible that the disconnection from clinical application at early stages of biomarker research may be a contributing factor of the relatively low rate of translation into

medical practice. Evaluation of the potential net benefit at these stages could be considered in conjunction with other measures of discriminatory accuracy. This chapter supports inference for net benefit from unmatched case-control samples.

5.2 Inference for Net Benefit from Unmatched Case-Control Samples

We consider three variants of unmatched case-control sampling schemes: a stand-alone case-control study (*cc*), a stand-alone case-control study augmented with an estimate of the outcome probability from an external cohort (*cc-c*), and an outcome-dependent secondary sample in a two-phase study conducted on a cohort (*tp*). As for the empirical estimator from cohort data, empirical estimators from case-control sampling schemes studied here are regular and asymptotically linear. Our main approach is to derive the influence function of the empirical estimator, from which the asymptotic distribution theory follows from the Central Limit Theorem. Even in cases where the limiting variance can be obtained without the influence function, the form of these influence functions are useful for their role in calculating the limiting variance for contrasts of net benefit as well as in constructing confidence bands for decision curves.

The empirical estimators and their corresponding influence functions will be connected to their underlying sampling scheme using superscripts, such as sNB_n^{cohort} for the previously introduced estimator defined on a cohort sample, and IF^{tp} for the influence function of sNB_n^{tp}, the estimator defined on data collected by a two-phase sampling design. Table 5.1 presents formulas for both the limiting variance and influence function for the various empirical estimators of sNB^N from unmatched case-control samples. Analogous formulas and estimators of sNB^A are derived similarly.

5.2.1 Case-Control Sample With Known Prevalence

We first consider estimation of sNB from a simple case-control sample with fixed numbers of cases and controls. In this case, true- and false-positive rates can be empirically estimated from the representative samples of n_1 cases and of n_0 controls. Each observation is of the

form $O_i = (D_i, X_i)$, where $X_i = R(W_i)$, and in the case of a clinical decision rule derived from a risk model, $X_i = I\left[r(W_i) > r_T\right]$. Since net benefit depends directly on the outcome probability, which cannot be estimated from case-control data, we assume that the event rate in the population, ρ_0, is known. The resulting empirical case-control estimator for net benefit is:

$$\mathrm{sNB}^{cc,\rho}_{n_1,n_0} := \mathrm{TPR}_{n_1} - \omega_0\, \mathrm{FPR}_{n0}, \tag{5.1}$$

where TPR_{n_1} is the empirical estimator of the true-positive rate based on n_1 cases, FPR_{n_0} is the empirical estimator of the true-positive rate based on n_0 controls, and $\omega_0 = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_0}{\rho_0}$. The centered and scaled estimator of the true-positive rate, $\sqrt{n_1}\left(\mathrm{TPR}_{n_1} - \mathrm{TPR}_0\right)$, has limiting variance $\mathrm{TPR}_0(1 - \mathrm{TPR}_0)$. Similarly, the centered and scaled estimator of the false-positive rate has limiting variance $\mathrm{FPR}_0(1 - \mathrm{FPR}_0)$. Since the cases and the controls are distinct sets of observations, in large samples the variability of $\mathrm{sNB}^{cc,\rho}_{n_1,n_0}$ is approximately:

$$\frac{1}{n_1}\mathrm{TPR}_{n_1}\left\{1 - \mathrm{TPR}_{n_1}\right\} + \frac{1}{n_0}\omega_0^2 \mathrm{FPR}_{n_0}\left(1 - \mathrm{FPR}_{n_0}\right). \tag{5.2}$$

Formally, the data unit is a set of J_0 controls and J_1 cases, where J_0 and J_1 are the smallest integers for which $J_0/J_1 = n_0/n_1$, i.e., $O^{cc,\rho} = (O_{11}, \ldots, O_{1J_1}, O_{01}, \ldots, O_{0J_0})$, where D_i equals 0 for components O_{0j} and 1 for components O_{1j}. The number of data units m satisfies $mJ_1 = n_1$ and $mJ_0 = n_0$. The asymptotic linearity of $\mathrm{sNB}^{cc,\rho}_{n_1,n_0}$, for which the asymptotics are with respect to $m \to \infty$, can be established by direct calculation, as follows:

$$\begin{aligned}
\mathrm{sNB}^{cc,\rho}_{n_1,n_0} - \mathrm{sNB}_0 &= \left(\mathrm{TPR}_{n_1} - \mathrm{TPR}_0\right) - \omega_0\left(\mathrm{FPR}_{n_0} - \mathrm{FPR}_0\right) \\
&= \frac{1}{mJ_1}\sum_{i=1}^{m}\sum_{k=1}^{J_1}\left(x_{1k,i} - \mathrm{TPR}_0\right) - \omega_0 \frac{1}{mJ_0}\sum_{i=1}^{m}\sum_{k=1}^{J_0}\left(x_{0k,i} - \mathrm{FPR}_0\right) \\
&= \frac{1}{m}\sum_{i}\left[\frac{1}{J_1}\sum_{k=1}^{J_1}\left(x_{1k,i} - \mathrm{TPR}_0\right) - \omega_0 \frac{1}{J_0}\sum_{k=1}^{J_0}\left(x_{0k,i} - \mathrm{FPR}_0\right)\right],
\end{aligned}$$

where the subscripts $1k, i$ indicate the k^{th} case within the i^{th} data unit and similarly for control observations. This expression reveals that $\mathrm{sNB}^{cc,\rho}_{n_1,n_0}$ is a linear estimator, the remainder

term is exactly 0, as opposed to only $o_P(n^{-1/2})$, and its influence function is:

$$IF_0^{cc,\rho}(o^{cc,\rho}) := \frac{1}{J_1}\sum_{k=1}^{J_1}(x_{1k,i} - \text{TPR}_0) - \omega_0\frac{1}{J_0}\sum_{k=1}^{J_0}(x_{0k,i} - \text{FPR}_0)$$

and simplifying $\sigma_0^{2,cc,\rho} := \mathbb{E}\left[IF^{cc,\rho}(O^{cc,\rho})\right]^2$, yields:

$$\sigma_0^{2,cc,\rho} = \frac{1}{J_1}\text{TPR}_0\left(1 - \text{TPR}_0\right) + \frac{1}{J_0}\omega_0^2\text{FPR}_0\left(1 - \text{FPR}_0\right),$$

which agrees with Equation 5.2. The factors of $\frac{1}{J_1}$ and $\frac{1}{J_0}$ account for the numbers of respective cases or controls within each of the m data units. Using the plug-in estimator for the limiting variance, the variance of $\text{sNB}_{n_1,n_0}^{cc,\rho}$ is approximately equal to $\frac{1}{m}\sigma_m^{2,cc}$, which is the expression in Equation 5.2.

5.2.2 Case-Control Sample with Estimated Event Rate

In practice, the outcome probability is generally not known and an estimate, perhaps from a very large sample, is employed as if it were known. Here we explicitly take this detail into account. Consider the existence of a cohort, on which the outcome D has been measured, that is independent from a stand-alone unmatched case-control sample on which the predictors W have been measured and the pre-specified decision rule $X = R(W)$ has been applied . Empirical estimators of the constituents of net benefit, ρ_0, TPR_0, and FPR_0, can each be constructed from the external cohort of size n_c and the samples of n_1 cases and n_0 controls, respectively. Together, they define an estimator of net benefit:

$$\text{sNB}_{n_1,n_0,n_c}^{cc-c} = \text{TPR}_{n_1} - \omega_{n_c}\,\text{FPR}_{n0}, \tag{5.3}$$

where TPR_{n_1} and FPR_{n_0} are the usual empirical estimators, and ρ_{n_c} is the empirical estimator of the event rate based on the external cohort, which appears in $\omega_{n_c} = \frac{B^{ctrl}}{B^{case}}\frac{1-\rho_{n_c}}{\rho_{n_c}}$.

We have previously observed that the three components of $\sigma_0^{2,cohort}$ (Equation 3.6) correspond to the uncertainty introduced by estimating each constituent. Since all three constituent estimators of $\text{sNB}_{n_1,n_0,n_c}^{cc-c}$ are based on distinct sets of observations, its approximate

variance can be empirically estimated by:

$$\frac{1}{n_1}\text{TPR}_{n_1}\left(1-\text{TPR}_{n_1}\right)+\frac{1}{n_0}\omega_{n_c}^2\text{FPR}_{n_0}\left(1-\text{FPR}_{n_0}\right)+\frac{1}{n_c}\frac{\omega_{n_c}^2}{(1-\rho_{n_c})\rho_{n_c}}\text{FPR}_{n_0}^2. \tag{5.4}$$

Analogous to estimation from a stand-alone case-control sample, the formal data unit is a set of J_1 cases, J_0 controls, and J_c observations from a cohort, which can be written: $O^{cc-c}=(O_{11},\ldots,O_{1J_1},O_{01},\ldots,O_{0J_0},O_1,\ldots,O_{cJ_c})$, where D_i equals 0 for components O_{0j}, 1 for components O_{1j} and is random for components O_{cj}; W is random for components O_{0j} and O_{1j}, and not recorded in O_{cj}. The number of data units m satisfies: $n_1=mJ_1$, $n_0=mJ_0$, and $n_c=mJ_c$. The asymptotic linearity of $\text{sNB}_{n_1,n_0,n_c}^{cc-c}$, for which the asymptotics are with respect to $m\to\infty$, can be expressed as follows:

$$\begin{aligned}\text{sNB}_n^{cc-c}-\text{sNB}_0 &= \left(\text{TPR}_{n_1}-\text{TPR}_0\right)-\omega_0\left(\text{FPR}_{n_0}-\text{FPR}_0\right)+\left(\omega_{n_c}-\omega_0\right)\text{FPR}_0+o_P(n^{-1/2})\\ &= \frac{1}{m}\sum_k IF_0^{cc-c}\left(O_i^{cc-c}\right)+o_P(n^{-1/2}),\end{aligned}$$

where the influence function is given by

$$IF_0^{cc-c}(o^{cc-c}):=\frac{1}{J_1}\sum_{k=1}^{J_1}\left(x_{1k}-\text{TPR}_0\right)-\omega_0\frac{1}{J_0}\sum_{k=1}^{J_0}\left(x_{0k}-\text{FPR}_0\right)+\omega_0\text{FPR}_0\frac{1}{J_c}\sum_{k=1}^{J_c}\frac{d_k-\rho_0}{\rho_0(1-\rho_0)}.$$

The asymptotically negligible term in the first line equals $-\left(\omega_{n_c}-\omega_0\right)\left(\text{FPR}_{n_0}-\text{FPR}_0\right)$, and in the second line, follows from application of the Continuous Mapping Theorem and Slutsky's Theorem. Simplifying the limiting variance, defined by $\sigma_0^{2,cc-c}:=\mathbb{E}\left[IF^{cc-c}(O^{cc-c})\right]^2$, yields:

$$\sigma_0^{2,cc-c}=\frac{1}{J_1}\text{TPR}_0\left(1-\text{TPR}_0\right)+\frac{1}{J_0}\omega_0^2\text{FPR}_0\left(1-\text{FPR}_0\right)+\frac{1}{J_c}\frac{\omega_0^2}{(1-\rho_0)\rho_0}\text{FPR}_0^2$$

which agrees with the claim of Equation 5.2. The factors $\frac{1}{J_1}$, $\frac{1}{J_0}$ and $\frac{1}{J_c}$ account for the numbers of respective cases, controls or cohort observations within each of the m data units.

These results describe the asymptotics of estimators from a case-control sample nested within a cohort, after adjustment of the denominators to reflect the relationship between data units and respective numbers of cases, controls, and observations contributing to estimation of the outcome probability. Nested case-control samples are an example of data missing by design and are appropriately treated in a missing data framework, which we consider next.

5.2.3 Two-Phase Samples

Two-phase sampling schemes, where phase-two sampling probabilities depend only on outcome, are analogous to unmatched case-control designs and have similar empirical estimators of sNB^N. We will consider two variants on the routine clinical covariates measured at phase one: (1) none, all variables other than clinical outcome are only available on the phase-two sample; (2) some clinical predictors, of finitely many values, are measured during phase one and additional biomarkers are measured in the second phase. We will also comment on a third scenario in which phase-one clinical variables are continuous.

The following developments assume Bernoulli sampling, i.e., the indicators of selection into phase two are independent binary rand variables with probability of success $\pi(d)$ conditionally upon outcome d. We assume that phase-two measurements are missing at random; the probability of being in the subsample is independent of the unmeasured biomarkers. One consequence is that the number of cases in a sample of size n, selected for phase-two measurement, will vary from sample to sample, and averages $n\pi(1)\rho_0$, and similarly for the number of controls. We denote measurement at phase two with the indicator variable Δ and the phase-two sampling probabilities by $\pi(d) = P(\Delta = 1 \mid D = d)$, which are known by design.

Nested case-control studies are a particular example of more general two-phase sampling design. The discrepancy between independent Bernoulli sampling and the dependence introduced by sampling fixed numbers of cases and controls from a finite cohort has been rigorously established to be asymptotically negligible by Ma (2010).

No predictors measured at Phase-I

We start by considering the scenario of evaluating clinical decision rules with predictors only measured in phase-II. The experimental data has the form $O^{tp,simple} = (\Delta W_2, \Delta, D) \sim P_0^{tp,simple}$, where D is measured on all subjects in phase one, and the phase-two indicator, Δ, equals 1 on the subsample of patients for whom covariates W_2 are measured and the clinical

decision rule evaluated.

In this special case where the sampling probabilities and the decision rule do not rely on any covariates available on the entire cohort, the usual empirical estimator

$$\text{sNB}_n^{tp,simple} := \text{TPR}_{n1} - \omega_n \text{FPR}_{n0},$$

is unbiased when the usual empirical estimators for the rates are constructed from the n_1 cases and n_0 controls measured at phase-II and the full cohort of size n. In general, the influence function for $\text{sNB}_n^{tp,simple}$ can be obtained from the influence function of the cohort estimator, sNB_n^N (Rose and van der Laan, 2011b). In this scenario, the result simplifies to:

$$\begin{aligned} IF_0^{tp,simple}(o^{tp,simple}) &= \frac{\delta}{\pi(d)} \left\{ \frac{d}{\rho_0}(x - \text{TPR}_0) - \omega_0 \frac{1-d}{1-\rho_0}(x - \text{FPR}_0) \right\} \\ &\quad + \frac{\omega_0}{(1-\rho_0)\rho_0} \text{FPR}_0 (d - \rho_0), \end{aligned}$$

where $x = R(w_2)$. As usual, the limiting variance equals the variance of an observation transformed by the influence function, which simplifies to:

$$\sigma_0^{2,tp,simple} = \frac{1}{\pi(1)\rho_0} \text{TPR}_0 (1 - \text{TPR}_0) + \frac{\omega_0^2}{\pi(0)(1-\rho_0)} \text{FPR}_0 (1 - \text{FPR}_0) + \frac{\omega_0^2}{\rho_0(1-\rho_0)} \text{FPR}_0^2$$

We note that the denominators $\pi(1)\rho_0$ and $\pi(0)(1-\rho_0)$ act to convert the total phase-one sample size into the expected number of phase-two sampled cases and the expected number of phase-two sampled controls, respectively. The entire phase-one cohort contributes to the estimation of the event rate and the third component of the limiting variance is the same as that for empirical estimators from data available on the full cohort.

Some predictors measured at phase two, categorical predictors measured at phase one

We now consider the scenario of evaluating clinical decision rules with some predictors measured in phase one, such as routine clinical variables, and some predictors measured in phase two, such as a novel biomarker. We assume that the routine clinical variables define finitely many strata. The observed data unit has the form $O^{tp} = (\Delta W_2, \Delta, W_1, D) \sim P_0^{tp}$, where D and W_1 are measured on all subjects in phase one, and the phase-two indicator, Δ, equals 1

on the subsample for whom additional covariates W_2 are measured. The clinical decision rule, $x = R(w_1, w_2)$, can only be evaluated for subjects with Phase-II measurements. Phase-two sampling probabilities are still only dependent on outcome.

We note that the simple two-phase estimator, $\text{sNB}_n^{tp,simple}$ is still an asymptotically linear estimator of sNB in the context of this more complex two-phase structure; it simply does not utilize the covariate information W_1 available on subjects not measured in the second phase. As we will see, in cases where the phase-one predictors are associated with the intervention decision, the simple estimator is not efficient. Working somewhat backwards, we will first calculate the efficient influence function for estimators of sNB that are efficient under this sampling scheme, and then introduce an efficient estimator.

Again, we apply the general result of Rose and van der Laan (2011b) to obtain the efficient influence function for sNB_n^{tp} from the influence function of the cohort estimator, sNB_n^{cohort}. In this scenario, the result simplifies to:

$$\begin{aligned} IF_0^{tp}(o^{tp}) &= \frac{\delta}{\pi(d)}\left[\frac{d}{\rho_0}\left\{R\left(w_1,w_2\right)-\text{TPR}_0(w_1)\right\}-\omega_0\frac{1-d}{1-\rho_0}\left\{R\left(w_1,w_2\right)-\text{FPR}_0(w_1)\right\}\right] \\ &+\frac{d}{\rho_0}\left\{\text{TPR}_0(w_1)-\text{TPR}_0\right\}-\omega_0\frac{1-d}{1-\rho_0}\left\{\text{FPR}_0(w_1)-\text{FPR}_0\right\}+\frac{\omega_0}{\rho_0(1-\rho_0)}\text{FPR}_0(d-\rho_0), \end{aligned}$$

where $x = R\left(w_1, w_2\right)$ is only available on observations sampled in phase two and $\text{TPR}_0(w_1)$ and $\text{FPR}_0(w_1)$ are the strata-specific true- and false-positive rates:

$$\text{TPR}_0(w_1) := \mathbb{E}_0\left[R(W_1,W_2) \mid W_1 = w_1, D = 1\right]$$

and

$$\text{FPR}_0(w_1) := \mathbb{E}_0\left[R(W_1,W_2) \mid W_1 = w_1, D = 1\right].$$

An influence function lies in $L_2^0(P_0)$ and in particular has mean equal to zero. Thus, we can employ the above influence function as an estimating equation which justifies defining a two-phase estimator of net benefit as:

$$\text{sNB}_n^{tp} := \text{TPR}_n^{tp} - \omega_n \text{FPR}_n^{tp},$$

where $\omega_n = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_n}{\rho_n}$ and ρ_n is the usual empirical estimator based on the full set of phase-one observations; the two-phase estimator of true-positive rate has the form:

$$\mathrm{TPR}_n^{tp} := \frac{1}{\sum_i D_i} \sum_i D_i \mathrm{TPR}_n^{pII}(W_1),$$

where

$$\mathrm{TPR}_n^{pII}(w_1) := \frac{1}{\sum_i \Delta_i D_i I[W_i = w_i]} \sum_i \Delta_i D_i I[W_i = w_i] X_i.$$

In words, the strata-specific true-positive rates, $\mathrm{TPR}^{pII}(w_1)$, are estimated empirically within each phase-one strata using all cases sampled at phase two. The estimate of the marginal true-positive rate is then obtained by averaging the strata-specific estimates over the empirical phase-one covariate distribution observed among all cases. The two-phase estimator of the false-positive rate is defined analogously using controls.

The limiting variance of sNB_n^{tp}, suitably centered and scaled, equals the variance of an observation transformed by the influence function, which simplifies to:

$$\begin{aligned}
\sigma_0^{2,tp} &= \frac{1}{\rho_0}\mathbb{E}_0\left[\frac{1}{\pi(1)}\mathrm{TPR}_0(W_1)\left(1-\mathrm{TPR}_0(W_1)\right) + \left\{\mathrm{TPR}_0(W_1) - \mathrm{TPR}_0\right\}^2 \mid D=1\right] \\
&\quad + \frac{1}{1-\rho_0}\omega_0^2\mathbb{E}\left[\frac{1}{\pi(0)}\mathrm{FPR}_0(W_1)\left(1-\mathrm{FPR}_0(W_1)\right) + \left\{\mathrm{FPR}_0(W_1) - \mathrm{FPR}_0\right\}^2 \mid D=0\right] \\
&\quad + \frac{\omega_0^2}{(1-\rho_0)\rho_0}\mathrm{FPR}_0^2
\end{aligned}$$

Using this expression, we can examine our previous claim regarding the efficiency difference between sNB_n^{tp} and $\mathrm{sNB}_n^{tp,simple}$. The difference of the two limiting variances, $\sigma_0^{2,tp,simple} - \sigma_0^{2,tp}$, equals:

$$\frac{1}{\rho_0}\left(\frac{1}{\pi(1)} - 1\right)\mathrm{VAR}_0\left\{\mathrm{TPR}_0\left(W_1\right) \mid D=1\right\} + \frac{\omega_0^2}{1-\rho_0}\left(\frac{1}{\pi(0)} - 1\right)\mathrm{VAR}_0\left\{\mathrm{FPR}_0\left(W_1\right) \mid D=0\right\}.$$

As expected, the difference is zero when all cases and all controls are measured in phase two, because both estimators reduce to the cohort estimator. The difference is also zero when the strata-specific true- and false-positive rates do not vary, which is also expected as this describes predictors that do not contribute to the discrimination between would-be cases

Sampling	$\mathcal{M}$odel	Variance Bound and Efficient Influence Function
Cohort	NP	$\sigma^2 = \frac{1}{\rho}\text{TPR}\,(1-\text{TPR}) + \omega^2\frac{1}{1-\rho}\text{FPR}\,(1-\text{FPR}) + \frac{\omega^2}{(1-\rho)\rho}\text{FPR}^2$
		$IF(o) = \frac{d}{\rho}\,(x-\text{TPR}) - \omega\frac{1-d}{1-\rho}\,(x-\text{FPR}) + \frac{\omega}{(1-\rho)\rho}\text{FPR}\,(d-\rho)$
	ρ known	$\sigma^2 = \frac{1}{\rho}\text{TPR}\,(1-\text{TPR}) + \omega^2\frac{1}{1-\rho}\text{FPR}\,(1-\text{FPR})$
		$IF(o) = \frac{d}{\rho}\,(x-\text{TPR}) - \omega\frac{1-d}{1-\rho}\,(x-\text{FPR})$
	The data unit is $o = (x, d)$.	
C-C	ρ known	$\sigma^2 = \frac{1}{J_1}\text{TPR}\,(1-\text{TPR}) + \frac{1}{J_0}\omega^2\text{FPR}\,(1-\text{FPR})$
		$IF(o) = \frac{1}{J_1}\sum_{k=1}^{J_1}(x_{1k,i} - \text{TPR}_0) - \omega_0\frac{1}{J_0}\sum_{k=1}^{J_0}(x_{0k,i} - \text{FPR}_0)$
	The data unit $o = (x_{11}, \ldots, x_{1J_1}, x_{01}, \ldots, x_{0J_0})$ is a set of J_1 cases and J_0 controls.	
C-C with External Cohort		$\sigma^2 = \frac{1}{J_1}\text{TPR}\,(1-\text{TPR}) + \frac{1}{J_0}\omega^2\text{FPR}\,(1-\text{FPR}) + \frac{1}{J_c}\frac{\omega^2}{(1-\rho)\rho}\text{FPR}^2$
		$IF(o) = \frac{1}{J_1}\sum_{k=1}^{J_1}(x_{1k} - \text{TPR}_0) - \omega_0\frac{1}{J_0}\sum_{k=1}^{J_0}(x_{0k} - \text{FPR}_0) + \omega_0\text{FPR}_0\frac{1}{J_c}\sum_{k=1}^{J_c}\frac{d_k-\rho_0}{\rho_0(1-\rho_0)}$
	The data unit $o = (x_{11}, \ldots, x_{1J_1}, x_{01}, \ldots, x_{0J_0}, d_1, \ldots, d_{J_c})$ is a set of J_1 cases, J_0 controls and J_c cohort outcomes.	
Two-phase Phase-II Predictors	$\pi(d)$ known	$\sigma^2 = \frac{1}{\pi(1)}\frac{1}{\rho}\text{TPR}\,(1-\text{TPR}) + \frac{1}{\pi(0)}\frac{1}{1-\rho}\omega^2\text{FPR}\,(1-\text{FPR}) + \frac{\omega^2}{(1-\rho)\rho}\text{FPR}^2$
		$IF(o) = \frac{\delta}{\pi(1)}\frac{d}{\rho}\{x-\text{TPR}\} - \frac{\delta}{\pi(0)}\omega\frac{1-d}{1-\rho}\{x-\text{FPR}\} + \frac{\omega}{(1-\rho)\rho}\text{FPR}\,(d-\rho)$
	The data unit is $o = (\delta x, \delta, d)$; δ indicates phase-II measurement; $\pi(1), \pi(0)$ are sampling probabilities.	
Phase-I & II Predictors	$\pi(d)$ known	$\sigma^2 = \frac{1}{\rho}\mathbb{E}\left[\frac{1}{\pi(1)}\text{TPR}(W_1)\,(1-\text{TPR}(W_1)) + \{\text{TPR}(W_1)-\text{TPR}\}^2 \mid D=1\right] + \frac{\omega^2}{(1-\rho)\rho}\text{FPR}^2$
		$+\frac{1}{1-\rho}\omega^2\mathbb{E}\left[\frac{1}{\pi(0)}\text{FPR}(W_1)\,(1-\text{FPR}(W_1)) + \{\text{FPR}(W_1)-\text{FPR}\}^2 \mid D=0\right]$
		$IF(o) = \frac{\delta}{\pi(d)}\left[\frac{d}{\rho}\{R(w_1,w_2)-\text{TPR}(w_1)\} - \omega\frac{1-d}{1-\rho}\{R(w_1,w_2)-\text{FPR}(w_1)\}\right]$
		$+\frac{d}{\rho}\{\text{TPR}(w_1)-\text{TPR}\} - \omega\frac{1-d}{1-\rho}\{\text{FPR}(w_1)-\text{FPR}\} + \frac{\omega}{\rho(1-\rho)}\text{FPR}(d-\rho)$
	The data unit is $o = (\delta w_2, \delta, w_1, d)$; $\text{TPR}(w_1)$ and $\text{FPR}(w_1)$ are strata-specific positive rates.	

Table 5.1: Asymptotic variance for empirical estimators of sNB^N under various unmatched sampling scenarios. These variances are conditional on prespecified clinical decision rule, R, which decides $X = R(W)$. In all scenarios, the weight $\omega = \frac{B^{ctrl}}{B^{case}}\frac{1-\rho}{\rho}$ and $\frac{J_0}{J_1}$ is the control-to-case ratio. Details are discussed in Section 5.2.

and would-be controls. Otherwise, this value is strictly positive and describes the absolute reduction in variability due to leveraging the distribution of predictors observed on the whole phase-one cohort.

Some predictors measured at phase two, continuous predictors measured at phase one

We continue the scenario of the previous section, in which second phase sampling probabilities are only a function of the outcome, and some predictors of the clinical decision rule are measured at phase one. In the previous section, the phase-one predictors were assumed to have finitely many values. When this is not the case, as for one or more continuous phase-one predictors, the strata-specific true- and false-positive rates, as previously defined, are no longer determined by a finite set of parameters than can be estimated empirically. Instead, a function with continuous support must be estimated. Though well-behaved empirical estimators are not available, non-parametric approaches can be employed. These approaches generally involve local smoothing, which introduces bias into the estimator and requires correction for valid inference. These techniques are beyond the scope of this part of the dissertation. We note that estimating sNB in this context could be addressed by the approach employed in the second part of the dissertation, where non-parametric approaches to estimating population-averaged conditional means is studied.

5.3 Optimal Control-to-Case Ratio

Some study constraints place strict limits on the size of the case-control sample. This can result from considerations such as the expense of assays used to measure novel biomarkers or from conservative approaches to distributing non-replenishable biological samples, that have been prospectively collected for biomarker research, across multiple studies.

For a fixed case-control sample size, in which the proportion of cases is denoted by π_1 and the proportion of controls by π_0, there is a control-to-case ratio $J = J_0/J_1$ that minimizes the variance of an empirical estimator of sNB. The limiting variances for empirical estimators

from the various case-control designs considered, all have the form:

$$\sigma_0^2 = c_1 \frac{1}{J_1} TPR_0 \left(1 - TPR_0\right) + c_1 \frac{1}{J_0} \omega_0^2 FPR_0 \left(1 - FPR_0\right) + c_2, \tag{5.5}$$

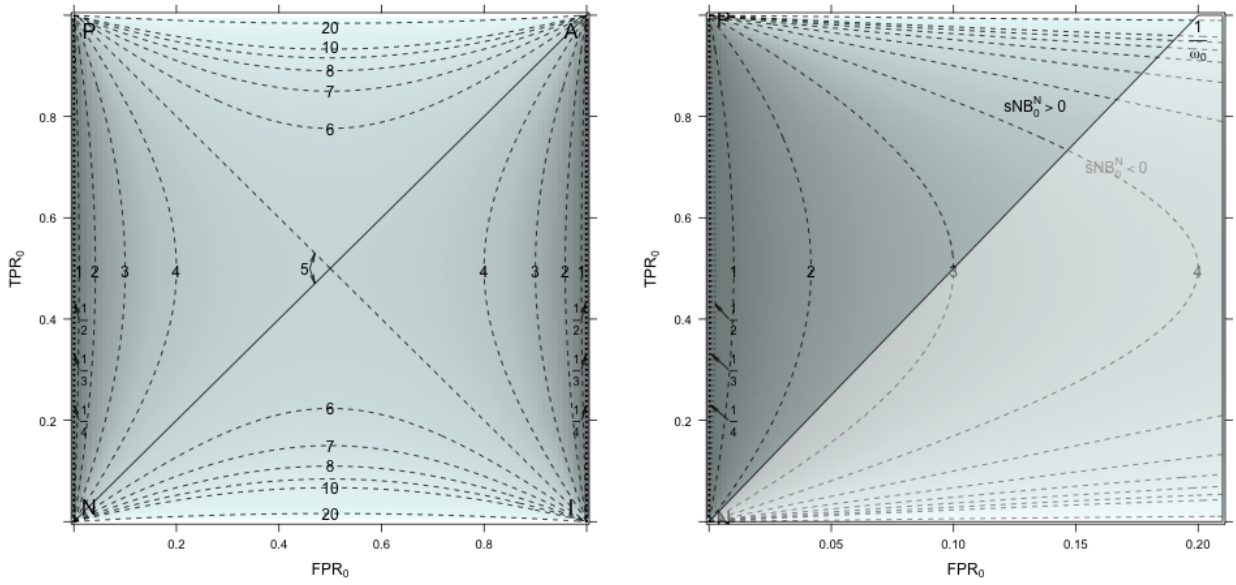
where the constants c_1 and c_2 vary according to which case-control sampling design has been adopted. For example, $c_1 = 1$ and $c_2 = 0$ for estimators from a stand-alone case-control sample. Both constants are independent of the control-to-case ratio and thus the ratio minimizing the variance is:

$$J^{opt} = \omega_0 \sqrt{\frac{FPR_0(1 - FPR_0)}{TPR_0(1 - TPR_0)}}. \tag{5.6}$$

This is optimal for all unmatched case-control sampling designs examined thus far. Within a clinical context, the optimal number of controls per case is driven by the relative variance of estimators of the positive rate among controls (false-positive) and the positive rate among cases (true-positive). The benefit ratio of the clinical context, ω_0, further weights the optimal ratio in accordance with the population-level benefits for treating controls correctly, relative to treating cases correctly. Estimation of TPR must be ω^2 times more variable than that of FPR, before the optimal ratio requires sampling more cases than controls. As the weight increases, this requirement becomes satisfied by fewer true- and false-positive rate combinations. In this sense, the optimal ratio for estimating the opt-in formulation of net benefit generally favors enrichment, relative to a 1:1 ratio, by controls.

Optimal Ratio as a Function of Classification Accuracies, Within a Cinical Context

Within a clinical context $C = \left(\rho_0, B^{ctrl}{:}B^{case}\right)$, the optimal ratio, like the measure of net benefit itself, is determined by the weight ω_0; any two clinical contexts yielding the same weight ω_0 have the same optimal ratio, when viewed as a function of the true and false-positive rates of possible rules $R = (r, r_T)$. Figure 5.1(a) represents the optimal control-to-case ratio, as a function of true- and false-positive rates, for clinical contexts in which ω_0 equals 5. Shading is darkest for rule characteristics optimized by the smallest control-to-case ratios and gradually becomes white over rule characteristics for which the optimal ratio is



(a) (b)

Figure 5.1: (a) Optimal control-to-case ratios, for estimating sNB^N, over possible accuracy characteristics of a rule, for clinical contexts with $\omega_0 = 5$. Dashed lines indicate contours of J^{opt}. (b) Same as (a), restricted to $\text{FPR}_0 \leq 0.2$, which focuses on classification accuracies yielding positive net benefit (region for which $\text{sNB}_0 < 0$ is faded out).

largest. Level sets for optimal ratios from 1-4 cases per control and from 1-20 controls per case are indicated by dashed lines.

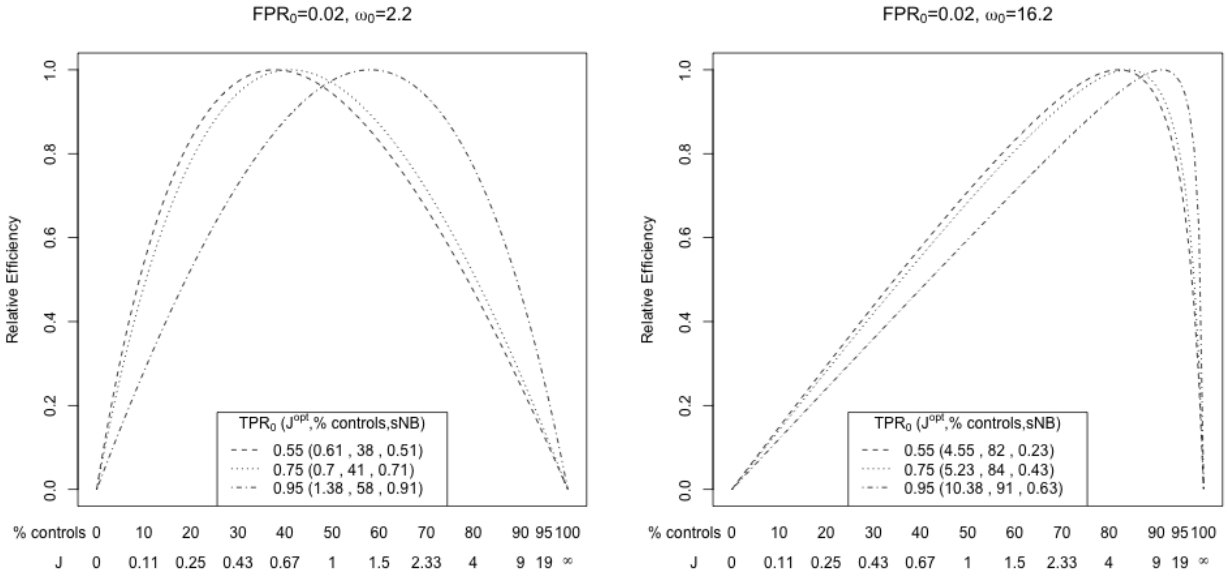
Algebraic manipulation of Equation 5.6 reveals that test characteristics with the same optimal ratio lie on hyperbolas with asymptotes equal to $TPR_0 = FPR_0$ or $TPR_0 = 1 - FPR_0$. For rules represented by points on the asymptotes the optimal ratio is ω_0, which is at least 1 when treating no-one is a rational decision and equals 5 in the figure. Characteristics for which more than ω_0 controls per case are favored lie in the upper and lower regions between asymptotes and the ratio increases moving towards more extreme true-positive rates. Characteristics for which less than ω_0 controls per case are favored lie in the left and right regions between asymptotes and the ratio decreases moving towards more extreme

false-positive rates.

In practice, only biomarkers contributing to rules with some anticipated net benefit would be pursued. Figure 5.1(b) focuses on the region of clinical interest, where $\text{sNB}^N > 0$. The optimal ratio can be quite sensitive to the underlying true- and false-positive rates, particularly when the false-positive rate is very small or the true-positive rate is very large. Distinct rules with the same net benefit can have performances for which estimation is optimized by drastically different ratios. For larger weights ω_0, the crescents at the extreme false-positive rates, regions for which the optimal ratio favors enrichment of cases, are smaller. As a rule of thumb, the greater the population-level benefit of correctly not treating all controls relative to the population-level benefit of correctly treating all cases (ω_0), the more likely variability will be reduced by enriching for controls; only promising rules with exceptionally small false-positive rates will benefit from evaluation on case-enriched samples.

Efficiency Loss for Suboptimal Ratios

Figure 5.2 shows the efficiency loss associated with sub-optimal case-control allocation for evaluating the net benefit of a decision rule in clinical contexts with ω_0 equal 2.2 (a) and 16.2 (b). These weights would apply, for example, when evaluating an intervention for a rare ($\rho = 0.03$) outcome that has a 1:15 or a 1:2 cost-benefit trade-off. Scenarios in which the false-positive rate is 2% and the true-positive rates are 55, 75, and 95% are examined. The relative efficiency equals 1 when the number of controls in the sample, as a percentage, corresponds to the optimal ratio and decreases monotonically to zero at 0% or 100% controls. The relative efficiency loss relies on both the population level trade-offs ω_0 as well as the underlying true- and false-positive rates of the clinical rule. The qualitative features of the plots appear to be driven primarily by the optimal ratio, which is in turn strongly driven by ω_0. Figure 5.2(a) illustrates scenarios in which the optimal ratio is near 1:1 and the relative efficiency curves are fairly rounded and symmetric. On the other hand, the optimal design for scenarios illustrated in Figure 5.2(b) require many controls per case; the curves are highly skewed and the efficiency loss associated with samples comprised of even greater portions of



(a) (b)

Figure 5.2: Efficiency of stand-alone case-control estimators across different control-to-case ratios (J), and corresponding number of controls (as a percentage), relative to the optimal ratio in clinical contexts with $\omega_0 = 2.2$ (a) and $\omega_0 = 16.2$ (b).

controls is dramatic.

The relative efficiencies examined in Figure 5.2 were calculated for estimation from stand-alone case-control samples. For estimators from case-control samples augmented with an independent cohort sample and estimators from two-phase samples, the component of variability due to the unknown event rate is non-zero and is not affected by the control-to case ratio. Hence, in any given scenario, the relative efficiency loss for suboptimal case-control allocation will be less in these studies when compared to a stand-alone case-control study. Since the optimal ratio is the same for each, the absolute efficiency loss due to sub-optimal case-control allocation will be the equal across all three study designs. Calculations and additional figures are in Appendix B.8.

Implication for Study Design

Consider evaluating a rule that has 2% FPR and 95% TPR using a routine 1:1 case-control sample. In the context of Figure 5.2(a), where the optimal design requires samples containing 58% controls, estimation based on samples with the sub-optimal 50% portion would lose roughly 2.5% efficiency. Estimation from samples with a 64% control portion, 8% more than the optimal 58%, would lose a similar amount of efficiency (2.7%). However, if the same rule were being evaluated in the clinical context of Figure 5.2(b), estimation from samples with 50% controls instead of the optimal 91% would result in losses of almost 40.5%. This is a direct consequence of the design being far from optimal. Considering designs that are 8% above or below optimal, estimation from samples with 83% or 99% composition by controls, would yield 4.6% and 38% loss in efficiency, which directly reflects the asymmetry of the relative efficiency curve. When the optimal design, based on assumed values for the event rate and rule performance, requires many more controls than cases, this asymmetry should be taken into account. Cautious approaches in setting the design ratio would favor lower ratios among those corresponding to anticipated ranges of the unknown parameters.

Connection with ROC Curves

A clinical decision rule based on a risk model has performance characteristics that corresponds to one point on the model's ROC curve; namely, the point corresponding to the threshold that defines the rule: $R(W) = I[r(W) > r_T]$. Baker et al. (2009), citing Metz (1978), have pointed out that the tangent line to the ROC curve at the point corresponding to the rational risk threshold, has slope equal to the weight ω_0. Hence, the optimal control-to-case ratio for estimating the standardized net benefit of a rational rule, stated in Equation 5.6, has the same form as the optimal ratio for estimating the true-positive rate corresponding to a fixed false-positive rate, as established by Janes and Pepe (2006). The role of false-positive rate that is fixed in the latter scenario, is played by the unknown false-positive rate corresponding to the rational risk rule in the net benefit scenario. For

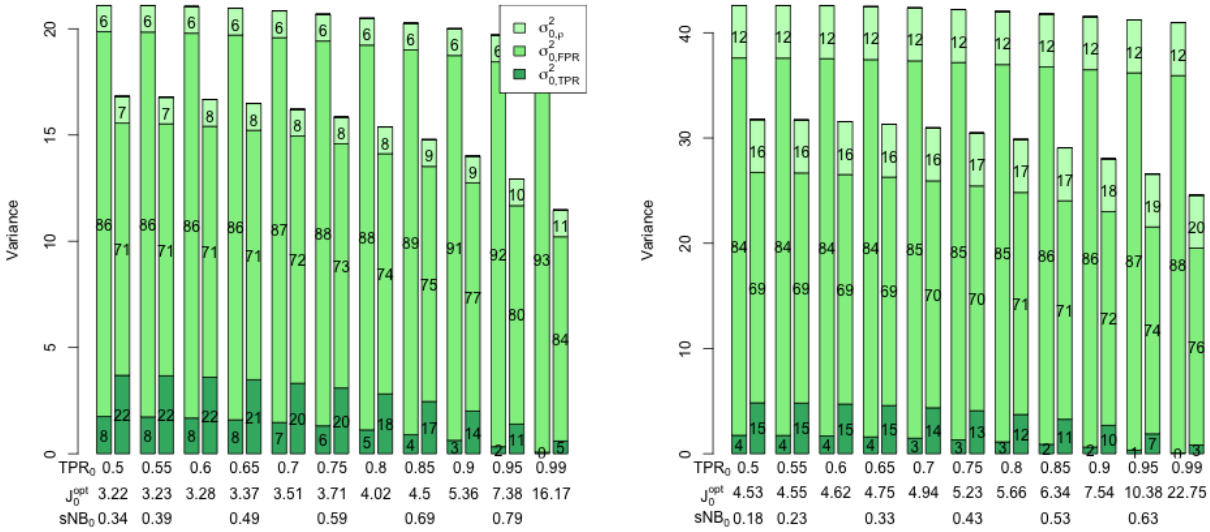
estimation of sNB^N, we view the role of ω_0 in J^{opt} as directly reflecting the population-level importance of controls over cases, rather than indirectly valuing controls through the threshold that achieves the fixed false-positive rate. Either way, the conclusion is the same, steepness of the ROC curve at the relevant point drives the need to sample more controls. The distinctions between the optimal ratio for estimating the ROC curve at a point and that for estimating the AUC, as observed by Janes and Pepe (2006), also apply to distinguishing optimal design of case-control studies for estimation of net benefit from that of AUC.

5.4 Implications of Assuming the Outcome Probability is Known

The limiting variance for empirical estimators of net benefit from stand-alone case-control samples is inherited from the limiting variance for empirical estimators from cohort data. In practice, exact knowledge of a feature of the data-generating mechanism, such as the event rate, is exceedingly rare. Typically, such assumptions are only approximately true, and in practice, an estimate is employed as if it were the true value. The variability decomposition for the empirical cohort estimator of net benefit, discussed in Section 3.5, should give one pause in doing so. The associated efficiency gain, when this assumption is true, is potentially an unaccounted source of variability when the assumption does not hold. Here, we examine the decomposition of the limiting variance for case-control studies nested in a cohort, where $\sigma^{2,tp}_{\text{sNB}} = \sigma^{2,tp}_{\text{T}} + \sigma^{2,tp}_{\text{F}} + \sigma^{2,tp}_{\text{P}}$.

The scenarios depicted in Figure 3.1 have a 3% outcome probability, which corresponds to roughly 32 controls per case in the cohort. Estimation from a 1:1 mix represents a significant shift in the relative case-control proportions of the sample and in the total variability due to estimation of the true- and false-positive rates from a case-control sample compared with a cohort. Estimation based on an optimal ratio of controls to cases will generally be in between, excepting the circumstances leading to enrichment by cases being optimal. Figure 5.3 illustrates the decomposition of variability for estimators of net benefit, under the same scenarios as in Figure 3.1, from case-control samples augmented with estimates of the event rate from nested designs. These plots show a pair of bars for each true-positive rate

corresponding to estimation from a routine 1-1 case-control subsample as well as a case-control subsample with optimal ratio. The limiting variances have been calculated for a case-control sub-sample of size 200 nested in a cohort of 1,000 subjects and apply equally to estimators from nested or two-phase case-control samples with the same relative sample sizes. For additional context, the corresponding value of standardized net benefit and the optimal ratio are labeled on the lower axis.



(a) (b)

Figure 5.3: Decomposition of the limiting variance, $\sigma_0^{2,ccc}$, for a clinical context defined by $\rho_0 = 0.03$, B^{ctrl}:B^{case} ($\omega_0 = 16.2$), when the rule has a 1% (a) or 2% (b) false-positive rate, across true-positive rates. Components are labeled with the corresponding percentage of the total. For each true-positive rate, routine 1-1 case-control sampling (left) and sampling with the optimal ratio (right) are represented.

Immediately noticeable is the efficiency gain achieved by using the optimal ratio, demonstrated by overall shorter bars. In the scenarios illustrated, going from a 1:1 design to one using J^{opt} produces absolute and relative increases in the variability component due to esti-

mation of the true-positive rate, whereas the contribution due to estimating the false-positive rate decreases. The control-to-case ratio parameter does not impact the estimation of the event rate, and the absolute contributions from estimating ρ_0, $\sigma_P^{2,tp}$, is the same across designs. However, $\sigma_{0,R}^{2,tp}$ represents a larger portion of the more efficient optimal design, e.g., when TPR=95%, $\sigma_{0,R}^{2,tp}$ equals 12% or 19% of the total depending on false-positive rates of 1 or 2%, respectively.

The formulas for nested case-control designs as well as more general two-phase and case-control samples augmented by an external cohort design are established in Table 5.1. There is no need to assume that the event rate is known when it is in fact estimated from one of these other designs. What could be an efficiency gain following from knowledge would otherwise become an unaccounted component of variability. The unaccounted variability may be all the more important when using optimal control-to-case ratios. Conversely, in the unusual case that the event rate is truly known, the efficiency gains can be substantial and all the more so when estimating from case-control samples with optimal control-to-case ratio.

5.5 Efficiency of Two-Phase Studies When Sampling All Cases

Many studies elect to analyze biomarker measurements on a sample consisting of all cases and a subset of controls. The choice to use these types of designs is motivated not by strict limits on the number of samples, but rather by other considerations, such as human effort and time or a limit on the number of cases available. Novel biomarkers under investigation may be measured using arrays that are also novel and require significant lab time to process. As biomarker candidates proceed through the development sequence, so do the assays. This produces a more flexible relationship between practical efficiency and statistical efficiency. Clearly, every additional control reduces the variability of estimates. The design question involves ascertaining the size of a control sample that achieves "enough" of the statistical efficiency possible, which is more subjective than an optimal control-to-case ratio.

The analytic formulas of the limiting variance enable one to evaluate the increase in

statistical efficiency as a function of the proportion of controls included in the case-control sample. In this section, we consider a two-phase study design with all cases included in the sub-sample, $\pi(1) = 1$, and all covariates measured only in phase-II. The efficiency of estimators based on only a portion $\pi(0)$ of the controls, relative to that of estimators based on all cases and controls, the full cohort estimator, equals:

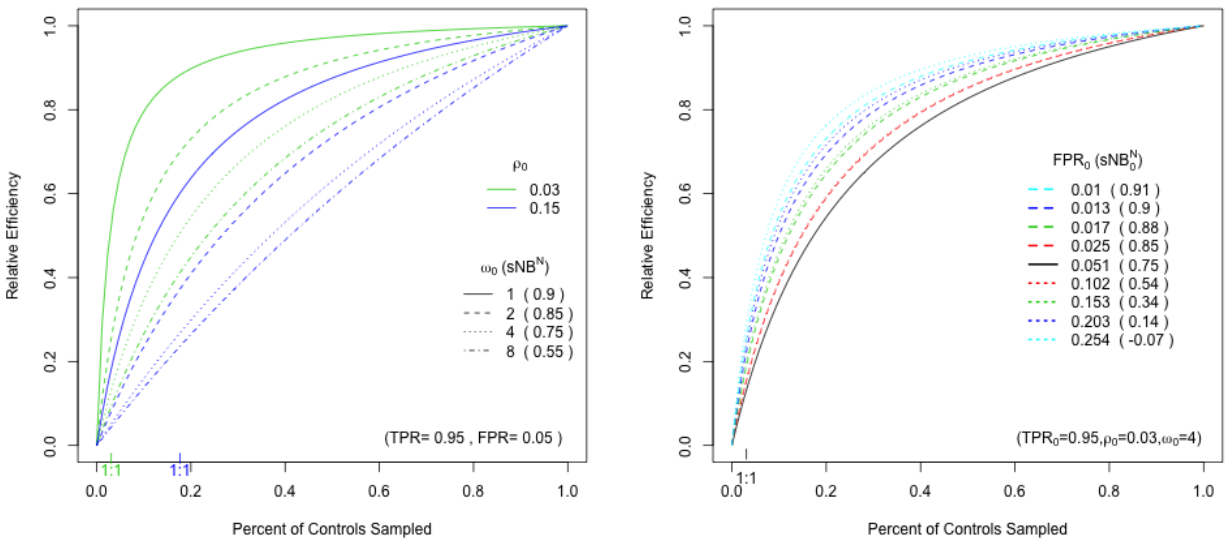
$$\text{RE} = \frac{\sigma^{2,cohort}}{\sigma^{2,tp}} = \frac{\sigma_T^2 + \sigma_F^2 + \sigma_R^2}{\sigma_T^2 + \frac{1}{\pi(0)}\sigma_F^2 + \sigma_R^2}.$$

We note that $1 \leq \frac{1}{\pi(0)} \leq \infty$ and hence the relative efficiency lies between 0 and 1, equaling zero in the degenerate case of no controls ($\pi(0) = 0$). Further, the expression implies that the relative efficiency will always be at least $\pi(0)$.

