Methods for Risk Markers that Incorporate Clinical Utility

Anupam Mishra

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Kathleen F. Kerr, Chair

Lurdes Y.T. Inoue, Chair

Robyn L. McClelland

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Methods for Risk Markers that Incorporate Clinical Utility

Anupam Mishra

Co-Chairs of the Supervisory Committee:
Associate Professor Kathleen F. Kerr
Department of Biostatistics

Professor Lurdes Y.T. Inoue
Department of Biostatistics

Risk markers are often used to help make clinical decisions. In this dissertation, we focus on developing statistical methods that account for the utility of a risk marker. We address problems of individualized decision-making, calibration, and combining multiple biomarkers when the ultimate goal is to use the combination for clinical decision-making. We review methods of estimating clinical utility from Bayesian and frequentist standpoints and draw connections between the two frameworks. We additionally consider the appropriateness of each framework to the individual decision-making problem. When existing risk models are applied to new populations, issues of miscalibration can arise. We propose two methods for recalibration that account for the clinical context in which the risk model will be used. Finally, we address the problem of combining risk markers into a single "composite" biomarker. We present a non-parametric method for developing linear combinations of risk markers that maximizes net benefit. We evaluate our methods using simulation studies and apply them to data from prostate cancer, cardiac disease, and diabetes studies.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would first like to thank my advisors Katie Kerr and Lurdes Inoue. I would not be the researcher I am today without their guidance and support. Thank you for your encouragement and patience. The successes I have experienced are a direct result of your amazing mentorship. I am grateful to my committee: Robyn McClelland, Holly Janes, and Aasthaa Bansal. Thank you for your time, insights, and enthusiasm.

I am absolutely indebted to Elizabeth Brown. She has created an amazing space for me to grow as a researcher. The time we have spent together has impacted me profoundly, and I am forever grateful for your mentorship. I am extremely privileged to have been able to work with Jen Balkus. I can only hope to work with such amazing and supportive collaborators in the future. I am thankful to Allison Meisner, Margaret Pepe, Tracey Marsh, Marshall Brown and other members of journal club. Our engaging discourse helped shape this work. I owe many thanks to Gitana Garofalo and the UW Biostatistics department. I also must acknowledge my college mentor Alicia Johnson, who helped me start this journey.

I will never be able to truly express how grateful I am to my family: Atul, Usha, Anjuli, Asha, and Krishna. Their love, support, and humor have never left me. Finally, I am forever thankful for my partner in life Jeremy. I could not have achieved any of this without you next to me.

DEDICATION

To Ram Badan Mishra (Dada),

Ramkumari Mishra (Dadi),

Rama Shankar Pandey (Bauji),

and Rajdevi Pandey (Nani)

Chapter 1

INTRODUCTION

Risk markers can help patients and clinicians make decisions, particularly in settings where an intervention is recommended based on risk. For example, clinicians may rely on risk scores or risk markers to summarize a patient's risk of disease, and help them assess the suitability of preventative medication or further testing. Similarly, risk markers may be incorporated into treatment policies or guidelines. In these settings, there is an inherent consideration of the potential harms and benefits associated with the intervention that, in tandem with the risk marker, shape clinical decision-making. Measures of *clinical utility* summarize the usefulness of a risk model by accounting for the clinical context in which it will be used and its harms and benefits. In this dissertation, we consider methods for risk markers that account for the clinical context in which they are used. Specifically, we consider settings where the risk marker is used to prescribe some intervention.

In Chapter 2 we review the notion of *clinical utility*, a measure of the usefulness of a risk marker while accounting for the harms and benefits associated with correctly or incorrectly prescribing the intervention. We review clinical utility measures that have been applied in both frequentist and Bayesian frameworks. A focus in this dissertation is the *net benefit* metric from the frequentist framework, and *expected utility theory* which comes from the Bayesian framework. We draw connections between the two methods, presenting analytic results that show when these frameworks are in agreement. Additionally, we discuss the appropriateness of each method to the application of decision-making problems for the individual.

Sometimes established risk scores that have been endorsed by bodies of experts are applied to new populations. In these instances, there may be issues of miscalibration, meaning that predicted risks do not accurately capture the event rate of disease. Recalibration methods can be used to correct issues of miscalibration. Existing methods of recalibration do not account for how a risk score will be used in practice. In Chapters 3 and 4 we propose two methods for recalibration of risk scores. The first, a weighted regression strategy, aims to ensure there is good calibration for the most impactful ranges of risk. The second, a constrained optimization approach, aims to produce risk models with high standardized net benefit, thereby producing models that are well-calibrated at the critical risk threshold.

We also consider the problem of combining multiple risk markers into a single "composite" risk marker, that can turned into a simple decision rule. Often, risk markers are combined by maximizing an objective function, such as the logistic likelihood, that does not have immediate clinical relevance. In Chapter 5, we present a method for linearly combining risk markers that maximizes a measure of clinical utility, net benefit. We propose a nonparametric method for finding both a linear combination of risk markers and a risk threshold that together comprise a decision rule.

In Chapter 2 we draw on illustrative examples from a prostate cancer screening and prenatal genetic testing setting. For Chapters 3 and 4 we consider an application in heart disease. The AHA-ACC-ASCVD is highly endorsed risk model used to assess 10-year risk of cardiac events (Goff et al., 2014). The AHA-ACC-ASCVD has been shown to overestimate risk in the ethnically diverse MESA cohort (DeFilippis et al., 2015). We implement the proposed methods to recalibrate the AHA-ACC-ASCVD risk model in the MESA cohort. In Chapter 5 we construct a combination of risk markers to predict diabetes among Pima Indian women. We close with a discussion of our work and future methods.

Chapter 2

COMPARISON OF METHODOLOGY FOR EVALUATING CLINICAL UTILITY AND APPLICATION TO INDIVIDUAL DECISION-MAKING PROBLEMS

2.1 Introduction

Clinical decision-making can be a complex process for patients and health care providers. Given several clinical alternatives with varying degrees of efficacy and harm, an individual's best option can be unclear. Even more, the best clinical option could differ depending on an individual's preferences. For example, older individuals may prefer options that maximize quality of life over a few years, while younger individuals may prefer the option that maximizes quality of life over a longer period of time. Utility evaluation methods provide researchers with a set of tools to assess the value of different clinical strategies. In particular, these methods provide a way of estimating the *clinical utility* of a clinical strategy. Measures of clinical utility quantify the usefulness of a clinical strategy while accounting for the potential harms and benefits associated with that strategy. In an era of "personalized medicine", health care providers appreciate both the heterogeneous nature of many diseases and variation among patient preferences. Understanding how utility evaluation methods can be used to help individuals make health care decisions is more pertinent than ever.

The statistical literature addressing clinical decision-making can be organized into Bayesian and frequentist methods. Under Bayesian decision-theoretic methods, the expected utility framework is used to compare clinical alternatives; under frequentist methods, net benefit (or transformations thereof) is one key measure of clinical utility. We aim to investigate how these approaches to decision-making are related and when they may lead to discrepant

results. We address the appropriateness of applying each framework to individual decision-making problems.

We are interested in addressing these questions in settings where a risk marker is utilized in the decision-making problem. As a heuristic, we consider two motivating examples. First, we consider a decision problem for a man with low-grade prostate cancer. Low-grade prostate cancer is defined as localized prostate cancer with low risk of progression (Klotz, 2005). Often, men can remain in relatively good health when diagnosed with low-grade prostate cancer. Some men with low-grade prostate cancer are put on active surveillance where they are monitored for any sign of progression or until they decide to start treatment. The monitoring of these patients is comprised of several diagnostic screenings a year, typically using the prostate specific antigen marker (PSA) (Klotz, 2005). We will refer to this as PSA-based active surveillance. Prostate cancer is a slow-progressing disease, so some men diagnosed may die from unrelated causes before cancer metastasis (Hsing and Devesa, 2001). In light of this, waiting for disease to present clinical symptoms (termed watchful waiting) is an alternative management strategy. Finally, the most aggressive clinical strategy is to undergo a radical prostatectomy (Heidenreich et al., 2014), which can help reduce the chances of cancer progression. For many men, the optimal decision is not always clear.

In a second example, we consider a woman who is pregnant and has the option to undergo prenatal genetic diagnostic testing. Invasive prenatal diagnostic tests, such as amniocentesis, are used to detect abnormal chromosomal patterns (termed aneuploidy) in the fetus which can manifest as disorders such as Down Syndrome (Wilson, 2000). Amniocentesis has been the standard prenatal genetic diagnostic, but has serious risks including miscarriage. Though all women are recommended to be offered invasive prenatal diagnostic tests, the American College of Medical Genetics (ACMG) additionally recommends that for women who are wary of the risks of amniocentesis, non-invasive maternal serum screening can be used to identify women who are at higher risk of aneuploidy (Driscoll and Gross, 2009). Women with in-

creased risk are recommended to have an invasive prenatal diagnostic testing performed as well as genetic counseling. Some women may not want to receive any genetic information, and only monitor their pregnancy via ultrasound.

In both the prostate cancer and prenatal screening examples, there is no consensus on which clinical management strategy is optimal and both risks and preferences vary between individuals (Mennuti et al., 2013; Dall-Era et al., 2012). We use these examples to illustrate the potential differences in optimal decisions when applying the two decision-making frameworks.

The rest of the chapter is organized as follows. In Section 2.2 we review the frequentist and Bayesian decision-theoretic methods, consider interpretations that will be useful for evaluating an individual's decision-making problem, and draw connections between the two frameworks. In Section 2.3 we argue the appropriateness of each paradigm applied to the individual-level decision making problem, and provide a theoretical result stating under what conditions we expect decision results to be in agreement between the two frameworks. We return to our motivating examples to provide heuristic illustration of these potential differences in decision making in Section 2.4, and close with a discussion.

2.2 Review of Frequentist and Bayesian Clinical Utility Evaluation Methods

2.2.1 Frequentist Decision-Theoretic Methods

We use the following notation for our review of the frequentist literature. We consider the setting where there is a binary outcome of interest, denoted by Y, such as progressing to high-grade prostrate cancer. We refer to those with $Y = 1$ as cases and $Y = 0$ as controls. Let M be some binary risk tool that is useful in predicting, screening for, or diagnosing Y. We refer to those with $M = 1$ as positives and $M = 0$ as negatives. Note that M can be derived from a continuous risk marker, Z, that has been dichotomized by comparing

the observed value to some threshold. For example Z could be PSA, and M could be an indicator of PSA above a certain threshold. The true positive rate and false positive rate of the risk maker M are denoted by TPR and FPR, respectively. Finally, we assume there is some intervention that can be applied in this setting, such as radical prostatectomy in the prostate cancer example, which has some associated harms and benefits.

Net Benefit and Standardized Net Benefit

The first measure of clinical utility, net benefit, is attributed to Peirce (1884). The net benefit (NB) for a clinical strategy is a function of the following population quantities: the expected benefit of the administering intervention to a case (B), the expected harm of administering intervention to a control (C), the prevalence of disease, and the proportion of cases and controls who are prescribed the intervention. The proportion of cases and controls who are prescribed intervention depends on the clinical strategy used. When evaluating the clinical utility of a marker M, the proportion of cases and controls who are prescribed the intervention is captured by the TPR and FPR of the marker, respectively. It is often of interest to also estimate the net benefit of two natural competing clinical strategies (Pauker and Kassirer, 1980; Vickers, 2008). One strategy is to administer intervention to all individuals, sometimes referred to as the "treat-all" strategy. Clearly, under this strategy the proportion of cases and controls who are prescribed the intervention is always 1. The other comparator strategy is to administer intervention to no one, sometimes referred to as the "treat-none" strategy, meaning the proportion of cases and controls who are prescribed the intervention is always 0. The NB of the strategies treat-all, use marker M, and treat-none are

$$NB(\text{treat-all}) = B \times P(Y = 1) - C \times P(Y = 0), \tag{2.1}$$

$$NB(M) = B \times P(Y = 1)TPR - C \times P(Y = 0)FPR, \tag{2.2}$$

$$NB(\text{treat-none}) = 0. \tag{2.3}$$

Under this formulation of NB, inherently the reference strategy is treat-none (Kerr et al.,

2019b). This implies the risk marker is used to identify high-risk individuals who "opt-in" for intervention. For other settings where the risk marker is used to identify low-risk individuals who "opt-out" of intervention, a reformulation of NB is appropriate and the reference strategy is treat-all (Kerr et al., 2019a).

Vickers and Elkin (2006) reformulated net benefit by scaling expressions (2.1) - (2.6) by B, $NB_B = \frac{NB}{B}$. An alternative version of net benefit, standardized net benefit (sNB), also appears in the literature (Pepe et al., 2015), which further scales NB by the probability of the outcome, $sNB = \frac{NB_B}{P(Y=1)}$. Then standardized net benefit (sNB) of the three clinical three strategies "treat-all", use marker M, and "treat-none" are

$$sNB(\text{treat all}) = 1 - \frac{C}{B}P(Y=0), \tag{2.4}$$

$$sNB(M) = TPR - \frac{C}{B}\frac{P(Y=0)}{P(Y=1)}FPR, \tag{2.5}$$

$$sNB(\text{treat none}) = 0. \tag{2.6}$$

Standardized net benefit rescales net benefit to attain a maximum of value of 1, and minimum of $-\frac{C}{B}\frac{P(Y=0)}{P(Y=1)}$.

Risk Threshold

The above expressions of NB and sNB parameterize the harms and benefits of the intervention with two parameters C and B. Pauker and Kassirer (1975) presented a framework for parameterizing the harms and benefits of intervention as a single summary measure R, where

$$R = \frac{C}{C+B}. \tag{2.7}$$

This simplification comes from the notion that prescribing the intervention is only rational when the benefit of the intervention outweighs its harms. We can elucidate this by considering

the prostate cancer example. Let $C > 0$ and $B > 0$ represent the harms and benefits associated with radical prostatectomy. The benefit of radical prostatectomy outweighs the harms when,

$$\underbrace{B \times P(Y=1)}_{\text{benefit of radical prostatectomy}} > \underbrace{C \times P(Y=0)}_{\text{harm of radical prostatectomy}} \tag{2.8}$$

Rearranging these terms yields

$$\begin{aligned} P(Y=1) &> \frac{C}{B} \times P(Y=0), \\ \frac{P(Y=1)}{P(Y=0)} &> \frac{C}{B}, \\ \Rightarrow P(Y=1) &> \frac{C}{C+B}. \end{aligned} \tag{2.9}$$

Therefore, prescribing radical prostatectomy would only be rational if the probability, or risk, of advancing to high-grade cancer is larger than $R = \frac{C}{C+B}$. Note that this relationship between R and harms and benefits does not require making comparisons between the benefit (or harm) of radical prostatectomy and use of the PSA based marker. This is because R characterizes the harms and benefits of the intervention, which we assume to be unchanged in the presence of the marker. We can interpret R as the *risk threshold* for which the benefit of the intervention is equal to the harms of the intervention. An advantage of using this formulation is only a single parameter, R, needs to be obtained, as opposed to harms and benefits separately. Additionally, this formulation can be easily applied to continuous risk scores (measured on the risk scale), as it would only be rational to prescribe the intervention if the risk score is larger than R. For a continuous risk score, Z, net benefit can be expressed

as

$$NB(Z) = P(Y=1)TPR_R - \frac{R}{1-R}P(Y=0)FPR_R, \tag{2.10}$$

$$= P(Y=1)P(Z>R|Y=1) - \frac{R}{1-R}P(Y=0)P(Z>R|Y=0). \tag{2.11}$$

Decision and Relative Utility Curves

Vickers and Elkin (2006) proposed the graphical decision tool, decision curves, to visualize the clinical utility of a continuous risk score Z. The decision curve plots the estimates of NB (according to equation (2.10)), for a range of risk thresholds R. The net benefit curves of comparator clinical strategies are also plotted. Figure 2.1 shows an example of a decision curve. In addition to comparing treat-all and treat-none strategies, decision curves can also be used to assess and compare the clinical utility of competing risk scores (Vickers and Elkin, 2006). Decision curves have been widely adopted in the literature (e.g Rouprêt et al. (2013), Shariat et al. (2011)). Kerr et al. (2016b) discuss the appropriate interpretation of a decision curve. In particular, the authors note two key assumptions of decision curves and net benefit. First, following the definition of NB, the harms and benefits associated with intervention are constant across the population, meaning everyone in the population shares a common R. Second, net benefit, and therefore a decision curve, is a population based metric. The authors caution against using net benefit or decision curves as an individual level decision making tool. We revisit these points in Section 2.3.

Baker et al. (2009) presented a different graphical tool, relative utility curves, for utility evaluation of a continuous risk score. The relative utility, RU, is plotted as a function of the

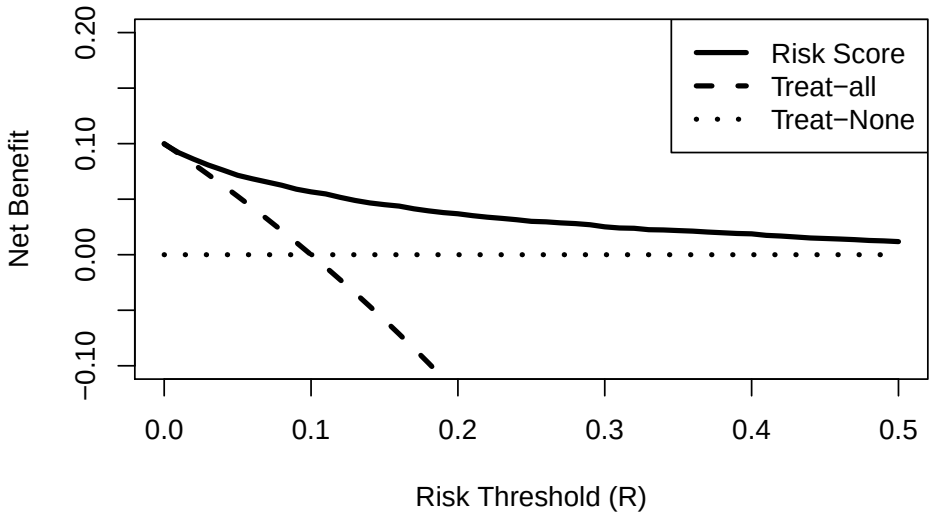


Figure 2.1: Example of a decision curve plot described by Vickers and Elkin (2006). The horizontal axis is the range of risk thresholds R, and the vertical axis shows the NB. In context of the prostate cancer example, the solid line shows the NB of the risk score (PSA based surveillance for all men with low-grade prostate cancer), the dashed shows the NB of the treat-all"strategy (radical prostatectomy for all men with low-grade prostate cancer), and the dotted line shows the NB of the treat-none strategy (watchful waiting for all men with low-grade prostate cancer). In this example, regardless of the risk threshold, the optimal treatment strategy is to use risk score since it has the largest net benefit. For this example, the the risk score is normally distributed conditional on disease status, with mean 1.5 for cases and 0 for controls. The variance for both populations is 1. The prevalence of the outcome is 10%.

.

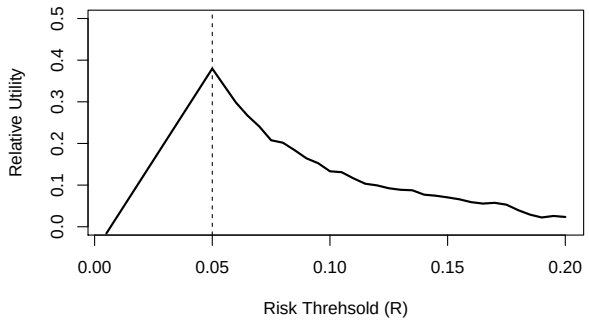


Figure 2.2: Example of a relative utility curve as described by Baker et al. (2009). The solid line gives the relative utility curve of a risk marker that is normally distributed conditional on disease status, with mean 1 for cases and 0 for controls. The variance is set to 1 for both populations. The dotted line indicates the prevalence of disease, 5%.

.

risk threshold R. That is, the following function is plotted,

$$RU = \begin{cases} 1 - FPR - (1 - TPR)\frac{P(Y=1)}{P(Y=0)}\frac{1-R}{R} : & R < P(Y=1) \\ TPR - \frac{P(Y=0)}{P(Y=1)}\frac{R}{1-R}FPR : & R > P(Y=1) \end{cases}$$

A key aspect of relative utility curves is the relevant region, which is represented by the piecewise function. The relevant region compares the clinical strategy of interest to the best uniform policy, which could either be the treat-all or treat-none policy. If the risk threshold is less than the prevalence, the standard of care is to treat-none. Then, it would only be of interest to determine the utility of the risk tool over no-treatment. If the risk threshold is greater than the prevalence the reference policy is treat-all, since it has higher clinical utility than treat-none. Figure 2.2 shows an example of a relative utility curve.

2.2.2 Bayesian Decision-Theoretic Methods

This section focuses on elements of the Bayesian-framework that are useful for clinical decision-making. To help draw comparisons between the frequentist and Bayesian frameworks we will continue to consider the prostate cancer example in the context of the population (e.g. competing clinical strategies radical prostatectomy for all men, watchful waiting for all men, or active surveillance for all men).

Preliminary Definitions: States of Nature, Actions, and Utilities

There are three key components of a Bayesian decision problem: *actions*, *states of nature* and *utilities*. Actions are clinical strategies presented to the decision-maker and are known at the outset of the problem. We denote a single action as a and the set of possible actions as $\mathcal{A}$. In the motivating prostate cancer example, the set of actions are radical prostatectomy for all men, watchful waiting for all men, or active surveillance for all men.

The states of nature are a collection of disjoint states that are unobserved at the outset of the decision problem. We denote a particular state of nature as θ, and set of states of nature Θ. Note that Θ can be a discrete or continuous set. A probability model $\pi(\theta)$, often referred to as a *prior probability*, is assigned to the set of states of nature to express the uncertainty about the true state of nature. The Bayesian framework offers an interpretation of probability called *subjective probability*, which differs from the frequentist definition (Anscombe et al., 1963; De Finetti, 1970). Subjective probability reflects a personal belief of the chance of the event. Subjective probability is most useful when the frequentist definition of risk does not apply (Berger, 2013). We discuss this distinction in more detail in Section 2.3.

The *consequence* of an action depends on the true state of nature. *Utility* is a real-valued representation of the preference for an action a and its consequence when the true state of nature is θ. Thus, the larger the utility the more preferable the consequence. Table 2.1 gives

Table 2.1: Notation for prostate cancer decision problem. We assume the utilities for outcomes occurring under active surveillance are the same as under the other actions.

Parameters	Notation	Description
States of Nature	θ_D	Moderate to high risk cancer (case)
	$\theta_{\bar{D}}$	Low risk cancer (control)
Actions	a_{RP}	Radical prostatectomy
	a_{WW}	Watchful waiting
	a_{AS}	PSA-based active surveillance
Utilities	$u(a_{RP}\|\theta_D)$	Utility of radical prostatectomy for a case
	$u(a_{RP}\|\theta_{\bar{D}})$	Utility of radical prostatectomy for a control
	$u(a_{WW}\|\theta_D)$	Utility of watchful waiting for a case
	$u(a_{WW}\|\theta_{\bar{D}})$	Utility of watchful waiting for a control

a summary of these elements in the context of the prostrate cancer example. For now, we assume that utilities for radical prostatectomy or watchful waiting conducted as a result of active surveillance (i.e., utilities for active surveillance) are the same as utilities for radical prostatectomy or watchful waiting clinical strategies on their own. Utility elicitation, especially in the clinical setting, can be a complex task. The standard gamble approach is a general method for utility elicitation (Farquhar, 1984; Gafni, 1994; Parmigiani and Inoue, 2009). We present a discussion of the standard gamble approach in Appendix A. Characteristics of utilities obtained with the standard gamble include the best outcome having utility 1, the worst outcome having utility 0, and all other utilities falling in between 0 and 1.

Expected Utility Principle

Similar to net benefit, expected utility is a measure of the clinical utility of an action. The

expected utility for a particular action a is

$$U_{\pi}(a) = \int_{\Theta} u(a|\theta)\pi(\theta)d\theta. \tag{2.12}$$

The expected utility principle states that a^* is the optimal (or Bayes) action if

$$a^* = \arg\max_{a} U_{\pi}(a), \tag{2.13}$$

(Von Neumann and Morgenstern, 1945).

Note that the formulation of expected utility given in equation (2.12) does not include data or risk markers. Let Z be a continuous risk marker. Given observed data Z (with density $f(z)$), the model capturing the uncertainty of θ can be updated using Bayes theorem. Then, the uncertainty of θ is captured by the posterior distribution, $\pi_z(\theta) = \pi(\theta|Z = z) = \frac{\pi(\theta)f(z|\theta)}{f(z)}$. The expected utility for action a, given observed data Z is

$$U_{\pi_z}(a) = \int_{\Theta} u(a|\theta)\pi(\theta|Z)d\theta. \tag{2.14}$$

This is referred to as the posterior expected utility of action a. By posterior, we mean, that the expected utility is updated, using the posterior distribution, to include the information captured by Z.

Expression (2.14) gives the expected utility of a given some realization of marker Z. Typically, it is of interest to measure the clinical utility of action a for the entire range of Z. To calculate the expected utility of action a, for all realizations of Z, the conditional expected utility is marginalized over the distribution of Z, $f(z)$, yielding

$$U_{\pi}(a) = \int_{\mathbb{R}} \left[\int_{\Theta} u(a|\theta)\pi(\theta|Z)d\theta \right] f(z)dz.$$

This is referred to as the marginal expected utility. The definition for the Bayes action remains the same, meaning the optimal action a is the one with largest marginal expected utility.

Decision Tree Representation

In settings where the states of nature and utilities are discrete and low-dimensional the decision problem can be visually represented by a decision tree. Solving the decision tree is equivalent to calculating the expected utility for each action and choosing the action with the highest expected utility. Figure 2.3 shows the decision tree representation for the prostate cancer problem where we ignore the PSA-surveillance action. That is, the patient is only considering the actions radical prostatectomy or watchful waiting. We use the notation in Table 2.1. The square on the tree represents the decision node, and from it extends the available actions. The circles nodes on the tree represent random events. We see in this tree the random, or uncertain event, is the underlying states of nature, high-grade cancer status. At the end point of the tree we see the utilities assigned to the consequences resulting from state of nature θ occurring given action a is chosen.

The decision tree is solved by taking a weighted average of utilities over random (circle) nodes (with with respect to the distribution $\pi(\theta)$), and maximizing the averaged utilities over decision (square) nodes. Let $\pi(\theta_D)$ represent the probability of progressing to high-grade prostate cancer, and therefore $1 - \pi(\theta_D)$ the probability of remaining at low-grade prostate cancer. Decision trees are solved backwards, (termed dynamic programming or backwards induction) by averaging over the random nodes furthest to the right, then maximizing over decision nodes moving from right to left (Bellman, 1966). Taking the weighted average at the two random nodes in this problem yields the expected utilities for radical prostatectomy

and watchful waiting,

$$U_\pi(a_{WW}) = u(a_{WW}|\theta_D)\pi(\theta_D) + u(a_{WW}|\theta_{\bar{D}})(1 - \pi(\theta_D)), \tag{2.15}$$

$$U_\pi(a_{RP}) = u(a_{RP}|\theta_D)\pi(\theta_D) + u(a_{RP}|\theta_{\bar{D}})(1 - \pi(\theta_D)), \tag{2.16}$$

respectively. Note that this matches the expression for expected utility as in expression (2.12). The final step is to maximize over the square node, meaning pick the action that has highest expected utility. This is equivalent to finding the Bayes action as in expression (2.13). Of course, the tree presented in Figure 2.3 is a very simple example. In section 2.4 we show and solve a multi-stage decision tree.

2.2.3 Relating net benefit and expected utility

Connections can be drawn between NB and expected utility when considering a clinical scenario such as the prostate cancer example. In this and the following sections we use the notation in Table 2.1 to help clarify these connections, and transition between the Bayesian and frequentist paradigms. First, we consider the connection between harms (C) and benefits (B), and utilities, $u(a|\theta)$. For now, we assume that utilities, like harms and benefits are constant across the population; we revisit this assumption in the next section. We can explicitly relate C, B, and $u(a|\theta)$ by adopting a regret theory approach (Tsalatsanis et al., 2010). Under regret theory utilities can be restated as losses, $u(a|\theta) = -L(a, \theta)$, and a regret loss is defined as

$$RL(a, \theta) = L(a, \theta) - inf_{a \in \mathcal{A}} L(a, \theta), \tag{2.17}$$

where $inf_{a \in \mathcal{A}} L(a, \theta)$ is the smallest negative utility among all actions when the state of nature θ is true. Assume that $u(a_{WW}|\theta_{\bar{D}}) > u(a_{RP}|\theta_{\bar{D}})$ and that $u(a_{RP}|\theta_D) > u(a_{WW}|\theta_D)$. Thus, if the state of nature is $\theta = \theta_{\bar{D}}$ (low-grade disease), the action with lowest loss (i.e.,

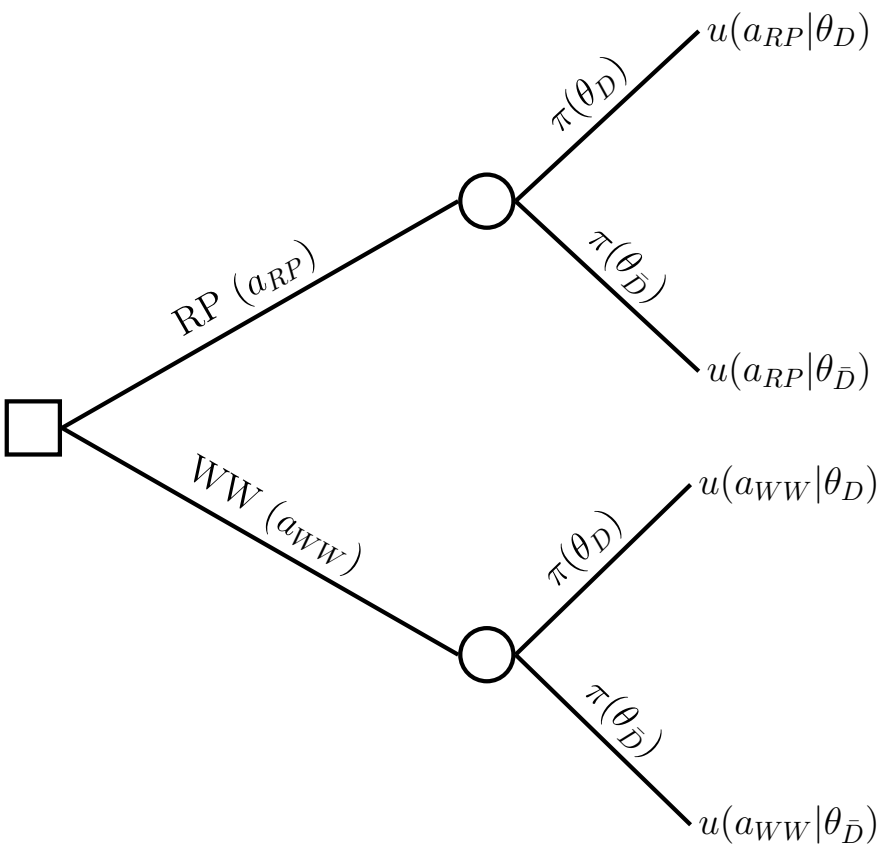


Figure 2.3: Bayesian decision tree for the decision problem summarized in Table 2.1.

highest utility) is watchful waiting. Then, $inf_{a\in\mathcal{A}}L(a,\theta_{\bar{D}}) = -u(a_{WW}|\theta_{\bar{D}})$. Therefore, the regret loss of each action given $\theta_{\bar{D}}$ is

$$RL(a_{WW}|\theta_{\bar{D}}) = 0,$$

$$RL(a_{RP}|\theta_{\bar{D}}) = u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})$$

$RL(a_{RP}|\theta_{\bar{D}})$ can be interpreted as the harm of administering radical prostatectomy to a control, which is the same as C under the net benefit formulation. Likewise, when $\theta = \theta_D$, the action with the lowest loss is radical prostatectomy so that $inf_{a\in\mathcal{A}}L(a,\theta_D) = -u(a_{RP}|\theta_D)$. Then, the regret loss of each action given θ_D is,

$$RL(a_{WW}|\theta_D) = u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D),$$

$$RL(a_{RP}|\theta_D) = 0$$

$RL(a_{WW}|\theta_D)$ can be interpreted as benefit of prescribing intervention to a case, which is the same as B under the net benefit formulation. Under the regret theory formulation, we can say,

$$B = RL(a_{WW}|\theta_D) = u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D),$$
$$C = RL(a_{RP}|\theta_{\bar{D}}) = u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}}).$$

Now we formalize the relationship between net benefit, NB, and expected utility, $U_\pi(a)$. Pauker and Kassirer (1980) presented a decision tree to represent the decision problem of choosing between a diagnostic test (M), a potentially harmful intervention, or no intervention. Figure 2.3 shows this decision tree for the motivating prostate cancer example. This decision tree and the risk threshold result given by Pauker and Kassirer (1980) motivate the decision curve analysis method (Vickers, 2008). Under this decision tree we have the following remark,

Remark 1. *Following the notation in Table 2.1, a_{RP} is the radical prostatectomy for all men action, a_{WW} is the watchful waiting for all men action, and a_{AS} is the PSA-based active surveillance for all men action. We use utilities as noted in Table in Table 2.1. Let $U_\pi(a)$ denote the expected utility of action a given prior distribution $\pi(\theta)$. We denote net benefit of action a as $NB(a)$. Let $B = u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)$ and $C = u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})$. We define*

$$\begin{aligned} NB(a_{RP}) &= P(Y=1) - \frac{C}{B}P(Y=0), \\ NB(a_{AS}) &= P(Y=1)TPR - \frac{C}{B}P(Y=0)FPR, \\ NB(a_{WW}) &= 0. \end{aligned}$$

Under the decision problem we have the following relationship

$$NB(a_{RP}) = \frac{U_\pi(a_{RP}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)} \tag{2.18}$$

$$NB(a_{AS}) = \frac{U_\pi(a_{AS}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)} \tag{2.19}$$

$$NB(a_{WW}) = \frac{U_\pi(a_{WW}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)}. \tag{2.20}$$

That is, the net benefit for action a is the expected utility of action a minus the expected utility of the reference action, a_{WW}, divided by the difference in utilities for radical prostatectomy and watchful waiting among cases (i.e., the benefit of treatment for a case, B).

Remark 1 says that for this decision tree, the net benefit of the action a is equal to the expected utility of that action a, minus the expected utility of the watchful waiting (or "treat-none") action, standardized by B. The proof of this remark is in Appendix A.

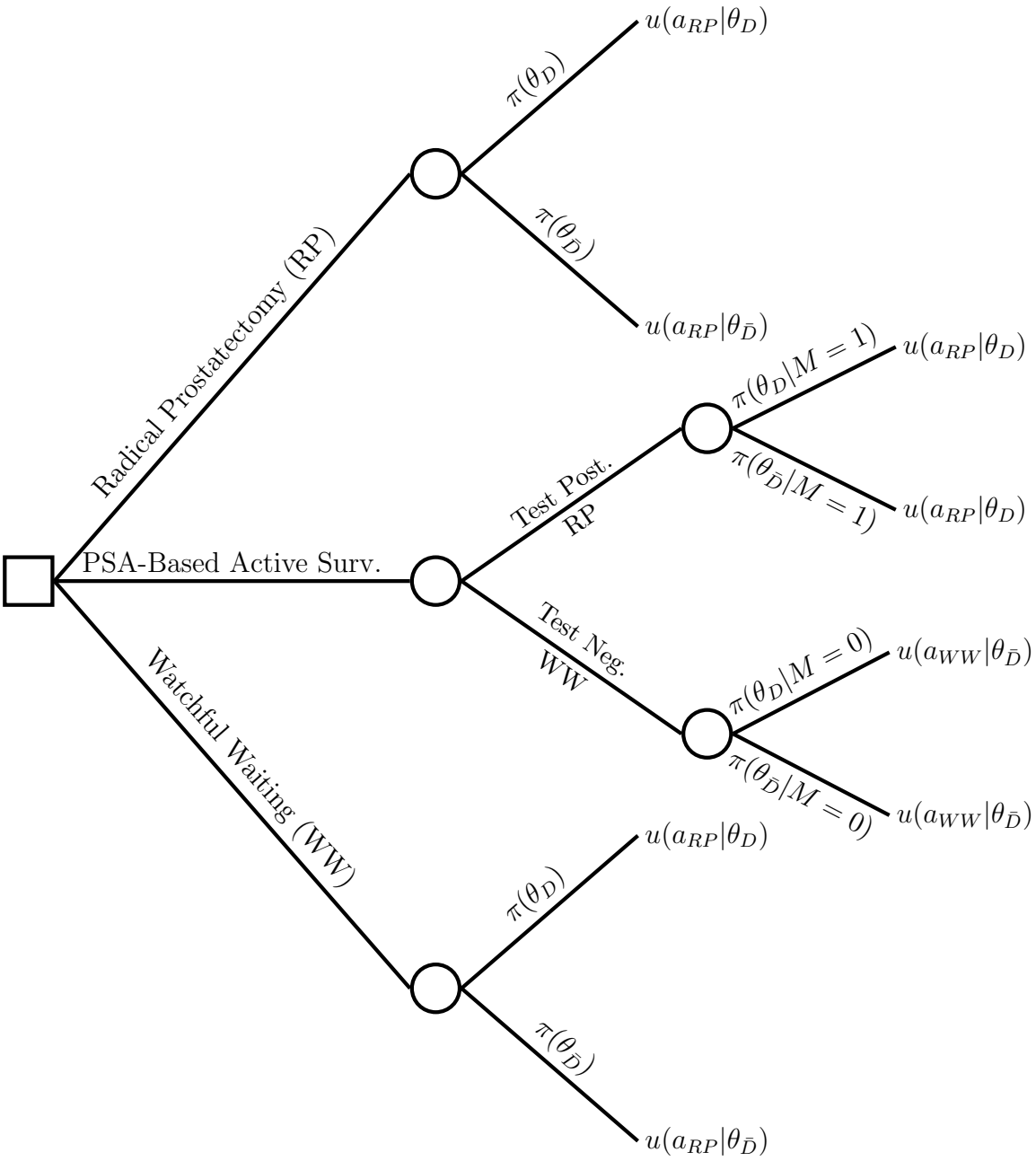


Figure 2.4: Decision problem presented in Pauker and Kassirer (1980) and revisited in Vickers (2008). M represents a binary diagnostic test based on PSA and other surveillance factors. WW stands for the watchful waiting clinical strategy, and RP stands for the radical prostatectomy clinical strategy

2.3 Individualized Decision Making

Now we turn to the individualized decision problem, where we would like to account for an individual's preferences when implementing a decision analysis. There has been discussion on the appropriate application of existing methods to individualized decision problems (Vickers and Elkin, 2006; Kerr et al., 2016b). Importantly, net benefit is a population measure (Kerr et al., 2016b, 2019b), and not naturally interpretable in the individual context. Net benefit is the benefit of intervention applied to the population minus the harms of intervention applied to the population. For risk score Z predicting outcome Y with true positive rate and false positive rate, TPR and FPR, respectively, the harms and benefits of applying the intervention to the population are

$$\text{Benefit of intervention to population} = B \times \pi(\theta_D)TPR, \tag{2.21}$$

$$\text{Harm of intervention to population} = C \times \pi(\theta_{\bar{D}})FPR. \tag{2.22}$$

$$\tag{2.23}$$

To quantify expected harms and benefits of the intervention to the population we need to estimate $\pi(\theta_D)$, which is the probability (or risk) or high-grade prostate cancer. The definition of risk in a frequentist setting is problematic, because it relies on some notion of repetition. For a binary disease outcome in the frequentist paradigm, the risk is the rate of that outcome. This means that the notion of risk is inherently tied to a population. It is unclear how to apply the notion of repeated trials when defining an individual's risk. An individual cannot re-live his or her life many times, so we cannot count the number of times the event occurs. At best we can find a population of similar individuals from whom risk can be estimated. Even in this case, however, the risk we estimate may depend on the characteristics we choose to inform our risk estimate (e.g. age, sex, etc.) (Kerr et al., 2016b).

Additionally, the other parameters needed for net benefit are also population measures.

TPR and FPR are interpreted as operating characteristics of the marker in a population. An assumption of net benefit is that the expected harms (C) and benefits (B) are constant across the population, meaning we have to assume the individual has the same preferences C and B as the population from which we estimate his risk. Though we may expect preferences to tend towards some average, it is a strong assumption that all individuals in the population, even those with same measurable characteristics, have the same preferences. Through a useful example Kerr et al. (2016b) show that decision curves can be misleading if the overall population is a combination of subpopulations with different risk distributions and preferences.

Comparatively, under the Bayesian framework there is flexibility in the interpretation of probabilities and utilities. We can adopt a subjective interpretation of probabilities and utilities. Under the subjective interpretation, the probability of the event is a reflection of one's personal degree of uncertainty. Utilities represent the preferences of the decision-maker, which can be an individual. Marker operating characteristics, TPR and FPR, are used to update the prior probabilities to a posterior probability, which is still interpretable in the individual sense. The use of subjective probability for risk of disease for an individual instead of the frequentist, rate-based probability is a fundamental difference between these two methods. Therefore, we argue that the Bayesian framework is the most appropriate for individualized decision-making.

2.3.1 Decisions trees for the individualized decision problem

The tree presented in Figure 2.4 can be used to consider the individualized decision making problem. Though we are focusing on the individual problem with Bayesian interpretations, the expected utilities derived under this tree can be related to the NB equations (as shown in Remark 1). Because of this, we can think of this tree as the Bayesian formulation of net benefit, for the individual.

Under this tree there are three actions to consider at the outset of the problem: use a watchful waiting approach for disease management, undergo radical prostatectomy, or use PSA-based active surveillance. In this tree, once PSA is measured, the following steps in clinical care are prescriptive, meaning there is no chance for the individual to update his choice. A high PSA measurement ($M = 1$) will result in radical prostatectomy, while a low PSA measurement ($M = 0$) will result in a watchful waiting approach. This may provide a sensible decision rule at the population level. However, after undergoing screening or diagnostic testing, an individual is likely to consult with his healthcare provider to discuss options, then revisit the initial decision problem with updated knowledge. The key point here is that after gaining additional information, such as PSA, there could be an updated decision making step. Figure 2.5 shows a decision tree with added decision nodes, which signify the ability of the decision-maker to re-evaluate the choice between watchful waiting and radical prostatectomy after PSA has been measured. As for the decision tree in Figure 2.4, we assume that utilities remain unchanged conditional on PSA-based surveillance, though we revisit this assumption later.

It is worthwhile to explore when decision making may differ between these two formulations of the problem. That is, when does the added decision node result in differential decision-making? Under either tree the expected utility of the watchful waiting or radical prostatectomy is the same. However, the expected utility of the PSA-based active surveillance action is not necessarily the same. In particular, the expected utility of the actions under the decision tree in Figure 2.4 is

$$\begin{aligned} U_\pi(a_{AS}) = &\left[u(a_{RP}|\theta_D)\pi(\theta_D)TPR + u(a_{RP}|\theta_{\bar{D}})\pi(\theta_{\bar{D}})FPR\right] \\ &+ \left[u(a_{WW}|\theta_D)\pi(\theta_D)(1-TPR) + u(a_{WW}|\theta_{\bar{D}})\pi(\theta_{\bar{D}})(1-FPR)\right]. \end{aligned} \tag{2.24}$$

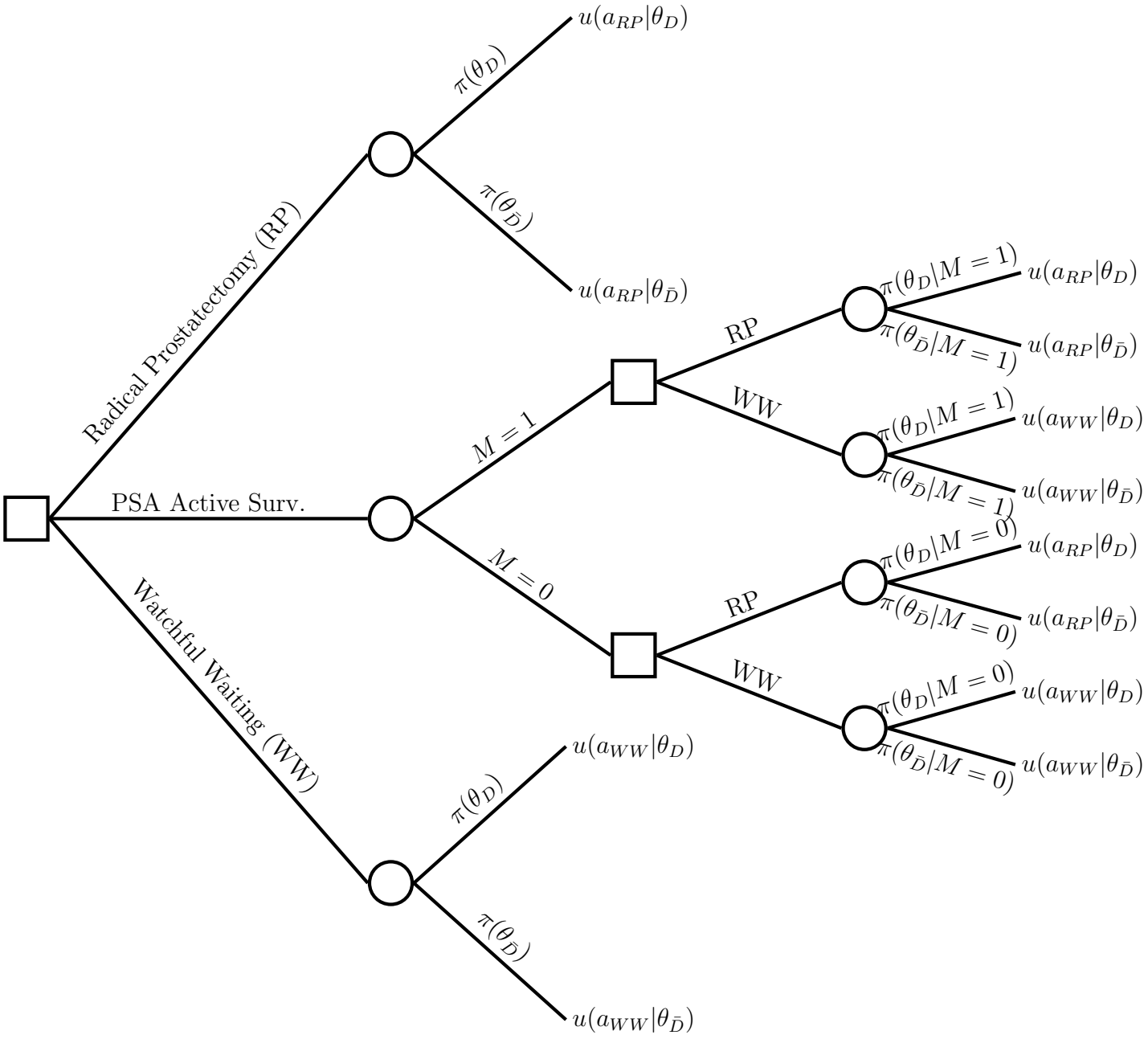


Figure 2.5: Bayesian decision tree for prostate cancer example with added decision node. M is a binary risk marker. RP stands for radical prostatectomy and WW stands for watchful waiting.

The expected utility of active surveillance under the decision tree in Figure 2.5 is

$$\begin{aligned} U_\pi(a_{AS}) = \max\{&u(a_{RP}|\theta_D)\pi(\theta_D)TPR + u(a_{RP}|\theta_{\bar{D}})\pi(\theta_{\bar{D}})FPR, \\ &u(a_{WW}|\theta_D)\pi(\theta_D)TPR + u(a_{WW}|\theta_{\bar{D}})\pi(\theta_{\bar{D}})FPR\} \\ + \max\{&u(a_{WW}|\theta_D)\pi(\theta_D)(1-TPR) + u(a_{WW}|\theta_{\bar{D}})\pi(\theta_{\bar{D}})(1-FPR), \\ &u(a_{RP}|\theta_D)\pi(\theta_D)(1-TPR) + u(a_{RP}|\theta_{\bar{D}})\pi(\theta_{\bar{D}})(1-FPR)\}, \end{aligned} \tag{2.25}$$

The added maximization steps in equation (2.25) reflects the added decision nodes.

Different expected utilities for the active surveillance action may not necessarily result in different optimal decisions. Remark A.3 states when the optimal or Bayes decision will differ between the two trees. Proof is given in Appendix A.