Figure 5.4 presents relative efficiency curves for estimators of sNB^N using all cases as a function of the percentage of controls sampled from the cohort. The first set of scenarios, Figure 5.4(a), applies to a clinical decision rule with classification accuracies ($\text{TPR}_0 = 95\%, \text{FPR}_0 = 5\%$) in clinical contexts defined by 3% and 10% outcome probabilities, and population level control-to-case benefit ratios that vary between 1 and 8. The second set of scenarios, Figure 5.4(b), apply to clinical contexts with 3% outcome probability, population level tradeoffs of 4, and a decision rule with 95% true-positive rate and false-positive rates varying between 1% and 25%.

Structure in the Relative Efficiency Curves

In Figure 5.4(a), we see that, for a rule with classification accuracies ($\text{TPR}_0 = 95\%, \text{FPR}_0 = 5\%$), the relative efficiency, plotted as a function of the percentage of controls sampled, approaches the $y = x$ line as the clinical contexts have higher outcome probability and increasing population-level benefit trade-offs. For any fixed outcome probability, weight (and hence control-to-case benefit ratio), and true-positive rate, there is a false-positive rate that achieves the minimum relative efficiency at any percentage of controls sampled. This follows from basic calculus that concludes with solving for the positive root of a quadratic formula. In the example of Figure 5.4(b), where $\rho_0 = 0.03$, $\omega_0 = 4$, and $\text{TPR}_0 = 0.95$, rules for which



(a) (b)

Figure 5.4: Efficiency of two-phase samples consisting of all cases and a subset of controls, relative to the using the whole phase-one cohort: (a) over various outcome probabilities ρ_0 and weights ω_0; (b) over various false-positive rates.

$\text{FPR}_0 = 0.051$ will require a greater portion of controls in phase two than any other in order to achieve any given relative efficiency level.

Implications for Study Design

Trends observed in Figure 5.4(a) can be stated as general considerations when designing a study subject to soft constraints. We observe that for more common outcomes, the relative efficiency curves are more gradual. For a given outcome probability, the same behavior is noticeable for larger weights, which correspond to larger control-to-case benefit ratios. Both cases describe clinical contexts in which controls are of relatively greater importance. The relative efficiency curves show that additional controls remain valuable, to varying extents,

across the spectrum of a particular portion being sampled. Conversely, for contexts in which the outcome is less common and the population level control-to-benefit trade-offs is smaller, a fair amount of efficiency is recouped with a small portion of controls. In the example of $C = (\rho_0 = 0.03, \omega_0 = 1)$, the relative efficiency curve rises quickly and plateaus; almost 85% of the efficiency available from sampling the full cohort is available from sampling only 20% of all controls, in addition to all cases.

5.6 Simulated Performance of $sNB^N_{cc,n}$

All simulations are based on the same super-population used in previous chapters.

We focus on evaluating the performance of pre-specified baseline and extended risk rules and their difference from nested case-control samples. The simulations presented mimic the model development step, which defines the risk model, followed by a validation step in which the developed model is evaluated. The origin of the pre-specified risk rule is not of central interest and we employ models that were fit on a cohort sample with 600 subjects. The pre-specified baseline risk model was defined by a logistic model for D, as predicted by two of the covariates W_1 and W_2 (PredMod), and the extended risk model included an additional novel biomarker W_3 (PredMod_{ext}) and was also logistic in form.

Validation samples with phase-one sample sizes N_{valid} equal to 1400, 2800, and 5600 were sampled from the remaining super-population and contribute values of D to the study. To remove the element of varying phase-two sample sizes across replicates, a nested case-control study with fixed phase-II sample sizes was employed. For each parent cohort, one-seventh of all observations were selected, in 1:1 control-to-case ratio, to contribute second phase covariate measurements (W_1, W_2, W_3) to the study. The net benefit for four possible clinical decision rules, combining a pre-specified risk model with risk thresholds of 10, 20, 30 or 40% (corresponding to control-to-case trade-offs of 1:9, 1:4, 1:2.5 and 2:3), is evaluated in the context of identifying high-risk individuals to target for an intervention that is not routinely administered.

To study the frequentist properties of the analytic variance and confidence intervals pro-

			Std. Dev.		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
Model 1	0.533	-0.26	0.055	0.055	94.6	94.7
Model 2	0.528	-0.23	0.057	0.057	94.7	94.9
Model 3	0.531	0.06	0.056	0.057	94.9	94.6
Model 4	0.530	-0.29	0.056	0.056	94.5	94.4

Table 5.2: Estimation of $sNB(r_T = 0.2)$ for 4 prespecified models (developed from 4 independent samples of N=600) over 5,000 replicates of two-phase sampled validation sets (N_I=2,800 and N_{II}=400).

posed for validating a clinical decision rule in terms of net benefit, selection of the validation samples and evaluation of standardized net benefit were conducted 5,000 times for each model developed. The true net benefit, sNB_0, of a pre-specified model is determined by using the predicted values for the entire super-population. General phase-two sampling results were used to the estimate the limiting variance, as was justified in Section 5.2. Estimates of variability made from bootstrapped distributions of estimators employed outcome-stratified re-sampling among phase-two observations, to capture the variability due to estimation of the true- and false-positive rates and re-sampling among the entire cohort for to capture the variability of estimation of the outcome probability. This bootstrapping approach does not emulate the study design. Re-sampling the whole cohort and then selecting which observations contribute in phase two is not viable in practice since the phase-two measurements are not available on the full cohort. The proposed method can not only be executed in practice, but will properly approximate the variability of sNB_n^{tp} because of the independence of the constituent estimators, or equivalently the orthogonality between the three components of the influence function IF^{tp}. This holds for each of the case-control estimators considered in this section.

Table 5.2 summarizes estimates of the net benefit, for each of 4 different pre-developed

models, from validation samples of Phase-I size 2,800 with 200 cases and 200 controls measured at Phase-II, for decision rules using a 20% risk threshold. Each independently developed model has a different true (model-dependent) value of sNB. Overall, the bias, as a percent of true sNB, is fairly small and analytic estimates of variance match the observed Monte Carlo variance of estimated performance quite well. Coverage of 95% confidence intervals is similarly on target with slightly greater coverage for bootstrap constructed compared to analytic Wald confidence intervals, but there is no clear trend on which is superior.

Table 5.4 examines in more depth the properties of validation for one pre-specified model (Model 1 in Table 5.2). We see similar behavior as for the cohort estimator, but with some previously noted weaknesses more apparent due to the smaller samples contributing to estimation of the true or false positive rates. For example, we expect good coverage of Wald confidence intervals to require large samples for the simple proportions represented by true- and false-positive rates, all the more so for rates closer to zero or one. For decision rules defined using a 40% high-risk threshold, the false-positive rate is roughly 2.6% and coverage in the smaller sample sizes considered is fairly low (88.6-93%) for both the bootstrapped and analytic confidence intervals. However, for the considered study design, phase-one cohorts of 1,400 or 2,800 participants translate into only 70 or 140 controls measured in phase two. In the simulations of Chapter 3, the smallest simple cohort considered had 400 participants, which corresponds to roughly 360 controls under the simulated 10% outcome probability. The coverage for confidence intervals for the true-positive rate, which equals roughly 84% for the rule using a 10% high-risk threshold, is low (92.8%) when only 70 cases are measured at phase two. This sample size is close to that of a simple cohort with 800 participants (roughly 80 cases) considered previously, and again the observed behavior is consistent with previous results.

Overall, and especially for the more moderate true- and false-positive rates and more moderate phase-II sample sizes, the bias is low, analytic estimates of variability match bootstrapped estimates, and coverage is quite good. As expected, estimators of sNB, and in particular, the coverage achieved by Wald confidence intervals, have better performance

			Std. Dev.		Coverage	
	ΔsNB$_0$	% Bias	MC	Ana	BS	Ana
Model 1	0.229	0.16	0.054	0.054	95.0	95.0
Model 2	0.228	0.27	0.054	0.055	95.0	95.4
Model 3	0.232	-0.50	0.054	0.054	94.7	94.7
Model 4	0.218	0.56	0.050	0.050	93.8	94.2

Table 5.3: Estimation of ΔsNB($r_T = 0.2$) for 4 prespecified models (developed from 4 independent samples of N=600) over 5000 replicates of two-phase sampled validation sets (N_I=2800, N_{II}=400).

			sNB				TPR				FPR				
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
								True Value							
			0.673	0.533	0.435	0.357	0.838	0.712	0.608	0.515	0.166	0.080	0.045	0.026	0.100
								% Bias							
1400	200		-0.13	-0.32	-0.12	0.07	0.07	0.09	0.12	0.17	0.19	0.54	-0.04	-0.32	0.02
2800	400		-0.12	-0.26	0.01	0.58	0.00	-0.08	-0.03	0.07	0.14	0.10	-0.49	-1.48	0.01
5600	800		-0.04	0.04	0.06	0.20	0.01	0.04	0.10	0.07	-0.09	-0.22	-0.07	-0.52	-0.07
								Standard Deviation							
1400	200	observed	0.055	0.079	0.097	0.110	0.037	0.045	0.049	0.050	0.037	0.027	0.021	0.016	0.008
		analytic	0.055	0.078	0.095	0.110	0.037	0.045	0.049	0.050	0.037	0.027	0.021	0.016	0.008
2800	400	observed	0.038	0.055	0.068	0.078	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011	0.006
		analytic	0.039	0.055	0.067	0.077	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011	0.006
5600	800	observed	0.027	0.039	0.047	0.054	0.018	0.022	0.024	0.025	0.019	0.014	0.010	0.008	0.004
		analytic	0.027	0.039	0.047	0.055	0.018	0.023	0.024	0.025	0.019	0.014	0.010	0.008	0.004
								95% Coverage							
1400	200	bootstrap	94.6	94.3	93.3	91.2	92.8	94.5	94.7	95.2	93.4	93.4	93.6	92.5	94.6
		analytic	94.2	94.1	93.4	90.7	92.8	95.6	94.7	95.7	94.4	90.0	94.3	93.0	94.6
2800	400	bootstrap	95.0	94.6	94.1	93.1	94.3	95.3	95.2	94.6	94.4	94.4	93.1	88.6	94.5
		analytic	94.8	94.7	93.8	92.4	94.2	95.3	95.3	93.9	93.9	95.5	94.0	88.6	94.5
5600	800	bootstrap	95.2	94.9	94.5	94.7	95.0	95.3	95.1	95.5	94.7	94.4	94.7	94.5	94.9
		analytic	95.1	94.8	94.4	94.4	95.2	94.9	94.9	95.6	95.3	95.1	92.9	94.6	95.1

Table 5.4: Simulation of validating(PredMod) from a a 1:1 case-control phase-two sample. The full phase-one cohort has N_I subjects, of which N_{II} are measured at phase two. Results for each validation sample size are based on: 5,000 replications.

			ΔsNB				ΔTPR				ΔFPR			
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
							True Value							
			0.164	0.229	0.267	0.289	0.075	0.159	0.229	0.289	-0.089	-0.031	-0.010	0.000
							% Bias							
1400	200		0.38	0.71	0.36	-0.26	-0.69	-0.23	-0.09	-0.11	0.58	2.06	2.31	326.28
2800	400		0.03	0.16	-0.51	-1.13	-0.53	0.16	-0.16	-0.29	0.14	-0.23	-2.94	1917.72
5600	800		-0.00	-0.06	-0.03	-0.08	-0.32	-0.19	-0.24	-0.03	-0.01	-0.02	1.03	108.40
							Standard Deviation							
1400	200	observed	0.053	0.079	0.100	0.124	0.036	0.044	0.048	0.050	0.036	0.028	0.022	0.018
		analytic	0.052	0.077	0.100	0.123	0.036	0.044	0.048	0.050	0.036	0.028	0.022	0.019
2800	400	observed	0.036	0.054	0.071	0.088	0.025	0.031	0.034	0.035	0.026	0.020	0.016	0.013
		analytic	0.037	0.054	0.070	0.087	0.026	0.031	0.034	0.035	0.026	0.020	0.016	0.013
5600	800	observed	0.026	0.038	0.050	0.061	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
		analytic	0.026	0.038	0.050	0.062	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
							95% Coverage							
1400	200	bootstrap	94.3	94.3	93.8	92.6	94.4	94.6	94.3	94.7	93.6	91.7	89.8	80.2
		analytic	94.4	94.7	95.0	95.6	94.7	94.3	94.2	94.2	93.8	93.6	93.7	93.3
2800	400	bootstrap	94.8	95.0	93.5	93.1	94.2	94.9	95.0	94.7	94.6	94.3	92.6	91.2
		analytic	94.9	95.0	94.2	94.7	94.8	95.1	94.9	94.6	94.5	94.5	95.0	94.7
5600	800	bootstrap	94.6	94.9	95.0	94.4	94.4	94.4	94.5	94.6	94.7	95.0	93.9	93.8
		analytic	94.7	95.1	95.3	95.0	94.5	94.4	94.6	94.6	94.8	95.1	94.9	95.3

Table 5.5: Simulation of validating(PredMod$_{ext}$-PredMod) from a 1:1 case-control phase-two sample. The full phase-one cohort has N_I subjects, of which N_{II} are measured at phase two. Results for each validation sample size are based on: 5000 replications.

than for the simple proportion constituents. Tables 5.3 and 5.5 present the same results for the contrast between the net benefits of the two risk models. Qualitatively, the results are consistent with the previously made conclusions. Taking into account the coverages of confidence intervals for estimators of the very small false-positive rates for the extended models (Appendix B.9), the coverages observed for the difference in false positive rate are generally better than those corresponding to the single extended risk models, but not always for the corresponding single base models. Accepting the known limitations, the performance of the variance estimators and their corresponding coverages, which are based on asymptotic approximations, appear to demonstrate quite reasonable finite sample performance on nested case-control samples of realistic sizes.

Extended results tables, including bootstrapped standard errors and corresponding Wald CI, results for the extended model, and simulations for estimation from stand-alone case-controls are in Appendix B.9.

5.7 Summary

In this chapter, we extended previous results for conducting inference on empirical estimators of net benefit from cohort studies to empirical estimators from three variants of unmatched case-control sampling schemes: (1) a stand-alone case-control study, (2) a case-control study augmented with an estimate of the outcome probability, and (3) an outcome-dependent secondary sample of a two-phase study conducted on a cohort. The stand-alone case-control study requires a known outcome probability and we examined the implications of this assumption. Design for case-control studies intended to assess the potential clinical utility of a decision rule was considered in two ways. First, the optimal control-to-case ratio, for a fixed sample-size, and efficiency loss associated with non-optimal ratios were presented. Second, the efficiency gain of additional controls for cases where there is not a strict limit on the case-control sample size were discussed. Performance of the proposed estimators was evaluated with a simulation study.

Chapter 6

ESTIMATION FROM A COHORT WITH CENSORED OUTCOMES

The focus of this chapter is to extend inference for standardized net benefit to settings in which the outcome is subject to censoring. This is of particular importance for assessing the potential clinical utility of proposed decision rules in contexts where the clinical outcome can occur within a period of time. We begin by introducing Framingham Risk Functions as motivation. An empirical estimator for standardized net benefit, under censored outcomes, is introduced. We then establish large sample distribution theory and examine the performance of the estimators in a simulation study.

6.1 Motivating Example: Framingham Risk Functions

The Framingham Heart Study was initiated in 1948 as a prospective study on cardiovascular disease (CVD) among residents of Framingham, Massachusetts. This project of the National Heart, Lung, and Blood Institute is long-term and remains active. Exam 32 for the original cohort of 5,209 participants was completed in 2014. Numerous additional studies and cohorts, such as those (1) focusing on children of previous cohort members and their spouses (Offspring and Third Generation Cohorts), (2) reflecting diversity of current residents (Omni Cohorts 1 and 2) and (3) expanding the clinical measurements obtained, such as post-mortem brain tissue and genotype data, continue to expand the scope of this project. (FHS, 2017)

Over the years, these studies have made numerous major findings on contributing factors of CVD. Additionally, predictive risk models, termed Framingham Risk Functions (FRF), have been published for a variety of cardiovascular outcomes and time-frames. For example,

the study website (https://www.framinghamheartstudy.org/risk-functions/index.php) gives access to calculators for the probability of congestive heart failure within 4 years, the 30 year risk of cardiovascular disease (CVD), or the two-year risk of a second coronary heart disease (CHD) event. Each risk calculator is intended for use in a specified population and relies on patient-specific variables including age, sex, BMI, clinical details on past disease, smoking history, and measurements such as cholesterol levels or blood pressure.

The Framingham Cardiovascular Risk Profile has been proposed for use in primary care (D'Agostino et al., 2008). The National Cholesterol Education Program guidelines for cholesterol management directly calls for assessment of 10-year risk of coronary heart disease from a Framingham Risk Function among patients with 2 or more risk factors for CHD. A predicted risk of greater than 20% results in more aggressive management recommendations and lower low density lipid thresholds for initiation of drug therapy (Panel et al., 2002).

Within the Tehran Lipid and Glucose Study (TLGS), the Framinghman 10-year cardiovascular risk score, developed on an American cohort, was found to be suitable for an Iranian cohort residing in Tehran (Khalili et al., 2012). In fact, the performance of the prespecfied FRF was essentially as good as the models re-fit to the Tehrani data. This study validated the clinical utility for settings where risk rules using 10 or 20% thresholds would be considered rational. It is well-poised to evaluate rules based on additional thresholds, once cost-benefit trade-offs for a well-defined intervention targeted to those high at risk for a cardiovascular event within 10 years have been examined. With a risk window of 10 years, any study is likely to have some amount of censored outcomes; the median follow-up among the 6,224 participants in the TLGS was 9.3 years.

6.2 Notation

We focus on estimation from a cohort and start with a set-up similar to that defined in Chapters 3. The clinical outcome of interest remains whether or not a clinical event occurs. In this setting, the clinical outcome includes a set time-frame, $[0, t_c]$, following patient evaluation. In the context of FRFs, this could be whether or not a hard coronary endpoint, such

as myocardial infarction, occurs within 10 years.

In the absence of censoring, this can be represented with a binary variable D for event occurred ($D = 1$) or not ($D = 0$) within the time-frame. With censored observations, we can only record follow-up time and whether or not the event was observed. Using T to represent the, perhaps unobserved, time of the clinical outcome and C to represent censoring time, the pair $(Y = T \wedge C, \Delta = I[Y = T])$ provides the usual observable data structure for the outcome; Y denotes the amount of time the patient was observed, and Δ is an indicator of whether or not the outcome was observed.

On each patient, additional baseline covariates W could be measured. The clinical decision rule, R, associates an intervention class $x = R(w)$ to each covariate level, where the decision is to intervene ($x = 1$) or not ($x = 0$). Given a risk function, r, the risk associated with a covariate value w is $r(w) := \Pr(D = 1 \mid W = w)$, which is meaningful for the patient and treatment decisions. Combined with a high-risk threshold, r_T, the resulting clinical decision rule is $x = R(w; r, r_T) = I[r(w) > r_T]$. When working with censored outcomes, it will be convenient to express the risk $r(w)$ as $\Pr(T \leq t_c \mid W = w)$ and the outcome D as $I[T \leq t_c]$. The cumulative incidence at time t_c is denoted by $\rho := \Pr(D = 1) = \Pr(T \leq t_c)$.

Each member of the cohort contributes an observation to the data set: $O^{cens} = (Y, \Delta, X)$ which we assume was sampled identically and independently from some unknown population distribution $O_1^{cens}, \ldots, O_n^{cens} \sim_{iid} P_0^{cens}$. The variable X inherits its randomness from W, $X = R(W)$, and at times it will be convenient to express the observation as (W, D), $(R(W), D)$, or in the case of a decision rule derived from a risk model, $(r(W), D)$ or $(I[r(W) > r_T], D)$.

6.3 Empirical Estimator of Net Benefit From Censored Outcomes

In the absence of censoring, empirical estimators for each constituent of net benefit, ρ, TPR, and FPR, were readily available using the observed outcome indicators. We will first establish empirical estimators, for each of these constituents, from data with censored outcomes. Under the assumption that censoring is independent of event time, the Kaplan-Meier curve at t_c, $S_n(t_c)$, provides a consistent estimator of the survival probability, $\Pr(T > t_c)$, and

consequently of the outcome probability through $\rho_n = 1 - S_n(t_c)$.

Let $X = R(W)$ denote the intervention decision for a patient with covariates W; in the case of a clinical risk rule, this is equivalent to high-risk classification $X = I[r(W) > r_T]$. We will use X to recast the true- and false-positive rates of the decision rule in terms of the observed data structure, through applying Bayes rule as follows:

$$TPR = \Pr(X = 1 \mid D = 1) = \frac{\Pr(D = 1 \mid X = 1) \Pr(X = 1)}{\Pr(D = 1)}$$

from which a survival-based estimator naturally follows:

$$TPR_n^{cens} \quad := \quad \rho_{X,n} \frac{1 - S_{1,n}(t_c)}{1 - S_n(t_c)}$$

where $\rho_{X,n} := \frac{1}{n} \sum_i X_i$ and $S_{1,n}(t_c)$ is the estimate from the Kaplan-Meier curve specific to patients with $X = 1$. The empirical estimator of the proportion of patients recommended intervention by the clinical decision rule, $\rho_{X,n}$, equals the probability of a high-risk classification when the clinical rule is derived from predicted risk.

An empirical estimator of the false-positive rate is defined similarly:

$$FPR_n^{cens} \quad := \quad \rho_{X,n} \frac{S_{1,n}(t_c)}{S_n(t_c)}$$

as well as an empirical estimator of the weight:

$$\omega_n \quad := \quad \frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}} \frac{S_n(t_c)}{1 - S_n(t_c)},$$

where $\frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}} = \frac{r_T}{1 - r_T}$ for a rational risk rule. Altogether these define sNB_n^{cens} and NB_n^{cens} that simplify to:

$$sNB_n^{cens} \quad := \quad \left[1 - \left(1 + \frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right) S_{1,n}(t_c)\right] \frac{\rho_{X,n}}{1 - S_n(t_c)}$$

$$NB_n^{cens} \quad := \quad \left[1 - \left(1 + \frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right) S_{1,n}(t_c)\right] \rho_{X,n}$$

respectively. Key elements of this approach were proposed by Heagerty et al. (2000) in the context of time-dependent ROC curves. This approach was promoted by Vickers et al. (2008) as a means of accounting for censoring when estimating net benefit, and applied by (Khalili et al., 2012) in a net benefit analysis within the Tehran Lipid and Glucose Study.

Assumptions

The estimators for true- and false-positive rates, as well as that for the composite measure of net benefit, rely on two Kaplan-Meier curves. Hence, these estimators inherit the assumptions of independent censoring from the Kaplan-Meier estimates, in order to ensure consistency. The overall Kaplan-Meier estimator, $S_n(t_c)$ requires censoring to be independent of event time: $C \perp\!\!\!\perp T$. The second Kaplan-Meier estimator, restricted to the population for whom intervention is assigned by the risk rule, i.e., the high-risk subpopulation in the case of an underlying risk rule, requires censoring to be independent of event time, conditional on high-risk status: $(C \perp\!\!\!\perp T) \mid X = R(W)$. The more transparent condition, that censoring is jointly independent of event time and the predictors in the decision rule, $C \perp\!\!\!\perp (T, W)$, is sufficient, though not necessary, to guarantee the required assumptions.

Alternate Formulations

The empirical estimator of net benefit, sNB_n^{cens}, and the empirical estimators of the constituent true- and false-positive rates, TPR_n^{cens} and FPR_n^{cens}, from data with censored outcomes were formulated in terms of the sub-population assigned to intervention, which for a risk rule is the high-risk sub-population. Noting that $\text{TPR} = 1 - \text{FNR}$ and $\text{FPR} = 1 - \text{TNR}$, it is clear that there are analogous formulations with respect to the sub-population recommended non-intervention by the rule, which for a risk rule is the low-risk sub-population. The two formulations are asymptotically equivalent, but in finite samples one formulation or the other, using $S_{1,n}$ or $S_{0,n}$, might perform better.

6.4 Inference for sNB$_n^{cens}$

Influence Functions

The influence function for the censored data formulation of sNB can be calculated via the delta method for influence functions. This yields the following expression in terms of the influence functions for the two Kaplan-Meier estimators, on the population as a whole, S_n,

and on the predicted high-risk sub-cohort, $S_{1,n}$, and for the overall probability of being predicted as high-risk, ρ_X:

$$\begin{aligned}\mathrm{IF}_{\mathrm{sNB}}^{cens}(o^{cens}) &= -\left(1+\frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right)\frac{\rho_X}{1-S(t_c)}\mathrm{IF}_{\mathrm{S}_1(t_c)}(o^{cens}) + \frac{1-\left(1+\frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right)S_1(t_c)}{\left(1-S(t_c)\right)^2}\rho_X\mathrm{IF}_{S(t_c)}(o^{cens}) \\ &\quad + \frac{1-\left(1+\frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right)S_1(t_c)}{1-S(t_c)}\mathrm{IF}_{\rho_X}(o^{cens}), \qquad (6.1)\end{aligned}$$

where o^{cens} is a realization of the observation is $O^{cens} = (Y, \Delta, X)$. Substituting in the influence function for a Kaplan-Meier estimator, as established by Reid (1981), and for a simple proportion yields:

$$\begin{aligned}\mathrm{IF}_{\mathrm{sNB}}^{cens}(o^{cens}) &= \left(1+\frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right)\frac{1}{1-S(t_c)}xS_1(t_c)\left\{\frac{\delta I[y\le t_c]}{S_{Y,1}(y)} - \int_0^{t_c\wedge y}\frac{1}{S_{Y,1}(u)}d\Lambda_1(u)\right\} \\ &\quad - \frac{1-\left(1+\frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right)S_1(t_c)}{\left(1-S(t_c)\right)^2}\rho_X S(t_c)\left\{\frac{\delta I[y\le t_c]}{S_Y(y)} - \int_0^{t_c\wedge y}\frac{1}{S_Y(u)}d\Lambda(u)\right\} \\ &\quad + \frac{1-\left(1+\frac{\mathrm{B}^{ctrl}}{\mathrm{B}^{case}}\right)S_1(t_c)}{1-S(t_c)}\left\{x-\rho_X\right\}.\end{aligned}$$

where S_Y and $S_{Y,1}$ are survival functions for membership in the risk set for the whole population and conditional on high-risk status, respectively. The measures $d\Lambda(u)$ and $d\Lambda_1(u)$ reflect the instantaneous hazard of $T = 1$ and $\Delta = 1$ at $t = u$ within the whole population and restricted to the high-risk subset.

The influence functions for TPR_n, FPR_n, and ρ_n are similarly derived and equal:

$$\begin{aligned}\mathrm{IF}_{\mathrm{TPR}}^{cens}(o^{cens}) &= -\frac{\rho_X}{1-S(t_c)}\mathrm{IF}_{\mathrm{S}_1(t_c)}^{cens}(o^{cens}) + \frac{1-S_1(t_c)}{\left(1-S(t_c)\right)^2}\rho_X\mathrm{IF}_{S(t_c)}^{cens}(o^{cens}) + \frac{1-S_1(t_c)}{1-S(t_c)}\mathrm{IF}_{\rho_X}^{cens}(o^{cens}) \\ \mathrm{IF}_{\mathrm{FPR}}^{cens}(o^{cens}) &= \frac{\rho_X}{S(t_c)}\mathrm{IF}_{\mathrm{S}_1(t_c)}^{cens}(o^{cens}) - \frac{S_1(t_c)}{S^2(t_c)}\rho_X\mathrm{IF}_{S(t_c)}^{cens}(o^{cens}) + \frac{S_1(t_c)}{S(t_c)}\mathrm{IF}_{\rho_X}^{cens}(o) \\ \mathrm{IF}_{\rho}^{cens}(o^{cens}) &= -IF_{S(t_c)}^{cens}(o^{cens}).\end{aligned}$$

Empirical Estimator of the Limiting Variance

The limiting variance of an asymptotically linear estimator is $\sigma_0^2 := \mathbb{E}\left[IF(O)^2\right]$, the variance of the zero-mean influence function. In previous sections, the limiting variance simplified to

an expression involving the same constituent parameters as the estimator, enabling use of a plug-in estimator. It is unclear that any such simplification will occur here. Instead, we define an empirical estimator of $\sigma_0^{2,cens}$ by taking the sample average of

$$\sigma_n^{2,cens} = \frac{1}{n}\sum_i \left\{IF_n^{cens}(O_i^{cens})\right\}^2,$$

where IF_n is the empirical estimator of the influence function obtained by plugging in estimators for each of the distributional features. For example, $\rho_{X,n}, S_{1,n}(t_c), S_n(t_c)$ can be calculated once and substituted into the formula of $\mathrm{IF}_{\mathrm{sNB}}(o)$. Likewise, for each value of y observed, $S_{Y,n}(y)$ and $S_{Y,1,n}(y)$ can be calculated and plugged into the influence function for observations with $Y = y$. Similarly, the compensator component of each Kaplan-Meier influence function, e.g., $\int_0^{t_c \wedge y} \frac{1}{S_Y(u)} d\Lambda(u)$, can be empirically estimated by $n \sum_{y_{(j)} \leq y \wedge t_c} \frac{d_{(j)}}{N_{(j)}^2}$, where the subscripts (j) reference an ordering of observation times. This last expression uses notation commonly employed when describing the Nelson-Aalen estimator $\Lambda_n(\mathrm{u})$; d_j is the number of events observed at the j^{th} observed follow-up time and N_j is the size of the corresponding risk set.

6.5 Simulated Performance of sNB_n^{cens}

Data Generation

We depart from the super-population employed in previous simulations and illustrations and instead opt to generate time-to-event data directly. The outcome event time T is modeled according to a Weibull distribution, with shape and location parameters dependent on the intervention assigned by the clinical decision rule $X = R(W)$. The censoring times follow an exponential distribution independent of both covariates W and event time T. The probability of being assigned to the intervention group is ρ_x.

Our choice of parameters are chosen to create a scenario similar to those considered in previous simulations. In the target population, 13% of the subjects would be classified high-risk by the implicit decision rule R. Among high-risk subjects, the shape and scale parameters are 4 and 7 respectively, whereas among low-risk subjects they are 8 and 12,

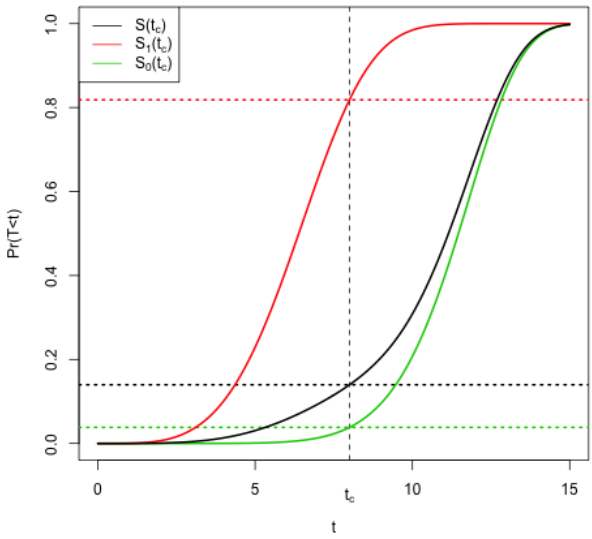


Figure 6.1: Underlying CDF of event times simulated for whole population (black), among high-risk (red) and among low-risk (green) sub-populations. The landmark time of $t_c = 8$ is designated with a vertical dashed line.

respectively. The landmark time is $t_c = 8$, which will we think of on the scale of years. The corresponding CDFs for event times across the whole population and limited to the high and low-risk sub-populations are presented in Figure 6.1. This data-generating mechanism implies an overall probability of the clinical outcome within 8 years, $\rho_0 = 1 - S(8)$, of roughly 14%, and the rule achieves classification accuracies of roughly $\text{TPR}_0 = 76.2\%$, and $\text{FPR}_0 = 2.7\%$. Employing a control-to-case benefit trade-off ratio equal to 2, yields weight $\omega \approx 12.3$ and $\text{sNB}_0 \approx 0.424$, in this setting.

Under no censoring (rate parameter equal to 0), in combination with the generated event times, we anticipate simulated data sets to have about 86% of participants observed through the entirety of the 8 year study period and occurrence of the outcome event among roughly 14%, who consequently had less than 8 years of observation. We used a rate of 0.01 in the exponential censoring distribution. In combination with the generated event times, we

anticipate simulated data sets to have roughly 79.4% of participants observed through the entirety of the 8 year study period. Of the roughly 20.6% with less than 8 years of observation, 63.9% are expected to have had the clinical outcome observed, with the remaining 36.1% (corresponding to $\approx 7.4\%$ of the study cohort) having incomplete follow-up. Three additional censoring scenarios with exponential rate parameters of 0.02, 0.04 and 0.07, respectively corresponding to roughly 14.3, 26.5 and 41.7% of the study cohort with incomplete follow-up, were also evaluated. Results were similar to those presented here; see Appendix B.10.

Results

In Table 6.1, the estimators of ρ_0, TPR_0, FPR_0, and sNB_0 and are evaluated in terms of bias and agreement between the estimated analytic and observed Monte Carlo standard deviations. Coverage of the proposed Wald confidence intervals is plotted in Figure 6.2. Observed behavior was ascertained from 5,000 replicates for sample sizes ranging between 2,000 and 40,000. The lower range of sample sizes considered represent the motivating examples of original Framingham cohort and that of the Tehran Glucose and Lipid Study. The upper range of sample sizes considered is sufficient to establish asymptotic behavior.

In the scenario considered, the proposed point and variance estimators behave quite well. Bias is negligible at the sample sizes considered and the average standard deviation, estimated using the proposed analytic formulas of the respective influence functions, agrees well with the Monte Carlo standard deviation observed across replicates for estimates of sNB_0 as well as the three constituents: TPR_0, FPR_0, and ρ_0. The observed coverage rates generally lie between 94.4 and 95.6% and deviations from . These values correspond to the upper and lower limits for the central 95% of coverage estimates, based on 5,000 replicates, when the true coverage is 95%. We will refer to this range, indicated by dashed horizontal lines in the figure, as the Monte Carlo range centered at 95%. Estimators for which coverage fell outside of the expected Monte Carlo range were primarily due to the known challenges associated with Wald-confidence intervals for a small probability, $\text{FPR}_0 = 2.7\%$, estimated

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-0.003	-0.000	-0.004	0.001
2000		0.003	0.001	-0.006	-0.001
5000		-0.000	-0.000	-0.002	-0.000
10000		0.001	0.000	-0.001	0.000
20000		0.000	-0.000	-0.000	0.000
40000		0.000	0.000	-0.000	0.000
			Standard Deviation		
1000	observed	0.088	0.039	0.006	0.011
	analytic	0.084	0.038	0.006	0.011
2000	observed	0.062	0.028	0.004	0.008
	analytic	0.060	0.027	0.004	0.008
5000	observed	0.038	0.017	0.003	0.005
	analytic	0.038	0.017	0.003	0.005
10000	observed	0.027	0.012	0.002	0.004
	analytic	0.027	0.012	0.002	0.004
20000	observed	0.019	0.009	0.001	0.003
	analytic	0.019	0.009	0.001	0.003
40000	observed	0.014	0.006	0.001	0.002
	analytic	0.014	0.006	0.001	0.002

Table 6.1: Evaluation of point and variance estimators for Kaplan-Meier based estimators, under a 0.01 exponential censoring rate.

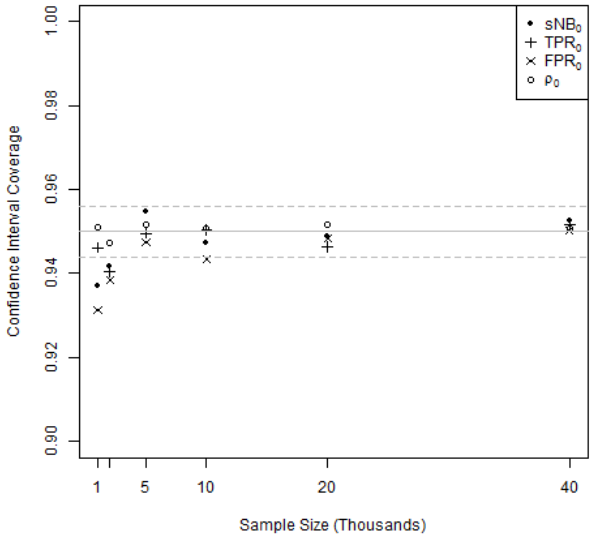


Figure 6.2: Coverage of 95% confidence intervals, for Kaplan-Meier based estimates, under a 0.01 exponential censoring rate.

from the smallest cohorts considered. At these sample sizes, estimation of the false-positive rate achieved 93.1 (N=1,000) and 93.8% (N=2,000), which translated to respective coverages of 93.7% and 94.2% by the estimator of net benefit, sNB_n^{cens}. All other estimator and sample size combinations performed at least as well.

6.6 Simulated Performance of Naive Use of sNB_n^N

Here we consider naive application of the estimator developed for uncensored cohort data, $\text{sNB}_{cohort,n}$, to a data set in which the outcomes are, in fact, subject to censoring. We employ the same data generation set-up as in Section 6.5. Once a case or control status is assigned to each observation, the estimator introduced in Chapter 3, can be calculated routinely. When the outcome event has been observed before the landmark time t_c, the observation is complete and can be correctly identified as a case subject. When the subject has been

observed through the landmark time, t_c, and no outcome event occurred, then the observation is also complete and can be correctly identified as a control subject. Observations censored before the landmark time and for whom no outcome event was observed, i.e., $\Delta = 0$ and $Z < t_c$, are incomplete in terms of case-control status.

We consider two approaches to conducting analyses, both of which effectively ignore the reality that outcomes were subject to censoring by either excluding incomplete observations or by treating incomplete observations as though the outcome did not occur during the unobserved follow-up time (i.e., assumes they are controls). Adoption of both of these approaches was observed in a systematic review of published analyses, for the purpose of assessing the clinical utility of biomarkers in decision making, using net reclassification statistics (Leening et al., 2014). The former strategy was cautioned against by Vickers et al. (2008) when using net benefit for similar purposes.

We show that , unlike the Kaplan-Meier based method considered in this Chapter, these analyses have undesirable properties when outcomes are in fact subject to censoring and we refer to the inappropriate use of sNB_n^{cohort} as "naive". As done previously, all simulations are based on 5,000 replicates of data sets that contain between 1,000 to 20,000 observations for which the rate is varied between 0 and 0.07 in the independent exponential censoring distribution.

6.6.1 Censored Observations Removed from Analysis

The simplest approach to dealing with censored observations is to simply ignore them. The so-called "complete-observation" analysis removes observations with incomplete follow-up from estimation entirely, and hence the effective sample size is smaller than the study cohort. In terms of the data collected, observations with $\Delta_i = 0$ and $Z_i < t_c$ are excluded from analysis. This approach omits both would-be controls and would-be cases.

Table 6.2 summarizes the average bias of the naive point estimates, as a percentage of the unknown estimand, across varying amounts of censoring. Because the observed bias was essentially the same across the sample sizes considered, results from datasets with 5,000

Censoring	N_{eff}	sNB	TPR	FPR	ρ
			True Value		
		0.42	0.76	0.03	0.14
			% Bias		
0.00	5000	-0	-0	0	0
0.01	4627.8	2	0	0	2
0.02	4283.6	4	1	0	3
0.04	3671.4	8	1	-0	7
0.07	2916.4	14	2	0	13

Table 6.2: Bias (% of estimand) of naive point estimators that exclude all censored observatons from analysis, across varying amounts of censoring. The effective sample size, N_{eff} is the average number of uncensored observations used to calculate the estimates.

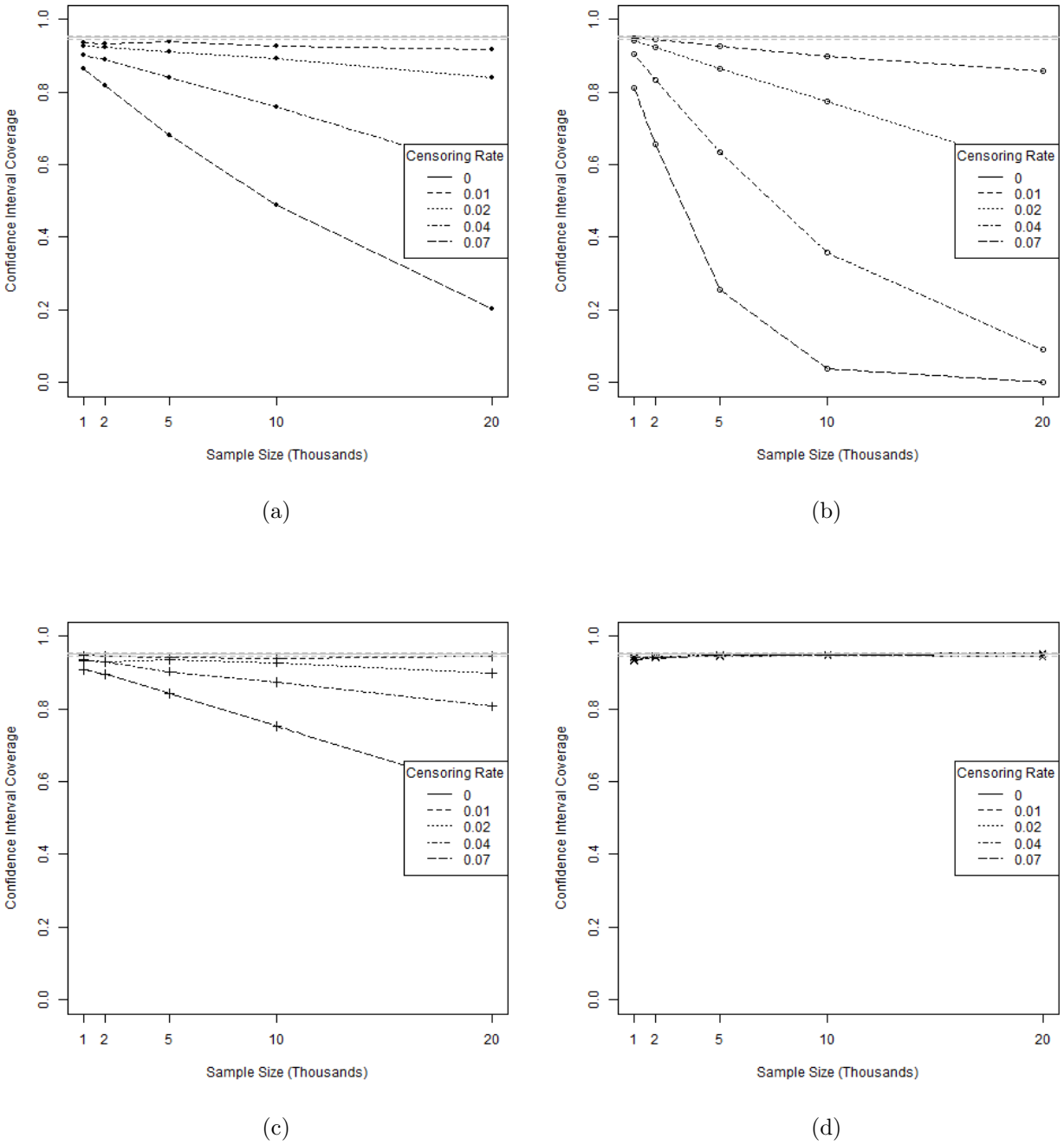
observations each are presented as representative of the asymptotic bias. The average number of observations contributing to estimation is reported under N_{eff}, the effective sample size. Under no censoring all 5,000 simulated data sets contained 5,000 complete obervations; when the censoring rate was 0.07, this dropped to an average of 2,916.4 contributing observations in each simulated dataset. This naive approach incorrectly omits some would-be cases and some would-be controls. The naive estimators for sNB, ρ and TPR are asymptotically biased when outcomes are censored, and increasingly so as the amount of censoring increases. However, the naive estimates of the false-positive rate are unbiased and appear to behave well.

In the context of the data generating mechanism used in this simulation, the observed results can be qualitatively explained. First we note that the probability of a would-be control, i.e. $T_i > t_c$, being incomplete in terms of case-control status, is constant and equals $Pr(C < t_C)$. Hence, the subset of uncensored controls contributing to estimation of the false-positive rate in this naive analysis is actually representative of the cohort of controls. Hence, the naive estimate of the false-positive rate should be unbiased as was observed.

However, the probability of a would-be case, *i.e.*, $T_i \leq t_c$, being censored is dependent on the would-be event time and equals $Pr(C < T_i)$, which is monotone increasing in the individual event time. Among would-be cases, the rule studied in this simulation begins to make errors in intervention-assignment increasingly as the would-be event time approaches the landmark time, see Figure 6.1. Greater censoring of such mis-identified would-be case observations leads to an inflated estimate of the true-positive rate, as observed in the simulation results. As there are more would-be controls in the population, the true outcome probability is 14%, each of whom has a greater likelihood of being censored than a would-be case, we reason that the set of observations excluded from this naive analysis contains more controls than cases. This would lead to an over-estimate of the outcome probability which is demonstrated by this simulation. The combination of a high outcome probability and a high true-positive rate combines to produce the observed positively biased naive estimator of standardized net benefit.

Figure 6.3 presents the coverage of 95% confidence intervals based on the naive use of the cohort estimators by excluding censored observations from analysis. For a given cohort size, confidence intervals were calculated for each simulated data set using the effective samples size N_{eff}, the number of uncensored observations, which varied across data sets. As anticipated from the results on asymptotic bias of the naive estimators, for all but the uncensored scenario, the coverage of the naive estimators for sNB, TPR, and ρ rapidly goes to zero as sample size increases and with increasing amounts of censoring. In addition to the asymptotic bias of the point estimates, asymptotic bias of the variance estimates is also expected to contribute to poor coverage properties. The behavior of these naive estimators is unacceptably poor in the scenarios considered and would be expected to be so generally. On the other hand, coverage of the unbiased false-positive rate estimates, as discussed above, behave well in this particular approach to ignoring censoring. More detailed results for each rate of censoring considered appear in Appendix B.10.

6.6.2 Censored Observations Assumed to be Controls



(a) (b)

(c) (d)

Figure 6.3: Asymptotic coverage of 95% confidence intervals for naive estimators of (a) sNB_0, (b) ρ_0, (c) TPR_0 and (d)FPR_0 that exclude censored observations from analysis.

Censoring	sNB	TPR	FPR	ρ
		True Value		
	0.42	0.76	0.03	0.14
		% Bias		
0.00	-0	-0	0	0
0.01	-26	0	24	-6
0.02	-54	1	46	-11
0.04	-114	1	86	-22
0.07	-219	2	135	-34

Table 6.3: Bias (% of estimand) of naive point estimators that assume all censored observations are controls, across varying amounts of censoring.

Another approach to naively analyzing censored data is to assume all incomplete observations are controls and then use a cohort estimator that does not account for censoring. In terms of the data collected, observations with $\Delta_i = 1$ and $Z_i \leq t_c$ are correctly identified cases and all other observations are assumed to be controls. This approach misidentifies, as controls, would-be cases for which the event time would have been observed before the landmark time, $T_i \leq t_c$, had the subject not been previously censored, i.e., $C_i < T_i$.

Table 6.3, summarizes the average bias of the naive point estimates, as a percentage of the unknown estimand, across varying amounts of censoring. Because the observed bias was essentially the same across the sample sizes considered, results from datasets with 5,000 observations each are presented as representative of the asymptotic bias. In this approach, all observations are used and the effective sample size is the same as the full cohort. All naive estimators of $\text{sNB}_{cohort,n}$ and its three constituents are asymptotically biased, increasingly so as the amount of censoring increases.

As this naive approach incorrectly counts some would-be cases as controls, we would expect that estimates of the outcome probability would be low and increasingly so as the

amount of censoring increases. This is clearly observed. As previously discussed in the context of the estimator that naively omits censored observations from analysis, this simulation scenario has censored would-be cases with greater event times and more frequent non-intervention classifications than uncensored cases. These otherwise censored cases are treated as controls and hence to not contribute to estimation of TPR. This again leads to overestimation of the true-positive rate by the naive application of $\text{TPR}_{cohort,n}$. Though these otherwise censored cases are more likely to be assigned no intervention by the decision rule than observed cases, it is reasonable to expect that they are more likely to be assigned intervention than the true would-be controls; recall $\text{TPR}_0 = 76\%$ and $\text{FPR}_0 = 3\%$. Thus, their treatment as controls, and inclusion in the estimation of FPR, would likely lead to an overestimation of the false-positive rate, which is observed. The combination of a low outcome rate and a high false-positive rate, with an only slightly high true-positive rate, combines to yield a naive estimator of standardized net benefit that is quite biased downward.

Figure 6.4 presents the coverage of 95% confidence intervals based on the naive use of the cohort estimators that treat all censored observations as controls. As anticipated from the results on asymptotic bias of the naive estimators the coverage of the naive estimators for sNB and its three consituents rapidly goes to zero as sample size increases and with increasing amounts of censoring. In addition to the asymptotic bias of the point estimates, asymptotic bias of the variance estimates is also expected to contribute to poor coverage properties. The behavior of these naive estimators is unacceptably poor in the scenarios considered and would be expected to be so generally. More detailed results for each rate of censoring considered appear in Appendix B.10.

6.7 Alternate Approaches

In this Chapter we provided asymptotic distribution theory and inference for an estimator of standardized net benefit that employs Kaplan-Meier estimates of survival at the landmark time. Key elements of this approach were proposed by Heagerty et al. (2000) in the context of time-dependent ROC curves and this approach was promoted by Vickers et al. (2008) as

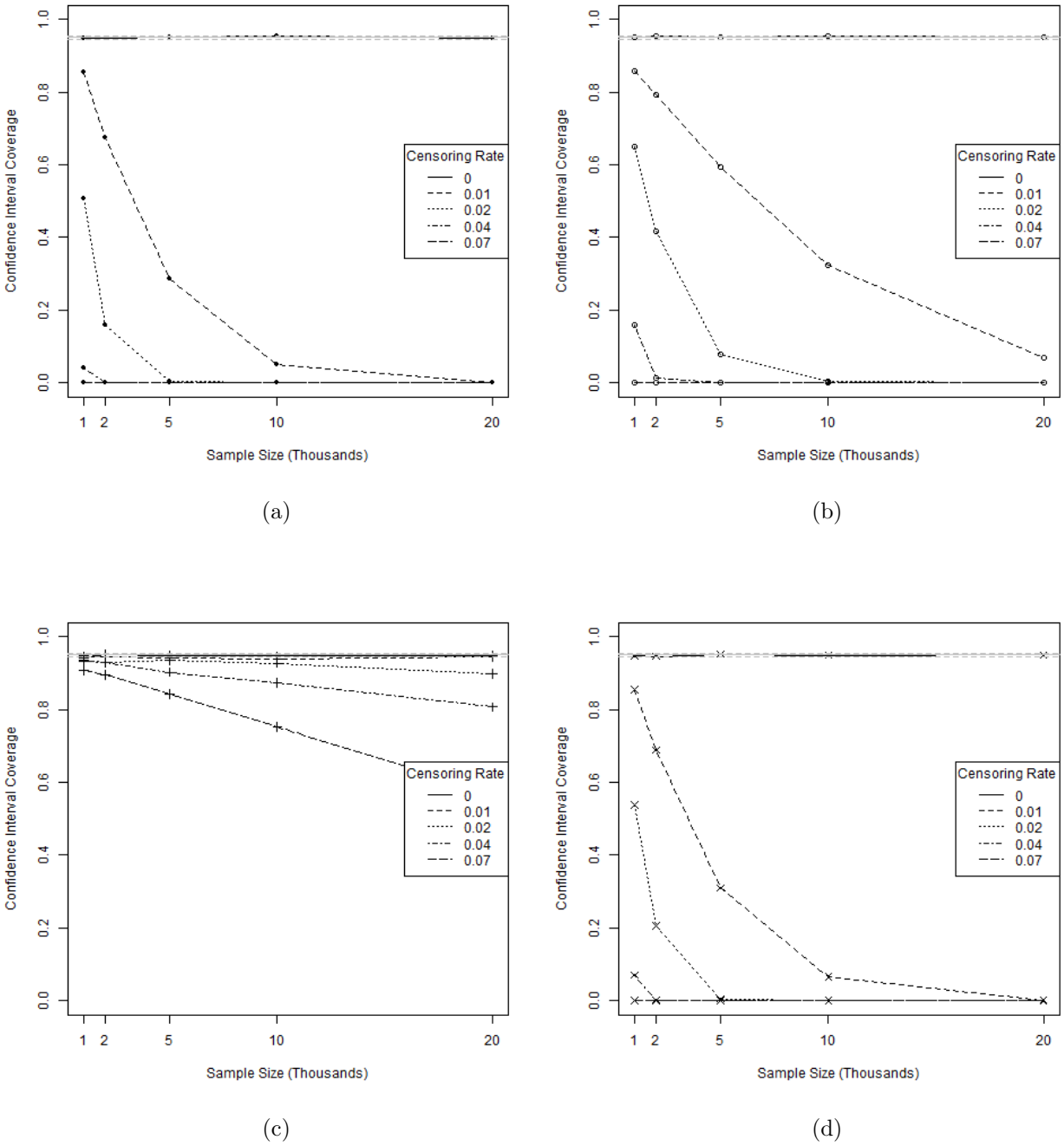


Figure 6.4: Asymptotic coverage of 95% confidence intervals for naive estimators of (a) sNB_0, (b) ρ_0, (c) TPR_0 and (d)FPR_0 that assume censored observations are controls.

a means of accounting for censoring when estimating net benefit. The use of Kaplan-Meier estimators has the benefit of being a well-accepted methodology supported by standard statistical software packages. Other approaches are possible.

To start, any method for estimating a distribution function at a point, in the presence of censoring, could be employed instead of Kaplan-Meier estimators. If such estimators were additionally asymptotically linear, then the delta method could also be used to establish asymptotic distribution theory and inference for the resulting estimators $\text{sNB}_n^{censAlt}$. The influence function of $\text{sNB}_n^{censAlt}$ would be obtained by simply replacing influence functions of the Kaplan-Meier estimators in Equation 6.1 with those of the alternate estimator. Of particular interest are estimators that impose less stringent assumptions on the censoring mechanism than do Kaplan-Meier estimators.

For failure time outcomes, such as time to a cardiovascular event, the derived binary disease-status has a dependence on the choice of landmark time that could be inherently of interest. This has inspired much research in assessing biomarker potential through time-varying ROC curves (Etzioni et al., 1999) and related concepts. As pointed out by Cai et al. (2006), cumulative incidence-based and instantaneous definitions of time-varying classification rates have arisen, and the cumulative rate can be obtained as an integral of the instantaneous. This dissertation focuses on (1) binary clinical decision rules that determine whether or not an intervention should be recommended and (2) assessments of the clinical utility of employing such rules when the current or would-be disease status is also binary. In this setting, cumulative incidence-based measures of time-varying true- and false-positive rates, evaluated at one particular time (the landmark time) are directly relevant. If the decision rule is derived from a dichotomized continuous measure, such as a risk score and fixed threshold, then any estimate of the time varying ROC curve, evaluated at the landmark time and threshold, could be used to assess the true- and false-positive rate constituents of net benefit. The assumptions on the censoring mechanism made by time-varying ROC methods that accommodate censored outcomes would be inherited by the resulting estimator. The approach considered in this chapter is one such example.

In the context of time-varying ROC curves, the Kaplan-Meier based estimators can produce single ROC curves that exhibit non-monotonic true- or false-positive rates. Heagerty et al. (2000) presented a second method for estimating time-varying ROC curves that corrects for this anomaly by directly modeling a joint survival curve. Though non-monotonicity is not a concern for estimating the net benefit of a single decision rule, the second method additionally allows the censoring distribution to depend on baseline covariates. Using the bivariate survival based estimator, evaluated at a single landmark time and threshold, would produce and estimator of net benefit with the same relaxed censoring assumptions. Other techniques that have been established for sensitivity and specificity of biomarkers, such as Cai et al. (2006), may be similarly applicable to estimation and inference for the net benefit of a single decision rule for managing occurrence of an outcome prior to a landmark time.

6.8 Summary

In this section we presented one method, employed in a published analysis, for estimating sNB when the clinical outcome is subject to censoring. Influence functions for the RAL estimators were calculated for empirical estimators of sNB as well as the true and false-positive rates and overall outcome probability. The influence functions were then used to define empirical estimators of the limiting variances and empirical approximations of asymptotically correct Wald confidence intervals. the frequentist properties of the point estimates, limiting variance estimates, and confidence intervals were evaluated in a small simulation study.

Chapter 7

CONCLUDING REMARKS

7.1 Summary

In this part of the dissertation, we have focused on establishing analytic inference for estimators of net benefit. As a measure originating in the discipline of decision theory, the foundations for statistical inference have not been previously established. We have argued that analytic expressions of variance surpass descriptions of uncertainty captured through bootstrapped estimator distributions because they provide insights on sources of variability and support the design of studies for assessing the utility of a clinical decision rule. The setting for inference considered throughout this dissertation is validation of pre-specified clinical decision rules. All considered estimators of net benefit are regular and asymptotically linear. Our primary technique is to calculate the influence function for each estimator and use distribution theory provided by the Central Limit Theorem, from which asymptotically correct Wald confidence intervals may be constructed.

We introduced two visual representations of standardized net benefit and used them to illustrate various interpretations of the measure. From the perspective of biomarker development, we oriented our investigation of the net benefit of clinical decision rules to the set of possible classification accuracies within a designated clinical context and often restricted attention to regions corresponding to positive net benefit.

We first address inference for estimators of net benefit from cohort studies. With a formula for the asymptotic variance in hand, we were able to explore how the variability of estimates of sNB relied on key features of the decision rule and clinical setting. Additionally, the natural decomposition of the asymptotic variance into components due to estimation of each of the three constituents in the measure, the outcome probability and the true and

false-positive rates of the rule, reveals the efficiency gain from making use of the outcome probability, when known. Considering the variability inherent in estimating net benefit over the set of all possible true- and false-positive rates of a rule, within a specified clinical context, gave insight into quick "back of the envelope' means of establishing the difficulty in assessing the potential utility of a rule over regions of minimal net benefit or regions of plausible classification accuracies. This led naturally to using the analytic formulas for variations of basic sample size calculations.

In earlier phases of biomarker development, the specific clinical application may not yet be identified or cost-benefit trade-offs for an intervention may not yet be conducted or consensus only established within a range of plausible trade-offs. For rules derived from risk models, decision curves and their standardized counterparts, relative utility curves, have been proposed as a means of assessing the potential net benefit over the family of rational risk rules indexed by varying the high-risk threshold. We present a numerical approach to constructing confidence bands for relative utility curves, and the difference between two relative utility curves, and provide the necessary analytic expressions of limiting variance and covariance.

Further support of earlier phase biomarker studies incorporating measures reflecting the possible clinical applications is provided through studying estimation of net benefit from case-control studies. We establish inference for estimators from three variants of unmatched case-control studies. The implications for study design are examined through 1) identifying an optimal control-to-case ratio, in studies where the total case- control sample size is fixed, and 2) examining the efficiency gain from additional controls in studies including all available cases along with a flexible number of controls.

The work of this thesis primarily focused on settings in which the clinical outcome was measured on all subjects and consequently observations could be distinguished as cases or controls. In the final chapter, we apply our techniques to establish inference for one particular estimator of net benefit, employing Kaplan-Meier curves, from cohort data on which the outcome status is independently censored.

7.2 Discussion

Net benefit is a weighted difference of the true- and false-positive rate of a decision rule and further, when the outcome probability is known, so is the weight. When we us a weight equal to 1, as if the control-to-case benefit ratio exactly counterbalanced the odds of the clinical outcome, then the net benefit of rule R equals the classic Youden's index (Youden, 1950). Hence, the inference results established for estimation of $\text{sNB}_n^{cohort,\rho}$ or $\text{sNB}_n^{cc,\rho}$ also apply to Youden's index and results for estimation of $\Delta_{\text{sNB}_n^{cohort,\rho}}$ or $\Delta_{\text{sNB}_n^{cc,\rho}}$ apply to a difference in Youden's index, between two decision rules. In the case of a single threshold, dividing risks into high and low categories, the Net Reclassification Index (NRI), proposed for evaluating the impact of a new rule over an old rule, is equal to the difference in Youden's indexes between the new and old rules (Van Calster et al., 2013; Pencina et al., 2011, 2008). Hence, the same results also provide inference for this particular realization of NRI.

The quick adoption of NRI by clinical researchers, including publications in high-impact clinical journals, has been well noted (Leening et al., 2014; Pepe et al., 2014) and continues to remain popular. An estimated 3,642 citations of the seminal paper that introduced net reclassification (Pencina et al., 2008), averaging almost 490 citations per year starting in 2011, was determined on July 29, 2017 through Google Scholar. Further, NRI was introduced with an accompanying z-statistic for testing the null hypothesis of zero net reclassification and is commonly reported as a point estimate accompanied by a p-value. However, the statistical properties of the test are unsatisfactory, in part stemming from issues with the variance formula (Kerr et al., 2014). Utilization of the inferential results established in this dissertation could provide substantial contribution to improved assessment of NRI by clinical researchers.

Distribution theory for the empirical estimator of net benefit, from a cohort study, could have been established by application of the delta method to the joint set of empirical estimators of the true- and false-positive rates and outcome probability. Details are available in Appendix B.2. We employed influence function techniques, to establish the same results, be-

cause of their versatility in establishing the asymptotics of differences in net benefit between two rules and the construction of approximate confidence bands. Another utility of influence functions is in establishing the joint asymptotics of two or more asymptotically linear estimators. For example, if differences in net benefit, between a baseline rule and two or more rules involving novel biomarkers were estimated, the joint asymptotic distribution, reflecting correlation between the performances of the multiple rules, could be used to calculate a joint confidence region for the estimates. This would be quite similar to the calculations used to support construction of confidence bands around an estimated decision curve.

It is well understood that good coverage performance by Wald confidence intervals for estimates of simple proportions generally require larger samples as the proportion becomes more extreme. This has been observed throughout the simulation studies conducted herein for two constituents of net benefit, the true- and false-positive rates, which becomes particularly relevant when these better the performance of the rule (TPR near 1 or FPR near 0). In such cases, any of the small-sample approaches available could be employed instead (Newcombe, 1998). Generally, coverage of the proposed confidence intervals for estimates of net benefit, a composite measure, was good in the scenarios examined. From small sample sizes with extreme true and/or false-positive rates, the coverage for estimates of net benefit may be brought into question. If the outcome probability is known, the small sample methods of Decrouez and Robinson (2012) may be of interest. These approaches, though expressly stated as applicable to estimates of net benefit, apply to a weighted sum of true and false-positive rates, where the weights are constant. As the weight used in net benefit generally relies on the outcome probability, which is generally unknown, their techniques are of limited applicability to conducting inference on net benefit.

Net benefit is established as a measure of clinical utility for a dichotomous decision rule. Thinking in terms of risk, many clinical decision making contexts have fairly agreed upon concepts of high-risk and low-risk patients, with well defined action and inaction, but a fair portion of the population naturally falls in between the two, in a mid-risk level. For example, this is an issue for the management of incidental lung cases. Extensions of net benefit to

this case could be quite useful. A derivation in terms of expected utilities is presented in Appendix B.11 and is analogous to the net benefit studied within, though now reliant on three ratios of differences in utility instead of two. Ostensibly, similar approaches to establishing inference for estimation of this extended measure of net benefit could be undertaken, though with expected increase in complexity.

7.3 Future Research

Many important clinical settings involve management of a possible clinical event that could occur at a future time. In Chapter 6 we examined an approach to evaluating the net benefit of a rule for managing the occurrence of an outcome prior to a landmark time that accommodated the practical issue of censoring. Methodology developed for time-varying ROC curves has interesting potential for defining net benefit as a time-varying concept. Though such an extension may not apply to validating a pre-specified clinical decision rule, it could be quite interesting for assessing the clinical potential of biomakers at earlier stages of the development pipeline.

Within this dissertation, we have only considered variations of case-control designs in which the case and control subsamples are each representative of the respective population of would-be cases or would-be controls. In practice, many studies are conducted on a simple sample from the case population and a sample of controls selected to have the same distribution of a variable as cases. In early phases of biomarker development, when simple measures of discrimination are evaluated, so-called matched designs are often used to make the cases and controls comparable, i.e., to reduce confounding. Slightly later phases of biomarker development may employ stratified case-control samples in which both the cases and controls are sampled according to a distribution of a stratifying variable that differs from that of cases and/or controls in the population of interest. Matched and stratified designs, when coupled with appropriate analyses, generally offer efficiency gains.

Quite generally, estimation of population-level parameters, from studies with matched or stratified designs, is more complex and may require more assumptions than their unmatched

counterparts. Further, whether and when efficiency gains follow from matching must be explored explicitly for each estimand. We are unaware of any such results for estimators of net benefit and are interested to explore this topic further. However, methods for estimation of two constituents of net benefit, true- and false-positive rates, from matched samples provide one reasonable starting point. Such methods include estimating a covariate adjusted ROC curve (Janes and Pepe, 2008; Pepe et al., 2013). Another starting point is the continued employment of general influence function techniques. Results for obtaining an influence function for an estimator from a matched case-control sample (van der Laan, 2008) or for a general two-phase sample (Rose and van der Laan, 2011b), from that for a cohort sample could be the basis of establishing inference for net benefit from studies with either design. The matched approach requires known strata-specific outcome probabilities and the two-phase approach assumes known sampling weights.

A key premise for the work presented within this part of the dissertation is pre-specification of the clinical decision rule being assessed in a validation context. At the time of decision rule development, researchers are typically interested in also assessing performance of the rule. An independent 'testing' set is often not available and performance measures, such as net benefit, are generally expected to be correlated with the criteria used to define the rule. Cross-validation techniques are often used to protect from the so-called over-optimism anticipated when assessing the performance of a rule from the same data on which it was 'fit'. Inference for cross-validated estimates of the absolute prediction error of a fitted regression model (Tian et al., 2007) and a general approach to inference for n-fold cross-validated estimates of performance are available (Hubbard et al., 2016). The latter approach applies to settings in which the estimators on each "fold" are asymptotically linear and employs influence function techniques as have been used throughout his dissertation. Results from this dissertation could be used in applying this method to give inference for n-fold cross validated estimates of the net benefit of a rational risk rule when the underlying risk model was fit on the same data. There are subtleties to cross-validation that are not always appreciated. For example, the estimand in the cross-validation setting is the average performance of the rules

fit to the n-folds, which is not exactly the same as the estimand in the validation setting, which is the performance of a single pre-specified rule. A commentary exploring these differences and illustrating inference for cross-validated estimates of net benefit would be a useful contribution to the practical evaluation of decision rules using clinical utility measures.

Throughout this part of the dissertation, we have used a general expression of net benefit that is well-defined for a general clinical decision rule, while pointing out the benefits to rules derived from risk models. The rational properties of clinical risk rules relies on the interpretation of the risk score as a probability, among individuals with similar predictors, of the clinical outcome. The concordance between predicted risks and observed outcome events can be examined empirically, such as overall in the population or by percentiles of predicted risk, and is described as the calibration of a risk model in a particular population. Often, a defined risk model will be re-calibrated to the observed population in a new study cohort, prior to assessment of its clinical utility for that population. This context lies somewhat in between assessing net benefit on the same data as complete model-fitting, as just discussed, and independent assessment of a pre-specified risk rule, as considered in this dissertation. A useful extension to our presented work would be to account for additional variability and possible bias introduced by re-calibrating a risk rule at the time of evaluation.

Numerous opportunities for applying more general nonparametric estimation and targeting techniques to the task of assessing the potential net benefit of a decision rule follow from the foundational contexts addressed in this the dissertation. For example, more sophisticated re- calibration techniques that use the pre-specified risk score as a predictor for a new risk model, that simultaneously achieves calibration in the population could be accomplished by applying the methods of Brooks et al. (2012). Another opportunity lies with relaxing the often unreasonable assumption of independent censoring required when using Kaplan-Meier estimates to estimate net benefit in the presence of censored outcomes. Yet another opportunity arises from estimation of net benefit within a two-phase sampling context and continuous phase-one clinical variables. As commented in Section 5.2, this realistic scenario requires estimation of the true- and false-positive rates as a function of phase-one clinical

variables. To do so in a model-free fashion would require local smoothing that would necessitate correction to regain asymptotic linearity which requires techniques similar to those employed in the next part of this dissertation (van der Laan and Rubin, 2006).

Part II

EFFICIENT ESTIMATION OF MARGINAL ADDITIVE INTERACTION FROM A STUDY NESTED WITHIN A TRIAL

Chapter 1

BACKGROUND

1.1 Gene-Environment Interaction Studies Nested Within Trials

An increasing number of scientific and clinical studies aim to establish whether individual variables modify the effects of interventions. Of particular interest to statistical genetics is the identification of gene-treatment and more generally gene-environment interactions. The success of genome-wide association studies in identifying susceptibility variants is contrasted with the lack of identified and reproduced gene-environment interactions. In response to these challenges, a "Gene-Environment Think Tank", sponsored by the National Cancer Institute, convened to discuss how to best pursue and include gene-environment interactions in cancer epidemiology. Among other points, the resulting report underscored the importance of pairing methods with scientific goals, gave attention to nested case-control study designs, and highlighted the importance of efficient methods. (Hutter et al., 2013). The methods presented in this part of the dissertation are developed to support characterizing the joint effects of an intervention and a candidate gene, or other biomarker, when the implications of differential effects for public health decisions are of central interest. The statistical focus is on the efficiency gain achieved by leveraging independence between treatment and any baseline biomarker, when the biomarker is only measured on a subsample of a randomized trial cohort.

Because randomized trials are a significant effort, there is a great incentive to conduct secondary studies on the high-quality data collected. Trials that also collect biological samples at baseline further facilitate additional studies by enabling the measurement of covariates not required for primary evaluation of the intervention. For example, the US National Human Genome Research Institute funded a consortium, the Genomics and Randomized

Trials Network (GARNET), to utilize data and samples from previously executed randomized, prospective trials to investigate genetic determinants of treatment response (Bookman et al., 2013). Bioassays are often expensive and consume a portion of the limited quantities of biological samples collected at baseline. Consequently, the blood-based biomarker or genetic information is often only measured on a subsample, typically a case-control or matched case-control subsample, of the full trial cohort. This produces a two-phase design comprised of phase-one variables, which include the outcome and inexpensive or previously collected covariates, measured on all participants, and the "expensive" phase-two biomarkers, measured only on the selected subsample.

Two examples arose within the hormone therapy trial component of the Women's Health Initiative (WHI). The trial was terminated early due to observed elevated risk of adverse events such as stroke and venous thromboembolism in both treatment arms and showed either no benefit from estrogen or elevated risk under estrogen and progestin therapy for the primary outcome of cardiovascular disease (CVD) (WHI Writing Group, 2002). These risks, in addition to known associations with breast cancer and dementia, were determined to outweigh observed reductions in fractures, colorectal cancer, and diabetes. Subsequent analyses were conducted to evaluate possible high-risk subgroups defined by blood biomarker levels (Wassertheil-Smoller et al., 2003). Additionally, through GARNET, a genome-wide interaction study (GWIS) was undertaken with data collected within WHI studies. Data from both the observational cohort and randomized trial participants were used to generate and test hypotheses regarding the modification of hormone therapy effects by common genetic variants on the risks of cardiovascular disease, stroke, venous thromboembolism and incident diabetes (GARNET-WHI).

A third example took place in the context of the North Central Cancer Treatment Group (NCCTG) N9831 trial to study the benefits of trastuzumab among women receiving adjuvant chemotherapy for HER2- positive breast cancer. A secondary study used surgical tissue samples from eligible N9831 participants to classify tumors according to intrinsic subtype and test the hypothesis that trastuzumab benefit varied across intrinsic subtypes (Perez

et al., 2017).

1.2 Established Estimators of Gene-Environment Interaction

Logistic Regression

In earlier stages of biomarker development, when studies often employ case-control samples, including an interaction term in a logistic regression model provides a convenient measure of interaction on the odds-ratio scale that has a prospective interpretation. The two variables of the interaction term are gene and treatment, or more generally, biomarker and exposure. When additional covariates are included in the model, then interpretation of the interaction parameter is conditional rather than marginal. Routine logistic regression does not accommodate incorporating structure of the covariate distribution. In particular, the known independence between treatment assignment and all other baseline covariates, guaranteed by randomization, is not leveraged by this approach.

Case-Only Estimator and Extensions

Since the introduction of the case-only estimator, which is well known for its efficiency gains over the case-control counterpart that does not make use of independence (Piegorsch et al., 1994), much effort has been devoted to developing methods that can exploit so-called gene-environment independence for efficiency gain. The case-only estimator provides a measure of interaction on the odds-ratio scale, relies on a rare-outcome assumption, does not accommodate additional covariates (Mukherjee and Chatterjee, 2008) and has proven sensitive to the gene-environment independence assumption (Albert et al., 2001).

The case-only estimator, which only provided estimation of the interaction, was extended by Umbach and Weinberg (1997) to provide estimators of all logistic regression parameters when the exposures are categorical. Chatterjee and Carroll (2005) introduced a semiparametric framework that (a) further extends the covariate structure to allow independence conditional on stratifying population covariates, (b) is free of the rare outcome assumption,

and (c) is applicable to case-control data. When gene-environment independence is assumed the covariate distribution is not constrained and hence Prentice and Pyke (1979) cannot be used to justify a prospective interpretation. Consequently, methods building upon logistic regression to exploit known independence, are limited to retrospective analyses.

Specific to the two-phase setting, Chatterjee and Chen (2007) extended profile likelihood methods to give an odds-ratio scale estimate of interaction for binary outcomes, and Dai et al. (2009) provided means to analyse interactions on various scales, for both continuous and binary outcomes, using generalized linear parametric models.

Interactions on the Additive Scale

Methods for testing interactions on the additive scale have been developed (Han et al., 2012; Tchetgen Tchetgen et al., 2014), but their direct estimation have received less attention than conditional quantities on other scales. Existing methods for estimation of additive scale interactions often target the relative excess risk due to interaction (RERI) as introduced by Rothman (1986). In large part, this choice is driven by convenience; RERI can be constructed from relative risk estimates which, when outcomes are rare, can be approximated by odds-ratios estimated from case-control data. To address concerns that incorrect specification of logistic or linear odds outcome models, including covariates in addition to the exposures, can lead to invalid inference (Skrondal, 2003; Greenland, 1993), VanderWeele and Vansteelandt (2011) introduced an inverse probability of treatment weighted estimator of RERI. The approach transfers the model specification from the outcome to the exposures and relies on a rare outcome assumption to justify estimation of the weights for the population using the controls only, in addition to approximating relative risks by odds ratios.

Union Models

All methods based on unsaturated parametric regression models are subject to model misspecification. The application of multiply robust union models to the estimation of interaction on various scales (Vansteelandt et al., 2008) offers one approach to addressing this risk,

by affording multiple chances for inference based on parametric models to be correct. Under known conditional independence, their estimators of conditional additive scale measures of interaction enjoy quadruple robustness whereas they are only triply robust without this knowledge.

1.3 Targeted Learning

Targeted learning is a very general approach to supporting scientific research through sound, science-driven statistics at stages where inference is of primary interest. A central tenet of this approach is to ensure that the assumptions of adopted estimation approaches reasonably apply to the unknown data-generating mechanism. In particular, parametric models are not typically used to define the parameter of interest. The starting point for targeted learning is flexible non-parametric or ensemble methods that may be modified to reflect knowledge about the data generating mechanism.

In the paradigm of targeted estimation, emphasis is placed on formulating quantities that are directly related to the primary scientific inquiry, rather than adapting the analysis to parameters of convenience. Estimable quantities are viewed as functions of probability distributions. The statistical model, over which estimation must discriminate, is a collection of possible data-generating mechanisms and each candidate distribution determines a particular value of the quantity of interest.

For a given estimand, targeted estimation is generally performed in two stages. The first stage defines an initial substitution estimator. This requires consistent estimators for features of the data generating mechanism needed to calculate the quantity of interest. The second stage involves correcting for bias in the initial substitution estimator, with the goal of recouping asymptotic linearity and possibly also asymptotic efficiency. The estimators needed in the first step are often required to meet minimum rates of consistency that are slower than $\sqrt{n}$. This facilitates the use of more flexible and nonparametric modeling approaches which can help mitigate model misspecification greatly, if not entirely.

In addition to general results for targeted estimation from case-control samples with

known prevalence (van der Laan, 2008) and more general two-phase samples (Rose and van der Laan, 2011b), a complete program for targeted learning is detailed in Rose and van der Laan (2011a).

1.4 Goals for Estimating Effect Modification from a Study Nested in a Trial

From a public health perspective, population-level, prospective quantities on the additive scale arguably provide the most pertinent measures of interaction (Skrondal, 2003; VanderWeele and Vansteelandt, 2011; Han et al., 2012). The relevance of the bivariate exposure-specific mean, in which the conditional mean-outcome has been averaged over the population distribution of non-exposure covariates, to this problem has been discussed in the literature (Rose and van der Laan, 2014a; VanderWeele and Vansteelandt, 2014; Rose and van der Laan, 2014b). In addition to providing direct estimation of a marginal measure of additive excess risk due to interaction (AERI), targeted learning approaches can reduce the reliance on parametric models. Here we present an approach that does not explicitly rely on parametric modeling for either defining the measure of interaction or its inference.

In the context of a study nested within a randomized trial, including additional covariates is of interest for increased precision and to ensure that the interaction estimand has an appropriate population interpretation by accounting for residual confounding in the interaction effect. Independence with baseline covariates, particularly with the phase-two biomarkers, is guaranteed by randomization of the intervention.

Freedom from the reliance on parametric models for the biomarker distribution and treatment-biomarker specific outcome probability distributions, conditional on baseline covariates, is a contribution of this approach to current interaction methodology. In this work, we implement a targeted learning approach to estimating such a measure of interaction and extend established two-phase results to reflect the constraint of known gene-environment independence. Inference for the proposed measure of additive interaction, with respect to the experimental data collected in a two-phase sampling scheme, is our focus with significant interest on efficiency gains following from utilizing the known independence.

Chapter 2

ADDITIVE EXCESS RISK DUE TO INTERACTION

2.1 Measures of AERI

The first goal is to define a nonparametric marginal measure of additive interaction between the treatment X and a binary biomarker B (e.g., high-low risk assignment based on a biomarker signature, presence/absence of a genetic variant) relative to a binary outcome Y. Within each strata defined by B and X we denote the probability of outcome, marginally over any baseline covariates, by $\mathrm{Q}_{jk} := \Pr(Y = 1 \mid B = j, X = k)$. The corresponding measure of additive excess risk due to interaction can be expressed as $\mathrm{AERI} := \mathrm{Q}_{11} - \mathrm{Q}_{10} - \mathrm{Q}_{01} + \mathrm{Q}_{00}$. An extension of AERI that allows for adjustment for baseline covariates W, while retaining a population-level interpretation, P-AERI, is naturally defined by conditioning the above outcome probabilities and then averaging over covariates. In symbols, we define:

$$\text{P-AERI} = \mathbb{E}_W \left[\mathrm{Q}_{11}(W) - \mathrm{Q}_{10}(W) - \mathrm{Q}_{01}(W) + \mathrm{Q}_{00}(W)\right] \tag{2.1}$$

where $\mathrm{Q}_{jk}(w) := \Pr(Y = 1 \mid B = j, X = k, W = w)$ are the outcome probabilities conditional on the biomarker-treatment levels and covariate value (Rose and van der Laan, 2011b).

This parameter retains a marginal interpretation for the population represented by the full cohort and is a population-averaged measure of the additive excess risk due to interaction that adjusts for baseline covariates. Further, we note that P-AERI is well-defined for any probability distribution of (Y, B, X, W), where Y, B, X are binary and W is a vector of random variables of any type, satisfying basic measure theoretic and finite moment conditions. P-AERI is the target of inference throughout this part of the dissertation.

2.2 Two-Phase Data Structure

Data Structure

Consider a study designed with a second phase of measurement conducted on a subsample of the cohort from a randomized controlled trial. For all trial participants, the outcome Y, randomized treatment assignment X, and vector of baseline covariates W is recorded. On the selected subsample, indicated by $\Delta = 1$, an additional biomarker B is measured. Each experimental observation can be expressed $O^E = (\Delta B, \Delta, Y, X, W)$ and the analysis sample consists of n independently and identically distributed observations O_i^E, collected on $i = 1, \ldots, n$, trial participants.

Throughout this work, Y, X, and B are binary random variables and W may consist of one or more variables that are either continuous or discrete. Notational abbreviations of E and F will be used throughout to differentiate data observed in the experiment just described, O^E, from the full data $O^F = (Y, B, X, W)$ that we would ideally have obtained. All observations collected within the study are assumed to be independent and identically distributed according to some true, but unknown, data-generating mechanism (P_0^E or P_0^F).

Statistical Models and Assumptions

Highlighting the nested structure of the design places analysis of the experiment in a missing data framework. We assume that the missingness is at random, i.e., $\Pr(\Delta = 1 \mid Y, B, X, W) = \Pr(\Delta = 1 \mid Y, X, W)$, known by design, and results from simple Bernoulli sampling of the phase-two subsample.

Other than standard finite moment and measure theoretic properties, we make no structural assumptions on the conditional mean outcome functions $\mathrm{Q}_{jk}(w)$ or on the distribution of covariates W. The bivariate exposure distribution factorizes as: $\Pr(B = j, X = k \mid W = w) = \Pr(B = j \mid X = k, W = w)\Pr(X = k \mid W = w)$ and the treatment assignment mechanism, $g_X(k; w) := \Pr(X = k \mid W = w)$, is known by design. The conditional distribution of the biomarker status is also left unspecified.

Our use of the term statistical model, denoted $\mathcal{M}$, refers to a set of probability distributions over which estimation must discriminate the true parameter value. Implicit in our methods is the assumption that the unknown truth lies in the statistical model: $P_0 \in \mathcal{M}$. We write $\mathcal{M}^E$ and $\mathcal{M}^F$ to distinguish the sets of data-generating mechanisms relevant to estimation based on experimental observations attained in the nested study from those relevant to estimation based on full observations that would have been obtained if biomarkers had been measured on all trial participants. These models may reflect the independence assumption and we use the superscript $\perp\!\!\!\perp$ to indicate the restriction to probability distributions in which the biomarker status is independent of treatment assignment. All probability distributions reflecting independence, $P \in \mathcal{M}^{\perp\!\!\!\perp}$, as guaranteed by randomization, satisfy: $\Pr(B = j \mid X = k, W = w) = \Pr(B = j \mid W = w)$. Notice that a restricted model is a subset of the model without the constraint, for example, $\mathcal{M}^{E,\perp\!\!\!\perp} \subset \mathcal{M}^E$.

A smaller model presents the possibility of efficiency gain for an estimator that utilizes the information implicit in the constraint. Knowledge of the treatment assignment and sampling mechanisms also reduces the set of possible probability distributions by eliminating the need to consider distributions with other conditional probability of treatment assignment or sampling in the second phase. As will be discussed in Section 3.1, these sources of knowledge do not impact the efficiency bounds for estimation of P-AERI and are suppressed in our notation. Complete knowledge of the conditional biomarker distribution could also be considered. For completion, we will comment on key features of inference reflecting this further constrained model, $\mathcal{M}^B$ and estimation contexts where it may be more realistic than for our motivating examples.

To simplify reference to related objects, we will overload the notation P of a joint probability distribution for a random observation O and the measure it induces to also refer to the distribution or measure corresponding to conditional random variables in the observation and rely on arguments to differentiate. For example, with $O^F = (Y, B, X, W) \sim P^F$, we write $P^F(b; x, w)$ for the measure induced by the random variable B conditional on $X = x, W = w$.

For brevity, we introduce functional notation for various conditional probabilities defined

by the data-generating mechanism. Let

$$\begin{aligned}
\pi(y,x) &= \Pr(\Delta = 1 \mid Y = y, X = x), \\
g(j,k \mid w) &= \Pr\left(B = j, X = k \mid W = w\right), \\
g_B(j \mid k,w) &= \Pr\left(B = j, \mid X = k, W = w\right), \text{ and} \\
g_X(k \mid w) &= \Pr\left(X = k \mid W = w\right).
\end{aligned}$$

2.3 Efficient Influence Functions for Estimators of P-AERI

We now formally consider our estimand P-AERI in terms of a real-valued function Ψ defined over the model $\mathcal{M}$ when P_0 is the unknown data generating mechanism; in symbols $\psi_0 := \Psi(P_0)$. The map $\Psi(P)$ is the usual linear combination of the population averaged exposure specific mean outcome functions $\Psi_{jk}(P) = \mathbb{E}_P[\mathrm{Q}_{jk}(W)]$, where the expectation over W and the parameter functions $\mathrm{Q}_{jk}(w)$ are determined by P. The gradient, a representation of the directional derivative of a map, also depends on evaluation at a probability distribution in a model, e.g., $D_0(O) := D(P_0)(O)$ where $O \sim P_0$. In this section we establish the gradients of the P-AERI map Ψ that are efficient relative to statistical models that do and do not reflect biomarker-treatment independence, and examine the efficiency gains that follow from utilizing that knowledge. The gradients will be employed in the construction of efficient RAL estimators in the next section. Henceforth, we will refer to an influence function of a map Υ as shorthand for the gradient of the map that equals the influence function of an asymptotically linear estimator of $\Upsilon(P_0)$.

By linearity, it suffices to focus on the efficient influence functions for a specific biomarker-treatment level parameter: $\psi_{jk} \equiv \mathbb{E}_W\left[\mathrm{Q}_{jk}(W)\right]$, where $j,k \in \{0,1\}$. This is the bivariate exposure analog of the treatment-specific mean outcome that has been studied in causal inference (Rose and van der Laan, 2011a). In Section 2.4 we will use these influence functions to construct asymptotically efficient estimators of ψ_0 with and without exploiting known independence.

Full Data Models

The efficient influence function of Ψ_{jk}, under sampling from an element P^F of the unconstrained full data model, is a natural extension of the influence function for the single exposure mapping and equals

$$IF_{jk}^{F,*}(P^F)(o^F) = \frac{I[B=j, X=k]}{g(j,k \mid w)} \left\{ y - \mathrm{Q}_{jk}(w) \right\} + \mathrm{Q}_{jk}(w) - \Psi_{jk}(P^F) \ , \tag{2.2}$$

where $g(j,k \mid w) = \Pr\left(B=j, X=k \mid W=w\right)$ and the asterix in the superscript denotes efficiency, in this case for estimation over the model $\mathcal{M}^F$. A full-data influence function, $D(P^F)$, is defined on the support of P^F.

Experimental Data Model Not Reflecting Independence

Each sampling mechanism and full data distribution pair $\left(\Pi, P^F\right)$ induces the distribution P^E of the corresponding experimental data. In terms of the experimental and full data structures defined, and suppressing the dependence on the probability distribution, the established correspondence between full data and experimental data influence functions (van der Laan and Robins, 2003) can be stated:

$$IF_{jk}^{E,*}(o^E) \ = \ \frac{\delta}{\pi(x,y)} IF_{jk}^{F,*}(o^F) + \left\{ \frac{\delta}{\pi(y,x)} - 1 \right\} \mathbb{E}_B \left[IF_{jk}^{F,*}(O^F) \mid Y=y, X=x, W=w \right],$$

where we note that full-data and experimental-data observations are equal when δ is one. The uniqueness of influence functions over unconstrained models ensures that the efficiency is preserved. The efficient influence function in the experimental data model is obtained by replacing $IF_{jk}^{F,*}(o^F)$ with the expression in Equation 2.2. It simplifies to

$$\begin{aligned} IF_{jk}^{E,*}(P^E)(o^E) = \ & \frac{\delta}{\pi(y,x)} \frac{I[X=k]}{g_X(k)} \left\{ \frac{I[B=j]}{g_B(j \mid k)} - \left(\frac{\mathrm{Q}_{jk}}{\overline{\mathrm{Q}}_k} \right)^y \left(\frac{1-\mathrm{Q}_{jk}}{1-\overline{\mathrm{Q}}_k} \right)^{(1-y)} \right\} \left(y - \mathrm{Q}_{jk} \right) \\ & + \frac{I[X=k]}{g_X(k)} \left(\frac{\mathrm{Q}_{jk}}{\overline{\mathrm{Q}}_k} \right)^y \left(\frac{1-\mathrm{Q}_{jk}}{1-\overline{\mathrm{Q}}_k} \right)^{(1-y)} \left(y - \mathrm{Q}_{jk} \right) + \mathrm{Q}_{jk} - \Psi_{jk}(P^E), \end{aligned} \tag{2.3}$$

where $g(j,k \mid w) = g_B(j \mid k,w) g_X(k \mid w)$ is a factorization of the conditional biomarker-treatment exposure probabilities and $\overline{\mathrm{Q}}_k(w)$ is the probability of outcome, marginal over

biomarker status and conditional on treatment assignment ($X = k$) and baseline covariates ($W = w$). The dependence of the functions $\pi, \mathrm{Q}_{jk}, \overline{\mathrm{Q}}_k, g_B$, and g_X on w has been supressed for brevity.

The above influence function is efficient for estimation over the completely unconstrained experimental model $\mathcal{M}^E$, the model reflecting known sampling weights, $\mathcal{M}^{E,\Pi}$, and the models reflecting known treatment assignment $\mathcal{M}^{E,X}, \mathcal{M}^{E,\Pi,X}$. This follows from $IF_{jk}^{E,*}(P^E)$ being P^E-uncorrelated with all scores corresponding to variations of these components of the data generating mechanism. In the parlance of Van der Laan and Robins (2003), the above efficient influence function lies in the orthogonal complement of the tangent space to the nuisance parameters $\pi(x, y, w)$ and $g_x(w)$. The models reflecting known treatment or sampling mechanism admit additional inefficient RAL estimators, but our interest is efficient estimation and these details are omitted from our notation and discussion.

2.3.1 Experimental Model Utilizing Known Independence

When estimation utilizes known conditional biomarker-treatment independence, the set of influence functions is indexed by the subset of scores that are orthogonal to variations within $\mathcal{M}^{\perp\!\!\!\perp}$. The natural correspondence between influence functions in the full and experimental data models does not generally preserve efficiency. Two general approaches to obtaining the influence function efficient for estimation relative to $\mathcal{M}^{E,\perp\!\!\!\perp}$ are to (1) calculate all influence functions and minimize in terms of $L^2(P^E)$-norm or (2) use the geometrical approaches described in Bickel et al. (1993), which involve calculating the information operator or characterizing the vector space of scores corresponding to variations within $\mathcal{M}^{E,\perp\!\!\!\perp}$ in order to directly project any influence function onto it. Both approaches can prove challenging in practice.

For the focus of this work, a third approach is available. The basic idea is that, for distributions of W with finite support, the estimator of $\psi_{jk,0}$ based on the maximum likelihood estimator of $\mathrm{Q}_{jk,0}$, $\psi_{jk,n} = \int \mathrm{Q}_{jk,n}(w) dP_0(w)$, can be analyzed to reveal, under some regularity assumptions, the influence function $D_{jk}^{E,*}$. Because likelihoods of binomial variables are

always correctly specified, the influence function of the MLE can be obtained in terms of the scores from a working likelihood that assumes the baselines covariates have finite support. Then, classical information theory provides a tractable means of executing the described projection.

Formally, the influence function obtained is only a candidate influence function for distributions with more general covariate distributions. The lack of reliance on the working assumption and the efficiency of the influence function can be established explicitly. The intuition behind this is twofold. In analytic terms, one can approximate any continuous distribution of covariates arbitrarily well with a discrete distribution (Chamberlain, 1987). In geometric terms, the component of the gradient that corresponds to the outcome regression is independent of the distribution of covariates on which it is conditioned and the component of the influence function that corresponds to the distribution of the baseline covariates, $\mathrm{Q}_{jk}(w) - \Psi_{jk}$, is the same in the full data model as in the observed-data model since the baseline covariates are always observed.

Conditional on W, the observation $[O^E \mid W = w]$ is a vector of binary variables and can be described by a likelihood function. For any given w, the conditional scores for $Q_{0j}(w), Q_{1j}(w)$, and $g_b(j, w)$, with $j \in \{0, 1\}$, are:

$$\begin{aligned} U_{0j}(o^E) &= \left[\delta(1-b)\frac{y - Q_{0j}}{Q_{0j}(1-Q_{0j})} + (1-\delta)g_{0j}\frac{(y-\overline{Q}_j)}{\overline{Q}_j(1-\overline{Q}_j)}\right] I\left[x = j\right] \\ U_{1j}(o^E) &= \left[\delta b\frac{y - Q_{1j}}{Q_{1j}(1-Q_{1j})} + (1-\delta)g_{1j}\frac{(y-\overline{Q}_j)}{\overline{Q}_j(1-\overline{Q}_j)}\right] I\left[x = j\right] \\ U_{gj}(o^E) &= \left[\delta\frac{b - g_{1j}}{g_{1j}g_{0j}} + (1-\delta)\frac{(y-\overline{Q}_j)}{\overline{Q}_j(1-\overline{Q}_j)}(Q_{1j} - Q_{0j})\right] I\left[x = j\right], \end{aligned} \tag{2.4}$$

respectively. Ordering the three scores corresponding to $x = 0$ first, followed by those corresponding to $x = 1$, produces a 6-vector of scores denoted by U_w. The dependence of $Q_{bx}(w)$, $g_{bx}(w)$, and the individuals scores on w has been suppressed. The conditional Fisher information is $I(w) = \mathbb{E}\left[U(O^E)U^T(O^E) \mid W = w\right]$.

Theorem 1 *Under sufficient regularity, the efficient influence function of Ψ_{jk}, with respect*

to estimation over $\mathcal{M}^E$, is of the form

$$IF^E_{jk}(P^E)(o^E) = I^{r_{jk}}(w)U(o^E) + Q_{jk}(w) - \Psi_{jk}(P^E) \tag{2.5}$$

where U is the vector of scores presented in Equation 2.4 and I its Fisher Information $\mathbb{E}[UU^T \mid W = w]$, both conditional on baseline covariates $W = w$, and we use $I^{r_{jk}}$ to denote the row of I^{-1} corresponding to the index of Q_{jk} in U.

Conceptually, the component of the influence function due to estimating the outcome regression parameters is expressed using classical information theory and the component following from using the empirical estimator of the covariate distribution remains unchanged. Additionally, consider the case when δ always equals 1 and each experimental observation is complete (is a full observation). Then, the second term in each score is identically 0 and pairwise the six scores are orthogonal. There is no information sharing between the parameters. In particular, there is no efficiency gain from knowledge pertaining to the nuisance parameters of the biomarker distribution with respect to estimation over $\mathcal{M}^F$. For estimation not exploiting independence, i.e., with respect to $\mathcal{M}^E$, this is an alternate expression of the efficient influence function stated in Equation 2.3. The utility of this form comes from its natural extension to estimation utilizing varying degrees of knowledge, such as independence, a the biomarker status distribution. The main result lies in the following theorem and its corollaries.