Remark 2. *Under the tree with the added decision node (Figure 2.5), use of the PSA-based risk marker is the optimal decision if*

$$\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{FPR}{TPR} \leq \frac{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)}{u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})} \leq \frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{1-FPR}{1-TPR} \tag{2.26}$$

or

$$\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{1-FPR}{1-TPR} \leq \frac{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)}{u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})} \leq \frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{FPR}{TPR}. \tag{2.27}$$

Under the simpler tree (Figure 2.4), use of the PSA-based risk marker is the optimal decision if

$$\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{FPR}{TPR} \leq \frac{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)}{u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})} \leq \frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{1-FPR}{1-TPR}. \tag{2.28}$$

Figure 2.6 illustrates this result. For fixed prevalence and utilities, a range of operating characteristics of the PSA-based screening tests are plotted. From this plot we can see the range of operating characteristics in which the two decision trees are concordant or discordant in optimal decision. Notably, when we assume utilities to be the same regardless of updating knowledge after the added decision step, the difference in optimal decision only occurs when the risk marker is not clinically appealing. Therefore, settings where the marker is known to have at least nominally acceptable performance, the added decision node may not be needed.

There are instances when this added decision node is important to more accurately capture the decision-making process. If the optimal decision when using the prior risk of disease differs from the optimal decision using posterior risk of disease, then it gives evidence that allowing individuals to update their choice at each stage of gained knowledge is useful. Measuring the marker has some informational value, and it is worth re-evaluating the entire decision problem (e.g. choosing between radical prostatectomy and watchful waiting), using posterior probabilities.

Another way in which the decision trees may produce differential optimal decisions relates to utilities. An important assumption of Remark A.3 is that the utilities for consequences under active surveillance are the same as utilities for consequences under the other actions. It could easily be the case that the utilities for consequences under active surveillance differ. The added costs of active surveillance (in terms of both monetary cost, time, and quality of life) can all impact the preference for consequences. The decision tree in Figure 2.5 allows for flexibility in specifying different utilities for outcomes under active surveillance, *a priori*. We examine this point in the next section.

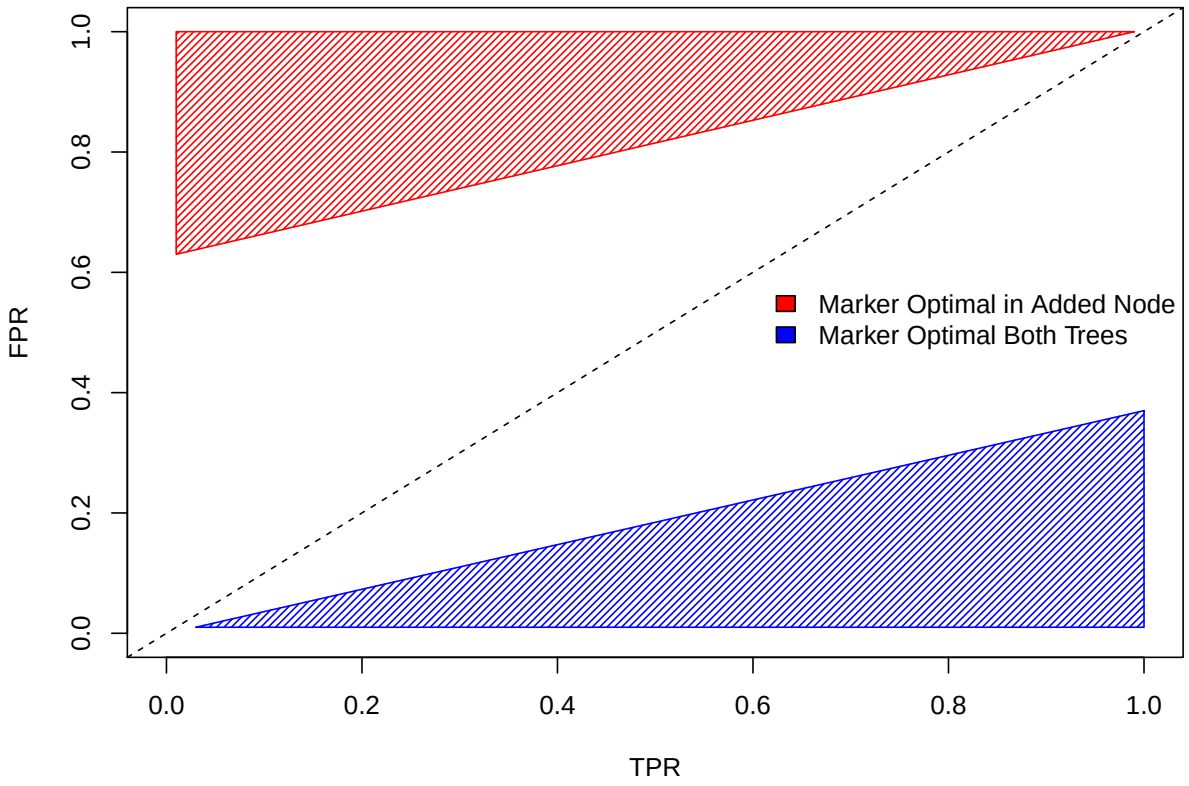


Figure 2.6: Difference in optimal decisions found under the decision trees presented in Figure 2.4 and Figure 2.5 (called added decision node tree) as a function of TPR/FPR. Prevalence is 15% and utilities are UTN = 1, UFN = 0, UTP = 0.85, and UFP = 0.6, which correspond to an optimal risk threshold of 32%. The red region shows when the PSA-based surveillance action is optimal under the added decision node tree but not the simpler tree. The blue region shows when the two trees agree that PSA-based surveillance is the optimal action. The white region indicates settings where both trees agree that the one of the other actions (radical prostatectomy or watchful waiting) is optimal.

2.4 Illustrative Examples for Individual Setting

2.4.1 Prostrate Cancer Surveillance

We revisit the prostate cancer example. Men with low-grade prostate cancer have PSA level less than 10 ng/mL, Gleason score < 6, and disease staging of T2a or lower (D'Amico et al., 1999). Management strategies for preventing men from progressing to high-grade prostate cancer include radical prostatectomy, active surveillance using PSA markers, and watchful waiting (Klotz, 2005). Because of the variability in preferences and disease among men no single strategy can be defined as the optimal strategy across the whole population (Dall-Era et al., 2012). We will use M to denote a diagnostic test result based on PSA surveillance.

We consider a man who has low-grade prostate cancer. His prior estimated risk of progressing to high-grade prostate cancer is slightly higher than the population, $\pi(\theta_D) = 0.3$ (Klotz et al., 2009). Estimates of the sensitivity and specificity of PSA based active surveillance programs are estimated to be between 45-83% and 40-82%, respectively (Iremashvili et al., 2012). For illustrative purposes, we will assume a TPR of 70% and FPR of 50%. Given these probabilities, the marginal probability of having high PSA screening test is $P(M = 1) = 0.56$, and posterior risks are $\pi(\theta_D|M = 1) = 0.38$ and $\pi(\theta_D|M = 0) = 0.20$.

We assume preferences are represented by the following utilities: $u(a_{WW}|\theta_{\bar{D}}) = 1$, $u(a_{WW}|\theta_D) = 0$, $u(a_{RP}|\theta_D) = 0.3$, and $u(a_{RP}|\theta_{\bar{D}}) = 0.5$. The utilities indicate the best outcome for a man is that he manages his disease with watchful waiting when the disease does not progress to high-grade. The worst outcome would be to use the watchful waiting approach but the disease progresses to high-grade. Because of the immediate potential complications of radical prostatectomy (such as surgical risk, impotence, incontinence, etc.) he has less preference for these outcomes. Suppose that a man has different preferences for consequences given PSA based surveillance, and in particular he is more open to radical prostatectomy if PSA test surveillance results are high. These different preferences are reflected in Figures 2.7 and 2.8.

Figure 2.7 shows the decision tree without the added node. Under this set-up, the optimal clinical strategy would be to use the PSA-based surveillance. The decision to have radical prostatectomy is determined by the PSA surveillance results. If PSA values are low, no radical prostatectomy will be performed, and if PSA values are high radical prostatectomy will be performed. Figure 2.8 shows the decision tree with added decision node. Given the prior probabilities, posterior probabilities of risk in light of PSA surveillance, and preferences, the optimal decision for the man under this setting would be to first opt for PSA-based surveillance, then have a radical prostatectomy, regardless of PSA reading. That is, after updating his risk to include information from the PSA-based markers, and accounting for the fact that given PSA preferences utilities are altered, the optimal decision for him is to have radical prostatectomy. This example illustrate that in some settings, re-evaluating the decision problem via the added decision node can result in a optimal decisions that are different to those found using the simpler tree in Figure 2.7.

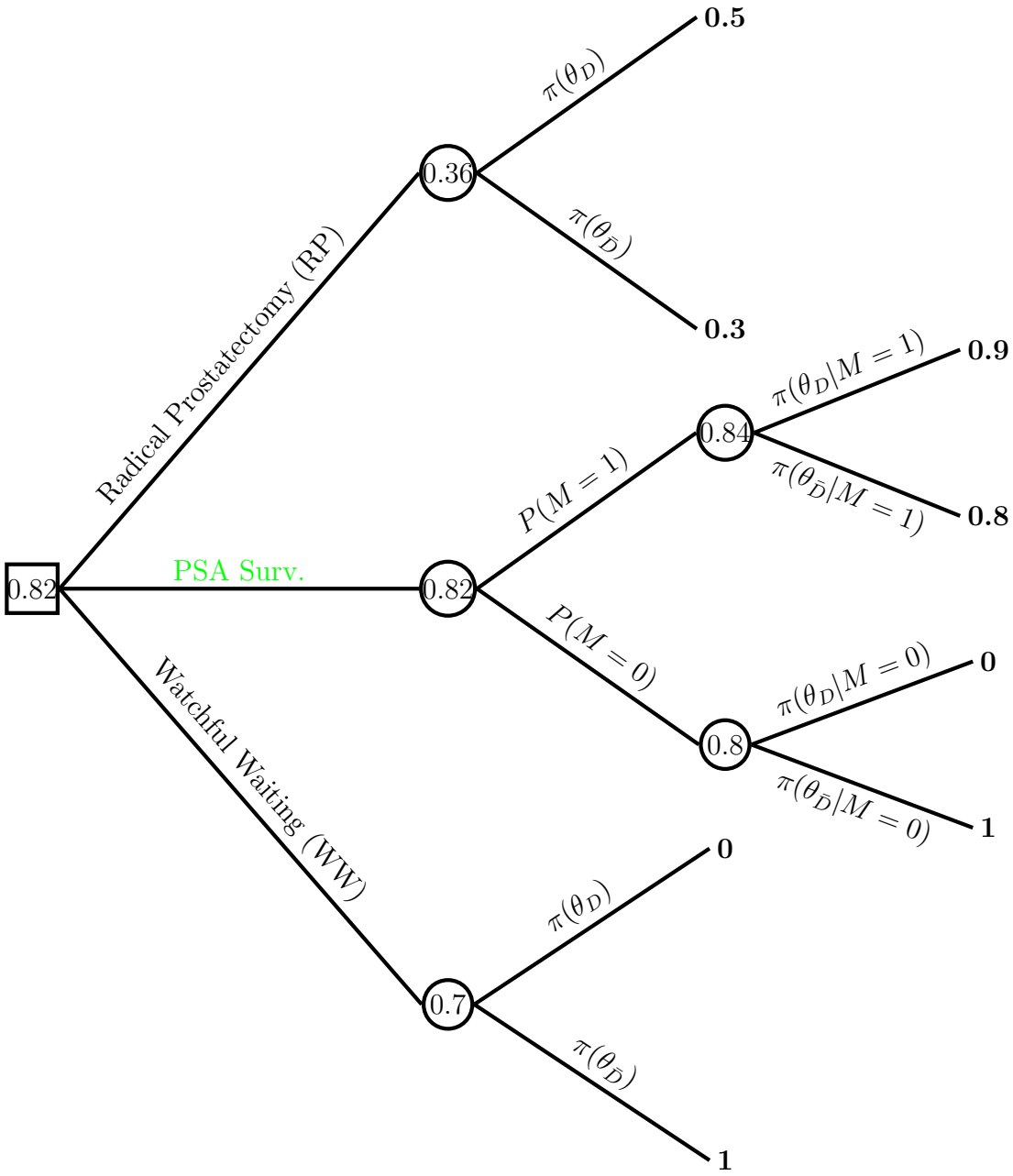


Figure 2.7: Solved Bayesian decision tree for radical prostatectomy example without added decision node. The bold values indicate utilities. The optimal decision is to use the prescriptive PSA surveillance. If $M = 1$ PSA surveillance recommends radical prostatectomy, if $M = 0$ no prostatectomy is performed. Posterior probabilities and probability of positive diagnostic test M are based on an assumed TPR of 70% and FPR of 50% for M, and prior probability $\pi(\theta_D) = 0.3$. The right-most circle show the posterior expected utilities. The left-most circles show the marginal expected utilities. The square node show maximums expected utility.

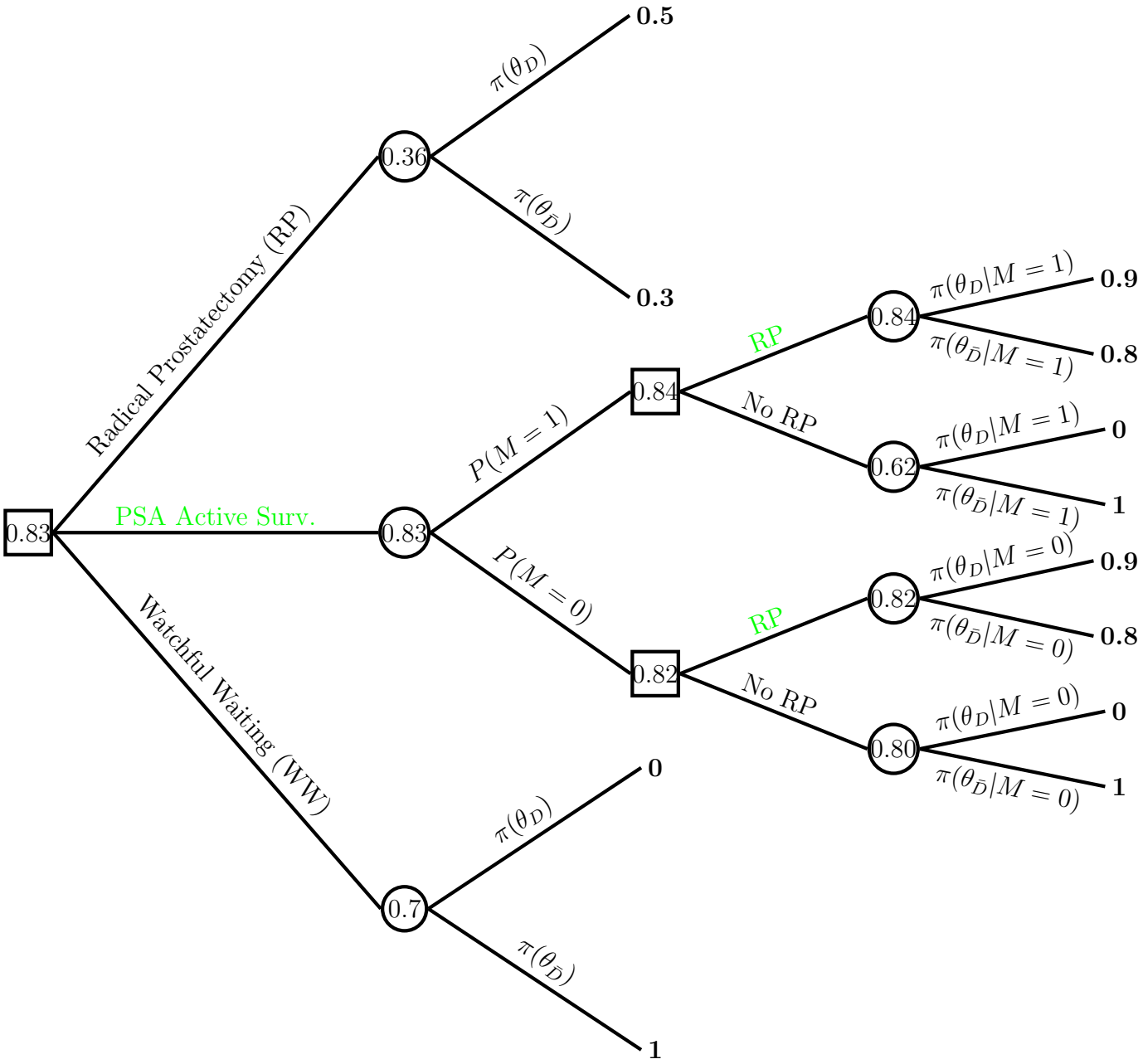


Figure 2.8: Solved Bayesian decision tree for radical prostatectomy example with added decision node.The bold values indicate utilities. The optimal decision is to have radical prostatectomy, but only after obtaining PSA results. Posterior probabilities and probability of positive diagnostic test are based on an assumed TPR of 70% and FPR of 50%, and prior probability $\pi(\theta_D) = 0.3$. Numbers within the right-most circle show the posterior expected utilities. The left-most circles show the marginal expected utilities. The square node show maximums expected utility.

2.4.2 Prenatal Genetic Testing

We consider a second example where an expectant mother is offered invasive prenatal genetic diagnostic testing. Invasive prenatal genetic testing is used to detect chromosomal abnormalities (fetal aneuploidy). Fetal aneuploidy results in birth defects that may affect the quality of life of the child. Amniocentesis is a common invasive prenatal genetic diagnostic test, typically offered to women in their second trimester of pregnancy. During the procedure a sample of amniotic fluid is drawn using a needle. Because of this procedure, there are serious risks of amniocentesis, including miscarriage and preterm labor. In 2009, the American College of Medical Genetics (ACGM) gave guidelines recommending that all women be offered amniocentesis, however acknowledged that many women may not want to undergo screening due to the associated risks (Driscoll and Gross, 2009).

ACMG recommends that certain screening tests, such as maternal serum screening tests, can be used to help identify women who have higher risk for fetal aneuploidy. Maternal serum screening combines blood biomarkers with characteristics of the carrier (e.g. age, race, fetus gestational age, etc.) to estimates risk of fetal aneuploidy. Women with high risk are recommended to follow-up with the amniocentesis diagnostic test. Maternal serum screening is offered for women within 15-20 weeks of pregnancy (Driscoll and Gross, 2008). The true positive rate of maternal serum screening ranges from 75-80% and the false positive rate is around 5% (Driscoll and Gross, 2009).

A downside of maternal serum screening is that it requires more time to process test results. Using maternal serum screening to help decide whether amniocentesis should be conducted, can delay the amniocentesis procedure, which typically needs to be conducted before 20 weeks of pregnancy (Wilson, 2000). Time is an important factor in diagnosis of fetal aneuploidy, as results have a large impact on future actions (such as pregnancy termination or major life changes to accommodate birth disorders). Further, the risk of death from

pregnancy termination dramatically increases for terminations conducted after 18 weeks of pregnancy (Zane et al., 2015). Giving the short and long-term implications of these results, receiving test results quickly can be very important to some women. Singer et al. (1999) noted that attitudes towards genetic testing vary greatly. Some women may be interested in receiving accurate test results quickly, accepting risks of amniocentesis, while other women may not want to know any genetic fetal information and opt to have only standard ultrasounds performed.

Table 2.2 presents the parameters used in this decision problem. Suppose the hypothetical woman has prior predicted risk of fetal birth defects $\pi(\theta_D) = 0.01$ (Hook, 1981). Given her risk factors for chromosomal abnormalities she is willing to consider the potential risks of amniocentesis, to be more certain of her results earlier in her pregnancy. The utilities for outcomes given amniocentesis or ultrasound are $u(a_U|\theta_{\bar{D}}) = 1$, $u(a_U|\theta_D) = 0$, $u(a_A|\theta_D) = 0.95$, and $u(a_A|\theta_{\bar{D}}) = 0.99$. However, her utilities conditional on maternal serum screening decrease significantly, as she may feel there is less time to address any genetic complications. This different utilities given maternal serum screening are depicted in Figures 2.9 and 2.10.

Table 2.2: Notation for fetal aneuploidy decision problem

Parameters	Notation	Description
States of Nature	θ_D	Fetal birth defects (case)
	$\theta_{\bar{D}}$	No defects (control)
Actions	a_A	Amniocentesis
	a_U	Ultrasound screening
	a_M	Maternal serum screening
Utilities	$u(a_A\|\theta_D)$	Utility of amniocentesis for a case
	$u(a_A\|\theta_{\bar{D}})$	Utility of amniocentesis for a control
	$u(a_U\|\theta_D)$	Utility of ultrasound for a case
	$u(a_U\|\theta_{\bar{D}})$	Utility of ultrasound for a control

We assume a $TPR = 0.75$ and $FPR = 0.05$ for maternal serum screening. Given these characteristics, the posterior probabilities are $\pi(\theta_D|M = 1) = 0.132$ and $\pi(\theta_D|M = 0) = 0.003$. The probability of a positive maternal serum screening test is $P(M = 1) = 0.057$. Figure 2.9 shows the decision tree without the added decision node. Under this tree the results of the maternal serum screening test completely determines if amniocentesis is conducted. If the test is positive amniocentesis is preformed, and if the test is negative only ultrasound is performed, without any additional input from the women.Figure 2.10 shows the optimal decision in a tree with added decision nodes.

Under both trees the optimal decision is to forego any genetic screening, or invasive diagnostic testing, and use only ultrasound to monitor pregnancy. This example illustrates that even when utilities conditional on the biomarker based screening strategy differ, it may not always result in differential decision making between the two trees. Therefore, the added node may be unnecessary in this problem. That is, different utilities associated with the screening test do not necessarily imply differential decision making.

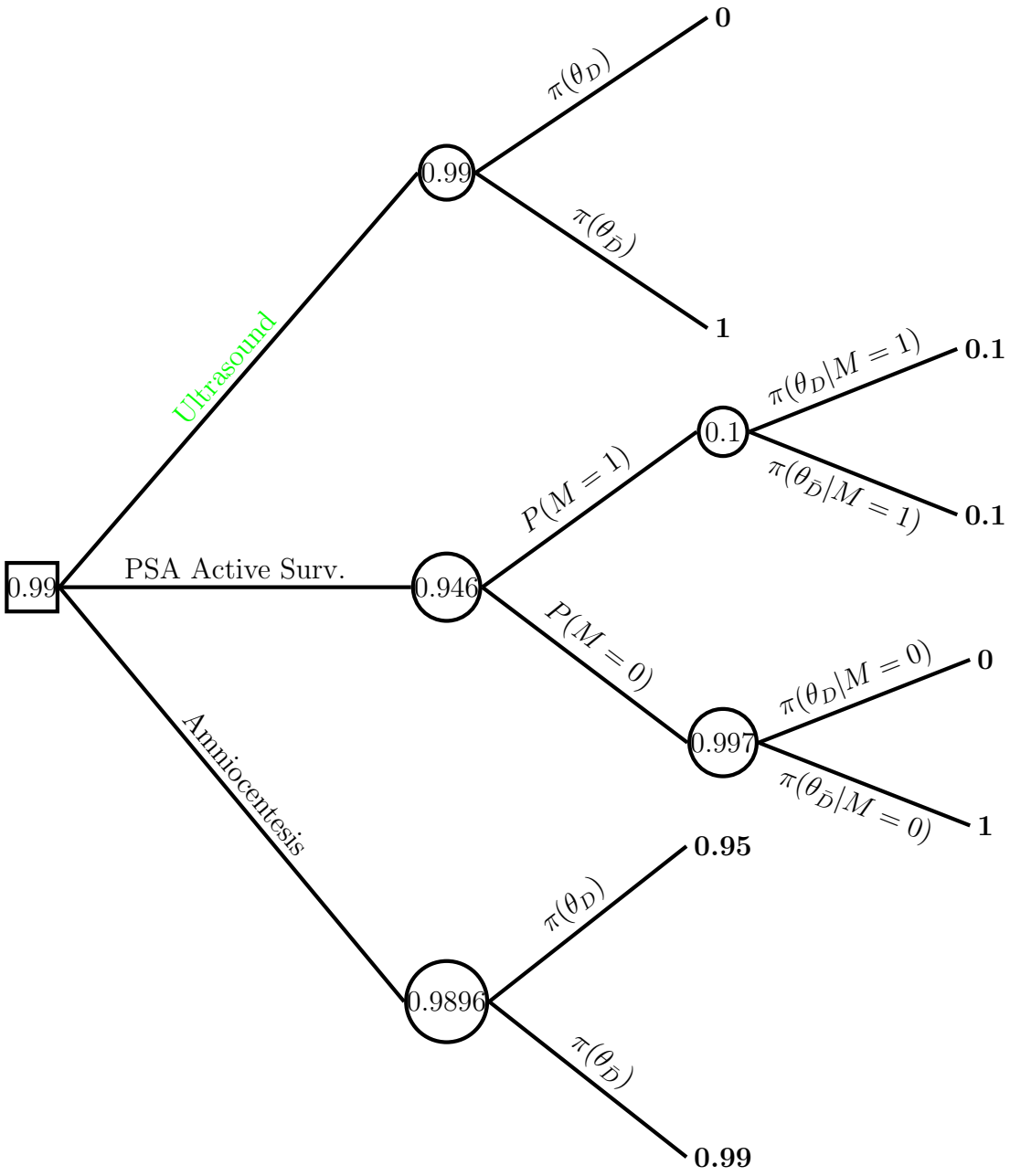


Figure 2.9: Solved Bayesian decision tree for prenatal example without added decision nodes. The bold values indicate utilities. Green text indicates the optimal pathway. The optimal decision is to use ultrasound screening. Posterior probabilities and probability of positive diagnostic test are based on an assumed TPR of 75% and FPR of 5%, and prior probability $\pi(\theta_D) = 0.01$. The right-most circles show posterior expected utilities. The left-most circles show the marginal expected utilities. The square node show maximums expected utility.

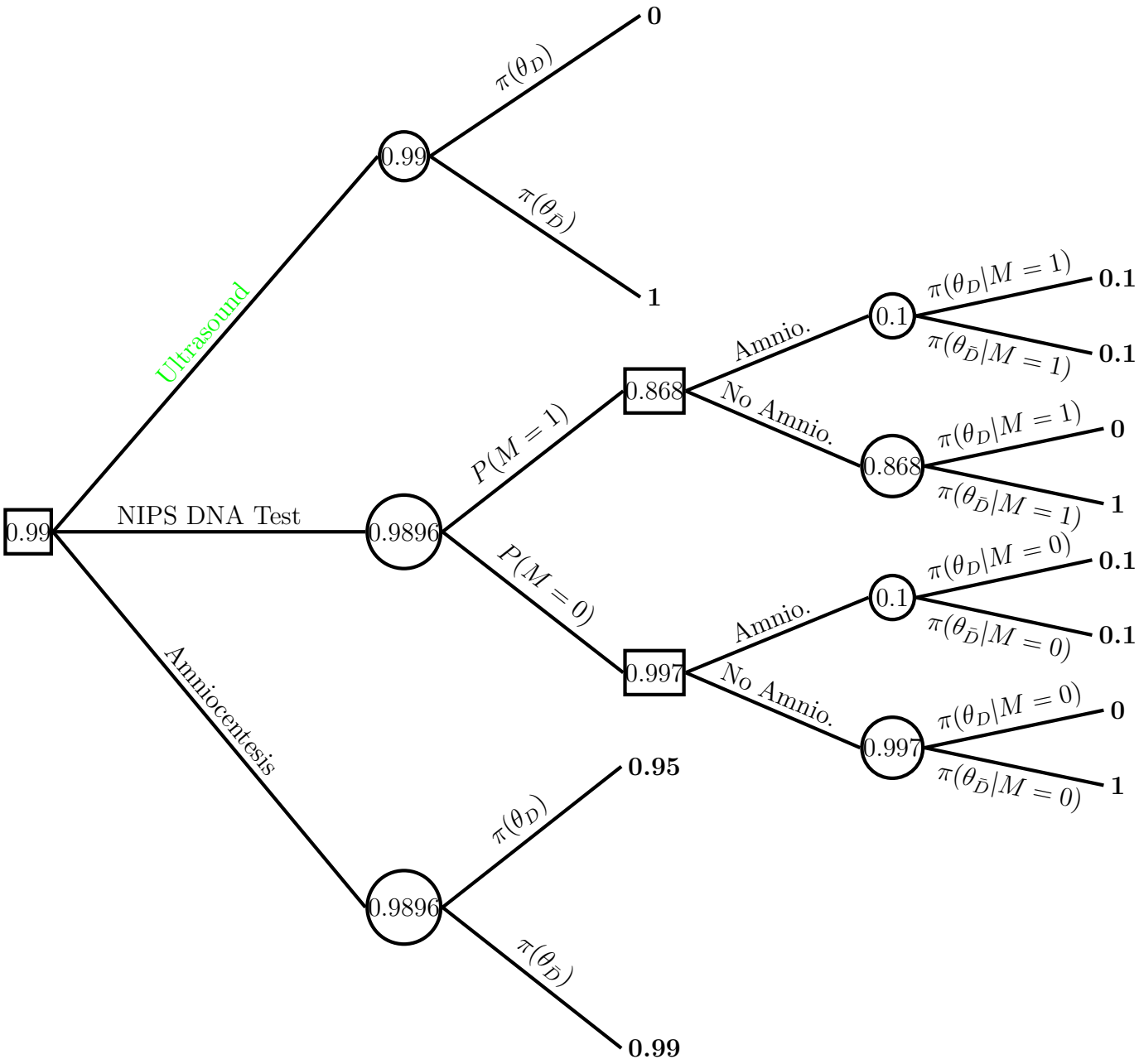


Figure 2.10: Solved Bayesian decision tree for prenatal screening example with added decision node.The bold values indicated utilities. Green text indicates the optimal pathway. The optimal decision is to avoid all screening tests and use ultrasounds. Estimates for posterior probability and probability of positive diagnostic test are based on an assumed TPR of 75% and FPR of 5%, and prior probability $\pi(\theta_D) = 0.01$. The right-most circles show the posterior expected utilities. The left-most circles show the marginal expected utilities. The square node show maximums expected utility.

2.5 Discussion

Decision making methods are powerful tools to help researchers, healthcare providers, and patients understand the different clinical options available to them. Characterizing the utility of clinical strategies or risk markers facilitates decision-making. We compared existing clinical utility frameworks in the Bayesian and frequentist settings, and connected decision trees that have been presented in both the frequentist and Bayesian literatures. We considered their applications to the individual decision-making problem, and argued that a Bayesian framework is needed to consider individual decision-making. Through examples we illustrated how the Bayesian framework can be applied to individual decision problems.

Though we contend that the Bayesian framework is most appropriate for individual decision making, we note there are some trade-offs to be made. The net benefit formulation of clinical utility allows a single summary measure R to represent harms and benefits. Comparatively, the Bayesian framework requires, using regret loss framework, requires specification of at least harms and benefits C and B. Though methods exist for utility evaluation, it is still a difficult and complex task. In many clinical settings, these methods may result in the same conclusion regarding the optimal decision. Regardless of the potential for increased difficulty in estimating clinical utility under the Bayesian method (with or without added decision node), it is still the appropriate method when evaluating individual's clinical options.

We emphasize that one should consider the goal of analysis and the assumptions of a population with homogenous utilities and risk when considering which method to use. Frequentist methods are appropriate for certain population settings. In a population with heterogenous utilities or risks, we may still be interested in determining a population level policy. Some methods to combine individual Bayesian decisions to elicit population level decisions have been proposed in the literature, but there is no clear consensus on the best method and warrants further study.

Chapter 3

WEIGHTED RECALIBRATION FOR IMPROVED CLINICAL UTILITY OF RISK SCORES

3.1 Introduction

A risk model that estimates the probability of a disease or a clinical outcome can help clinicians and patients make health-care decisions. In some clinical contexts, recommendations for intervention are based on comparing estimated risks to a predefined risk threshold. Ideally, this risk threshold is chosen to balance the benefits of the intervention for those who truly have the disease (cases) with the harms of the intervention for those who do not have the disease or will not experience the outcome (controls). An implicit assumption in this decision-making process is that the risk model accurately predicts risk.

The *calibration* of a risk model refers to the agreement between predicted risks and rates of events. When an established risk model is applied to a new population we are particularly concerned that predicted risks may not be well-calibrated. If a model is poorly calibrated and development of a new model is infeasible or undesirable, then it may be prudent to use statistical methods to *recalibrate* the risk model so observed event rates more closely match predicted risks. The goal of recalibration is to correct issues of miscalibration by transforming the risk score via some functional, such as the linear-logistic model.

Existing methods of risk model recalibration do not account for how the risk model will be applied in practice. If medical decisions are based on comparing predicted risks to a risk threshold, good calibration near the risk threshold is more clinically impactful than good calibration elsewhere. This can be illustrated in the context of cardiovascular risk models. The

ACC-AHA-ASCVD risk calculator was developed to estimate the 10-year risk of atherosclerotic cardiovascular disease (ASCVD). A joint panel of experts from the American College of Cardiology (ACC) and American Heart Association (AHA) gave guidelines recommending that individuals with estimated 10-year risk of ASCVD greater than 7.5% receive statin therapy (Goff et al., 2014). In this context, it is most critical to have well-calibrated risks for individuals with 10-year ASCVD risk near 7.5%. Issues of miscalibration are less critical far from the risk threshold 7.5%, because minor miscalibration will not affect decisions regarding intervention.

We propose two methods of recalibration when the intended use of a risk model is to prescribe an intervention using a pre-defined risk threshold. The first approach uses a weighted logistic regression framework to prioritize calibration near the risk threshold. The second approach (Chapter 3) seeks to maximize a measure of clinical utility of the recalibrated risk model, which in turn leads to improved calibration near the risk threshold.

The remaining sections are organized as follows. In Section 2.2 we review relevant literature on calibration, clinical utility, and their intersection. In Section 2.3 we propose a graphical tool for assessing the potential for recalibration to improve the clinical utility of risk scores, and present the weighted recalibration approach. In Section 2.4, we present results from simulation studies, and in Section 2.5 we apply the proposed method to a risk model predicting ASCVD events using data from a prospective cohort study. We close with a discussion of the proposed methods, results, and future work in Section 2.6.

3.2 Background

3.2.1 Calibration

Roughly, calibration refers to the agreement between predicted risks and event rates. Different precise notions of calibration have been defined in the literature and have been ordered

in a hierarchy of weakest to strongest (Van Calster et al., 2016). Calibration-in-the-large is considered the weakest definition of calibration, followed by weak/logistic calibration, moderate calibration, and strong calibration having the most stringent definition.

The weakest form of calibration, calibration-in-the-large, exists when the overall event rate equals the average predicted risk. Weak or logistic calibration, is met if there is no systematic under- or over-fitting in the predicted risks (Cox, 1958). To assess if a risk score is calibrated in the weak sense a calibration slope (α_1) and intercept (α_0) are estimated by fitting the following logistic model to a sample of data with predicted risks r_i and known binary outcome Y_i,

$$\log\left\{\frac{P(Y_i=1)}{P(Y_i=0)}\right\} = \alpha_0 + \alpha_1 \log\left\{\frac{r_i}{1-r_i}\right\}, \tag{3.1}$$

A risk model is well calibrated in the weak sense if $\alpha_0 = 0$ and $\alpha_1 = 1$. Departures from these values indicate over- or under-fitting of the original risk score.

A risk score is calibrated in the moderate sense if the average predicted risks equal the event rates in subgroups that have similar predicted risks. Moderate calibration of a model can be assessed by examining the observed event rate in groups with similar predicted risks visually, via a calibration curve. Hosmer-Lemeshow plots are a conventional way of depicting calibration (Hosmer Jr et al., 2013). In such plots, predicted risks are grouped by deciles and the corresponding average predicted risks and event rates for the decile subgroups are plotted. Alternatively, smoothing functions (such as a LOESS smoother) can be used to generate calibration curves. Figure 3.1 shows an example of a calibration curve. Following Harrell (2015), we use a span of 2/3 and degree of 1 in the smoothing procedure. A calibration curve near the identity line implies that subgroup predicted risks match observed event rates for all subgroups, and suggests that the risk score is well-calibrated in the moderate sense. Strong calibration exists if observed event rate is equal to the predicted risk for subgroups

with the same risk factors (i.e $X = x$).

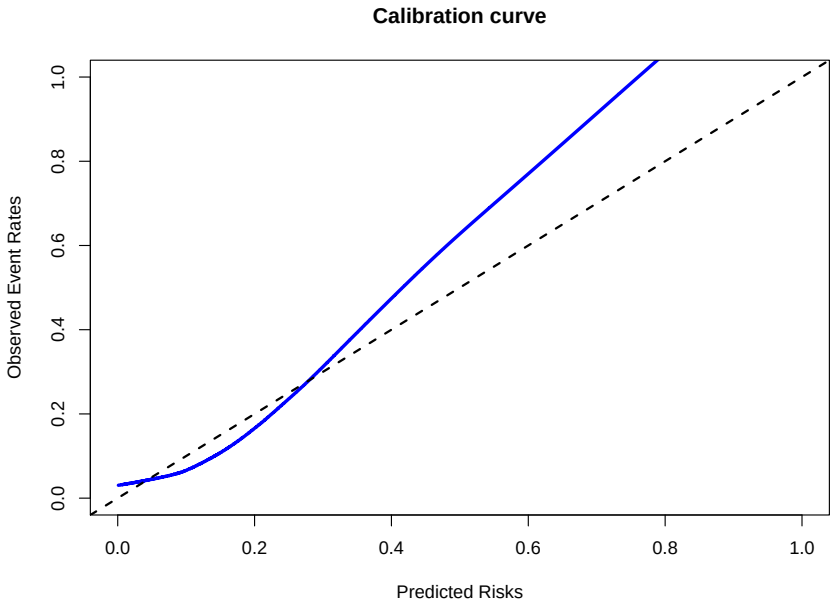


Figure 3.1: Example of calibration curve generated from a LOESS smoother. A calibration curve equal to the identity line (dotted line) suggests perfect calibration in the moderate sense. The blue curve shows a risk score that is calibrated in the moderate sense for predicted risks near 5% and 30%, but not calibrated in the moderate sense elsewhere.

In practice, definitions of weak/logistic calibration and moderate calibration are typically used when assessing the calibration of a risk model (Harrell, 2015). In this dissertation we take calibration to mean calibration in the moderate sense:

Definition 3.2.1 (Calibration)**.** *We say that risk model r is calibrated at t if $P(Y|r = t) = t$, where $t \in [0, 1]$. A risk model is calibrated in the moderate sense if $P(Y|r = t) = t \ \forall \ t$.*

Recalibration refers to methods that attempt to correct miscalibration by transforming the risk score via some functional. Few methods of recalibration exist in the literature. The most prominent method of recalibration, logistic recalibration, was proposed by Cox (1958).

Parameters α_0 and α_1 are estimated by fitting the logistic model given in expression (3.1). α_0 and α_1 are the recalibration intercept and slope, respectively. Recalibrated risk scores are generated by scaling logit-transformed risks, $\text{logit}(r_i)$, by $\hat{\alpha_1}$, shifting by $\hat{\alpha}_0$, then applying the expit function.

Dalton (2013) developed a more flexible framework that models the log observed-to-expected odds ratio, as a complex function of the risk score. Example of functions are splines, indicator functions with pre-specified cut-points, or polynomials. Recalibration parameters $\vec{\alpha}_0$ and $\vec{\alpha}_1$ are estimated from apply the following linear-logistic model to a sample of data,

$$\log\left(\frac{\frac{P(Y_i=1)}{1-P(Y_i=1)}}{\frac{r_i}{1-r_i}}\right) = \vec{\alpha}_0 + \mathbf{H}\vec{\alpha}_1,$$

where $\mathbf{H}$ is a $n \times k$ dimensional basis expansion of the risk score, r . Though Dalton's method can detect and correct more complex issues of miscalibration, the results are largely dependent on the functional of the model, and limited guidance for appropriate function choice is provided. Without careful consideration of choice of function $\mathbf{H}$, recalibrated risks scores can have non-desirable properties, such as non-monotonicity. Additionally, increasing the complexity of the recalibration model may lead to overfitting.

Other methods of recalibration have been developed to tackle calibration issues that result from machine-learning models. Binning as a method of recalibration was proposed by Zadrozny and Elkan (2001). Similar to the Hosmer-Lemeshow plot, predicted risks are grouped into bins, usually deciles. The recalibrated risk assigned to each observation is the event rate within the bin. A limitation of this approach is that results strongly depend on the number and specification of bins. Too few bins can lead to bias and too many bins can lead to overfitting. Isotonic regression (Zadrozny and Elkan, 2002) is a subsequent non-parametric approach which overcomes the limitations of binning. Under isotonic regression the pairwise adjacent violators (PAV) algorithm is used to find a function that minimizes least squares

error, under the constraint that the function must enforce monotonicity. However, isotonic regression has been shown to overfit to training data (Jiang et al., 2011).

Platt et al. (1999) developed Platt Scaling to recalibrate risk models developed from support vector machines (SVM), and Niculescu-Mizil and Caruana (2005) extended methods to handle other machine learning models. Similar to logistic recalibration, recalibration parameters are estimated via maximum likelihood techniques. Unlike logistic regression, the outcome for the model used in Platt Scaling, t_i, is a transformation of the binary outcome Y_i; $t_i = \frac{Y_i+1}{2}$. For some sigmoid function of risk factors, $g(\mathbf{X})$ (such as those obtained from SVM) and transformed outcome t_i, recalibration parameters are estimated by minimizing the negative binomial log-likelihood

$$-\ell(\vec{\alpha}) = -\sum_i t_i log(p_i) + (1 - t_i) log(1 - p_i)$$

with link function

$$p_i = \frac{1}{1 + exp(\alpha_0 + \alpha_1 g(\mathbf{X}))}.$$

Platt scaling has been shown to perform poorly for risks scores that have high density at the extremes, which is sometimes the case in medical risk prediction models (Jiang et al., 2011). Jiang et al. (2011) showed that for some settings binning, isotonic regression, and Platt scaling do not produce well-calibrated risks for risk models obtained from logistic regression.

3.2.2 Standardized Net Benefit

When a model is used to recommend an intervention to high-risk patients, measures of clinical utility summarize the usefulness of the risk model by considering the consequences of prescribing intervention. The standardized net benefit (sNB) of a model is a measure of clinical utility that combines the benefit of the intervention to patients who will have the outcome and the harms of the intervention to patients who will not have the outcome (Vick-

ers et al., 2007). For the formulation of sNB that we will use, a key assumption is that the risk threshold, R, is rational; that is, R reflects the relative harms and benefits associated with the intervention (Pauker and Kassirer, 1980). In particular, the ratio of benefits to harms is equal to $\frac{1-R}{R}$. In the ACC-AHA-ASCVD example, the harm of intervention is the burden of taking statins, such as side effects, for individuals who will not have an ASCVD event within 10 years. The benefit of the intervention is the reduction of ASCVD events for individuals who would have had an ASCVD event within 10 years without statins. The risk threshold of 7.5% implies the benefit of statins to a patient who would otherwise have an ASCVD event is about twelve times greater than the harm of statins to a patient who will not have an ASCVD event in 10 years. If R has been rationally-selected and used as the threshold to prescribe intervention, standardized net benefit summarizes the impact of the treatment policy to the patient population.

Given harms and benefits benefits, C and B respectively, and threshold t, the standardized net benefit, sNB, of risk model r which predicts event Y,

$$sNB_t\left(r\right) = TPR_t(r) - \frac{C}{B}\frac{1-P(Y=1)}{P(Y=1)}FPR_t(r), \tag{3.2}$$

Under the assumption that a risk threshold, R, has been selected such that it represents the harms and benefits (i.e. $\frac{R}{1-R} = \frac{C}{B}$), sNB can be written as

$$sNB_R(r) = TPR_R(r) - \frac{R}{1-R}\frac{1-P(Y=1)}{P(Y=1)}FPR_R(r), \tag{3.3}$$

where TPR_R and FPR_R are the true and false positive rates of risk model r, for risk threshold R.

sNB is a proper scoring rule, meaning that for a risk score r that estimates true risks $\tilde{r}$, $sNB(\tilde{r}) \geq sNB(r)$ (Pepe et al., 2015). We expect that risk scores that do not approximate

the true risks well, such as those that are miscalibrated, have reduced sNB. Van Calster and Vickers (2015) used simulated data to illustrate this. For some examples of a risk model that systematically under- or over-estimates risks, the authors found that using the risk model could have lower sNB compared to a uniform decision rule, such as not administering intervention to anyone in the population. Kerr et al. (2016a) emphasize that miscalibration can harm clinical utility of a risk model even in situations where a miscalibrated risk model out-performs the uniform treatment policies (treat-none and treat-all).

3.2.3 Theoretical Results Relating Calibration and Standardized Net Benefit

To formalize the relationship between calibration and clinical utility, we first review relevant definitions and theoretical results that exist in the literature. We start with results about the ROC curve and monotonicity of a risk score. Pepe (2003) gave the following result,

Lemma 3.2.1. *The slope of the ROC curve at cut-point t is*

$$\frac{\delta ROC(t)}{\delta t} = \frac{f_r(t|Y=1)}{f_r(t|Y=0)},$$

where $f_r(t|Y=1)$ and $f_r(t|Y=0)$ are the probability distribution functions of r among cases and controls, respectively.

Lemma 3.2.1 can be interpreted as the slope of the ROC curve at point t is equal to the likelihood ratio for the risk score r at t. Metz (1978) and Baker et al. (2012) showed that sNB of the risk score is maximized (among all possible FPRs of the risk score) using the cut-point t where the slope of the ROC curve for r equals $\frac{P(Y=0)}{P(Y=1)}\frac{R}{1-R}$. Baker et al. (2012) proves this result by treating the risk score as discrete. Metz (1978) proved this result for a different using measure of clinical utility. We formally state and prove this result in Appendix B for a continuous risk score and sNB measure.

Definition 3.2.2. *(Monotonicity) We say that the risk score, r, is monotonically non-decreasing if*

$$r_i > r_j \Rightarrow P(Y = 1|r_i) \geq P(Y = 1|r_j).$$

The risk score r is monotonic if it is monotonically non-decreasing.

Lemma 3.2.2. *If r is monotonically non-decreasing then r has concave ROC curve.*

The proof of Lemma 3.2.2 is given in Appendix B. Building upon these results, we show that sNB of a risk model, using risk threshold R, has maximum sNB among all recalibrated versions of the risk model (referring to equation 3.1) if and only if the risk model is well calibrated at the risk threshold R. Versions of this result have been shown elsewhere (Baker et al., 2012); however, Theorem 3.2.3 presents the most generalized version of this result that we have found in the literature. Proof of Theorem 3.2.3 is given in Appendix B.

Theorem 3.2.3. *(Calibration of r at R and maximized sNB.) Let r be a monotonically non-decreasing risk score predicting binary outcome Y. Let R be the risk threshold that represents harms and benefits of the intervention associated with the risk model. r has maximum sNB among any risk model of the form*

$$\log\left\{\frac{P(Y=1)}{P(Y=0)}\right\} = \alpha_0 + \alpha_1 \log\left\{\frac{r}{1-r}\right\},$$

if and only if r is calibrated at R.

It follows from Theorem 3.2.3 that improving calibration at the risk threshold will increase clinical utility, and vice versa. In the following section we propose a method that prioritizes calibration near the risk threshold as a means to increase clinical utility of the risk marker.

3.3 Methods

We propose a method for recalibration that targets the most clinically impactful region of risk. We adapt the logistic recalibration framework to prioritize good calibration near the

clinically important risk threshold. For some instances of miscalibration, standard logistic recalibration results in a risk score with high sNB, leaving little room for improvement in sNB for specialized methods. We propose a graphical tool to help assess whether our proposed method has the potential to produce higher sNB than standard logistic recalibration.

3.3.1 A Graphical Tool for Assessing the Impact of Recalibration on sNB

Due to the fact that sNB is a ranked-based measure, changes in sNB will only be observed for recalibration parameters, $\vec{\alpha}$, that shift subjects across the risk threshold, thereby changing TPR and FPR. Alternatively, changes in TPR and FPR (and therefore sNB) can be achieved by fixing $\vec{\alpha}$ and the harm-benefit ratio, and varying the risk threshold. Figures 3.2a - 3.2c show an example of how the same sNB can be achieved via recalibration or by changing the risk threshold. We formalize this relationship with Remark 3.

Remark 3. *Given a risk score, r, harm-benefit ratio $\frac{C}{B}$, and threshold t, there exists (possibly non-unique) recalibration parameters α_0 and α_1, and some other decision threshold t^* such that*

$$sNB_t\left(expit(\alpha_0 + \alpha_1 Z)\right) = sNB_{t^*}\left(r\right), \tag{3.4}$$

where $Z = logit(r)$.