Corollary 1 *Under the same regularity conditions as Theorem 1, the efficient influence function of Ψ_{jk}, with respect to estimation over $\mathcal{M}^{E,\perp\!\!\!\perp}$ or $\mathcal{M}^{E,B}$, is of the same form as in Equation 2.5, but with I and U replaced by $I_{\perp\!\!\!\perp}$ and $U_{\perp\!\!\!\perp}$, or I_B and U_B, respectively.*

The conditional likelihood in the unrestricted observed model has 6 parameters relevant to the efficient influence function: $q_B(1; k, w), \mathrm{Q}_{0k}(w)$, and $\mathrm{Q}_{1k}(w)$ each by two levels of treatment $k \in \{0, 1\}$, 6 corresponding scores in U, and I is a 6-by-6 matrix. When the biomarker status parameter is known, this reduces to a vector of 4 scores U_B and a 4-by-4 matrix I_B. The independence constraint reflects an intermediate amount of knowledge since

the two $q_B(1;k,w)$ collapse into one $q_B(1;w)$ and a 5-dimensional analog in terms of $U_{\perp\!\!\!\perp}$ and $I_{\perp\!\!\!\perp}$ holds. Efficiency gains due to knowledge pertaining to the conditional biomarker distribution follow from a reduction in parameters.

Full details and expressions for U and I are presented in Appendix C.1.

2.4 Targeted Estimators of P-AERI

Our goal is the construction of asymptotically efficient estimators. The proposed estimator will involve two steps: (1) construction of consistent estimators of relevant features of the data-generating mechanism that define the initial estimator by substitution into Ψ, followed by (2) a correction to the initial estimator to recover asymptotic efficiency. Justification for our approach arises from analysis of the first-order expansion of Ψ_{jk}. In this step, we rely on the fact that the function of interest Ψ and each of its terms Ψ_{jk} are pathwise differentiable as a function over each of the statistical models: $\mathcal{M}^F, \mathcal{M}^E, \mathcal{M}^{F,\perp\!\!\!\perp}$, and $\mathcal{M}^{E,\perp\!\!\!\perp}$ that we consider herein.

Initial Substitution Estimator

The first step in our construction of an estimator is an initial substitution estimator $\Psi_{jk}(P_n)$, where P_n is an estimate of the salient features of P_0 from an *i.i.d.* sample $O_i \sim P_0$ of size n. For the biomarker-treatment specific mean, $\psi_{jk,0}$, the salient distributional features are the outcome probability function $\mathrm{Q}_{jk,0}$ and the covariate distribution. We use the empirical distribution of baseline covariates, $\mathbb{P}_{W,n}$, and must construct an estimate of $\mathrm{Q}_{jk,0}$. This can be accomplished with any flexible approach that yields a consistent estimator. For example, local linear regression could be employed, with bandwidth selected to minimize a cross-validated risk, to produce $\mathrm{Q}_{jk,n}$. The initial substitution estimator follows as $\psi^{sub}_{jk,n} = \Psi_{jk}(P_n) = \mathbb{E}_{W,n}\left[\mathrm{Q}_{jk,n}(W)\right]$, the average of $\mathrm{Q}_{jk,n}$ over the observed covariates W_i.

When working with experimental observations, estimation techniques able to account for the two-phase design must be employed. For example, inverse sampling-probability weighted Nadaraya-Watson estimation could be used to construct $\mathrm{Q}_{jk,n}$. In this case, only observations

sampled in the second phase contribute to the initial estimator. Otherwise, the procedure is the same for estimation over $\mathcal{M}^E$ as for $\mathcal{M}^F$. In either case, each subject only contributes estimation of the conditional outcome probability, $Q_{jk}(w)$, in accordance with their treatment assignment and biomarker level.

The substitution estimate of the population-averaged measure of additive interaction, P-AERI, will rely on consistent estimates of all four $\mathrm{Q}_{jk,0}$ features and the empirical distribution of covariates; $\psi^{sub} := \Psi(\mathrm{Q}_{11,n}, \mathrm{Q}_{10,n}, \mathrm{Q}_{01,n}, \mathrm{Q}_{00,n}, \mathbb{P}_{W,n}) = \psi^{sub}_{11,n} - \psi^{sub}_{10,n} - \psi^{sub}_{01,n} + \psi^{sub}_{00,n}$.

Motivation for Approach

Study of the asymptotic behavior of these initial estimators requires algebraic manipulation of the first-order expansion of the maps Ψ_{jk}, around P_0, in terms of their gradients. The following Theorem shows the resulting formula which both motivates the targeted approach and implies regularity conditions on the estimated features of the data-generating mechanism that comprise P_n.

Theorem 2 *Plug-in estimators of the biomarker-treatment specific components of the interaction parameter of interest, can be expanded around the estimand $\Psi_{jk}(P_0)$ as*

$$\Psi_{jk}(P_n) - \Psi_{jk}(P_0) = \frac{1}{n}\sum_i IF_{jk}(P_0)(O_i) - \frac{1}{n}\sum_i IF_{jk}(P_n)(O_i) + R_{jk}(P_n, P_0) + o_p(n^{-1/2})$$

so long as the centered influence functions, $\{IF_{jk}(P_n) - IF_{jk}(P_0)\}$, fall within a Donsker class.

The Donsker condition is sufficient to establish that an empirical process term is negligible at standard rates as correctly accounted for by the $o_P(n^{-1/2})$ term. Our planned estimation approach additionally relies on asymptotic neglibility of the remainder term, which will impose rate requirements on estimators of the unknown distributional features. In Section 3.3 we will see that, for example, using Nadaraya-Watson estimates of the conditional outcome and biomarker status distributions will meet the rate requirements for W having up to three continuous components.

Bias Correction

Assuming the initial consistent estimators meet the rate requirements, Theorem 2 can be simplified as

$$\Psi_{jk}(P_n) - \Psi_{jk}(P_0) = \frac{1}{n}\sum_i IF_{jk}(P_0)(O_i) - \frac{1}{n}\sum_i IF_{jk}(P_n)(O_i) + o_p(n^{-1/2}) \tag{2.6}$$

which, up to the empirical average of the influence function at P_n, meets the definition of asymptotic linearity. The process of eliminating this term, a potential source of asymptotic bias, is referred to as targeting the parameter of interest. In this article, we will construct and study one-step corrected estimators which are obtained by adding the bias term to the initial estimator:

$$\psi^{OS}_{jk,n} = \Psi_{jk}(P_n) + \frac{1}{n}\sum_i IF_{jk}(P_n)(O_i)$$

Another option is to use targeted maximum likelihood, which produces a substitution estimator by updating P_n such that the bias term evaluates to zero (van der Laan and Rubin, 2006). While this method is a bit more involved than the one-step correction, the benefit of a substitution estimator can deliver improved finite sample behavior, especially for a binary outcome. The two correction approaches produce asymptotically equivalent estimators of P-AERI.

The only features of the data-generating mechanism salient to the parameter map were $\mathrm{Q}_{11,0}$ and the covariate distribution. However, additional features are necessary for estimation of the influence function. The sampling weights and the treatment assignment distribution, both known by design, are required. The biomarker distribution is not known and a consistent estimate is needed. When leveraging known independence between the biomarker and treatment status, as in Equations 2.3, this is only one pooled distributional feature $g_{B,0}$, that is a function of w, and to which observations of either treatment assignment group contribute. When known independence is not being leveraged, as in Equations 2.2 , consistent estimates for each of the two treatment-specific biomarker distributions are required. As for $\mathrm{Q}_{jk,n}$, flexible approaches to constructing consistent estimates of $g_{B,n}$ can be taken. The esti-

mated influence function is defined by substitution. For estimation exploiting independence, $IF^{E,\perp\!\!\!\perp}(P_n) = IF(\mathrm{Q}_{11,n}, \mathrm{Q}_{10,n}, \mathrm{Q}_{01,n}, \mathrm{Q}_{00,n}, g_{B,n}, g_{X,0}, \mathbb{P}_{W,n})$, and for estimation not exploiting independence $IF^{E}(P_n)$ is defined similarly.

The one-step corrected estimator of P-AERI is defined:

$$\psi_n^{OS} = \psi_{11,n}^{OS} - \psi_{10,n}^{OS} - \psi_{01,n}^{OS} + \psi_{00,n}^{OS},$$

as expected.

2.4.1 Statistical Inference for ψ_n^{OS}

The limiting behavior of an asymptotically linear estimator is governed by the Central Limit Theorem, which implies that

$$\sqrt{n}\left(\gamma_n - \gamma_0\right) \rightarrow_d N\left(0, \sigma_0^2\right)$$

where $\sigma_0^2 = \int IF^2(P_0)(o)dP_0(o)$ is the variance of the influence function $IF(P_0)(o)$ of the estimator γ_n of γ_0 constructed under sampling from P_0. A natural estimator of the limiting variance is $\sigma_n^2 = \frac{1}{n}\sum_i IF^2(P_n)(O_i)$ where P_n is an estimator of P_0 (or the relevant features of P_0). Approximate Wald confidence intervals with confidence level α can be constructed as $\gamma_n \pm \sqrt{\frac{\sigma_n^2}{n}} z_{\alpha/2}$. The efficient estimators ψ_n and $\psi_{jk,n}$ proposed in this work are asymptotically linear and Wald confidence intervals can be contructed as detailed above using the corresponding efficient influence functions established previously.

2.5 Analysis of a Nested Gene-Treatment Interaction Study

The GWIS two-phase analyses of the hormone therapy trial within the Women's Health Initiative (WHI) identified a single nucelotide polymorphism (SNP) that came quite close to meeting the genome-wide significance level and was "weakly" replicated in an independent sample from the observational cohort of the WHI (unpublished). Here, we demonstrate our method by reanalysing the trial data as if conducting a validation of the SNP identified

as 'rs9909279' as a candidate genetic modifier of the effect of hormone therapy on incident diabetes.

Let Y indicate diabetes incident during the trial study period, the genetic exposure B indicate whether or not the subject had any adenine alleles at 'rs9909279', the treatment variable X indicate assignment of estrogen, with or without progestin and W denote age and BMI covariates collected at baseline. The phase-one analysis population consisted of 20,938 postmenopausal women (after 1,092 were removed for missing diabetes status (976), baseline BMI (108), or both (8)) who were randomized to receipt of hormone therapy containing estrogen (50%) or to placebo (50%), independently of baseline covariates. The phase-two population (N=3,129) was comprised of a case-control subsample matched on age ($\pm$ 5 years) and administrative variables (e.g., enrollment date and length of follow-up). We approximated the known sampling weights from the data: the simple percentage of cases in phase two gives $\pi(y = 1, w_{age}) = 0.876$ and the probability of controls being sampled depended on age, with $\pi(y = 0, w_{age} = age)$ estimated by assuming a 2:1 control to case ratio and distributing the probability of selection among match candidates. Sampling probabilities for controls varied between 3.1 and 11.5 percent, depending on age.

Initial estimates of features of the unknown data-distributing mechanism were obtained using weighted local-linear regression. The estimates were constructed from observations with complete phase-two data weighted by their inverse probability of sampling. The smoothing parameter α was chosen to minimize the negative log-likelihood risk estimated via 10-fold cross-validation. The features of the data generating mechanism required to construct the one-step estimator are: (1) the four probability of incident diabetes functions, conditional on gene-treatment exposure and covariates, $\mathrm{Q}_{jk}(w)$, and (2) the probability of having at least one copy of adenine at 'rs9909279', conditional on hormone therapy receipt and covariates, $q_B\,(1; k, w)$, as well as (3) the pooled version applicable under known gene-treatment independence, $q_B\,(1; w)$. Estimates of P-AERI and its components are presented in Table 2.1 for the uncorrected (Naive) initial estimators, and the one-step (OS) corrected estimators efficient relative to using and ignoring known independence. Both estimates show a statisti-

		$\mathcal{M}^E$		$\mathcal{M}^{E,\perp\!\!\!\perp}$	
Parameter	ψ_n^{Naive}	ψ_n^{OS} $(95\%CI)$	SE	ψ_n^{OS} $(95\%CI)$	SE
ψ_{11}	0.039	0.042 (0.036,0.047)	0.0029	0.044 (0.038,0.05)	0.003
ψ_{10}	0.067	0.065 (0.057,0.074)	0.0042	0.061 (0.054,0.068)	0.0037
ψ_{01}	0.065	0.065 (0.056,0.073)	0.0045	0.06 (0.053,0.068)	0.0038
ψ_{00}	0.052	0.055 (0.048,0.062)	0.0036	0.058 (0.051,0.065)	0.0037
$\psi_{1\Delta}$	-0.028	-0.024 (-0.034,-0.014)	0.0051	-0.017 (-0.026,-0.009)	0.0045
$\psi_{0\Delta}$	0.013	0.009 (-0.002,0.021)	0.0057	0.002 (-0.008,0.012)	0.0049
ψ	-0.041	-0.033 (-0.05,-0.016)	0.0086	-0.019 (-0.033,-0.006)	0.0069

Table 2.1: Estimated modification by 'rs9909279' of hormone therapy effect on incident diabetes among white participants of the WHI trial. In addition to the gene-treatment specific components, ψ_{gx} and P-AERI estimand, ψ, we present gene-specific treatment effects $\psi_{\jmath\Delta} := \psi_{j1} - \psi_{j0}$.

cally significant negative estimate of P-AERI which indicates that hormone therapy and an adenine variant at 'rs9909279' combine to make a protective contribution against incident diabetes. The reduction in standard errors of the additive measure of interaction, from 0.086 to 0.0069, corresponds to an estimated 55% efficiency gain from utilizing known independence.

However, the smaller P-AERI estimate when utilizing independence, -0.019 compared with -0.033 for the estimate not utilizing independence, leads to a p-value of 0.0053, which is larger than that obtained when ignoring the knowledge (0.0001). The difference in point estimates can be attributed to two differences in estimation over the model reflecting independence. First, the efficient influence functions differ over the two models and hence so does the additive correction made to the naive estimate. Another difference in the two corrections is that a pooled single initial estimator of genetic exposure independent of treatment assignment replaces the two separate estimates of treatment specific genetic exposure, so the two influence functions D^* and $D^{\perp\!\!\!\perp,*}$ are evaluated at distributions P_n and $P_{\perp\!\!\!\perp,n}$ that likely

differ. Though both methods provide asymptotically unbiased estimates, the difference can be apparent in finite samples. Estimation over $\mathcal{M}^{E,\perp}$ is the more appropriate analysis.

In addition to conducting inference for the P-AERI interaction estimand, the method introduced in this paper naturally provides additional quantities that are useful for characterizing the joint effects of hormone therapy and the variant at 'rs9909279' on incident diabetes. Estimates and confidence intervals of the probability of outcome, standardized to the study population joint age and BMI distribution, for each gene-treatment level can be reported with minimal additional effort. Similarly, estimates and confidence intervals for the two treatment effects, among those with the genetic variant of interest and without, are provided and denoted by $\psi_{b\Delta}$ for $b \in 0, 1$. This analysis supports that the detected interaction is pure, in the sense that presence of the genetic variant appears necessary for a treatment effect. There is no evidence for a treatment effect in a population with no copies of the adenine variant at 'rs9909279'. However, with one or more adenine allele at 'rs9909279', the analysis supports a statistically significant protective effect of hormone therapy. This is true whether utilizing the known independence or not. This additional information is often not available when using methods that only support hypothesis testing or inference for the interaction parameter.

While the analysis performed by the GARNET GWAS study focused on different measures of interaction, namely conditional measures of interaction on the odds ratio scale that considered all three genetic variants (0,1, or 2 alleles of adenine), the qualitative conclusion of additional benefit of treatment among those with at least one and no effect among those without any adenine variants at 'rs9909279' was observed.

Chapter 3

PROPERTIES AND PERFORMANCE OF ESTIMATORS OF ADDITIVE EXCESS RISK DUE TO INTERACTION

3.1 Efficiency Gain Due to Independence

The variances of the efficient influence functions calculated in the previous sections, which make use of varying degrees of knowledge on the biomarker status distribution, determine the efficiency bounds for suitably centered and scaled RAL estimators of the $\psi_{jk,0}$ and ψ_0 estimands. The bounds rely on features of the unknown underlying data-generating mechanism P_0 and have the form

$$\sigma^2_{jk,0} = \mathbb{E}\left[I_0^{r,r}(W)\right] + \mathbb{E}\left[\left\{\mathrm{Q}_{jk,0}(W) - \psi_{jk,0}\right\}^2\right]$$

where r denotes the index in the vector of conditional scores U corresponding to the distributional parameter Q_{jk}, and I_0 is the conditional Fisher Information matrix describing the covariances of the scores in U. These quantities are the same as those employed in the derivation of the influence functions and complete details can be found in Appendix C.1.

With these expressions we can start to explore the potential for and features of efficiency gain under various amounts of knowledge. Only the first term of $\sigma^2_{jk,0}$ changes with different levels of knowledge on the biomarker status distribution. In scenarios where the outcome probability varies significantly and the second term dominates, relative efficiency gains could be small. The two terms in the variance formula can be attributed to estimation of the conditional outcome probability and estimation of the covariate distribution, respectively. We recall that the estimand is defined as the average of the conditional probability of a binary outcome and point out the similarity of the above expression and the law of total variance for a binary random variable Y.

For estimation based on full-data observations the scores in U^F are pairwise orthogonal and no asymptotic efficiency gains result from knowledge on the bivariate exposure distribution. In other words, the exposure distribution, sometimes referred to as the propensity score, is an orthogonal nuisance to the parameters ψ_{jk}, and the orthogonal projection of the efficient influence function onto the span of scores corresponding to the exposure distribution is zero. Consequently, any efficiency gain due to knowledge on the exposure distribution, such as conditional independence between the exposures, observed in estimation based on the experimental data is a direct result of the missing data structure induced by the two-phase sampling scheme.

Intuitively, this is reasonable since observations not measured in the second phase only contribute to the marginal outcome regression, which involves the genetic exposure distribution. The scores, as functions of the experimental observation unit, for outcome and biomarker status parameters are not orthogonal, and the biomarker component of the bivariate exposure distribution is not an orthogonal nuisance parameter in the observed models; knowledge about it has the potential to reduce efficiency bounds. In the context presented here, the second exposure, treatment, is measured at phase one. Since this variable is not subject to missingness, its distribution remains an orthogonal nuisance parameter and estimation of $\psi_{jk,0}$ over $\mathcal{M}^E$ and over $\mathcal{M}^{E,X}$ have the same efficiency bound.

As an example, we considered a scenario in which the conditional outcome and biomarker status probabilities were logistic-linear in covariates, treatment assignment was 50% for all subjects, and the covariates followed a truncated bivariate normal distribution modeled loosely on the age-BMI distribution of the WHI trial cohort. Figure 3.1 presents an examination of efficiency bounds under unmatched case-control and simple random phase-two samples for estimation reflecting differing amounts of knowledge about the genetic-exposure distribution: no knowledge, conditional independence with treatment, and completely known genetic exposure distribution. The efficiency bounds are relative to that of estimation using the full data, hence always less than 1 when using experimental data, and were approximated from the empirical variance of the efficient influence function in a very large (N=one

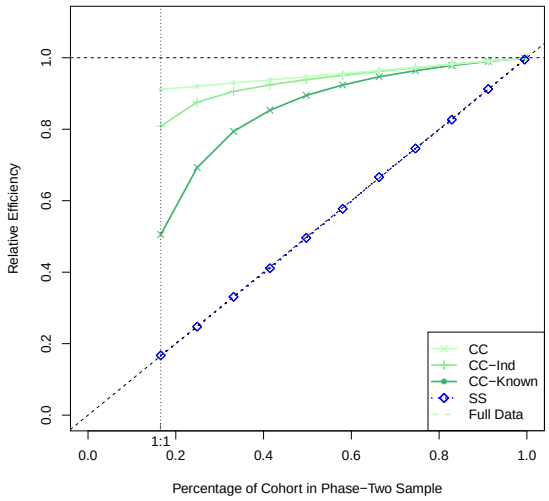


Figure 3.1: Efficiency gains from utilizing various amounts of knowledge on the biomarker-treatment distribution when estimating $P-$AERI. Levels of knowledge: none, independence (Ind), and complete (Known) are considered for two phase-two sampling schemes: case-control (CC) including all cases and simple random samples (SS) of the same size. Efficiencies in estimation are all taken relative to that using full data in which all observations are measured at phase two.

million) simulated data set. All relative efficiencies clearly increase and approach one as the phase-two sample size increases and the experimental model approaches the full data model.

In the 1:1 case-control scenario, with all cases sampled, the efficiency gains attributable to the exploiting knowledge is readily seen. As the ratio increases the observed gain decreases to none, which is expected since the missingness also goes to zero. For any fixed phase-two sample size, the case-control samples support more efficient estimation than the simple random sample; this is consistent with conventional wisdom that cases are the most informative observations. For simple outcome-independent phase-two sampling, the efficiencies for estimation utilizing each level of knowledge are essentially the same and appear indistinguishable on the plot in FIgure 3.1, and the label 'SS' applies to all three simple-sampling scenarios.

Algebraically, this behavior is directly related to entries of the information matrix. In

particular, the inner product between the scores corresponding to the outcome and the genetic-exposure parameters contains a term proportional to $\pi(y=1,x,w)-\pi(y=0,x,w)$ which is 0 for a simple random sampling scheme and quite different from 0 when all of the cases and only a small proportion of controls are sampled. As is generally the case, sampling design impacts both overall efficiency in unconstrained models as well as the potential for efficiency gain in constrained models.

3.2 Efficiency Gain from Precision Variables

Genetic variants are often considered to be uncorrelated with variables commonly measured at baseline, such as age and BMI. Hence, when the biomarker under investigation is a gene variant, the motivation to account for baseline covariates in estimation of additive interaction lies with the potential for efficiency gain. We now show that the gains available to P-AERI$_n^{OS}$ are similar in nature to those achieved through the inclusion of so-called precision variables by estimators of conditional parameters of regression models. As before, by linearity, it will suffice to examine one component of P-AERI, and we will do so for the full-data structure.

First, in the absence of covariates, the natural quantity for comparison with ψ_{jk} is the previously introduced marginal quantity $Q_{jk} := Pr(Y=1 \mid B=j, X=k)$, which is an unadjusted version of ψ_{jk} that we call ψ_{jk}^{unadj}. In the full data model, this quantity has a simple empirical estimator: $\psi_{jk,n}^{unadj} := \frac{1}{n_{jk}}\sum_i I[B=j,X=k]Y_i$, where n_{jk} is the size of the sample with exposure level (j,k). The limiting variance of $\psi_{jk,n}^{unadj}$ is $\sigma_{jk}^{2,unadj} = \frac{1}{g_{jk}}\text{VAR}\left(Y \mid B=j, X=k\right)$, where $g_{jk} := Pr(B=k, X=j)$ is the unconditional bivariate exposure probability.

The adjusted version $\psi_{jk,n}$, when complete information is available for the full phase-one cohort, has asymptotic variance $\sigma_{jk}^2 := \mathbb{E}\left[IF_{jk}^2(O^F)\right]$, which can be expressed as:

$$\sigma_{jk}^2 \;=\; \mathbb{E}\left[\frac{1}{g_{jk}(W)}\text{VAR}\left(Y \mid B=j, X=x, W\right)\right] + \text{VAR}\left(\mathbb{E}\left[Y \mid B=j, X=x, W\right]\right).$$

If the covariates are precision variables, then W is necessarily independent of the bivariate exposure level (B,X). This implies that the probability of having exposure status $(B,X) =$

(j,k) is constant; $g_{jk}(w) = g_{jk}$, and the expression simplifies:

$$\sigma^2_{jk} \;=\; \frac{1}{g_{jk}}\mathbb{E}\left[\mathrm{VAR}\left(Y \mid B=j, X=x, W\right)\right] + \mathrm{VAR}\left(\mathbb{E}\left[Y \mid B=j, X=x, W\right]\right).$$

Employing the conditional variance formula, we can re-express this in terms of the absolute reduction in variance:

$$\sigma^{2,unadj}_{jk} - \sigma^2_{jk} \;=\; \left(\frac{1}{g_{jk}} - 1\right)\mathrm{VAR}\left(\mathbb{E}\left[Y \mid B=j, X=k, W\right]\right)$$

which reveals that the absolute reduction in variance is proportional to the variance of the conditional mean, $\mathrm{VAR}\left(\mathbb{E}\left[Y \mid B=j, X=k, W\right]\right)$, a familiar measure of how strongly the covariates and outcome are associated. This quantity is zero if the probability of outcome does not vary over baseline covariate values and positive otherwise. The constant of proportionality, $\left(\frac{1}{g_{jk}} - 1\right)$, is positive whenever P-AERI is identifiable.

The variances of the corresponding two-phase estimators

$$\sigma^{2,tp} = \mathbb{E}_W\left[\widetilde{I}^{r,r}_0(W)\right] + VAR\left(\mathbb{E}\left[Y|B=j, X=k\right]\right)$$

have the same general form as above and this basic property should also apply to the two-phase estimators.

125

3.3 Double Robustness and Rate Requirements

First Order Expansion

With P_ϵ a regular parametric submodel of $\mathcal{M}$, the first order expansion of a map Υ centered at P_0 is of the form: $\Upsilon\left(P_\epsilon\right) - \Upsilon\left(P_0\right) = -\int IF_\Upsilon\left(P_\epsilon\right)(o)\, dP_0(o) + R\left(P_\epsilon, P_0\right)$, where IF_Υ is an influence function of Υ relative to estimation over $\mathcal{M}$, and R is the remainder term. In the first-order expansion of the biomarker-treatment specific mean, Ψ_{jk}, as a function on $\mathcal{M}^F$ with efficient influence function $IF^{F,*}_{jk}$ defined in Equation 2.2, the remainder is directly calculated as:

$$R^F_{jk}\left(P^F_\epsilon, P^F_0\right)\left(o^F\right) = \mathbb{E}^0_W\left[\frac{g^0(j,k;W) - g^\epsilon(j,k;W)}{g^\epsilon(j,k;W)}\left\{\mathrm{Q}^\epsilon_{jk}\left(W\right) - \mathrm{Q}^0_{jk}\left(W\right)\right\}\right], \tag{3.1}$$

where the superscripts ϵ and 0 on the parameter functions g and Q_{jk} associate them to their corresponding probability distribution (P_ϵ and P_0, respectively).

The domain of the function, the statistical model $\mathcal{M}$, determines the parametric submodels over which the expansion holds. The efficient influence function, and hence the remainder, has the same form with respect to any of the full data models constrained to reflect any amount of knowledge on the gene-treatment exposure distribution: $\mathcal{M}^F, \mathcal{M}^{F,\Pi}, \mathcal{M}^{F,\Pi,g_x}, \mathcal{M}^{F,\perp}$ and so forth. Each element of the experimental model, $\mathcal{M}^E$, is naturally identified an element of the full data model. The parameter map naturally passes to the experimental model unchanged, and the remainder term is of the same form.

Theorem 3 *(**Preservation of remainder.**) The remainder term of the first-order expansion of the pathwise differentiable function Ψ, defined on the full data model $\mathcal{M}^F$ remains unchanged in the expansion relative to the corresponding observed model $\mathcal{M}^E$ under the association induced by the known sampling mechanism. That is: $R^F\left(P_\epsilon^F, P_0^F\right) = R^E\left(P_\epsilon^E, P_0^E\right)$*

Regular and asymptotically linear estimators are governed by the above deterministic result when sequences of probability distributions, determined by empirical samples, replace regular parametric submodels P_ϵ. We shall use P_n to denote such a distribution and note that it is generally distinct from the empirical distribution $\mathbb{P}_n$. For example, P_n is sufficiently determined by $(Q_{00,n}(w), Q_{01,n}(w), Q_{10,n}(w), Q_{11,n}, g_{B,n}(1;w), g_{B,n}(0;w), g_X(w), \mathbb{P}_{W,n})$ for the purposes of estimating P-AERI equal to $\Psi(P_0)$. These remainder terms govern the asymptotics of estimators of $\psi_{jk,0}$.

We expect the remainders from a first-order expansion of a map to be of higher order. In particular, the remainders $R_{jk}^F(P_n^F, P_0^F)$ are second-order and consist of the product of two differences between an estimator and its estimand. The two estimators are both used either in the construction of the initial substitution estimator, namely the biomarker-treatment specific outcome probability function, or in the influence function used to correct the substitution estimator, the biomarker distribution function or functions, depending on whether or not independence is being leveraged.

The remainder $R^F_{jk}(P^F_n, P^F_0)$ in Equation 3.1 has been expressed in complete generality and applies to the each element of the completely unconstrained model. When the treatment mechanism is known by design, $g^\epsilon_X = g^0_X$, and the factor involving the conditional biomarker-treatment distribution simplifies to a factor involving only the conditional biomarker status components g^ϵ_B and g^0_B. In cases where the conditional gene-treatment distribution is completely known, the remainder would be identically zero. By Theorem 3, these statements apply equally to the full or experimental data models.

Required Rates of Initial Estimators

In presenting the approach to constructing targeted estimators of P-AERI, we examined as expression, Equation 2.6, derived from the first-order expansion. To achieve asymptotic linearity, the remainder term, given by Equation 3.1, must be $o_P(n^{-1/2})$. The second-order nature of the remainder term is what allows some degree of flexibility in constructing consistent initial estimators of the outcome and biomarker probability functions Q_{jk} and g_B. The rate requirement on the remainder is equivalently viewed as a rate requirement on the product of a pair of estimators. The required rate on the remainder is achieved if the initial estimates $\mathrm{Q}_{jk,n}$ and $g_{B,n}$ are each consistently converging to $\mathrm{Q}_{jk,0}$ and $g_{B,0}$, respectively, at rates faster than $n^{-1/4}$. In the case of Nadaraya-Watson estimators, with bandwidth selected through cross-validation, this is achievable for at most three continuous baseline covariates, unless higher-order kernels are employed. Many other estimation techniques are available to an analyst. In practice, ensemble methods such as Super Learner are recommended (Van der Laan et al., 2007).

Double Robustness

If the remainder itself is asymptotically negligible, but at rates slower than standard, the targeted estimator will still be consistent, albeit also at rates slower than standard. From this perspective, the second-order nature of the remainder provides two-opportunities for the targeted estimator to be consistent. For estimation of P-AERI, and any of the consituent

biomarker-treatment specific mean outcome parameters, this double robustness property is ensured when either the outcome probability function or the biomarker probability function are consistently estimated. Consistent estimation is also guaranteed when both functions are consistently estimated, but not at rates sufficient to give asymptotic linearity. Without asymptotic linearity, our techniques for establishing inference do not apply.

The double robustness property of RAL estimators of the univariate treatment-specific mean is well- established. The remainder for the bivariate treatment-biomarker specific mean outcome is directly analogous. The previous statement regarding the preservation of the remainder furtehr established that this property extends to RAL estimators for the described two-phase experiments.

3.4 Simulated Performance of P-AERI$_n^{OS}$

We conducted simulations to evaluate the finite sample behavior of our proposed etimator and to investigate the efficiency gain from utilizing known independence. We emulate a two-phase sampling scheme where a binary outcome Y, treatment assignment X, and age and BMI measurements $W = (W_{age}, W_{BMI})$ are collected on all subjects. A binary biomarker status variable $B-$ is observed on the selected second phase subsample.

We used the two-phase GWAS study conducted on the WHI hormone therapy trial cohort as guidance for generating baseline covariates and selecting sample sizes and phase-two sampling scheme. Analyses on the subcohort of 'white' subcohort of participants identified a variant at a single nucleotide polymorphism (SNP) that potentially modified the effect of hormone therapy on incident diabetes. Informed by this dataset we generated two baseline covariates, age and BMI, jointly using a Normal distribution with mean $(63.88, 28.61)$ and covariance defined by $\sigma^2_{age} = 7.16^2$, $\sigma^2_{BMI} = 5.79^2$ and $\rho = 0.082$, truncated to ages $[50, 81]$ and BMI values in $[13.83, 69.4]$. Phase-one samples of 20,000 subjects were generated for all simulations presented in Tables 3.1 and 3.2. An unmatched 1:1 case-control sub sample was selected by Bernoulli sampling and the probability of phase-two observation was defined

separately for cases and controls, independent of other covariates, to achieve an average of 750 cases and 750 controls in phase two. Treatment assignment followed a Bernoulli(p=0.5) distribution. For each scenario, 5,000 replicates of the two-phase study were simulated.

The remaining details of our simulation are loosely modeled on the simulation study by Dai et al. (2009). The binary outcome was generated using a logistic model with linear predictor: $\beta_0 + \beta_G G + \beta_X X + \beta_{age} W_{age} + \beta_{BMI} W_{BMI} + \beta_{GX} GX$. Parameter values $\beta_B = 0.1$, $\beta_X = 0.2$, $\beta_{age} = 0.2/30$, $\beta_{BMI} = 0.2/55$ were fixed for all simulations and three interaction strengths, corresponding to the odds-ratio interaction parameter, β_{GX} being 0, 0.5, and 1, were considered. The intercept was set to achieve overall prevalences of outcome, in the three interaction scenarios examined, of roughly $5, 6$, and 7 percent.

The binary genetic exposure variable was also generated using a logistic model with linear predictor: $\alpha_0 + \alpha_{age} W_{age} + \alpha_{BMI} W_{BMI}$ Parameter values of $\alpha_{age} = 0.5/30$, $\alpha_{BMI} = 0.5/55$ were fixed and the intercept selected so that about 25% of the simulated subjects had $G = 1$.

Initial estimates of the conditional mean outcome parameters $\mathrm{Q}_{jk}(w)$ were constructed via correctly specified logistic regression, weighted by known inverse probability of sampling on the phase-two subsample. The empirical average of the regression functions, evaluated at the covariate values of each phase-one observations, defined the uncorrected (naive) estimators of the $\psi_{jk,0}$ and hence ψ_0 targets. One-step corrected estimators, $\psi_{jk,n}$ were calculated by adding the empirical average of the appropriate influence function, evaluated at the initial estimates of the mean outcome and biomarker status parameters, to the naive estimates. Estimation of the biomarker status parameters, g_B were constructed similar to the estimates of the conditional mean outcome parameters, either with or without utilizing the known independence according to the estimation scenario. That is, for estimators not exploiting known independence, the naive estimates of the probability of $B = 1$ could vary by treatment level for a common covariate value, whereas for estimators exploiting independence they are equal.

Table 3.1 summarizes the bias and sample variances of the one-step estimators of the various gene-treatment level specific parameters ψ_{jk} and the population averaged AERI pa-

rameter $\psi = \psi_{11} - \psi_{10} - \psi_{01} + \psi_{00}$ under the three strength of interaction scenarios considered. Results from estimation with and without using known independence are presented. With efficient inference as our focus, Monte Carlo bias and standard deviation are accompanied by a ratio that compares the scaled Monte Carlo variance with the asymptotic bound which was approximated from a large simulated data set (N=ten million). The gain in efficiency from using independence is presented in terms of observed Monte Carlo variances and in terms of the asymptotic bounds. The constructed estimators all demonstrate consistency and the expected efficiency gains from using independence. The gains vary between 10 and 46 percent for individual gene-treatment level parameters, and combine to yield gains of 86 to 94 percent for estimation of P-AERI. This magnitude is comparable to the two-fold gains reported in studies of other methods for additive and multiplicative interaction (Dai et al., 2009).

The Monte Carlo variances are close to but generally larger than what would be expected from the limiting bounds (ratios are close to, but slightly less than 1). While all simulations are based on $N_I = 20,000$ observations, only $N_{II} \approx 1,500$ are used to construct the initial estimators and contribute complete information to the correction. $N_{II,jk}$ reports the average number of observations in phase two with $X = j$ and $B = k$. The distribution of the phase-two observations across (B, X) strata changes with interaction strength as a result of the outcome dependent sampling; this is most noticeable in the enrichment of $B = 1$ observations.

We evaluate the estimated variances in Table 3.2. Overall, the mean of the estimated limiting variances are fairly close to the observed Monte Carlo variances, scaled by sample size, which in turn agree quite well with the asymptotic bounds. Aproximate 95% Wald confidence intervals behave well with observed coverages falling between 94.5 and 95.7 percent.

In addition to detailed examination of estimators based on a realistically sized study, we examine the asymptotic behavior of our estimators of P-AERI in Table 3.3. For these simulations, data was generated by the same approach described above, just scaled to create phase-one sample sizes of $5, 10, 20, 40, 60$ and 80 thousand observations and phase-two sam-

ples consisting of 7.5% of phase-one subjects. We confirm that the observed Monte Carlo variances and efficiency gains settle to those dictated by the asymptotic bounds as sample sizes increase, as do the estimators of the asymptotic variance bound. The bias and coverages behave quite well for all sample sizes considered.

Table 3.1: Evaluation of Point Estimators With and Without Using Independence

			$\mathcal{M}^E$			$\mathcal{M}^{E,\perp\!\!\!\perp}$			Efficiency	
	$\psi_{ij,0}$	$N_{II,jk}$	Bias	SD	$\frac{\sigma_0^2}{Var}$	Bias	SD	$\frac{\sigma_0^2}{Var}$	MC	Bounds
$\psi \approx 0$										
ψ_{00}	0.048	525.0	0.0001	0.003	1.01	0.0001	0.003	1.02	1.12	1.11
ψ_{10}	0.052	191.6	0.0007	0.007	0.93	0.0005	0.006	0.97	1.29	1.24
ψ_{01}	0.057	572.7	0.0001	0.003	1.02	0.0000	0.003	1.04	1.15	1.12
ψ_{11}	0.063	210.1	0.0007	0.008	0.95	0.0005	0.007	0.98	1.31	1.27
ψ	0.001		0.0000	0.013	0.94	0.0001	0.009	0.99	1.92	1.82
$\psi \approx 0.04$										
ψ_{00}	0.048	506.5	0.0001	0.003	1.01	0.0001	0.003	1.02	1.12	1.11
ψ_{10}	0.052	184.3	0.0006	0.007	0.92	0.0005	0.006	0.95	1.28	1.23
ψ_{01}	0.057	549.9	0.0001	0.003	1.02	0.0000	0.003	1.04	1.13	1.11
ψ_{11}	0.100	258.5	0.0011	0.011	0.94	0.0009	0.009	0.98	1.41	1.35
ψ	0.038		0.0006	0.016	0.93	0.0005	0.011	0.99	1.98	1.86
$\psi \approx 0.09$										
ψ_{00}	0.048	484.3	0.0001	0.003	1.01	0.0001	0.003	1.02	1.11	1.10
ψ_{10}	0.052	175.7	0.0006	0.007	0.93	0.0004	0.006	0.97	1.26	1.22
ψ_{01}	0.057	522.8	0.0000	0.003	1.02	0.0000	0.003	1.04	1.10	1.08
ψ_{11}	0.155	316.8	0.0018	0.015	0.93	0.0014	0.012	0.98	1.50	1.41
ψ	0.093		0.0013	0.019	0.92	0.0011	0.014	0.98	1.99	1.86

The estimators are evaluated by the sample mean of errors (Bias), the Monte Carlo standard deviation (SD), and the ratio of the limiting variance to the scaled Monte Carlo variance. The observed Monte Carlo (MC) efficiency gains from exploiting independence are presented and ratio of asymptotic bounds (Bounds). $N_{II,jk}$ is the average number of observations with biomarker-treatment level $(B, X) = (j, k)$ in the phase-two subsample.

Table 3.2: Evaluation of Variance Estimators With and Without Using Independence

			$\mathcal{M}^E$				$\mathcal{M}^{E,\perp\!\!\!\perp}$			
	$\psi_{ij,0}$	$N_{II,jk}$	$\text{Var}(\widehat{\psi})$	$\widehat{\text{Var}}(\widehat{\psi})$	σ_0^2	Cov	Var	$\widehat{\text{Var}}$	σ_0^2	Cov
$\psi \approx 0$										
ψ_{00}	0.048	525.0	0.17	0.18	0.17	0.949	0.15	0.16	0.16	0.957
ψ_{10}	0.052	191.6	0.94	0.91	0.88	0.948	0.73	0.72	0.71	0.950
ψ_{01}	0.057	572.7	0.21	0.22	0.22	0.953	0.19	0.19	0.19	0.952
ψ_{11}	0.063	210.1	1.17	1.17	1.11	0.949	0.89	0.91	0.87	0.952
ψ	0.001		3.32	3.22	3.11	0.953	1.73	1.75	1.71	0.952
$\psi \approx 0.04$										
ψ_{00}	0.048	506.5	0.18	0.18	0.18	0.948	0.16	0.16	0.16	0.954
ψ_{10}	0.052	184.3	0.99	0.94	0.91	0.944	0.77	0.75	0.74	0.946
ψ_{01}	0.057	549.9	0.22	0.23	0.23	0.955	0.20	0.21	0.20	0.954
ψ_{11}	0.100	258.5	2.31	2.29	2.16	0.951	1.64	1.68	1.61	0.950
ψ	0.038		4.82	4.65	4.47	0.948	2.44	2.47	2.40	0.953
$\psi \approx 0.09$										
ψ_{00}	0.048	484.3	0.18	0.18	0.18	0.948	0.16	0.17	0.17	0.955
ψ_{10}	0.052	175.7	1.02	0.99	0.95	0.946	0.81	0.80	0.78	0.946
ψ_{01}	0.057	522.8	0.24	0.25	0.25	0.952	0.22	0.23	0.23	0.953
ψ_{11}	0.155	316.8	4.51	4.41	4.18	0.950	3.02	3.08	2.96	0.952
ψ	0.093		7.54	7.20	6.92	0.948	3.79	3.84	3.73	0.950

The estimators are evaluated by the scaled Monte Carlo variance (Var), mean asymptotic variance estimate ($\widehat{\text{Var}}$), and the coverage of 95% confidence intervals (Cov). The true limiting variances σ_0^2 are included for reference. $N_{II,jk}$ is the average number of observations with biomarker-treatment level $(B, X) = (j, k)$ in the phase-two subsample.

Table 3.3: Asymptotic Behavior of Estimators Exploiting Known Independence

N_I	N_{II}	Bias	Var	$\frac{\sigma_0^2}{Var}$	$\widehat{\text{Var}}$	Cov	RE
$\psi \approx 0$							
5000	374.7	0.000	1.999	0.856	1.938	0.947	2.000
10000	750.0	0.000	1.793	0.955	1.810	0.952	1.939
20000	1499.4	0.000	1.729	0.990	1.748	0.952	1.917
40000	2999.9	0.000	1.666	1.027	1.730	0.956	1.868
60000	4497.6	0.000	1.766	0.970	1.718	0.945	1.868
∞			1.712	1.000	1.712	0.950	1.817
$\psi \approx 0.04$							
5000	374.6	0.002	2.889	0.832	2.747	0.946	2.051
10000	749.8	0.001	2.555	0.941	2.562	0.952	1.997
20000	1499.2	0.000	2.435	0.987	2.472	0.953	1.979
40000	2999.1	0.000	2.376	1.012	2.443	0.955	1.915
60000	4496.3	0.000	2.434	0.988	2.427	0.948	1.912
∞			2.404	1.000	2.404	0.950	1.858
$\psi \approx 0.09$							
5000	374.8	0.005	4.561	0.817	4.274	0.947	2.070
10000	750.0	0.002	3.971	0.939	3.981	0.949	1.986
20000	1499.6	0.001	3.788	0.984	3.841	0.950	1.991
40000	2999.2	0.000	3.731	0.999	3.796	0.953	1.908
60000	4496.7	0.000	3.788	0.984	3.770	0.951	1.911
∞			3.729	1.000	3.729	0.950	1.855

The estimators are evaluated by the sample mean of errors (Bias), N_I times the Monte Carlo variance (Var), mean variance estimate ($\widehat{\text{Var}}$), ratio of the limiting variance to the scaled Monte Carlo variance, mean variance estimate ($\widehat{\text{Var}}$), coverage of 95% confidence intervals (Cov), and the observed Monte Carlo efficiency gain from utilizing independence (RE). Asymptotic limits are labeled as $N_I = \infty$ and included for reference, N_I is the phase-one sample size and N_{II} is the average phase-two sample size (7.5% of N_I).

Chapter 4

CONCLUSIONS AND FUTURE RESEARCH

4.1 Summary

In this part of the dissertation, we focused on a measure of additive interaction, P-AERI, defined as a linear combination of bivariate versions of the so-called treatment-specific mean. This estimand has a characteristic profile seemingly unique among the methodology available for assessment of gene-treatment, and more generally biomarker- treatment interactions. We considered estimation of P-AERI from a two-phase sample for which phase one is a trial cohort. The primary contribution of this work was establishing efficient inference for P-AERI, reflecting the independence between randomized treatment and baseline covariates.

Using general results on the correspondence between efficient gradients for full-data and for two-phase experimental data, we calculated the influence function for RAL estimators efficient in unconstrained models. Under a working finiteness assumption on the covariate distribution, we used likelihood methods to compute an influence function for RAL estimators efficient in the model constrained to reflect independence. The candidate influence function was argued to be the same as that for models with continuous covariates.

One-step corrected substitution estimators, $\psi_n^{OS,NP}$ and $\psi_n^{OS,\perp\!\!\!\perp}$, efficient in either the unconstrained or constrained ($\perp\!\!\!\perp$) models were proposed. Each estimator relies on consistent estimates of distributional features in accordance with the statistical model relevant to the estimation context. These features can be estimated using flexible techniques and resulting bias corrected by using the respective influence functions.

We employed these estimators in a re-analysis of a "weak" finding from a GWIS conducted on data from the hormone-therapy trial component of the WHI. Though based on a different estimand, we observed qualitatively similar results, showing modest evidence for a

small protective effect of hormone therapy on incident diabetes was seen among women with an adenine variant at 'rs9909279', and none otherwise, to those observed in the GWIS analysis. In addition to observing expected efficiency gains, when accounting for independence between randomized hormone therapy and gene variant, the re-analysis provided opportunity to highlight the efficient inference for the four treatment-variant specific constituents as well as the two gene variant specific treatment effects estimands that are established in the course of conducting inference on P-AERI. The proposed approach provides a considerable amount of information for characterizing the joint effect of an intervention and biomarker.

We presented three important characteristics of the proposed estimators: (1) the potential for efficiency gain due to exploiting independence, (2) the potential for efficiency gain due to adjusting for baseline precision variables, and (3) double robustness. The frequentist performance of the one-step estimators was examined through simulation study. Overall, results showed that the bias, variance estimates, and coverage of Wald confidence intervals of the one-step estimators were well-behaved under the scenarios considered. Of particular interest, our results exhibited efficiency gains, due to leveraging known independence, of magnitude similar to those reported for other interaction estimands.

4.2 Discussion

While the focus of this work has been on estimation of population-averaged additive excess risk due to interaction (P-AERI), from data collected under a two-phase sampling structure in which one of the exposures is missing by design, these techniques are applicable to less specific scenarios. To start, estimators of other parameters similarly constructed out of the same building blocks, ψ_{jk}, follow immediately from application of the delta method. For example, a population-level measure of multiplicative interaction (P-MERI) could be defined as $\frac{\psi_{11}\psi_{00}}{\psi_{10}\psi_{01}}$ and similarly P-RERI as $\frac{\text{P-AERI}}{\psi_{00}}$ which are analogous to the multiplicative interaction in a saturated log-linear model and RERI estimands, respectively. Unlike for P-AERI, the lack of linearity in P-MERI and P-RERI prohibits their interpretation as the population-averaged covariate-specific multiplicative, or respectively relative, excess risk due

to interaction. Efficiency gains following from exploiting independence in estimation of each component should be inherited by the final estimand.

Another simple adaptation can be used to estimate the main effect between outcome and a biomarker, or other exposure, measured only at phase two, $\psi_1 - \psi_0$. This follows immediately from the omission of the treatment variable in all calculations. Of course, no concept analogous to conditional independence between biomarker and treatment exists in this application, but knowledge of the single exposure distribution could be analogously utilized for efficiency gain. Finally, in any sample for which the biomarker and treatment, or other two exposures, have known independence conditional on covariates, the methods presented here can be safely applied. However, we emphasize that for a study nested within a trial, independence is guaranteed. We caution against application of these methods in contexts where independence is plausible but not truly known, as we expect them to be equally sensitive to this assumption as the case-only estimator.

The biomarker, treatment, and outcome were all taken to be binary in the work presented. Extension of the two binary exposures, biomarker status and treatment assignment, to two categorical variables is immediate. As in Han et al. (2012), where there would be JK interaction estimators, where exposures $bx = 11, 10, 01, 00$ generalize to $bx = jk, j0, 0k, 00$ for each of the J non-referent levels of B and K non-referent levels of X.

Adjusting for heritage is often important in epigenetic studies. In the two-phase experiment setting, baseline self-report data could be accounted for as a stratified variable. When full genetic sequencing is available, accounting for the top three continuous principle components is a preferred practice. The current methods do not lend themselves to adjusting for continuous covariates that are only measured at phase two. Though this is a desirable extension of the work presented herein, it pushes up against the limitations of estimating continuous variables with missingness without modeling assumptions.

The techniques presented here facilitate using flexible approaches to estimation that can ameliorate model mis-specification concerns. As long as a second order reminder term in the asymptotic expansion of the proposed estimator remains asymptotically negligible,

the correction step can account for non-negligible bias introduced by a certain amount of smoothing, which is a key aspect of many nonparametric estimation techniques. While this work is not presented as a high-dimensional method, it does afford completely non-parametric approaches to adjusting for up to three continuous covariates, which can often meet the needs of genetic epidemiology and biomarker development studies.Initial estimates of distributional parameters constructed from ensemble methods could afford further flexibility.