Varying the risk threshold between 0 and 1 yields all the different estimates of sNB that can be achieved from monotonically shifting the risk distribution via logistic recalibration and using a fixed risk threshold. Specifically, for fixed r, and harms to benefit ratio, $\frac{C}{B}$, and $t \in [0, 1]$ we plot an estimate of

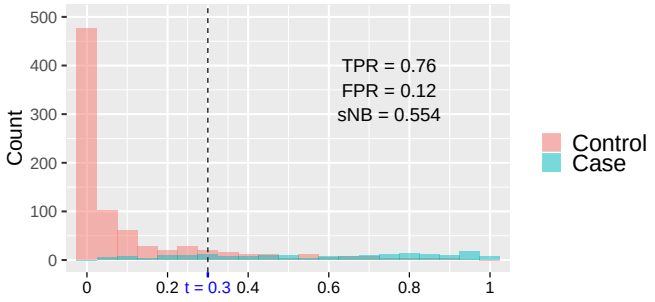
$$sNB_t(r) = TPR_t(r) - \frac{C}{B}\frac{1 - P(Y = 1)}{P(Y = 1)}FPR_t(r). \tag{3.5}$$

Figure 3.3 shows an example of a plot of the potential sNB under recalibration. The

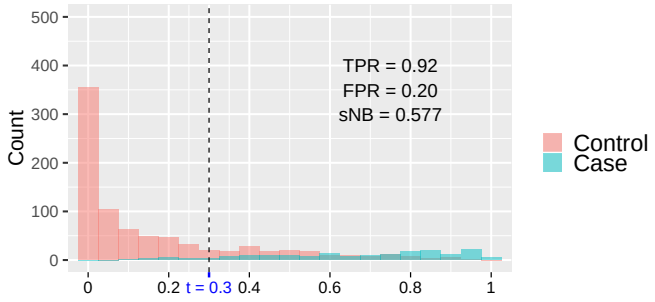
horizontal axis gives all possible thresholds t. The vertical axis gives the estimate of sNB given the threshold t, and fixed harm-benefit ratio. The maximum of the curve estimates the maximum sNB that can be achieved under logistic recalibration of the risk score. The sNB of the original risk score and the standard logistic recalibrated risk score are noted on the curve, which can be compared to the maximum. If the estimated sNB of the original risk score is not near the maximum of the curve, then there is potentially some recalibration parameters $\{\alpha_0, \alpha_1\} \neq \{0, 1\}$ that can increase clinical utility. If the estimated sNB of the standard logistic recalibrated risk score is not near the maximum of the curve, other methods of recalibration may produce a risk model with higher sNB than the risk score recalibrated with standard logistic recalibration.

In light of sampling variability, it may be unclear whether the estimated sNB for a risk score is "close" to the maximum possible value. As a helpful guide, standard error bars around the maximum are included on the curve. Following Friedman et al. (2001) we suggest using a "one-standard error" rule to decide if sNB of a risk score is near the maximum. The left panel of Figure 3.3 shows an example of the sNB curve where the sNB under standard logistic recalibration is outside the one-standard error bandwidth. This suggests there may be different recalibration parameters, $\vec{\alpha}$, that could produce a risk score with higher sNB. In particular, the proposed approach, weighted logistic recalibration, could produce a risk score with higher sNB. The right panel of Figure 3.3 shows an example where the sNB under standard logistic recalibration is near the maximum of the curve, and in particular within one standard error of the maximum. For this example, we might judge that it is not worthwhile to pursue alternatives to standard logistic recalibration.

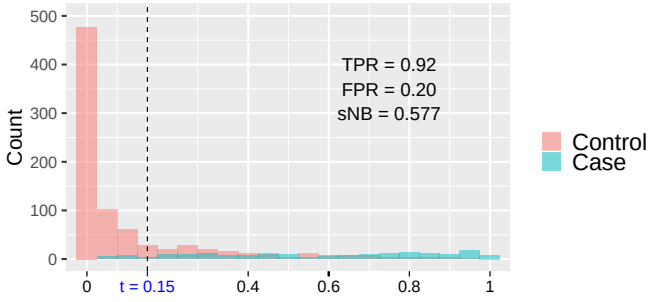
Proposition 1 gives the analytic variance formula for sNB, assuming a fixed risk model, intervention threshold t, and harm-benefit ratio $\frac{C}{B} = \frac{R}{1-R}$. A proof of Proposition 1 is presented in Appendix B.



(a) Original Risk Score



(b) Recalibrated Risk Score: $\alpha_0 = 0.5$, $\alpha_1 = 0.8$



(c) Original Risk Score

Figure 3.2: Figures 3.2a and 3.2b show the distribution of risk scores among those with outcome (cases) and those without outcome (controls). The dotted vertical line indicates the risk threshold. We set $\frac{C}{B} = \frac{1-0.3}{0.3}$. The original risk score is shown in Figure 3.2a and a risk score after recalibration is applied is shown in Figure 3.2b. Using a risk threshold of $t = 0.3$ yields $\hat{sNB} = 0.554$ under the original risk score and $\hat{sNB} = 0.577$ under the recalibrated risk score. Figure 3.2c shows that if a different threshold, $t = 0.15$, is used with the original risk model, the same sNB can be obtained as that estimated after recalibrating the risk model.

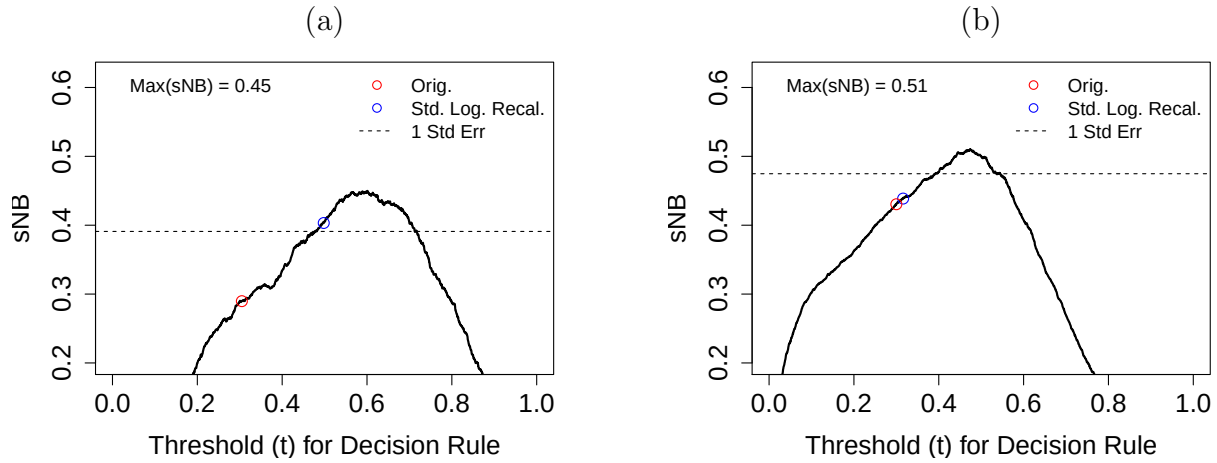


Figure 3.3: Potential sNB under recalibration. The dotted line shows a one-standard error lower bound of the estimated maximum possible sNB. In both (a) and (b), the estimated sNB for the original risk model is more than one standard error lower than the estimated maximum possible sNB, indicating that a recalibrated risk score could yield higher sNB. In (a) the estimated sNB for the risk model after standard logistic recalibration is near the maximum value. Alternative methods of recalibration may not be worth pursuing in this setting. In (b) the risk score produced by standard logistic recalibration yields estimated sNB more than one standard error lower than the estimated maximum possible sNB, suggesting alternative recalibration methods may be useful

Proposition 1. *Let r be a risk model predicting binary random variable Y. Assume that the harm to benefit ratio of the intervention associated with the outcome is $\frac{C}{B}$. For a random sample of i individuals from the relevant population, the estimated standardized net benefit of the policy that treats individuals with $r > t$ is*

$$\widehat{sNB}_t(r) = \widehat{TPR}_t(r) - \frac{C}{B}\frac{\hat{P}(Y=0)}{\hat{P}(Y=1)}\widehat{FPR}_t(r) \tag{3.6}$$

Define the sample proportions:

$$\hat{p}_{11} = \frac{1}{n}\mathbf{1}[r \geq t, Y = 1] \qquad \hat{p}_{01} = \frac{1}{n}\mathbf{1}[r < t, Y = 1],$$
$$\hat{p}_{10} = \frac{1}{n}\mathbf{1}[r \geq t, Y = 0] \qquad \hat{p}_{00} = \frac{1}{n}\mathbf{1}[r < t, Y = 0].$$

The estimator in 3.6 can be expressed as a function of $\vec{\hat{p}} = \{\hat{p}_{11}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{00}\}$*, when* $\hat{p}_{11}+\hat{p}_{01} > 0$ *and* $\hat{p}_{10} + \hat{p}_{00} > 0$ *(since we can never divide by 0),*

$$\widehat{sNB}(\vec{\hat{p}}) = \frac{\hat{p}_{11}}{\hat{p}_{11} + \hat{p}_{01}} - \frac{C}{B}\left(\frac{\hat{p}_{10} + \hat{p}_{00}}{\hat{p}_{11} + \hat{p}_{01}}\right)\frac{\hat{p}_{10}}{\hat{p}_{10} + \hat{p}_{00}}.$$

Then, the variance of $\widehat{sNB}(\vec{\hat{p}})$ *given harm to benefit ratio,* $\frac{C}{B}$*, is*

$$V\left[\widehat{sNB}\left(\vec{\hat{p}}\right)\right] = \left(\frac{1}{p_{11} + p_{01}}\right)^2 \left[\left(\frac{C}{B}\right)^2 p_{10} + \frac{p_{11}\left(p_{01} + \frac{C}{B}p_{10}\right)^2}{(p_{11} + p_{01})^2} + \frac{p_{01}(\frac{C}{B}p_{10} - p_{11})^2}{(p_{11} + p_{01})^2}\right], \quad (3.7)$$

where

$$p_{11} = P(r \geq t, Y = 1) \qquad p_{01} = P(r < t, Y = 1),$$
$$p_{10} = P(r \geq t, Y = 0) \qquad p_{00} = P(r < t, Y = 0),$$

and $p_{11} + p_{01} > 0$ *and* $p_{10} + p_{00} > 0$*.*

3.3.2 Weighted Logistic Recalibration

Let Y_i, $i = 1, \ldots, n$, indicate the event of interest and let r_i, $i = 1, \ldots, n$, be the set of estimated risks obtained from applying the risk model to a sample from the population of interest. Let $Z_i = \text{logit}(r_i)$, be the logit transformed risk score. As described above, R is the risk threshold.

We propose the weighted recalibration intercept α_0^* and slope α_1^* are estimated by maximizing the weighted likelihood function

$$L(\alpha_0^*, \alpha_1^* | Y_i, z_i) = \prod_{i=1}^{n} \left(\frac{e^{\alpha_0^* + \alpha_1^* z_i}}{1 + e^{\alpha_0^* + \alpha_1^* z_i}}\right)^{w_i Y_i} \left(\frac{1}{1 + e^{\alpha_0^* + \alpha_1^* z_i}}\right)^{w_i(1-Y_i)}. \quad (3.8)$$

We suggest two functions to produce the weights w_i. The first weight function has the form

of an exponential decay weight,

$$w_{1,i} = \exp\left(-\frac{(o(r_i) - R)^2}{\lambda}\right), \tag{3.9}$$

where $o(r_i)$ is a smoothed observed event rate, obtained via LOESS regression of Y_i on the risk score r_i, using a span of 2/3 and degree of 1 in the smoothing procedure. Notation reflects the dependence of observed event rate on the risk model r_i. $o(r_i)$ are presented on the vertical axis of the calibration plot. Under this weighting scheme, observations with event rates at or near the risk threshold have the largest contribution to the likelihood. Since the clinical decisions are not likely to change for observations with event rates far from the risk threshold, the method down-weights their contribution to the likelihood. λ is a tuning parameter and controls the degree of differential weighting. As λ increases all weights tend to 1, and the weighted recalibration method approaches standard logistic recalibration.

The weight function w_1 can be generalized to prioritize calibration over a range of risks instead of a single risk threshold. In the motivating example, additional guidelines and current practices in cardiology indicate 5%-10% is an interval of critical predicted risks that may impact clinical decisions. To accommodate this, we generalize w_1 so that observations outside an interval around R can be further down-weighted. Given an interval $[R_l, R_u]$ (with $R_l < R < R_u$) we propose the weight

$$w_{2,i} = \begin{cases} \exp\left(-\frac{(o(r_i)-R)^2}{\lambda}\right) & : o(r_i) \in [R_l, R_u] \\ \delta & : o(r_i) \notin [R_l, R_u] \end{cases}. \tag{3.10}$$

The parameter δ prescribes how much weight is assigned to observations outside the critical risk interval. δ is bounded below by 0 and bounded above by the infimum of the weights within the interval $[R_l, R_u]$. Weight w_1 (3.9) is a special case of the weight w_2 (3.10) with $[R_l, R_u] = [0, 1]$. For settings where good calibration is important for the interval $[R_l, R_u]$, λ can be fixed large enough that weights within interval are all close to 1 (e.g. $\lambda \geq 10$), and

only specification of δ is needed. For settings where good calibration at the risk threshold R is most important, weight w_1 can be used and only specification of λ is needed. For w_1, small values of λ down-weight observations far from R. Figure 1 gives examples of the shape of the two weight functions.

As more down-weighting is applied, more discounting of observations is applied in the recalibration procedure. In a sense, we are using less data to achieve a more targeted calibration, and therefore there is a trade-off between the improved calibration at or near the risk threshold (and therefore also clinical utility) and the precision of estimated sNB. The variability of the estimates of sNB can be attributed to two sources: (1) the variability in estimation of TPR and FPR and (2) the variability in estimates of $\vec{\alpha}$ used to construct the risk score (estimated in the training sample). If heavy down-weighting is applied there is increased variability in $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$.

To account for the reduction in data used to fit the model, we propose reporting the effective sample proportion. The effective sample proportion is the effective sample size used to estimate weighted recalibration parameters divided by the total sample size. Since all weights are bounded between 0 and 1, the effective sample proportion can be calculated as the average weights,

$$Eff = \frac{1}{n}\sum_{i=1}^{n} w_i.$$

In the case where no down-weighted is applied, the entire sample is assigned a weight of 1, and the effective sample proportion is 1.

3.3.3 Selecting Tuning Parameter λ or δ

The tuning parameter λ controls the degree of differential weighting of observations away from the risk threshold. Similarly, for the more general weight w_2, δ controls the degree of differential weighting for observations outside the critical risk interval. Neither λ nor δ have

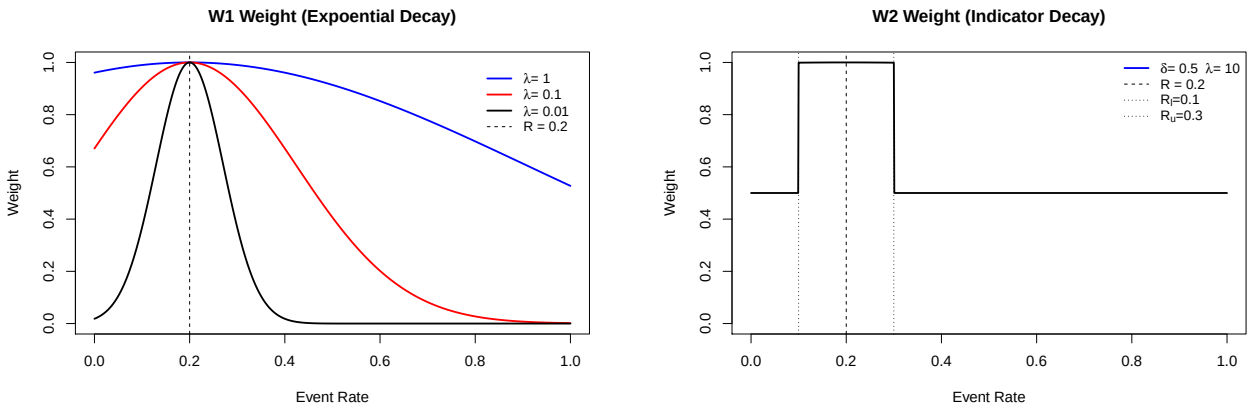


Figure 3.4: Example of weight functions used in weighting scheme.

clinical interpretation, so appropriate choice of λ or δ may not be clear. We define a more interpretable measure, *the relative average weight* (RAW), that can be used to elicit or select an appropriate value of λ or δ. RAW compares the average weight given to observations within an interval (defined by R_l and R_u) around R compared to the average weight of observations outside the interval. Define n_R ($n_{\bar{R}}$) to be the sample size within (outside) the interval around R, in a dataset for recalibration. For weight w_2 the critical risk interval is already defined. For weight w_1 a similar interval should be chosen only for the purpose of eliciting a RAW value. We define the relative average weight to be

$$RAW = \frac{\frac{1}{n_{\bar{R}}} \sum_{i=1}^{n_{\bar{R}}} w_i \mathbf{1}[o(r_i) \notin (R_l, R_u)]}{\frac{1}{n_R} \sum_{i=1}^{n_R} w_i \mathbf{1}[o(r_i) \in (R_l, R_u)]}. \tag{3.11}$$

Weights are bounded by 0 and 1. For w_1, weights within the predefined interval will always be larger than weights outside the interval. This is also true for w_2 because we impose the requirement that δ is bounded above by the the infimum of the weights. Therefore, RAW is bounded by (0,1). A RAW value close to 1 implies that, on average, observations far from R have similar weights as observations near R. RAW close to 1 implies little re-weighting and corresponds to large values of λ or δ. RAW near 0 implies that the average weight for

an observation outside the interval is much smaller than the average weight for observations within the interval. RAW close to 0 implies substantial down-weighting of observations far from the risk threshold and corresponds to a small value of λ or δ.

If there is a prior notion of how much observations outside the interval should be down-weighted (i.e RAW value), λ or δ can be elicited. For weight w_1, given the data and *a priori* defined RAW, R, R_l and R_u, equation (3.11) can be solved for λ numerically via the `uniroot` function in `R`. For weight w_2, when λ is fixed, RAW can similarly be used to elicit appropriate δ. This may be useful in applications where all observations within the interval should be similarly weighted close to 1, but the choice of weight δ for observations outside the interval is unknown.

If RAW is not defined *a priori*, a range of suitable RAW values can be used to obtain a grid of λ or δ. Once a suitable grid of tuning parameter (λ or δ) is obtained, we propose a k-fold cross-validation procedure to select the tuning parameter. We designed the cross-validation procedure to select the tuning parameter that yields highest sNB when weighted recalibration is employed.

For a given tuning parameter k-fold cross-validation is implemented to estimate $sNB(\lambda)$ (or $sNB(\delta)$). This process is repeated for all values in the grid of tuning parameters. Since a single round of cross-validation can be noisy, we propose that cross-validation be repeated multiple times with independent random partitions and the results be averaged. We have found that 25 CV replications of 5-fold cross-validation leads to a suitable reduction in noise. We used $cv.sNB(\lambda)$ to denote the average cross-validated estimate of sNB for tuning parameter λ. The cross-validation procedure steps are outlined in Algorithm 1.

Following Friedman et al. (2001) we implement a "one-standard error" rule for tuning parameter selection. Under the "one-standard error" rule the selected tuning parameter for

Algorithm 1: Repeated Cross-Validation Procedure for Selecting Tuning Parameter

1 Given a sequence of RAW, values obtain grid of tuning parameters λ (δ) via `uniroot` function.

2 Randomly partition data into K-folds of roughly equal size.

3 Implement k-fold cross-validation for the grid of RAW values. Define $sNB_{k,1}(\lambda)$ to be the estimated sNB of λ within fold k. Define $cv.sNB_m(\lambda) = \frac{1}{K}\sum_{i=1}^{K} cv.sNB_{i,1}(\lambda)$ to be the cross-validated estimate of sNB from one round of k-fold cross-validation for tuning parameter λ.

4 Repeat steps (2)-(3) M times, with a new random partitioning each time. Define $cv.sNB(\lambda) = \frac{1}{M}\sum_{i=1}^{M} cv.sNB_m(\lambda)$ to be the repeated cross-validated estimate of sNB for tuning parameter λ.

5 Let λ^* the tuning parameter with largest $cv.sNB(\lambda)$. Select the largest tuning parameter that has $cv.sNB(\lambda)$ within one standard error of $cv.sNB(\lambda^*)$, that is select

$$\lambda = \max\{\lambda : cv.sNB(\lambda) \leq cv.sNB(\lambda^*) - \hat{\sigma}(cv.sNB(\lambda^*))\}.$$

the weighting scheme is the largest λ such that $cv.sNB(\lambda)$ is within one-standard deviation of the maximum $cv.sNB(\lambda)$, across all λ. This provides protection against overfitting the data and selecting a tuning parameter that is too extreme. The one-standard error rule also ensures that for instances when weighted logistic recalibration does not provide any gains in sNB beyond standard logistic recalibration, the largest tuning parameter is chosen, and the weighted logistic solution approximates the standard logistic recalibration solution. To estimate the standard deviation of $cv.sNB(\lambda)$, $\sigma(cv.sNB(\lambda))$, we must account for dependence between the different partitions of cross-validation. We adapt a procedure presented by Gelman et al. (1992) to obtain a variance estimate of $\sigma(cv.sNB)$. We describe this procedure in more detail in Appendix B.

3.4 Simulation Examples

In this section we compare the proposed method to standard logistic recalibration using simulated data. We present four different simulation examples that represent different types of miscalibration. For all examples we consider a risk threshold of $R = 0.3$, and use the symmetric interval around the risk threshold [0.25, 0.35] for selecting tuning parameter λ for weight w_1. Tuning parameters are selected using 5-fold cross-validation with 25 cross-validation replications. Recalibration parameters are estimated from a training set of size 500, 1000, 5000, and 10000. We use a large independent validation dataset of size 100,000 to evaluate the true (rather than estimated) performance of risk scores before and after recalibration.

To simulate risk score data, we implement the following procedure. First, true risks (p_i) are generated from a mixture Beta distribution, comprised of 3 sub-distributions. The sub-distributions are defined by tendency to have low, medium, or high true risks. Beta hyperparameters and mixing proportions vary by example. Next, outcomes Y_i are generated from a $Bern(p_i)$ distribution. Under this simulation set-up the overall event rate is

$$\begin{aligned} E[Y_i] &= E[E[Y_i|p_i]] \\ &= E[p_i] \\ &= \sum_{m=1}^{M} \pi_m \frac{\alpha_m}{\alpha_m + \beta_m}, \end{aligned}$$

where M is the number of subpopulations, α_m and β_m are the Beta hyperparameters, and π_m is mixing proportion for sup-population m. Finally, we induce miscalibration by applying a piecewise polynomial function to the true risk model. We vary the type of miscalibration by example to consider scenarios where consistent overestimation or underestimation of risk occurs, or type of miscalibration differs by sub-population. Full simulation details for each scenario are presented in Appendix B. All comparisons between sNB under the proposed approach and standard logistic recalibration are additive (rather than multiplicative)

differences.

3.4.1 Simulation Example 1: Overestimation for Moderate Risk Groups

For the first simulation example we consider a risk score that is miscalibrated for observations with moderate risk. Figure 3.5 shows the sNB that could potentially be achieved under recalibrations of the risk score. The original risk score is not near the maximum of the curve, indicating that recalibration could improve clinical utility. Standard recalibration appears to offer little improvement. This indicates that there is an opportunity for weighted logistic recalibration to improve the clinical utility of the original risk score.

The red curve in Figure 3.6 shows the calibration curve for the original risk score. Moderate risk scores overestimate risk, while low and high risk scores underestimate risk. In particular, the risk score is poorly calibrated near the risk threshold, $R = 0.3$. The bulk of risk scores are near 0. The disease prevalence for this example is 0.23.

Table 3.1 compares calibration and clinical utility measures of the original risk score, risk score under standard recalibration, and the risk score under the weighted logistic recalibration. For training sample size of $n = 10,000$ the proposed method increases clinical utility by slightly over 5%. Additionally, under weighted logistic recalibration the event rate in a symmetric interval around the risk threshold is closer to the risk threshold (0.272) than under standard recalibration (0.163). This improved calibration is depicted in Figure 3.6, with the calibration curve under the weighted approach closer to the identity line at the risk threshold.

As the training sample size decreases the risk score estimated under the weighted logistic recalibration method has smaller gains in clinical utility and calibration over standard logistic recalibration. The weighted recalibration solution more closely approximates the standard recalibration solution for small sample sizes. As the sample size decreases, the estimated standard error used for the one-standard error rule increases. This leads to larger RAW

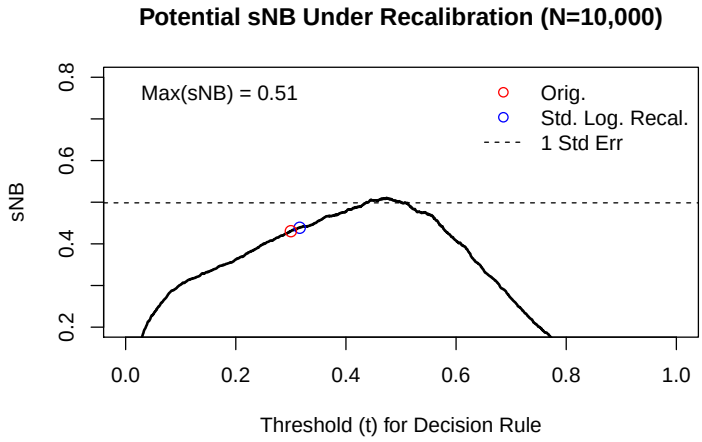


Figure 3.5: Plot of potential sNB, for simulation example 1. The risk threshold is $R = 0.3$. Estimates of sNB and standard error are obtained from sample size $n = 10,000$. The dotted line shows a 1 standard error lower bound for the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error interval, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

values selected during the cross-validation procedure. For example, for $n = 500$, the least amount of weighting is applied, with an effective sample proportion nearly 100%. Curves displaying potential sNB under recalibration and calibration curves for training sample size of $n = 500$, $n = 1000$, and $n = 5000$ are shown in Appendix B.

Standard logistic recalibration prioritizes improved calibration for areas where there are many observations. In Figure 3.6 the histogram of the original risk score shows the majority of the risk scores are near 0, and the next largest cluster of observations is near 0.5. The risk score estimated from standard recalibration is fairly well calibrated near 0 and 0.5. In constrast, weighted logistic recalibration trades off better calibration near the risk threshold, 0.3. The trade-off in this example is worse calibration for the observations whose original predicted risk was near 0.5. This is arguably a trade-off since poor calibration away from the risk threshold will not impact clinical decisions.

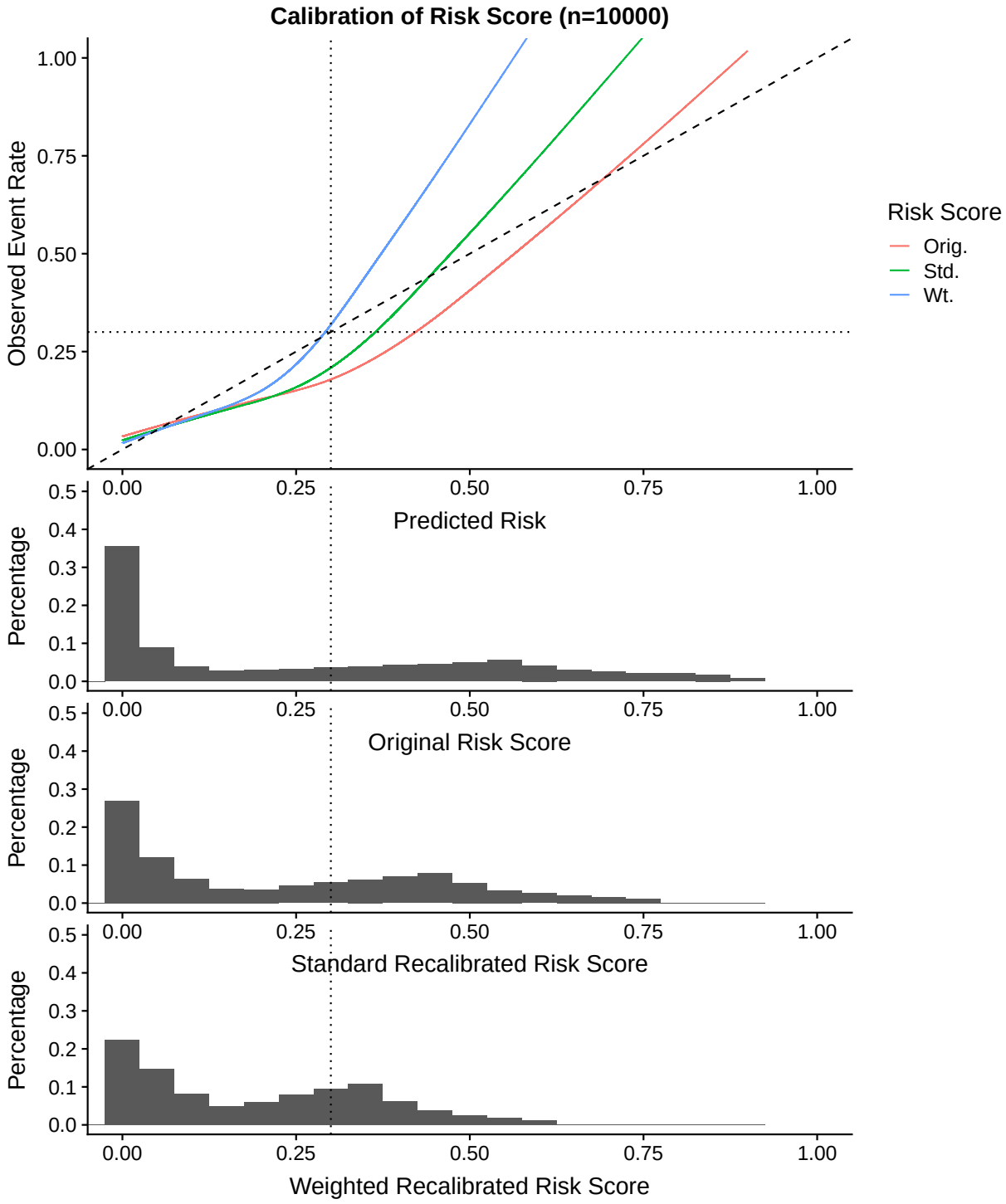


Figure 3.6: Calibration curves and histograms of predicted risks for simulation example 1, with training sample size $N = 10,000$. Calibration of predicted risks estimated under standard logistic recalibration and weighted logistic recalibration are compared.

Table 3.1: Comparison Original, Standard Recalibrated, and Weighted Recalibrated Risk Models for Simulation Example 1

Sample Size	Measure	Orig. (No Recal.)	Std. Recal	Wt. Recal[1]
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.31, 0.69)	(-0.69, 0.53)
	sNB	0.440	0.449	0.502
	TPR	0.844	0.836	0.757
	FPR	0.282	0.271	0.178
	Event Rate for $r_i \in [0.25, 0.35]$	0.161	0.174	0.258
	Prop. Assigned Intervention	0.41	0.40	0.31
	CV-Selected RAW (k)	-	-	0.57 (0.096)
	Effective Sample Proportion (%)	100	100	59.6
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.30, 0.76)	(-0.54, 0.64)
	sNB	0.440	0.458	0.484
	TPR	0.844	0.830	0.795
	FPR	0.282	0.260	0.217
	Event Rate for $r_i \in [0.25, 0.35]$	0.161	0.179	0.215
	Prop. Assigned Intervention	0.40	0.39	0.35
	CV-Selected RAW (k)	-	-	0.73 (0.200)
	Effective Sample Proportion (%)	100	100	74.5
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.36, 0.58)	(-0.46, 0.54)
	sNB	0.440	0.440	0.457
	TPR	0.844	0.844	0.831
	FPR	0.282	0.282	0.262
	Event Rate for $r_i \in [0.25, 0.35]$	0.161	0.170	0.184
	Prop. Assigned Intervention	0.41	0.41	0.39
	CV-Selected RAW (k)	-	-	0.89 (0.526)
	Effective Sample Proportion (%)	100	100	89.8
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.38, 0.66)	(-0.39, 0.65)
	sNB	0.440	0.459	0.461
	TPR	0.844	0.828	0.827
	FPR	0.282	0.258	0.256
	Event Rate for $r_i \in [0.25, 0.35]$	0.161	0.182	0.183
	Prop. Assigned Intervention	0.41	0.39	0.39
	CV-Selected RAW (k)	-	-	0.99 (6.211)
	Effective Sample Proportion (%)	100	100	99.1

[1] Weighted recalibration using exponential decay form of weight.

3.4.2 Simulation Example 2: Overestimation of Risk Across all Risk Levels

Next, we consider an example where risks are overestimated across all levels of risk. The calibration curve in Figure 3.8 shows the original risk score overestimates risks for all risk levels, and risk is most severely overestimated for those with moderate predicted risks. For this setting, the prevalence is 0.16

In a training sample size of 10,000, the original risk score has median risk near 10%, with the highest density of risk scores clustered near 7%. The clinical utility of the original risk score in the training data is 0.242, which is nearly 20% lower than the maximum sNB, as illustrated in Figure 3.7. Standard recalibration improves the clinical utility of the risk score, increasing sNB estimated in the training set to 0.393. Given the large sample size, the standard error around the maximum is fairly small. Therefore, the clinical utility of the risk score estimated from standard logistic recalibration is slightly below the 1 standard error lower bound. Figure B.10 in Appendix B shows as sample size decreases, the standard error for maximum sNB under recalibration increases and the estimated sNB under standard recalibration is within one standard error of the maximum.

Table 3.2 compares the risk scores estimated under standard recalibration and the proposed method. For training sample size of $n = 10,000$ the proposed method increases clinical utility by 1.8%. Figure 3.8 shows that calibration at the risk threshold is slightly improved under the proposed method. Both the standard recalibrated risk score and weighted recalibrated risk score have similar calibration for low risks. For large sample sizes, weighted recalibration trades off worse calibration at high risks, where there are relatively few observations, in favor of better calibration near the risk threshold.

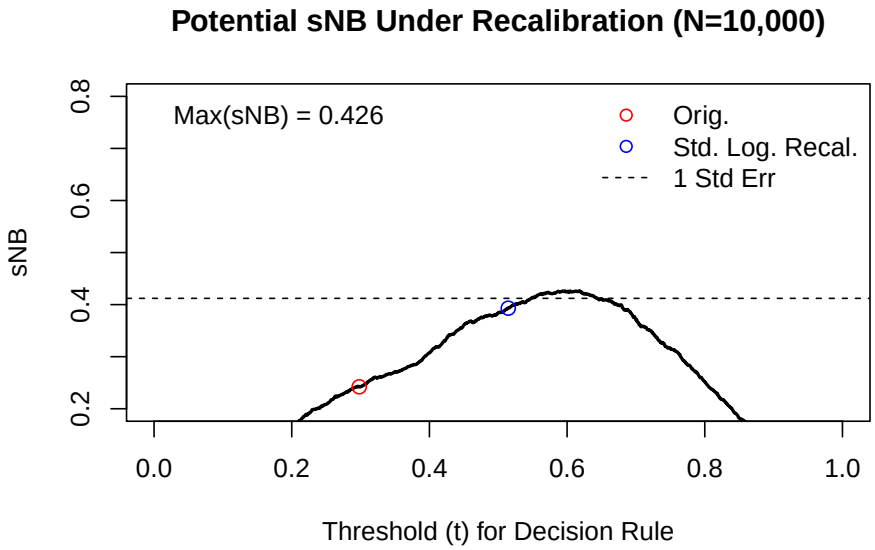


Figure 3.7: Plot of potential sNB, for simulation example 2. The risk threshold is $R = 0.3$. Estimates of sNB and standard error are obtained from sample size $n = 10,000$. The dotted line shows a 1 standard error lower bound for the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error interval, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

As the training sample size decreases the weighted solution tends towards the standard recalibration solution. For sample sizes $N = 500$ and $N = 1000$ the estimated recalibration parameters under the proposed approach very closely approximate the estimated recalibration parameters obtained from standard recalibration. The CV-selected RAW values are 0.99, indicating that little down-weighting is applied outside the RAW interval. In this example, when the sample size of the data available for recalibration is small, there is too little data near the risk threshold to support the weighted approach. Therefore, weighted recalibration approaches standard logistic recalibration.

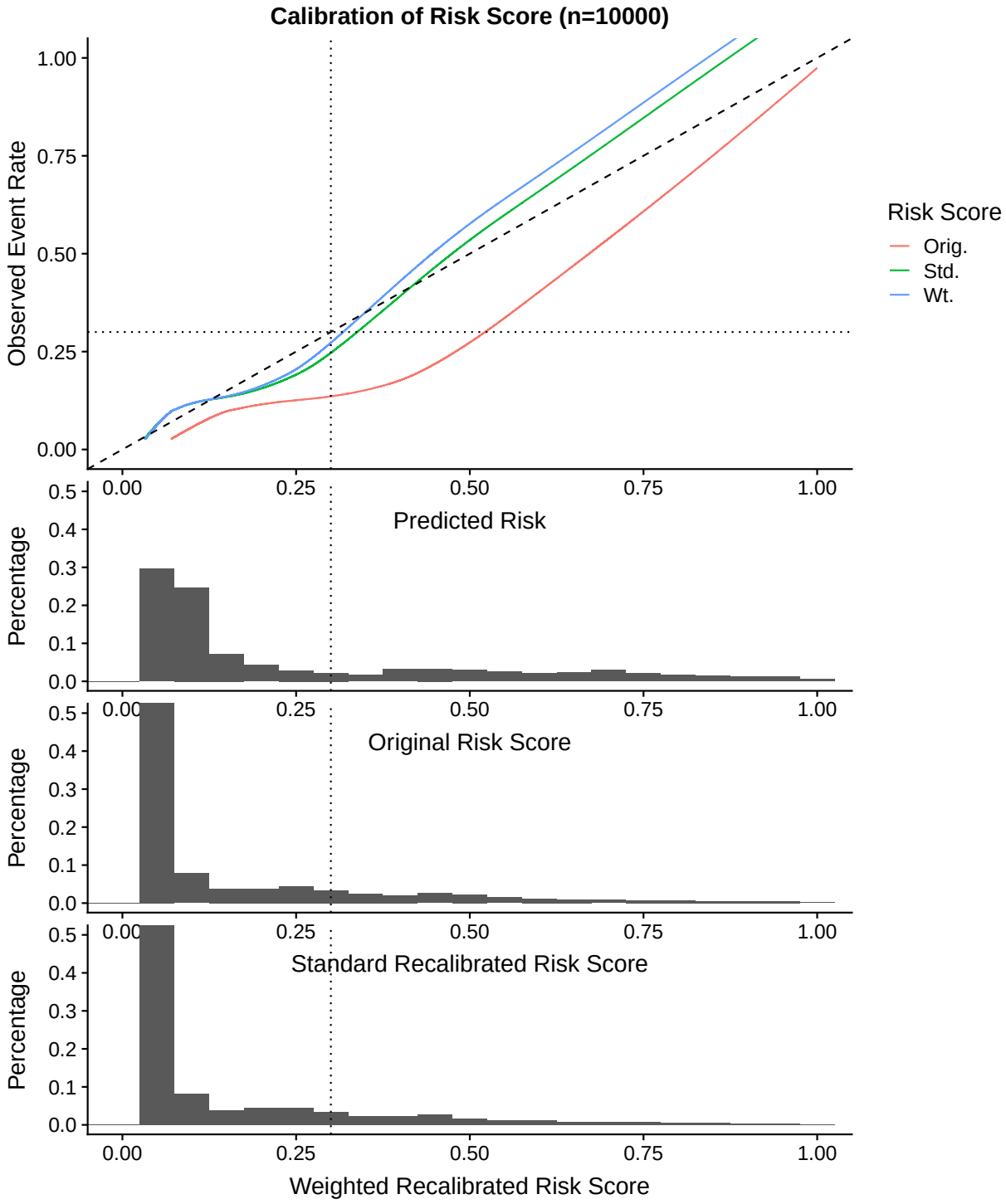


Figure 3.8: Calibration curves and histograms of predicted risks for simulation example 2, with $N = 10,000$. Calibration of predicted risks estimated under standard logistic recalibration and weighted logistic recalibration are compared.

Table 3.2: Comparison Original, Standard Recalibrated, and Weighted Recalibrated Risk Models for Simulation Example 2

Sample Size	Measure	Orig. (No Recal.)	Std. Recal	Wt. Recal[1]
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.90, 0.97)	(-0.98, 0.92)
	sNB	0.282	0.422	0.430
	TPR	0.759	0.641	0.622
	FPR	0.212	0.097	0.086
	Event Rate for $r_i \in [0.25, 0.35]$	0.137	0.215	0.233
	Prop. Assigned Intervention	0.30	0.18	0.17
	CV-Selected RAW (k)	-	-	0.76 (0.212)
	Effective Sample Proportion (%)	100	100	76.7
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.84, 0.99)	(-0.90, 0.95)
	sNB	0.282	0.413	0.421
	TPR	0.759	0.652	0.641
	FPR	0.212	0.106	0.098
	Event Rate for $r_i \in [0.25, 0.35]$	0.137	0.199	0.214
	Prop. Assigned Intervention	0.30	0.19	0.19
	CV-Selected RAW (k)	-	-	0.86 (0.413)
	Effective Sample Proportion (%)	100	100	86.4
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.87, 0.81)	(-0.87, 0.81)
	sNB	0.282	0.417	0.417
	TPR	0.759	0.647	0.647
	FPR	0.212	0.102	0.102
	Event Rate for $r_i \in [0.25, 0.35]$	0.137	0.205	0.206
	Prop. Assigned Intervention	0.30	0.19	0.19
	CV-Selected RAW (k)	-	-	0.99 (5.101)
	Effective Sample Proportion (%)	100	100	99.0
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.87, 0.97)	(-0.88, 0.97)
	sNB	0.282	0.417	0.418
	TPR	0.759	0.646	0.645
	FPR	0.212	0.102	0.101
	Event Rate for $r_i \in [0.25, 0.35]$	0.137	0.207	0.208
	Prop. Assigned Intervention	0.30	0.19	0.19
	CV-Selected RAW (k)	-	-	0.99 (6.414)
	Effective Sample Proportion (%)	100	100	99.0

[1] Weighted recalibration using exponential decay form of weight.

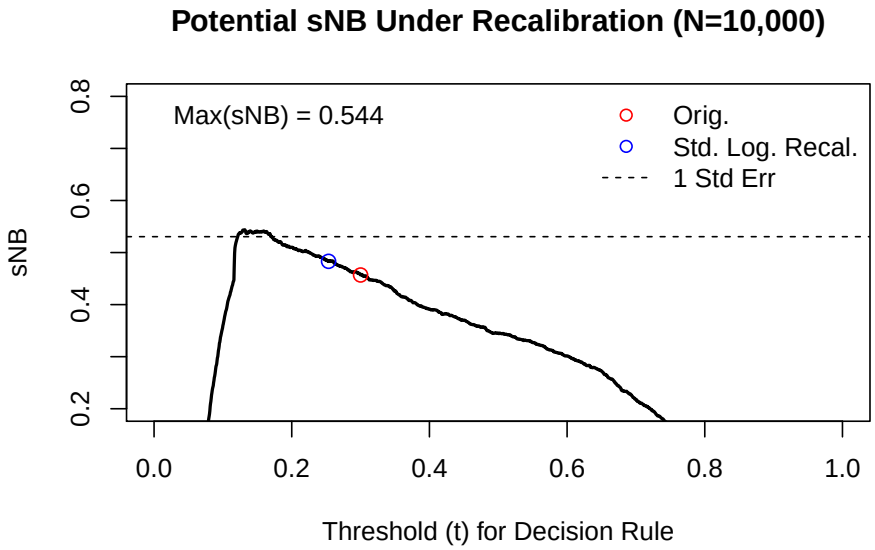


Figure 3.9: Plot of potential sNB, for simulation example 3. The risk threshold is $R = 0.3$. Estimates of sNB and standard error are obtained from sample size $n = 10,000$. The dotted line shows a 1 standard error lower bound for the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error interval, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

3.4.3 Simulation Example 3: Underestimation for moderate risk groups

We now consider a miscalibration example where there is underestimation of risk near the risk threshold, and overestimation elsewhere. In this example, the population prevalence is 0.19. Figure 3.10 displays a histogram of risks for the original risk score. Risks are clustered between 0 and 0.10, and have low density elsewhere. For a training sample size of 10,000 the $\widehat{sNB}$ of the original risk score, estimated in the training set, is 0.456 and the $\widehat{sNB}$ under standard recalibration, estimated in the training set, is 0.483. As illustrated in Figure 3.9, both these estimates are more than one standard error below of the estimated maximum sNB achievable under logistic recalibration.

The calibration curves shown in Figure 3.10 show the improved calibration at the risk threshold under the proposed approach. Under weighted logistic recalibration, risks are well-calibrated at the risk threshold. Under standard recalibration risks are underestimated at the risk threshold. Comparisons of sNB, TPR , and FPR are presented in Table 3.3. For training sample size of $N = 10,000$ weighted logistic recalibration increases clinical utility by 4.2% compared to standard logistic recalibration, and by 6.9% compared to the original risk score. Weighted recalibration substantially down-weights data far from the risk threshold. The CV-selected RAW value is 0.24, meaning the average weight for observations with smoothed event rate, $\hat{o}_i$, in the interval [0.25, 0.35] is about 4 times larger than the average weight for observations outside that interval. Figures B.19-B.21 in Appendix B compare calibration curves for different training sample sizes. For a training sample sizes of $n = 1000$ and $n = 500$ calibration at the risk threshold is similar under standard logistic recalibration and the proposed approach.

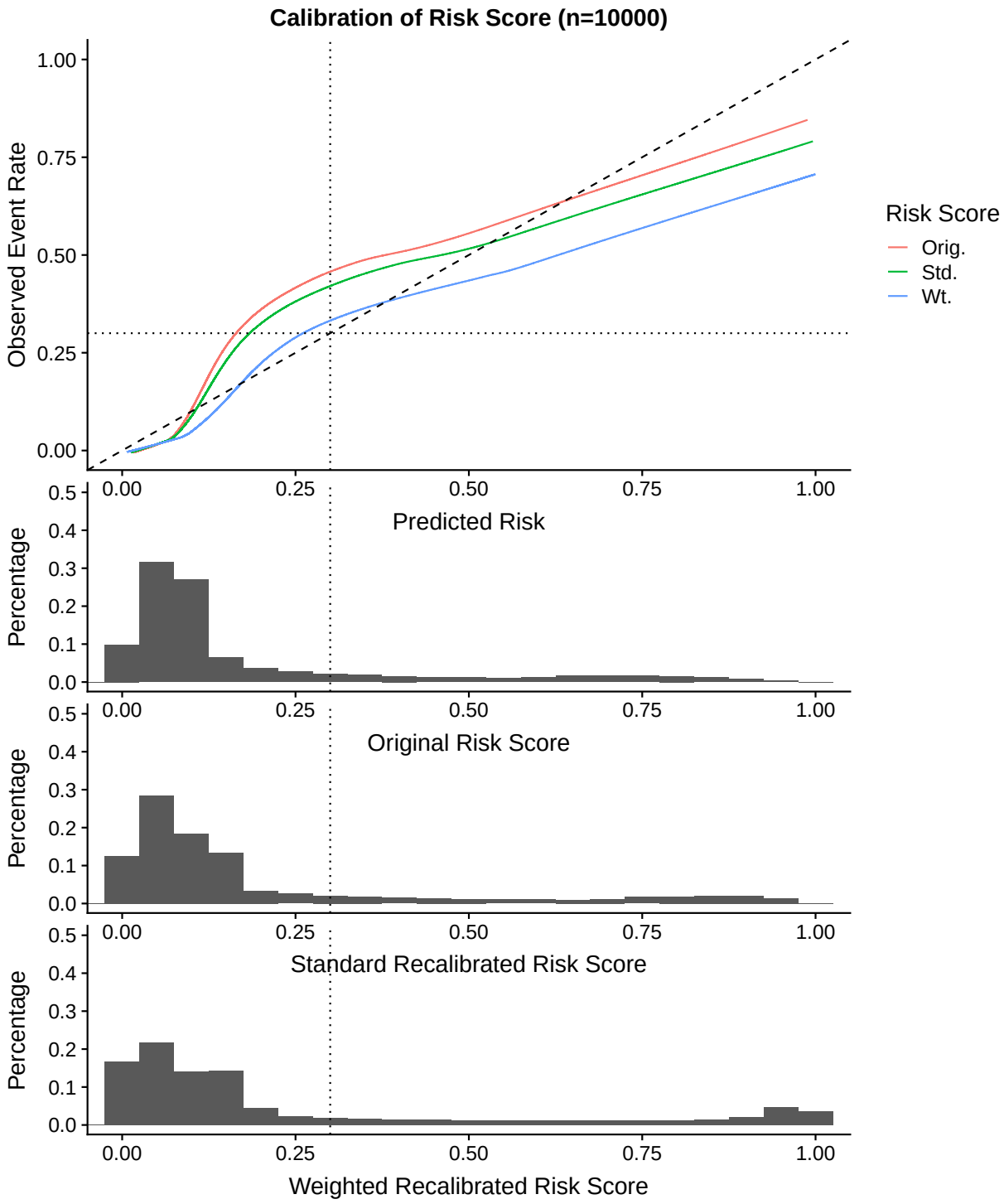


Figure 3.10: Calibration curves and histograms of predicted risks for simulation example 3, with $N = 10,000$. Calibration of predicted risks estimated under standard logistic recalibration and weighted logistic recalibration are compared.

Table 3.3: Comparison Original, Standard Recalibrated, and Weighted Recalibrated Risk Models for Simulation Example 3

Sample Size	Measure	Orig. (No Recal.)	Std. Recal	Wt. Recal[1]
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.40, 1.16)	(1.60, 1.59)
	sNB	0.430	0.457	0.499
	TPR	0.581	0.634	0.745
	FPR	0.083	0.097	0.135
	Event Rate for $r_i \in [0.25, 0.35]$	0.475	0.442	0.377
	Prop. Assigned Intervention	0.18	0.20	0.25
	CV-Selected RAW (k)	-	-	0.24 (0.034)
	Effective Sample Proportion (%)	100	100	26.1
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.34, 1.12)	(1.46, 1.45)
	sNB	0.430	0.455	0.503
	TPR	0.581	0.629	0.759
	FPR	0.083	0.096	0.140
	Event Rate for $r_i \in [0.25, 0.35]$	0.475	0.450	0.371
	Prop. Assigned Intervention	0.18	0.20	0.26
	CV-Selected RAW (k)	-	-	0.18 (0.024)
	Effective Sample Proportion (%)	100	100	21.0
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.19, 1.07)	(0.19, 1.07)
	sNB	0.430	0.445	0.445
	TPR	0.581	0.608	0.609
	FPR	0.083	0.090	0.090
	Event Rate for $r_i \in [0.25, 0.35]$	0.475	0.456	0.457
	Prop. Assigned Intervention	0.18	0.19	0.19
	CV-Selected RAW (k)	-	-	0.99 (6.893)
	Effective Sample Proportion (%)	100	100	99.9
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.23, 1.10)	(0.22, 1.10)
	sNB	0.430	0.446	0.446
	TPR	0.581	0.610	0.610
	FPR	0.083	0.090	0.090
	Event Rate for $r_i \in [0.25, 0.35]$	0.475	0.455	0.454
	Prop. Assigned Intervention	0.18	0.19	0.19
	CV-Selected RAW (k)	-	-	0.99 (6.442)
	Effective Sample Proportion (%)	100	100	99.0

[1] Weighted recalibration using exponential decay form of weight.

3.4.4 Simulation Example 4: Underestimation for all risk groups

Finally, we consider an example where risks are underestimated across all predicted risks. In this example the distribution of predicted risks is similar to example 3. The disease prevalence in this example is 0.23. Figure 3.11 shows the estimated maximum sNB that can be achieved under logistic recalibration, indicating one standard error below the maximum. The estimated sNB for the risk score estimated under standard recalibration is slightly below this lower bound. This indicates the potential for weighted logistic recalibration to produce a risk score with higher sNB. Figure B.26 in Appendix B shows that for smaller samples the estimated sNB under standard recalibration is within or near the boundary of the 1 standard error lower bound.

In this example, we see some improvement in calibration at the risk threshold and sNB when the training sample size is large. The calibration curves shown in Figure 3.12 show slight improvement in calibration at the risk threshold under the proposed approach compared to standard logistic recalibration. As shown in Table 3.4, for training sample size of $N = 10,000$ the estimated standardized net benefit under the weighted logistic recalibration is 1% higher than the standardized net benefit under standard logistic recalibration.

Figures B.27-B.29 in Appendix B compare calibration curves for different training sample sizes. As the training sample size decreases the gains in the proposed method over standard recalibration lessen. In particular, when $N = 500$ or $N = 1000$ the estimated recalibration parameters under the proposed approach approximate the estimates under standard logistic recalibration. In this simulation example, there are very few observations near the risk threshold for sample sizes of $N = 500$ and $N = 1000$. When $N = 500$ there are fewer than 5 observations in the RAW interval [0.25,0.35], and 12 observations when $N = 1000$. Given the limited number observations near the risk threshold for small sample sizes, the standard error estimates for $cv.sNB(\lambda)$ are quite large, leading to selection of a large RAW value.

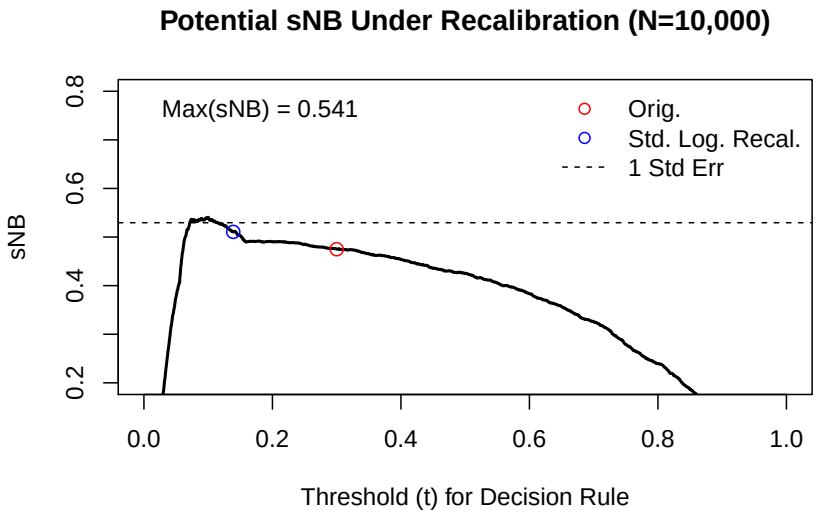


Figure 3.11: Plot of potential sNB, for simulation example 4. The risk threshold is $R = 0.3$. Estimates of sNB and standard error are obtained from sample size $n = 10,000$. The dotted line shows a 1 standard error lower bound for the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error interval, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

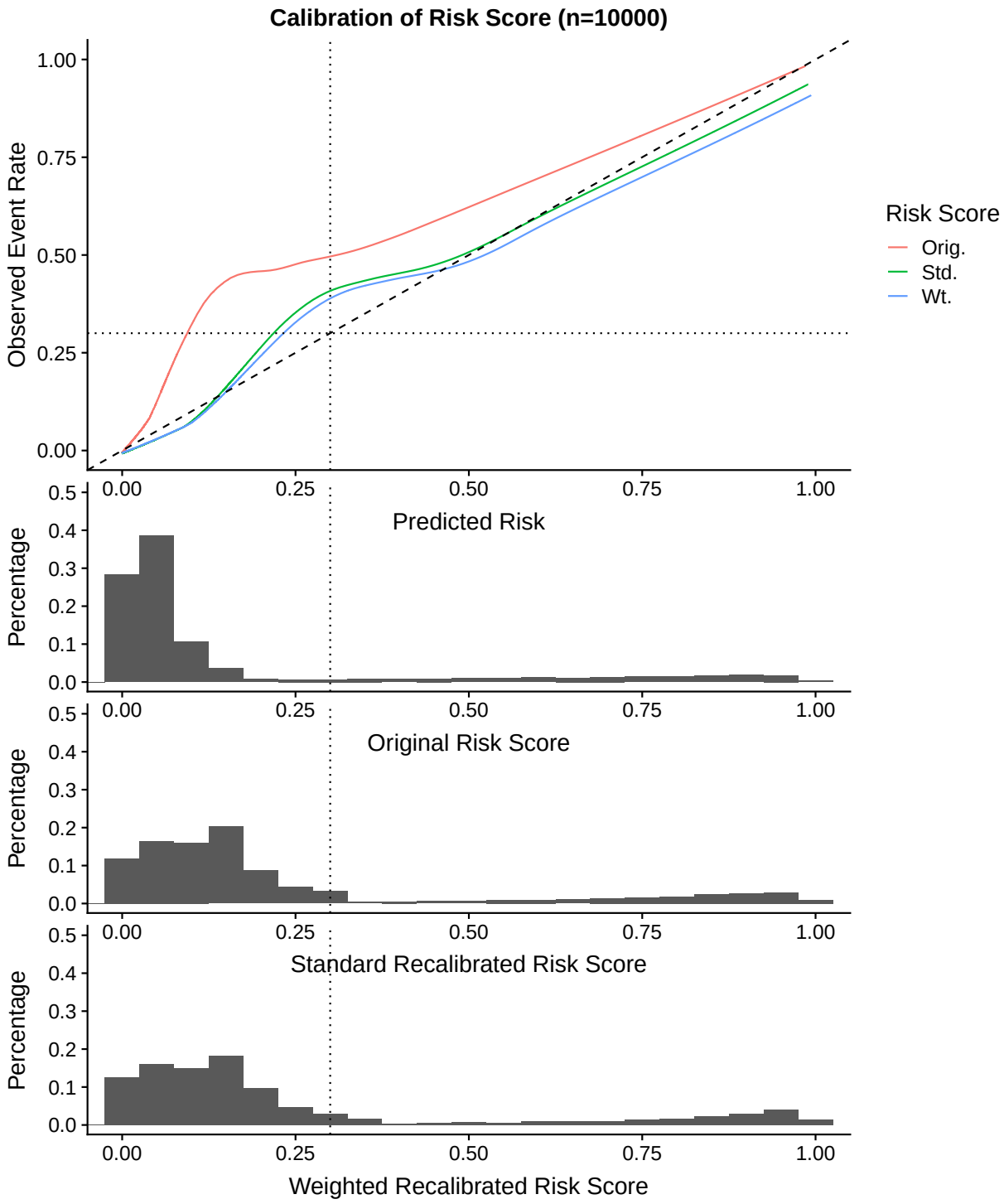


Figure 3.12: Calibration curves and histograms of predicted risks for simulation example 4, with $N = 10,000$. Calibration of predicted risks estimated under standard logistic recalibration and weighted logistic recalibration are compared.

Table 3.4: Comparison Original, Standard Recalibrated, and Weighted Recalibrated Risk Models for Simulation Example 4

Sample Size	Measure	Orig. (No Recal.)	Std. Recal	Wt. Recal[1]
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.75, 0.89)	(0.99, 0.96)
	sNB	0.475	0.511	0.521
	TPR	0.546	0.615	0.640
	FPR	0.052	0.077	0.088
	Event Rate for $r_i \in [0.25, 0.35]$	0.510	0.429	0.410
	Prop. Assigned Intervention	0.17	0.21	0.22
	CV-Selected RAW (k)	-	-	0.70 (0.195)
	Effective Sample Proportion (%)	100	100	71.6
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.64, 0.82)	(0.86, 0.88)
	sNB	0.475	0.511	0.523
	TPR	0.546	0.619	0.644
	FPR	0.052	0.080	0.089
	Event Rate for $r_i \in [0.25, 0.35]$	0.510	0.424	0.413
	Prop. Assigned Intervention	0.17	0.21	0.22
	CV-Selected RAW (k)	-	-	0.74 (0.214)
	Effective Sample Proportion (%)	100	100	75.2
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.51, 0.84)	(0.52, 0.84)
	sNB	0.475	0.491	0.491
	TPR	0.546	0.581	0.581
	FPR	0.052	0.066	0.066
	Event Rate for $r_i \in [0.25, 0.35]$	0.510	0.448	0.447
	Prop. Assigned Intervention	0.17	0.19	0.19
	CV-Selected RAW (k)	-	-	0.99 (7.557)
	Effective Sample Proportion (%)	100	100	99.0
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.77, 0.90)	(0.86, 0.93)
	sNB	0.475	0.510	0.513
	TPR	0.546	0.616	0.624
	FPR	0.052	0.077	0.082
	Event Rate for $r_i \in [0.25, 0.35]$	0.510	0.432	0.423
	Prop. Assigned Intervention	0.17	0.21	0.21
	CV-Selected RAW (k)	-	-	0.90 (0.775)
	Effective Sample Proportion (%)	100	100	90.5

[1] Weighted recalibration using exponential decay form of weight.

3.5 Recalibration of ACC-AHA-ASCVD Risk Score in the MESA Cohort

In 2013, the American College of Cardiology (ACC) and American Heart Association (AHA) published guidelines recommending that individuals with estimated 10-year risk of atherosclerotic cardiovascular disease (ASCVD) greater than 7.5% receive statin therapy (Goff et al., 2014). Paired with the guidelines, the joint panel developed the ACC-AHA-ASCVD risk calculator to estimate the 10-year risk of ASCVD events. The ACC-AHA-ASCVD risk calculator has been shown to overestimate the risk of ASCVD by 37-154% in men and by 8-57% in women (DeFilippis et al., 2015). Combining overestimated risks with risk-based treatment guidelines implies over-treatment in the population. This miscalibration can be a serious issue with large public health impact (Ridker and Cook, 2013).

MESA is a prospective study of cardiovascular disease in a large, nationwide, multi-ethnic cohort. Men and women who were free of clinical cardiovascular disease were enrolled into the study (Bild et al., 2002). Demographic and clinical data were collected at baseline, and participants were monitored for over 10-years for cardiovascular clinical events. Recalibrating the ACC-AHA-ASCVD tool for the MESA cohort, while taking into account the treatment threshold (i.e., an estimated 10-year risk of at least 7.5%), could give clinicians and patients greater confidence in using the ACC-AHA-ASCVD tool and improve the clinical utility of the risk tool for the population.

Figure 3.13 shows the estimated potential sNB of the ACC-AHA-ASCVD risk score. The estimated sNB of the risk score after applying standard recalibration is near the maximum of the curve, suggesting that weighted logistic recalibration may not offer additional gains in sNB or improved calibration at the risk threshold. The calibration plots for the risk score presented in Figure 3.14 support this by showing good calibration near the clinically relevant risk threshold of 7.5%.

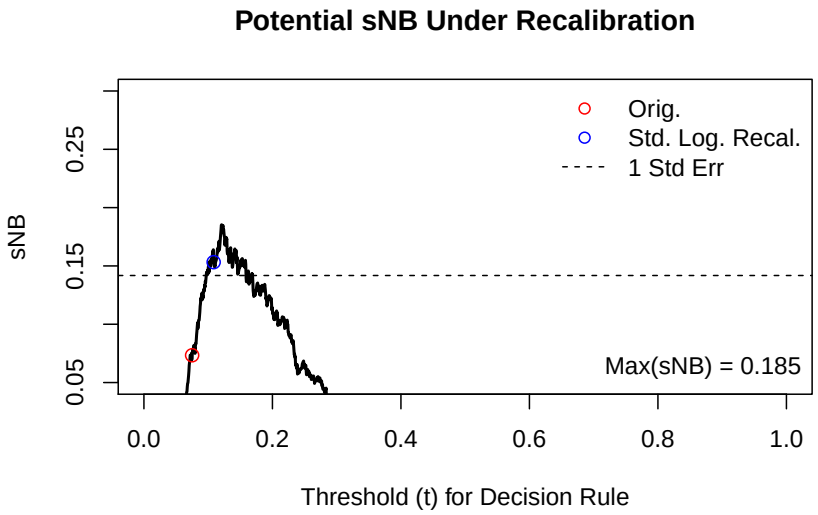


Figure 3.13: Plot of potential sNB for ACC-AHA-ASCVD risk score in MESA cohort. The dotted lines indicates one standard error lower-bound of the maximum of the curve. The red dot shows the estimated sNB for the ACC-AHA-ASCVD risk score. The blue dot shows the estimated sNB for the recalibrated risk score after standard logistic recalibration.

We compare the risk models recalibrated under standard logistic recalibration to the weighted approach using the indicator function approximation weight, w_2. Some MESA participants were not a part of the population that the ACC-AHA-ASCVD risk score was developed for and were excluded from analysis. Inclusion criteria and final sample size are shown in Figure B.34 in Appendix B. We are interested in correcting for issues of miscalibration across the range of clinically relevant risks. The clinically critical risk interval is $[2.5\%, 10\%]$. Therefore we fix large $\lambda = 10$, so that there is very little exponential decay of weights within the clinically relevant risk interval. The RAW-tuning parameter for weighting is chosen *a priori*, so cross-validation is not needed. We select $RAW = 0.01$, so that observations outside the risk interval have little influence in estimating recalibration parameters. We use a bootstrap approach to correct for the optimistic bias that arises from estimating recalibration parameters and measures of performance in the same data (Harrell, 2015). We used 500 bootstrap replications to estimate the optimism and to estimate 95% confidence

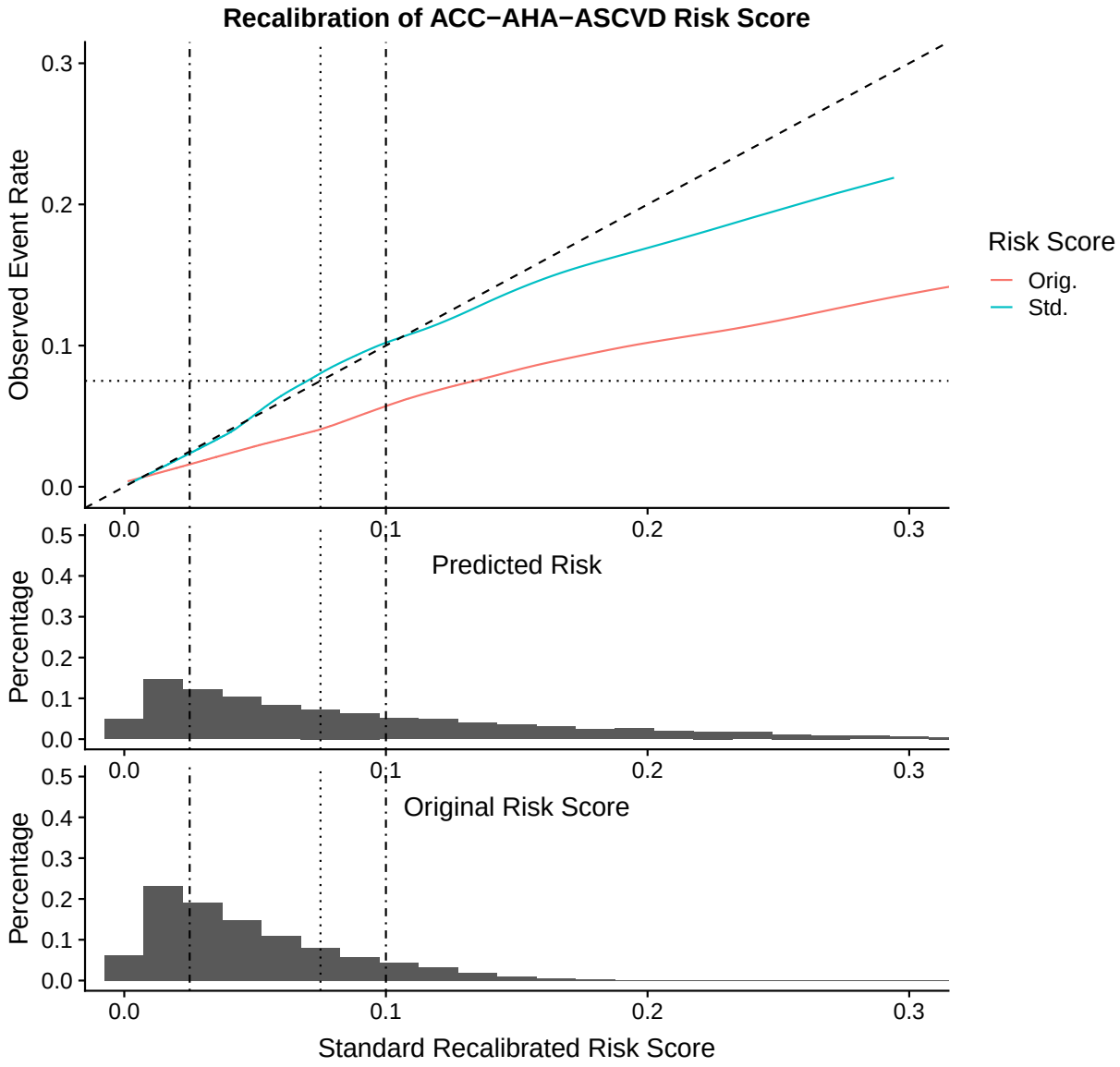


Figure 3.14: Calibration curve of the ACC-AHA-ASCVD risk score before and after standard recalibaration in MESA cohort. The dotted line indicates the risk threshold, 7.5%. The dotted-dashed lines indicate the clinically relevant risk interval, [2.5%, 10.0%], where clinicians are most concerned about good calibration.

intervals for sNB.

In accordance with treatment guidelines and clinical practice, we use a risk threshold of 7.5% with risk interval $[2.5\%, 10\%]$. Given the results in Figure 3.13, we expect the weighted recalibration solution approximates the standard recalibration solution. Though the risk threshold should be selected based on the benefits and harms of the intervention, we additionally consider two other risk thresholds, 5% and 15% to illustrate the method. For indicator weight definition, we pair the 5% risk threshold with the risk interval $[0\%, 7.5\%]$ and the 15% risk threshold with risk interval $[10\%, 20\%]$.

Table 3.5 shows the results of applying standard recalibration and the weighted method. For risk thresholds 5% and 7.5% the standard and weighted recalibration methods produce a risk score with similar $\widehat{sNB}$, $\widehat{TPR}$, and $\widehat{FPR}$ compared to standard logistic recalibration. Figures 3.15 - 3.17 show the calibration curves after standard and weighted recalibration for the different risk thresholds. For $R = 5\%$ or 7.5%, the recalibrated risk scores estimated via standard and weighted recalibration are similar within the clinically relevant risk interval.

For $R = 0.05$ or $R = 0.075$ it is expected that standard recalibration improves calibration at the risk threshold sufficiently, and the weighted approach approximates standard recalibration. Standard recalibration tends to focus recalibration efforts for ranges of risk scores with high density. As shown in Figure 3.15, risk scores are overestimated for the entire range of risks. The fitted recalibration parameters that correct miscalibration for very low risks, also correct the of miscalibration for higher risks, thereby producing good calibration at the risk threshold.

When $R = 0.15$ we see that there is a 1% increase in $\widehat{sNB}$ for the recalibrated risk score obtained from the weighted approach compared to standard logistic recalibration. The calibration curve in Figure 3.17 also shows improved calibration at the risk threshold under the

weighted approach compared to standard logistic recalibration. For this setting, the weighted recalibration approach makes a trade-off of improving calibration at the risk threshold for worse calibration for low-risk scores. However, as shown in Table 3.5 the optimism-corrected estimates of sNB are below 0, indicating that given a risk threshold of 15% and its implied harms to benefit ratio, an alternative clinical strategy, namely treat-none, may have higher clinical utility.

Due to the fact that MESA is an ethnically diverse cohort, and the differential miscalibration of the ACC-AHA-ASCVD risk observed in men and women by (DeFilippis et al., 2015), there is interest in evaluating and correcting miscalibration of the ACC-AHA-ASCVD risk score within different subgroups defined by sex and/or ethnicity. Figures B.35 -B.38 in Appendix B show the calibration curves and potential sNB for male, female, black, and hispanic cohorts, and their intersections. Among these, there is potential to increase sNB and improve calibration at the risk threshold beyond standard logistic recalibration in the black male cohort, as shown in Figure B.39. We applied standard and weighted logistic recalibration and weighted logistic recalibration to the Black, male MESA cohort. We found that the weighted approach may offer small improvement in $\widehat{sNB}$ over standard recalibration. However, we caution that the sample size of the Black, male cohort is small. Full results are presented in Appendix B.

Table 3.5: Comparison of recalibration methods in MESA for RAW = 0.01, for different risk thresholds. An indicator weight is used for the weighted recalibration approach

Measure	Orig.	Std. Recal.	Wt. Recal.
		R = 0.05	
$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-1.08, 0.81)	(-0.90, 0.88)
Effective Sample Proportion %	100	100	76
$\widehat{sNB}$ (95% CI)[1]	0.274 (0.184, 0.364)	0.333 (0.258, 0.404)	0.332 (0.254, 0.400)
$\widehat{sNB}^2$	-	0.324 (0.255 0.401)	0.329 (0.250, 0.396)
$\widehat{TPR}^2$	0.873	0.714	0.717
$\widehat{FPR}^2$	0.442	0.374	0.379
Event Rate for $r_i \in [0.025, 0.075]$	0.02	0.03	0.03
		R = 0.075	
$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-1.08, 0.81)	(-0.67, 0.99)
Effective Sample Proportion %	100	100	53
$\widehat{sNB}$ (95% CI)[1]	0.073 (-0.035, 0.181)	0.153 (0.065 0.240)	0.161 (0.074 0.246)
$\widehat{sNB}^2$	-	0.152 (0.063, 0.239)	0.157 (0.069, 0.242)
$\widehat{TPR}^2$	0.771	0.470	0.518
$\widehat{FPR}^2$	0.584	0.201	0.228
Event Rate for $r_i \in [0.025, 0.10]$	0.03	0.06	0.05
		R = 0.15	
$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-1.08, 0.81)	(-1.32, 0.63)
Effective Sample Proportion %	100	100	13
$\widehat{sNB}$ (95% CI)[1]	-0.200 (-0.312, -0.086)	-0.004 (-0.028, 0.032)	0.005 (-0.012, 0.036)
$\widehat{sNB}^2$	-	-0.007 (-0.032, 0.029)	-0.001 (-0.019, 0.030)
$\widehat{TPR}^2$	0.449	0.053	0.029
$\widehat{FPR}^2$	0.584	0.017	0.008
Event Rate for $r_i \in [0.10, 0.20]$	0.08	0.12	0.12

[1]Delta-method derived std error used for original risk score 95% CI calculation.
Bootstrap used with 500 replications used for standard and weighted recalibration

[2] Optimism corrected estimates

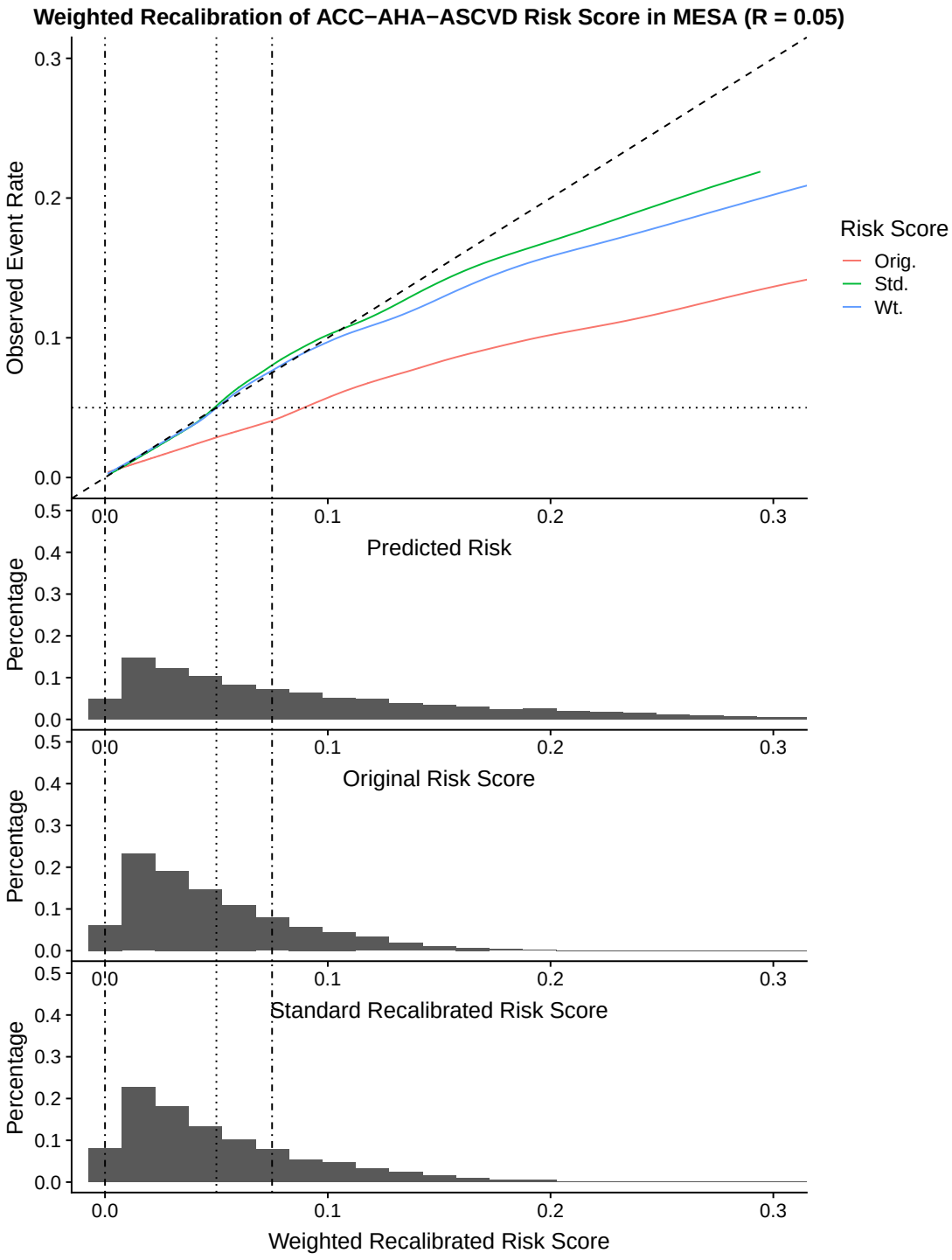


Figure 3.15: Calibration curve of the ACC-AHA-ASCVD risk score in MESA cohort, after standard and weighted recalibration approaches. The dotted line indicates the **risk threshold of 5%**. The dotted-dashed lines indicate the clinically relevant risk interval, [0%, 7.5%], where clinicians are also concerned about good calibration.

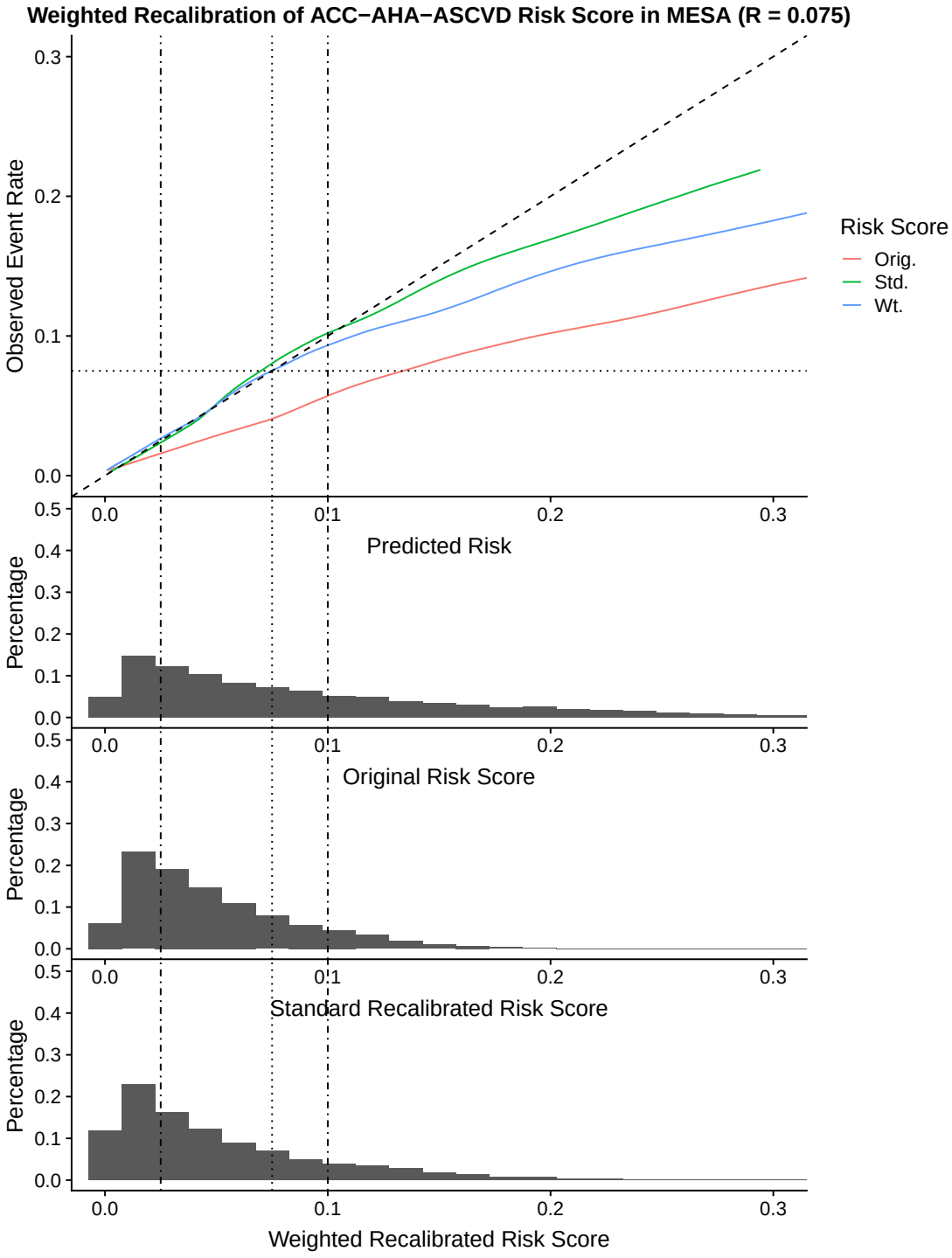


Figure 3.16: Calibration curve of the ACC-AHA-ASCVD risk score in MESA cohort, after standard and weighted recalibration approaches. The dotted line indicates the **risk threshold of 7.5%**. The dotted-dashed lines indicate the clinically relevant risk interval, [2.5%, 10.0%], where clinicians are most concerned about good calibration.

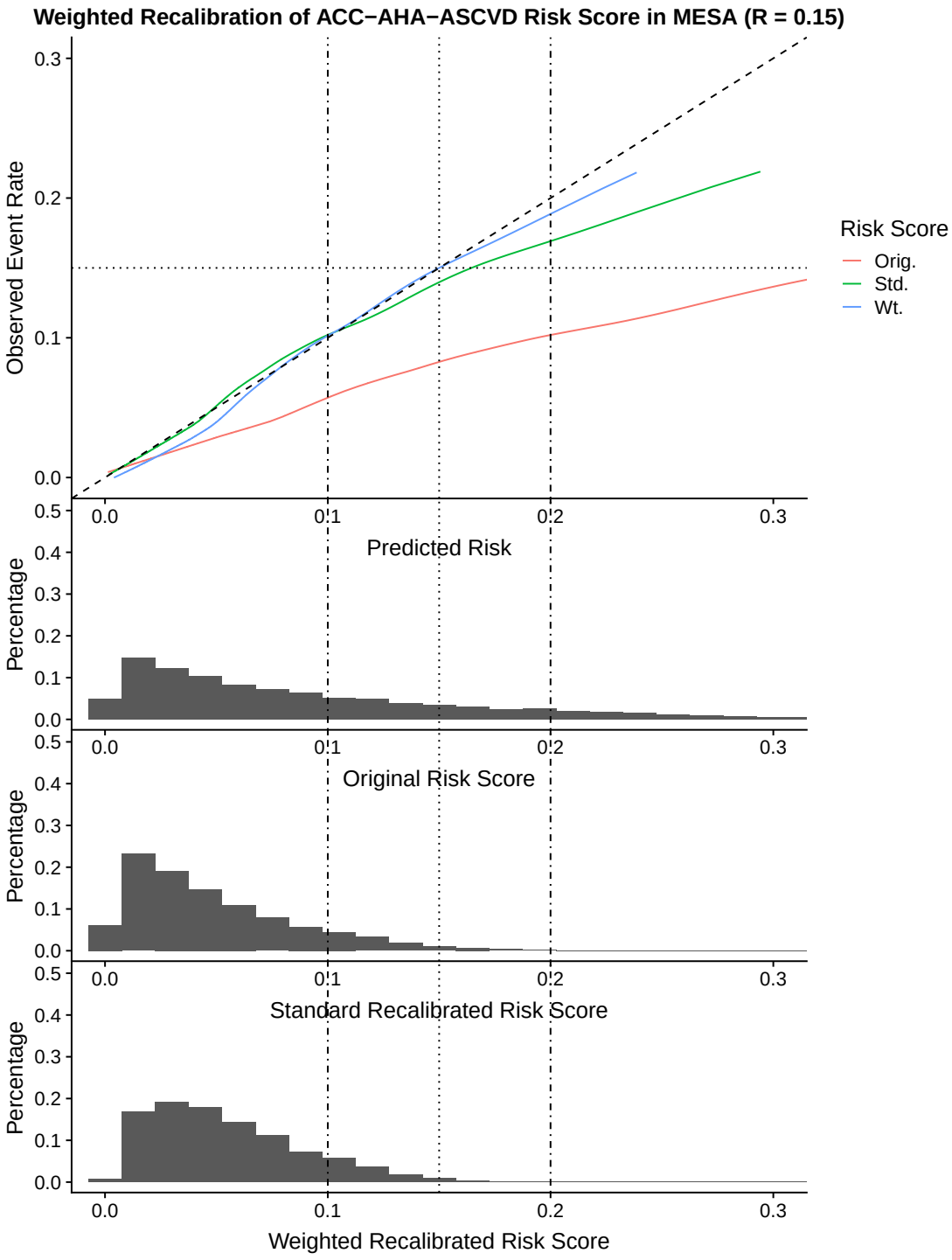


Figure 3.17: Calibration curve of the ACC-AHA-ASCVD risk score in MESA cohort, and after standard and weighted recalibration. The dotted lines indicate the **risk threshold of 15%**. The dotted-dashed lines indicate the clinically relevant risk interval, [10%, 20%], where clinicians are most concerned about good calibration.

3.6 Discussion

We presented an approach for recalibrating risk models that prioritizes good calibration near critical risk thresholds that are used to make clinical decisions. In addition to the examples presented, we conducted extensive simulations illustrating examples where weighted logistic recalibration produces better calibration at the risk threshold and recalibrated risks with higher clinical utility. Compared to standard logistic recalibration, the weighted approach can produce sizable improvements in the clinical utility of recalibrated risks. We applied the proposed methods to the ACC-AHA-ASCVD risk model using the MESA study data.

The proposed method can be considered a generalized version of standard logistic recalibration. Unlike standard logistic recalibration, the proposed approach requires tuning parameters. To aid in selecting a tuning parameter, we proposed the measure RAW, which has a simple interpretation. If RAW is not specified in advance then one can use a cross-validation procedure to optimize the standardized net benefit of the recalibrated risk model, which adds to the computational cost compared to standard recalibration. We have suggested guidelines, such as the one standard error rule, to protect against overfitting.

Under the weighted approach, observations are down-weighted, meaning that effectively less data are used to estimate recalibration parameters. We suggested reporting the effective sample proportion to gauge the impact of applying weighting. Bootstrap confidence intervals for sNB can be used to gauge variability in estimates of sNB. When there are few events, applying heavy down-weighting may be undesirable. In these instances, the cross-validation procedure paired with a one-standard error rule will indicate that the data do not support the weighted approach, and a large RAW value will be selected.

There are some settings when standard logistic recalibration sufficiently improves calibration at the risk threshold. The MESA application illustrated one setting when this occurs.

The ACC-AHA-ASCVD risk score overestimated risk across the the entire range of predicted risks.

Standard logistic recalibration will tend to achieve good calibration for regions with the highest density of risk scores. When the nature of the miscalibration at the risk threshold is similar to the miscalibration that exists for the bulk of observations, standard recalibration may adequately improve calibration at the risk threshold. In fact, even if the miscalibration patterns at the risk threshold differ for different ranges of predicted risks, standard recalibration could still improve calibration near the risk threshold. For example, suppose a risk model has high density of observations for ranges of low risk scores and is paired with a large risk threshold (i.e., the risk threshold is not near the majority of risk scores). Further suppose the risk model is "over-fitting", meaning low risk scores are underestimated and high risk score are overestimated. In this setting, standard recalibration could still correct both issues of miscalibration by estimating a recalibration slope that is less than 1.

To help identify settings in which one might consider weighted logistic recalibration, we presented a graphical device which, which captures the potential improvement in clinical utility recalibration can offer. For a pre-defined risk score, with known risk threshold we have provided a variance formula to estimate the variability sNB. From this variance formula, a guideline based on standard deviation of sNB is useful for distinguishing settings were weighted logistic recalibration may not provide additional improvement in clinical utility beyond standard recalibration and may not be worthwhile to pursue.

Chapter 4

CONSTRAINED LOGISTIC RECALIBRATION FOR IMPROVED CLINICAL UTILITY OF RISK SCORES

4.1 Introduction

In Chapter 3 we connected calibration of a risk model at the clinically relevant risk threshold and the clinical utility of the risk model. Under certain assumptions, a risk model that is well-calibrated at the risk threshold has maximum sNB among all logistically recalibrated risk models. The converse is also true; a risk model that has maximum sNB among all logistically recalibrated risk models is calibrated at the risk threshold. The former statement motivated the weighted logistic recalibration method presented in Chapter 3. The latter statement motivates the recalibration method presented in this chapter.

We propose a recalibration method that aims to produce a recalibrated risk model with maximum standardized net benefit among all logistically recalibrated risk models. To achieve this, we introduce a constraint to the logistic likelihood optimization problem. A side-effect of maximizing standardized net benefit is good calibration near the risk threshold.

The relevant background is the same as for Chapter 3. The remaining sections are organized as follows. In Section 4.2 we present the constrained recalibration approach and discuss implementation details. In Section 4.3 we give examples using simulated data. Section 4.4 applies the proposed method to the MESA study cohort introduced in Chapter 3. We close with a discussion of the method, results, and future work in Section 4.5.

4.2 Methods

4.2.1 Constrained Logistic Recalibration

Following the notation of Chapter 3, let r_i, be the predicted risk obtained from a model predicting binary outcome Y_i, for $i = 1, \ldots, n$. According to the risk model, r_i is a function of a set of risk predictors X (i.e $r_i = \hat{P}[Y_i = 1|X_i]$). Let $Z_i = \text{logit}(r_i)$ be the logit transformed risk score. Let R be the risk threshold used to prescribe intervention, which represents harms associated with prescribing the intervention to a control and the benefits associated with prescribing the intervention to a case. Let α_0 and α_1 be the recalibration intercept and slope.

Given a risk score r and risk threshold R, the standardized net benefit of the recalibrated risk score (obtained via logistic recalibration) as function of the recalibration parameters is

$$sNB(\alpha_0, \alpha_1) = P\left(\frac{e^{\alpha_0+\alpha_1 Z}}{1+e^{\alpha_0+\alpha_1 Z}} > R \middle| Y = 1\right) - \frac{P(Y=0)}{P(Y=1)}\frac{R}{1-R}P\left(\frac{e^{\alpha_0+\alpha_1 Z}}{1+e^{\alpha_0+\alpha_1 Z}} > R \middle| Y = 0\right). \tag{4.1}$$

A plug-in estimator for sNB is

$$\widehat{sNB}(\alpha_0, \alpha_1) = \frac{1}{n_D}\sum_{i=1}^{n_D} \mathbb{1}\left[\frac{e^{\alpha_0+\alpha_1 Z_i}}{1+e^{\alpha_0+\alpha_1 Z_i}} > R\right] - \frac{\hat{P}(Y=0)}{\hat{P}(Y=1)}\frac{R}{1-R}\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}} \mathbb{1}\left[\frac{e^{\alpha_0+\alpha_1 Z_i}}{1+e^{\alpha_0+\alpha_1 Z_i i}} > R\right], \tag{4.2}$$

where n_D and $n_{\bar{D}}$ are the number of cases and controls, respectively, in the sample. To find recalibration parameters, $\vec{\alpha}$, that maximize $\widehat{sNB}$ (among all logistic recalibrated risk scores), one could directly optimize (4.2). However, this can be a difficult task since (4.2) is a non-smooth function. Moreover, since sNB is a rank-based measure, there may not be a unique $\vec{\alpha}$ that maximizes sNB.

We address these challenges by reformulating the problem with sNB as the constraint rather than the objective function, and by selecting a more well-behaved objective function. We propose using a binomial log-likelihood objective function, the same objective function used in standard logistic recalibration. The parameter space is restricted to only include recalibration parameters $\vec{\alpha}$ that produce recalibrated risk scores with high sNB. As in Chapter 3, we assume the risk score is monotonically non-decreasing and therefore restrict the α_1-parameter space to positive real numbers. Let $\mathbb{R}^+$ denote the set of positive real numbers. We propose estimating recalibration parameters $\vec{\alpha} = (\alpha_0, \alpha_1)$ via the following constrained maximization problem.

$$\begin{aligned}(\alpha_0^*, \alpha_1^*) = &\underset{(\alpha_0,\alpha_1)\in\mathbb{R}\times\mathbb{R}^+}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} Y_i(\alpha_0 + \alpha_1 Z_i) - \log\left(1 + e^{\alpha_0+\alpha_1 Z_i}\right) \\ &\text{subject to} \quad \widehat{sNB}(\alpha_0, \alpha_1) \geq \widehat{sNB}_{\max} - \hat{\sigma}(\widehat{sNB}_{max}),\end{aligned} \tag{4.3}$$

where $\widehat{sNB}_{\max}$ is the estimated maximum achievable sNB among all risk scores of the form

$$r_i^* = \frac{e^{\alpha_0+\alpha_1 Z_i}}{1 + e^{\alpha_0+\alpha_1 Z_i}}.$$

Following Remark 3 in Chapter 3, for a fixed harm-benefit ratio $\frac{C}{B}$, $\widehat{sNB}_{\max}$ can be found by varying decision threshold t. That is, $\widehat{sNB}_{\max}$ can be found by solving the 1-dimensional optimization problem

$$\widehat{sNB}_{\max}(r_i) = \max_{t\in[0,1]} \left\{ \widehat{TPR}_t(r_i) - \frac{\hat{P}(Y_i = 0)}{\hat{P}(Y_i = 1)} \frac{C}{B} \widehat{FPR}_t(r_i) \right\}.$$

Recall that if the risk threshold, R, is selected rationally, $\frac{C}{B} = \frac{R}{1-R}$.

Acknowledging that there is variability in $\widehat{sNB}_{\max}$, we use a "one-standard-error" type of rule in the inequality constraint. This rule is often used when tuning conventional penalized regression methods (Friedman et al., 2001). The constrained parameter space includes all

parameters $\vec{\alpha}$ that produce risk scores with $\widehat{sNB}$ within one standard error of $\widehat{sNB}_{\max}$. Following Proposition 1, the plug-in estimator of $\sigma(\widehat{sNB}_{max})$ with $\frac{R}{1-R} = \frac{C}{B}$ is

$$V\left[\widehat{sNB}_t\left(r\right)\right] = \left(\frac{1}{\hat{p}_{11}+\hat{p}_{01}}\right)^2 \left[\left(\frac{C}{B}\right)^2 \hat{p}_{10} + \frac{\hat{p}_{11}\left(p_{01}+\frac{C}{B}\hat{p}_{10}\right)^2}{(\hat{p}_{11}+\hat{p}_{01})^2} + \frac{\hat{p}_{01}(\frac{R}{1-R}\hat{p}_{10}-\hat{p}_{11})^2}{(\hat{p}_{11}+\hat{p}_{01})^2}\right].$$

where

$$\hat{p}_{11} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[r_i \geq R, Y_i = 1], \quad \hat{p}_{01} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[r_i < R, Y_i = 1],$$
$$\hat{p}_{10} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[r_i \geq R, Y_i = 0], \quad \hat{p}_{00} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[r_i < R, Y_i = 0].$$

Figure 4.1 shows an example of the $\vec{\alpha}$-parameter space with a polygon indicating the set of recalibration parameters that have $\widehat{sNB}$ within one standard error of $\widehat{sNB}_{max}$. The top panel of Figure 4.1 shows an example where the standard logistic recalibration solution is outside the the constrained parameter space, meaning the recalibrated risk score obtained from constrained logistic recalibration should have higher $\widehat{sNB}$ compared to standard logistic recalibration. $\vec{\alpha}$ parameters on or within the blue polygon produce a risk score that satisfy the $\widehat{sNB}(\alpha_0, \alpha_1)$ constraint. Of that set, the parameters with the largest log-likelihood lie on the top boundary line. The bottom panel shows a setting when standard logistic recalibration produces a recalibrated risk score with $\widehat{sNB}$ within one standard error of $\widehat{sNB}_{max}$, and therefore the constrained parameter space contains the standard logistic recalibration solution. For the example, the solution under standard recalibration and constrained logistic recalibration will be the same.