We used a one-step corrected estimator to demonstrate the utility and properties of targeted methods. Of related interest is the substitution corrected estimator, a targeted maximum likelihood estimator, as it has been established to have improved finite sample performance for binary outcomes.

4.3 Future Research

This part of the dissertation focuses on assessment of a treatment-biomarker interaction. Though the particular estimand introduced was motivated by population-level considerations, such as retaining a marginal interpretation, a secondary biomarker study nested within a trial has aims characteristic of early- phase biomarker research. For example, the presented data re-analysis was originally conducted in the context of a GWIS. Incorporating the presented results in a framework for multiple testing would be likely prove useful for studies interested in estimating biomarker-treatment interactions.

Accounting for loss to follow-up, an important consideration in many long-term trials, has not been addressed and would be another practical extension.

The method presented here relied on simple Bernoulli sampling of the phase-two subset. Bernoulli sampling produces data that is independently and identically distributed, whereas common fixed sample size practices create dependent sampling. Recent work by Ma (2010) established a first-order equivalence between fixed count and random case-control samples. It is of interest to examine to what extent our estimators inherit this good approximation property, when evaluated on case-control phase-two samples selected via stratified matching with fixed counts.

Many nested case-control samples employ matching schemes that are complex and do not obviously fit a standard statistical framework. It would be of practical value to assess the feasibility of approximating such approaches by a simple two-phase design with Bernoulli sampling and then to evaluate estimation of P-AERI by applying the results of this dissertation assuming the approximated sampling scheme.

In a two-phase study nested within a trial, the sampling probabilities are under investigator control. A natural question is whether the efficient influence functions can guide decisions for study design. The influence functions involve many features of the unknown data generating mechanism and their variance requires averaging over the covariate distribution. Even so, numerical study is easily supported and useful if one is willing to make informed assumptions on these parameters. It would be of value to find more tractable expressions that yield higher-level analytic insights to efficient design.

BIBLIOGRAPHY

Paul S. Albert, Duminda Ratnasinghe, Joseph Tangrea, and Sholom Wacholder. Limitations of the Case-only Design for Identifying Gene-Environment Interactions. *American Journal of Epidemiology*, 154(8):687–693, 2001.

Thomas Atwater, Christine M Cook, and Pierre P Massion. The Pursuit of Noninvasive Diagnosis of Lung Cancer. *Seminars in Respiratory and Critical Care Medicine*, 37(5): 670–680, 2016.

Stuart G Baker. Putting risk prediction in perspective: relative utility curves. *Journal of the National Cancer Institute*, 101(22):1538–1542, 2009.

Stuart G Baker, Nancy R Cook, Andrew Vickers, and Barnett S Kramer. Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):729–748, 2009.

Stuart G. Baker, Barnett S. Kramer, and Sudhir Srivastava. Markers for early detection of cancer: Statistical guidelines for nested case-control studies. *BMC Medical Research Methodology*, 2(1):4, 2002.

Stuart G Baker, Ben Van Calster, and Ewout W Steyerberg. Evaluating a new marker for risk prediction using the test tradeoff: an update. *The international journal of biostatistics*, 8 (1):1–37, 2012.

Seetharaman Balasenthil, Ying Huang, Suyu Liu, Tracey Marsh, Jinyun Chen, Sanford A. Stass, Debra KuKuruga, Randall Brand, Nanyue Chen, Marsha L. Frazier, J. Jack Lee, Sudhir Srivastava, Subrata Sen, and Ann McNeill Killary. A Plasma Biomarker Panel

to Identify Surgically Resectable Early-Stage Pancreatic Cancer. *JNCI: Journal of the National Cancer Institute*, 109(8):djw341, 2017.

OE Barndorff-Nielsen, DR Cox, and N Reid. The role of differential geometry in statistical theory. *International Statistical Review/Revue Internationale de Statistique*, pages 83–96, 1986.

James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Springer, 1993.

Ebony B. Bookman, Corina Din-Lovinescu, Bradford B. Worrall, Teri A. Manolio, Siiri N. Bennett, Cathy Laurie, Daniel B. Mirel, Kimberly F. Doheny, Garnet L. Anderson, Kate Wehr, Richard Weinshilboum, and Donna T. Chen. Incidental genetic findings in randomized clinical trials: recommendations from the Genomics and Randomized Trials Network (GARNET). *Genome Medicine*, 5(1):7, 2013.

Patrick M Bossuyt, Johannes B Reitsma, David E Bruns, Constantine A Gatsonis, Paul P Glasziou, Les M Irwig, David Moher, Drummond Rennie, Henrica CW De Vet, and Jeroen G Lijmer. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical chemistry*, 49(1):7–18, 2003a.

Patrick M Bossuyt, Johannes B Reitsma, David E Bruns, Constantine A Gatsonis, Paul P Glasziou, Les M Irwig, Jeroen G Lijmer, David Moher, Drummond Rennie, and Henrica CW De Vet. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Clinical chemistry and laboratory medicine*, 41(1):68–73, 2003b.

Jordan Brooks, Mark J van der Laan, and Alan S Go. Targeted maximum likelihood estimation for prediction calibration. *The international journal of biostatistics*, 8(1), 2012.

Tianxi Cai, Margaret Sullivan Pepe, Yingye Zheng, Thomas Lumley, and Nancy Swords Jenny. The sensitivity and specificity of markers for event times. *Biostatistics*, 7(2): 182–197, 2006.

Gregory Campbell. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13(5-7):499–508, 1994.

Peter R Casson, Stephen A Krawetz, Michael P Diamond, Heping Zhang, Richard S Legro, William D Schlaff, Christos Coutifaris, Robert G Brzyski, Gregory M Christman, Nanette Santoro, and Ester Eisenberg. Proactively Establishing a Biologic Specimens Repository for Large Clinical Trials: An Idea Whose Time has Come. *Systems biology in reproductive medicine*, 57(5), 2011. Syst Biol Reprod Med.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305 – 334, 1987.

Nilanjan Chatterjee and Raymond J Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92(2):399–418, 2005.

Nilanjan Chatterjee and Yi-Hau Chen. Maximum likelihood inference on a mixed conditionally and marginally specified regression model for genetic epidemiologic studies with two-phase sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):123–142, 2007.

Ralph B D'Agostino, Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General cardiovascular risk profile for use in primary care the Framingham Heart Study. *Circulation*, 117(6):743–753, 2008.

James Y. Dai, Michael LeBlanc, and Charles Kooperberg. Semiparametric Estimation Exploiting Covariate Independence in Two-Phase Randomized Trials. *Biometrics*, 65(1): 178–187, 2009.

Geoffrey Decrouez and Andrew P Robinson. Confidence intervals for the weighted sum of two independent binomial proportions. *Australian & New Zealand Journal of Statistics*, 54(3):281–299, 2012.

David A Degras. Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, pages 1735–1765, 2011.

Sandrine Dudoit, Mark J van der Laan, and Katherine S Pollard. Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–69, 2004.

E. B. Elkin and A. J Vickers. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):pp. 565+, 2006.

Ruth Etzioni, Margaret Pepe, Gary Longton, Chengcheng Hu, and Gary Goodman. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*, 19(3):242–251, 1999.

J. Scott Ferguson, Ryan Van Wert, Yoonha Choi, Michael J. Rosenbluth, Kate Porta Smith, Jing Huang, and Avrum Spira. Impact of a bronchial genomic classifier on clinical decision making in patients undergoing diagnostic evaluation for lung cancer. *BMC Pulmonary Medicine*, 16(1):66, 2016.

FHS. Framingham Heart Study. https://www.framinghamheartstudy.org, 2017.

FORUM. Stratified, personalised or P4 medicine: a new direction for placing the patient at the centre of healthcare and health education. Meeting report, The Academy of Medical Sciences, 2015.

Alexandre M Furman, Jihane Zaza Dit Yafawi, and Ayman O Soubani. An update on the evaluation and management of small pulmonary nodules. *Future Oncology*, 9(6):855–865, 2013.

GARNET-WHI. Garnet Study, Women's Health Initiative. https://www.whi.org/researchers/data/WHIStudies/StudySites/M13/pages/home.aspx, 2015.

Genomics England. 100,000 Genomes Project. https://www.genomicsengland.co.uk/the-100000-genomes-project, 2012.

Paul L. F. Giangrande. The history of blood transfusion. *British Journal of Haematology*, 110(4):758–767, 2000.

Francis M Giardiello, John I Allen, Jennifer E Axilbund, C Richard Boland, Carol A Burke, Randall W Burt, James M Church, Jason A Dominitz, David A Johnson, Tonya Kaltenbach, et al. Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-Society Task Force on Colorectal Cancer. *Gastroenterology*, 147(2):502–526, 2014.

RD Gill, MJ van der Laan, and JA Wellner. Inefficient estimators of the bivariate survival function for three models. In *Annales de l'Institut Henri Poincare.*, 1995.

Michael K Gould, Jessica Donington, William R Lynch, Peter J Mazzone, David E Midthun, David P Naidich, and Renda Soylemez Wiener. Evaluation of individuals with pulmonary nodules: When is it lung cancer?: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines. *CHEST Journal*, 143(5_suppl):e93S–e120S, 2013.

Sander Greenland. Additive risk versus additive relative risk models. *Epidemiology*, pages 32–36, 1993.

Wen Gu and Margaret Pepe. Measures to summarize and compare the predictive capacity of markers. *The International Journal of Biostatistics*, 5(1), 2009.

Summer S. Han, Philip S. Rosenberg, Montse Garcia-Closas, Jonine D. Figueroa, Debra Silverman, Stephen J. Chanock, Nathaniel Rothman, and Nilanjan Chatterjee. Likelihood Ratio Test for Detecting Gene (G)-Environment (E) Interactions Under an Additive Risk Model Exploiting G-E Independence for Case-Control Data. *American Journal of Epidemiology*, 176(11):1060–1067, 2012.

Daniel F. Hayes, Robert C. Bast, Christopher E. Desch, Herbert Fritsche, Jr., Nancy E. Kemeny, J. Milburn Jessup, Gershon Y. Locker, John S. Macdonald, Robert G. Mennel, Larry Norton, Peter Ravdin, Sheila Taube, and Rodger J. Winn. Tumor Marker Utility Grading System: a Framework to Evaluate Clinical Utility of Tumor Markers. *JNCI: Journal of the National Cancer Institute*, 88(20):1456, 1996.

Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

Alan E Hubbard, Sara Kherad-Pajouh, and Mark J van der Laan. Statistical Inference for Data Adaptive Target Parameters. *The international journal of biostatistics*, 12(1):3–19, 2016.

Carolyn M. Hutter, Leah E. Mechanic, Nilanjan Chatterjee, Peter Kraft, Elizabeth M. Gillanders, and on behalf of the NCI Gene-Environment Think Tank. Gene-Environment Interactions in Cancer Epidemiology: A National Cancer Institute Think Tank Report. *Genetic Epidemiology*, 37(7):643–657, 2013.

ICH. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. http://www.ich.org/home.html, 1990.

ICH Guideline E16. BIOMARKERS RELATED TO DRUG OR BIOTECHNOLOGY PRODUCT DEVELOPMENT: CONTEXT, STRUCTURE AND FORMAT OF QUALIFICATION SUBMISSIONS. Technical report, International Council of Harmonisation, 2010.

ICH Guideline E8. GENERAL CONSIDERATIONS FOR CLINICAL TRIALS. Technical report, International Council of Harmonisation, 1997.

Holly Janes and Margaret Pepe. The optimal ratio of cases to controls for estimating the classification accuracy of a biomarker. *Biostatistics*, 7(3):456–468, 2006.

Holly Janes and Margaret Pepe. Re:"clinical usefulness of the Framingham cardiovascular risk profile beyond its statistical performance: the Tehran Lipid and Glucose Study". *American journal of epidemiology*, 177(8):864–865, 2013.

Holly Janes and Margaret S Pepe. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. *American Journal of Epidemiology*, 168(1):89–97, 2008.

Robert E. Kass. The Geometry of Asymptotic Inference. *Statistical Science*, 4(3):pp. 188–219, 1989.

Fay Kastrinos, Rohit P Ojha, Celine Leenen, Carmelita Alvero, Rowena C Mercado, Judith Balmaña, Irene Valenzuela, Francesc Balaguer, Roger Green, Noralane M Lindor, et al. Comparison of prediction models for Lynch syndrome among individuals with colorectal cancer. *Journal of the National Cancer Institute*, 108(2):djv308, 2016.

Kathleen F Kerr, Marshall D Brown, Kehao Zhu, and Holly Janes. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *Journal of Clinical Oncology*, page JCO655654, 2016.

Kathleen F Kerr, Zheyu Wang, Holly Janes, Robyn L McClelland, Bruce M Psaty, and Margaret S Pepe. Net reclassification indices for evaluating risk-prediction instruments: a critical review. *Epidemiology (Cambridge, Mass.)*, 25(1):114, 2014.

Davood Khalili, Farzad Hadaegh, Hamid Soori, Ewout W Steyerberg, Mohammadreza Bozorgmanesh, and Fereidoun Azizi. Clinical Usefulness of the Framingham Cardiovascular

Risk Profile Beyond Its Statistical Performance The Tehran Lipid and Glucose Study. *American journal of epidemiology*, page kws204, 2012.

Chris A. J. Klaassen. Consistent Estimation of the Influence Function of Locally Asymptotically Linear Estimators. *The Annals of Statistics*, 15(4):1548–1562, 1987.

Andrej N Kolmogorov. Sulla determinazione empirica di una leggi di distribuzione. *Giorn. 1st it lit Ital. Attuari*, 91(4):83, 1933.

Maarten JG Leening, Moniek M Vedder, Jacqueline CM Witteman, Michael J Pencina, and Ewout W Steyerberg. Net Reclassification Improvement: Computation, Interpretation, and ControversiesA Literature Review and Clinician's Guide. *Annals of internal medicine*, 160(2):122–131, 2014.

John A Lewis. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Statistics in medicine*, 18(15):1903–1942, 1999.

Hui Li. Cancer Precision Medicine in China. *Genomics, Proteomics and Bioinformatics*, 14 (5):325 – 328, 2016. SI: Big Data and Precision Medicine.

Yanyuan Ma. A semiparametric efficient estimator in case-control studies. *Bernoulli*, 16(2): 585–603, 2010.

Heber MacMahon, John H. M. Austin, Gordon Gamsu, Christian J. Herold, James R. Jett, David P. Naidich, Jr Edward F. Patz, and Stephen J. Swensen. Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society. *Radiology*, 237(2):395–400, 2005. PMID: 16244247.

Heber MacMahon, David P. Naidich, Jin Mo Goo, Kyung Soo Lee, Ann N. C. Leung, John R. Mayo, Atul C. Mehta, Yoshiharu Ohno, Charles A. Powell, Mathias Prokop, Geoffrey D. Rubin, Cornelia M. Schaefer-Prokop, William D. Travis, Paul E. Van Schil, and Alexander A. Bankier. Guidelines for Management of Incidental Pulmonary Nodules Detected on

CT Images: From the Fleischner Society 2017. *Radiology*, 284(1):228–243, 2017. PMID: 28240562.

H. Moch, P. R. Blank, M. Dietel, G. Elmberger, K. M. Kerr, J. Palacios, F. Penault-Llorca, G. Rossi, and T. D. Szucs. Personalized cancer medicine and the future of pathology. *Virchows Archiv*, 460(1):3–8, 2012.

Virginia A Moyer. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *Annals of internal medicine*, 160(5):330–338, 2014.

Bhramar Mukherjee and Nilanjan Chatterjee. Exploiting Gene-Environment Independence for Analysis of Case–Control Studies: An Empirical Bayes-Type Shrinkage Estimator to Trade-Off between Bias and Efficiency. *Biometrics*, 64(3):685–694, 2008.

Marjolein A. M. Mulders, Monique M. J. Walenkamp, Bente F. H. Dubois, Annelie Slaar, J. Carel Goslings, and Niels W. L. Schep. External validation of clinical decision rules for children with wrist trauma. *Pediatric Radiology*, 47(5):590–598, 2017.

David P Naidich, Alexander A Bankier, Heber MacMahon, Cornelia M Schaefer-Prokop, Massimo Pistolesi, Jin Mo Goo, Paolo Macchiarini, James D Crapo, Christian J Herold, John H Austin, et al. Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology*, 266(1):304–317, 2013.

Robert G Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, 17(8):857–872, 1998.

Office of the Press Secretary. FACT SHEET: President Obama's Precision Medicine Initiative. https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative, 2015.

National Cholesterol Education Program NCEP Expert Panel et al. Third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation,

and treatment of high blood cholesterol in adults (Adult Treatment Panel III) final report. *Circulation*, 106(25):3143, 2002.

Stephen G Pauker and Jerome P Kassirer. Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine*, 293(5):229–234, 1975.

Charles S Peirce. The numerical measure of the success of predictions. *Science*, 4(93): 453–454, 1884.

Michael J Pencina, Ralph B D'Agostino, and Ewout W Steyerberg. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in medicine*, 30(1):11–21, 2011.

Michael J Pencina, Ralph B D'Agostino, and Ramachandran S Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.

Margaret Pepe and Holly Janes. Methods for evaluating prediction performance of biomarkers and tests. In *Risk Assessment and Evaluation of Predictions*, pages 107–142. Springer, 2013.

Margaret S. Pepe. Problems With Risk Reclassification Methods for Evaluating Prediction Models. *American Journal of Epidemiology*, 2011.

Margaret S Pepe, Ziding Feng, Holly Janes, Patrick M Bossuyt, and John D Potter. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *Journal of the National Cancer Institute*, 100(20):1432–1438, 2008.

Margaret S Pepe, Holly Janes, and Christopher I Li. Net risk reclassification p values: valid or misleading? *JNCI: Journal of the National Cancer Institute*, 106(4), 2014.

Margaret S Pepe, Holly Janes, Christopher I Li, Patrick M Bossuyt, Ziding Feng, and Jørgen

Hilden. Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? *Clinical chemistry*, 62(5):737–742, 2016.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA, 2003.

Margaret Sullivan Pepe, Ruth Etzioni, Ziding Feng, John D Potter, Mary Lou Thompson, Mark Thornquist, Marcy Winget, and Yutaka Yasui. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93(14):1054–1061, 2001.

Margaret Sullivan Pepe, Jing Fan, and Christopher W Seymour. Estimating the receiver operating characteristic curve in studies that match controls to cases on covariates. *Academic radiology*, 20(7):863–873, 2013.

Edith A. Perez, Karla V. Ballman, Afshin Mashadi-Hossein, Kathleen S. Tenner, Jennifer M. Kachergus, Nadine Norton, Brian M. Necela, Jennifer M. Carr, Sean Ferree, Charles M. Perou, Frederick Baehner, Maggie Chon U. Cheang, and E. Aubrey Thompson. Intrinsic Subtype and Therapeutic Response Among HER2-Positive Breast Tumors from the NCCTG (Alliance) N9831 Trial. *JNCI: Journal of the National Cancer Institute*, 109(2):1, 2017.

Johann Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer Science Business Media, 1982.

Walter W. Piegorsch, Clarice R. Weinberg, and Jack A. Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13(2):153–162, 1994.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

Ross L. Prentice. Tumor Marker Utility Grading System. *JNCI: Journal of the National Cancer Institute*, 88(20):1424, 1996.

Robert N Proctor. The history of the discovery of the cigarette–lung cancer link: evidentiary traditions, corporate denial, global toll. *Tobacco Control*, 21(2):87–91, 2012.

Nancy Reid. Influence functions for censored data. *The Annals of Statistics*, pages 78–92, 1981.

Sherri Rose and Mark van der Laan. *Causal Inference for Observational and Experimental Data Series: Springer Series in Statistics van der Laan, Mark J., Rose, Sherri Targeted Learning; Causal Inference for Observational and Experimental Data Series*. Springer, 2011a.

Sherri Rose and Mark van der Laan. A Double Robust Approach to Causal Effects in Case-Control Studies. *American Journal of Epidemiology*, 179(6):663–669, 2014a.

Sherri Rose and Mark van der Laan. Rose and van der Laan Respond to "Some Advantages of the Relative Excess Risk due to Interaction". *American Journal of Epidemiology*, 179 (6):672–673, 2014b.

Sherri Rose and Mark J van der Laan. A targeted maximum likelihood estimator for two-stage designs. *The international journal of biostatistics*, 7(1):1–21, 2011b.

Liana S Rosenthal, Daniel Drake, Roy N Alcalay, Debra Babcock, F DuBois Bowman, Alice Chen-Plotkin, Ted M Dawson, Richard B Dewey, Dwight C German, Xuemei Huang, et al. The NINDS Parkinson's disease biomarkers program. *Movement Disorders*, 2015.

KJ Rothman. *Modern Epidemiology*. Boston: Little, Brown, 1986.

Sonali Sethi and Scott Parrish. Incidental nodule management—should there be a formal process? *Journal of Thoracic Disease*, 8(Suppl 6):S494, 2016.

Gerard A. Silvestri, Anil Vachani, Duncan Whitney, Michael Elashoff, Kate Porta Smith, J. Scott Ferguson, Ed Parsons, Nandita Mitra, Jerome Brody, Marc E. Lenburg, and Avrum Spira. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *New England Journal of Medicine*, 373(3):243–251, 2015. PMID: 25981554.

Tania Simoncelli, Lisa Barclay, Khaled Bouri, Kathleen Burns, Kate Cook, Ross Filice, James Fuscoe, Francis Kalush, Chava Kimchi-Sarfaty, Sheryl Kochman, Siyeon Lee, Ernest Litwack, Peter Lurie, William Maisel, Elizabeth Mansfield, Peter Marks, Donna Mendrick, Karen Midthun, Baitang Ning, Michael Pacanowski, Barbara Parsons, Karen Riley, Zuben Sauna, Jeffrey Shuren, William Slikker, Jr., Stephen Spielberg, Julie Tierney, Weida Tong, Jill Warner, Carolyn Wilson, Janet Woodcock, Denise Zavagno, and Issam Zineh. Paving the Way for Paving the Way for Personalized Medicine, FDA's Role in a New Era of Medical Product Development. https://www.ucsf.edu/news/2016/06/403221/white-house-gates-foundation-summit-explores-applying-precision-medicine-public, 2013.

Anders Skrondal. Interaction as Departure from Additivity in Case-Control Studies: A Cautionary Note. *American Journal of Epidemiology*, 158(3):251–258, 2003.

Annelie Slaar, Monique MJ Walenkamp, Abdelali Bentohami, Mario Maas, Rick R van Rijn, Ewout W Steyerberg, L Cara Jager, Nico L Sosef, Romuald van Velde, Jan M Ultee, et al. A clinical decision rule for the use of plain radiography in children after acute wrist injury: development and external validation of the Amsterdam Pediatric Wrist Rules. *Pediatric radiology*, 46(1):50–60, 2016.

Sudhir Srivastava and Barnett S Kramer. Early detection cancer research network. *Laboratory Investigation*, 80(8):1147–1148, 2000.

Ewout W Steyerberg, Michael J Pencina, Hester F Lingsma, Michael W Kattan, Andrew J Vickers, and Ben Van Calster. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *European journal of clinical investigation*, 42(2):216–228, 2012.

Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128, 2010.

Kyle Strimbu and Jorge A Tavel. What are biomarkers? *Current Opinion in HIV and AIDS*, 5(6):463, 2010.

Stephen J Swensen, Marc D Silverstein, Duane M Ilstrup, Cathy D Schleck, and Eric S Edell. The probability of malignancy in solitary pulmonary nodules: application to small radiologically indeterminate nodules. *Archives of internal medicine*, 157(8):849–855, 1997.

Bethany B. Tan, Kevin R. Flaherty, Ella A. Kazerooni, and Mark D. Lannettoni. The solitary pulmonary nodule. *Chest*, 123(1suppl):89S–96S, 2003.

Sheila E Taube, Gary M Clark, Janet E Dancey, Lisa M McShane, Caroline C Sigman, and Steven I Gutman. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. *Journal of the National Cancer Institute*, 2009.

Eric Tchetgen Tchetgen, Tamar Sofer, and Benedict HW Wong. A General Approach to Detect Gene (G)-environment (E) Additive Interaction Leveraging GE Independence in Case-control Studies. *Bepress*, 2014.

Lu Tian, Tianxi Cai, Els Goetghebeur, and LJ Wei. Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika*, 94(2):297–311, 2007.

Athanasios Tsalatsanis, Iztok Hozo, Andrew Vickers, and Benjamin Djulbegovic. A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. *BMC Medical Informatics and Decision Making*, 10(1): 1, 2010.

David M Umbach and Clarice R Weinberg. Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in medicine*, 16(15):1731–1743, 1997.

University of California at San Francisco, Bill and Melinda Gates Foundation, White House Office of Science and Technology Policy. *Precision Public Health: The First 1,000 Days*, https://www.ucsf.edu/news/2016/06/403221/white-house-gates-foundation-summit-explores-applying-precision-medicine-public, 2016.

Anil Vachani, Duncan H. Whitney, Edward C. Parsons, Marc Lenburg, J. Scott Ferguson, Gerard A. Silvestri, and Avrum Spira. Clinical utility of a bronchial genomic classifier in patients with suspected lung cancer. *Chest*, 150(1):210–218, 2016.

Ben Van Calster, Andrew J Vickers, Michael J Pencina, Stuart G Baker, Dirk Timmerman, and Ewout W Steyerberg. Evaluation of markers and risk prediction models overview of relationships between NRI and decision-analytic measures. *Medical Decision Making*, 33 (4):490–501, 2013.

Mark J van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, 4(1):1–57, 2008.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Mark J van der Laan and James M Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2003.

Mark J Van der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning, 2006.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.

Tyler J. VanderWeele and Stijn Vansteelandt. A Weighting Approach to Causal Effects and Additive Interaction in Case-Control Studies: Marginal Structural Linear Odds Models. *American Journal of Epidemiology*, 174(10):1197–1203, 2011.

Tyler J. VanderWeele and Stijn Vansteelandt. Invited Commentary: Some Advantages of the Relative Excess Risk due to Interaction (RERI)—Towards Better Estimators of Additive Interaction. *American Journal of Epidemiology*, 179(6):670–671, 2014.

Stijn Vansteelandt, Tyler J. VanderWeele, Eric J. Tchetgen, and James M. Robins. Multiply Robust Inference for Statistical Interactions. *Journal of the American Statistical Association*, 103(484):1693–1704, 2008.

Andrew J Vickers, Angel M Cronin, Elena B Elkin, and Mithat Gonen. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC medical informatics and decision making*, 8(1):53, 2008.

S Wassertheil-Smoller, S Hendrix, M Limacher, and et al. Effect of estrogen plus progestin on stroke in postmenopausal women: The women's health initiative: a randomized trial. *JAMA*, 289(20):2673–2684, 2003.

WHI Writing Group. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the women's health initiative randomized controlled trial. *JAMA*, 288(3):321–333, 2002.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

Appendix A

METHODOLOGICAL FRAMEWORK

Reference texts for this material are found in Bickel et al. (1993); Van der Vaart (2000) and Pfanzagl (1982).

A.1 Asymptotically Linear Estimators

Point Estimators

An estimator ψ_n of ψ_0, under $i.i.d$ sampling of O_i from P_0 is asymptotically linear if it satisfies

$$\psi_n - \psi_0 = \frac{1}{n}\sum_i \mathrm{IF}_0(O_i) + o_P(n^{-\frac{1}{2}}).$$

for a function, IF_0, of the observation that has zero-mean and finite variance, under P_0. In words, the centered estimator, up to an asymptotically negligible term, can be written as a sample average of the suitably transformed observations. The transformation, IF, is called the influence function of the estimator. The influence function is a score[1] that captures the dominant behavior of asymptotic equivalence classes of estimators and provides a convenient method for calculating the asymptotics of compositions of estimators.

The empirical estimator ρ_n, is an asymptotically linear estimators of $\rho_0 = Pr(D = 1)$.

[1]It is an element of $\mathcal{L}(P_0)_2^0$, meaning is has zero mean and finite second moment: $\sigma_0^2 < \infty$, with respect to the true data generating mechanism P_0. Technically, it could be the limit of a sequence of scores.

This can be seen via direct algebraic manipulation:

$$\begin{aligned}\rho_n - \rho_0 &= \frac{1}{n}\sum_i D_i - \rho_0 \\ &= \frac{1}{n}\sum_i (D_i - \rho_0) \\ &= \frac{1}{n}\sum_i \mathrm{IF}_{\rho,0}(D_i).\end{aligned}$$

The influence function of ρ_n is $\mathrm{IF}_{\rho,0}(d) = d - \rho_0$. By inspection, the remainder term is not just $o_P(n^{-\frac{1}{2}})$, it is exactly zero.

In general, the remainder term follows from what is conceptually a sophisticated Taylor approximation:

$$\Psi(P_n) - \Psi(P_0) = -\int IF(P_n)(o)dP_0\,(o) + R(P_n, P_0).$$

Estimators of the Limiting Variance

The Central Limit Theorem governs the asymptotics of such estimators

$$\sqrt{n}\,(\psi_n - \psi_0) \xrightarrow{d} N\left(0, \sigma_0^2\right),$$

The limiting variance, of the centered and scaled estimator, is the variance of the observation transformed by the influence function: $\sigma_0^2 = \mathbb{E}_0[\mathrm{IF}_0^2(O)]$.

For some estimators the limiting variance can be calculated explicitly and simplifies to a simple expression defined in terms of some key features of the underlying data generating mechanism. In the example of using a simple sample mean to estimate a Bernoulli probability:

$$\begin{aligned}\sigma_0^2 &= \mathbb{E}_0[(D - \rho_0)^2] \\ &= \mathbb{E}_0[D^2] - \rho_0^2 \\ &= \rho_0\,(1 - -\rho_0)\,,\end{aligned}$$

as expected. An empirical estimator σ_n^2 of σ_0^2 is defined by simply substituting ρ_n for its unknown counterpart in the above expression.

In some cases where this in not feasible, one can define σ_n^2 using the empirical distribution in place of P_0 and empirical estimators in place of features of the unknown data generating meachanishm that appear in the influence function. For ρ_n, this equals:

$$\begin{aligned} \sigma_n^2 &:= \mathbb{E}_n[(D-\rho_n)^2] \\ &= \frac{1}{n}\sum_i (D_i-\rho_n)^2 \end{aligned}$$

which simplifies to the same expression established previously.

A.2 Mathematical Framework

The mathematical discipline of differential geometry was developed in order to generalize the tools of calculus and analysis to abstract spaces. Statistical inference, following theoretical physics, is an important application of this theory (Kass, 1989; Barndorff-Nielsen et al., 1986).

The abstract space of statistical inference is a statistical model, $\mathcal{M}$, which is defined as a set of candidate probability distributions. The tangent space to a statistical model $\mathcal{M}$ at a point $P \in \mathcal{M}$ is defined as the closed linear span of scores of parametric submodels, P_ϵ through P at $\epsilon = 0$ and is denoted $T_P\mathcal{M}$. All scores have mean 0 and finite variance with respect to the probability distribution P, thus $T_P\mathcal{M} \subseteq L_2^0(P)$, the mean zero subspace of the vector space of all functions square integrable with respect to the measure induced by P. For nonparametric models, $T_P\mathcal{M}^{NP} = \mathcal{L}_2^0(P)$. All other statistical models have tangent spaces that are sub-spaces of the $\mathcal{L}_2^0(P)$ vector space. Parametric models have finite dimensional tangent spaces. More generally, tangent spaces are Hilbert Spaces.

We now consider an estimand, ψ_0, as the value of a parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$, from a, to the reals, evaluated at a paricular point: P_0, the unknown truth. In symbols we write $\psi_0 := \Psi(P_0)$. A sufficiently smooth map Ψ from $\mathcal{M}$ to $\mathbb{R}^d$ has a derivative denoted $\dot{\Psi}$. At any $P \in \mathcal{M}$, $\dot{\Psi}(P)$ is a bounded linear operator on the tangent space to $\mathcal{M}$ at P. This linear operator has a representative called the canonical gradient, $D^*(P) \in T_P\mathcal{M}$, which satisfies $\langle D^*(P), s\rangle_P = \dot{\Psi}s,\ \forall s \in T_P\mathcal{M}$, where $\langle\cdot,\cdot\rangle_P$ is the inner product in $\mathcal{L}_2(P)$ induced

by P through expectation: $\langle f, g \rangle_P := \int f(o)g(o)dP(o)$. For any element t of $\mathcal{L}_2^0(P)$ that is orthogonal to all $s \in T_P\mathcal{M}$, $D(P) = D^*(P) + t$ is also a gradient of Ψ, albeit not the canonical one; it represents the derivative of Ψ in all directions determined by fluctuations in $\mathcal{M}$.

These concepts assume the existence of sufficient structure on the model, fluctuations, and parameter mappings (e.g., differentiability in quadratic mean and Haddamard differentiability).

A.3 Regular Asymptotically Linear Estimators

Regular Estimator

A notion of regularity connects the use of the mathematical framework to the study of asymptotic inference of estimators. A locally regular estimator is one for which the limiting distribution of $\sqrt{n}\,(\psi_n - \psi_0)$ is the same as that of $\sqrt{n}\,(\psi_n - \psi_{0n})$ where ψ_{0n} is a sequence of values approacing ψ_0 and defined by any fluctuation P_ϵ through P_0, with a well-defined direction, according to $\psi_{0n} = \Psi(P_{n^{-1/2}})$. When this holds uniformly over the statistical model the estimator is simply called regular.

RAL

Simply put, the mathematical framework applies to the study of regular asymptotically linear (RAL) estimators. For example, there is an efficiency theory for RAL estimators: of all RAL estimators of a given estimand over a statistical model, the influence function with the smallest variance determines the variance bound and all of its corresponding estimators are efficient. There is a correspondence between the influence function of a RAL estimator, ψ_n under sampling from P_0 and a representative of the derivative of Ψ evaluated at P_0 (Klaassen, 1987). This provides a mechanism for calculating influence functions for classes of estimators.

Formally, a gradient represents the derivative of a deterministic function defined on a set of probability measures (i.e., a statistical model). An asymptotically linear statistical

estimator of $\Psi(P_0)$ has an influence function which governs its asymptotic behavior. Under regularity the two notions merge and influence functions of RAL estimators of $\Psi(P_0)$ are also gradients of Ψ, evaluated at P_0. Under this identification, canonical gradients correspond to efficient influence functions and efficiency, like the defintiion of canonical, is always with respect to estimation over a particular statistical model. For brevity, and to reinforce that our interest is in construction of RAL estimators, we will use the single term influence function throughout, even when the term gradient should rightly be used.

Efficiency Theory

Over the set of all regular and asymptotically linear (RAL) estimators, the influence function yielding the smallest variance is called the efficient influence function and the corresponding estimators are efficient. An asterix '*' is used throughout to highlight efficiency, e.g., D^*. Detailed treatment of these concepts can be found in Bickel et al. (1993).

A.4 Missing Data Framework

For clarity, constructs relevent to general missing data structures are stated in terms of the data structure introduced in the article.

Let $\mathcal{M}^F$ denote the set of possible data-generating mechanisms for the full data $O^F = (Y, G, X, W) \sim P^F$ that one would ideally observe and $\mathcal{M}^{exp}$ the analog for the observed experimental data units, $O^{exp} = (\Delta, \Delta G, Y, X, W) \sim P^{exp}$. Implicitly, $\mathcal{M}^{exp}$ relies on a model for the set of probability distributions describing the possible missingness mechanisms Π, which here we explicitly denote $\mathcal{M}^{\Delta}$. The coarsening of a full data probability distribution into an experimental data probability distribution is captured by a bijective map between the statistical models.

$$f_c : \mathcal{M}^{\Delta} \times \mathcal{M}^F \to \mathcal{M}^{exp}$$

defined by $f_c\left(\Pi, P^F\right) = P^{exp}$ and stated in terms of the induced probability measures

$$dP^{exp}\left(o^{exp}\right) = d\Pi\left(\delta; y, x, w\right) dP^F(y, g, x, w)^{\delta} dP^F(y, x, w)^{1-\delta}$$

where $(\Delta, O^F) \sim (\Pi, P^F)$ and $O^{exp} = (\Delta, f_c(O^F)) \sim P^{exp}$, where f_c is also used to denote coarsening of the random variables. Missingness at random is reflected through the independence of Π on g, the variable subject to missingness.

Under sufficient smoothness of the coarsening map, a regular parametric submodel P_{ϵ}^F of $\mathcal{M}^F$ induces a regular parametric submodel $f_c\left(P_{\epsilon}^F\right) = P_{\epsilon}^{exp}$ of $\mathcal{M}^{exp}$. At any point P, the push-forward, in this context also called the score operator, of this map,

$$f_c^* : T_P\mathcal{M}^{\Delta} \oplus T_P\mathcal{M}^F \to T_{f_c(P)}\mathcal{M}^{exp}$$

takes the score of P_{ϵ} to the score of $f_c\left(P_{\epsilon}\right)$ at $\epsilon = 0$. For two-phase gene-environment data collection, the score operator can be expressed:

$$f_c^* s^F(o^F) = \delta s^F(o^F) + (1 - \delta)\, \mathbb{E}_G\left[s^F(O^F) \mid Y = y, X = x, W = w\right].$$

Surjectivity of the push-forward follows from smooth invertibility of the coarsening map. This provides a characterization of the tangent space to $\mathcal{M}^{exp}$ at $f_c(P)$ as the closure of the

image of the tangent space to $T_P\mathcal{M}^{\Delta} \oplus T_P\mathcal{M}^{F}$ at P under f_c^*. The push- forward relies solely on the relationship between the full data and observed data models; it should not be expected to take efficient influence functions to efficient influence functions of any function Ψ defined on the model.

Appendix B

CLINICAL DECISION RULES

B.1 sNB Variants

Here we calculate the influence functions for variants of net benefit measures. All influence functions and their corresponding variance are listed in Table 3.1.

sNB^N This is the formulation of sNB for evaluating an opt-in rule. It implicitly contrasts against the universal decision not to treat (N): $\text{sNB}_N^N = 0$. The influence function for sNB_n^N was established in Appendix **??**.

$$IF_{\text{sNB},0}(o) = IF_{\text{TPR},0}(o) - \frac{r_T}{1-r_T}\frac{1-\rho}{\rho}IF_{\text{FPR},0}(o) - \frac{r_T}{1-r_T}\text{FPR}_0\frac{\partial}{\partial\rho}\left(\frac{1-\rho}{\rho}\right)IF_{\rho_0} \quad \text{(B.1)}$$

$$= \frac{d}{\rho_0}\left\{I[r\,(y) > r_T] - \text{TPR}_0\right\} - \frac{r_T}{1-r_T}\frac{1-\rho_0}{\rho_0}\frac{1-d}{1-\rho_0}\left\{I\left[r\,(y) > r\right] - \text{FPR}_0\right\}$$

$$+ \frac{r_H}{1-r_H}\frac{1}{\rho_0^2}\text{FPR}_0\,(d-\rho_0) \quad \text{(B.2)}$$

which is akin to applying multivariate chain rule to obtain a total derivative of Ψ. The influence function for the unstandardized measure of net benefit can be obtained from the relationship $\text{NB} = \rho \cdot \text{sNB}$ between the two measures. By a product rule, the delta method for influence functions yields:

$$\begin{aligned} IF_{\text{NB}_0^N}(o) &= \rho_0 IF_{\text{sNB}_0} + IF_{\rho_0} sNB_0 \\ &= d\left\{I[r\,(y) > r_T] - \text{TPR}_0\right\} - \frac{B^{ctrl}}{B^{case}}(1-d)\left\{I\left[r\,(y) > r_T\right] - \text{FPR}_0\right\} \\ &\quad + \frac{B^{ctrl}}{B^{case}}\frac{1}{\rho_0}\text{FPR}_0\,(d-\rho_0) \end{aligned}$$

and again, one may explicitly check that $\mathbb{E}\left[IF_{\text{NB}}^2(O)\right]$ matches the previously stated formula for the limiting variance of $\sqrt{n}\,(\text{NB}_n - \text{NB}_0)$.

sNB^A This is the formulation of sNB for evaluating an opt-out rule. It implicitly contrasts against the universal treat all (A) decision: $\text{sNB}_A^A = 0$. The defintion of sNB^A for an arbitrary rule R, can be manipulated as follows:

$$\begin{aligned} \text{sNB}^A &:= \text{TNR} - \frac{1}{\omega}\text{FNR} \\ &= 1 - \frac{1}{\omega} + \frac{1}{\omega}\left(\text{TPR} - \omega\text{FPR}\right) \\ &= 1 - \frac{1}{\omega} + \frac{1}{\omega}\left(\text{sNB}^N\right) \end{aligned}$$

which reflects the relationships between positive and negative rates: $TNR = 1 - FPR$ and $FNR = 1 - TPR$. Applying the delta method for influence functions and expressing the result in terms of negative-rates yields:

$$\begin{aligned}
IF_{\mathrm{sNB}^A,0}(o) &= \frac{1}{\omega_0^2}\frac{\partial \omega_0}{\partial \rho} IF_{\rho,0}(o)\left\{1 - \mathrm{sNB}^N\right\} + \frac{1}{\omega_0} IF_{\mathrm{sNB}^N,0}(o) \\
&= \frac{1}{\omega_0}\frac{d}{\rho_0}\left\{I[r\,(y) > r_T] - \mathrm{TPR}_0\right\} - \frac{1-d}{(1-\rho_0)}\left\{I\left[r\,(y) > r_T\right] - \mathrm{FPR}_0\right\} \\
&\quad -\frac{1}{\omega_0}\frac{1}{\rho_0(1-\rho_0)}\left(1 - \mathrm{TPR}\right)(d - \rho_0) \\
&= \frac{1-d}{(1-\rho_0)}\left\{I\left[r\,(y) \leq r_T\right] - \mathrm{TNR}_0\right\} - \frac{1}{\omega_0}\frac{d}{\rho_0}\left\{I[r\,(y) \leq r_T] - \mathrm{FNR}_0\right\} \\
&\quad -\frac{1}{\omega_0}\frac{1}{\rho_0(1-\rho_0)}\mathrm{FNR}\,(d - \rho_0)
\end{aligned}$$

which is akin to applying multivariate chain rule to obtain a total derivative of Ψ and could also have been derived directly from the definition, in terms of influence functions for the true and false-negative rates. The influence function for the corresponding unstandardized measure of net benefit can be obtained from the relationship $\mathrm{NB}^A = (1-\rho)\cdot \mathrm{sNB}^A$ between the two measures. By a product rule, the delta method for influence functions yields:

$$\begin{aligned}
IF_{\mathrm{NB}_0^A}(o) &= (1-\rho_0) IF_{\mathrm{sNB}_0^A} - IF_{\rho_0}\mathrm{sNB}_0^A \\
&= (1-d)\left\{I[r\,(y) \leq r_T] - \mathrm{TNR}_0\right\} - \frac{1}{\omega_0}\frac{1-\rho_0}{\rho_0} d\left\{I\left[r\,(y) \leq r_T\right] - \mathrm{FNR}_0\right\} \\
&\quad + \left\{\frac{1}{\omega_0}\frac{1}{\rho_0}\mathrm{FNR}_0 - \mathrm{sNB}_0^A\right\}(d - \rho_0)
\end{aligned}$$

B.2 Alternate Variance Derivation for sNB^N

The asymptotics of net benefit can be derived by viewing the estimand, sNB_0, as a function of the joint probabilities $p_{11,0} = P_0\left[r\left(Y\right) > r_T \;\&\; D = 1\right]$ and $p_{10,0} = P_0\left[r\left(Y\right) > r_T \;\&\; D = 0\right]$, and the event rate (incidence or prevalence) $\rho_0 = P_0(D = 1)$, evaluated at the parameters of the unknown data generating mechanism. That is:

$$\begin{aligned}
\text{sNB}_0(r, r_T) &= \text{TPR}_0(r, r_T) - \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_0}{\rho_0} \text{FPR}_0(r, r_T) \\
&= \frac{p_{11,0}}{\rho_0} - \frac{B^{ctrl}}{B^{case}} \frac{1-\rho_0}{\rho_0} \cdot \frac{p_{10,0}}{1-\rho_0} \\
&:= f\left(p_{11,0}, p_{10,0}, \rho_0\right)
\end{aligned}$$

An empirical estimator of sNB_0 is then defined from f by plugging in empirical estimates of the joint probabilities and event rate: $\text{sNB}_n(r, r_T) = f\left(p_{11,n}, p_{10,n}, \rho_n\right)$. Jointly,

$$\sqrt{n}\left(\left(p_{11,n}, p_{10,n}, \rho_n\right)^T - \left(p_{11,0}, p_{10,0}, \rho_0\right)^T\right) \to N\left(0, \Sigma_0\right)$$

where the covariance matrix is:

$$\Sigma_0 = \begin{bmatrix}
p_{11,0}(1-p_{11,0}) & -p_{11,0}p_{10,0} & p_{11,0}(1-\rho_0) \\
-p_{11,0}p_{10,0} & p_{10,0}(1-p_{10,0}) & -p_{10,0}\rho_0 \\
p_{11,0}(1-\rho_0) & -p_{10,0}\rho_0 & \rho_0(1-\rho_0)
\end{bmatrix}$$

Evaluating the gradient of the function f gives:

$$\begin{aligned}
\nabla f_0 &= \nabla f(p_{11,0}, p_{10,0}, \rho_0) \\
&= \left(\frac{1}{\rho_0}, -\frac{B^{ctrl}}{B^{case}}\frac{1}{\rho_0}, -\frac{1}{\rho^2}\left\{p_{11,0} - \frac{r_T}{1-r_T}p_{10,0}\right\}\right)^T
\end{aligned}$$

which used in an application of the delta method yields:

$$\sqrt{n}\left(\text{sNB}_n - \text{sNB}_0\right) \;\to\; N\left(0, \nabla f_0^T \Sigma_0 \nabla f_0\right)$$

which simplifies to the expressions presented in the Equation 3.6.

Formally, results of the delta method apply to the particular estimator for which they were calculated. Technically this requires new calculations of correlation and the variance

transformation for every estimator and corresponding sampling context. A more elegant alternative to this brute force approach is provided through influence function methods. These are powerful methods which reveal the dominant behaviour of classes of estimators, extend naturally to other sampling scenarios, and can be used to establish the joint asymptotics of two or more estimators. The latter applies to contrasting estimates of net benefit between two models and to constructing confidence bands. Will we see these benefits throughout

B.3 Decomposition of Variability - Cohort Samples

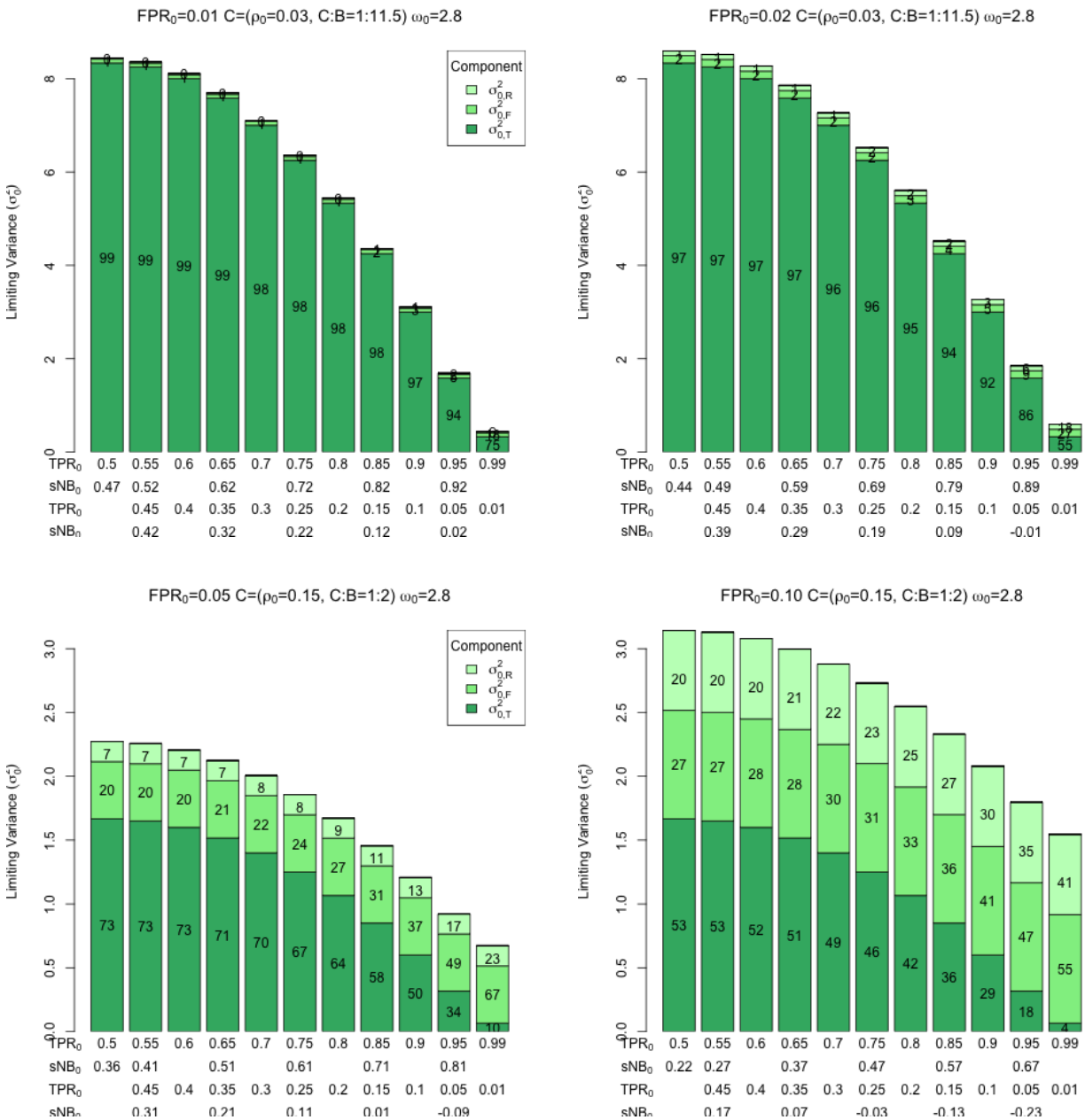


Figure B.1: Companion to Figure 3.1, additional scenarios.

B.4 Variance Level Sets

We work with the formulation of net benefit developed for evaluation of opt-in rules, sNB^N, that implicitly contrasts the expected utility of a proposed rule against the universal treat-none decision. All calculations performed in this section are from the perspective of a fixed clinical context, i.e., ρ and C:B are fixed (and consequently so are r_T and ω). We examine the limiting variance of empirical estimators of sNB^N as a function of the underlying true TPR and FPR values.

Levels sets of variance are hyperbolas

Fixing the limiting variance at some value $\sigma^2 > 0$, and substituting x and y for FPR and TPR into the variance formula (3.6), respectively produces:

$$\begin{aligned}\sigma^2 &= \frac{1}{\rho}y\,(1-y) + \omega^2\frac{1}{1-\rho}x\,(1-x) + \frac{1}{(1-\rho)\rho}\omega^2x^2 \\ &= \frac{\left(x + \frac{\rho}{2(1-\rho)}\right)^2}{\frac{\rho}{\omega^2}} - \frac{(y-0.5)^2}{\rho} + \frac{1}{4\rho} - \frac{\omega^2\rho}{4(1-\rho)^2}\end{aligned}$$

by completing two squares and is of the claimed form:

$$\frac{(x-x_0)^2}{a^2} - \frac{(y-y_0)^2}{b^2} = c^2$$

when

$$\begin{aligned}x_0 &= -\frac{\rho}{2(1-\rho)} \\ y_0 &= 0.5 \\ a &= \sqrt{\frac{\rho}{\omega^2}} \\ b &= \sqrt{\rho}\end{aligned}$$

and

$$c = \sigma^2 - \frac{1}{4\rho} + \frac{\omega^2\rho}{4(1-\rho)^2}.$$

The hyperbolas are centered at $(\text{FPR} = x, \text{TPR} = y)$ and have two asymptotes that go through this point with slopes $\pm\frac{b}{a} = \pm\omega$ (corresponding to $c = 0$).

For any pair of true- and false-positive rates (FPR, TPR) on the asymptote with positive slope (asy+), both the net benefit and the limiting variance of its empirical estimators are constant and equal:

$$\text{sNB}^N_{\text{asy+}} = \frac{1}{2}\left(1 + \frac{\rho}{1-\rho}\omega\right) = \frac{1}{2}\left(1 + \frac{r_T}{1-r_T}\right),$$

and

$$\sigma^2_{\text{asy+}} = \frac{1}{4\rho}\left\{1 - \frac{\omega^2\rho^2}{(1-\rho)^2}\right\} = \frac{1}{4\rho}\left\{1 - \left(\frac{r_T}{1-r_T}\right)^2\right\}$$

respectively.

Level sets of smaller limiting variance occur in horizontal pairs, with one in the upper quadrant and one in the lower quadrant defined by the two asymptotes; the common value decreases as the curves move away from the TPR=50% symmetry line. Level sets of greater limiting variance occur as singletons in the right quadrant defined by both asymptotes; the common value decreases as the curves move away from the $FPR = -\frac{\rho}{2(1-\rho)}$ symmetry line. The matching pair is over negative x-values which do not correspond to false-positive rates.

Variance over FPR-TPR unit square The first and second order partial derivatives of $\sigma^2_{\text{sNB}^N}$ with respect to $y = \text{TPR}$ demonstrates the symmetry about TPR=50% that has already been commented on.

$$\begin{aligned}\frac{\partial\sigma^2}{\partial y} &= \frac{1}{\rho}(1-2y)\\ \\ \frac{\partial^2\sigma^2}{\partial^2 y} &= -2\frac{1}{\rho}\end{aligned}$$

For any fixed value of $x = \text{FPR}$ the limiting variance achieves its maximum at $y = \text{TPR} = 50\%$. The mixed second order partial is 0; all additional understanding of limiting variablility comes from examining its dependence on $x = \text{FPR}$.

$$\begin{aligned}\frac{\partial\sigma^2}{\partial x} &= \frac{\omega^2}{1-\rho} + \frac{2x\omega^2}{\rho}\\ \\ \frac{\partial^2\sigma^2}{\partial^2 x} &= \frac{2\omega^2}{\rho}\end{aligned}$$

The first and second order derivatives are both strictly positive which is consistent with the limiting variance increasing as a function of $x = \text{FPR}$ and increasingly so. Geometrically, the vertical level sets, those with limiting variance greater than that of the asymptote with positive slope, become closer to each other as their crossing with the $y = \text{TPR} = 50\%$ line increases.

Upper bound on limiting variance in a region of minimum net benefit

A region of minimum net benefit corresponds to an upper triangle of the unit square defined by (x, y) points on or above the line $y = \omega x + d$ where d is the minimum sNB^N under consideration. For any given $y = \text{TPR}$ value - the maximum limiting variance occurs at the largest value of $x = \text{FPR}$ in the region; i.e., it lies on the boundary line. We can use the derivatives of the limiting variance, restricted to such as line, to determine the point establishing an upper bound on the limiting variances for all FPR-TPR combinations yielding $\text{sNB}^N \geq d$. The equation for the limiting variance restricted to such a line, σ_d^2, equals:

$$\frac{1}{\rho}\left(\omega x + d\right)\left(1 - \omega x - d\right) + \frac{\omega^2}{1-\rho}\left(x + \frac{1-\rho}{\rho}x^2\right)$$

and its first order derivative, with respect to $x = \text{FPR}$ equals:

$$\begin{aligned}\frac{\partial \sigma_d^2}{\partial x} &= \frac{1}{\rho}\left\{1 - 2\left(\omega x + d\right)\right\} + \frac{\omega^2}{1-\rho} + 2x\frac{\omega^2}{\rho} \\ &= \frac{1}{\rho(1-\rho)}\left\{\omega\left(1-\rho\right)\left(1-2d\right) + \omega^2\rho\right\}\end{aligned}$$

The derivative equals 0 when $\omega = 0$ (not of practical relevance) and when $d = \frac{1}{2} + \frac{1}{2}\frac{\rho}{1-\rho}\omega$, which is the value of the net benefit, constant over the asymptote with positive slope. Further examination leads to the conclusion, that for all regions defined by d less than this value, the derivative is positive and the limiting variance reaches its maximum at the point for which $\left(x = FPR = \frac{1-d}{\omega}, y = TPR = 1\right)$, which is the upper right corner of the region of minimum net benefit d. This value equals,

$$\sigma_{d,max}^2 = \frac{(1-d)^2}{\rho} + \omega\frac{1-d}{1-\rho}$$

For all regions defined by d equal to this value, the maximum limiting variance is achieved by all points on the boundary line. For values of d greater than this value, the maximum limiting variance is achieved by the lower left corner of the boundary. In this case, the derivative decreases then increases, as x increases from 0. A minimum occurs somewhere on the interior of the line, but as the true- and false-positive rates are not design-dependent parameters, this is of no practical interest. The upper bound on the limiting variance could be achieved at either the lower left corner of the the region ($x = FPR = 0, y = TPR = d$) or the upper right corner. The limiting variance at the lower left corner is $\frac{1}{\rho}d(1-d)$. When $d > \frac{1}{2} + \frac{1}{2}\frac{\rho}{1-\rho}\omega$, it follows that $2d(1-\rho) > 1 - \rho + \rho\omega$ and, by putting the variances associated with each corner on a common denominator, this establishes that the maximum limiting variance over the line is achieved at the lower corner.

B.5 Efficiency Gain Due to Known Event Rate

An influence function captures the first order behavior of an asymptotic linear estimator. The efficiency of an estimator is tied to the set of potential data generating mechanisms (the statistical model). Efficiency theory exists for regular asymptotically linear estimators (RAL); all of the empirical estimators introduced thus far are RAL estimators. The influence function for the empirical estimator of net benefit from cohort data (Table 5.1) was calculated with respect to estimation over the nonparametric model, i.e., the set of candidate distributions is unrestricted other than standard measure theoretic requirements. Over nonparametric models, all RAL estimators are asymptotically equivalent whereas over constrained models there can be distinct classes of asymptotically equivalent estimators, some of which are efficient and some of which are not. Of all RAL estimators the ones with the smallest asymptotic variance are called efficient, for estimation with respect to the given model, and their corresponding influence function is called the efficient influence function.

There are scenarios in which using an estimate rather than a known feature of the data generating mechanism produces an efficiency gain. A well-known example arises when employing propensity score weighting methods for estimation of counterfactual means. The characteristics particular to producing this counterintuitive result are that the proposed estimators are inefficient to begin with, and that the knowledge (e.g., the probability of exposure conditional on covariates has logistic behaviour in truth) pertains to a feature of the data generating mechanism that is an orthogonal nuisance of the estimand. In the case of estimating net benefit, the event rate is not a nuisance parameter and utilizing knowledge about it should (and does) provide an efficiency gain (Rose and van der Laan, 2011a, Thm. 2.3).

B.6 Chapter 3 Simulation Study - Complete Results

Here we collect a complete set of results for the simulation described in Section 3.10. For completion, some tables are duplicated. The set of results for the extended model, analogous to the base model, are included here. Additionally, the tables exploring validation of one model in detail have been split into two: one focusing on evaluation of the point estimator and one on the variance estimator. Evaluation of the variance estimator has been extended to consider coverage of Wald confidence intervals constructed using a bootstrapped estimate of the variance. The coverages are generally similar to that of the Wald confidence intervals based on analytic estimates of the asymptotic variance. This comparison supports the conclusion that compromised coverage of the analytically estimated intervals is due to poorness of the Normal approximation rather than poor estimation of the asymptotic variance.

Base Model - PredMod

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
r_1	0.531	-0.40	0.060	0.060	94.7	94.6
r_2	0.526	-0.41	0.059	0.060	95.4	95.2
r_3	0.517	-0.68	0.062	0.062	94.9	94.6
r_4	0.530	-0.55	0.060	0.061	95.2	94.9
r_5	0.532	-0.06	0.061	0.060	95.1	94.4

Table B.1: Estimation of sNB(rT=0.2) for 5 prespecified models developed from 5 independent samples of N=400) over 5000 replicates of external validation sets (N=800).

		sNB				TPR				FPR				
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
							True Value							
		0.671	0.531	0.431	0.353	0.840	0.732	0.643	0.563	0.168	0.090	0.055	0.035	0.100
							% Bias							
400		-0.48	-1.05	-1.53	-1.95	0.14	-0.03	-0.09	-0.17	-0.01	0.05	0.24	0.18	-0.02
800		-0.32	-0.40	-0.56	-0.86	0.04	0.02	-0.03	-0.15	-0.08	-0.38	-0.46	-0.44	-0.18
2000		0.01	0.01	0.01	0.05	0.02	0.02	0.03	0.06	-0.17	-0.17	-0.13	-0.13	0.26

Table B.2: Simulation of validating a prespecified model (PredMod) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

		sNB				TPR				FPR				
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
	Estimand:	0.671	0.531	0.431	0.353	0.840	0.732	0.643	0.563	0.168	0.090	0.055	0.035	0.100
								Standard Deviation						
400	observed	0.069	0.086	0.098	0.106	0.058	0.071	0.078	0.079	0.020	0.015	0.012	0.010	0.015
	bootstrap	0.071	0.089	0.101	0.110	0.058	0.071	0.077	0.079	0.020	0.015	0.012	0.010	0.015
	analytic	0.068	0.086	0.098	0.106	0.058	0.070	0.076	0.078	0.020	0.015	0.012	0.010	0.015
800	observed	0.048	0.060	0.068	0.074	0.041	0.050	0.054	0.056	0.014	0.011	0.008	0.007	0.011
	bootstrap	0.049	0.061	0.069	0.075	0.041	0.050	0.054	0.056	0.014	0.011	0.008	0.007	0.011
	analytic	0.048	0.060	0.068	0.074	0.041	0.049	0.054	0.055	0.014	0.011	0.008	0.007	0.011
2000	observed	0.030	0.038	0.043	0.046	0.026	0.032	0.034	0.035	0.009	0.007	0.005	0.004	0.007
	bootstrap	0.030	0.038	0.043	0.047	0.026	0.031	0.034	0.035	0.009	0.007	0.005	0.004	0.007
	analytic	0.030	0.038	0.043	0.046	0.026	0.031	0.034	0.035	0.009	0.007	0.005	0.004	0.007
								95% Coverage						
400	bootstrapP	94.6	95.3	95.1	94.8	93.2	94.5	94.3	94.8	94.8	94.7	94.7	94.5	94.3
	bootstrapW	94.0	95.5	95.3	95.4	92.0	93.5	93.7	94.2	94.8	94.2	94.5	93.4	94.7
	analytic	93.6	94.7	94.7	94.7	91.8	93.0	93.6	93.9	94.8	94.2	94.4	93.4	94.7
800	bootstrapP	95.1	94.7	94.6	94.8	94.5	94.4	94.9	94.7	95.7	94.6	94.6	94.8	94.5
	bootstrapW	94.6	95.1	95.1	95.0	93.8	93.8	94.5	94.3	95.6	94.6	94.3	94.0	94.7
	analytic	94.4	94.6	94.8	94.6	93.6	93.8	94.4	94.2	95.6	94.4	94.4	94.0	95.0
2000	bootstrapP	95.0	94.8	95.2	95.2	94.6	94.4	94.5	94.8	94.5	94.8	94.9	94.1	94.5
	bootstrapW	94.8	94.7	95.2	95.1	94.4	94.3	94.4	94.7	94.5	94.9	94.6	93.9	94.5
	analytic	94.8	94.7	95.1	95.1	94.3	94.1	94.3	94.5	94.5	94.9	94.7	93.9	94.7

Table B.3: Simulation of validating a prespecified model (PredMod) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

Extended Model - $PredMod_{ext}$

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
r_1	0.761	-0.24	0.044	0.044	94.3	93.8
r_2	0.759	-0.21	0.045	0.045	95.3	94.6
r_3	0.743	-0.36	0.045	0.046	95.0	94.2
r_4	0.758	-0.14	0.044	0.044	95.2	95.1
r_5	0.745	0.05	0.046	0.045	94.3	93.5

Table B.4: Estimation of sNB(rT=0.2) for 5 prespecified models developed from 5 independent samples of N=400) over 5000 replicates of external validation sets (N=800).

		sNB				TPR				FPR				
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
							True Value							
		0.836	0.761	0.699	0.642	0.918	0.876	0.838	0.804	0.083	0.051	0.036	0.027	0.100
							% Bias							
400		-0.48	-0.66	-0.86	-1.04	-0.17	-0.15	-0.10	-0.11	0.37	0.64	1.05	0.95	-0.02
800		-0.21	-0.24	-0.21	-0.27	-0.05	-0.01	0.07	0.06	0.06	0.01	-0.01	-0.15	-0.18
2000		-0.00	0.01	0.00	0.00	0.01	0.05	0.03	0.07	-0.03	0.08	-0.02	0.11	0.26

Table B.5: Simulation of validating a prespecified model (PredMod$_{ext}$) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

		sNB				TPR				FPR				
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
	Estimand:	0.836	0.761	0.699	0.642	0.918	0.876	0.838	0.804	0.083	0.051	0.036	0.027	0.100
								Standard Deviation						
400	observed	0.049	0.063	0.074	0.087	0.044	0.053	0.058	0.063	0.014	0.011	0.010	0.009	0.015
	bootstrap	0.050	0.064	0.077	0.091	0.044	0.053	0.059	0.064	0.015	0.012	0.010	0.009	0.015
	analytic	0.049	0.063	0.075	0.088	0.044	0.052	0.058	0.063	0.015	0.012	0.010	0.009	0.015
800	observed	0.034	0.044	0.053	0.062	0.031	0.037	0.042	0.045	0.010	0.008	0.007	0.006	0.011
	bootstrap	0.034	0.044	0.053	0.062	0.031	0.037	0.041	0.045	0.010	0.008	0.007	0.006	0.011
	analytic	0.034	0.044	0.052	0.061	0.031	0.037	0.041	0.044	0.010	0.008	0.007	0.006	0.011
2000	observed	0.021	0.027	0.032	0.037	0.019	0.023	0.026	0.028	0.007	0.005	0.004	0.004	0.007
	bootstrap	0.021	0.028	0.033	0.038	0.019	0.023	0.026	0.028	0.006	0.005	0.004	0.004	0.007
	analytic	0.021	0.027	0.033	0.038	0.019	0.023	0.026	0.028	0.006	0.005	0.004	0.004	0.007
								95% Coverage						
400	bootstrapP	93.5	94.2	94.7	94.4	92.2	93.0	93.8	94.1	95.2	94.5	93.9	94.7	94.3
	bootstrapW	92.1	93.9	94.4	94.8	88.9	91.3	92.4	93.0	95.0	94.3	94.1	91.8	94.7
	analytic	91.6	93.4	93.8	94.1	88.2	90.8	92.0	92.8	95.1	94.3	94.1	91.8	94.7
800	bootstrapP	94.4	94.3	94.5	94.2	93.5	94.0	94.1	94.2	94.8	94.3	94.1	93.4	94.5
	bootstrapW	93.6	94.2	94.5	94.4	92.0	93.1	93.5	93.4	94.9	94.0	93.6	93.2	94.7
	analytic	93.5	93.8	94.2	93.9	91.9	92.8	93.4	93.2	94.9	94.0	93.8	93.1	95.0
2000	bootstrapP	94.8	95.0	95.1	95.2	94.5	94.6	94.5	94.9	95.1	94.7	95.1	94.8	94.5
	bootstrapW	94.5	95.0	94.9	95.0	93.8	94.3	94.3	94.9	95.1	94.8	95.1	94.7	94.5
	analytic	94.4	94.9	94.8	95.0	93.7	94.2	94.3	94.9	95.1	94.9	95.1	94.8	94.7

Table B.6: Simulation of validating a prespecified model (PredMod$_{ext}$) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

Difference in Net Benefit between PredMod and PredMod$_{ext}$

			Variance		Coverage	
	ΔsNB$_0$	% Bias	MC	Ana	BS	Ana
r_1	0.230	0.13	0.056	0.057	95.3	95.2
r_2	0.233	0.25	0.055	0.056	94.8	94.8
r_3	0.226	0.38	0.051	0.052	95.0	94.9
r_4	0.228	0.80	0.052	0.052	94.9	94.5
r_5	0.213	0.34	0.050	0.050	94.4	94.5

Table B.7: Estimation of ΔsNB(rT=0.2) for 5 prespecified models developed from 5 independent samples of N=400) over 5000 replicates of external validation sets (N=800).

		ΔsNB				ΔTPR				ΔFPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
						True Value							
		0.164	0.230	0.268	0.289	0.079	0.144	0.196	0.241	-0.086	-0.038	-0.019	-0.008
						% Bias							
400		-0.49	0.23	0.22	0.07	-3.45	-0.74	-0.15	0.02	-0.38	-0.75	-1.32	-2.40
800		0.25	0.13	0.34	0.45	-0.96	-0.19	0.39	0.57	-0.22	-0.91	-1.32	-1.41
2000		-0.06	-0.01	-0.01	-0.06	-0.02	0.16	0.04	0.07	-0.31	-0.50	-0.35	-0.96

Table B.8: Simulation of validating a prespecified model (PredMod - PredMod$_{ext}$) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

		ΔsNB				ΔTPR				ΔFPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
	Estimand:	0.164	0.230	0.268	0.289	0.079	0.144	0.196	0.241	-0.086	-0.038	-0.019	-0.008
							Standard Deviation						
400	observed	0.065	0.081	0.093	0.105	0.059	0.072	0.077	0.081	0.020	0.016	0.013	0.011
	bootstrap	0.065	0.082	0.095	0.108	0.059	0.071	0.076	0.080	0.020	0.016	0.013	0.011
	analytic	0.064	0.080	0.093	0.106	0.059	0.070	0.076	0.079	0.020	0.016	0.013	0.011
800	observed	0.045	0.056	0.065	0.073	0.042	0.049	0.054	0.056	0.014	0.011	0.009	0.008
	bootstrap	0.046	0.057	0.066	0.075	0.042	0.050	0.054	0.056	0.014	0.011	0.009	0.008
	analytic	0.045	0.057	0.065	0.074	0.042	0.049	0.053	0.056	0.014	0.011	0.009	0.008
2000	observed	0.029	0.036	0.041	0.046	0.026	0.032	0.034	0.035	0.009	0.007	0.006	0.005
	bootstrap	0.028	0.036	0.041	0.046	0.026	0.031	0.034	0.036	0.009	0.007	0.006	0.005
	analytic	0.028	0.036	0.041	0.046	0.026	0.031	0.034	0.035	0.009	0.007	0.006	0.005
							95% Coverage						
400	bootstrapP	94.0	94.4	94.5	94.9	92.8	93.1	94.3	94.3	94.6	94.4	94.1	94.6
	bootstrapW	94.1	94.6	95.1	95.6	92.7	93.2	94.1	94.1	94.6	94.8	94.5	95.0
	analytic	93.5	94.1	94.5	95.0	92.4	92.7	93.8	93.8	94.6	94.7	94.5	95.0
800	bootstrapP	94.9	95.3	95.1	94.9	94.6	94.6	94.3	94.8	95.4	95.2	95.2	94.5
	bootstrapW	95.3	95.4	95.1	95.2	94.8	94.8	94.2	94.7	95.5	95.4	95.2	95.1
	analytic	95.1	95.2	95.0	95.1	94.6	94.7	94.1	94.6	95.5	95.3	95.2	95.1
2000	bootstrapP	94.8	94.4	94.3	94.8	94.7	94.0	94.5	94.3	94.5	94.9	95.4	95.2
	bootstrapW	94.8	94.5	94.5	95.1	94.9	94.0	94.5	94.3	94.6	95.0	95.4	95.5
	analytic	94.8	94.3	94.4	95.0	94.9	93.9	94.3	94.3	94.5	95.1	95.5	95.4

Table B.9: Simulation of validating a prespecified model (PredMod - PredMod$_{ext}$) that was developed on a sample of size 400. Results for each validation sample size are based on: 5000 replications.

B.7 Confidence Bands

We start by considering standardized net benefit as a function of allowed thresholds. Pointwise, any fixed $t \in [0,1)$, asymptotic normality has been established. Dependence on the pre-specified risk model is suppressed in the notation that follows.

The covariance $\mathrm{Cov}\,(t_1,t_2) := \mathrm{Cov}\,(sNB(t_1), sNB(t_2))$ can be calculated using the influence function as follows:

$$\begin{aligned}
\mathrm{Cov}\,(t_1,t_2) &= \mathbb{E}\left[IF_{t_1,0}(O)IF_{t_2,0}(O)\right] && \text{(B.3)}\\
&= \mathbb{E}\left[\frac{D}{\rho_0^2}\left\{X_1-\mathrm{TPR}_1\right\}\left\{X_2-\mathrm{TPR}_2\right\}\right] && \text{(B.4)}\\
&\quad+\frac{t_1}{1-t_1}\frac{t_2}{1-t_2}\left(\frac{1-\rho_0}{\rho_0}\right)^2\mathbb{E}\left[\frac{1-D}{(1-\rho_0)^2}\left\{X_1-\mathrm{FPR}_1\right\}\left\{X_2-\mathrm{FPR}_2\right\}\right]\\
&\quad+\frac{t_1}{1-t_1}\frac{t_2}{1-t_2}\frac{1}{\rho_0^4}\mathrm{FPR}_1\mathrm{FPR}_2\mathbb{E}\left[(D-\rho_0)^2\right]\\
&= \frac{1}{\rho_0}(\mathrm{TPR}_{1\wedge2}-\mathrm{TPR}_1\mathrm{TPR}_2)\\
&\quad+\frac{t_1}{1-t_1}\frac{t_2}{1-t_2}\frac{1-\rho_0}{\rho_0^2}\left(\mathrm{FPR}_{1\wedge2}-\mathrm{FPR}_1\mathrm{FPR}_2\right)\\
&\quad+\frac{t_1}{1-t_1}\frac{t_2}{1-t_2}\frac{1-\rho_0}{\rho_0^3}\mathrm{FPR}_1\mathrm{FPR}_2
\end{aligned}$$

where $X_i := I\left[r\left(Y\right) > t_i\right]$, $\mathrm{FPR}_i := \mathrm{FPR}_0(t_i)$ and $\mathrm{FPR}_{1\wedge2} := \mathrm{FPR}_1 \wedge \mathrm{FPR}_2 = \mathrm{FPR}_0(t_1 \vee t_2)$ etc. have been used for brevity. Pairwise orthogonality of the components of each influence function ensures that the analogous pairwise correlations are also 0.

In Section 4.4, a numerical algorithm for constructing approximate confidence bands was described. The algorithm relies on approximating the limiting process whose existence was heuristically argued. The weak convergence of the process $\sqrt{n}\left(\mathrm{sNB}_n - \mathrm{sNB}_0\right)$ can be established rigorously by using asymptotic linearity of the estimator sNB_n which is established by the expansion:

$$\sqrt{n}\left\{\mathrm{sNB}_n(t)-\mathrm{sNB}_0(t)\right\} = \frac{1}{\sqrt{n}}\sum_i IF_{sNB,0}(O_i;t)+\sqrt{n}R(P_n,P_0,t)$$

where the influence function for $\mathrm{sNB}_n(t)$ is:

$$\begin{aligned} IF_{t,0}(o) &= \tfrac{d}{\rho_0}\left\{I[r(y)>t]-\mathrm{TPR}_0\right\}-\tfrac{t}{1-t}\tfrac{1-\rho_0}{\rho_0}\tfrac{1-d}{1-\rho_0}\left\{I\left[r(y)>t\right]-\mathrm{FPR}_0\right\} \\ &\quad +\tfrac{t}{1-t}\tfrac{1}{\rho_0^2}\mathrm{FPR}_0\left(d-\rho_0\right), \end{aligned}$$

which was originally stated in Equation 3.5 and has been re-expressed to highlight the dependence on the risk threshold t. The remainder term,

$$\begin{aligned} R\left(P_n,P_0,t\right) &= \left\{\mathrm{TPR}_n(t)-\mathrm{TPR}_0(t)\right\}\left(1-\frac{\rho_0}{\rho_n}\right) \\ &\quad -\frac{t}{1-t}\frac{\rho_n-1}{\rho_n^2}\left(\rho_0-\rho_n\right)\left\{\mathrm{FPR}_n(t)-\mathrm{FPR}_0(t)\right\} \\ &\quad +\frac{t}{1-t}\left(\rho_0-\rho_n\right)^2\frac{1}{\rho_n^2\rho_0}\mathrm{FPR}_0(t) \end{aligned}$$

is second order at every $t\in[0,1)$; each term is the product of two $O_p(n^{-1/2})$ factors.

To establish weak convergence of $\sqrt{n}\left(\mathrm{sNB}_n-\mathrm{sNB}_0\right)$ to a Gaussian process with the covariance structure given above, it suffices to establish the following two conditions:

$$\sup_{t\in\mathcal{T}}\sqrt{n}\left|R\left(P_n,P_0,t\right)\right|\to_p 0$$

and

$$\left\{IF_{t,0}:t\in\mathcal{T}\right\}\subset\mathcal{F}\text{ a }P_0\text{-Donsker class}$$

We first establish that the condition on the remainder is satisfied for the standardized net benefit process. Consistent with recommended practice for decision and relative utility curves, we only consider risk thresholds less than $1-\epsilon$ for some value of $\epsilon\in(0,1)$. Consequently, the odds of t is bounded by $(1-\epsilon)/\epsilon$. For any $t\in[0,1-\epsilon]$, the remainder satisfies:

$$\begin{aligned} \sqrt{n}\left|R\left(P_n,P_0,t\right)\right| &\le \left|\sqrt{n}\left(1-\frac{\rho_0}{\rho_n}\right)\right|\left|\mathrm{TPR}_n(t)-\mathrm{TPR}_0(t)\right| \\ &\quad +\frac{1-\epsilon}{\epsilon}\left|\sqrt{n}\left(\rho_0-\rho_n\right)\frac{\rho_n-1}{\rho_n^2}\right|\left|\left\{\mathrm{FPR}_n(t)-\mathrm{FPR}_0(t)\right\}\right| \\ &\quad +\frac{1-\epsilon}{\epsilon}\left|\sqrt{n}\left(\rho_0-\rho_n\right)^2\frac{1}{\rho_n^2\rho_0}\right|, \end{aligned}$$

where the last term reflects the fact that FPR_0 is always between 0 and 1. The third term is $o_p(1)$ and independent of t. The fist and second terms are products of a factor that is $O_p(1)$ and a factor that is governed by Glivenko-Cantelli's Theorem. Hence, the supremum over all $t \in \mathcal{T}$ goes to 0 in probability.

A sufficient, though not necessary condition for the second requirement is that the set of influence functions be bounded in uniform sectional variation norm (Gill et al., 1995). That is:

$$\sup_{t \in \mathcal{T}} \{\|IF_t\|_V^*\} < \infty,$$

where $\|IF_t\|_V^* = \max \left\{ \|IF_t(o)\|_\infty, \sup_{o_2} \int |IF_t(do_1, o_2)|, \sup_{o_1} \int |IF_t(o_1, do_2)|, \int |IF_t(do_1, do_2)| \right\}$, where $o_1 = x$ and $o_2 = d$ have been used for clarity. By virtue of both variables being binary, this condition is straightforward to confirm. In particular, for any $t \in [0, 1-\epsilon]$,

$$\begin{aligned}
\|IF_t(o)\|_\infty &\leq \frac{1}{\rho_0} + \frac{1-\epsilon}{\epsilon}\frac{1}{\rho_0} + \frac{1-\epsilon}{\epsilon}\frac{1}{\rho_0} \\
\sup_{o_2} \int |IF_t(do_1, o_2)| &\leq \max\left\{ \frac{1}{\rho_0}, \frac{\omega_0}{1-\rho_0} \right\} \rho_X \\
\sup_{o_1} \int |IF_t(o_1, do_2)| &\leq \max\left\{ \left| \frac{1-\text{TPR}_0}{\rho_0} + \frac{\omega_0}{1-\rho_0}(1-\text{FPR}_0) + \frac{\omega_0}{(1-\rho_0)\rho_0}\text{FPR}_0 \right|, \right. \\
&\qquad \left. \left| -\frac{\text{TPR}_0}{\rho_0} - \frac{\omega_0}{1-\rho_0}\text{FPR}_0 + \frac{\omega_0}{(1-\rho_0)\rho_0}\text{FPR}_0 \right| \right\} \rho_0 \\
\int |IF_t(do_1, do_2)| &= \left(1 + \frac{B^{ctrl}}{B^{case}} \right) \rho_X,
\end{aligned}$$

where $\rho_X := Pr(X = 1)$ is the probability of high-risk classification in the population. Since all four of these quantities are are bounded by finite values that do not rely on the threshold t, the indexed set of influence functions is bounded in uniform sectional variation norm. In conclusion:

$$\sqrt{n}\left(\text{sNB}_n - \text{sNB}_0\right) \rightsquigarrow \mathbb{G}\left(\Sigma\right),$$

the standardized net benefit process converges weakly to a Gaussian process with covariance structure Σ defined by Equation B.3.

B.8 Optimal Unmatched Control-Case Ratio

Derivation For a fixed sample size, $n = n_0 + n_1$, there is choice of control-to-case ratio $J = n_0/n_1$ which minimizes the variance of the empirical case-control estimator. The finite sample variance is approximately:

$$\frac{\sigma_0^2}{n} = c_1 \frac{J+1}{n} TPR_0 \left(1 - TPR_0\right) + c_1 \frac{J+1}{Jn} \omega_0^2 FPR_0 \left(1 - FPR_0\right) + c_2,$$

where the constants c_1 and c_2 vary according to which case-control sampling design has been adopted. Both constants are independent of the control-to-case ratio and without loss of generality, the ratio derived for the stand-alone case-control sample ($c_1 = 1$ and $c_2 = 0$) is also optimal for all unmatched case-control sampling designs examined thus far.

One can aim to minimize the variance by designing the study with ratio that is the solution of:

$$\frac{\partial}{\partial J} \frac{\sigma_0^{2,cc}}{n} = \frac{1}{n} TPR_0 \left(1 - TPR_0\right) - \frac{1}{J^2 n} \omega_0^2 FPR_0 \left(1 - FPR_0\right) = 0$$

Convexity of the original function is clear with a second derivative. Solving for J yields the optimal ratio:

$$J_0^{opt} = \sqrt{\omega_0^2 \frac{FPR_0(1 - FPR_0)}{TPR_0(1 - TPR_0)}} = \omega_0 \sqrt{\frac{FPR_0(1 - FPR_0)}{TPR_0(1 - TPR_0)}}. \tag{B.5}$$

Level Sets of J^{opt} Fix a clinical context $(\rho, \frac{B^{ctrl}}{B^{case}})$ and note that this defines the relevant weight $\omega = \frac{B^{ctrl}}{B^{case}} \frac{1-\rho}{\rho}$. The optimal control-to-case ratio equals:

$$J^{opt} = \omega \sqrt{\frac{FPR(1 - FPR)}{TPR(1 - TPR)}}$$

where TPR and FPR are the unknown true- and false-positive rates associated with a particular risk rule $R = (r, r_T)$. For rational risk rules, $r_T = \frac{B^{ctrl}}{B^{ctrl} + B^{case}}$. We note that the level sets, over the $x = FPR$-$y = TPR$ unit square, of optimal ratios, are hyperbolas with asymptotes equal to the $TPR = FPR$ and $TPR = 1 - FPR$ lines intersecting at the point $(FPR = 0.5, TPR = 0.5)$.

By considering $J^{opt} = c$ to be any particular ratio $c > 0$, and some algebraic manipulation leads to

$$\begin{aligned} \frac{c^2}{\omega^2} y(1-y) &= x(1-x) \\ \frac{c^2}{\omega^2}(y^2 - y + 0.25 - 0.25) &= (x^2 - x + 0.25 - 0.25) \\ \frac{c^2}{\omega^2}(y-0.5)^2 - (x^2-0.5)^2 &= \frac{1}{4}\left(\frac{c^2}{\omega^2} - 1\right) \end{aligned}$$

which has three cases

$$\begin{cases} \frac{(y-0.5)^2}{\frac{1}{4}\left(1-\frac{\omega^2}{c^2}\right)} - \frac{(x-0.5)^2}{\frac{1}{4}\left(\frac{c^2}{\omega^2}-1\right)} = 1 & c > \omega \\ \frac{(x-0.5)^2}{\frac{1}{4}\left(1-\frac{c^2}{\omega^2}\right)} - \frac{(y-0.5)^2}{\frac{1}{4}\left(\frac{\omega^2}{c^2}-1\right)} = 1 & c < \omega \\ y = x \text{ or } y = 1-x & c = \omega \end{cases}$$

according to the relationship between the ratio and the weight. Values of FPR and TPR that yield optimal ratio $J^{opt} = c > \omega$ lie on hyperbolas above and below the $TPR = 0.5$ axis of symmetry (with increasing ratio as TPR becomes more extreme); that yield optimal ratio $J^{opt} = c = \omega$ lie on the asymptotes; that yield optimal ratio $J^{opt} = c < \omega$ lie on hyperbolas to the left and rtight of the $FPR = 0.5$ axis of symmetry (with decreasing ratio as FPR becomes more extreme).

Relative Efficiency Figure B.2 presents relative efficiency plots expanding on the set presented in Figure 5.2 of Section 5.3.

We established that the limiting variance of the case-control empirical estimator of sNB^N is

$$\sigma^{2,cc}_{0,J} = \frac{J+1}{1}\text{TPR}_0\left(1-\text{TPR}_0\right) + \frac{J+1}{J}\omega_0^2\text{FPR}_0\left(1-\text{FPR}_0\right)$$

where J is the control-to-case ratio. This ratio corresponds to the sample containing $\frac{J^{opt}}{1+J^{opt}}$ controls (as a percentage).

Under the optimal design, and again using x and y in place of FPR and TPR, this

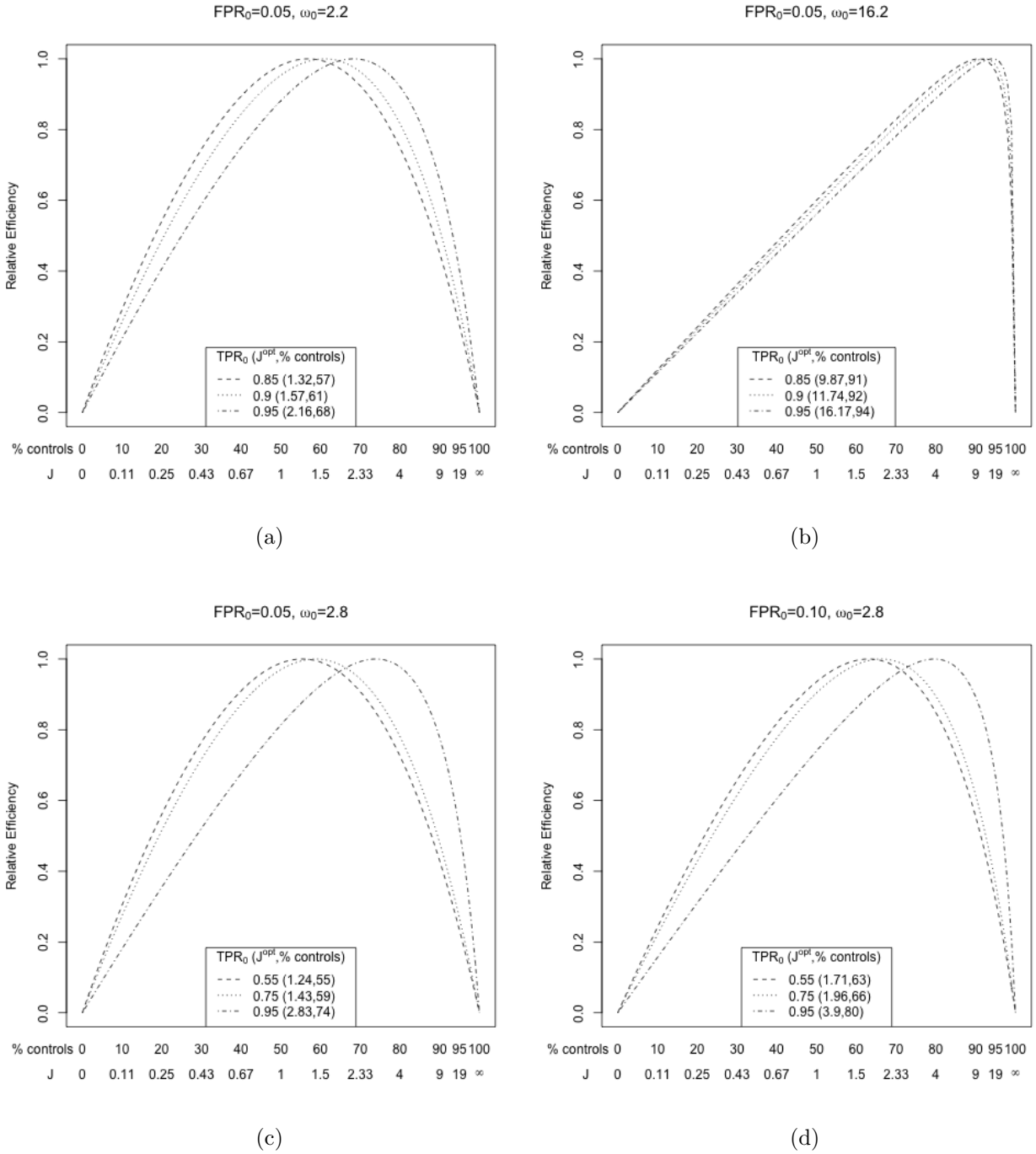


Figure B.2: Efficiency of stand-alone case-control estimators across different control-to-case ratios (J), and corresponding number of controls (as a percentage), relative to the optimal ratio. Plots (a) and (b) parallel those in the document, for a larger false-positive rate (5% vs 2%) and consequently larger true-positive rates. Plots (c) and (d) highlight the commonality with plot (a) related to similar values of ω_0 and show the subtle impacts of the decision rule performance characteristics on the shape.

expression simplifies to:

$$\begin{aligned}
\sigma^{2,cc}_{J^{opt}} &= \frac{J^{opt}+1}{1} y\,(1-y) + \frac{J^{opt}+1}{J^{opt}} \omega^2 x\,(1-x) \\
&= J^{opt} y\,(1-y) + y\,(1-y) + \omega^2 x\,(1-x) + \frac{1}{J^{opt}} \omega^2 x\,(1-x) \\
&= \omega \sqrt{\frac{x(1-x)}{y(1-y)}}\, y\,(1-y) + y\,(1-y) + \omega^2 x\,(1-x) + \frac{1}{\omega} \sqrt{\frac{y(1-y)}{x(1-x)}}\, \omega^2 x\,(1-x) \\
&= y\,(1-y) + \omega^2 x\,(1-x) + 2\,\omega \sqrt{x(1-x)y(1-y)} \\
&= \left(\sqrt{y\,(1-y)} + \omega \sqrt{x\,(1-x)} \right)^2 .
\end{aligned}$$

Similarly, the limiting variance for any design (i.e., any ratio J) can be manipulated as follows:

$$\begin{aligned}
\sigma^{2,cc}_{J} &= \frac{J+1}{1} y\,(1-y) + \frac{J+1}{J} \omega^2 x\,(1-x) \\
&= J y\,(1-y) + y\,(1-y) + \omega^2 x\,(1-x) + \frac{1}{J} \omega^2 x\,(1-x) \\
&= \left(\sqrt{y\,(1-y)} + \omega \sqrt{x\,(1-x)} \right)^2 - 2\,\omega \sqrt{x(1-x)y(1-y)} + J y\,(1-y) + \frac{1}{J} \omega^2 x\,(1-x) \\
&= \sigma^{2,cc}_{J^{opt}} + J y\,(1-y) \left(1 - \frac{J^{opt}}{J} \right)^2
\end{aligned}$$

and hence the relative efficiency can be expressed:

$$\mathrm{RE}\left(J, J^{opt}\right) = \frac{\sigma^{2,cc}_{J^{opt}}}{\sigma^{2,cc}_{J}} = \frac{\sigma^{2,cc}_{J^{opt}}}{\sigma^{2,cc}_{J^{opt}} + J y\,(1-y) \left(1 - \frac{J^{opt}}{J}\right)^2}$$

The above expression pertains to estimation from a stand-alone case-control sample (where the outcome rate ρ_0 is assumed known). We have seen that extensions to case-control samples augmented with an external estimate of ρ_0, nested within a cohort or of a more general two-phase design have empirical estimators with limiting variance of the form:

$$\sigma^2_J = c_1 \sigma^{2,cc}_J + c_2$$

where c_1, c_2 are constants that account for the relative sample sizes of the case-control portion and the cohort portion (possibly overlapping) as well as the contribution due to estimating

the outcome rate. Only the first component is impacted by the design parameter J and quite often the cohort contribution (whether external or containing the case-control sample as a subset) exists outside of the particular study. Hence, the above relative efficiency equation is more generally of the form:

$$\mathrm{RE}\left(J, J^{opt}\right) = \frac{\sigma^2_{J^{opt}}}{\sigma^2_J} = \frac{c_1\sigma^{2,cc}_{J^{opt}} + c_2}{c_1\sigma^{2,cc}_{J^{opt}} + c_2 + c_1 Jy\left(1-y\right)\left(1-\frac{J^{opt}}{J}\right)^2}$$

B.9 Chapter 5 Simulation Study - Complete Results

Here we collect a complete set of results for the simulation described in Section 5.6 for estimation from a stand-alone case-control sample, using known outcome probability, and an unmatched case-control study nested within a cohort. For completion, some tables are duplicated. The set of results for the extended model, analogous to the base model, are included here. Additionally, the tables exploring validation of one model in detail have been split into two: one focusing on evaluation of the point estimator and one on the variance estimator. Evaluation of the variance estimator has been extended to consider coverage of Wald confidence intervals constructed using a bootstrapped estimate of the variance. The coverages are generally similar to that of the Wald confidence intervals based on analytic estimates of the asymptotic variance. This comparison supports the conclusion that compromised coverage of the analytically estimated intervals is due to poorness of the Normal approximation rather than poor estimation of the asymptotic variance.