4.2.2 Implementation

Because of the low-dimensional setting ($p = 2$), this problem is well-suited for optimization methods that ensure a global maximum is found. To solve the optimization problem in (4.3) we use the DIRECT optimization algorithm (Jones et al., 1993; Gablonsky and

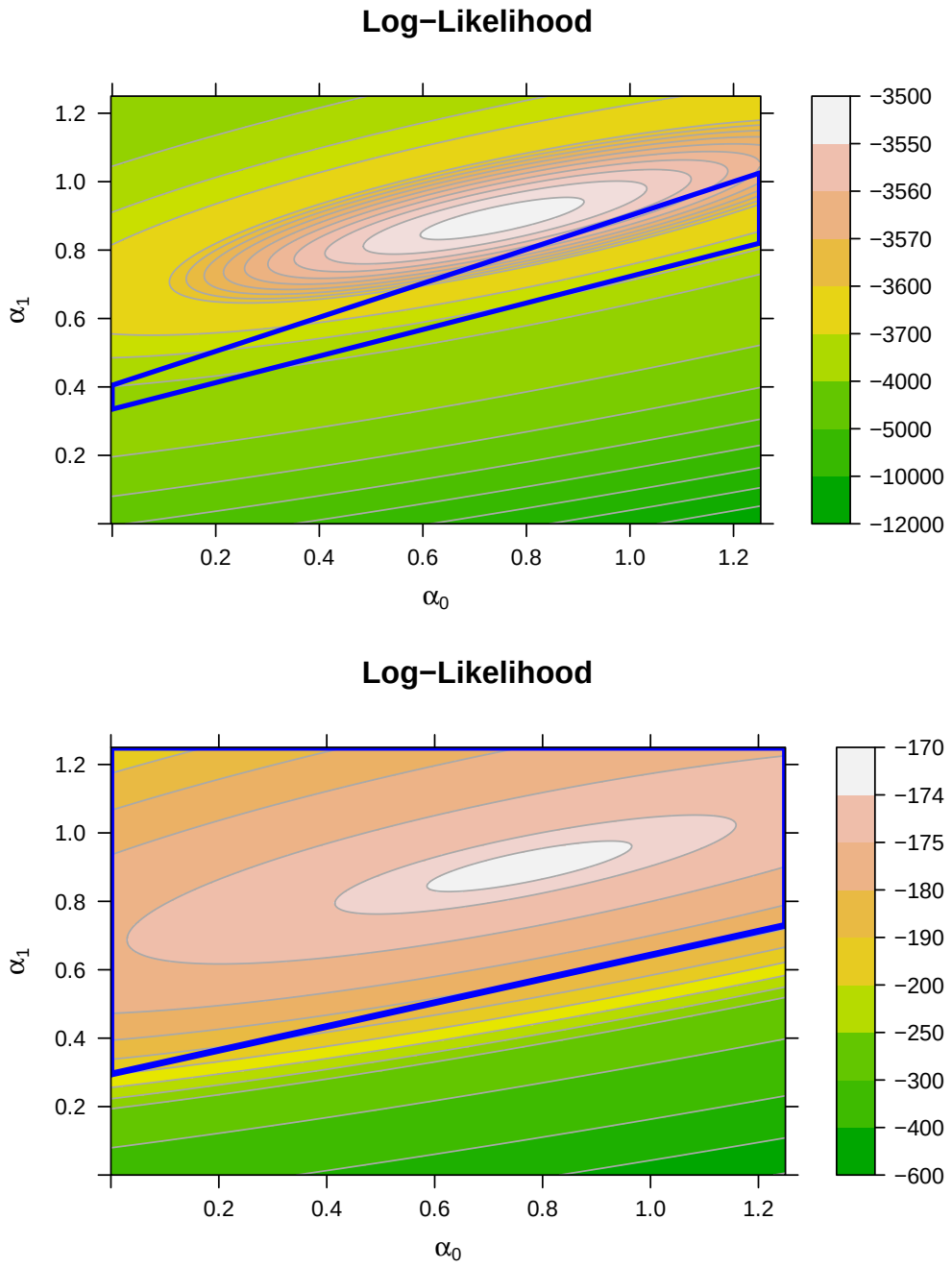


Figure 4.1: Example of log-likelihood objective function for different recalibration parameters $\vec{\alpha}$. The region within the blue polygon shows the constrained parameter space, where $\widehat{sNB}(\vec{\alpha})$ is within one standard error of $\widehat{sNB}_{max}$. The white region is where the standard recalibration solution lies. In the top panel, the constraint region does not included the standard recalibration solution, and therefore constrained logistic recalibration method will have higher $\widehat{sNB}(\vec{\alpha})$ compared to logistic recalibration. In the bottom panel, the constraint region includes the standard recalibration solution, and therefore constrained logistic recalibration will be the same as standard logistic recalibration.

Kelley, 2001). The algorithm searches the parameter space by dividing the domain into rectangles and performing a deterministic search. Importantly, the DIRECT algorithm allows non-linear, non-smooth constraint functions. We implement the algorithm via the `nloptr` package in `R` (Johnson, 2014). Implementation of the optimization algorithm requires the parameter space be bounded with a "box" constraint, which initializes the parameter search. If r is monotonically decreasing, it follows that $\alpha_1 > 0$. Given the magnitude of recalibration parameters, a large box constraint should not impact solutions. We have successfully used the box constraint $\{\alpha_0, \alpha_1 : \alpha_0 \in [-10, 10], \alpha_1 \in [0, 10]\}$ with success.

The optimization software requires specification of two stopping rules for convergence. The first convergence criterion is based on the objective function f, the binomial log-likelihood. The optimization routine will converge if $\frac{|\Delta f|}{|f|}$ falls below some tolerance, where $|f|$ is the absolute value of the log-likelihood for a given optimization step and $|\Delta f|$ is the absolute change in log-likelihood between two optimization steps. We have found the convergence criterion $\frac{|\Delta f|}{|f|} < 1 \times 10^{-8}$ to work well in simulations. The second convergence criterion is based on the parameters $\vec{\alpha}$. We set a stopping rule of $\frac{||\Delta\vec{\alpha}||}{||\vec{\alpha}||} < 1 \times 10^{-4}$, where $||\vec{\alpha}||$ is the $L2$ norm of $\vec{\alpha}$ at a given optimization step and $||\Delta\vec{\alpha}||$ represents the absolute change in proposed $\vec{\alpha}$ values optimization steps. The optimizer ends its search when either criterion is met.

4.3 Simulation Examples

In this section, we use simulated data to compare constrained logistic recalibration to standard logistic recalibration, and the previously proposed weighted logistic recalibration. We re-visit the four examples from Chapter 3, and again consider a risk threshold of $R = 0.3$. Appendix B gives complete details. Recalibration parameters are estimated from training sets of sizes $n = 500, 1000, 5000$, and 10000. We compare sNB and calibration at the risk threshold between the methods in an independent validation dataset of size $n = 100,000$ to evaluate the true (rather than estimated) performance of risk scores. Additionally, we com-

pare the constrained optimization approach to the weighted recalibration approach presented in Chapter 3. All comparisons between sNB under the proposed approach and standard logistic recalibration are additive (rather than multiplicative) differences.

4.3.1 Simulation Example 1: Overestimation for Moderate Risk Groups

First, we consider an example where the risk score is overestimated for observations with moderate predicted risk. In the training data, we estimated $\widehat{sNB}_{max} = 0.507$, with estimated standard error of 0.012. Therefore, the lower-bound used to constrain the parameter space is $\widehat{sNB}(\vec{\alpha}) = 0.495$. Table 4.1 shows the estimated recalibration parameters and sNB under standard, weighted, and constrained logistic recalibration. For the largest training sample size, the risk score estimated from constrained logistic recalibration has 5.4% higher sNB compared to standard logistic recalibration. As sample size decreases, constrained logistic recalibration has sustained improvements in sNB compared to standard logistic recalibration.

As illustrated in Figure 4.2, when $n = 10000$ the constrained logistic recalibrated risk score is well-calibrated at the risk threshold. Figures B.6 - B.8 show the calibration curves for estimated recalibration parameters for smaller sample sizes. For training samples sizes of $n = 1000$ and $n = 5000$ the calibration curves show improved calibration at the risk threshold compared to standard logistic recalibration. For $n = 500$, the proposed approach has little improvement in calibration at the risk threshold compared to standard logistic recalibration.

For large sample sizes, the sNB of the estimated risk model produced by constrained recalibration is similar to the sNB of the estimated risk model produced by weighted recalibration. As sample size decreases, the sNB for weighted recalibration is similar to that under standard logistic recalibration, while the constrained optimization approach has sustained increased sNB compared to standard logistic recalibration.

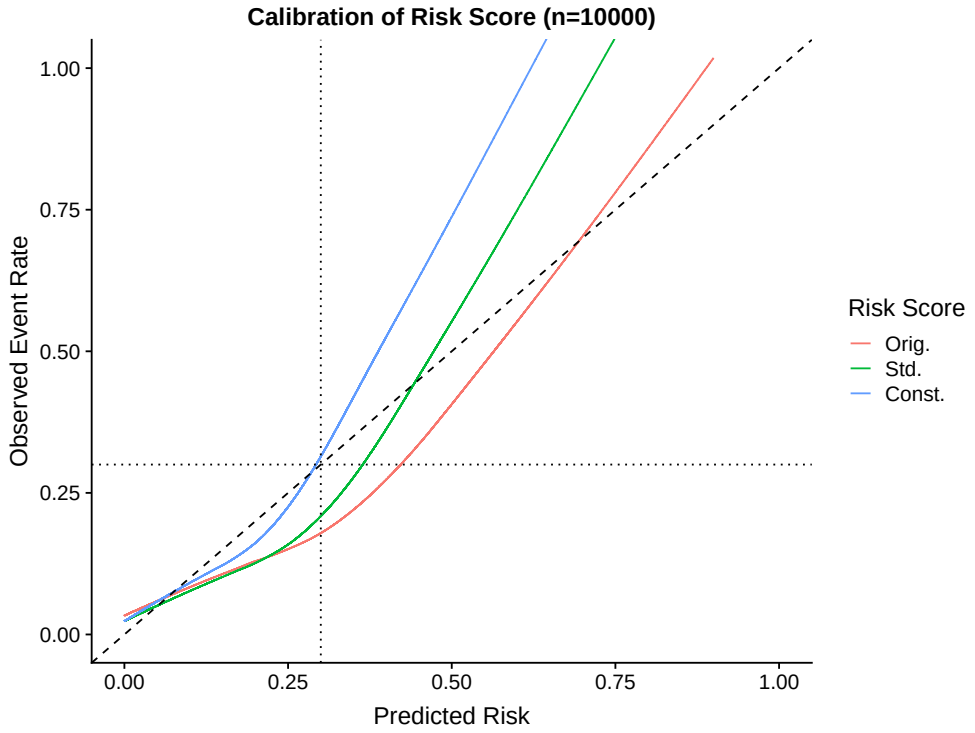


Figure 4.2: Calibration curves comparing standard logistic recalibration and constrained logistic recalibration for simulation example 1. Recalibration parameters were estimated in a training set of size 10,000.

Table 4.1: Comparison of Constrained Recalibration to Other Recalibration Methods for Simulation Example 1.

Training Sample Size	Measure	Orig. (No Recal.)	Std. Recal.	Wt. Recal.	Const. Recal.
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.31, 0.69)	(-0.69, 0.53)	(-0.66, 0.66)
	sNB	0.440	0.449	0.502	0.503
	TPR	0.844	0.836	0.757	0.751
	FPR	0.282	0.271	0.173	0.178
	Prop. Assigned Intervention	0.41	0.40	0.32	0.31
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.30, 0.76)	(-0.54, 0.64)	(-0.49, 0.79)
	sNB	0.440	0.458	0.484	0.488
	TPR	0.844	0.830	0.795	0.789
	FPR	0.282	0.260	0.217	0.210
	Prop. Assigned Intervention	0.41	0.39	0.35	0.34
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.36, 0.57)	(-0.46, 0.54)	(-0.52, 0.58)
	sNB	0.440	0.440	0.457	0.476
	TPR	0.844	0.844	0.831	0.808
	FPR	0.282	0.282	0.262	0.232
	Prop. Assigned Intervention	0.41	0.41	0.39	0.36
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.38, 0.66)	(-0.39, 0.65)	(-0.49, 0.63)
	sNB	0.440	0.459	0.461	0.475
	TPR	0.844	0.828	0.827	0.810
	FPR	0.282	0.258	0.256	0.232
	Prop. Assigned Intervention	0.41	0.39	0.39	0.36

4.3.2 Simulation Example 2: Overestimation of Risk Across all Risk Levels

Next, we consider an example where there is systematic overestimation of risk. For training sample size $n = 10,000$, constrained logistic recalibration yields a 1% improvement in sNB compared to standard logistic recalibration. Similarly, in Figure 4.3 calibration at the risk threshold for the risk score obtained from constrained logistic recalibration is slightly improved over standard logistic recalibration. Figures B.14 - B.16 in Appendix B show the calibration curves for recalibration parameters estimated in training sets with smaller sample sizes. For training sample sizes of $n = 1,000$ and $n = 500$ the recalibration parameters estimated under constrained logistic recalibration are the same as those estimated under standard logistic recalibration. For these settings, the constrained parameter space includes the logistic recalibration solution.

As shown in Table 4.2, for this example, the weighted and constrained recalibration

Table 4.2: Comparison of Constrained Recalibration to Other Recalibration Methods for Simulation Example 2.

Training Sample Size	Measure	Orig. (No Recal.)	Std. Recal.	Wt. Recal.	Const. Recal.
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.90, 0.97)	(-0.98, 0.92)	(-1.01, 0.95)
	sNB	0.282	0.422	0.430	0.431
	TPR	0.759	0.641	0.622	0.618
	FPR	0.212	0.097	0.086	0.083
	Prop. Assigned Intervention	0.30	0.18	0.17	0.17
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.84, 0.99)	(-0.90, 0.95)	(-1.11, 1.11)
	sNB	0.282	0.413	0.421	0.435
	TPR	0.759	0.652	0.641	0.605
	FPR	0.212	0.106	0.098	0.076
	Prop. Assigned Intervention	0.30	0.19	0.19	0.16
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.87, 0.81)	(-0.87, 0.81)	(-0.87, 0.81)
	sNB	0.282	0.417	0.417	0.417
	TPR	0.759	0.647	0.647	0.647
	FPR	0.212	0.102	0.102	0.102
	Prop. Assigned Intervention	0.30	0.19	0.19	0.19
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.87, 0.97)	(-0.88, 0.97)	(-0.87, 0.97)
	sNB	0.282	0.417	0.418	0.417
	TPR	0.759	0.646	0.645	0.646
	FPR	0.212	0.102	0.101	0.102
	Prop. Assigned Intervention	0.30	0.19	0.19	0.19

methods tend to perform similarly. When the sample size is large, both methods produce a risk model with sNB nearly 1% larger than the risk model estimated under standard logistic recalibration. For training sample sizes $n = 500$ and $n = 1000$ both methods approximate the standard logistic recalibration solution.

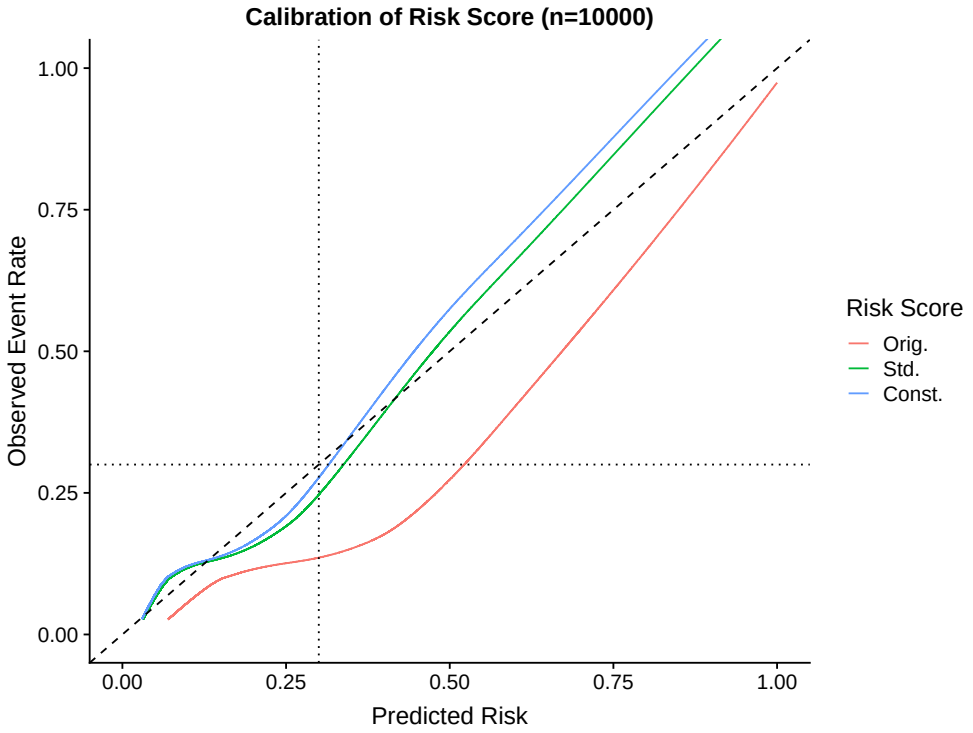


Figure 4.3: Calibration curves comparing standard logistic recalibration and constrained logistic recalibration for simulation example 2. Recalibration parameters were estimated in a training set of size 10,000.

4.3.3 Simulation Example 3: Underestimation for Moderate Risk Groups

Now we consider a risk score that underestimates risk for observations with moderate predicted risk. Table 4.3 shows the sNB for the risk score estimated under the different recalibration methods. For the largest training sample size, sNB is 4.4% higher under constrained logistic recalibration compared to standard logistic recalibration. Figure 4.4 shows the calibration of the constrained logistic recalibrated risk score. Notably, the calibration at the risk threshold is perfect for predicted risks estimated under the constrained logistic recalibration method, while standard logistic recalibration has substantial miscalibration. For small training sample sizes, there are sustained gains in sNB for risk scores estimated under constrained logistic recalibration compared to standard logistic recalibration (see Figures B.22 - B.30 in Appendix B). As the training sample size decreases, there is a reduction in calibration at the risk threshold. This reduction in sNB can be a result of the larger constraint space that occurs as a result of small sample sizes. When the sample size is small, the standard error around $\widehat{sNB}_{max}$ increases, and therefore the constraint region increases to include recalibration parameters that may have higher log-likelihood but lower net benefit.

When the training sample size is large, $n = 5000$ or $n = 10000$, the constrained and weighted recalibration approaches produce risk scores with similar sNB. For smaller sample sizes, the weighted approach approximates the standard recalibration solution. In contrast, the gains in sNB of the risk model obtained from constrained optimization compared to standard recalibration remain even for small sample size. Notably, when $n = 1000$ the weighted approach has the same sNB as standard logistic recalibration, while the constrained approach has nearly 4% higher sNB.

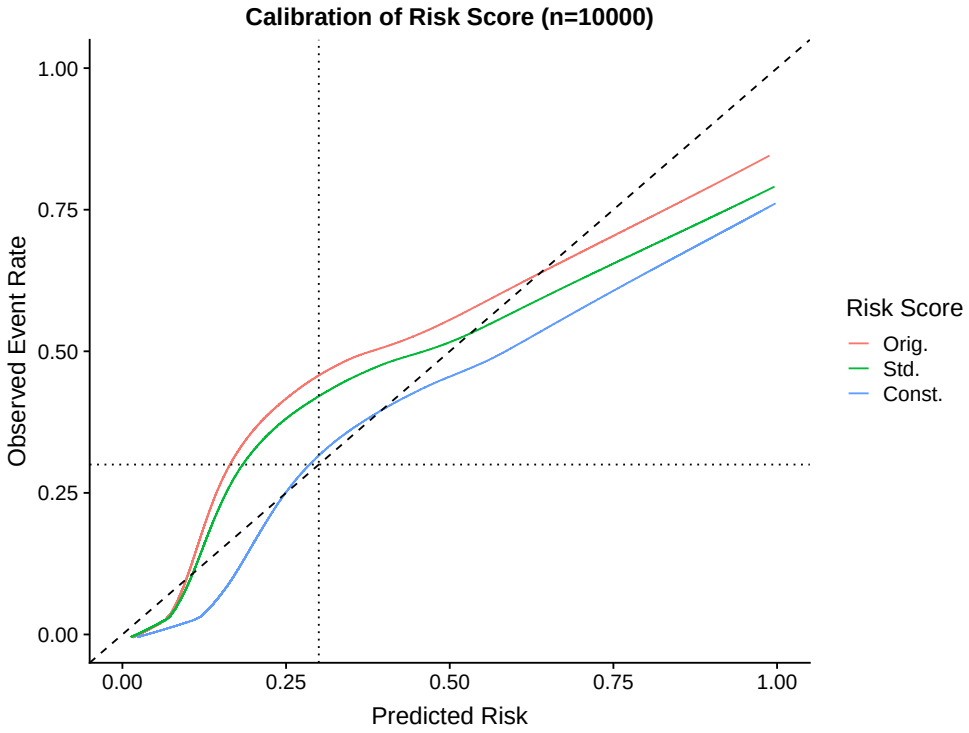


Figure 4.4: Calibration curves comparing standard logistic recalibration and constrained logistic recalibration for simulation example 3. Recalibration parameters were estimated in a training set of size 10,000.

Table 4.3: Comparison of Constrained Recalibration to Other Recalibration Methods for Simulation Example 3.

Training Sample Size	Measure	Orig. (No Recal.)	Std. Recal.	Wt. Recal.	Const. Recal.
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.40, 1.16)	(1.60, 1.59)	(0.96, 1.15)
	sNB	0.430	0.457	0.499	0.501
	TPR	0.581	0.634	0.745	0.751
	FPR	0.083	0.097	0.135	0.138
	Prop. Assigned Intervention	0.18	0.20	0.25	0.26
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.34, 1.12)	(1.46, 1.45)	(0.96, 1.12)
	sNB	0.430	0.455	0.503	0.504
	TPR	0.581	0.629	0.759	0.761
	FPR	0.083	0.096	0.140	0.141
	Prop. Assigned Intervention	0.18	0.20	0.26	0.26
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.19, 1.07)	(0.19, 1.07)	(0.57, 1.08)
	sNB	0.430	0.445	0.445	0.481
	TPR	0.581	0.608	0.609	0.688
	FPR	0.083	0.090	0.090	0.114
	Prop. Assigned Intervention	0.18	0.19	0.19	0.22
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.22, 1.10)	(0.22, 1.10)	(0.36, 1.11)
	sNB	0.430	0.446	0.446	0.458
	TPR	0.581	0.610	0.610	0.636
	FPR	0.083	0.090	0.090	0.098
	Prop. Assigned Intervention	0.16	0.18	0.18	0.20

4.3.4 Simulation Example 4: Underestimation of Risk Across all Risk Levels

Last, we present an example where predicted risks are consistently underestimated. As shown in Table 4.4, sNB is 1.8% higher under constrained logistic recalibration compared to standard logistic recalibration for training sample size of $n = 10,000$. Figure 4.5 shows the calibration curves for standard logistic recalibration and constrained logistic recalibration. Though constrained logistic recalibration does not result in perfect calibration at the risk threshold, there is improved calibration at the risk threshold compared to standard logistic recalibration.

As sample size decreases the improvement in net benefit over standard logistic recalibration decreases. When $n = 1,000$, the recalibration parameters estimated under constrained recalibration are the same as those estimated under standard logistic recalibration. Therefore estimated standardized net benefit of the logistic recalibrated risk score, in the training

Table 4.4: Comparison of Constrained Recalibration to Other Recalibration Methods for Simulation Example 4.

Training Sample Size	Measure	Orig. (No Recal.)	Std. Recal.	Wt. Recal.	Const. Recal.
n =10,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.75, 0.89)	(0.99, 0.96)	(0.97, 0.90)
	sNB	0.475	0.511	0.521	0.525
	TPR	0.546	0.615	0.640	0.650
	FPR	0.052	0.077	0.088	0.092
	Prop. Assigned Intervention	0.17	0.21	0.22	0.23
n =5,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.64, 0.82)	(0.86, 0.88)	(0.75, 0.82)
	sNB	0.475	0.511	0.523	0.524
	TPR	0.546	0.619	0.644	0.647
	FPR	0.052	0.080	0.089	0.091
	Prop. Assigned Intervention	0.17	0.21	0.22	0.22
n =1,000	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.51, 0.84)	(0.52, 0.84)	(0.51, 0.83)
	sNB	0.475	0.491	0.491	0.491
	TPR	0.546	0.581	0.581	0.581
	FPR	0.052	0.066	0.066	0.066
	Prop. Assigned Intervention	0.17	0.19	0.19	0.19
n =500	$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(0.77, 0.90)	(0.86, 0.93)	(0.99, 0.94)
	sNB	0.475	0.510	0.513	0.524
	TPR	0.546	0.616	0.624	0.647
	FPR	0.052	0.077	0.082	0.091
	Prop. Assigned Intervention	0.17	0.21	0.21	0.22

set, is within one standard error of $\widehat{sNB}_{max}$.

Both weighted and constrained logistic recalibration yield risk scores with larger sNB compared to standard logistic recalibration for all training sample sizes, except $n = 1000$. Weighted recalibration and constrained logistic recalibration tend to perform similarly, with slightly higher sNB under the constrained approach. For the smallest training sample size, $n = 500$, the constrained logistic recalibration approach has over 1% higher sNB compared to standard recalibration, while the weighted approach has only 0.3% higher sNB compared to standard logistic recalibration.

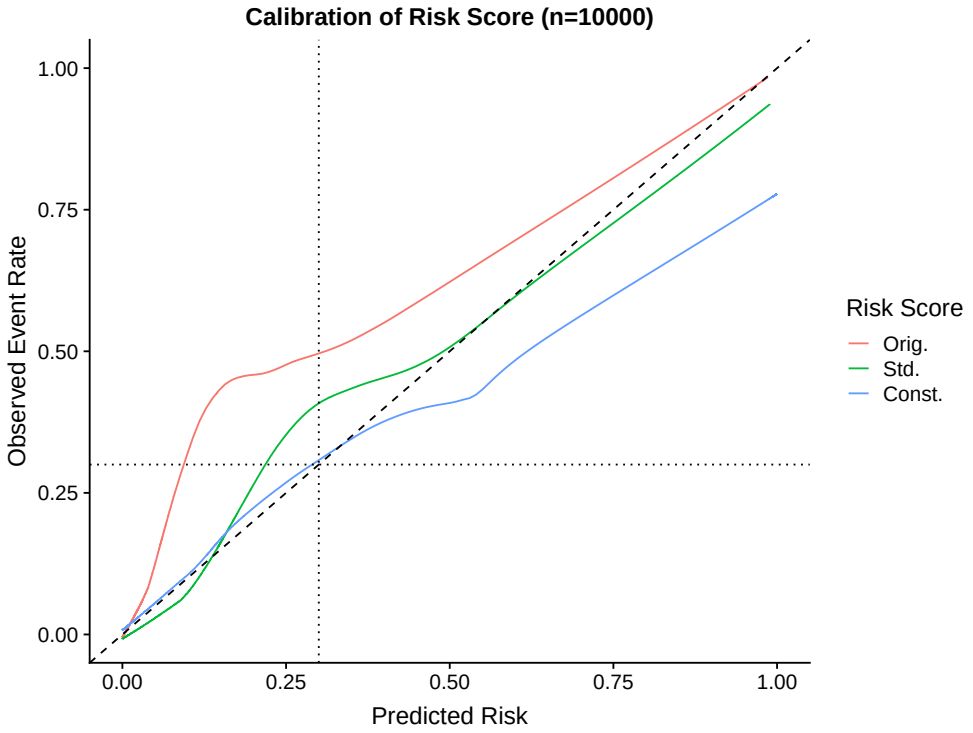


Figure 4.5: Calibration curves comparing standard logistic recalibration and constrained logistic recalibration for simulation example 4. Recalibration parameters were estimated in a training set of size 10,000.

4.4 Application to ACC-AHA-ASCVD Risk Score in MESA Cohort

As in Chapter 3, we compare recalibration methods for the ACC-AHA-ASCVD risk score for predicting 10-year risk of atherosclerotic cardiovascular disease (ASCVD) in MESA cohort. Recall that ACC and AHA guidelines recommend statins for individuals with 10-year risk of ASCVD greater than 7.5%. DeFilippis et al. (2015) showed that the ACC-AHA-ASCVD risk score overestimated risk in the the MESA cohort.

We compare the standard and constrained logistic recalibrated ACC-AHA-ASCVD risk score using the MESA cohort for recalibration. Following the treatment guidelines and clinical practice we use a risk threshold of 7.5%. As in Chapter 3, we additionally consider risk thresholds of 5% and 15% to illustrate the method. To correct for optimistic bias incurred by estimating the recalibration parameters and sNB in the same sample, we use bootstrap optimization bias correction methods (Harrell, 2015). Bootstrap confidence intervals were used to obtain 95% confidence intervals for $\widehat{sNB}$. For all bootstrap methods we used 500 replications. We additionally examine application of recalibration methods in the Black, male MESA cohort and present results in Appendix B.

Table 4.5 shows the results of applying standard, weighted, and constrained recalibration methods. Figures 4.6 - 4.8 show the calibration curves after standard and weighted recalibration for the different risk thresholds. For all risk thresholds, the constrained logistic recalibrated risk score has similar $\widehat{sNB}$ and calibration at the risk threshold as the standard logistic recalibrated risk score. Standard recalibration tends to focus efforts of recalibration near the the bulk observations lie. As discussed in Chapter 3, the type of miscalibration near the risk thresholds, 5% and 7.5%, is similar to the type of miscalibration where the bulk of observations lie, so again standard recalibration can sufficiently correct estimates of recalibration near the risk threshold.

Unlike the weighted approach, constrained logistic recalibration does not show much improvement over standard recalibration when the risk threshold is $R = 0.15$. This result is related to the lower bound used in the constraint function. Recall, under constrained recalibration the parameter space is constrained to α parameters that are within 1 standard error of the estimated maximum achievable sNB. When $R = 0.15$ the estimated maximum achievable sNB is 0.007, with corresponding estimated standard error 0.013. Therefore, the lower bound used in the constraint is -0.006, which includes the sNB estimated under standard recalibration. Using a less conservative rule for the lower bound constraint (e.g. α parameters with $\widehat{sNB}$ within 0.5 standard error of maximum) could produce a risk score that differs from the standard recalibrated risk score and with potentially higher sNB

Table 4.5: Comparison of recalibration methods in MESA for RAW = 0.01, for different risk thresholds. An indicator weight is used for the weighted recalibration approach

Measure	Orig.	Std. Recal.	Wt. Recal.	Const. Recal.
		R = 0.05		
$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-1.08, 0.81)	(-0.90, 0.88)	(-1.09 0.80)
Effective Sample Proportion %	100	100	76	100
$\widehat{sNB}$ (95% CI)[1]	0.274 (0.184, 0.364)	0.333 (0.258, 0.404)	0.332 (0.254, 0.400)	0.333 (0.254, 0.406)
$\widehat{sNB}^2$	-	0.324 (0.255 0.401)	0.329 (0.250, 0.396)	0.331 (0.251, 0.404)
$\widehat{TPR}^2$	0.873	0.714	0.717	0.714
$\widehat{FPR}^2$	0.442	0.374	0.379	0.374
		R = 0.075		
$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-1.08, 0.81)	(-0.67, 0.99)	(-1.21, 0.75)
Effective Sample Proportion %	100	100	53	100
$\widehat{sNB}$ (95% CI)[1]	0.073 (-0.035, 0.181)	0.153 (0.065 0.240)	0.161 (0.074 0.246)	0.152 (0.068, 0.232)
$\widehat{sNB}^2$	-	0.152 (0.063, 0.239)	0.157 (0.069, 0.242)	0.150 (0.074, 0.237)
$\widehat{TPR}^2$	0.771	0.470	0.518	0.448
$\widehat{FPR}^2$	0.584	0.201	0.228	0.188
		R = 0.15		
$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-1.08, 0.81)	(-1.32, 0.63)	(-1.06, 0.82)
Effective Sample Proportion %	100	100	13	100
$\widehat{sNB}$ (95% CI)[1]	-0.200 (-0.312, -0.086)	-0.004 (-0.028, 0.032)	0.005 (-0.012, 0.036)	-0.005 (-0.004, 0.036)
$\widehat{sNB}^2$	-	-0.007 (-0.032, 0.029)	-0.001 (-0.019, 0.030)	-0.009 (-0.012 0.029)
$\widehat{TPR}^2$	0.449	0.053	0.029	0.054
$\widehat{FPR}^2$	0.584	0.017	0.008	0.019

[1]Delta-method derived std error used for original risk score 95% CI calculation.
Bootstrap used with 500 replications used for standard and weighted recalibration
[2] Optimism corrected estimates using 500 bootstrap replications

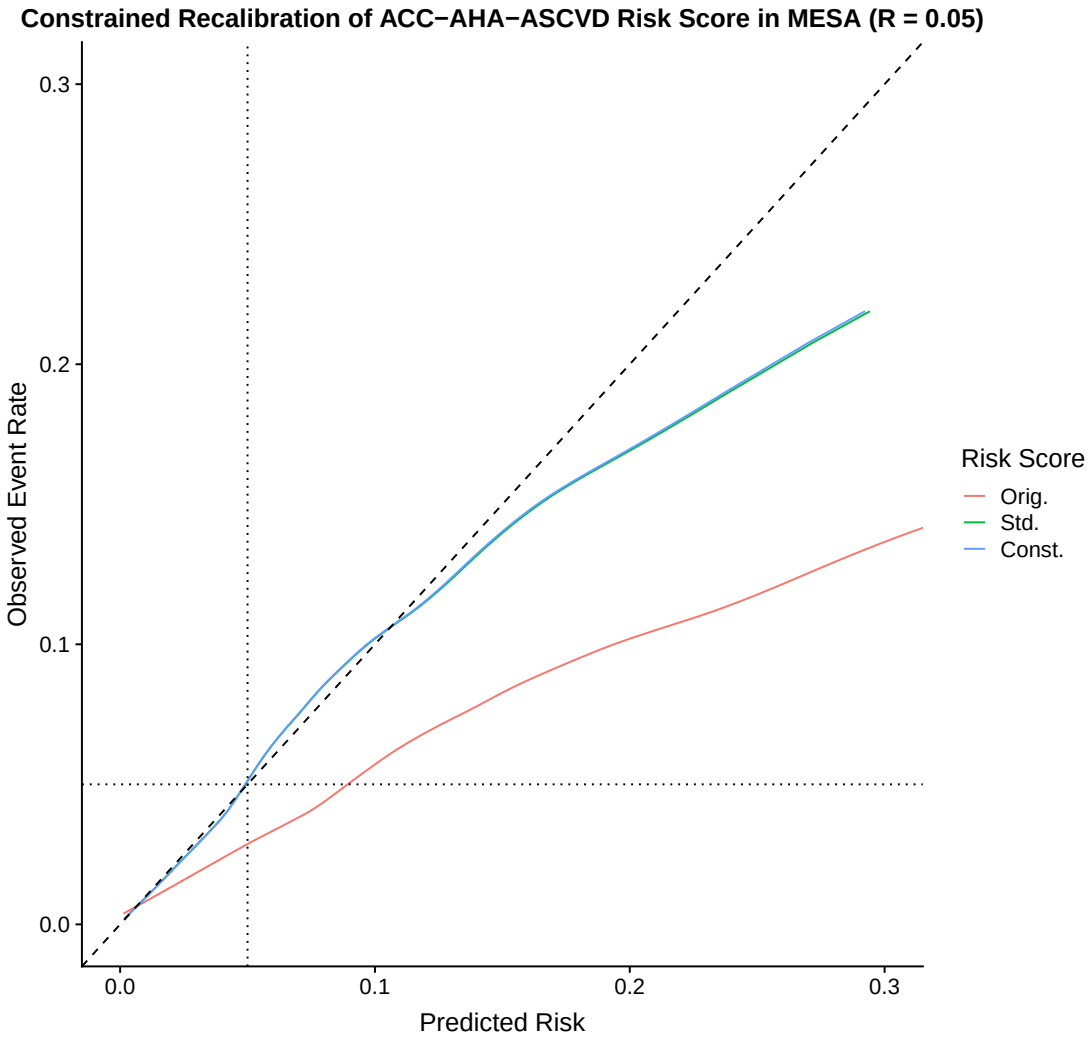


Figure 4.6: Calibration curve of the ACC-AHA-ASCVD risk score in MESA cohort, after standard and weighted recalibration approaches. The dotted line indicates the **risk threshold of 5%**.

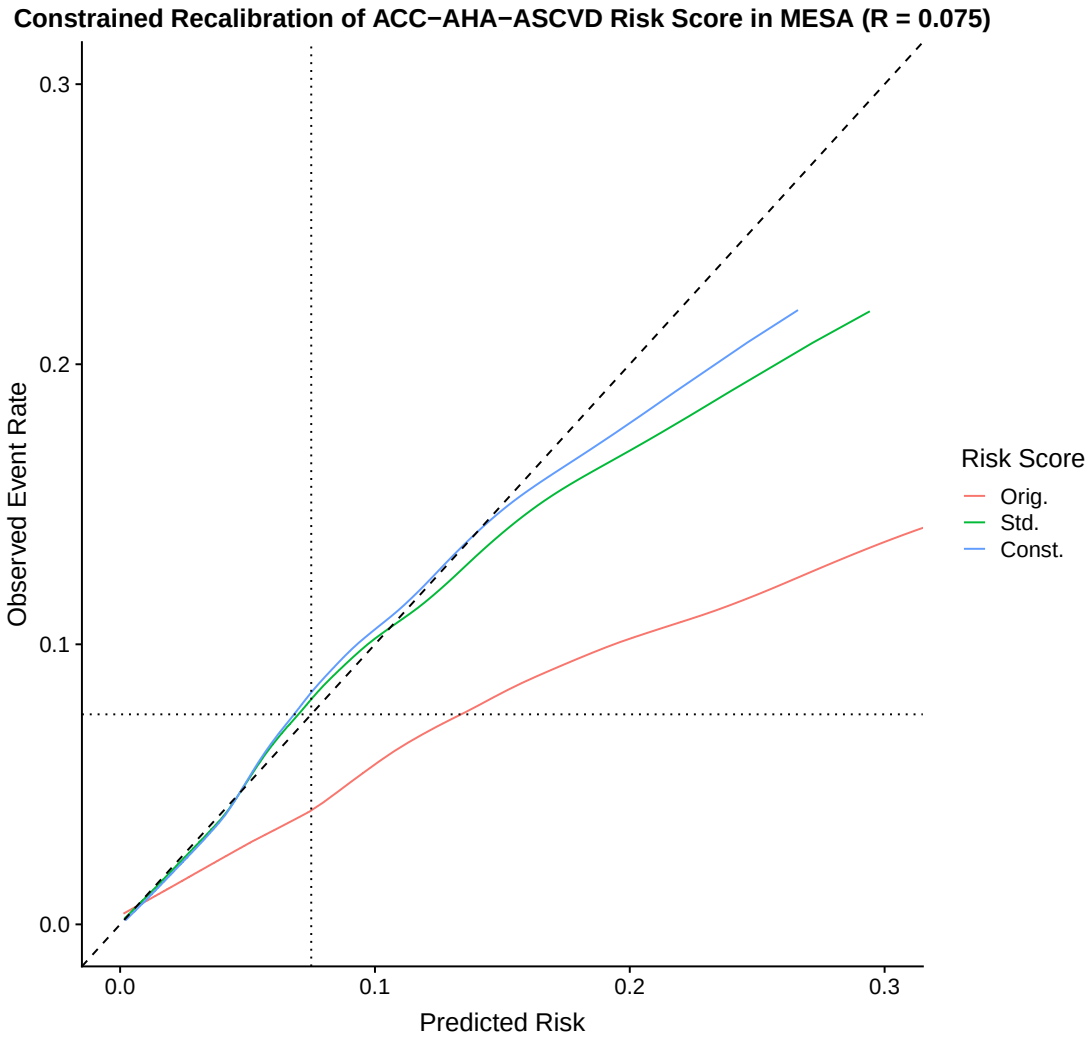


Figure 4.7: Calibration curve of the ACC-AHA-ASCVD risk score in MESA cohort, after standard and weighted recalibration approaches. The dotted line indicates the **risk threshold of 7.5%**.

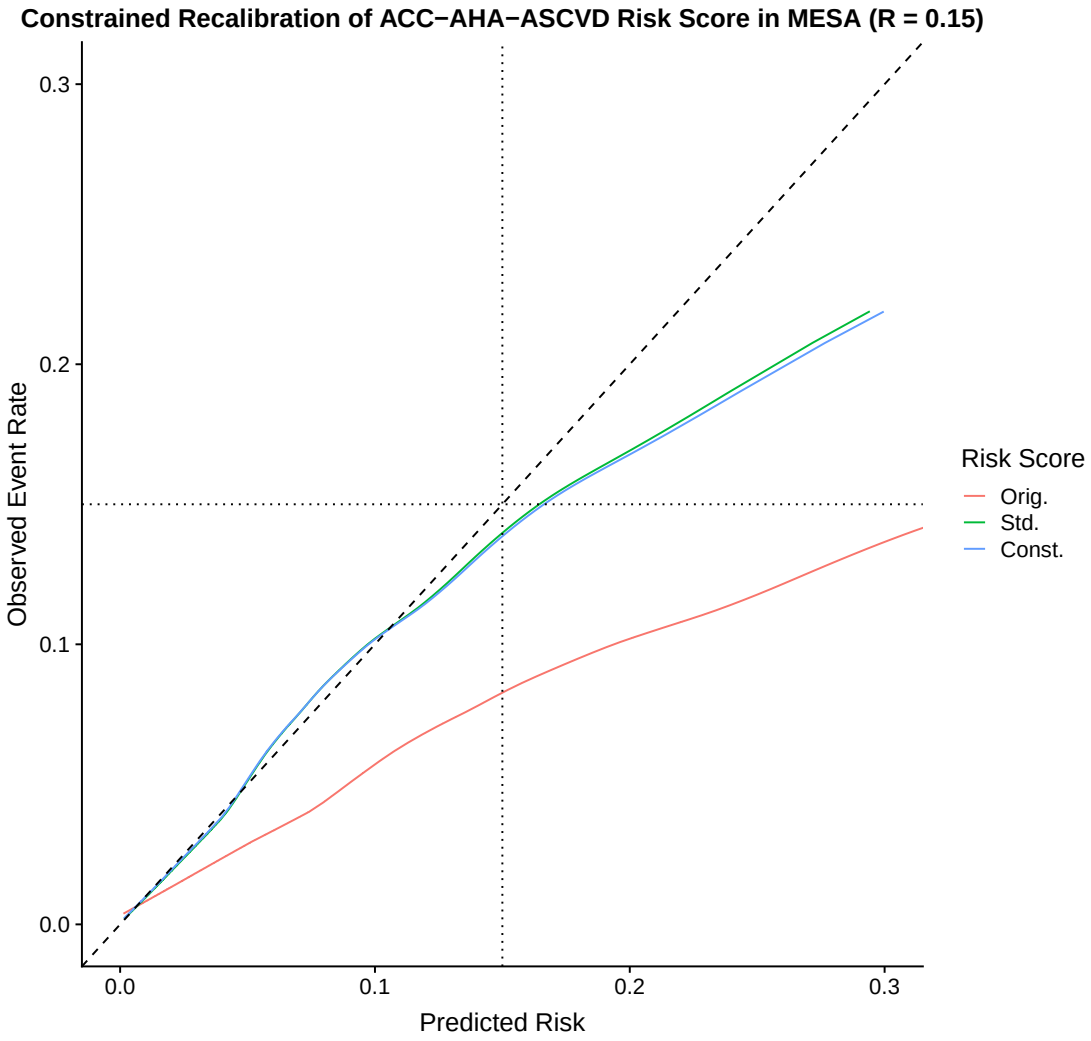


Figure 4.8: Calibration curve of the ACC-AHA-ASCVD risk score in MESA cohort, after standard and weighted recalibration approaches. The dotted line indicates the **risk threshold of 15%**.

4.5 Discussion

We presented a method for recalibration that aims to produce risk models with increased standardized net benefit. Improved clinical utility of a risk model also results in improved calibration at the risk threshold, where decision-making has the highest clinical impact. We considered different miscalibration patterns in our examples using simulated data and found the constrained approach can produced sizable improvements in the clinical utility of recalibrated risks compared to standard logistic recalibration. We also applied the constrained approach to the ACC-AHA-ASCVD risk model using the MESA study data, and found that clinical utility of the risk score improved when applying the proposed method compared to standard logistic recalibration.

In some instances standard logistic recalibration works well, such as when there is a large proportion of observations with event rate near the risk threshold, or when standard recalibration is able to achieve good calibration across all levels of risk. For these settings, the standardized net benefit under standard logistic recalibration will likely be close to the maximum achievable net benefit under any form of logistic recalibration. Under the proposed framework, the constrained logistic solution can revert to the logistic recalibration solution in these situations. For this reason, logistic recalibration can be thought of a special case of the constrained approach.

Unlike the approach presented in Chapter 3, the constrained approach does not apply any weighting and therefore there is no decrease in the effective sample size. This can be useful in problems were the sample size (or event rate) is low, and down-weighting may be undesirable. However, the weighted approach from Chapter 3 can provide additional flexibility in targeted calibration, such as targeting calibration over a region of risks as opposed to specifically at the risk threshold. The objective function in the constrained approach could be augmented to include weights, combining the both proposed recalibration methods. Further study is needed to assess how these approaches could be combined.

Chapter 5

COMBINING BIOMARKERS TO MAXIMIZE NET BENEFIT

5.1 Introduction

For many diseases, multiple risk factors are predictive of the outcome. Evaluating each of these risk factors separately makes the assessment of risk difficult, and makes the decision-making process murky. A single measure that combines risk factors can simplify the decision-making process, especially if paired with a risk threshold that yields a decision rule. Such decision rules are helpful for identifying which patients should be prescribed a particular intervention.

Often, risk scores are built using logistic regression models (Steyerberg and Vergouwe, 2014). After the model is constructed, it may be paired with a risk threshold to create a decision rule with some desirable properties (e.g. certain sensitivity or specificity). In settings where the risk score will ultimately be used to prescribe an intervention, it is desirable to account for this knowledge of the intervention's harms and benefits when developing the linear combination. Logistic regression, which is fit by maximizing the binomial log-likelihood, does not do this.

We propose a method for finding linear combinations of risk markers by directly maximizing standardized net benefit (sNB). Standardized net benefit is a measure of the population impact of a treatment policy that accounts for the harms and benefits of the intervention. Without making any distributional assumptions, we directly maximize a smooth approximation of sNB. Through simulation studies we show the direct maximization method can produce linear combinations of risk markers with higher sNB than conventional risk model-

ing methods for a range of data settings. We apply our method to find a linear combination of risk markers that can be used to screen for diabetes in Pima Indian women.

The rest of the chapter is organized as follows. In Section 5.2 we review the relevant literature on methods for constructing linear combinations of risk markers. In Section 5.3 we present the proposed direct maximization method and give details for implementation. In Section 5.4 we give the results of our simulation study and in Section 5.5 we present the results for the diabetes example. We close with a discussion of the method, limitations, and areas of future study in Section 5.6.

5.2 Background

There is a rich literature on methods for combining risk markers to produce a univariate risk score. Combinations of risk markers are sometimes referred to as *composite biomarkers* (FDA, 2014). Su and Liu (1993) presented a seminal result; under settings where risk markers have disease-conditional bivariate normal distributions with equal variance, the Fisher linear discriminant has highest sensitivity across all ranges of specificity compared to any other linear combination (i.e., has dominating ROC curve). McIntosh and Pepe (2002) give a more general result that risk score based on markers $\mathbf{X}$ (i.e., $P(Y = 1|\mathbf{X})$) is the optimal way to combine risk markers $\mathbf{X}$, in the sense that the risk score yields the highest sensitivity for any given specificity. In practice, we do not know the true form of $P(Y = 1|\mathbf{X})$, so finding methods for combining biomarkers that have good operating characteristics is a question of interest.

Though logistic regression is often used to develop risk models, methods that maximize particular metrics of discrimination (AUC, TPR/FPR, partial area under the ROC curve, etc.) may be appealing since the optimization criterion and evaluation criterion are the same. Area under the ROC curve (AUC) is often used as a target metric and several methods have been proposed to find combinations of risk markers by directly maximizing AUC (Pepe et al.,

2006; Ma and Huang, 2007).

In this work we consider settings in which a composite risk marker will be used with a threshold to decide whether or not to recommend an intervention. In these settings, it may be of interest to only consider composite markers that achieve some sensitivity or specificity that are relevant to the clinical setting. For this reason, partial area under the ROC curve ($pAUC$) may be considered a useful target measure for finding linear combinations of risk markers. $pAUC$ summarizes the area under the ROC curve for FPR within a predefined range (McClish, 1989). Relying on assumptions of normality, Hsu and Hsueh (2013) analytically derived the derivative of $pAUC$. Given the complex form of the derivative the authors use an algorithmic approach to find linear combinations of risk markers that maximize $pAUC$. An obvious drawback of parametric approaches is that they may perform poorly when their distributional assumptions fail to hold. Pepe and Thompson (2000) proposed a distribution-free grid search for finding linear combinations of markers that maximize $pAUC$. However, this method is computationally prohibitive when more than two markers are considered. Smooth approximations of AUC and $pAUC$ have been proposed to avoid parametric assumptions without incurring prohibitive computational costs (Lin et al., 2011). Fong et al. (2016) and Yan et al. (2018) used a kernel based approximation of $pAUC$ as a target measure. However, the authors note that these methods may be sensitive to issues of local maxima. Liu et al. (2011) used a "min-max" approach for finding linear combinations of markers that maximize $pAUC$. The min-max approach aims to reduce the dimensionality of the linear combination problem by finding a linear combination of $\min(\mathbf{X})$ and $\max(\mathbf{X})$, where $\mathbf{X}$ is the set of available risk markers. This dimension reduction may be unappealing because only a subset of risk markers remain in the model.

A drawback of targeting $pAUC$, is that a decision rule is not obtained since a threshold for prescribing intervention is not defined (McIntosh and Pepe, 2002). Comparatively, methods that optimize TPR directly produce both a linear combination of markers and a

threshold, resulting in a decision rule. Gao et al. (2008) proposed a parametric method for finding linear combinations of markers by maximizing TPR for a fixed FPR. Meisner et al. (2017) proposed a distribution-free method for finding linear combinations of risk markers that maximize a smooth approximation of TPR while constraining FPR to be below some acceptable upper bound.

Finally, some authors proposed methods for finding linear combinations of risk markers by maximizing the Youden's index (Yin and Tian, 2014; Xu et al., 2015). The Youden's index estimates the difference in TPR and FPR at a given threshold, t, (Youden, 1950). Notably, the Youden's index is a special case of sNB, when the harm-benefit ratio equals the odds of the outcome (i.e. $\frac{C}{B} = \frac{P(Y=1)}{P(Y=0)}$). Following the approach by Liu et al. (2011), Yin and Tian (2014) used a min-max approach to find a linear combination of markers that optimize the Youden's index. The authors also proposed a kernel-based method. Since a normal kernel function is used the authors found that it did not perform well when the underlying data distribution is far from normal. Xu et al. (2015) extended the kernel-based method to find non-linear combinations of biomarkers that maximize the Youden's index. Hsu and Hsueh (2013) noted that linear combinations are preferable over non-linear combinations because they are easier to interpret.

In many clinical settings a composite risk marker is derived when an intervention is already known. Additionally, the intervention may have some known expected benefit when prescribed for a case, and known expected harm when prescribed to a control. It is desirable to explicitly account for the benefits and harm when building the linear combination. We therefore propose a method that uses standardized net benefit (sNB) as the objective function. We refer the reader to Chapter 3 for a review of standardized net benefit.

5.3 Methods

Let $\mathbf{X}$ be a p-dimensional vector of risk markers that are of interest for predicting binary outcome Y. We call observations with $Y = 1$ cases, and with $Y = 0$ controls. We are interested in constructing a linear combination of markers, $\theta^T\mathbf{X}$, that is predictive of Y and can be used to help recommend an intervention. For example, in the Pima Indian diabetes application we are interested in constructing a composite risk marker that is predictive of diabetes. Those who are at increased risk for diabetes will be prescribed some intervention, such as behavior or diet modification, medication, or some combination thereof (Bantle et al., 2008). Further, suppose there is some ratio, $\frac{C}{B}$, that represents the harms and benefits of the intervention. C represents the harms of prescribing intervention to a control, and B represents the benefit of prescribing intervention to a case. Given an intervention and known harm-benefit ratio, we are interested in developing a decision rule that incorporates this knowledge.

We propose a method for constructing linear combinations of risk markers, $\theta^T\mathbf{X}$ that maximizes standardized net benefit. Particularly, we are interested in estimating parameters (θ, t), where θ is a p-dimensional vector of marker coefficients and t is the decision threshold. The linear combination, $\theta^T\mathbf{X}$ paired with threshold, t, give a decision rule for prescribing intervention. The following expression shows standardized net benefit expressed as a function parameter (θ, t),

$$sNB(\theta, t) = P\left(\theta^T\mathbf{X} > t | Y = 1\right) - \frac{C}{B}\frac{P(Y=0)}{P(Y=1)}\left(\theta^T\mathbf{X} > t | Y = 0\right), \tag{5.1}$$

for given harm-benefit ratio $\frac{C}{B}$. The empirical estimate of (5.1) is

$$\widehat{sNB}(\theta, t) = \frac{1}{n_D}\sum_{i=1}^{n_D}\mathbf{1}(\theta^T X > t) - \frac{C}{B}\frac{\hat{P}(Y=0)}{\hat{P}(Y=1)}\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}}\mathbf{1}(\theta^T X > t), \tag{5.2}$$

where n_D and $n_{\bar{D}}$ are the number of cases and controls, respectively.

One could find a linear combination that accounts for the clinical context of the risk model by directly maximizing equation (5.2). However, this is a difficult task given that (5.2) is a non-smooth function and therefore derivative-based methods are infeasible. A common solution to this problem is to use a smooth approximation of the objective function (Lin et al., 2011; Meisner et al., 2017; Yan et al., 2018). Following the methodology of Lin et al. (2011) and Meisner et al. (2017) we propose using the following smooth approximation

$$\widetilde{sNB}(\theta, t) = \frac{1}{n_D} \sum_{i=1}^{n_D} \Phi\left(\frac{\theta^T X_i - t}{s_n}\right) - \omega \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n_{\bar{D}}} \Phi\left(\frac{\theta^T X_i - t}{s_n}\right), \tag{5.3}$$

where s_n is a tuning parameter that controls the approximation to an indicator function, and $\Phi(x)$ is the standard normal cumulative distribution function. As s_n approaches zero, $\Phi(x/s_n)$ more closely approximates $\mathbf{1}[x > 0]$. We discuss choice of tuning parameter in more detail below. To simplify expressions we use $\omega = \frac{C}{B} \frac{\hat{P}(Y=0)}{\hat{P}(Y=1)}$.

We propose finding a linear combination of risk markers by directly optimizing expression (5.3). sNB is ranked-based and invariant to scalings θ. Therefore, we include the constraint $||\theta||^2 = 1$ (Meisner et al., 2017; Fong et al., 2016). Given these considerations we estimate (θ, t) by solving the optimization problem

$$(\hat{\theta}, \hat{t}) = \underset{\theta \in \mathbb{R}^p, t \in \mathbb{R}}{\arg\max} \ \widetilde{sNB}(\theta^T X, t) \tag{5.4}$$

$$\text{subject to} \ \ ||\theta||^2 = 1. \tag{5.5}$$

This problem can be solved using gradient-based optimization methods, such as those implemented in the the `Rsolpn` package in `R`. Initial values are required for software implementation. Fong et al. (2016) and Meisner et al. (2017) proposed using normalized coefficients estimated from robust logistic regression for initial values (Bianco and Yohai, 1996). In simulations studies we found instances in which standard logistic regression produces linear

combinations with higher sNB than robust logistic regression, and vice versa. For this reason, we propose that first sNB be estimated for the risk scores obtained from both standard and robust logistic regression. For the proposed method, we choose initial values from either standard or robust logistic regression, depending on which method produces a risk model with larger estimated sNB. Initial values are rescaled to satisfy the constraint. Estimating sNB for a fitted regression requires a risk threshold. To find such a threshold for models fit with standard or robust logistic regression, we solve the one-dimensional optimization problem

$$t_\beta = \underset{t\in[0,1]}{\arg\max}\ \widetilde{sNB}(t|\text{expit}(\beta^T\mathbf{X})) \tag{5.6}$$

$$= \underset{t\in[0,1]}{\arg\max}\ \frac{1}{n_D}\sum_{i=1}^{n_D}\Phi\left(\frac{\text{expit}(\beta^T X_i)-t}{s_n}\right) - \omega\frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}}\Phi\left(\frac{\text{expit}(\beta^T X_i)-t}{s_n}\right). \tag{5.7}$$

Let (θ_0, t_0) represent the initial values required for the proposed method. The normalized coefficients from the logistic regression model that has larger $\widehat{sNB}$ are used as θ_0. Once the θ_0 is found, then it is left to find t_0. Similar to solving the one-dimensional optimization problem in (5.7), t_0 can be found by finding the t that maximizes $\widetilde{sNB}(t|\theta_0^T\mathbf{X})$.

Finally, we must choose the tuning parameter s_n for the smooth approximation of empirical estimator. Lin et al. (2011) and Meisner et al. (2017) proposed using a tuning parameter of the form $s_n = \frac{\hat{\sigma}(\theta_0^T\mathbf{X})}{n^h}$, where $\hat{\sigma}(\theta_0^T\mathbf{X})$ is the estimated standard deviation of the linear combination obtained using initial values, θ_0. As h increases, s_n decreases, and the smooth function more closely approximates the empirical measure. In simulation study, we considered tuning parameters with $h = \frac{1}{3}, \frac{1}{2}$, and 1. We found that $h = \frac{1}{2}$ yields the best approximation without resulting in convergence issues. This is the same tuning parameter used in Meisner et al. (2017).

5.4 Simulation Study

In this section we present compare the proposed method to standard and robust logistic regression using simulated data. We consider the three following simulation settings: (i) a bivariate marker drawn from disease-conditioned normal distributions, (ii) markers drawn from distributions with propensity for outliers, and (iii) markers drawn from skewed distributions, specifically log-normal distributions.

The proposed direct maximization method estimates a decision rule, by estimating both coefficients (θ) of the linear combination and the decision threshold (t). Standard or robust logistic regression, alone, does not produce a decision rule without a threshold. In order to compare sNB of the composite risk marker obtained from the different methods, we estimate a decision threshold t for the linear combination constructed by logistic regression in two ways. First, we estimate the threshold t as a function of the harm-benefit ratio, relying on the relationship that $\frac{C}{B} = \frac{t}{1-t}$ (Pauker and Kassirer, 1980). We refer to this as $\hat{t}_{CB}$, and expect this to be a good decision threshold if the risk model produced by logistic regression is well calibrated. Second, we estimate t by directly optimizing $\widetilde{sNB}(t|\hat{\beta}^T\mathbf{X})$, where $\hat{\beta}^T\mathbf{X}$ is the estimated linear combination from standard or robust logistic regression. We refer to this estimate as $\hat{t}_{opt}$.

For all simulation scenarios, linear combinations and decision thresholds are estimated in training sample sizes of $n = 500, 800, 1000$ and 5000 to consider both small and large sample size properties. We use a large independent validation dataset of size $n = 100,000$ to evaluate the true (rather than estimated) performance of composite risk marker and decision rule obtained from the proposed direct maximization method. Means and standard deviation of sNB are obtained from 500 Monte Carlo simulations. For each example we fix the harm-benefit ratio such that such that $\omega = \frac{C}{B}\frac{\hat{P}(Y=0)}{\hat{P}(Y=1)}$ is approximately 0.5, 1, and 2. $\frac{\hat{P}(Y=0)}{\hat{P}(Y=1)}$ varies across simulation replications, whereas $\frac{C}{B}$ is fixed. Percentage comparisons of sNB, TPR,

and FPR are additive (rather than multiplicative) differences.

5.4.1 Simulation 1: Bivariate Normal

We first consider two instances of the disease-conditioned bivariate normal data generating mechanism. In both settings instances we use a prevalence of 10% and simulate outcome data from $Y \sim Bern(0.1)$. We use two markers to predict the outcome. First, we consider the setting where markers for cases are drawn from a bivariate normal distribution with mean $\vec{\mu}_D = (1, 1)$, and identity covariance matrix. The markers for controls are drawn from a standard bivariate normal distribution. Results are presented in Appendix C. Under this setting the logistic-linear model is correctly specified and produces the optimal linear combination of risk markers (McIntosh and Pepe, 2002). Therefore, we do not expect the proposed method to find a linear combination of markers that out-performs the risk score obtained from logistic regression. Because the true optimal linear combination of markers is known in this setting, we view this example as providing evidence that the method produces good results.

We also consider a setting where the disease-conditioned covariance matrices are non-proportional, and therefore the logistic-linear does not hold. We sample markers from

$$\mathbf{X}|Y = 1 \sim N\left(\begin{bmatrix}2.0\\2.0\end{bmatrix}, \begin{bmatrix}9.0 & 12.6\\12.6 & 36.0\end{bmatrix}\right) \quad \mathbf{X}|Y = 0 \sim N\left(\begin{bmatrix}0.0\\0.0\end{bmatrix}, \begin{bmatrix}7.3 & 3.8\\3.8 & 4.0\end{bmatrix}\right).$$

Table 5.1 compares methods under this setting when $\frac{C}{B} = \frac{1}{9}$, so $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)}$ is on average 1. When $n = 5,000$, robust logistic regression performed better than standard logistic regression in 100% of simulation replications. Therefore, initial values for the direct maximization method were taken from robust logistic regression. For the largest sample size, the proposed direct maximization method and robust logistic regression using t_{opt} produce risk scores with the same average sNB. As sample size decreases, robust logistic regression using t_{opt}

performs the best of the three methods. Results for $\frac{C}{B} = \frac{2}{9}$ ($\omega \approx 2$) and $\frac{C}{B} = \frac{0.5}{9}$ ($\omega \approx 0.5$) are shown in Appendix C. Results are similar for $\omega \approx 2$. When the harm-benefit ratio is $\frac{C}{B} = \frac{0.5}{9}$, indicating increased preference for treating cases with higher harm for treating controls, the average sNB tends to be higher under direct maximization compared to either logistic regression method.

Table 5.1: Mean and (standard deviation) of sNB, TPR, and FPR, for non-proportional covariance bivariate normal setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 1$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB}).

	Direct Maximization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			sNB		
5000	0.363 (0.009)	0.312 (0.016)	0.270 (0.010)	0.363 (0.007)	0.311 (0.009)
1000	0.346 (0.040)	0.308 (0.034)	0.272 (0.023)	0.355 (0.025)	0.308 (0.029)
800	0.342 (0.038)	0.308 (0.039)	0.272 (0.028)	0.352 (0.032)	0.308 (0.032)
500	0.332 (0.049)	0.300 (0.054)	0.270 (0.043)	0.342 (0.046)	0.298 (0.051)
			TPR		
5000	0.431 (0.025)	0.460 (0.039)	0.607 (0.022)	0.430 (0.019)	0.531 (0.013)
1000	0.429 (0.070)	0.463 (0.074)	0.603 (0.046)	0.415 (0.036)	0.530 (0.032)
800	0.421 (0.073)	0.447 (0.075)	0.597 (0.051)	0.408 (0.041)	0.527 (0.035)
500	0.423 (0.099)	0.448 (0.102)	0.599 (0.073)	0.393 (0.056)	0.523 (0.055)
			FPR		
5000	0.068 (0.024)	0.148 (0.044)	0.337 (0.031)	0.067 (0.018)	0.221 (0.023)
1000	0.084 (0.087)	0.157 (0.090)	0.334 (0.064)	0.060 (0.028)	0.224 (0.051)
800	0.080 (0.078)	0.140 (0.087)	0.328 (0.070)	0.056 (0.030)	0.220 (0.051)
500	0.095 (0.112)	0.154 (0.111)	0.336 (0.093)	0.053 (0.035)	0.230 (0.071)

5.4.2 Simulation 2: Outliers

Next, we consider a simulation setting used in Fong et al. (2016) and Meisner et al. (2017), where outliers are present. Performance of standard logistic regression may be diminished in the presence of outliers; one expects robust logistic regression to be less sensitive to outliers. To induce a moderate case of outliers we first simulate a binary indicator $\Delta \sim Bern(0.1)$. Then we simulate risk marker $\mathbf{X}$

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = (1-\Delta) \times Z_1 + \Delta \times Z_2,$$

where

$$\mathbf{Z}_1 \sim N\left(\begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}, \begin{bmatrix} 0.20 & 0.18 \\ 0.18 & 0.20 \end{bmatrix}\right) \qquad \mathbf{Z}_2 \sim N\left(\begin{bmatrix} 0.00 \\ 0.00 \end{bmatrix}, \begin{bmatrix} 2.00 & 0.00 \\ 0.00 & 2.00 \end{bmatrix}\right)$$

Outcomes are drawn from $Y \sim Bern(p)$, where p are true risks generated from $p = \text{expit}(-1 + 4X_1 - 3X_2 - 0.8(X_1 - X_2)^3)$. Based on this mean model, the prevalence of disease is approximately 30%. Figure 5.1 shows the distribution of the two biomarkers by disease status. Note the slight bimodal nature of the distribution of markers among cases.

Table 5.2 compares sNB and discriminatory measures when $\frac{C}{B} = \frac{3}{7}$ ($\omega \approx 1$). For a training sample size of $n = 5000$, the average sNB of the composite marker obtained via direct maximization is about 2% higher compared to standard or robust logistic regression. For all training sample sizes, the average sNB under direct maximization is larger than the average sNB obtained from standard or logistic regression; however, as sample size decreases the improvement in sNB under direct maximization diminishes. Results for the same setting with $\frac{C}{B} = \frac{2}{7}$ ($\omega \approx 2$) and $\frac{C}{B} = \frac{0.5}{7}$ ($\omega \approx 0.5$) are shown in Appendix C. When $\omega \approx 2$ methods perform comparably. We also see lower variability in sNB for larger harm-benefit ratio.

Table 5.2: Mean and (standard deviation) of sNB, TPR, and FPR, for outlier simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 1$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB}).

	Direct Maximization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			sNB		
5000	0.187 (0.054)	0.166 (0.014)	0.168 (0.013)	0.168 (0.015)	0.170 (0.014)
1000	0.179 (0.057)	0.165 (0.032)	0.169 (0.030)	0.169 (0.035)	0.173 (0.034)
800	0.185 (0.063)	0.164 (0.036)	0.168 (0.034)	0.170 (0.043)	0.175 (0.042)
500	0.177 (0.064)	0.162 (0.045)	0.169 (0.042)	0.170 (0.054)	0.176 (0.051)
			TPR		
5000	0.581 (0.093)	0.584 (0.067)	0.601 (0.033)	0.584 (0.066)	0.600 (0.032)
1000	0.572 (0.153)	0.574 (0.116)	0.600 (0.077)	0.574 (0.112)	0.600 (0.074)
800	0.570 (0.161)	0.566 (0.129)	0.594 (0.091)	0.565 (0.126)	0.594 (0.088)
500	0.571 (0.184)	0.564 (0.150)	0.589 (0.110)	0.561 (0.148)	0.590 (0.107)
			FPR		
5000	0.403 (0.092)	0.427 (0.069)	0.443 (0.035)	0.424 (0.068)	0.439 (0.034)
1000	0.406 (0.154)	0.422 (0.121)	0.445 (0.082)	0.418 (0.117)	0.440 (0.078)
800	0.398 (0.167)	0.414 (0.129)	0.438 (0.094)	0.405 (0.124)	0.431 (0.090)
500	0.410 (0.191)	0.415 (0.149)	0.436 (0.113)	0.404 (0.147)	0.429 (0.110)

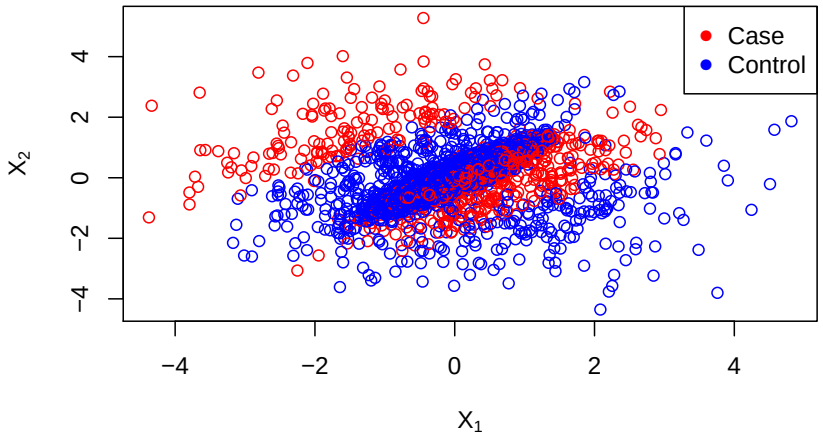


Figure 5.1: Distribution of bivariate risk markers for outliers scenario when $\beta = -1$, which results in approximately 30% prevalence.

5.4.3 Simulation 3: Skewed distributions

Yan et al. (2018) note that logistic regression may perform poorly (as measured by $pAUC$) when marker data come from skewed distributions. To explore this, we consider a setting where risk markers are drawn from a log-normal distribution. We simulate three risk markers, drawn from log-normal distributions conditioned on disease status. Two markers $\mathbf{X} = (X_1, X_2)$ are correlated and are informative of disease status.

$$\mathbf{X}|Y = 1 \sim LogNorm\left(\begin{bmatrix}1.10\\1.10\end{bmatrix}, \begin{bmatrix}0.04 & 0.09\\0.09 & 0.50\end{bmatrix}\right) \quad \mathbf{X}|Y = 0 \sim LogNorm\left(\begin{bmatrix}1.00\\1.00\end{bmatrix}, \begin{bmatrix}0.05 & 0.02\\0.02 & 0.05\end{bmatrix}\right).$$

The remaining risk marker, X_3 is uncorrelated to the previous two, and has the same distribution among cases and controls,

$$X_3|Y = 1 \equiv X_3|Y = 0 \sim LogNorm(1.65, 4.66).$$

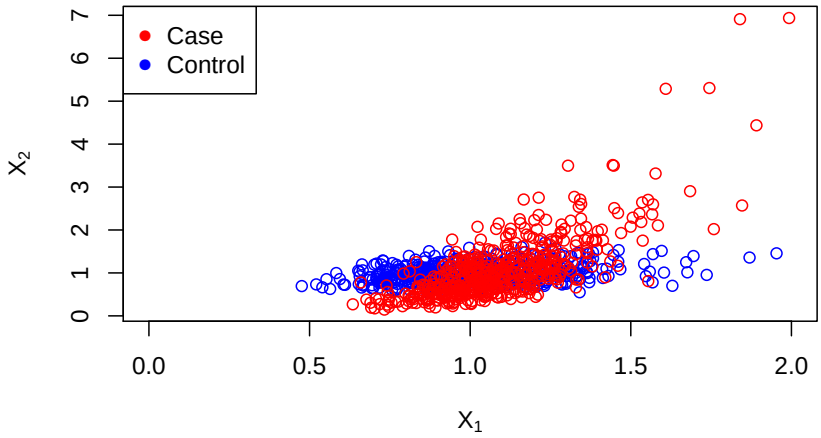


Figure 5.2: Distribution for the two log-normal risk markers (X_1, X_2) that are correlated with risk of disease. The prevalence is 50%.

The outcome Y follows as $Bern(0.5)$ distribution. This scenario is representative of settings where the composite risk marker may be used to help guide clinical decisions for high-risk populations, or settings where individuals who already have disease are being monitored for disease progression. Figure 5.2 shows the distribution of the first two markers by disease status. Note there is a large overlap between cases and controls.

Table 5.3 compares sNB and discriminatory measures when $\frac{C}{B} = 1$, so $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)}$ is on average 1. For all sample sizes, we see an higher average sNB for the linear combination produced by the direct maximization method compared to standard or robust logistic regression. The gains in average sNB are smaller for smaller sample sizes. For the smallest sample size there is still a 1% increase in average sNB. FPR is similar across all three methods, while TPR is, on average, highest under direct maximization. Results for the same setting with $\frac{C}{B} = 2$ ($\omega \approx 2$) are presented in Table 5.4. In this setting the cost of treating a control is larger that the benefit of treating a case. When $\omega \approx 2$ the average sNB of the linear

Table 5.3: Mean and (standard deviation) of sNB, TPR, and FPR, for skewed data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 1$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB}).

	Direct Maximization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			sNB		
5000	0.267 (0.018)	0.240 (0.019)	0.207 (0.018)	0.248 (0.019)	0.217 (0.019)
1000	0.255 (0.048)	0.237 (0.043)	0.207 (0.043)	0.244 (0.043)	0.217 (0.045)
800	0.253 (0.049)	0.234 (0.048)	0.205 (0.049)	0.242 (0.049)	0.215 (0.050)
500	0.248 (0.063)	0.231 (0.057)	0.204 (0.060)	0.239 (0.058)	0.215 (0.061)
			TPR		
5000	0.815 (0.043)	0.782 (0.041)	0.583 (0.028)	0.796 (0.039)	0.612 (0.028)
1000	0.783 (0.086)	0.764 (0.091)	0.582 (0.065)	0.780 (0.078)	0.610 (0.066)
800	0.778 (0.092)	0.760 (0.095)	0.580 (0.073)	0.777 (0.087)	0.607 (0.074)
500	0.763 (0.135)	0.749 (0.115)	0.580 (0.089)	0.763 (0.106)	0.606 (0.088)
			FPR		
5000	0.549 (0.038)	0.542 (0.037)	0.376 (0.019)	0.548 (0.035)	0.395 (0.019)
1000	0.531 (0.076)	0.530 (0.075)	0.379 (0.046)	0.538 (0.066)	0.397 (0.045)
800	0.527 (0.075)	0.528 (0.080)	0.379 (0.052)	0.537 (0.073)	0.396 (0.052)
500	0.520 (0.107)	0.522 (0.095)	0.382 (0.063)	0.528 (0.088)	0.398 (0.061)

combination obtained from direct maximization is larger than 0, while standard and robust logistic regression produce risk models with average sNB less than 0. A risk model with standardized net benefit less than 0 implies that a "treat-none" policy has clinical utility. Therefore, direct optimization produces a composite risk marker and risk threshold that has positive clinical utility, while logistic regression produces a composite marker that does not have clinical utility. Results for when $\frac{C}{B} = \frac{1}{2}$ ($\omega \approx 0.5$) are shown in Appendix C.

Table 5.4: Mean and (standard deviation) of sNB, TPR, and FPR, for skewed data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 2$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB}).

	Direct Maximization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$s\hat{N}B$		
5000	0.005 (0.030)	-0.005 (0.017)	-0.059 (0.013)	-0.010 (0.026)	-0.075 (0.012)
1000	0.026 (0.061)	-0.017 (0.031)	-0.058 (0.029)	-0.022 (0.039)	-0.072 (0.028)
800	0.024 (0.077)	-0.018 (0.033)	-0.058 (0.030)	-0.024 (0.039)	-0.072 (0.029)
500	0.018 (0.088)	-0.023 (0.040)	-0.056 (0.039)	-0.026 (0.043)	-0.069 (0.038)
			$T\hat{P}R$		
5000	0.008 (0.035)	0.011 (0.038)	0.093 (0.018)	0.018 (0.049)	0.114 (0.021)
1000	0.055 (0.104)	0.055 (0.101)	0.100 (0.042)	0.067 (0.128)	0.122 (0.050)
800	0.065 (0.115)	0.061 (0.112)	0.105 (0.049)	0.077 (0.137)	0.127 (0.058)
500	0.092 (0.153)	0.088 (0.131)	0.112 (0.059)	0.093 (0.150)	0.134 (0.069)
			$F\hat{P}R$		
5000	0.001 (0.003)	0.008 (0.028)	0.077 (0.012)	0.014 (0.037)	0.095 (0.015)
1000	0.015 (0.054)	0.038 (0.068)	0.080 (0.029)	0.047 (0.087)	0.099 (0.035)
800	0.021 (0.071)	0.042 (0.075)	0.083 (0.033)	0.053 (0.092)	0.101 (0.039)
500	0.040 (0.111)	0.058 (0.087)	0.086 (0.040)	0.063 (0.100)	0.104 (0.048)

5.5 Predicting Diabetes in the Pima Indian Population

We apply the proposed direct maximization method to construct a linear combination of risk factors for predicting Type 2 diabetes in Pima Indian women. Pima Indians have one of the highest rates of Type 2 diabetes in the United States (McLaughlin, 2010). Diabetes treatment includes lifestyle changes and medication, such as antihyperglycemic oral agents. Medication can help diabetic patients control their disease and prevent diabetes related complications; however, many antihyperglycemic agents have potential side effects including weight gain, hypoglecemia, and in some instances heart failure (Inzucchi et al., 2012). Following risk thresholds selected for existing diabetes risk models used for recommending antihyperglycemics, we set the harm-benefit ratio to be $\frac{C}{B} = \frac{1}{4}$ (Lindström and Tuomilehto, 2003; Griffin et al., 2000).

We used the publicly available Pima Indian Diabetes dataset for composite marker construction. Diabetes risk factors are measured on 768 women aged 21 years or older. Diabetes risk factors included in the dataset are: number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, body mass index (BMI), 2-hour serum insulin, diabetes pedigree function, and age. Triceps skin fold thickness is a predictor of body density in women. Diabetes pedigree function summarizes family history of diabetes, with higher values indicating diabetes in closely related family members (Smith et al., 1988). We did not include triceps skin fold thickness or 2-hour serum insulin when constructing a linear combination of risk markers because over 30% of observations had missing measurements on either one or both variables. We removed 44 women from analysis due to missing data in the variables that were included in construction of a composite risk marker. The final sample size used for analysis was 724 women.

We constructed a linear combination of the risk markers under the proposed direct maximization method and compared it to linear combinations obtained from standard and robust

Table 5.5: Summary of risk markers in the Pima Indian training data

Risk Factors	Overall (N = 434) Mean (SD)	Cases (N=114) Mean (SD)	Controls (N=290) Mean (SD)
Number of pregnancies	3.7 (3.3)	4.9 (3.7)	3.3 (2.9)
Plasma glucose concentration (mmol/L)	122.6 (30.9)	141.4 (30.0)	111.3 (25.4)
Diastolic blood Pressure (mmHg)	72.1 (12.2)	74.6 (12.8)	70.6 (11.6)
BMI	32.5 (7.2)	35.4 (7.0)	30.8 (6.7)
Diabetes Pedigree Function	0.5 (0.4)	0.6 (0.4)	0.4 (0.3)
Age (Years)	33.1 (11.4)	36.4 (10.9)	31.1 (11.3)

logistic regression. For logistic regression models, we optimized sNB given the fitted regression coefficients to find a decision threshold. Linear combinations were estimated in a training set and estimates of sNB were obtained from an independent validation dataset, using a 60-40 training/validation split. The training sample size was 434 with a 35% event rate, and the validation sample size was 290 with 30% event rate.

Table 5.5 gives a summary of the risk markers in the training set by case and control status. The mean for all markers tend to be higher in cases compared to controls. Notably, average plasma glucose for cases is about 1 standard deviation larger than controls. Table 5.6 shows the estimated sNB, TPR, and FPR for fitted combinations and risk threhsolds in the validation datasets. The proposed method produces a combination with 3.6% higher $\widehat{sNB}$ compared to the linear combination obtained from either logistic regression method. The $\widehat{TPR}$ of the risk score obtained from standard or robust logistic regression is 1. The 3.5% lower in $\widehat{TPR}$ for the proposed method comes with a substantially lower $\widehat{FPR}$ (18.2% lower than standard logistic regression and 17.2% lower than robust logistic regression).

Table 5.6: Estimated sNB and Discriminatory Measures in Test Data

Method	Direct Maximization	Logistic Regression	Robust Logistic Regression
$\widehat{sNB}$	0.745	0.709	0.705
$\widehat{TPR}$	0.965	1.000	1.000
$\widehat{FPR}$	0.529	0.701	0.711

5.6 Discussion

When several risk factors are predictive of disease, it is useful to summarize them as a single composite risk marker, particularly in settings where the composite risk marker will be used for prescribing intervention. It is compelling to use methods to combine markers that account for the clinical setting in which the combination will be used.

We have presented a non-parametric approach for linearly combining risk markers by directly maximizing standardized net benefit. A crucial component of this method is that harms and benefits of the intervention are known at the outset of the problem. We showed, through simulation study and a real-data example, that the proposed direct maximization method can produce linear combinations of markers with higher sNB compared to conventional methods of combining markers.

There are some important limitations of this method to consider. The formulation of the problem is non-convex. There is no guarantee the solution found from this method is the optimal combination of risk markers that maximize sNB. In simulation studies, we have found there can be some sensitivity to starting value for the optimization routine. However, our prescribed method of using either normalized robust or standard logistic recalibration as the initial values has worked well across many different settings. Some other smooth approximation based methods, such as those in Fong et al. (2016), reformulate the problem using

more complex but convex approximations (e.g. ramp function). Alternatively, for settings where the number of parameters is not too large ($p \approx 5$) a global optimizer, such as the one in Chapter 4 could be used. We applied a global optimization routine in our simulation studies and found the results to be consistent with the proposed method. Adapting the propose method to address limitations due to local maxima is an area of future work.

The proposed direct optimization method produces a composite marker and threshold that can help in clinical decision making. However, unlike logistic regression, the composite marker is not on the risk scale, and therefore can make communicating risk difficult. Adjusting the method to produce a combination that can be used to communicate risk as well as guide clinical decision making is an interesting area of future work. Finally, one could consider extending the proposed to non-linear combinations or incorporating feature selection.

Chapter 6

CONCLUDING REMARKS

In this dissertation we proposed statistical methodology for risk markers that integrate measures of clinical utility. In each chapter we examined an existing area of methodological research, and consider how it could be augmented or added to, so that the clinical context of the risk model is embedded in the statistical methodology.

With new biomedical and information technology it is becoming easier to measure many biomarkers on individuals and retain a long history of personalized medical information. Additionally, clinicians and patients are opening doors of communication to ensure that individuals' preferences are accounted for in clinical decisions. As medicine trends this way, interest in finding optimal decision for individuals will continue to increase. By reviewing the foundations of decision-theoretic measures from both a frequentist and Bayesian perspective, we contend that a Bayesian interpretation can allow one to parameterize and solve an individualized decision problem. When considering the population clinical utility of a risk marker, Bayesian or frequentist interpretation is appropriate. One difficulty in measuring population clinical utility is how to consider heterogeneous preferences in the population. An interesting area of future work is to evaluate decision-making at a population level, in populations that have heterogenous preferences.

We developed methods for recalibration that account for the clinical utility of the risk model. We proposed two distinct approaches: (i) a weighted recalibration approach that aims to improve calibration at the risk threshold, and (ii) a constrained optimization method that seeks to maximize standardized net benefit. Through simulation study and data example we

showed that, compared to standard methods or recalibration, both methods can produce risk scores with higher estimates of utility and improved calibration at the clinically relevant risk threshold. The weighted recalibration approach requires specification of additional tuning parameters for the weighting function. We provided a way to give these tuning parameters clinical interpretability, and proposed a cross-validation approach when tuning parameters are difficult to elicit. A drawback of the cross-validation approach is that it may not be sensitive when there are few events, which is particularly an issue for small datasets. In these settings, our proposed constrained optimization method for recalibration may be an appealing alternative. In some settings there are more than one risk threshold for classifying patients in to high, moderate, and low risk groups with different management strategies. Extending our methodology to ensure risk models are calibrated at all clinically relevant risk thresholds could be an interesting area of future work.

Finally, we proposed a method of linearly combining risk markers that maximize a clinically relevant metric, standardized net benefit. This method does not make any distributional assumptions, and requires only a single tuning parameter to be specified. We have shown through simulation study and a real data example that the proposed method can produce biomarker combinations with higher sNB compared to conventional risk modeling approaches. However, a limitation of this method is that there is no guarantee it finds the global maximum, rather than a local maximum. This is an immediate area of future work. Additionally, adapting methodology to produce a linear combination that is on the risk scale can be useful for communicating risk in clinical settings. There are also many interesting extensions of this problem that could be considered, such as finding non-linear combinations or incorporating feature selection.

The methods we proposed will be developed into `R` packages to help facilitate their application.

BIBLIOGRAPHY

Anscombe, F. J., Aumann, R. J., et al. (1963). A definition of subjective probability. *Annals of mathematical statistics*, 34(1):199–205.

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

Baker, S. G. and Kramer, B. S. (2007). Peirce, youden, and receiver operating characteristic curves. *The American Statistician*, 61(4):343–346.

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

Bantle, J. P., Wylie-Rosett, J., Albright, A. L., Apovian, C. M., Clark, N. G., Franz, M. J., Hoogwerf, B. J., Lichtenstein, A. H., Mayer-Davis, E., Mooradian, A. D., et al. (2008). Nutrition recommendations and interventions for diabetes: a position statement of the american diabetes association. *Diabetes care*, 31:S61–S78.

Bellman, R. (1966). Dynamic programming. *Science*, 153(3731):34–37.

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

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods*, pages 17–34. Springer.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., Greenland, P., JacobsJr, D. R., Kronmal, R., Liu, K., et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology*, 156(9):871–881.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562–565.

Dall-Era, M. A., Albertsen, P. C., Bangma, C., Carroll, P. R., Carter, H. B., Cooperberg, M. R., Freedland, S. J., Klotz, L. H., Parker, C., and Soloway, M. S. (2012). Active surveillance for prostate cancer: a systematic review of the literature. *European urology*, 62(6):976–983.

Dalton, J. E. (2013). Flexible recalibration of binary clinical prediction models. *Statistics in medicine*, 32(2):282–289.

D'Amico, A. V., Whittington, R., Malkowicz, S. B., Fondurulia, J., Chen, M.-H., Kaplan, I., Beard, C. J., Tomaszewski, J. E., Renshaw, A. A., Wein, A., et al. (1999). Pretreatment nomogram for prostate-specific antigen recurrence after radical prostatectomy or external-beam radiation therapy for clinically localized prostate cancer. *Journal of Clinical Oncology*, 17(1):168–172.

De Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica*, 34:129–145.

DeFilippis, A. P., Young, R., Carrubba, C. J., McEvoy, J. W., Budoff, M. J., Blumenthal, R. S., Kronmal, R. A., McClelland, R. L., Nasir, K., and Blaha, M. J. (2015). An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort. *Annals of internal medicine*, 162(4):266–275.

Driscoll, D. A. and Gross, S. J. (2008). First trimester diagnosis and screening for fetal aneuploidy. *Genetics in Medicine*, 10(1):73.

Driscoll, D. A. and Gross, S. J. (2009). Screening for fetal aneuploidy and neural tube defects. *Genetics in Medicine*, 11(11):818.

Farquhar, P. H. (1984). State of the art utility assessment methods. *Management science*, 30(11):1283–1300.

FDA, U. (2014). Guidance for industry and fda staff: qualification process for drug development tools. *Federal Register*, pages 83100–2.

Fong, Y., Yin, S., and Huang, Y. (2016). Combining biomarkers linearly and nonlinearly for classification using the area under the roc curve. *Statistics in medicine*, 35(21):3792–3809.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Gablonsky, J. M. and Kelley, C. T. (2001). A locally-biased form of the direct algorithm. *Journal of Global Optimization*, 21(1):27–37.

Gafni, A. (1994). The standard gamble method: what is being measured and how it is interpreted. *Health Services Research*, 29(2):207.

Gao, F., Xiong, C., Yan, Y., Yu, K., and Zhang, Z. (2008). Estimating optimum linear combination of multiple correlated diagnostic tests at a fixed specificity with receiver operating characteristic curves. *Journal of Data Science*, 6(1):105–123.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Goff et al. (2014). 2013 acc/aha guideline on the assessment of cardiovascular risk. *Circulation*, 129(25 suppl 2):S49–S73.

Griffin, S., Little, P., Hales, C., Kinmonth, A., and Wareham, N. (2000). Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. *Diabetes/metabolism research and reviews*, 16(3):164–171.

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Heidenreich, A., Bastian, P. J., Bellmunt, J., Bolla, M., Joniau, S., van der Kwast, T., Mason, M., Matveev, V., Wiegel, T., Zattoni, F., et al. (2014). Eau guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent?update 2013. *European urology*, 65(1):124–137.

Hook, E. B. (1981). Rates of chromosome abnormalities at different maternal ages. *Obstetrics and gynecology*, 58(3):282–285.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Hsing, A. W. and Devesa, S. S. (2001). Trends and patterns of prostate cancer: what do they suggest? *Epidemiologic reviews*, 23(1):3–13.

Hsu, M.-J. and Hsueh, H.-M. (2013). The linear combinations of biomarkers which maximize the partial area under the roc curves. *Computational Statistics*, 28(2):647–666.

Inzucchi, S. E., Bergenstal, R. M., Buse, J. B., Diamant, M., Ferrannini, E., Nauck, M., Peters, A. L., Tsapas, A., Wender, R., and Matthews, D. R. (2012). Management of hyperglycemia in type 2 diabetes: a patient-centered approach: position statement of the american diabetes association (ada) and the european association for the study of diabetes (easd). *Diabetes care*, 35(6):1364–1379.

Iremashvili, V., Pelaez, L., Manoharan, M., Jorda, M., Rosenberg, D. L., and Soloway, M. S. (2012). Pathologic prostate cancer characteristics in patients eligible for active surveillance: a head-to-head comparison of contemporary protocols. *European urology*, 62(3):462–468.

Jensen, N. E. (1967). An introduction to bernoullian utility theory: I. utility functions. *The Swedish journal of economics*, pages 163–183.

Jiang, X., Osl, M., Kim, J., and Ohno-Machado, L. (2011). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274.

Johnson, S. G. (2014). The nlopt nonlinear-optimization package.

Jones, D. R., Perttunen, C. D., and Stuckman, B. E. (1993). Lipschitzian optimization without the lipschitz constant. *Journal of optimization Theory and Applications*, 79(1):157–181.

Kerr, K. F., Brown, M., and Janes, H. (2016a). Reply to aj vickers et al. *Journal of Clinical Oncology*, 35(4):473–475.

Kerr, K. F., Brown, M. D., Marsh, T. L., and Janes, H. (2019a). Assessing the clinical impact of risk models for opting out of treatment. *Medical Decision Making*, 39(2):86–90.

Kerr, K. F., Brown, M. D., Zhu, K., and Janes, H. (2016b). Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *Journal of Clinical Oncology*, page JCO655654.

Kerr, K. F., Marsh, T. L., and Janes, H. (2019b). The importance of uncertainty and opt in vs. opt out: best practices for decision curve analysis. *To be published in Medical Decision Making*.

Klotz, L. (2005). Active surveillance for prostate cancer: for whom? *Journal of Clinical Oncology*, 23(32):8165–8169.

Klotz, L., Zhang, L., Lam, A., Nam, R., Mamedov, A., and Loblaw, A. (2009). Clinical results of long-term follow-up of a large, active surveillance cohort with localized prostate cancer. *Journal of Clinical Oncology*, 28(1):126–131.

Lin, H., Zhou, L., Peng, H., and Zhou, X.-H. (2011). Selection and combination of biomarkers using roc method for disease classification and prediction. *Canadian Journal of Statistics*, 39(2):324–343.

Lindström, J. and Tuomilehto, J. (2003). The diabetes risk score: a practical tool to predict type 2 diabetes risk. *Diabetes care*, 26(3):725–731.

Liu, C., Liu, A., and Halabi, S. (2011). A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 30(16):2005–2014.

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using roc. *Biometrics*, 63(3):751–757.

McClish, D. K. (1989). Analyzing a portion of the roc curve. *Medical Decision Making*, 9(3):190–195.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics*, 58(3):657–664.

McLaughlin, S. (2010). Traditions and diabetes prevention: a healthy path for native americans. *Diabetes Spectrum*, 23(4):272–277.

Meisner, A., Carone, M., Pepe, M., and Kerr, K. F. (2017). Combining biomarkers by maximizing the true positive fate for a fixed false positive rate. *UW Biostatistics Working Paper Series*, 420.

Mennuti, M. T., Cherry, A. M., Morrissette, J. J., and Dugoff, L. (2013). Is it time to sound an alarm about false-positive cell-free dna testing for fetal aneuploidy? *American journal of obstetrics and gynecology*, 209(5):415–419.

Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. ACM.

Parmigiani, G. and Inoue, L. (2009). *Decision theory: principles and approaches*, volume 812. John Wiley & Sons.

Pauker, S. G. and Kassirer, J. P. (1975). Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine*, 293(5):229–234.

Pauker, S. G. and Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, 302(20):1109–1117.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, 4(93):453–454.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229.

Pepe, M. S., Fan, J., Feng, Z., Gerds, T., and Hilden, J. (2015). The net reclassification index (nri): a misleading measure of prediction improvement even with independent test data sets. *Statistics in biosciences*, 7(2):282–295.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Ridker, P. M. and Cook, N. R. (2013). Statins: new american guidelines for prevention of cardiovascular disease. *The Lancet*, 382(9907):1762–1765.

Rouprêt, M., Hupertan, V., Seisen, T., Colin, P., Xylinas, E., Yates, D. R., Fajkovic, H., Lotan, Y., Raman, J. D., Zigeuner, R., et al. (2013). Prediction of cancer specific survival

after radical nephroureterectomy for upper tract urothelial carcinoma: development of an optimized postoperative nomogram using decision curve analysis. *The Journal of urology*, 189(5):1662–1669.

Shariat, S. F., Savage, C., Chromecki, T. F., Sun, M., Scherr, D. S., Lee, R. K., Lughezzani, G., Remzi, M., Marberger, M. J., Karakiewicz, P. I., et al. (2011). Assessing the clinical benefit of nuclear matrix protein 22 in the surveillance of patients with nonmuscle-invasive bladder cancer and negative cytology: A decision-curve analysis. *Cancer*, 117(13):2892–2897.

Singer, E., Corning, A. D., and Antonucci, T. (1999). Attitudes toward genetic testing and fetal diagnosis, 1990-1996. *Journal of Health and Social Behavior*, pages 429–445.

Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association.

Steyerberg, E. W. and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an abcd for validation. *European heart journal*, 35(29):1925–1931.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355.

Tsalatsanis, A., Hozo, I., Vickers, A., and Djulbegovic, B. (2010). 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):51.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., and Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*.

Van Calster, B. and Vickers, A. J. (2015). Calibration of risk prediction models impact on decision-analytic performance. *Medical Decision Making*, 35(2):162–169.

Vickers, A. J. (2008). Decision analysis for the evaluation of diagnostic tests, prediction models, and molecular markers. *The American Statistician*, 62(4).

Vickers, A. J. and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574.

Vickers, A. J., Kattan, M. W., and Sargent, D. J. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*, 8(1):14.

Von Neumann, J. and Morgenstern, O. (1945). *Theory of games and economic behavior*. Princeton University Press Princeton, NJ.

Wilson, R. D. (2000). Amniocentesis and chorionic villus sampling. *Current Opinion in Obstetrics and Gynecology*, 12(2):81–86.

Xu, T., Fang, Y., Rong, A., and Wang, J. (2015). Flexible combination of multiple diagnostic biomarkers to improve diagnostic accuracy. *BMC medical research methodology*, 15(1):94.

Yan, Q., Bantis, L. E., Stanford, J. L., and Feng, Z. (2018). Combining multiple biomarkers linearly to maximize the partial area under the roc curve. *Statistics in medicine*, 37(4):627–642.

Yin, J. and Tian, L. (2014). Optimal linear combinations of multiple diagnostic biomarkers based on youden index. *Statistics in medicine*, 33(8):1426–1440.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM.

Zane, S., Creanga, A. A., Berg, C. J., Pazol, K., Suchdev, D. B., Jamieson, D. J., and Callaghan, W. M. (2015). Abortion-related mortality in the united states 1998-2010. *Obstetrics and gynecology*, 126(2):258.

Appendix A

APPENDIX A: SUPPLEMENTARY MATERIAL FOR CHAPTER 2

A.1 Expected Utility Framework and Standard Gamble Approach

The expected utility framework is attributed to Von Neumann and Morgenstern (1945) which says that we can compare preferences for actions by comparing expected utilities. This theorem relies on some necessary and sufficient conditions that ensure the expected utility principle holds. Parmigiani and Inoue (2009) or Jensen (1967) provide full statement of conditions and proof of expected utility principle, which we do not restate here. The von Neumann and Morgenstern expected utility theory also provides a framework for eliciting utilities by using the "standard gamble" which we discuss next via an example. The standard gamble approach has been discussed in many different contexts including health care (Farquhar, 1984; Gafni, 1994; Parmigiani and Inoue, 2009).

A classic illustration of the standard gamble approach is an individual who is contemplating treatment in the form of surgery for a disease that causes chronic pain. This individual has two actions to choose from: undergo surgery or avoid surgery. Suppose there are two potential consequences for this surgery: success in the form of alleviation from pain (with probability π), or death (with probability $1 - \pi$). If the patient chooses the second action, to avoid surgery the resulting consequence will be to remain in chronic pain, with probability of 1. Therefore, the patient is tasked with choosing between the "sure thing", that is, with probability one of avoiding treatment and remaining in chronic pain or the "gamble" of undergoing surgery with uncertain outcomes. The goal of using the standard gamble for utility elicitation is to identify the probability π (of a successful surgery) at

which the decision-maker is indifferent between the sure-thing (no surgery) and the gamble (surgery).