Unmatched Case-Control - Known Prevalence

Base Model - PredMod

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
Model 1	0.532	0.04	0.055	0.054	94.8	94.4
Model 2	0.533	-0.02	0.053	0.053	94.5	94.6
Model 3	0.528	-0.05	0.056	0.055	94.4	94.3
Model 4	0.531	-0.08	0.055	0.055	94.7	94.5
Model 5	0.530	0.07	0.054	0.055	94.8	94.9

Table B.10: Estimation of sNB(rT=0.2) for 5 prespecified models (developed from 5 independent samples of N=600) over 5000 replicates of external validation sets (N=400).

		sNB				TPR				FPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
						True Value							
		0.673	0.533	0.435	0.357	0.838	0.712	0.608	0.515	0.166	0.080	0.045	0.026
						% Bias							
200		0.04	0.25	0.29	-0.25	0.06	0.17	0.13	0.04	0.15	-0.08	-0.26	0.68
400		-0.12	-0.02	-0.14	-0.05	-0.06	-0.09	-0.17	-0.12	0.16	-0.29	-0.23	-0.26
800		0.08	0.09	0.31	0.31	0.01	0.00	-0.01	-0.06	-0.28	-0.25	-0.82	-0.90
1600		0.04	0.08	0.02	-0.16	0.02	0.03	0.03	0.02	-0.08	-0.11	0.06	0.44

Table B.11: Simulation of validating(PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

		sNB				TPR				FPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
	Estimand:	0.673	0.533	0.435	0.357	0.838	0.712	0.608	0.515	0.166	0.080	0.045	0.026
						Standard Deviation							
200	observed	0.051	0.074	0.092	0.108	0.036	0.045	0.048	0.049	0.037	0.027	0.020	0.016
	bootstrap	0.052	0.075	0.093	0.108	0.037	0.045	0.049	0.050	0.037	0.027	0.021	0.016
	analytic	0.052	0.075	0.093	0.108	0.037	0.045	0.049	0.050	0.037	0.027	0.021	0.016
400	observed	0.037	0.053	0.066	0.077	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011
	bootstrap	0.037	0.053	0.066	0.076	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011
	analytic	0.037	0.053	0.066	0.076	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011
800	observed	0.026	0.038	0.046	0.054	0.018	0.023	0.024	0.025	0.019	0.013	0.010	0.008
	bootstrap	0.026	0.038	0.047	0.054	0.018	0.023	0.024	0.025	0.019	0.014	0.010	0.008
	analytic	0.026	0.038	0.047	0.054	0.018	0.023	0.024	0.025	0.019	0.014	0.010	0.008
						95% Coverage							
200	bootstrapP	94.9	94.5	93.5	91.5	93.6	94.4	94.5	95.3	93.4	94.0	93.0	92.3
	bootstrapW	95.5	94.2	93.1	91.0	93.6	95.1	94.3	95.3	94.2	90.4	93.6	93.0
	analytic	95.5	94.3	93.1	90.8	93.6	95.1	94.6	95.7	94.4	90.0	93.6	92.8
400	bootstrapP	94.4	94.5	94.6	93.8	94.5	95.0	95.1	95.0	94.3	94.3	93.1	89.5
	bootstrapW	94.4	94.5	94.2	92.9	94.5	95.0	95.2	94.6	94.0	95.0	93.9	89.5
	analytic	94.5	94.6	94.2	92.7	94.4	95.0	95.2	94.6	93.8	95.3	93.9	89.5
800	bootstrapP	94.9	94.8	94.5	94.0	94.1	94.4	95.1	95.1	95.1	94.5	95.3	93.9
	bootstrapW	94.7	94.7	94.3	93.7	94.5	94.3	94.9	94.9	95.1	94.9	93.3	94.1
	analytic	94.7	94.6	94.4	93.8	94.5	94.2	94.9	94.9	95.6	95.0	93.2	94.0

Table B.12: Simulation of validating(PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

Extended Model - $PredMod_{ext}$

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
Model 1	0.758	0.12	0.041	0.041	93.7	93.6
Model 2	0.762	-0.11	0.041	0.041	94.5	94.3
Model 3	0.756	0.01	0.043	0.042	94.0	93.6
Model 4	0.763	-0.22	0.041	0.041	94.7	94.6
Model 5	0.748	0.22	0.042	0.042	93.9	93.6

Table B.13: Estimation of sNB(rT=0.2) for 5 prespecified models (developed from 5 independent samples of N=600) over 5000 replicates of external validation sets (N=400).

		sNB				TPR				FPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
						True Value							
		0.837	0.762	0.702	0.646	0.913	0.871	0.837	0.804	0.077	0.048	0.035	0.026
						% Bias							
200		0.04	-0.03	-0.12	-0.32	0.08	0.05	0.06	0.05	0.56	0.60	1.01	1.58
400		-0.08	-0.11	-0.09	-0.01	-0.07	-0.09	-0.09	-0.08	0.08	0.06	-0.12	-0.36
800		0.00	0.02	-0.00	0.02	-0.00	0.00	-0.01	-0.00	-0.06	-0.13	-0.04	-0.09
1600		-0.03	-0.02	-0.11	-0.18	0.00	0.02	-0.02	-0.00	0.28	0.29	0.46	0.70

Table B.14: Simulation of validating(PredMod_ext) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

		sNB				TPR				FPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
	Estimand:	0.837	0.762	0.702	0.646	0.913	0.871	0.837	0.804	0.077	0.048	0.035	0.026
						Standard Deviation							
200	observed	0.038	0.059	0.081	0.105	0.027	0.033	0.037	0.039	0.027	0.022	0.019	0.016
	bootstrap	0.038	0.059	0.080	0.104	0.028	0.033	0.037	0.039	0.027	0.021	0.018	0.016
	analytic	0.038	0.059	0.080	0.104	0.028	0.033	0.037	0.040	0.027	0.021	0.018	0.016
400	observed	0.027	0.041	0.057	0.074	0.020	0.024	0.026	0.028	0.019	0.015	0.013	0.011
	bootstrap	0.027	0.041	0.056	0.073	0.020	0.024	0.026	0.028	0.019	0.015	0.013	0.011
	analytic	0.027	0.041	0.056	0.073	0.020	0.024	0.026	0.028	0.019	0.015	0.013	0.011
800	observed	0.019	0.029	0.039	0.052	0.014	0.017	0.019	0.020	0.013	0.011	0.009	0.008
	bootstrap	0.019	0.029	0.040	0.052	0.014	0.017	0.018	0.020	0.013	0.011	0.009	0.008
	analytic	0.019	0.029	0.040	0.052	0.014	0.017	0.018	0.020	0.013	0.011	0.009	0.008
						95% Coverage							
200	bootstrapP	93.6	93.0	91.7	91.2	93.0	93.4	93.2	95.4	93.5	93.6	85.9	92.3
	bootstrapW	92.9	92.8	91.2	90.2	93.2	93.9	93.2	93.8	94.5	85.9	86.5	92.7
	analytic	92.9	92.8	91.2	90.2	93.2	93.9	93.2	93.1	94.5	85.6	86.6	92.6
400	bootstrapP	95.2	94.5	93.9	93.4	94.8	94.3	94.3	94.8	93.2	94.6	94.7	88.7
	bootstrapW	95.0	94.3	93.4	92.0	93.5	95.3	95.0	94.9	93.5	91.1	91.4	88.7
	analytic	95.3	94.3	93.4	92.2	95.1	95.3	95.3	94.9	93.5	91.1	91.6	88.7
800	bootstrapP	95.3	94.8	94.7	94.8	95.3	94.8	94.7	95.0	94.6	94.2	95.2	94.8
	bootstrapW	95.5	94.7	94.1	93.9	94.4	95.0	94.5	94.4	94.8	94.3	93.3	94.9
	analytic	95.6	94.8	94.1	93.8	94.3	94.7	94.0	94.3	94.7	94.3	93.4	94.9

Table B.15: Simulation of validating(PredMod_ext) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

Difference in Net Benefit between PredMod and PredMod$_{ext}$

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
Model 1	0.226	0.32	0.057	0.057	94.6	94.8
Model 2	0.229	-0.34	0.054	0.054	95.1	95.1
Model 3	0.228	0.16	0.056	0.054	94.0	94.2
Model 4	0.232	-0.53	0.053	0.053	94.9	95.0
Model 5	0.218	0.59	0.049	0.049	94.3	94.5

Table B.16: Estimation of sNB(rT=0.2) for 5 prespecified models (developed from 5 independent samples of N=600) over 5000 replicates of external validation sets (N=400).

		ΔsNB				ΔTPR				ΔFPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
						True Value							
		0.164	0.229	0.267	0.289	0.075	0.159	0.229	0.289	-0.089	-0.031	-0.010	0.000
						% Bias							
200		0.05	-0.66	-0.79	-0.41	0.35	-0.46	-0.14	0.08	-0.20	-1.12	-4.75	1122.00
400		0.08	-0.34	-0.00	0.03	-0.09	-0.12	0.10	-0.02	0.23	-0.84	-0.64	-118.95
800		-0.29	-0.14	-0.50	-0.34	-0.09	-0.00	0.01	0.10	-0.46	-0.44	-3.58	1005.96
1600		-0.30	-0.26	-0.32	-0.20	-0.18	-0.05	-0.14	-0.05	-0.40	-0.71	-1.36	325.09

Table B.17: Simulation of validating(PredMod_ext-PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

		ΔsNB				ΔTPR				ΔFPR			
N	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
	Estimand:	0.164	0.229	0.267	0.289	0.075	0.159	0.229	0.289	-0.089	-0.031	-0.010	0.000
							Standard Deviation						
200	observed	0.051	0.075	0.100	0.124	0.035	0.044	0.048	0.049	0.036	0.028	0.023	0.019
	bootstrap	0.051	0.076	0.098	0.122	0.036	0.044	0.048	0.050	0.036	0.028	0.022	0.019
	analytic	0.051	0.076	0.098	0.122	0.036	0.044	0.048	0.050	0.036	0.028	0.022	0.019
400	observed	0.036	0.054	0.070	0.087	0.025	0.031	0.034	0.036	0.026	0.020	0.016	0.013
	bootstrap	0.036	0.054	0.070	0.086	0.026	0.031	0.034	0.035	0.026	0.020	0.016	0.013
	analytic	0.036	0.054	0.070	0.086	0.026	0.031	0.034	0.035	0.026	0.020	0.016	0.013
800	observed	0.025	0.038	0.049	0.061	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
	bootstrap	0.026	0.038	0.049	0.061	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
	analytic	0.026	0.038	0.049	0.061	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
							95% Coverage						
200	bootstrapP	94.7	94.3	93.2	92.2	94.7	94.1	94.9	94.8	94.1	92.3	88.6	80.6
	bootstrapW	94.9	94.6	94.3	95.3	94.9	94.0	94.9	94.7	93.8	94.4	93.0	93.6
	analytic	94.8	94.7	94.4	95.3	95.1	93.9	94.9	94.6	93.5	94.4	92.8	93.6
400	bootstrapP	94.8	95.1	94.4	94.0	94.6	94.6	95.0	94.3	94.4	93.9	92.7	91.6
	bootstrapW	94.7	95.1	94.9	95.1	95.1	94.4	95.0	94.3	94.6	94.4	94.6	95.4
	analytic	94.7	95.1	95.0	95.0	95.2	94.2	94.9	94.3	94.5	94.4	94.8	95.4
800	bootstrapP	95.0	94.6	94.8	94.5	94.7	94.6	94.9	94.8	94.8	94.5	94.5	93.7
	bootstrapW	95.2	94.8	95.0	95.2	94.8	94.6	95.0	94.7	94.8	94.7	94.8	94.9
	analytic	95.1	94.7	95.0	95.2	94.8	94.6	95.0	94.7	94.7	94.7	94.9	95.2

Table B.18: Simulation of validating(PredMod_ext-PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

Unmatched Case-Control - Nested

Base Model - PredMod

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
Model 1	0.533	-0.26	0.055	0.055	94.6	94.7
Model 2	0.528	-0.23	0.057	0.057	94.7	94.9
Model 3	0.531	0.06	0.056	0.057	94.9	94.6
Model 4	0.530	-0.29	0.056	0.056	94.5	94.4

Table B.19: Estimation of sNB(rT=0.2) for 5 prespecified models (developed from 5 independent samples of N=600) over 5000 replicates of external validation sets (N=2800).

			sNB				TPR				FPR				
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
							True Value								
			0.673	0.533	0.435	0.357	0.838	0.712	0.608	0.515	0.166	0.080	0.045	0.026	0.100
							% Bias								
1400	200		-0.13	-0.32	-0.12	0.07	0.07	0.09	0.12	0.17	0.19	0.54	-0.04	-0.32	0.02
2800	400		-0.12	-0.26	0.01	0.58	0.00	-0.08	-0.03	0.07	0.14	0.10	-0.49	-1.48	0.01
5600	800		-0.04	0.04	0.06	0.20	0.01	0.04	0.10	0.07	-0.09	-0.22	-0.07	-0.52	-0.07

Table B.20: Simulation of validating(PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

			sNB				TPR				FPR				
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
		Estimand:	0.673	0.533	0.435	0.357	0.838	0.712	0.608	0.515	0.166	0.080	0.045	0.026	0.100
								Standard Deviation							
1400	200	observed	0.055	0.079	0.097	0.110	0.037	0.045	0.049	0.050	0.037	0.027	0.021	0.016	0.008
		bootstrap	0.055	0.079	0.096	0.111	0.037	0.045	0.049	0.050	0.037	0.027	0.021	0.016	0.008
		analytic	0.055	0.078	0.095	0.110	0.037	0.045	0.049	0.050	0.037	0.027	0.021	0.016	0.008
2800	400	observed	0.038	0.055	0.068	0.078	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011	0.006
		bootstrap	0.039	0.055	0.067	0.077	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011	0.006
		analytic	0.039	0.055	0.067	0.077	0.026	0.032	0.034	0.035	0.026	0.019	0.015	0.011	0.006
5600	800	observed	0.027	0.039	0.047	0.054	0.018	0.022	0.024	0.025	0.019	0.014	0.010	0.008	0.004
		bootstrap	0.027	0.039	0.048	0.055	0.018	0.023	0.024	0.025	0.019	0.014	0.010	0.008	0.004
		analytic	0.027	0.039	0.047	0.055	0.018	0.023	0.024	0.025	0.019	0.014	0.010	0.008	0.004
								95% Coverage							
1400	200	bootstrapP	94.6	94.3	93.3	91.2	92.8	94.5	94.7	95.2	93.4	93.4	93.6	92.5	94.6
		bootstrapW	94.4	94.3	93.5	90.9	92.8	95.6	94.2	95.1	94.0	90.4	94.3	93.0	94.7
		analytic	94.2	94.1	93.4	90.7	92.8	95.6	94.7	95.7	94.4	90.0	94.3	93.0	94.6
2800	400	bootstrapP	95.0	94.6	94.1	93.1	94.3	95.3	95.2	94.6	94.4	94.4	93.1	88.6	94.5
		bootstrapW	94.9	94.8	93.9	92.5	94.3	95.3	95.3	94.0	94.2	95.2	94.0	88.6	94.5
		analytic	94.8	94.7	93.8	92.4	94.2	95.3	95.3	93.9	93.9	95.5	94.0	88.6	94.5
5600	800	bootstrapP	95.2	94.9	94.5	94.7	95.0	95.3	95.1	95.5	94.7	94.4	94.7	94.5	94.9
		bootstrapW	95.1	94.7	94.4	94.4	95.2	95.0	94.9	95.6	94.9	95.0	92.9	94.6	95.0
		analytic	95.1	94.8	94.4	94.4	95.2	94.9	94.9	95.6	95.3	95.1	92.9	94.6	95.1

Table B.21: Simulation of validating(PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

Difference in Net Benefit between PredMod and PredMod$_{ext}$

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
Model 1	0.762	-0.14	0.042	0.042	94.8	94.5
Model 2	0.756	-0.08	0.043	0.043	94.1	93.9
Model 3	0.763	-0.11	0.042	0.042	94.4	94.2
Model 4	0.748	-0.04	0.043	0.043	94.2	94.0

Table B.22: Estimation of sNB(rT=0.2) for 5 prespecified models (developed from 5 independent samples of N=600) over 5000 replicates of external validation sets (N=2800).

			sNB				TPR				FPR				
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
								True Value							
			0.837	0.762	0.702	0.646	0.913	0.871	0.837	0.804	0.077	0.048	0.035	0.026	0.100
								% Bias							
1400	200		-0.03	-0.01	0.06	-0.08	0.01	0.03	0.06	0.07	-0.26	-0.44	-0.70	-0.06	0.02
2800	400		-0.09	-0.14	-0.19	-0.18	-0.04	-0.04	-0.07	-0.06	0.15	0.30	0.21	0.06	0.01
5600	800		-0.03	0.01	0.02	0.08	-0.02	-0.00	0.00	0.03	-0.18	-0.35	-0.38	-0.43	-0.07

Table B.23: Simulation of validating(PredMod_ext) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

			sNB				TPR				FPR				
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	ρ
		Estimand:	0.837	0.762	0.702	0.646	0.913	0.871	0.837	0.804	0.077	0.048	0.035	0.026	0.100
								Standard Deviation							
1400	200	observed	0.039	0.059	0.080	0.106	0.028	0.032	0.036	0.039	0.026	0.021	0.018	0.016	0.008
		bootstrap	0.040	0.060	0.082	0.107	0.028	0.033	0.037	0.039	0.026	0.021	0.018	0.016	0.008
		analytic	0.039	0.060	0.081	0.106	0.028	0.033	0.037	0.039	0.026	0.021	0.018	0.016	0.008
2800	400	observed	0.028	0.042	0.057	0.074	0.020	0.024	0.026	0.028	0.019	0.015	0.013	0.011	0.006
		bootstrap	0.028	0.042	0.058	0.075	0.020	0.024	0.026	0.028	0.019	0.015	0.013	0.011	0.006
		analytic	0.028	0.042	0.057	0.074	0.020	0.024	0.026	0.028	0.019	0.015	0.013	0.011	0.006
5600	800	observed	0.020	0.030	0.040	0.052	0.014	0.017	0.018	0.020	0.013	0.011	0.009	0.008	0.004
		bootstrap	0.020	0.030	0.040	0.053	0.014	0.017	0.018	0.020	0.013	0.011	0.009	0.008	0.004
		analytic	0.020	0.030	0.040	0.052	0.014	0.017	0.018	0.020	0.013	0.011	0.009	0.008	0.004
								95% Coverage							
1400	200	bootstrapP	94.6	93.7	92.4	91.0	93.3	94.4	94.2	95.7	93.6	94.0	85.6	91.7	94.6
		bootstrapW	94.3	92.9	91.6	89.7	93.4	94.6	94.2	94.0	94.4	86.0	86.3	92.2	94.7
		analytic	94.2	92.7	91.5	89.6	93.4	94.6	94.2	93.5	94.4	85.7	86.3	92.1	94.6
2800	400	bootstrapP	95.1	94.8	94.3	93.9	94.9	93.9	94.3	94.9	93.7	94.9	95.3	89.9	94.5
		bootstrapW	95.0	94.7	93.9	93.0	93.5	94.7	94.7	95.0	93.8	91.8	92.5	89.9	94.5
		analytic	94.9	94.5	93.8	93.0	95.1	94.7	94.9	95.0	93.8	91.8	92.6	89.9	94.5
5600	800	bootstrapP	94.7	94.9	94.7	93.8	94.4	94.7	94.8	94.9	94.7	93.4	94.9	93.8	94.9
		bootstrapW	94.7	94.9	94.5	93.4	93.5	94.8	94.6	94.6	94.9	93.5	92.6	93.9	95.0
		analytic	94.7	94.8	94.5	93.4	93.5	94.6	93.9	94.6	94.9	93.5	92.8	93.9	95.1

Table B.24: Simulation of validating(PredMod_ext) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

Difference in Net Benefit between PredMod and PredMod$_{ext}$

			Variance		Coverage	
	sNB_0	% Bias	MC	Ana	BS	Ana
Model 1	0.229	0.16	0.054	0.054	95.0	95.0
Model 2	0.228	0.27	0.054	0.055	95.0	95.4
Model 3	0.232	-0.50	0.054	0.054	94.7	94.7
Model 4	0.218	0.56	0.050	0.050	93.8	94.2

Table B.25: Estimation of sNB(rT=0.2) for 5 prespecified models (developed from 5 independent samples of N=600) over 5000 replicates of external validation sets (N=2800).

			ΔsNB				ΔTPR				ΔFPR			
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
							True Value							
			0.164	0.229	0.267	0.289	0.075	0.159	0.229	0.289	-0.089	-0.031	-0.010	0.000
							% Bias							
1400	200		0.38	0.71	0.36	-0.26	-0.69	-0.23	-0.09	-0.11	0.58	2.06	2.31	326.28
2800	400		0.03	0.16	-0.51	-1.13	-0.53	0.16	-0.16	-0.29	0.14	-0.23	-2.94	1917.72
5600	800		-0.00	-0.06	-0.03	-0.08	-0.32	-0.19	-0.24	-0.03	-0.01	-0.02	1.03	108.40

Table B.26: Simulation of validating(PredMod_ext-PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

			ΔsNB				ΔTPR				ΔFPR			
N_I	N_{II}	r_T:	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
		Estimand:	0.164	0.229	0.267	0.289	0.075	0.159	0.229	0.289	-0.089	-0.031	-0.010	0.000
							Standard Deviation							
1400	200	observed	0.053	0.079	0.100	0.124	0.036	0.044	0.048	0.050	0.036	0.028	0.022	0.018
		bootstrap	0.052	0.078	0.101	0.125	0.036	0.044	0.048	0.050	0.036	0.028	0.022	0.019
		analytic	0.052	0.077	0.100	0.123	0.036	0.044	0.048	0.050	0.036	0.028	0.022	0.019
2800	400	observed	0.036	0.054	0.071	0.088	0.025	0.031	0.034	0.035	0.026	0.020	0.016	0.013
		bootstrap	0.037	0.055	0.071	0.087	0.026	0.031	0.034	0.035	0.026	0.020	0.016	0.013
		analytic	0.037	0.054	0.070	0.087	0.026	0.031	0.034	0.035	0.026	0.020	0.016	0.013
5600	800	observed	0.026	0.038	0.050	0.061	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
		bootstrap	0.026	0.038	0.050	0.062	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
		analytic	0.026	0.038	0.050	0.062	0.018	0.022	0.024	0.025	0.018	0.014	0.011	0.009
							95% Coverage							
1400	200	bootstrapP	94.3	94.3	93.8	92.6	94.4	94.6	94.3	94.7	93.6	91.7	89.8	80.2
		bootstrapW	94.5	94.9	95.2	95.9	94.6	94.4	94.2	94.3	94.0	93.6	93.9	93.3
		analytic	94.4	94.7	95.0	95.6	94.7	94.3	94.2	94.2	93.8	93.6	93.7	93.3
2800	400	bootstrapP	94.8	95.0	93.5	93.1	94.2	94.9	95.0	94.7	94.6	94.3	92.6	91.2
		bootstrapW	94.9	95.2	94.4	94.8	94.7	95.1	95.1	94.6	94.5	94.6	94.8	94.7
		analytic	94.9	95.0	94.2	94.7	94.8	95.1	94.9	94.6	94.5	94.5	95.0	94.7
5600	800	bootstrapP	94.6	94.9	95.0	94.4	94.4	94.4	94.5	94.6	94.7	95.0	93.9	93.8
		bootstrapW	94.6	95.1	95.3	95.2	94.6	94.5	94.6	94.6	94.9	95.1	94.7	95.2
		analytic	94.7	95.1	95.3	95.0	94.5	94.4	94.6	94.6	94.8	95.1	94.9	95.3

Table B.27: Simulation of validating(PredMod_ext-PredMod) from a case-control sample with J=1. Results for each validation sample size are based on: 5000 replications.

B.10 Chapter 6.5 Simulation Study - Additional Results

Here we present additional simulation results for estimation from cohort data subject to censored outcomes. With the exception of the amount of censoring, all other parameters are the same as considered in Chapter 6.5. We present the same figure (coverage of 95% CI) and table (bias of point estimators and comparison of analytic variance estimates with observed Monte Carlo variance) as in the main text. We evaluated estimation under rates of: 0.01, 0.02, 0.04, 0.07, and 0.10 in the exponential censoring distribution. Observed behavior was ascertained from 5,000 replicates for sample sizes ranging between 1,000 and 40,000.

Under no censoring (rate parameter equal to 0), in combination with the generated event times, we anticipate simulated data sets to have about 86% of participants observed through the entirety of the 8 year study period and occurrence of the outcome event among roughly 14%, who consequently had less than 8 years of observation.

Kaplan-Meier Based Estimators

Censoring Parameter: 0.01

In this scenario, a rate parameter equal to 0.01 was used in the exponential censoring distribution. In combination with the generated event times, we anticipate simulated data sets to have about 79.4% of participants observed through the entirety of the 8 year study period. Of the roughly 20.6% with less than 8 years of observation, 63.9% are expected to have had the clinical outcome observed, with the remaining 36.1% (corresponding to $\approx$ 7.4% of the study cohort), lost to follow-up.

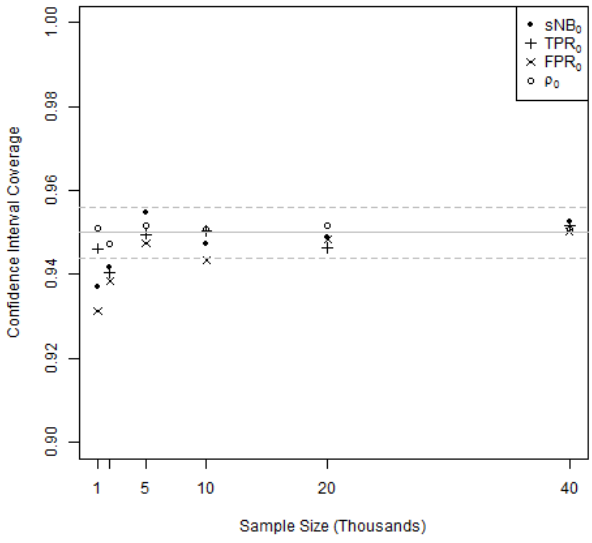


Figure B.3: Coverage of confidence intervals for Kaplan-Meier based estimators, under a 0.01 exponential censoring rate.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-0.003	-0.000	-0.004	0.001
2000		0.003	0.001	-0.006	-0.001
5000		-0.000	-0.000	-0.002	-0.000
10000		0.001	0.000	-0.001	0.000
20000		0.000	-0.000	-0.000	0.000
40000		0.000	0.000	-0.000	0.000
			Standard Deviation		
1000	observed	0.088	0.039	0.006	0.011
	analytic	0.084	0.038	0.006	0.011
2000	observed	0.062	0.028	0.004	0.008
	analytic	0.060	0.027	0.004	0.008
5000	observed	0.038	0.017	0.003	0.005
	analytic	0.038	0.017	0.003	0.005
10000	observed	0.027	0.012	0.002	0.004
	analytic	0.027	0.012	0.002	0.004
20000	observed	0.019	0.009	0.001	0.003
	analytic	0.019	0.009	0.001	0.003
40000	observed	0.014	0.006	0.001	0.002
	analytic	0.014	0.006	0.001	0.002

Table B.28: Evaluation of point and variance estimators for Kaplan-Meier based estimators, under a 0.01 exponential censoring rate.

Censoring Parameter: 0.02

In this scenario, a rate parameter equal to 0.02 was used in the exponential censoring distribution. In combination with the generated event times, we anticipate simulated data sets to have about 73.3% of participants observed through the entirety of the 8 year study period. Of the roughly 26.7% with less than 8 years of observation, 46.4% are expected to have had the clinical outcome observed, with the remaining 53.6% (corresponding to $\approx$ 14.3% of the study cohort), lost to follow-up.

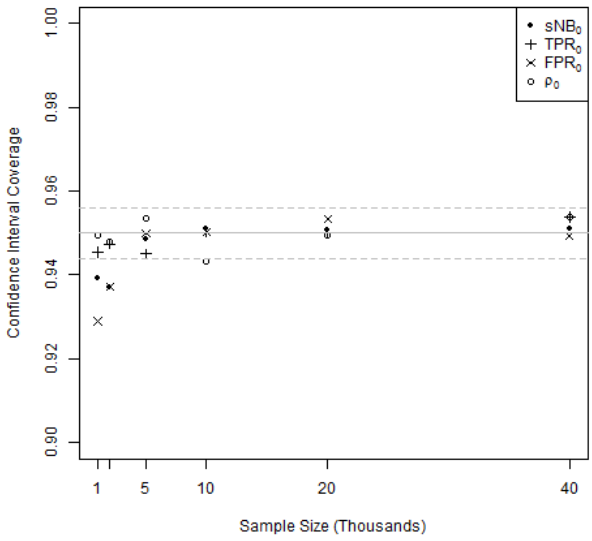


Figure B.4: Coverage of confidence intervals or Kaplan-Meier based estimators, under a 0.02 exponential censoring rate.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-0.005	0.000	-0.003	-0.001
2000		-0.001	0.000	-0.002	-0.000
5000		0.000	0.001	-0.000	0.000
10000		-0.001	-0.000	0.001	0.000
20000		-0.000	0.000	0.000	-0.000
40000		0.000	0.000	-0.001	-0.000
			Standard Deviation		
1000	observed	0.091	0.042	0.006	0.012
	analytic	0.087	0.041	0.006	0.012
2000	observed	0.065	0.029	0.004	0.008
	analytic	0.063	0.029	0.004	0.008
5000	observed	0.040	0.018	0.003	0.005
	analytic	0.040	0.018	0.003	0.005
10000	observed	0.028	0.013	0.002	0.004
	analytic	0.028	0.013	0.002	0.004
20000	observed	0.020	0.009	0.001	0.003
	analytic	0.020	0.009	0.001	0.003
40000	observed	0.014	0.006	0.001	0.002
	analytic	0.014	0.007	0.001	0.002

Table B.29: Evaluation of point and variance estimators for Kaplan-Meier based estimators, under a 0.02 exponential censoring rate.

Censoring Parameter: 0.04

In this scenario, a rate parameter equal to 0.04 was used in the exponential censoring distribution. In combination with the generated event times, we anticipate simulated data sets to have about 62.5% of participants observed through the entirety of the 8 year study period. Of the roughly 37.5% with less than 8 years of observation, 29.2% are expected to have had the clinical outcome observed, with the remaining 70.8% (corresponding to $\approx$ 26.5% of the study cohort), lost to follow-up.

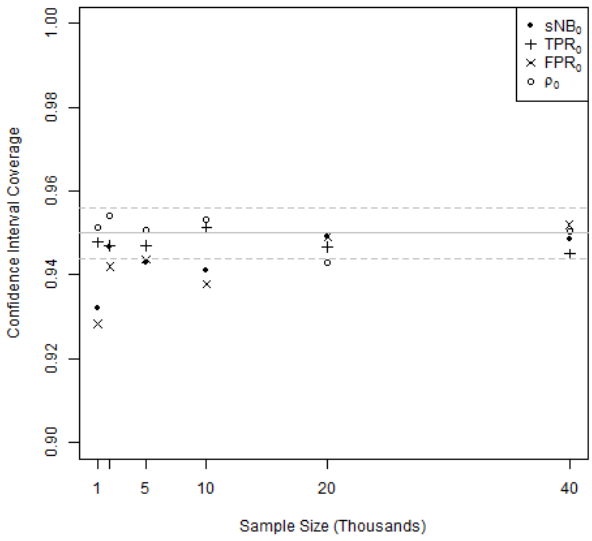


Figure B.5: Coverage of confidence intervals for Kaplan-Meier based estimators, under a 0.04 exponential censoring rate.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-0.002	0.001	-0.005	0.002
2000		-0.006	0.001	0.004	-0.001
5000		0.000	0.000	-0.002	0.000
10000		0.000	0.000	-0.001	0.000
20000		-0.000	0.000	-0.000	-0.000
40000		0.000	-0.000	-0.000	0.000
			Standard Deviation		
1000	observed	0.098	0.047	0.006	0.012
	analytic	0.094	0.046	0.006	0.012
2000	observed	0.069	0.033	0.004	0.009
	analytic	0.068	0.033	0.004	0.009
5000	observed	0.043	0.021	0.003	0.005
	analytic	0.043	0.021	0.003	0.006
10000	observed	0.031	0.015	0.002	0.004
	analytic	0.030	0.015	0.002	0.004
20000	observed	0.022	0.010	0.001	0.003
	analytic	0.022	0.010	0.001	0.003
40000	observed	0.015	0.007	0.001	0.002
	analytic	0.015	0.007	0.001	0.002

Table B.30: Evaluation of point and variance estimators for Kaplan-Meier based estimators, under a 0.04 exponential censoring rate.

Censoring Parameter: 0.07

In this scenario, a rate parameter equal to 0.07 was used in the exponential censoring distribution. In combination with the generated event times, we anticipate simulated data sets to have roughly 49.1% of participants observed through the entirety of the 8 year study period. Of the roughly 50.9% with less than 8 years of observation, 18% are expected to have had the clinical outcome observed, with the remaining 82% (corresponding to $\approx$ 41.7% of the study cohort), lost to follow-up.

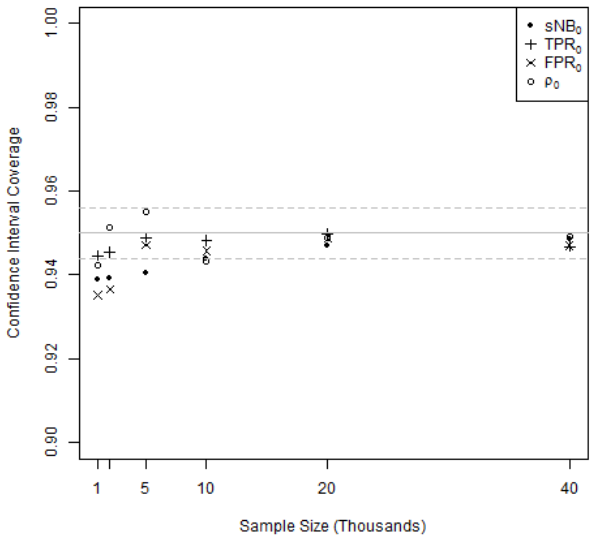


Figure B.6: Coverage of confidence intervals for Kaplan-Meier based estimators, under a 0.07 exponential censoring rate.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-0.012	0.002	0.005	0.000
2000		-0.004	0.002	0.002	-0.001
5000		0.000	0.001	-0.001	0.000
10000		-0.002	-0.000	0.001	0.001
20000		-0.001	0.000	-0.000	-0.001
40000		-0.000	0.000	0.000	-0.000
			Standard Deviation		
1000	observed	0.110	0.055	0.007	0.014
	analytic	0.105	0.054	0.007	0.014
2000	observed	0.078	0.039	0.005	0.010
	analytic	0.075	0.038	0.005	0.010
5000	observed	0.048	0.025	0.003	0.006
	analytic	0.048	0.024	0.003	0.006
10000	observed	0.034	0.017	0.002	0.004
	analytic	0.034	0.017	0.002	0.004
20000	observed	0.024	0.012	0.002	0.003
	analytic	0.024	0.012	0.002	0.003
40000	observed	0.017	0.009	0.001	0.002
	analytic	0.017	0.009	0.001	0.002

Table B.31: Evaluation of point and variance estimators for Kaplan-Meier based estimators, under a 0.07 exponential censoring rate.

Naive Estimators - Exclude Censored Observations from Analysis

Here we present results for the naive approach to analyzing censored data that assumes omits all incomplete observations from analysis. As done previously, all simulations are based on 5,000 replicates of data sets that contain between 1,000 to 20,000 observations for which the rate is varied between 0.01 and 0.07 in the independent exponential censoring distribution. Because the effective sample size, the number of observations contributing to estimation, varies across simulated data sets of the same cohort size, the usual estimates of variance have been reported on the asymptotic scale (i.e., the observed Monte Carlo variance times n or σ_n^2 instead of $\sqrt{\frac{\sigma_n^2}{n}}$, are reported).

Censoring Parameter: 0.1

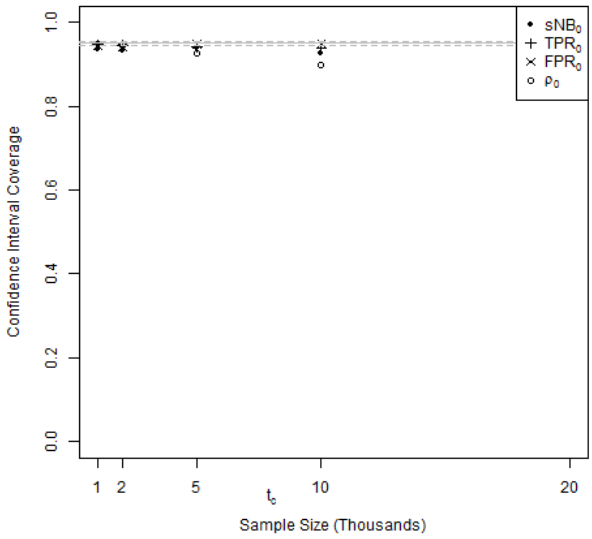


Figure B.7: Coverage of confidence intervals, under a 0.01 exponential censoring rate, for estimates that naively exclude censored observations from analysis.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		0.014	0.003	-0.002	0.015
2000		0.015	0.003	0.003	0.015
5000		0.020	0.004	0.001	0.017
10000		0.021	0.003	0.000	0.018
20000		0.022	0.003	-0.002	0.016
			Asymptotic Variance		
1000	observed	7.487	1.344	0.033	0.135
	analytic	6.917	1.271	0.031	0.122
2000	observed	7.572	1.380	0.035	0.126
	analytic	6.833	1.270	0.031	0.122
5000	observed	7.275	1.379	0.033	0.134
	analytic	6.747	1.267	0.031	0.122
10000	observed	7.457	1.436	0.034	0.132
	analytic	6.719	1.268	0.031	0.122
20000	observed	7.198	1.321	0.033	0.131
	analytic	6.715	1.269	0.031	0.122

Table B.32: Evaluation of naive point and variance estimators, under a 0.01 exponential censoring rate, that exclude all censored observatons from analysis.

Censoring Parameter: 0.2

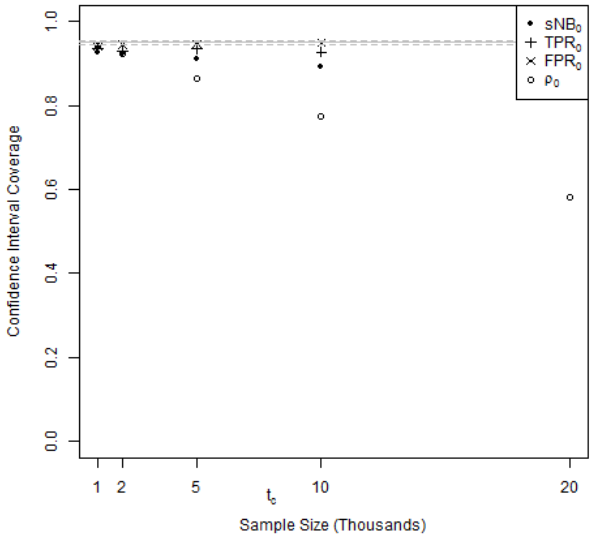


Figure B.8: Coverage of confidence intervals, under a 0.02 exponential censoring rate, for estimates that naively exclude censored observations from analysis.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		0.033	0.006	0.000	0.032
2000		0.038	0.006	0.001	0.034
5000		0.040	0.006	0.000	0.033
10000		0.041	0.007	-0.000	0.033
20000		0.042	0.007	-0.000	0.034
			Asymptotic Variance		
1000	observed	7.858	1.466	0.036	0.144
	analytic	6.680	1.241	0.031	0.123
2000	observed	7.685	1.460	0.037	0.137
	analytic	6.561	1.239	0.031	0.124
5000	observed	7.807	1.450	0.035	0.146
	analytic	6.519	1.240	0.031	0.123
10000	observed	7.535	1.399	0.036	0.144
	analytic	6.498	1.239	0.031	0.123
20000	observed	7.500	1.456	0.036	0.146
	analytic	6.483	1.239	0.031	0.124

Table B.33: Evaluation of naive point and variance estimators, under a 0.02 exponential censoring rate, that exclude all censored observatons from analysis.

Censoring Parameter: 0.4

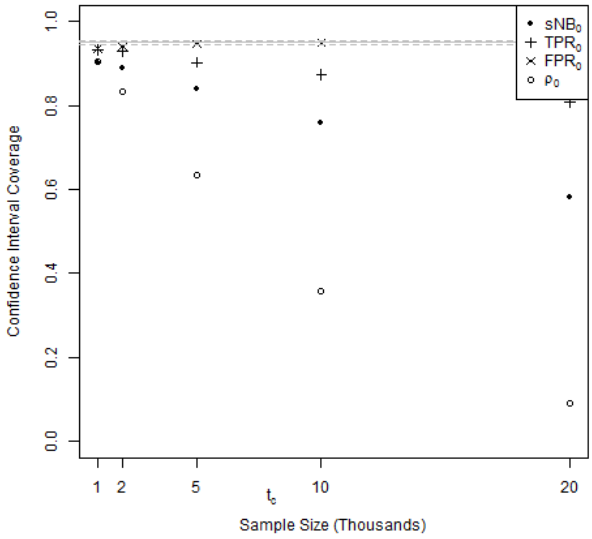


Figure B.9: Coverage of confidence intervals, under a 0.04 exponential censoring rate, for estimates that naively exclude censored observations from analysis.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		0.072	0.013	0.006	0.069
2000		0.081	0.012	-0.005	0.069
5000		0.082	0.013	-0.001	0.068
10000		0.081	0.012	-0.000	0.069
20000		0.083	0.013	0.000	0.069
			Asymptotic Variance		
1000	observed	8.574	1.649	0.045	0.169
	analytic	6.240	1.179	0.032	0.127
2000	observed	8.078	1.628	0.042	0.176
	analytic	6.092	1.184	0.031	0.127
5000	observed	8.142	1.589	0.043	0.171
	analytic	6.062	1.180	0.031	0.127
10000	observed	8.231	1.632	0.042	0.171
	analytic	6.044	1.182	0.031	0.127
20000	observed	7.842	1.579	0.041	0.172
	analytic	6.031	1.181	0.031	0.127

Table B.34: Evaluation of naive point and variance estimators, under a 0.04 exponential censoring rate, that exclude all censored observatons from analysis.

Censoring Parameter: 0.7

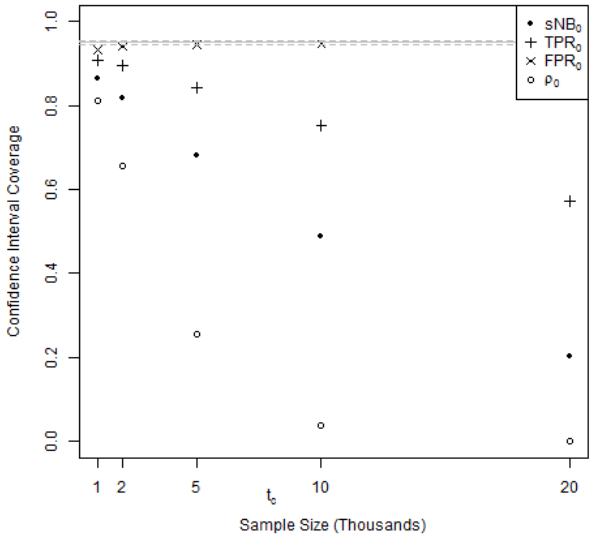


Figure B.10: Coverage of confidence intervals, under a 0.07 exponential censoring rate, for estimates that naively exclude censored observations from analysis.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		0.134	0.022	-0.001	0.123
2000		0.137	0.022	-0.000	0.122
5000		0.141	0.023	0.001	0.125
10000		0.142	0.022	0.000	0.125
20000		0.144	0.022	-0.001	0.125
			Asymptotic Variance		
1000	observed	9.495	1.915	0.054	0.224
	analytic	5.611	1.098	0.032	0.132
2000	observed	9.733	1.879	0.056	0.221
	analytic	5.519	1.099	0.032	0.132
5000	observed	9.260	1.875	0.054	0.227
	analytic	5.436	1.096	0.032	0.132
10000	observed	9.350	1.854	0.054	0.223
	analytic	5.418	1.097	0.032	0.132
20000	observed	9.323	1.875	0.055	0.224
	analytic	5.397	1.096	0.032	0.132

Table B.35: Evaluation of naive point and variance estimators, under a 0.07 exponential censoring rate, that exclude all censored observatons from analysis.

Naive Estimators - Assume Censored Observations are Controls

Here we present results for the naive approach to analyzing censored data that assumes all incomplete observations are controls. As done previously, all simulations are based on 5,000 replicates of data sets that contain between 1,000 to 20,000 observations for which the rate is varied between 0.01 and 0.07 in the independent exponential censoring distribution. To be consistent with the summaries used in the naive approach that omits censored observations from analysis, the usual estimates of variance have been reported on the asymptotic scale (i.e., the observed Monte Carlo variance times n or σ_n^2 instead of $\sqrt{\frac{\sigma_n^2}{n}}$, are reported).

Censoring Parameter: 0.1

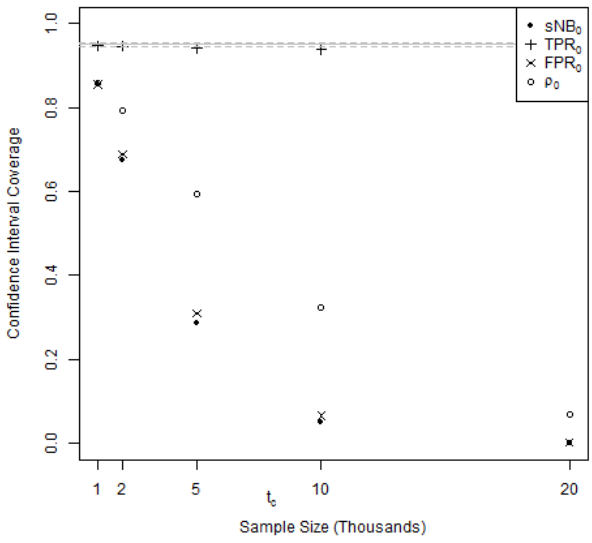


Figure B.11: Coverage of confidence intervals, under a 0.01 exponential censoring rate, for estimates that naively assume censored observations are controls.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-0.268	0.003	0.238	-0.061
2000		-0.266	0.003	0.243	-0.060
5000		-0.261	0.004	0.243	-0.059
10000		-0.258	0.003	0.241	-0.058
20000		-0.258	0.003	0.240	-0.059
			Asymptotic Variance		
1000	observed	10.077	1.344	0.037	0.117
	analytic	10.073	1.373	0.038	0.114
2000	observed	10.091	1.380	0.039	0.109
	analytic	9.936	1.372	0.038	0.114
5000	observed	9.723	1.379	0.037	0.116
	analytic	9.822	1.369	0.038	0.114
10000	observed	10.258	1.436	0.040	0.115
	analytic	9.767	1.370	0.038	0.114
20000	observed	9.717	1.321	0.038	0.113
	analytic	9.768	1.372	0.038	0.114

Table B.36: Evaluation of naive point and variance estimators, under a 0.01 exponential censoring rate, that assume all censored observations are controls.

Censoring Parameter: 0.2

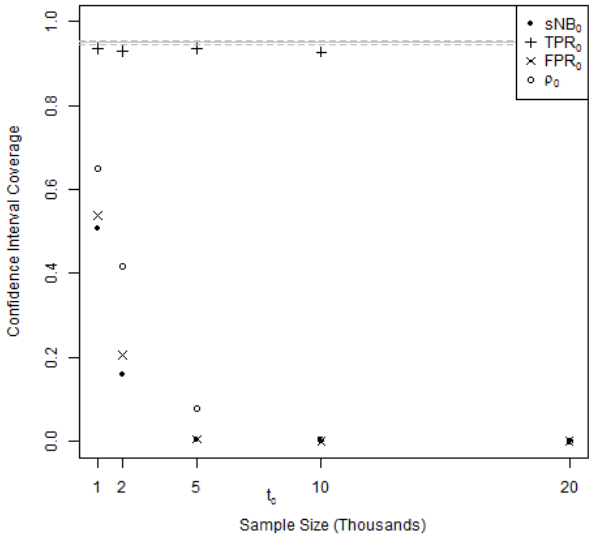


Figure B.12: Coverage of confidence intervals, under a 0.02 exponential censoring rate, for estimates that naively assume censored observations are controls.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-0.546	0.006	0.461	-0.116
2000		-0.537	0.006	0.462	-0.114
5000		-0.535	0.006	0.462	-0.115
10000		-0.534	0.007	0.463	-0.115
20000		-0.534	0.007	0.464	-0.114
			Asymptotic Variance		
1000	observed	13.734	1.466	0.044	0.108
	analytic	13.713	1.448	0.044	0.108
2000	observed	13.259	1.460	0.045	0.103
	analytic	13.436	1.446	0.044	0.108
5000	observed	13.715	1.450	0.043	0.109
	analytic	13.344	1.447	0.044	0.108
10000	observed	12.850	1.399	0.042	0.108
	analytic	13.303	1.447	0.044	0.108
20000	observed	13.278	1.456	0.043	0.110
	analytic	13.281	1.446	0.044	0.108

Table B.37: Evaluation of naive point and variance estimators, under a 0.02 exponential censoring rate, that assume all censored observations are controls.

Censoring Parameter: 0.4

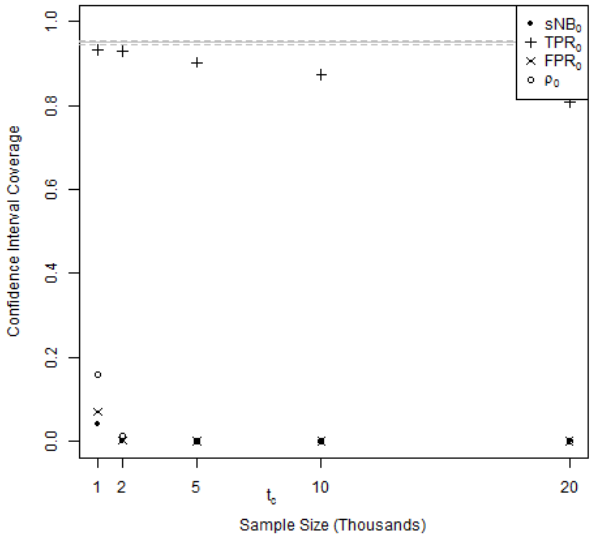


Figure B.13: Coverage of confidence intervals, under a 0.04 exponential censoring rate, for estimates that naively assume censored observations are controls.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-1.155	0.013	0.862	-0.215
2000		-1.142	0.012	0.856	-0.215
5000		-1.140	0.013	0.860	-0.215
10000		-1.139	0.012	0.861	-0.215
20000		-1.136	0.013	0.860	-0.215
			Asymptotic Variance		
1000	observed	24.002	1.649	0.056	0.096
	analytic	23.896	1.605	0.054	0.098
2000	observed	22.379	1.628	0.053	0.099
	analytic	23.359	1.612	0.054	0.098
5000	observed	22.553	1.589	0.052	0.097
	analytic	23.199	1.607	0.054	0.098
10000	observed	22.598	1.632	0.054	0.096
	analytic	23.106	1.610	0.054	0.098
20000	observed	22.223	1.579	0.053	0.097
	analytic	23.024	1.608	0.054	0.098

Table B.38: Evaluation of naive point and variance estimators, under a 0.04 exponential censoring rate, that assume all censored observations are controls.