The standard gamble approach to eliciting utilities goes as follows. Assuming the axioms required for the von Neumann Morgenstern theorem hold, the patient should be able to rank preferences for the potential consequences. In this simple example, assume that the best outcome is a successful surgery (represented by z^0) and the worst outcome is death (represented by z_0), while the outcome of living in chronic pain (represented by z) falls somewhere in between. The best outcome is assigned a utility of 1, so $u(z^0) = 1$; the worst outcome is assigned a utility of 0, so $u(z_0) = 0$. Let $u(z) = u$ be the remaining unknown utility associated with living with chronic pain. Figure A.1 shows a decision tree which represents this problem. The expected utility for each action is,

$$\begin{aligned} U_\pi(\text{surgery}) &= u(\text{success}) \times \pi(\text{sucess}) + u(\text{death})(1 - \pi(\text{death})) \\ &= u(z^0)\pi + u(z_0)(1 - \pi) \\ &= 1 \times \pi + 0(1 - \pi) \\ &= \pi \end{aligned}$$

$$\begin{aligned} U_\pi(\text{no surgery}) &= u(\text{chronic pain}) \times \pi(\text{chronic pain}) \\ &= u(z) \times 1 \\ &= u(z) \end{aligned}$$

If the decision-maker is indifferent between the sure-thing (no surgery) and the gamble (surgery), then the expected utilities of these two actions should be the same. Thus,

$$\begin{aligned} U_\pi(\text{surgery}) &= U_\pi(\text{no surgery}) \\ u(z) &= \pi. \end{aligned}$$

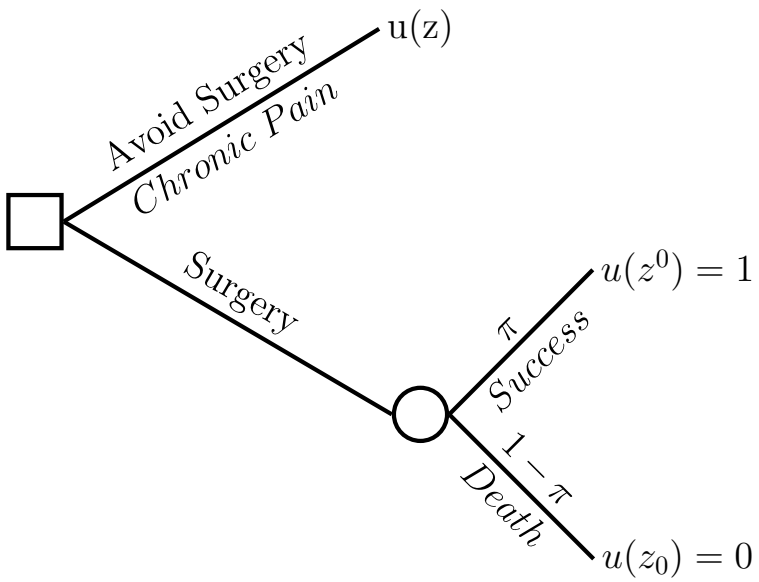


Figure A.1: Decision tree for the standard gamble illustration. The decision-maker can choose to undergo surgery with probability of success π, or avoid surgery but have no uncertainty about having chronic pain. $u(z^0)$ represents the utility of the best outcome, successful surgery, and $u(z_0)$ represents the utility of the worst outcome, death from surgery. $u(z)$ represents the utility of living with chronic pain.

Therefore, it must be that the utility of living in chronic pain is equal to $\pi(\theta)$. For a risk-averse person the probability of a successful surgery might have to be quite high before he or she becomes indifferent between a risky surgery and living with chronic pain; for a risk-tolerant person that same probability could be low. For a decision problem with a larger number of discrete outcomes the process can be iterated comparing any three sets of ordered outcomes and finding the point of indifference as described above.

A.2 Connections between net benefit and expected utility

Remark. *Following the notation in Table 2.1, a_{RP} is the radical prostatectomy for all men action, a_{WW} is the watchful waiting for all men action, and a_{AS} is the PSA-based active surveillance for all men action. We use utilities as noted in Table in Table 2.1. Let $U_\pi(a)$ denote the expected utility of action a given prior distribution $\pi(\theta)$. We denote net benefit of action a as $NB(a)$. Let $B = u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)$ and $C = u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})$. We define*

$$\begin{aligned} NB(a_{RP}) &= P(Y=1) - \frac{C}{B}P(Y=0), \\ NB(a_{AS}) &= P(Y=1)TPR - \frac{C}{B}P(Y=0)FPR, \\ NB(a_{WW}) &= 0. \end{aligned}$$

Under the decision problem we have the following relationship

$$NB(a_{RP}) = \frac{U_\pi(a_{RP}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)} \tag{A.1}$$

$$NB(a_{AS}) = \frac{U_\pi(a_{AS}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)} \tag{A.2}$$

$$NB(a_{WW}) = \frac{U_\pi(a_{WW}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)}. \tag{A.3}$$

That is, the net benefit for action a is the expected utility of action a minus the expected utility of the reference action, a_{WW}, divided by the difference in utilities for radical prostatectomy and watchful waiting among cases (i.e., the benefit of treatment for a case, B).

Proof. Let M represent the PSA based screening test, with $M = 1$ being a positive and $M = 0$ a negative. The let $P(M = 1)$ be the probability of a positive test, and let $P(M = 1|\theta_D)$ and $P(M = 1|\theta_{\bar{D}})$ denote TPR and FPR of M. Under the decision tree without added decision node (given in Figure 2.4) the expected utility of the three actions are

$$\begin{aligned}
U_\pi(a_{RP}) &= u(a_{RP}|\theta_D)\pi(\theta_D) + u(a_{RP}|\theta_{\bar{D}})\pi(\theta_{\bar{D}}) \\
U_\pi(a_{WW}) &= u(a_{WW}|\theta_D)\pi(\theta_D) + u(a_{WW}|\theta_{\bar{D}})\pi(\theta_{\bar{D}}) \\
U_\pi(a_{AS}) &= P(M = 1|\theta_D)\pi(\theta_D)u(a_{RP}|\theta_D) + P(M = 1|\theta_{\bar{D}})\pi(\theta_{\bar{D}})u(a_{RP}|\theta_{\bar{D}}) \\
&\quad + P(M = 0|\theta_D)\pi(\theta_D)u(a_{WW}|\theta_D) + P(M = 0|\theta_{\bar{D}})\pi(\theta_{\bar{D}})u(a_{WW}|\theta_{\bar{D}})
\end{aligned}$$

First we prove equality (A.1) holds. The difference in expected utility between action a_{RP} and a_{WW} is

$$\begin{aligned}
U_\pi(a_{RP}) - U_\pi(a_{WW}) &= u(a_{RP}|\theta_D)\pi(\theta_D) + u(a_{RP}|\theta_{\bar{D}})\pi(\theta_{\bar{D}}) \\
&\quad - u(a_{WW}|\theta_D)\pi(\theta_D) - u(a_{WW}|\theta_{\bar{D}})\pi(\theta_{\bar{D}}) \\
&= \pi(\theta_D)[u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)] - \pi(\theta_{\bar{D}})[u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})] \\
&= \pi(\theta_D) \times B - \pi(\theta_{\bar{D}}) \times C \\
\Rightarrow \frac{U_\pi(a_{RP}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D) \times C} &= \frac{\pi(\theta_D) \times B - \pi(\theta_{\bar{D}})}{B} \\
&= NB(a_{RP}).
\end{aligned}$$

Next we prove equality (A.2). The difference in expected utility between action a_{AS} and

a_{WW} is

$$\begin{aligned}
U_\pi(a_{AS}) - U_\pi(a_{WW}) &= P(M=1|\theta_D)\pi(\theta_D)u(a_{RP}|\theta_D) + P(M=1|\theta_{\bar{D}})\pi(\theta_{\bar{D}})u(a_{RP}|\theta_{\bar{D}}) \\
&\quad + P(M=0|\theta_D)\pi(\theta_D)u(a_{WW}|\theta_D) + P(M=0|\theta_{\bar{D}})\pi(\theta_{\bar{D}})u(a_{WW}|\theta_{\bar{D}}) \\
&\quad - u(a_{WW}|\theta_D)\pi(\theta_D) - u(a_{WW}|\theta_{\bar{D}})\pi(\theta_{\bar{D}}) \\
&= \pi(\theta_D)\left[P(M=1|\theta_D)u(a_{RP}|\theta_D)\right. \\
&\qquad \left. + (1 - P(M=1|\theta_D))u(a_{WW}|\theta_D) - u(a_{WW}|\theta_D)\right] \\
&\quad + \pi(\theta_{\bar{D}})\left[P(M=1|\theta_{\bar{D}})u(a_{RP}|\theta_{\bar{D}})\right. \\
&\qquad \left. + (1 - P(M=1|\theta_{\bar{D}}))u(a_{WW}|\theta_{\bar{D}}) - u(a_{WW}|\theta_{\bar{D}})\right] \\
&= \pi(\theta_D)[P(M=1|\theta_D)u(a_{RP}|\theta_D) + u(a_{WW}|\theta_D) \\
&\quad - P(M=1|\theta_D)u(a_{WW}|\theta_D) - u(a_{WW}|\theta_D)] \\
&\quad + \pi(\theta_{\bar{D}})[P(M=1|\theta_{\bar{D}})u(a_{RP}|\theta_{\bar{D}}) + u(a_{WW}|\theta_{\bar{D}}) \\
&\quad - P(M=1|\theta_{\bar{D}})u(a_{WW}|\theta_{\bar{D}}) - u(a_{WW}|\theta_{\bar{D}})] \\
&= \pi(\theta_D)[P(M=1|\theta_D)u(a_{RP}|\theta_D) - P(M=1|\theta_D)u(a_{WW}|\theta_D)] \\
&\quad + \pi(\theta_{\bar{D}})[P(M=1|\theta_{\bar{D}})u(a_{RP}|\theta_{\bar{D}}) - P(M=1|\theta_{\bar{D}})u(a_{WW}|\theta_{\bar{D}})] \\
&= \pi(\theta_D)P(M=1|\theta_D)[u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)] \\
&\quad + \pi(\theta_{\bar{D}})P(M=1|\theta_{\bar{D}})[u(a_{RP}|\theta_{\bar{D}}) - u(a_{WW}|\theta_{\bar{D}})] \\
&= \pi(\theta_D)P(M=1|\theta_D)B + \pi(\theta_{\bar{D}})P(M=1|\theta_{\bar{D}})(-C) \\
&= \pi(\theta_D)TPR \times B - \pi(\theta_{\bar{D}})FPR \times C \\
\Rightarrow \frac{U_\pi(a_{AS}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)} &= \frac{\pi(\theta_D)TPR \times B - \pi(\theta_{\bar{D}})FPR \times C}{B} \\
&= NB(a_{AS}).
\end{aligned}$$

Finally, we prove (A.3). The difference in expected utility between action a_{WW} and a_{WW} is

$$\begin{aligned} U_\pi(a_{WW}) - U_\pi(a_{WW}) &= 0 \\ \Rightarrow \frac{U_\pi(a_{WW}) - U_\pi(a_{WW})}{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)} &= \frac{0}{B} \\ &= NB(a_{WW}). \end{aligned}$$

□

A.3 Optimality of risk marker with and without added decision node

Remark. *Following the notation in Table 2.1, a_{RP} is the radical prostatectomy for all men action, a_{WW} is the watchful waiting for all men action, and a_{AS} is the PSA-based active surveillance for all men action. We use utilities as noted in Table in Table 2.1. TPR and FPR represent the true and false positive rate of PSA marker M. Under the tree with the added decision node (Figure 2.5), use of the PSA-based active surveillance is the optimal decision if*

$$\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{FPR}{TPR} \leq \frac{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)}{u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})} \leq \frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{1-FPR}{1-TPR} \tag{A.4}$$

or

$$\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{1-FPR}{1-TPR} \leq \frac{u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)}{u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})} \leq \frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{FPR}{TPR}. \tag{A.5}$$

Proof. We will refer to the tree with the added decision nodes as the "Bayes" tree and the tree without the added decision nodes as the "Simple" tree. The expected utility of the

active surveillance under the decision-theoretic framework is

$$\begin{aligned} U_{\pi,Bayes}(a_{AS}) = \max\{&\pi(\theta_D)TPR\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{RP}|\theta_{\bar{D}}), \\ &\pi(\theta_D)TPR\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{ww}|\theta_{\bar{D}})\} \\ + \max\{&\pi(\theta_D)(1-TPR)\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{RP}|\theta_{\bar{D}}), \\ &\pi(\theta_D)(1-TPR)\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{ww}|\theta_{\bar{D}})\} \end{aligned}$$

Under the tree with the added decision nodes active surveillance is optimal if (1) $U_{\pi,Bayes}(a_{AS}) \geq U_{\pi}(a_{RP})$ and (2) $U_{\pi,Bayes}(a_{AS}) \geq U_{\pi}(a_{WW})$. We will evaluate when inequalities (1) and (2) are true under both the intuitive case (radical prostatectomy if $M = 1$, no radical prostatectomy if $M = 0$) and the non-intuitive case (radical prostatectomy if $M = 0$, no radical prostatectomy if $M = 1$)

Intuitive Case:

Under the intuitive case, where radical prostatectomy occurs if $M = 1$, no radical prostatectomy occurs if $M = 0$, the expected utility of active surveillance is

$$\begin{aligned} U_{\pi,Bayes}(a_{AS}) = &\pi(\theta_D)TPR\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{RP}|\theta_{\bar{D}}) \\ &+ \pi(\theta_D)(1-TPR)\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{WW}|\theta_{\bar{D}}) \end{aligned} \quad \text{(A.6)}$$

$$\begin{aligned} U_{\pi}(a_{RP}) &\leq U_{\pi,Bayes}(a_{AS}) \\ \pi(\theta_D)\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})\ u(a_{RP}|\theta_{\bar{D}}) &\leq \pi(\theta_D)TPR\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{RP}|\theta_{\bar{D}}) \\ &\quad + \pi(\theta_D)(1-TPR)\ u(a_{WW}|\theta_D) \\ &\quad + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{WW}|\theta_{\bar{D}}) \\ (1-TPR)\pi(\theta_D)\ u(a_{RP}|\theta_D) + (1-FPR)\pi(\theta_{\bar{D}})\ u(a_{RP}|\theta_{\bar{D}}) &\leq \pi(\theta_D)(1-TPR)\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{WW}|\theta_{\bar{D}}) \\ (1-TPR)\pi(\theta_D)[u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)] &\leq (1-FPR)\pi(\theta_{\bar{D}})[u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})] \\ (1-TPR)\pi(\theta_D)\ B &\leq (1-FPR)\pi(\theta_{\bar{D}})\ C \\ \frac{B}{C} &\leq \frac{1-FPR}{1-TPR}\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}. \end{aligned}$$

The inequality (2) is true when

$$\begin{aligned}
U_\pi(a_{WW}) &\leq U_{\pi,Bayes}(a_{AS}) \\
\pi(\theta_D)\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})\ u(a_{WW}|\theta_{\bar{D}}) &\leq \pi(\theta_D)TPR\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{RP}|\theta_{\bar{D}}) \\
&\quad + \pi(\theta_D)(1-TPR)\ u(a_{WW}|\theta_D) \\
&\quad + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{WW}|\theta_{\bar{D}}) \\
\pi(\theta_D)TPR\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{WW}|\theta_{\bar{D}}) &\leq \pi(\theta_D)TPR\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{RP}|\theta_{\bar{D}}) \\
FPR\ \pi(\theta_{\bar{D}})[u(a_{WW}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})] &\leq TPR\ \pi(\theta_D)[u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)] \\
FPR\ \pi(\theta_{\bar{D}})\ C &\leq TPR\ \pi(\theta_D)\ B \\
\frac{FPR}{TPR}\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)} &\leq \frac{B}{C}.
\end{aligned}$$

Therefore, for the intuitive setting (radical prostatectomy if $M = 1$, no radical prostatectomy if $M = 0$) the PSA based screening is optimal under the Bayes tree if

$$\frac{FPR}{TPR}\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)} \leq \frac{B}{C} \leq \frac{1-FPR}{1-TPR}\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}. \tag{A.7}$$

Non-intuitive Case:

Next we will consider the non-intuitive case (radical prostatectomy if $M = 0$, no radical prostatectomy if $M = 1$). In this setting the expected utility of using PSA-based active surveillance is

$$\begin{aligned}
U_{\pi,Bayes}(a_{AS}) &= \pi(\theta_D)TPR\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{WW}|\theta_{\bar{D}}) \\
&\quad + \pi(\theta_D)(1-TPR)\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{RP}|\theta_{\bar{D}}).
\end{aligned} \tag{A.8}$$

First we show when (1) is true.

$$
\begin{aligned}
U_\pi(a_{RP}) &\le U_{\pi,Bayes}(a_{AS}) \\
\pi(\theta_D)\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})\ u(a_{RP}|\theta_{\bar{D}}) &\le \pi(\theta_D)TPR\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{WW}|\theta_{\bar{D}}) \\
&\quad + \pi(\theta_D)(1-TPR)\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{RP}|\theta_{\bar{D}}), \\
\pi(\theta_D)TPR\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{RP}|\theta_{\bar{D}}) &\le \pi(\theta_D)TPR\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{ww}|\theta_{\bar{D}}), \\
\pi(\theta_D)TPR[u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)] &\le \pi(\theta_{\bar{D}})FPR[u(a_{ww}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})], \\
\pi(\theta_D)TPR \times B &\le \pi(\theta_{\bar{D}})FPR \times C, \\
\frac{B}{C} &\le \frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{FPR}{TPR},
\end{aligned}
$$

Next we show (2) is true when

$$
\begin{aligned}
U_\pi(a_{WW}) &\le U_{\pi,Bayes}(a_{AS}) \\
\pi(\theta_D)\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})\ u(a_{ww}|\theta_{\bar{D}}) &\le \pi(\theta_D)TPR\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{ww}|\theta_{\bar{D}}) \\
&\quad + \pi(\theta_D)(1-TPR)\ u(a_{RP}|\theta_D) \\
&\quad + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{RP}|\theta_{\bar{D}}), \\
\pi(\theta_D)(1-TPR)\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{ww}|\theta_{\bar{D}}) &\le \pi(\theta_D)(1-TPR)u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)u(a_{RP}|\theta_{\bar{D}}), \\
\pi(\theta_{\bar{D}})(1-FPR)[u(a_{ww}|\theta_{\bar{D}}) - u(a_{RP}|\theta_{\bar{D}})] &\le \pi(\theta_D)(1-TPR)[u(a_{RP}|\theta_D) - u(a_{WW}|\theta_D)], \\
\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{(1-FPR)}{(1-TPR)} &\le \frac{B}{C}.
\end{aligned}
$$

Therefore, the PSA based screening is optimal under the Bayes tree in the non-intuitive case if

$$
\frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{(1-FPR)}{(1-TPR)} \le \frac{B}{C} \le \frac{\pi(\theta_{\bar{D}})}{\pi(\theta_D)}\frac{FPR}{TPR} \tag{A.9}
$$

Under the Bayes tree, the PSA based active surveillance action will be optimal if either (A.7) or (A.9) is true.

Under the Simple tree the expected utility of PSA based active surveillance is

$$U_{\pi,Simple}(a_{AS}) = \pi(\theta_D)TPR\ u(a_{RP}|\theta_D) + \pi(\theta_{\bar{D}})FPR\ u(a_{RP}|\theta_{\bar{D}}) \\ + \pi(\theta_D)(1-TPR)\ u(a_{WW}|\theta_D) + \pi(\theta_{\bar{D}})(1-FPR)\ u(a_{WW}|\theta_{\bar{D}}). \quad \text{(A.10)}$$

Note this is the same as the expected utility of active surveillance under the Bayes tree for the intuitive case (as presented in A.6). Therefore for the Simple tree, the PSA based active surveillance action will be optimal if (A.7) is true. □

Appendix B

APPENDIX B: SUPPLEMENTARY MATERIAL FOR CHAPTER 3 AND CHAPTER 4

B.1 Theoretical Results

To prove that calibration at the risk threshold implies that sNB is maximized using the risk threshold R, and vice versa, we first prove two lemmas. First, a condition needed for the main result is that the risk score r has concave ROC curve. Lemma 3.2.2 states that a risk score that is monotonically non-decreasing, as defined in Definition 3.2.2, has concave *ROC* curve. In Lemma B.1.2 we relate maximized sNB at R and the slope of the ROC curve. Versions of this result have been proved in Metz (1978), Baker and Kramer (2007), and Baker et al. (2012). Finally, building off these results, we prove Theorem 3.2.3 formally stating the connection between calibration and sNB.

B.1.1 Monotonicity of the risk score implies concave ROC curve

Lemma. *If r is monotonically non-decreasing, meaning*

$$r_i > r_j \Rightarrow P(Y = 1|r_i) \geq P(Y = 1|r_j), \tag{B.1}$$

then r has concave ROC curve.

Proof. Let r be a risk score that is monotonically non-decreasing. Let $f_r(t)$ be the density of r, and let $f_r(t|Y = 1)$ and $f_r(t|Y = 0)$ be the density of r among cases and controls, respectively. Starting with Lemma 3.2.1, we have,

$$ROCSLOPE(t) = \frac{f_r(t|Y = 1)}{f_r(t|Y = 0)}.$$

Using Bayes rule,

$$= \frac{P(Y=1|r=t)f_r(t)}{P(Y=1)} \frac{P(Y=0)}{P(Y=0|r=t)f_r(t)}$$
$$= \frac{P(Y=1|r=t)}{P(Y=0|r=t)} \frac{P(Y=0)}{P(Y=1)}$$
$$= \frac{P(Y=1|r=t)}{1-P(Y=1|r=t)} \frac{1-P(Y=1)}{P(Y=1)}$$

The ROC curve for r is concave if $ROCSLOPE(t)$ is non-decreasing as $FPR = P(r \geq t|Y = 0)$ decreases; as the risk threshold t increases. Since $\frac{P(Y=1)}{1-P(Y=1)}$ is a constant with respect to t, it suffices to show $\frac{P(Y=1|r=t)}{1-P(Y=1|r=t)}$ is non-decreasing in t. $P(Y = 1|r = t)$ is non-decreasing by definition, when r is monotonically non-decreasing. Since the transformation $\frac{P(Y=1|r=t)}{1-P(Y=1|r=t)}$ is an increasing transformation, $\frac{P(Y=1|r=t)}{1-P(Y=1|r=t)}$ is non-decreasing in t. □

B.1.2 Risk threshold maximizes sNB

Lemma. *Assume that r is monotonically non-decreasing. Let $\frac{C}{B} = \frac{R}{1-R}$ represent the harm-benefit ratio of intervention associated with r. sNB is maximized over all possible thresholds, t, if and only if*

$$ROCSLOPE(t) = \frac{P(Y=0)}{P(Y=1)} \frac{R}{1-R}.$$

Proof. For a given harms-benefit ratio, $\frac{R}{1-R}$, the standardized net benefit of the risk score, r, is

$$sNB(r) = TPR_t(r) - \frac{P(Y=0)}{P(Y=1)} \frac{R}{1-R} FPR_t(r). \tag{B.2}$$

Taking the derivative of (B.2) with respect to FPR yields

$$\begin{aligned}\frac{\partial \ sNB}{\partial \ FPR} &= \frac{\partial \ TPR}{\partial \ FPR} - \frac{P(Y=0)}{P(Y=1)}\frac{R}{1-R} \\ &= ROCSLOPE(t) - \frac{P(Y=0)}{P(Y=1)}\frac{R}{1-R} \end{aligned} \tag{B.3}$$

Setting expression (B.3) equal to 0 yields

$$\begin{aligned}\frac{\partial \ sNB}{\partial \ FPR} &\stackrel{set}{=} 0 \\ \Rightarrow ROCSLOPE(t) &= \frac{P(Y=0)}{P(Y=1)}\frac{R}{1-R}\end{aligned}$$

Since r is monotonically non-decreasing, r has concave ROC curve. A concave ROC curve implies that the $ROCSLOPE(t)$ is a non-increasing function of FPR, and therefore $\frac{\partial^2 \ sNB(r)}{\partial \ FPR^2}$ is negative and the critical value found above is a maximum. Therefore sNB is maximized across all thresholds t if and only if the slope of the ROC curve at threshold t equals $\frac{P(Y=0)}{P(Y=1)}\frac{R}{1-R}$. □

B.1.3 Calibration and Clinical Utility

Theorem. *Let r be a monotonically non-decreasing risk score predicting binary outcome Y. Let R be the risk threshold that represents harms and benefits of the intervention associated with the risk model. r has maximum sNB among any risk model of the form*

$$\log\left\{\frac{P(Y=1)}{P(Y=0)}\right\} = \alpha_0 + \alpha_1 \log\left\{\frac{r}{1-r}\right\}, \tag{B.4}$$

if and only if r is calibrated at R.

Proof. Max $sNB \Rightarrow$ calibration of r at R

Let $f_r(R)$ be the density of r at threshold R. Let $f_r(R|Y=1)$ and $f_r(R|Y=0)$ be the density of r among cases and controls, respectively, for threshold R. From Lemma 3.2.1, the

slope of the ROC curve for risk model r at threshold R is

$$\begin{aligned} ROCSLOPE(R) &= \frac{f_r(R|Y=1)}{f_r(R|Y=0)} \\ &= \frac{P(Y=1|r=R)f_r(R)}{P(Y=1)} \frac{P(Y=0)}{P(Y=0|r=R)f_r(R)} \\ &= \frac{P(Y=1|r=R)}{P(Y=0|r=R)} \frac{P(Y=0)}{P(Y=1)}. \end{aligned} \tag{B.5}$$

Since R is the threshold that maximizes sNB it follows (from Lemma B.1.2) that

$$ROCSLOPE(R) = \frac{R}{1-R} \frac{P(Y=0)}{P(Y=1)}. \tag{B.6}$$

Setting (B.5) and (B.6) equal to each other yields,

$$\begin{aligned} \frac{P(Y=1|r=R)}{P(Y=0|r=R)} \frac{P(Y=0)}{P(Y=1)} &= \frac{R}{1-R} \frac{P(Y=0)}{P(Y=1)} \\ \frac{P(Y=1|r=R)}{1-P(Y=1|r=R)} &= \frac{R}{1-R} \\ P(Y=1|r=R) &= R. \end{aligned}$$

Calibration of r at $R \Rightarrow$ Max sNB

The slope of the ROC curve for risk score r at risk threshold R is

$$\begin{aligned} ROCSLOPE(R) &= \frac{f_r(R|Y=1)}{f_r(R|Y=0)} \\ &= \frac{P(Y=1|r=R)}{P(Y=0|r=R)} \frac{P(Y=0)}{P(Y=1)}. \end{aligned} \tag{B.7}$$

Suppose r is calibrated at R, meaning $P(Y=1|r=R) = R$. Substituting this equality into

(B.7) yields

$$ROCSLOPE(R) = \frac{R}{1-R}\frac{P(Y=0)}{P(Y=1)} \tag{B.8}$$

By Lemma B.1.2 it follows that risk threshold R maximizes sNB. □

B.1.4 Variance of $\widehat{sNB}$ for given risk score

Proposition. *Let r be a risk model predicting binary random variable Y. Assume that the harm to benefit ratio of the intervention associated with the outcome is $\frac{C}{B}$. For a random sample of i individuals from the relevant population, the estimated standardized net benefit of the policy that treats individuals with $r > t$ is*

$$\widehat{sNB}\,(r) = \widehat{TPR}_t(r) - \frac{C}{B}\frac{\hat{P}(Y=0)}{\hat{P}(Y=1)}\widehat{FPR}_t(r) \tag{B.9}$$

Define the sample proportions:

$$\hat{p}_{11} = \frac{1}{n}\mathbf{1}[r \geq t, Y=1] \qquad \hat{p}_{01} = \frac{1}{n}\mathbf{1}[r < t, Y=1], \tag{B.10}$$

$$\hat{p}_{10} = \frac{1}{n}\mathbf{1}[r \geq t, Y=0] \qquad \hat{p}_{00} = \frac{1}{n}\mathbf{1}[r < t, Y=0]. \tag{B.11}$$

The estimator in B.9 can be expressed as a function of $\vec{\hat{p}} = \{\hat{p}_{11}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{00}\}$, when $\hat{p}_{11}+\hat{p}_{01} > 0$ and $\hat{p}_{10} + \hat{p}_{00} > 0$ (since we can never divide by 0),

$$\widehat{sNB}(\vec{\hat{p}}) = \frac{\hat{p}_{11}}{\hat{p}_{11}+\hat{p}_{01}} - \frac{C}{B}\left(\frac{\hat{p}_{10}+\hat{p}_{00}}{\hat{p}_{11}+\hat{p}_{01}}\right)\frac{\hat{p}_{10}}{\hat{p}_{10}+\hat{p}_{00}}.$$

Then, the variance of $\widehat{sNB}_t(\vec{\hat{p}})$ given harm to benefit ratio, $\frac{C}{B}$, is

$$V\left[\widehat{sNB}\left(\vec{\hat{p}}\right)\right] = \left(\frac{1}{p_{11}+p_{01}}\right)^2\left[\left(\frac{C}{B}\right)^2 p_{10} + \frac{p_{11}\left(p_{01}+\frac{C}{B}p_{10}\right)^2}{(p_{11}+p_{01})^2} + \frac{p_{01}(\frac{C}{B}p_{10}-p_{11})^2}{(p_{11}+p_{01})^2}\right], \tag{B.12}$$

where

$$p_{11} = P(r \geq t, Y = 1) \qquad p_{01} = P(r < t, Y = 1), \tag{B.13}$$

$$p_{10} = P(r \geq t, Y = 0) \qquad p_{00} = P(r < t, Y = 0), \tag{B.14}$$

and $p_{11} + p_{01} > 0$ *and* $p_{10} + p_{00} > 0$.

Proof. For $i = 1, \ldots, n$ define

$$N_{11} = \sum_{i=1}^{n} \mathbf{1}[r_i \geq t, Y_i = 1] \qquad N_{01} = \sum_{i=1}^{n} \mathbf{1}[r_i < t, Y_i = 1]$$

$$N_{10} = \sum_{i=1}^{n} \mathbf{1}[r_i \geq t, Y_i = 0] \qquad N_{00} = \sum_{i=1}^{n} \mathbf{1}[r_i < t, Y_i = 0]$$

with $N_{11} + N_{10} + N_{01} + N_{00} = n$. The counts N_{11}, N_{10}, N_{01}, and N_{00} follow a multinomial distribution with corresponding probabilities,

$$p_{11} = P(r \geq t, Y = 1) \qquad p_{01} = P(r < t, Y = 1), \tag{B.15}$$

$$p_{10} = P(r \geq t, Y = 0) \qquad p_{00} = P(r < t, Y = 0), \tag{B.16}$$

which have maximum likelihood estimator $\hat{p}_{lk} = \frac{N_{lk}}{n}$ for $l = 1, 2$ and $j = 1, 2$. By the multivariate central limit theorem

$$\sqrt{n} \begin{Bmatrix} \hat{p}_{11} - p_{11} \\ \hat{p}_{10} - p_{10} \\ \hat{p}_{01} - p_{01} \\ \hat{p}_{00} - p_{00} \end{Bmatrix} \xrightarrow{d} N\left(\vec{0}, \Sigma\right),$$

where

$$\Sigma = \begin{Bmatrix} p_{11}(1-p_{11}) & -p_{11}p_{10} & -p_{11}p_{01} & -p_{11}p_{00} \\ -p_{11}p_{10} & p_{10}(1-p_{10}) & -p_{10}p_{01} & -p_{10}p_{00} \\ -p_{11}p_{01} & -p_{10}p_{01} & p_{01}(1-p_{01}) & -p_{01}p_{00} \\ -p_{11}p_{00} & -p_{10}p_{00} & -p_{01}p_{00} & p_{00}(1-p_{00}) \end{Bmatrix}.$$

The estimand of (B.9) can be expressed as a function of $\vec{p} = \{p_{11}, p_{10}, p_{01}, p_{00}\}$,

$$\begin{aligned} sNB(\vec{p}) &= \frac{p_{11}}{p_{11}+p_{01}} - \frac{C}{B}\left(\frac{p_{10}+p_{00}}{p_{11}+p_{01}}\right)\frac{p_{10}}{p_{10}+p_{00}} && \text{(B.17)} \\ &= \frac{p_{11}}{p_{11}+p_{01}} - \frac{C}{B}\frac{p_{10}}{p_{11}+p_{01}} \\ &= \frac{1}{p_{11}+p_{01}}[p_{11} - \frac{C}{B}p_{10}]. && \text{(B.18)} \end{aligned}$$

The gradient of (B.18) is

$$\nabla sNB_t(\vec{p}) = \begin{Bmatrix} \frac{p_{01}+\frac{C}{B}p_{10}}{(p_{11}+p_{01})^2} \\ \frac{-\frac{C}{B}}{p_{11}+p_{01}} \\ \frac{-p_{11}+\frac{C}{B}p_{10}}{(p_{11}+p_{01})^2} \\ 0 \end{Bmatrix}$$

Then by the invariance property of the maximum likelihood estimator and invoking the δ-method,

$$\sqrt{n}\left(\widehat{sNB}(\hat{\vec{p}}) - sNB(\vec{p})\right) \xrightarrow{d} N(0, \Sigma_2),$$

where

$$\begin{aligned}\Sigma_2 &= \nabla sNB(\vec{p})^T \Sigma \nabla sNB(\vec{p}) \\ &= \left(\frac{1}{p_{11}+p_{01}}\right)^2 \left[\left(\frac{C}{B}\right)^2 p_{10} + \frac{p_{11}\left(p_{01}+\frac{C}{B}p_{10}\right)^2}{(p_{11}+p_{01})^2} + \frac{p_{01}(\frac{C}{B}p_{10}-p_{11})^2}{(p_{11}+p_{01})^2}\right].\end{aligned}$$

□

B.2 Additional Simulation Results

B.2.1 Simulation Example 1

Table B.1 gives the parameters for Beta distributions used to simulate true risks used in simulation example 1. Events are simulated from a $Bern(p_i)$ distributions. The miscalibrated risk score is obtained by applying piecewise polynomial $f_1(p)$. Figure B.1 shows the miscalibration pattern.

Table B.1: True Risk Parameters for Simulation Example 1

Subpopulation	π_m	$E[p_i]$	α_m	β_m
Subpop. 1	0.34	0.05	1	19
Subpop. 2	0.33	0.15	1.5	8.5
Subpop. 3	0.33	0.5	1	1
Overall prevalence = 0.23				

$$f_1(p_i) = \begin{cases} 139p_i^{3.2} & : p_i \in [0, 0.1) \\ -0.1p_i^{-0.83} - 0.1 & : p_i \in [0.1, 0.54) \\ 0.4p_i^{2.5} + 0.5 & : p_i \in [0.52, 1] \end{cases}$$

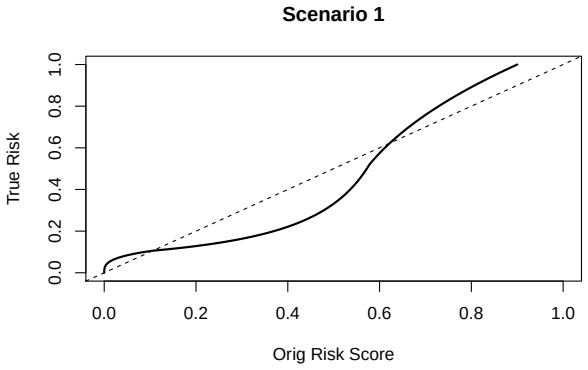


Figure B.1: Miscalibration setting for Simulation Example 1. Moderate risk scores are overestimated, while low and high risk scores tend towards underestimation.

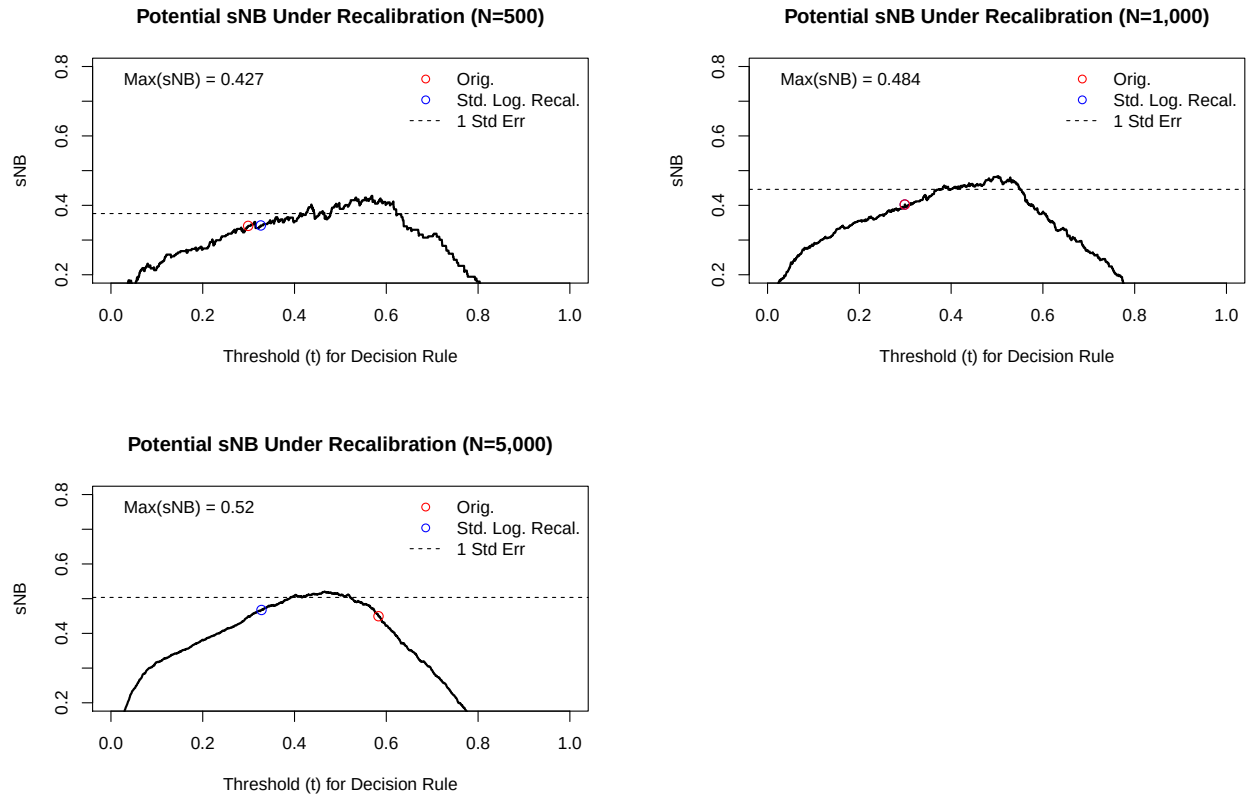


Figure B.2: Plot of potential sNB for simulation example 1, for differing training sample sizes. The dotted line shows a bandwidth of 1 standard error around the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error bandwidth, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

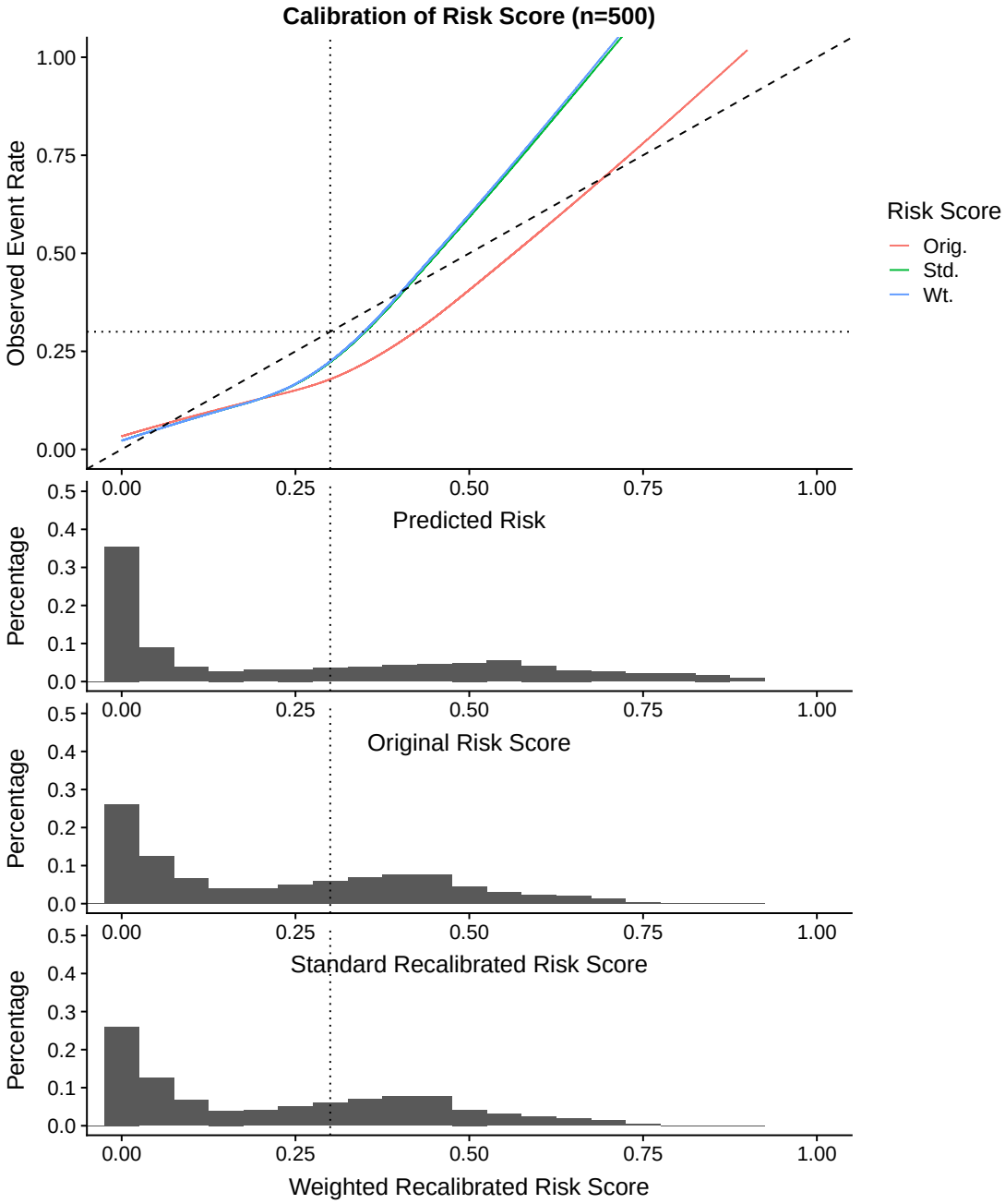


Figure B.3: Calibration curves for simulation example 1 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**.

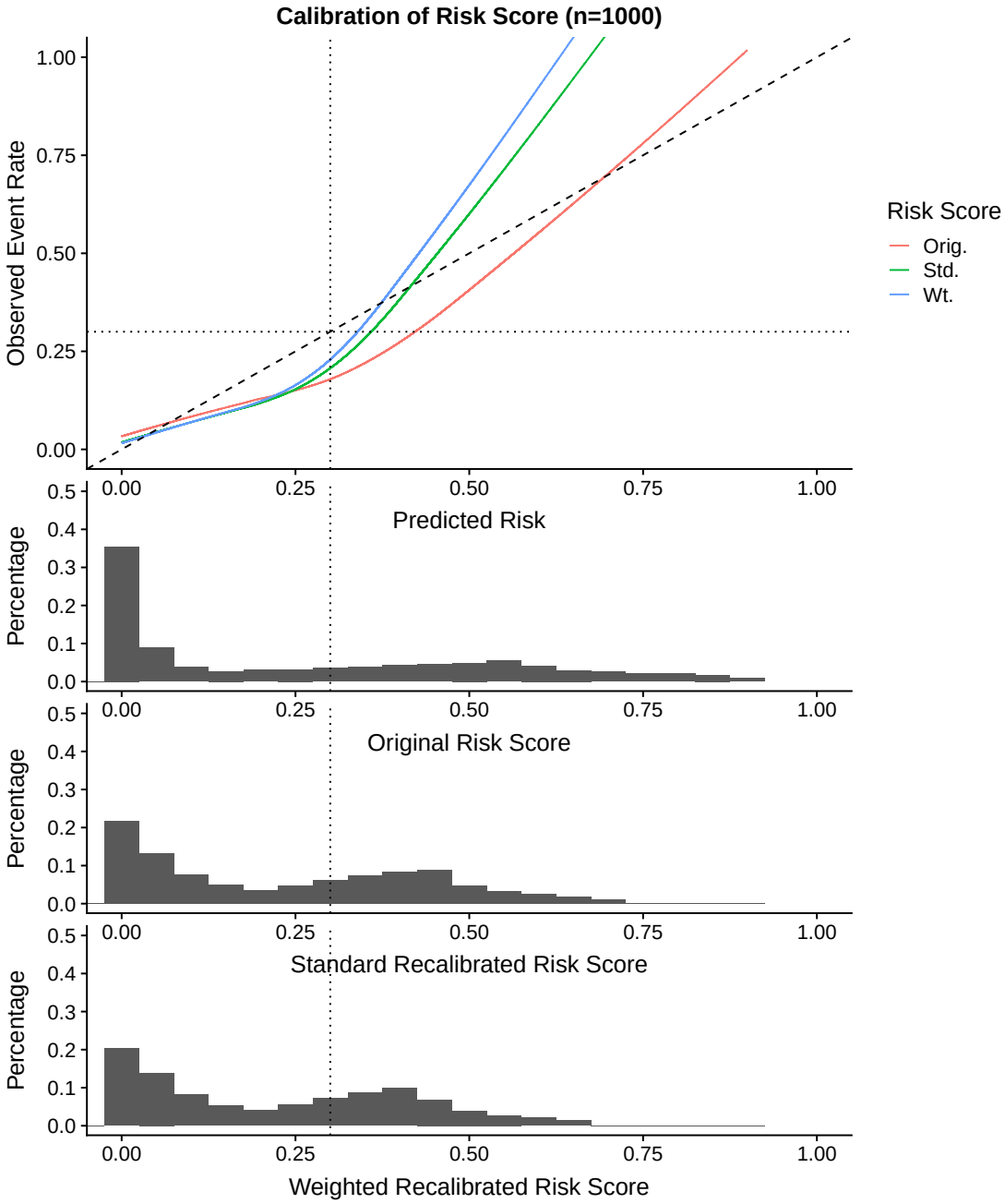


Figure B.4: Calibration curves for simulation example 1 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**.

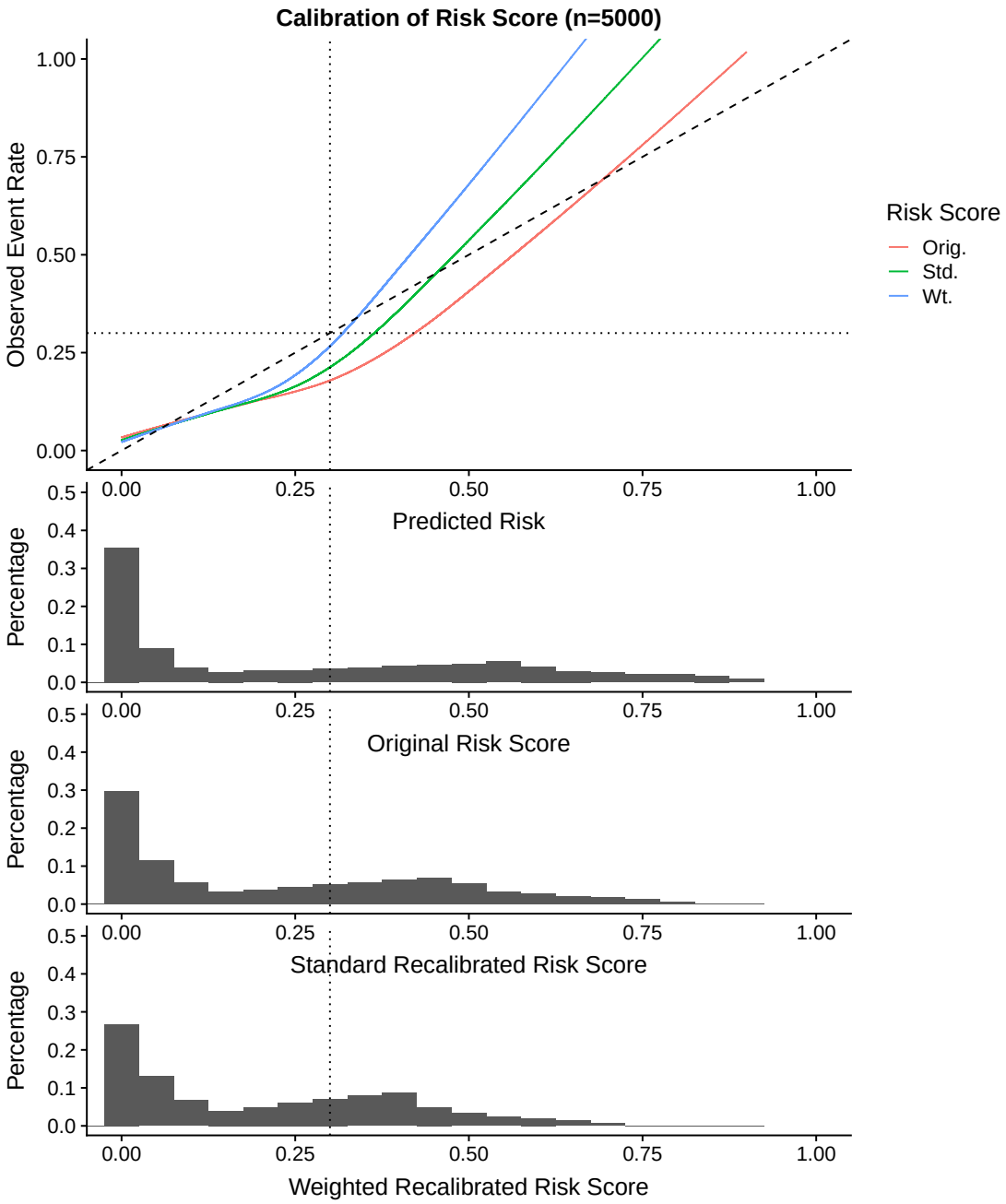


Figure B.5: Calibration curves for simulation example 1 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**.