Censoring Parameter: 0.7

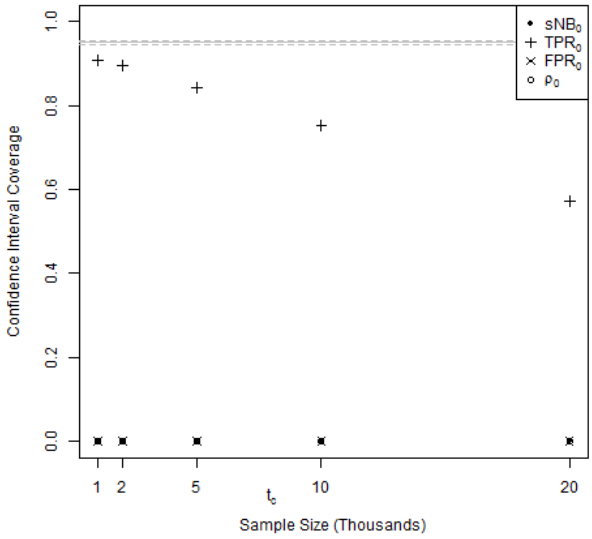


Figure B.14: Coverage of confidence intervals, under a 0.07 exponential censoring rate, for estimates that naively assume censored observations are controls.

N		sNB	TPR	FPR	ρ
			True Value		
		0.424	0.762	0.027	0.140
			% Bias		
1000		-2.233	0.022	1.357	-0.346
2000		-2.207	0.022	1.353	-0.345
5000		-2.188	0.023	1.353	-0.344
10000		-2.190	0.022	1.355	-0.344
20000		-2.184	0.022	1.353	-0.344
			Asymptotic Variance		
1000	observed	48.554	1.915	0.066	0.082
	analytic	50.595	1.886	0.067	0.083
2000	observed	48.505	1.879	0.067	0.081
	analytic	49.108	1.884	0.066	0.083
5000	observed	48.288	1.875	0.066	0.084
	analytic	48.112	1.879	0.066	0.083
10000	observed	46.181	1.854	0.065	0.082
	analytic	48.025	1.882	0.067	0.083
20000	observed	47.933	1.875	0.068	0.081
	analytic	47.783	1.880	0.066	0.083

Table B.39: Evaluation of naive point and variance estimators, under a 0.07 exponential censoring rate, that assume all censored observations are controls.

B.11 Standardized Net Benefit for a Trichotomous Decision Rule

We extend the derivation of standardized net benefit, in terms of underlying utilities, to the scenario of a dichotomous clinical decision rule. We now denote the three actions with letters: $\mathcal{X} = \{l, m, h\}$, which conceptually represent actions appropriate to low, medium, and high-risk patients. We continue to denote case-control status as $d \in \{0, 1\}$, and the corresponding 6 utilities by u_{dx}.

The universal decision to treat no-one N is now interpreted as the universal decision of treating everyone as low-risk. The perfect oracle rule P is now interpreted as the decision to treat cases as high-risk and controls as low-risk. The expected utility of this decision can be calculated as follows:

$$\begin{aligned} \mathbb{E}U_N &= \rho u_{1l} + (1-\rho)u_{0l} \\ \mathbb{E}U_P &= \rho u_{1h} + (1-\rho)u_{0l} \end{aligned}$$

Classification into one of the risk-based actions, depends on covariates, $X = R(W)$ and is characterized through:

$$\begin{aligned} HR_d &= Pr[X = h \mid D = d] \\ MR_d &= Pr[X = m \mid D = d] \\ LR_d &= Pr[X = l \mid D = d], \end{aligned}$$

and we note that $HR_d + MR_d + LR_d = 1$. These quantities represent the case or control-specific rates of being classified as high, medium or low-risk, and consequently recommended to take the corresponding high, medium, or low-risk treatment course, based on the rule R. The expected utility of the rule is thus:

$$\mathbb{E}U_R = \rho\left\{HR_1u_{1h} + MR_1u_{1m} + LR_1u_{1l}\right\} + (1-\rho)\left\{HR_0u_{0h} + MR_0u_{0m} + LR_0u_{0l}\right\}$$

and the differences in expected utility between the rule R and either N or P are:

$$\begin{aligned}
\mathbb{E}U_R - \mathbb{E}U_N &= \rho\{HR_1u_{1h} + MR_1u_{1m} + (LR_1 - 1)u_{1l}\} + (1-\rho)\{HR_0u_{0h} + MR_0u_{0m} + (LR_0 - 1)u_{0l}\} \\
&= \rho\{HR_1(u_{1h} - u_{1l}) + MR_1(u_{1m} - u_{1l})\} + (1-\rho)\{HR_0(u_{0h} - u_{0l}) + MR_0(u_{0m} - u_{0l})\} \\
&= \rho\{HR_1B_{hl}^{case} + MR_1B_{ml}^{case}\} - (1-\rho)\left\{HR_0B_{lh}^{ctrl} + MR_0B_{lm}^{ctrl}\right\},
\end{aligned}$$

where we have introduced the notation $B_{ml}^{case} = u_{1m} - u_{1l}$ to denote the benefit to treating as medium risk over low-risk, $B_{hl}^{case} = u_{1h} - u_{1l}$ to denote the benefit to treating as high risk over low-risk and analogously for the benefits to treating controls. These benefits are potentially gained among cases and lost among controls, in accordance with the classification accuracies of the rule. The increase in expected utility achieved by a perfect rule is:

$$\begin{aligned}
\mathbb{E}U_P - \mathbb{E}U_N &= \rho u_{1h} + (1-\rho)U_{0l} - \rho u_{1l} - (1-\rho)u_{0l} \\
&= \rho B_{hl}^{case}
\end{aligned}$$

which is achieved by correctly treating the cases as high risk and not changing the treatment of controls as low-risk. The relative utility, and hence standardized net benefit for a trichotomous rule, follows as:

$$\begin{aligned}
\frac{\mathbb{E}U_P - \mathbb{E}U_N}{\mathbb{E}U_R - \mathbb{E}U_N} &= \frac{1}{\rho B_{hl}^{case}}\left\{\rho\left(HR_1B_{hl}^{case} + MR_1B_{ml}^{case}\right) - (1-\rho)\left(HR_0B_{lh}^{ctrl} + MR_0B_{lm}^{ctrl}\right)\right\} \\
&= HR_1 + MR_1\frac{B_{ml}^{case}}{B_{hl}^{case}} - \frac{(1-\rho)}{\rho}\left(HR_0\frac{B_{lh}^{ctrl}}{B_{hl}^{case}} + MR_0\frac{B_{lm}^{ctrl}}{B_{hl}^{case}}\right).
\end{aligned}$$

As stated in the Discussion, this measure has reduced the need for 6 absolute utilities down to three ratios of differences in utilities. Qualitatively, the expression is quite similar to the net benefit of a dichotomous rule, only now there is the consideration of getting a case or control "partially" right, in that treating a control as medium-risk is less undesirable than treating a control as high-risk, and analogously for cases.

Appendix C

ADDITIVE INTERACTION

C.1 Candidate Influence Function

Here we calculate the influence function of $\psi_{jk,n}$, the empirical average of the MLE of $\theta_{jk,0}(w) = Pr_0(Y = 1|G = j, X = k, W = w)$, as an estimator of $\psi_{jk,0} = \mathbb{E}^0_W \theta_{jk,0}(W)$ based on an i.i.d. sample of size n, with each observation $O^E \sim P^E_0$. The approach is to 1) calculate $\theta_{jk,n}$, the MLE of $\theta_{jk,0}$, under the working assumption that the covariates have a finite discrete distribution, 2) simplify the expansion of $\psi_{jk,n}$ around $\psi_{jk,0}$ to reveal the influence function, and 3) establish that the candidate (under the working assumption) influence function is also an influence function under general covariate structures.

1. MLE of $\theta_{jk,0}$ The likelihood, conditional on covariates, is the product of the usual conditional Bernoulli likelihoods. With $dP_0(w) = q_W(w)dw$, the product of its density with respect to a measure, the full likelihood may be written explicitly:

$$L(o^E) = \begin{cases} \pi(y,x,w)\theta_{gx}(w)^y \{1-\theta_{gx}(w)\}^{(1-y)} q_G(g;x,w)q_X(x;w)q_W(w) & \delta = 1 \\ \{1-\pi(y,x,w)\}\,\overline{\theta}_x(w)^y \left\{1-\overline{\theta}_x(w)\right\}^{(1-y)} q_X(x;w)q_W(w) & \delta = 0 \end{cases}$$

where π is the probability of $\Delta = 1$ conditional on (Y, X, W), $\overline{\theta}_x$ is the probability of $Y = 1$ conditional on $(X = x, W)$, q_G is the probability of G conditional on (X, W), and q_X is the probability of X conditional on W.

Assuming a discrete working model for the baseline covariates, we can write $q_W(w) = \sum_m p_k I[w = w_m]$ where $\{w_1, \ldots, w_M\}$ are the M levels of W with probability masses $\{p_1, \ldots, p_M\}$. Given an observation, the likelihood is then a function of a finite number of parameters. The vector of influence functions for the MLE of the vector of likelihood parameters can be calculated classically as $I^{-1}U$ where I is the Fisher information matrix and U is a vector of scores. Orthogonality between the scores for the π and q_X nuisance parameters, with each of the other likelihood parameters, and pairwise between scores for parameters pertaining to different values of w, reduces the dimension of the linear algebra; the MLE of $\theta_{jk,0}(w_m)$ can be expressed in terms of the submatrix of the Fisher information and the subvector of scores corresponding to the parameters for $\theta_{gx}(w_m)$ and $q_G(g;x,w_m)$.

The scores, for parameters conditional on $X = k$ and $W = w_m$ are:

$$
\begin{aligned}
U_{0k,m}(o^E) &= \left[\delta(1-g)\frac{y-\theta_{0k}(w)}{\theta_{0k}(w)(1-\theta_{0,k}(w))} + (1-\delta)q_G(0;k,w)\frac{(y-\overline{\theta}_k(w))}{\overline{\theta}_j k(w)(1-\overline{\theta}_k(w))}\right] Ind_{km} \\
U_{1k,m}(o^E) &= \left[\delta g\frac{y-\theta_{1k}(w)}{\theta_{1k}(w)(1-\theta_{1k}(w))} + (1-\delta)q_G(1;k,w)\frac{(y-\overline{\theta}_k(w))}{\overline{\theta}_k(w)(1-\overline{\theta}_k(w)}\right] Ind_{km} \\
U_{Gk,m}(o^E) &= \left[\delta\frac{g-q_G(1;k,w)}{q_G(1;k,w)q_G(0;k,w)} + (1-\delta)\frac{y-\overline{\theta}_k(w)}{\overline{\theta}_j(w)(1-\overline{\theta}_k(w))}(\theta_{1k}(w)-\theta_{0k}(w))\right] Ind_{km}
\end{aligned}
$$

where Ind_{km} abbreviates the indicator for a treatment-covariate level $[x = k, w = w_m]$, $\overline{\theta}_j(w_m) = q_G(1;k,w_m)\theta_{1k}(w_m) + q_G(0;k,w_m)\theta_{0j}(w_m)$, and the scores $(U_{0k,m}, U_{1k,m}, U_{Gk,m})$ correspond to the parameters $\{\theta_{0k}(w_m), \theta_{0k}(w_m), q_G(1;k,w_m)\}$.

The Fisher information, corresponding to the likelihood parameters for a particular covariate level, can be calculated in terms of the above scores. To emphasize the likelihood parameters, we use subscripts to denote the corresponding pair of scores and the argument to denote the treatment-covariate value; e.g., $I_{10}(k,m)$ is the inner product of the scores $U_{1k,m}$ and $U_{0k,m}$, which would correspond to the $(4,4)$ entry of the information matrix using the complete 6-vector of scores ordered as $U_m = \{U_{00,m}, U_{10,m}, U_{G0,m}, U_{01,m}, U_{11,m}, U_{G1,m}\}$. Treatment-specific submatrices are defined by:

$$
\begin{aligned}
I_{00}(k,m) &= \left[X_0(k,m) + X(k,m)q_G(0;k,w_m)\right] q_G(0;k,w_m)q_X(k;w_m)p_m \\
I_{10}(k,m) &= X(k,m)q_G(1;k,w_m)q_G(0;k,w_m)q_X(k;w_m)p_m \\
I_{11}(k,m) &= \left[X_1(k,m) + X(k,m)q_G(1;k,w_m)\right] q_G(1;k,w_m)q_X(k;w_m)p_m \\
I_{0G}(k,m) &= \left[\pi(0,k,m) - \pi(1,k,m) + q_G(0;k,w_m)X(k,m)\left\{\theta_{1k}(w_m) - \theta_{0k}(w_m)\right\}\right] q_X(k;w_m)p_m \\
I_{1G}(k,m) &= \left[\pi(1,k,m) - \pi(0,k,m) + q_G(1;k,w_m)X(k,m)\left\{\theta_{1k}(k) - \theta_{0k}(w_m)\right\}\right] q_X(k;w_m)p_m \\
I_{GG}(k,m) &= \left[X(k,m)\left\{\theta_{1k}(w_m) - \theta_{0k}(w_m)\right\}^2 + \frac{\pi(1,k,m)\theta_{1k}(w_m) + \pi(0,k,m)\left\{1-\theta_{1k}(w_m)\right\}}{q_G(1;k,w_m)}\right. \\
&\quad \left. + \frac{\pi(1,k,m)\theta_{0k}(w_m) + \pi(0,k,m)\left\{1-\theta_{0k}(w_m)\right\}}{q_G(0;k,w_m)}\right] q_X(k;w_m)p_m
\end{aligned}
$$

where:

$X_0(k,m) = \frac{\pi(1,k,m)}{\theta_{0k}(w_m)} + \frac{\pi(0,k,m)}{1-\theta_{0k}(w_m)}$, and

$X_1(k,m) = \frac{\pi(1,k,m)}{\theta_{1k}(w_m)} + \frac{\pi(0,k,m)}{1-\theta_{1k}(w_m)}$, and
$X(k,m) = \frac{1-\pi(1,k,m)}{\bar{\theta}_k(w_m)} + \frac{1-\pi(0,k,m)}{1-\bar{\theta}_k(w_m)}$.

The full 6-by-6 information matrix has block diagonal form with the 3-by-3 upper block corresponding to $j = 0$ (comparison treatment) and the lower 3-by-3 block to $j = 1$ (treatment of study). In a model with known genetic-exposure distribution, the U_{Gj} parameters are eliminated and the 4-by-4 information matrix has block diagonal structure with 2-by-2 blocks defined by I_{00}, I_{01}, I_{11} for each treatment level. In the model with known independence, the two parameters $q_G(1;0,w_m), q_G(1;1,w_m)$ consolidate into one $q_G(1;w_m)$, the vector of scores becomes $U_m = \{U_{00,m}, U_{10,m}, U_{G,m}, U_{01,m}, U_{11,m}\}$, with $U_G = U_{G0} + U_{G1}$, the 5-by-5 information matrix no longer has block diagonal form, and $I_{GG}(j,w_m) = I_{GG}(1,w_m) + I_{GG}(0,w_m)$. We note that all non-zero entries of the information matrix contain a factor of p_k and it can be written $I(w_m) = \widetilde{I}(w_m)p_m$.

It will be useful to express the influence functions for the MLE of $\theta_{jk}(w_m)$ as:

$$IF_{jk,m}(o^E) \;=\; \sum_l I^{rl}(w_m)U_{m,l}(o^E) = \sum_l \frac{1}{p_m}\widetilde{I}^{rl}(w_m)U_{m,l}(o^E)$$

which uses superscripts to denote elements of the inverse of the matrix $I(w_m)$ and r to indicate the index of $U_{jk,m}$ in the vector of scores U_m and equals the usual Euclidean inner product between the rth row of $I^{-1}(w_m)$ and the vector of scores U_m.

2. Expansion of ψ_{jk} In the expansion of $\psi_{jk,n}$ about $\psi_{jk,0}$, we express expectation through integration, using $dP_0(w)$ to denote the measure induced by the component of the joint distribution P_0^E governing the covariates and $dP_n(w)$ the empirical covariate distribution,

this gives:

$$\begin{aligned}
\psi_{jk,n} - \psi_{jk,0} &= \int \theta_{jk,n}(w)\, dP_n(w) - \psi_{jk,0} \\
&= \int [\theta_{jk,n}(w) - \theta_{jk,0}(w)]\, dP_0(w) + \int \theta_{jk,0}(w) dP_n(w) - \psi_{jk,0} \\
&\quad + \int [\theta_{jk,n}(w) - \theta_{jk,0}(w)]\, d(P_n - P_0)(w) \\
&= \int [\theta_{jk,n}(w) - \theta_{jk,0}(w)]\, dP_0(w) + \frac{1}{n}\sum_i \theta_{jk,0}(W_i) - \psi_{jk,0} + o_P(n^{-\frac{1}{2}})
\end{aligned}$$

When the empirical process term, $\int [\theta_{jk,n}(w) - \theta_{jk,0}(w)]\, d(P_n - P_0)(w)$ is $o_P(n^{-\frac{1}{2}})$.

Under the working assumption of finite covariate levels, $dP_0(w) = \sum_k p_k I[w = w_m]$ and we can use the developed expression for the MLE of $\theta_{jk,0}$, $\theta_{jk,n} = I^{-1}U$. Continuing the above expansion yields:

$$= \sum_m [\theta_{jk,n}(w_m) - \theta_{jk,0}(w_m)]\, p_{m,0} + \frac{1}{n}\sum_i \theta_{jk,0}(W_i) - \psi_{jk,0} + o_P(n^{-\frac{1}{2}})$$

and we now only need focus on the first term, which is summed over the values (finite) of W. Using $\widetilde{U}(o^E)\mathrm{I}[W = w_m] = U_m(o^E)$ and denoting the lth entry of $\widetilde{U}$ by $\widetilde{U}_l$

$$\begin{aligned}
\sum_m [\theta_{jk,n}(w_m) - \theta_{jk,0}(w_m)]\, p_{k,0} &= \sum_m \left[\frac{1}{n}\sum_i \frac{1}{p_{m,0}} \widetilde{I}^{rl}(w_m) U_{m,l}(O_i^E) + o_P(n^{-1/2})\right] p_{m,0} \\
&= \sum_m \left[\frac{1}{n}\sum_i \widetilde{I}^{rl}(w_m)\widetilde{U}_l(O_i^E)\mathrm{I}[W_i = w_m]\right] + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_i \sum_m \widetilde{I}^{rl}(w_m)\widetilde{U}_l(O_i^E)\mathrm{I}[W_i = w_m] + o_P(n^{-1/2}) \\
&= \frac{1}{n}\sum_i \widetilde{I}^{rl}(W_i)\widetilde{U}_l(O_i^E) + o_P(n^{-1/2})
\end{aligned}$$

Plugging this into the original expansion, now yields:

$$\begin{aligned}
\psi_{jk,n} - \psi_{jk,0} &= \frac{1}{n}\sum_i \widetilde{I}^{rl}(W_i)\widetilde{U}_l(O_i^E) + \frac{1}{n}\sum_i \theta_{jk,0}(W_i) - \psi_{jk,0} + o_P(n^{-\frac{1}{2}}) \\
&= \frac{1}{n}\sum_i \left[\widetilde{I}^{rl}(W_i)\widetilde{U}_l(O_i^E) + \theta_{jk,0}(W_i) - \psi_{jk,0}\right] + o_P(n^{-\frac{1}{2}})
\end{aligned}$$

Which establishes asymptotic linearity of $\psi_{jk,n}$ with influence function

$$D^{obs,cand}(P_0^E)(o^E) = \sum_l \widetilde{I}^{rl}(w)\widetilde{U}_l(o^E) + \theta_{jk,0}(w) - \Psi_{jk}(P_0^E)$$

by definition. All likelihood parameter functions involved in $\widetilde{I}$ and $\widetilde{U}$ are determined by P^F through P_0^E. At this point, the influence function derived under the working assumption that the baseline covariates had finite support, is only a candidate influence function for more general covariate distributions. There are three candidate influence functions corresponding to efficient estimation over the three levels of knowledge: none (unconstrained model), independence and known genetic-distribution, all of which have the same form.

3. Candidate is Efficient Influence Function We first establish that the candidate efficient influence functions are in the tangent space of any their corresponsing observed modeIs: $T_P\mathcal{M}^E$, $T_P\mathcal{M}^{E,\perp\!\!\!\perp}$ or $T_P\mathcal{M}^{E,G}$, by establishing that it lies in the image of the push-forward of the corresponding full model tangent space. Over the unconstrained model, the candidate influence function is expressed as a linear combination of scores $\tilde{U}(w) = (\tilde{U}_{00}, \tilde{U}_{10}, \tilde{U}_{G0}, \tilde{U}_{01}, \tilde{U}_{11}, \tilde{U}_{G1})(w)$ and $\theta_{jk}(w) - \Psi_{jk}$.

The function $\theta_{jk}(w) - \Psi_{jk}$ is mapped to itself under f_c^* as it only involves variables that are always observed. We can confirm directly that the following functions

$$\begin{aligned}
\widetilde{U}_{0k}^F(o^F) &= \frac{y - \theta_{0k}(w)}{\theta_{0k}(w)(1-\theta_{0k}(w))} I[X = k] \\
\widetilde{U}_{1k}^F(o^F) &= \frac{y - \theta_{1k}(w)}{\theta_{1k}(w)(1-\theta_{1k}(w))} I[X = k] \\
\widetilde{U}_{Gk}^F(o^F) &= \frac{a - q_G(1;k,w)}{q_G(1;k,w)(1-q_G(1;k,w))} I[X = k]
\end{aligned}$$

for $k \in \{0, 1\}$, belong to $T_P\mathcal{M}^F$ and are mapped onto components of $\tilde{U}(w)$ by the push-forward f_c^*. This first condition can be stated as an assumption that the functions lie in $\mathcal{L}_2^0(P)$, which imposes required regularity on the underlying data generating mechanism. It's clear that they are all of mean 0; the assumption of finite variance effectively limits how wildly the conditional mean outcome and conditional exposure functions, in any of our statistical models, can vary over the covariates.

The complete 6-vector of scores in the full data model will be denoted $\widetilde{U}^F(o^F)$ and is analagous to the vector $\widetilde{U}(o^F)$ in the tangent space of the observed model and similarly $\widetilde{I}_F(w)$ denotes the 'conditional Information matrix' assocaited with a full data probability distribution. Recall the use of r to denote the index within the vector of scores corresponding to the $\theta_{jk}(w)$ parameter. In particular, the candidate influence function for Ψ_{jk} over the unconstrained model is the image of $\tilde{I}^{rk}(w)\widetilde{U}^F(o^F)$ under the score operator. Similarly, the candidate influence function over the model reflecting known genetic exposure distribution, is the image of $\tilde{I}^{rk}_G(w)\widetilde{U}^F_{G,m}(w)$, where $\widetilde{U}^F_{G,m}$ is the 4-subvector of $\widetilde{U}^F_m$ after removing the two scores for the genetic distribution parameters. The candidate influence function over the model reflecting known gene-treatment independence, is the image of $\tilde{I}^{rk}_{\perp\!\!\!\perp}(w)s^F_{\perp\!\!\!\perp,m}(w)$, is the 5-vector of scores where $\widetilde{U}^F_G = \frac{a-q_G(1;k,w)}{q_G(1;k,w)(1-q_G(1;k,w))} = \widetilde{U}^F_{G0} + \widetilde{U}^F_{G1}$ is the pooled score, under the push-forward that replaces the two treatment-specific genetic distribution parameters. Once the candidate is established as a true influence function, efficiency will follow immediately by the above argument.

Formal confirmation that they remain gradients when W has continuous support follows from: identifiability of the parameter on the model space, i.e., $\Psi(P^F) = \Psi^E(f_c(P^F)), \forall P^F \in \mathcal{M}^F$ and the relationship of the score operator and the coarsening function: $\langle D^F, s^F\rangle_{P^F} = \dot{\Psi} s^F = \frac{d}{d\epsilon}\Psi(P^F_\epsilon)\big|_{\epsilon=0} = \frac{d}{d\epsilon}\Psi^E(f_c(P^F_\epsilon))\big|_{\epsilon=0} = \dot{\Psi}^E(f_c^* s^F) = \langle D^E(P^E), f_c^* s^F\rangle_{f_c(P^F)}, \forall s^F \in T_{P^F}\mathcal{M}^F$. The efficient influence function in the full data models can be expressed $D^F_{jk}(P^F) = \widetilde{I}^{rk}_F(w)\widetilde{U}^F_m(o^F) + \theta_{jk}(w) - \Psi_{jk}$. Since $T_{P^F}\mathcal{M}^F_{[W]} \equiv T_{P^F}\mathcal{M}^F_{[W]}$ the two gradients (full-data and observed data) are equal on this component, $\theta_{jk}(w) - \Psi_{jk}$; it suffices to check the components of $\widetilde{U}^F(o^F) \in T_{P^F}\mathcal{M}^F_{[Y,G,X|W]}$, for any covariate value.

$$\langle D^F_{jk}(P^F), \widetilde{U}^F_m\rangle_{P^F} = \langle \sum_l \widetilde{I}^{rl}_F \widetilde{U}^F_l, \widetilde{U}^F_m\rangle_{P^F} = \mathbb{E}_W\left[\sum_l \widetilde{I}^{rl}(W)\widetilde{I}_{lm}(w)\right] = I[m=r]$$

which equals 1 when acting on $\widetilde{U}_r(o^F)$ and 0 otherwise. This follows from the orthogonality of the six conditional scores that span $[Y, G, X|W = w]$. Now in the observed model,

$$\langle D^E_{jk}, f_c^*\widetilde{U}^F_m\rangle_{P^E} = \langle \sum_l \widetilde{I}^{rl}\widetilde{U}_l, \widetilde{U}_m\rangle_{P^E} = \mathbb{E}_W\left[\sum_l \widetilde{I}^{rl}(W)\widetilde{I}_{lm}(W)\right] = I[m=r]$$

As expected. Hence, D^E (which until now was truly $D^{obs,cand}$) is an influence function. The previous fact that it lies in the tangent space establishes it as the efficient influence function: $D^{obs,*}$. Analagous arguments applies to the candidates over the two other experimental data models that correspond to known independence or complete knowledge of the genetice exposure distribution.

C.2 Remainder Under Missingness at Random

Covariate average conditional mean outcome parameters

Here we present direct calculation of the remainder for expansion of $\Psi_{jk} = \mathbb{E}_W\left[\theta_{jk}(W)\right]$ as a function of the unrestricted model $\mathcal{M}^E$. Dependence on baseline covariates, W, is suppressed until the end and notation differentiating P^E from P^F is suppressed entirely. P_ϵ represents the image of a regular path, or fluctuation, through P_0 in $\mathcal{M}^E$. The form of the remainder follows from what is conceptually a sophisticated Taylor approximation:

$$\Psi_{jk}(P_\epsilon) - \Psi_{jk}(P_0) = -\int D_{jk}^{E,*}(P_\epsilon)(o) dP_0(o) + R_{jk}(P_\epsilon, P_0).$$

The remainder can be isolated and calculated as follows:

$$R_{jk}(P_\epsilon, P_0) = \Psi_{jk}(P_\epsilon) - \Psi_{jk}(P_0) + \int D_{jk}^{E,*}(P_\epsilon)(o) dP_0(o)$$

And substituting in the closed form expression for $D_{jk}^{E,*}$ evaluated at P_ϵ:

$$\begin{aligned}
R_{jk}(P_\epsilon, P_0) &= \Psi_{jk}(P_\epsilon) - \Psi_{jk}(P_0) + \int \theta_{jk}^\epsilon(w)\, dP_0(w) - \Psi_{jk}(P_\epsilon) \\
&\quad + \int \frac{\delta}{\pi^\epsilon(y,x)} \frac{I[x=k]}{q_X^\epsilon(k)} \left(\frac{I[g=j]}{q_G^\epsilon(j;k)} - \left(\frac{\theta_{jk}^\epsilon}{\overline{\theta_k^\epsilon}}\right)^y \left(\frac{1-\theta_{jk}^\epsilon}{1-\overline{\theta_k^\epsilon}}\right)^{(1-y)} \right) \left(y - \theta_{jk}^\epsilon\right) dP_0(o) \\
&\quad + \int \frac{I[x=k]}{q_X^\epsilon(k)} \left(\frac{\theta_{jk}^\epsilon}{\overline{\theta_k^\epsilon}}\right)^y \left(\frac{1-\theta_{jk}^\epsilon}{1-\overline{\theta_k^\epsilon}}\right)^{(1-y)} \left(y - \theta_{jk}^\epsilon\right) dP_0(o) \\
&= \int \left\{\theta_{jk}^\epsilon(w) - \theta_{jk}^0(w)\right\} dP_0(w) + \int \frac{I[x=k]}{q_X^\epsilon(k)} \frac{I[g=j]}{q_G^\epsilon(j;k)} \left(y - \theta_{jk}^\epsilon\right) dP_0(o) \\
&\quad + \int \left\{ \frac{\pi^0(y,k)}{\pi^\epsilon(y,k)} - 1 \right\} \frac{I[x=k]}{q_X^\epsilon(k)} \left\{ \frac{I[g=j]}{q_G^\epsilon(j;k)} - \left(\frac{\theta_{jk}^\epsilon}{\overline{\theta_k^\epsilon}}\right)^y \left(\frac{1-\theta_{jk}^\epsilon}{1-\overline{\theta_k^\epsilon}}\right)^{(1-y)} \right\} \left(y - \theta_{jk}^\epsilon\right) dP_0(o) \\
&= \int \left[\theta_{jk}^\epsilon(w) - \theta_{jk}^0(w) + \frac{q_X^0(k) q_G^0(j;k)}{q_X^\epsilon(k) q_G^\epsilon(j;k)} \left\{\theta_{jk}^0(w) - \theta_{jk}^\epsilon(w)\right\} \right] dP_0(w) + \\
&\quad \int \left\{ \frac{\pi^0(y,k)}{\pi^\epsilon(y,k)} - 1 \right\} \frac{q_X^0(k)}{q_X^\epsilon(k)} \left\{ \frac{q_G^0(j;k)}{q_G^\epsilon(j;k)} - 1 \right\} \left(\frac{\theta_{jk}^0}{\overline{\theta_k^0}}\right)^y \left(\frac{1-\theta_{jk}^0}{1-\overline{\theta_k^0}}\right)^{(1-y)} \left(y - \theta_{jk}^\epsilon\right) dP_0(o) + \\
&\quad \int \left\{ \frac{\pi^0(y,k)}{\pi^\epsilon(y,k)} - 1 \right\} \frac{q_X^0(k)}{q_X^\epsilon(k)} \left\{ \left(\frac{\theta_{jk}^0}{\overline{\theta_k^0}} - \frac{\theta_{jk}^\epsilon}{\overline{\theta_k^\epsilon}}\right)^y \left(\frac{1-\theta_{jk}^0}{1-\overline{\theta_k^0}} - \frac{1-\theta_{jk}^\epsilon}{1-\overline{\theta_k^\epsilon}}\right)^{(1-y)} \right\} \left(y - \theta_{jk}^\epsilon\right) dP_0(o)
\end{aligned}$$

When the sampling distribution is known, $\pi^\epsilon = \pi^0$ and this simplifies considerably to:

$$R_{jk}(P_\epsilon, P_0) = \mathbb{E}^0_W \left[\left\{ \frac{q^0_X(k;w) q^0_G(j;k,w)}{q^\epsilon_X(k;w) q^\epsilon_G(j;k,w)} - 1 \right\} \left\{ \theta^0_{jk}(w) - \theta^\epsilon_{jk}(w) \right\} \right]$$

Since the baseline covariates are always observed, expectation with respect to W is the same over either a full data probability distribution or its corresponding experimental data distribution. The remainder calculated above is exactly the same as that when working over full data distributions in $\mathcal{M}^F$.

Over Unrestricted Models In General

A sufficiently smooth function Ψ^F defined over a model $\mathcal{M}^F$ has a corresponding function Ψ^E defined over $\mathcal{M}^E$, where $\Psi^E \{f_C(P)\} = \Psi^F(P)$. The natural correpondence between the models given by the coarsening map allows us to supress the domain of the funciton in the notation, i.e., we just use Ψ. The remainders of the expansions of Ψ with respect to the influence functions unique to estimation over the respective models are:

$$\begin{aligned} R^{F,*}(P^F_\epsilon, P^F_0) &= \Psi(P^F_\epsilon) - \Psi(P^F_0) + \int D^{F,*}(P^F_\epsilon)(o^F) dP^F_0\left(o^F\right) \\ R^{E,*}(P^E_\epsilon, P^E_0) &= \Psi(P^E_\epsilon) - \Psi(P^E_0) + \int D^{E,*}(P^E_\epsilon)(o^E) dP^E_0\left(o^E\right) \end{aligned}$$

which can be combined using the equivalence of $\Psi(P^F) = \Psi(P^E)$ to give the general relationship between remainders of the expansion over $\mathcal{M}^E$ in terms of that over $\mathcal{M}^F$:

$$R^{E,*}(P^E_\epsilon, P^E_0) = R^{F,*}(P^F_\epsilon, P^F_0) - \int D^{F,*}(P^F_\epsilon)(o^F) dP^F_0\left(o^F\right) + \int D^{E,*}(P^E_\epsilon)(o^E) dP^E_0\left(o^E\right)$$

Plugging in the known relationship between the efficient influence function over $\mathcal{M}^F$ and that of $\mathcal{M}^E$, which relies on the MAR assumption, yields:

$$\begin{aligned} R^{E,*}(P^E_\epsilon, P^E_0) &= R^{F,*}(P^F_\epsilon, P^F_0) - \int D^{F,*}(P^F_\epsilon)(o^F) dP^F_0\left(o^F\right) + \int \frac{\delta}{\pi^\epsilon(o^F)} D^{F,*}(P^E_\epsilon)(o) dP^E_0\left(o^E\right) \\ &\quad + \int \left\{ \frac{\delta}{\pi(o^E)} - 1 \right\} \mathbb{E}^F\left[D^{F,*}(P^E_\epsilon)(o^F) \mid o^E \right] dP^E_0\left(o^E\right) \\ &= R^{F,*}(P^F_\epsilon, P^F_0) + \\ &\quad \int \left\{ \frac{\pi^0(o^E)}{\pi^\epsilon(o^E)} - 1 \right\} \left\{ D^{F,*}(P^F_\epsilon)(o^F) - \mathbb{E}^F\left[D^{F,*}(P^E_\epsilon)(o^F) \mid o^E \right] \right\} dP^E_0\left(o^E\right) \end{aligned}$$

It follows immediately, that when the missingness mechanism is known, the corresponding remainders are equal.

Restrictions on the Orthogonal Nuisance Parameter in $\mathcal{M}^F$

We first establish that the remainder of an expansion over the full data model, is independent of choice of influence function for estimation over models $\mathcal{M}^{F,nuis}$ with restrictions to the nuisance parameter orthogonal to Ψ. Any influence function D^F with respect to one of these models can be expressed as $D^{F,*} + s^F$ where s^F belongs to the orthogonal complement of the tangent space (and is an element of the nuisance tangent space).

$$\begin{aligned} R^F(P_\epsilon^F, P_0^F) &= \Psi(P_\epsilon^F) - \Psi(P_0^F) + \int D^F(P_\epsilon^F)(o^F) dP_0^F\left(o^F\right) \\ &= \Psi(P_\epsilon^F) - \Psi(P_0^F) + \int D^{F,*}(P_\epsilon^F)(o^F) dP_0^F\left(o^F\right) + \int s_\epsilon^F(o^F) dP_0^F\left(o^F\right) \\ &= R^{F,*}(P_\epsilon^F, P_0^F) + \int s_\epsilon^F(o^F) dP_0^F\left(o^F\right) \end{aligned}$$

An expansion over a restricted model need only hold for parametric submodels into that restricted model. In particular, all parametric submodels into $\mathcal{M}^{F,nuis}$ have scores orthogonal to s^F and do not vary in the component that governs it's expectation. That is $\mathbb{E}^0\left[s_\epsilon^F(O^F)\right] = \mathbb{E}^\epsilon\left[s_0^F(O^F)\right] = 0$ (which does not assume $s_\epsilon^F = s_0^F$), and hence $R^F(P_\epsilon^F, P_0^F) = R^{F,*}(P_\epsilon^F, P_0^F)$. Combined with the previous result, gives:

$$\begin{aligned} R^E(P_\epsilon^E, P_0^E) &= R^F(P_\epsilon^F, P_0^F) + \\ &\quad \int \left\{\frac{\pi^0(o^E)}{\pi^\epsilon(o^E)} - 1\right\} \left\{D^{F,*}(P_\epsilon^F)(o^F) - \mathbb{E}^F\left[D^{F,*}(P_\epsilon^E)(o^F) \mid o^E\right]\right\} dP_0^E\left(o^E\right) \\ &\quad \int \left\{\frac{\pi^0(o^E)}{\pi^\epsilon(o^E)} - 1\right\} \left\{s^{F,*}(P_\epsilon^F)(o^F) - \mathbb{E}^F\left[s^{F,*}(P_\epsilon^E)(o^F) \mid o^E\right]\right\} dP_0^E\left(o^E\right) \\ &= R^{F,*}(P_\epsilon^F, P_0^F) + \\ &\quad \int \left\{\frac{\pi^0(o^E)}{\pi^\epsilon(o^E)} - 1\right\} \left\{D^{F,*}(P_\epsilon^F)(o^F) - \mathbb{E}^F\left[D^{F,*}(P_\epsilon^E)(o^F) \mid o^E\right]\right\} dP_0^E\left(o^E\right) \\ &\quad \int \left\{\frac{\pi^0(o^E)}{\pi^\epsilon(o^E)} - 1\right\} \left\{s^{F,*}(P_\epsilon^F)(o^F) - \mathbb{E}^F\left[s^{F,*}(P_\epsilon^E)(o^F) \mid o^E\right]\right\} dP_0^E\left(o^E\right) \end{aligned}$$

Again, it follows immediately, that when the missingness mechanism is known, the corresponding remainders are equal. Under restricted models, the remainders may simplify by virtue of being evaluated only at distributions with particular properties. This is seen in the

remainder of the population averaged mean outcome, over distributions with fixed bivariate exposure, the remainder is exactly zero.

C.3 Expansion of Plug-in Estimator

Once the parameter of interest has been linearized:

$$\Psi(P_n) - \Psi(P_0) \;=\; -\int D(P_n)(o)dP_0\,(o) + R(P_n, P_0)$$

where we use P_n to denote an estimate of, perhaps only the required features of, P_0, based on a set of n observations under sampling from P_0, the expansion can be further manipulated into a sum of terms, one of which is the empricial average of a tranformation of the observations. For brevity, we will employ operator notation in which $Pf := \int f(o)dP(o)$. Adding and subtracting $\mathbb{P}_n \{D^*(P_n) - D^*(P_0)\} + P_0 D^*(P_0)$ gives:

$$\begin{aligned}
\Psi(P_n) - \Psi(P_0) &= (\mathbb{P}_n - P_0)\, D^*(P_0) - \mathbb{P}_n D^*(P_n) + (\mathbb{P}_n - P_0)\,\{D^*(P_n) - D^*(P_0)\} + R(P_n, P_0) \\
&= \frac{1}{n}\sum_i D^*(P_0)(O_i) - \frac{1}{n}\sum_i D^*(P_n)(O_i) + \underbrace{(\mathbb{P}_n - P_0)\,(D^*(P_n) - D^*(P_0))}_{EP} + R(P_n, P_0)
\end{aligned}$$

where $\mathbb{P}_n$ denoted the empirical distribution. The fact that influence functions have mean zero, i.e., $D(P) \in \mathcal{L}_2^0(P)$, has been used to simplify the first term in the last line.

The remainder term is established as $o_p(n^{-\frac{1}{2}})$ by direct examination as was discussed in Section C.2. The second to last term is an empirical process term.

Sufficient conditions for this term to be asymptotically neglible are that the functions $D(P_n)$ are $\mathcal{L}_2(P_0)$ consistent and that there is a P_0-Donsker class $\mathcal{F}$ which contains the functions $D(P_n)$ with probability going to 1 as n goes to infinity.

When both of these terms are asymptotically neglible then writing

$$\Psi(P_n) - \Psi(P_0) \;=\; \frac{1}{n}\sum_i D^*(P_0)(O_i) - \frac{1}{n}\sum_i D^*(P_n)(O_i) + o_p(n^{-\frac{1}{2}})$$

is justified. Absorbing the empirical average term into the estimator as:

$$\widehat{\psi}_{OS,n} - \Psi(P_0) = \Psi(P_n) + \frac{1}{n}\sum_i D^*(P_n)(O_i) - \Psi(P_0) = \frac{1}{n}\sum_i D^*(P_0)(O_i) + o_p(n^{-\frac{1}{2}})$$

and establishes the asymptotic linearity of the One-Step estimator $\widehat{\psi}_{OS,n}$ of $\psi_0 := \Psi(P_0)$.

C.4 Details for Simulations and Example

Here we provide the details of data generation that were employed in constructing the example of potential efficiency gains and the simulation study. We present: 1) our use of the cohort from the WHI Hormone Therapy trial to model sample sizes and baseline covariate distributions, 2) the complete data generation process for Example 3.1, and 3) the complete data generation processes used in the simulation study and evaluated in Tables 3.1 and 3.2.

1. WHI Hormone Therapy Trial Cohort Data collected in the hormone therapy trial conducted within the Women's Health Initiative was was analysed in a post-trial GWAS study funded by GARNET. The analysis of interaction between receipt of hormone therapy and genetic SNP 'rs9909279' variant on incident diabetes provides a real study population on which we loosely model our examples and simulations. Complete description of the cohort and primary analyses have been described in detail elsewherecite.

The complete observation analysis for this single interaction consisted of data from 20,938 postmenopausal women of at least 50 years. A subsample of 3,129 women had genetic analysis performed on their blood samples which were collected at baseline. This second phase of data collection was performed on cases and controls in roughly a 1:2 ratio. The overall prevalence of incident diabetes cases in the cohort was about 5.5%.

In addition to age, body mass index (BMI) was also collected at baseline. A two-dimensional histogram (Figure C.1) provides a coarse view of the joint distribution of age and BMI inspires our choice of a truncated bivariate normal distribution as a reasonable distribution of age and BMI. Using the first and second moments estimated from the cohort, age and bmi were simulated according to:

$$(W_{age}, W_{BMI}) \sim NormT\left(\mu = \begin{bmatrix} 63.88 \\ 28.61 \end{bmatrix}, \Sigma = \begin{bmatrix} 7.16^2 & 0.082 \cdot 7.16 \cdot 5.79 \\ 0.082 \cdot 7.16 \cdot 5.79 & 5.79^2 \end{bmatrix}\right)$$

truncated to support $[50, 81] \times [13.83, 69.4]$.

Letting Δ be an indicator of phase-two data collection, ΔG the variant of SNP 'rs9909279'

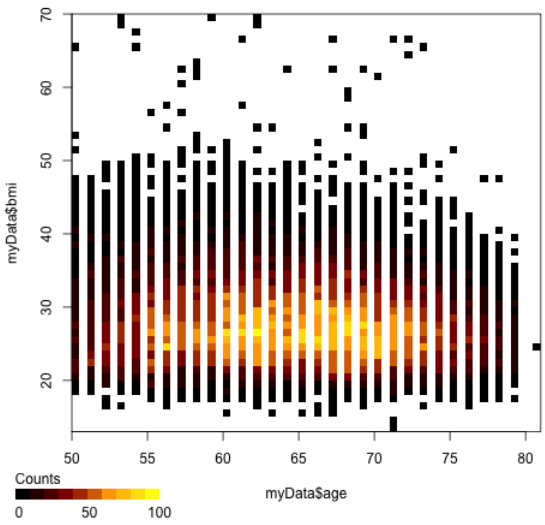


Figure C.1: Histogram representation of joint age-BMI distribution.

with a dominant-recessive coding, and X indicate the receipt of any hormone therapy. Y is the outcome of diabetes incident during the trial. Table C.1 shows the distribution of the analysis cohort across the experimental variables.

Δ	ΔG	X	$\%Y = 1$	N
0	0	0	0.009	8,755
0	0	1	0.007	9,054
1	0	0	0.320	798
1	0	1	0.353	767
1	1	0	0.359	779
1	1	1	0.268	785
	Overall:		0.055	20,938

Table C.1: Sample breakdown in terms of experimental covariates.

Treatment assignment was a simple random $50 - 50$ assignment

$$X \sim Bernoulli(p = 0.5)$$

made independently of all covariates.

2. Efficiency Example The random variable for genetic exposure, $[G \mid W_{age}, W_{bmi}]$, was generated as a Bernoulli variable with probability following a logistic model:

$$\text{logit}\left(\mathbb{E}\left[G \mid W_{age}, W_{bmi}\right]\right) = \alpha_0 + \alpha_{age} W_{age} + \alpha_{BMI} W_{BMI}$$

where $\alpha_{age} = 0.5/30$ and $\alpha_{BMI} = 0.5/55$ were selected to contribute a difference of 0.5 to the linear predictor over the span of variable values. The intercept -1.06, corresponding to a basleine probability of roughly 25%, was adjusted to effectively center the distribution, that is $\alpha_0 = -1.06 - \alpha_{age}\mu_{age} - \alpha_{BMI}\mu_{BMI}$.

The random variable for the incident diabetes outcome, $[Y \mid G, X, W_{age}, W_{bmi}]$, was generated as a Bernoulli variable with probability following a logistic model:

$$\text{logit}\left(\mathbb{E}\left[Y \mid G, X, W_{age}, W_{bmi}\right]\right) = \beta_0 + \beta_{age} W_{age} + \beta_{BMI} W_{BMI} + \beta_X X + \beta_G G + \beta_{GX} GX$$

where $\beta_{age} = 0.5/30$ and $\beta_{BMI} = 0.5/55$ were selected to contribute a difference of 0.5 to the linear predictor over the span of variable values. The treatment effect was $\beta_X = -0.85$ (protective) and the genetic risk effect was $\alpha_G = 0.05$ and the i interaction $\beta_{GX} = 0.2$. The intercept -2.06, corresponding to a basleine probability of roughly 12%, was adjusted to effectively center the covariate distribution, that is $\beta_0 = -2.06 - \beta_{age}\mu_{age} - \beta_{BMI}\mu_{BMI}$.

In the case-control sampling scenarios considered, all cases ($\pi(1, x, w) = 1$), and on average J controls ($\pi(0, x, w) = J * \widehat{\rho}/(1 - \widehat{\rho})$ where $\widehat{\rho}$ is the observed percentage of cases) for each case were selected for phase-two data collection. The simple sampling scenarios considered were designed to yield similar sample sizes as each case-control ratio considered ($\pi(1, x, w) = \pi(0, x, w) = (1 + J)\widehat{\rho}$).

3. Simulation Study The random variable for genetic exposure, $[G \mid W_{age}, W_{bmi}]$, was generated as a Bernoulli variable exactly as for the efficiency gain example. The random variable for the incident diabetes outcome, $[Y \mid G, X, W_{age}, W_{bmi}]$, was generated as a Bernoulli variable with probability following a logistic model:

$$\text{logit}\left(\mathbb{E}\left[Y \mid G, X, W_{age}, W_{bmi}\right]\right) = \beta_0 + \beta_{age} W_{age} + \beta_{BMI} W_{BMI} + \beta_X X + \beta_G G + \beta_{GX} GX$$

similarly to that used in the efficiency gain example, with some different choices of model coefficients. The covariate effects used were $\beta_{age} = 0.2/30$ and $\beta_{BMI} = 0.2/55$, which were selected to contribute a difference of 0.2 to the linear predictor over the span of variable values. The treatment effect was $\beta_X = 0.2$ (harmful) and the genetic risk effect was $\alpha_G = 0.1$ and the interaction β_{GX} varied between $0, 0.5$, and 1. An intercept -3, corresponding to a basleine probability of roughly 4.7%, was adjusted to effectively center the covariate distribution, that is $\beta_0 = -3 - \beta_{age}\mu_{age} - \beta_{BMI}\mu_{BMI}$. The overall prevalence of incident diabetes was roughly $5, 6$, and 7% under the three interaction values considered.

In the case-control sampling scenarios considered, sampling probabilities were set to achieve a phase-two sample consisting of roughly 3.75% of phase-one subjects at a 1-to-1 case:control ratio ($\pi(1, x, w) = min(0.375/\widehat{\rho}, 1)$,$\pi(0, x, w) = min(0.375/(1 - \widehat{\rho}), 1)$.