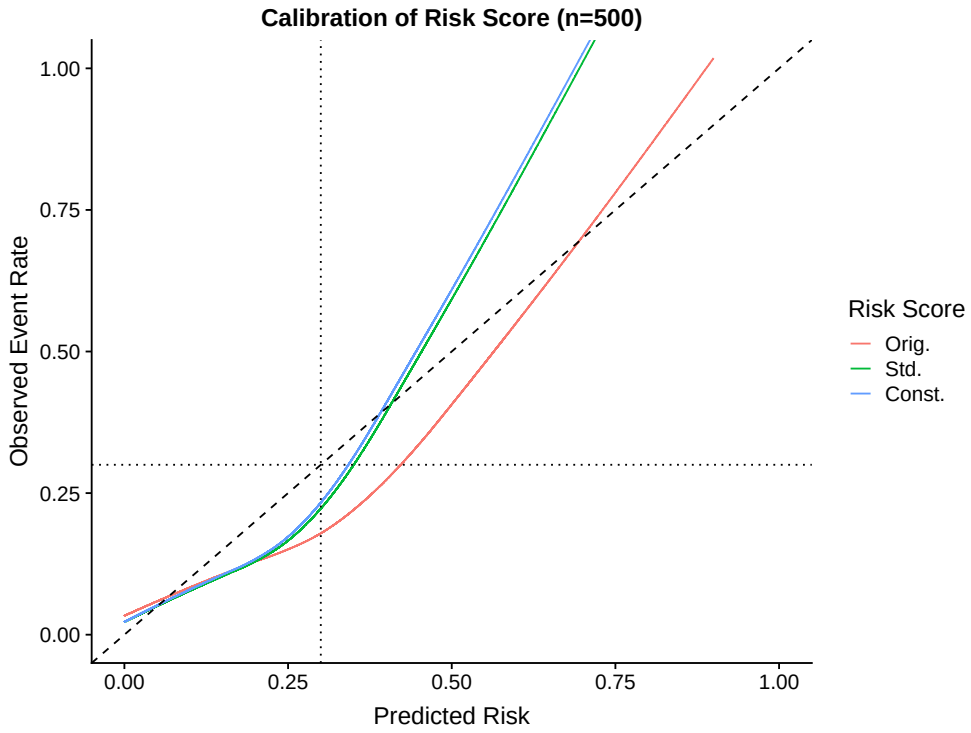


Figure B.6: Calibration curves for simulation example 1 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

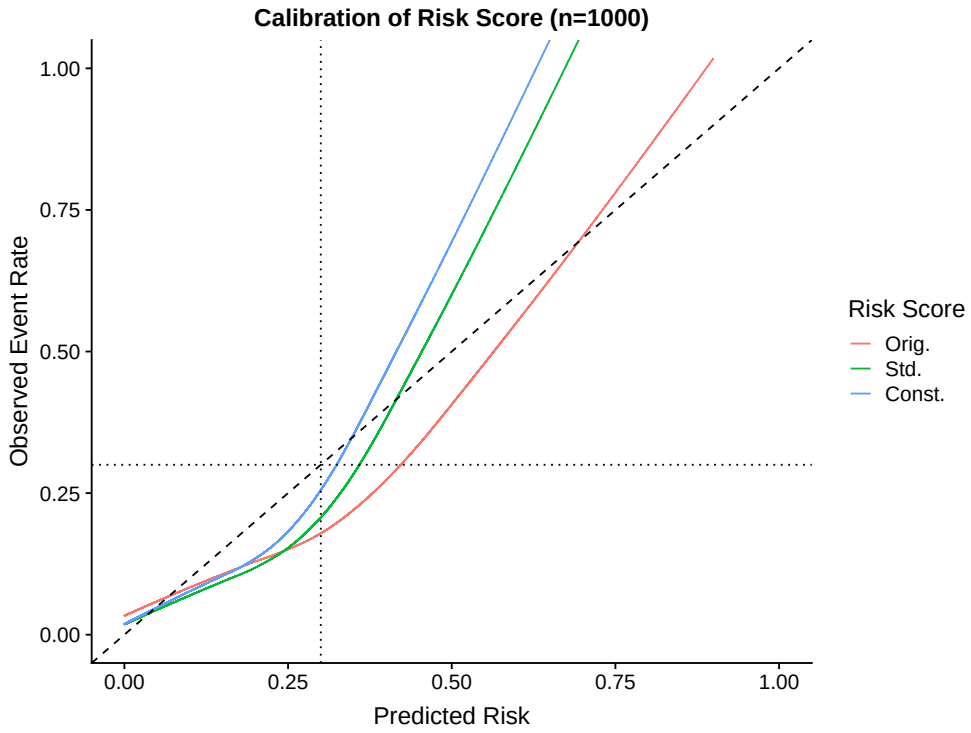


Figure B.7: Calibration curves for simulation example 1 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

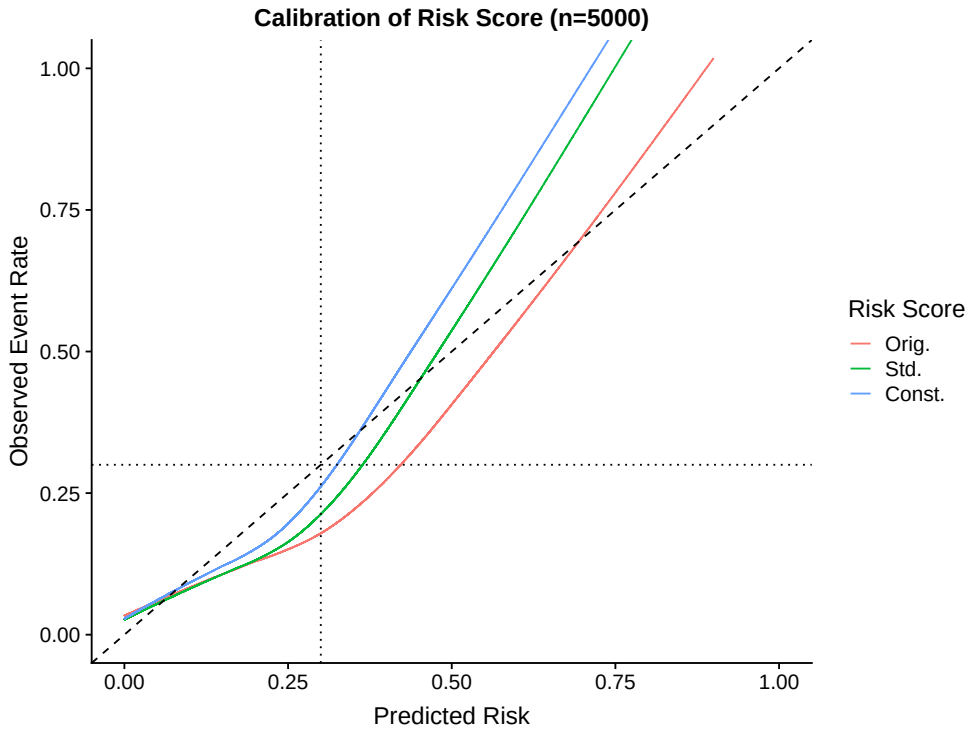


Figure B.8: Calibration curves for simulation example 1 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

B.2.2 Simulation Scenario 2

Table B.2 gives the parameters for Beta distributions used to simulate true risks used in simulation example 2. Events are simulated from a $Bern(p_i)$ distributions. The miscalibrated risk score is obtained by applying piecewise polynomial $f_2(p)$. Figure B.9 shows the miscalibration pattern.

Table B.2: True Risk Parameters for Simulation Example 2

Subpopulation	π_m	$E[p_i]$	α_m	β_m
Subpop. 1	0.4	0.05	1	19
Subpop. 2	0.4	0.1	2	18
Subpop. 3	0.2	0.5	1	1
Overall prevalence = 0.16				

$$f_2(p_i) = \begin{cases} 139p_i^{3.2} + 0.07 & : p_i \in [0, 0.15) \\ -0.1p_i^{-0.83} + 0.869 & : p_i \in [0.15, 0.54) \\ 0.38p_i^2.5 + 0.62 & : p_i \in [0.54, 1] \end{cases}$$

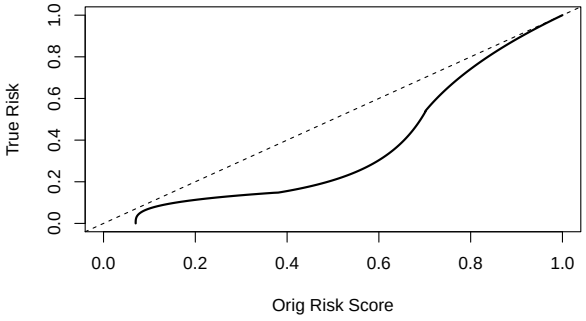


Figure B.9: Miscalibration setting for simulation example 2. All risk are overestimated, with moderate risks having worse miscalibration

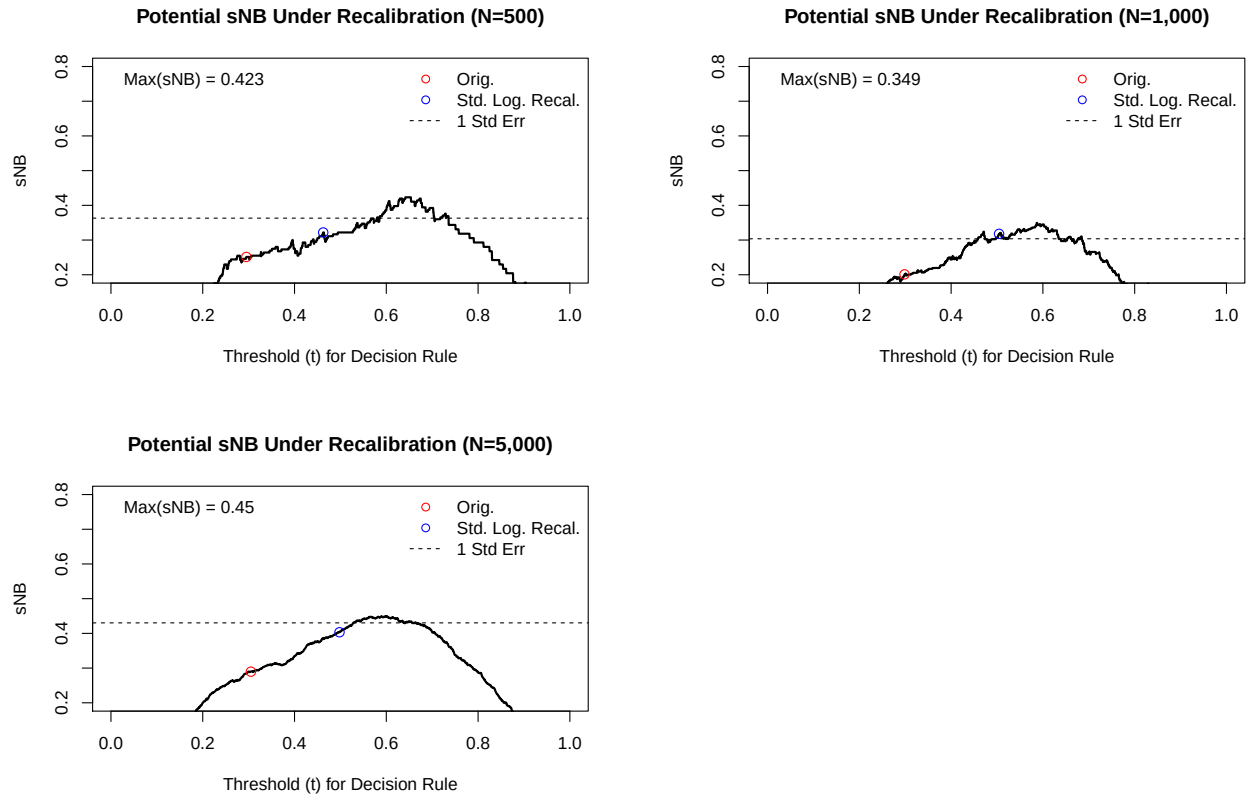


Figure B.10: Plot of potential sNB for simulation example 2, for differing training sample sizes. The dotted line shows a bandwidth of 1 standard error around the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error bandwidth, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

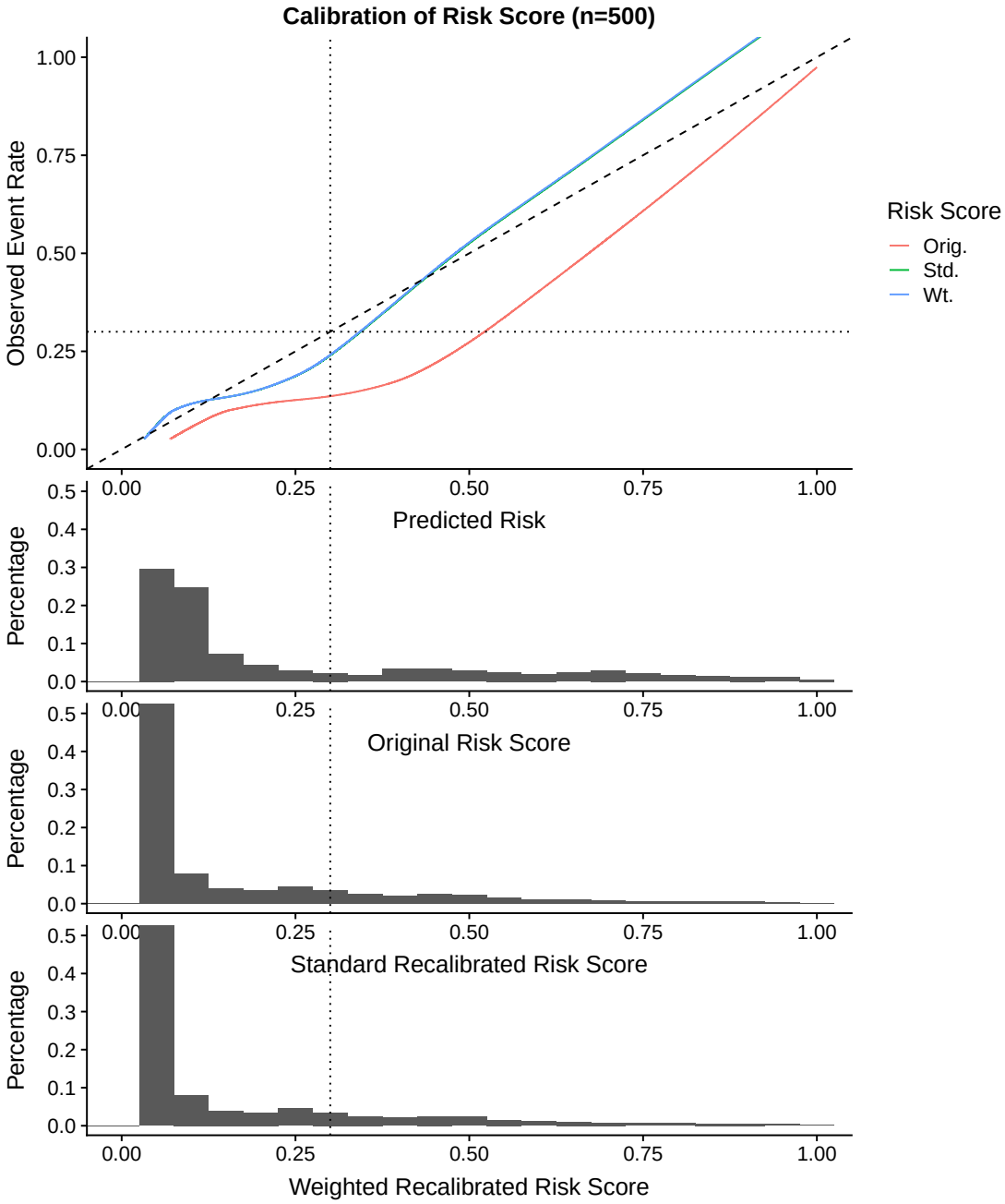


Figure B.11: Calibration curves for simulation example 2 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely.

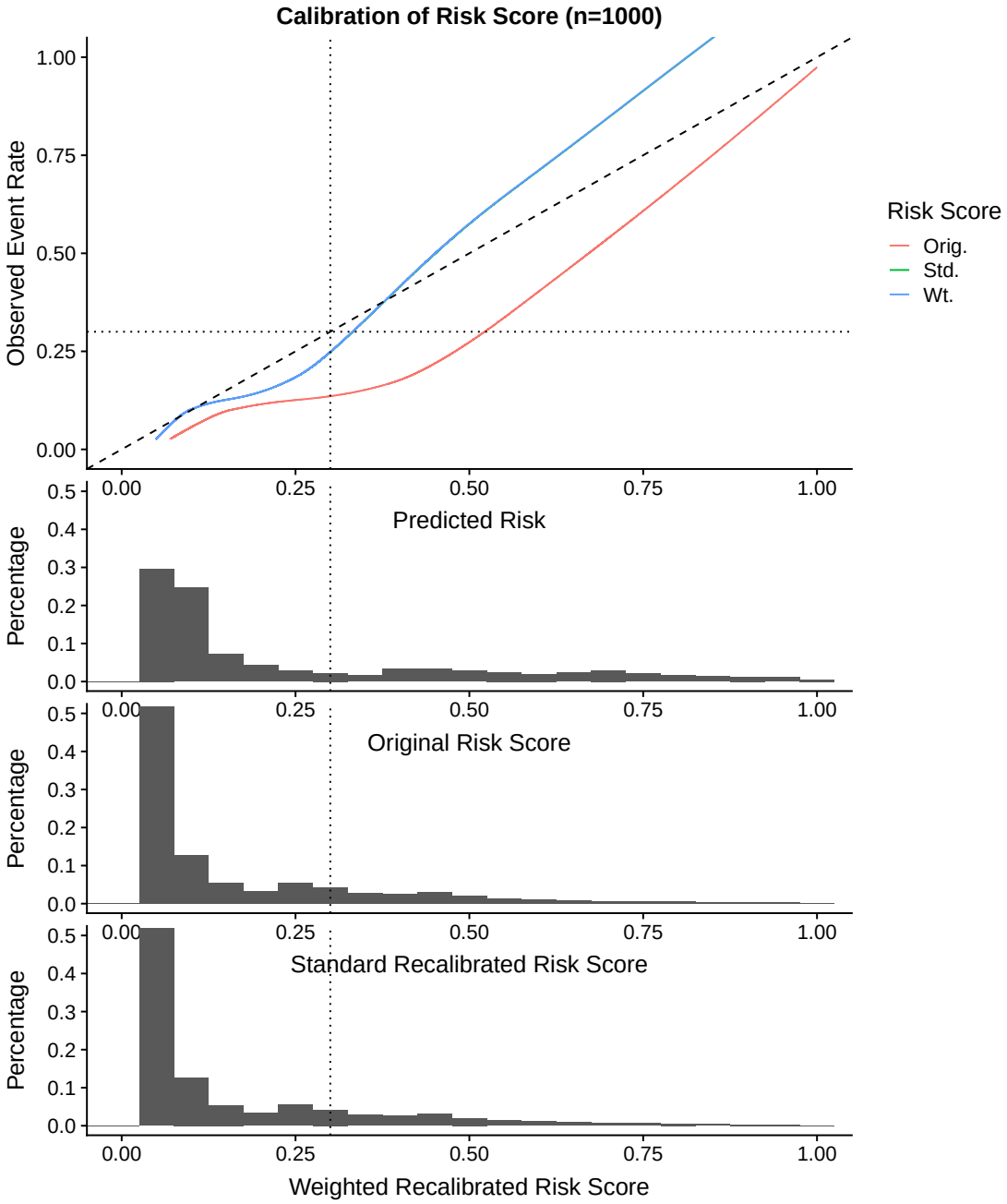


Figure B.12: Calibration curves for simulation example 2 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely.

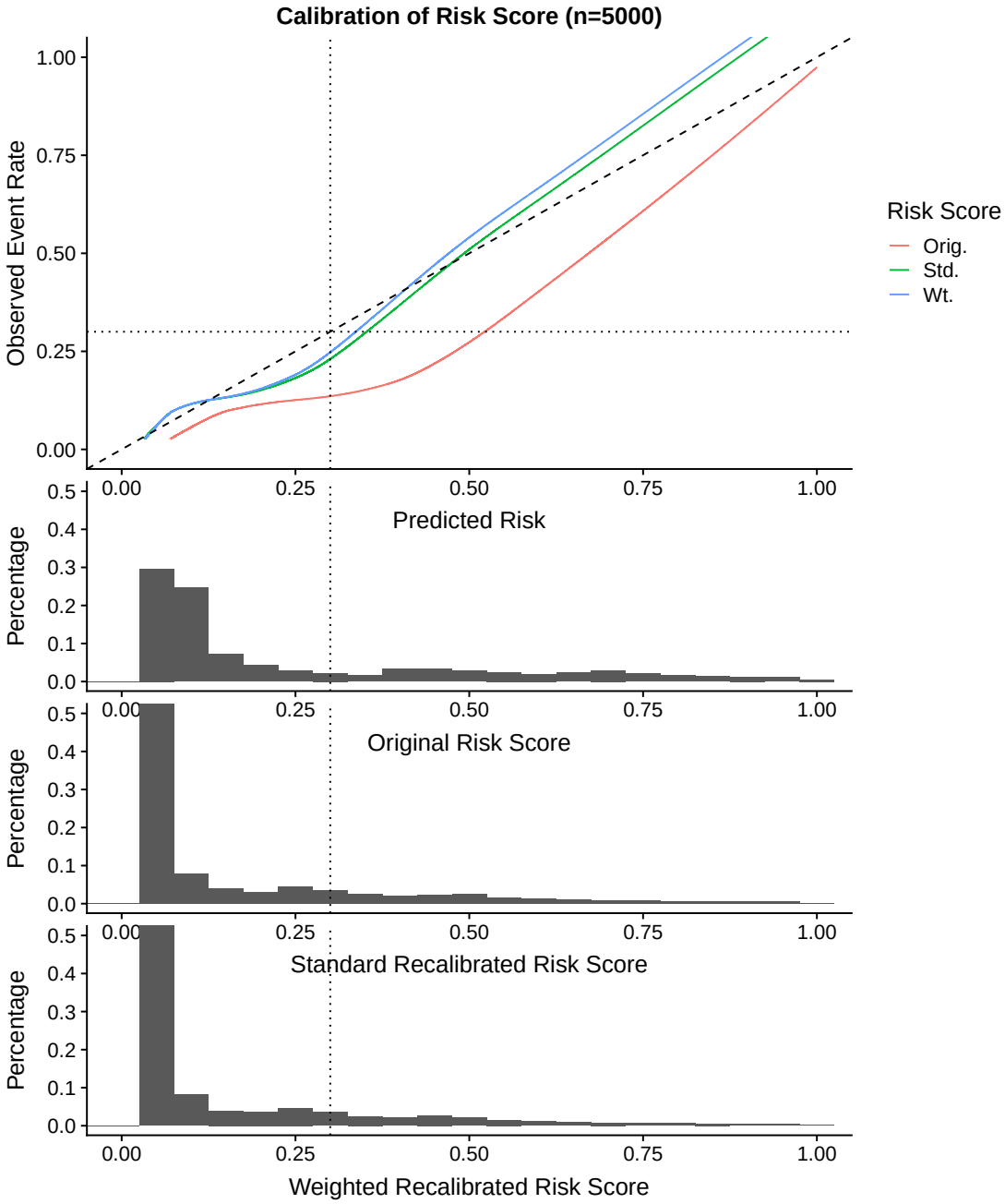


Figure B.13: Calibration curves for simulation example 2 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**.

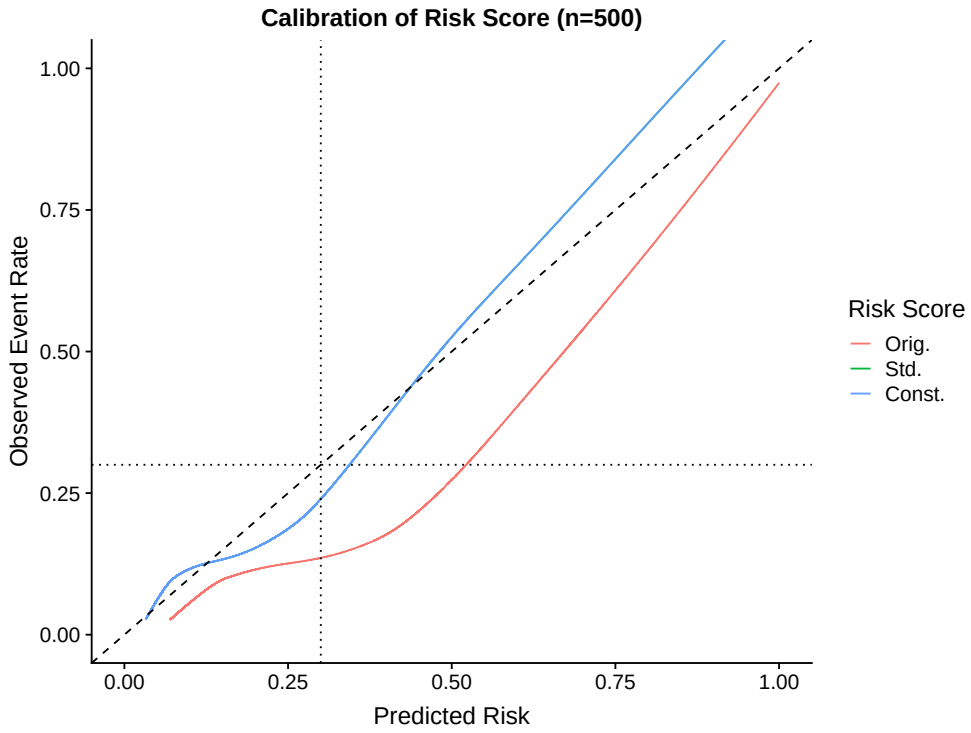


Figure B.14: Calibration curves for simulation example 2 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely.

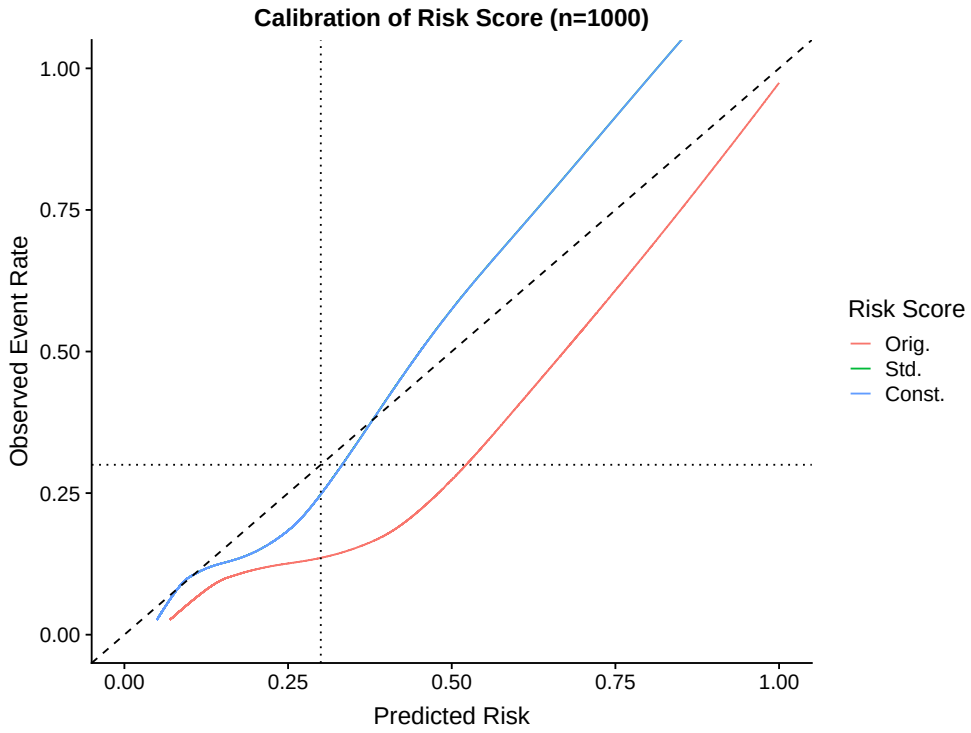


Figure B.15: Calibration curves for simulation example 2 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely.

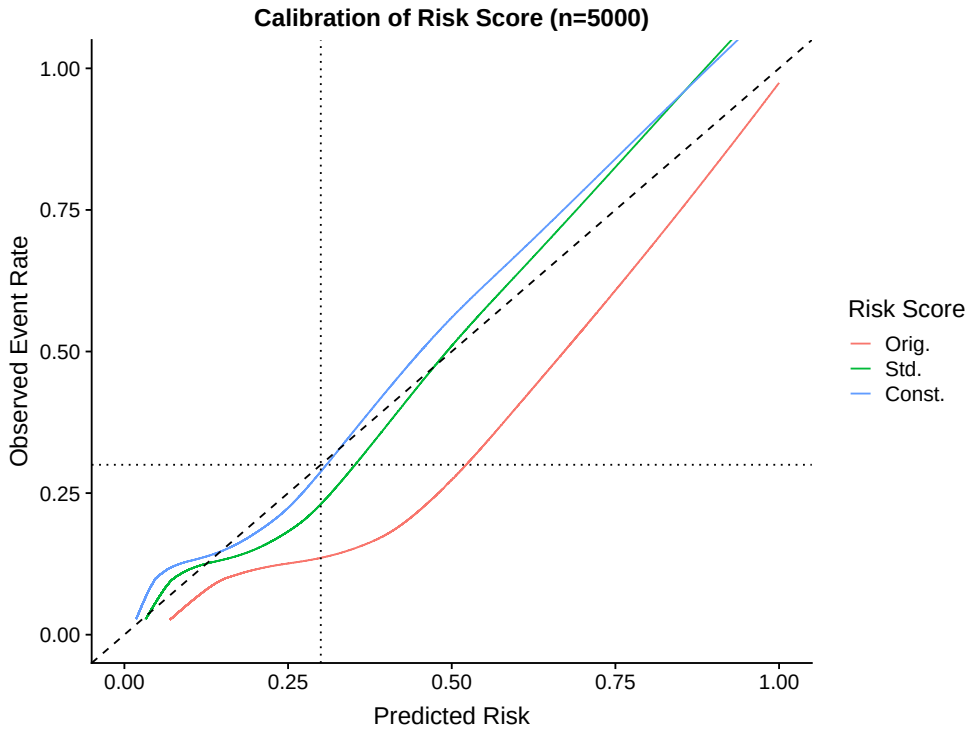


Figure B.16: Calibration curves for simulation example 2 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

B.2.3 Simulation Example 3

Table B.3 gives the parameters for Beta distributions used to simulate true risks used simulation example 3. Events are simulated from a $Bern(p_i)$ distributions. The miscalibrated risk score is obtained by applying piecewise polynomial $f_3(p)$. Figure B.17 shows the miscalibration pattern.

Table B.3: True Risk Parameters for Simulation Example 3

Subpopulation	π_m	$E[p_i]$	α_m	β_m
Subpop. 1	0.33	0.03	0.6	19.4
Subpop. 2	0.34	0.05	0.5	9.5
Subpop. 3	0.33	0.5	4	4
Overall prevlance = 0.19				

$$f_3(p_i) = \begin{cases} 0.3(p+0.003)^{0.26} - 0.05 & : p_i \in [0, 0.1) \\ 4.5p^{4.37} + 0.116 & : p_i \in [0.1, 0.6) \\ -0.2p^{-2.15} + 1.2 & : p_i \in [0.6, 1] \end{cases}$$

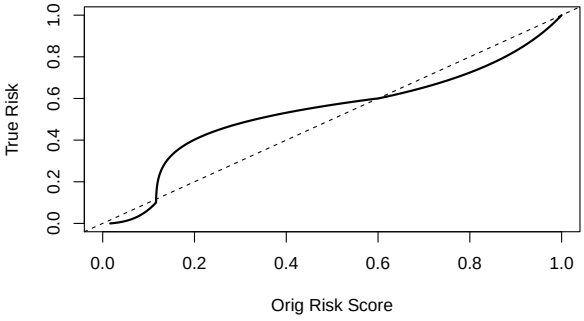


Figure B.17: Miscalibration setting for simulation example 3. Moderate risks are underestimated, while high and low risks are overestimated

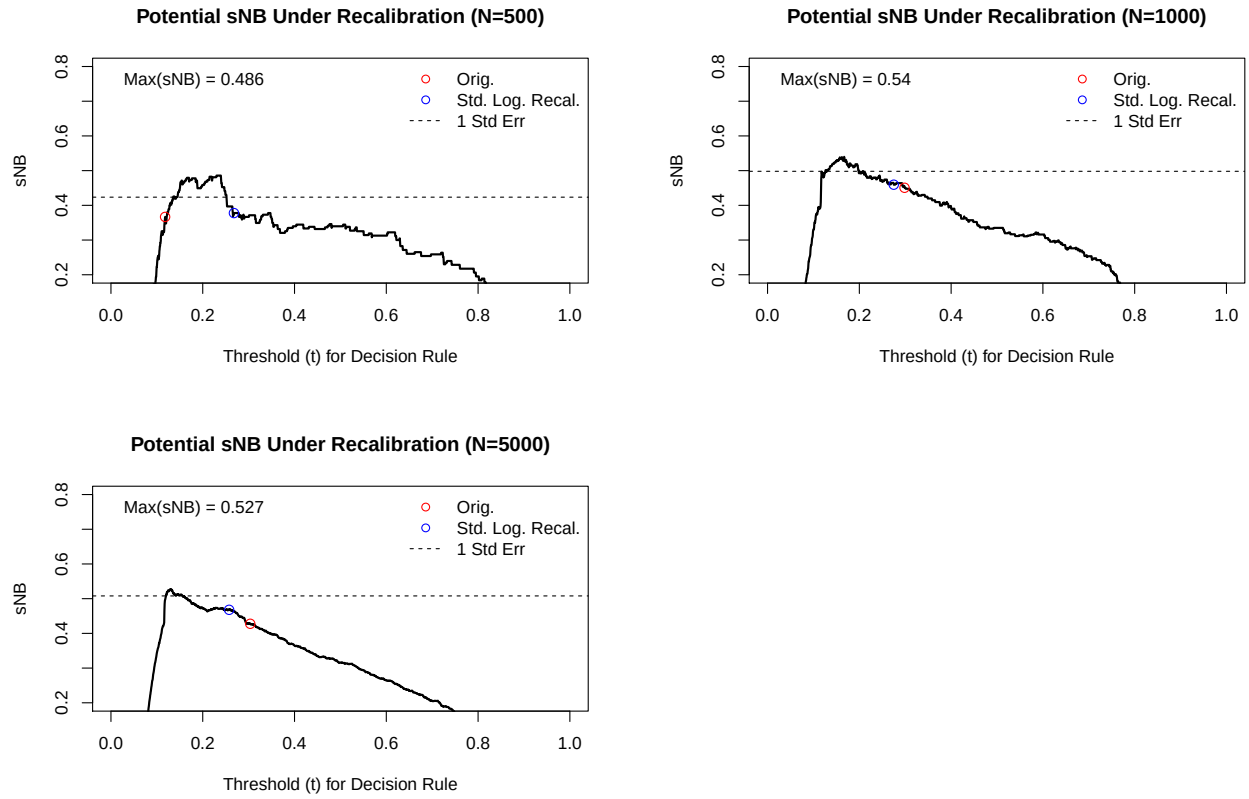


Figure B.18: Plot of potential sNB for simulation example 4, for differing training sample sizes. The dotted line shows a bandwidth of 1 standard error around the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error bandwidth, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

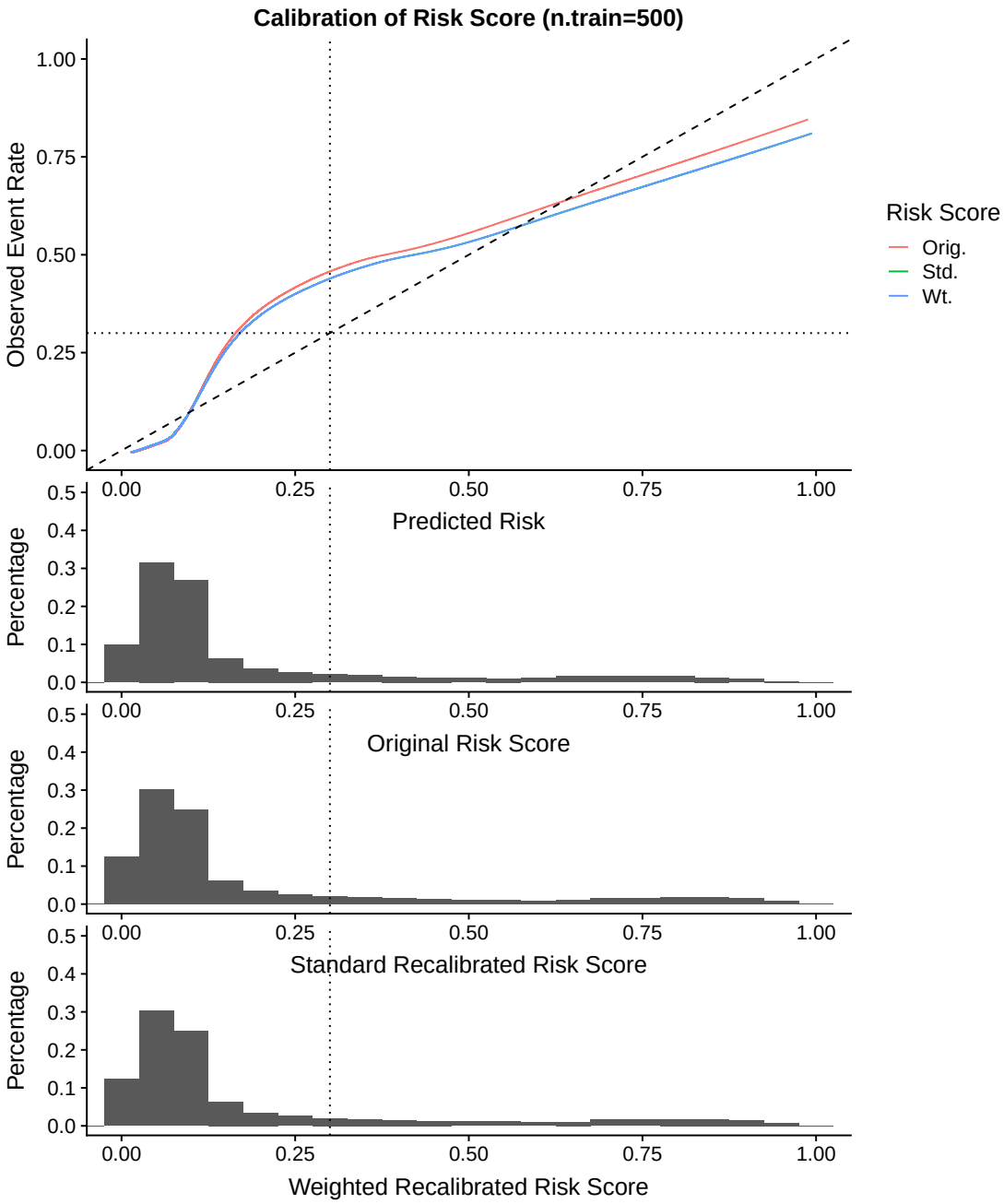


Figure B.19: Calibration curves for simulation example 3 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**. The calibration curves for standard logistic recalibration (green) and the original risk score (red) overlap completely.

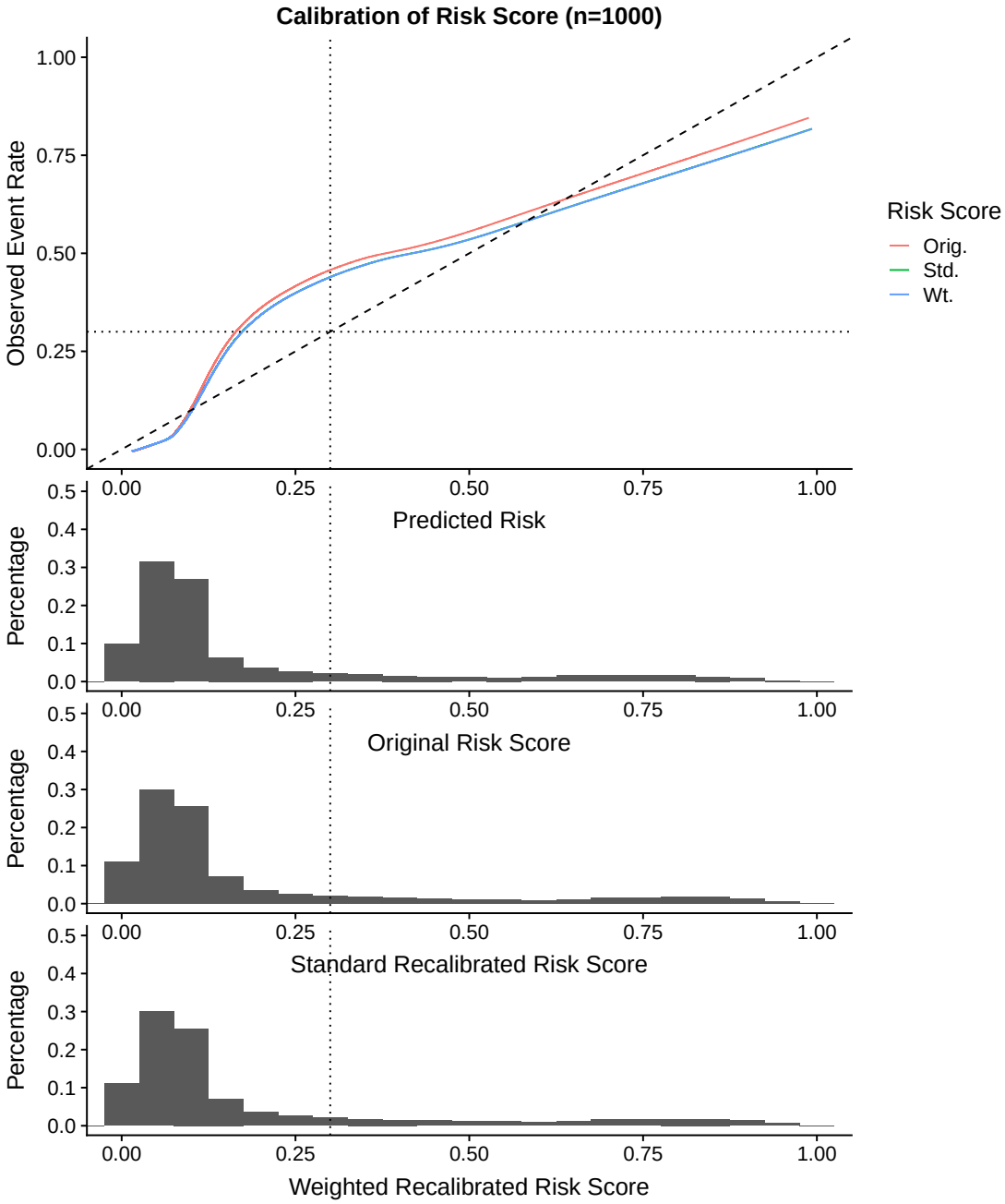


Figure B.20: Calibration curves for simulation example 3 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely

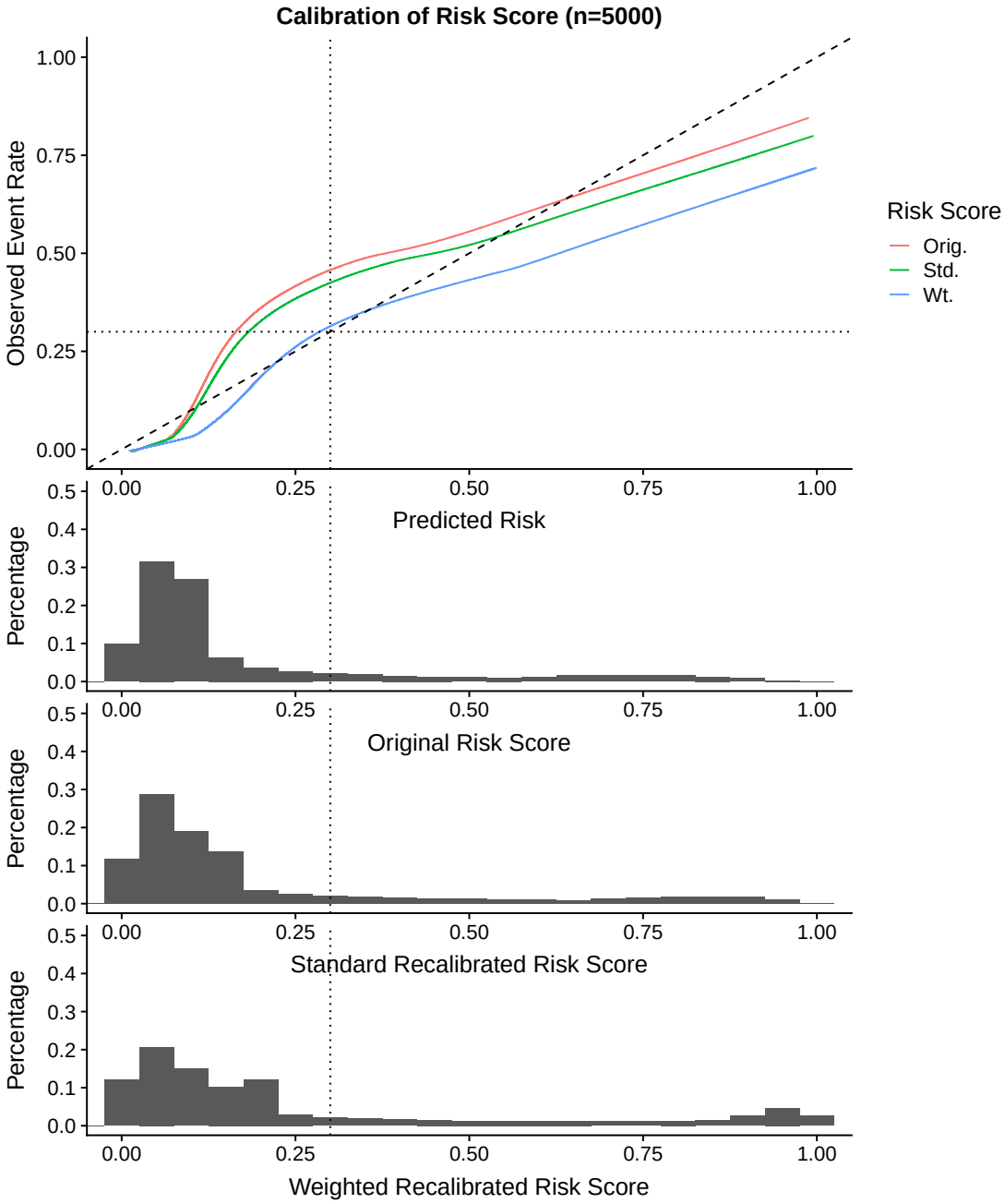


Figure B.21: Calibration curves for simulation example 3 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**.

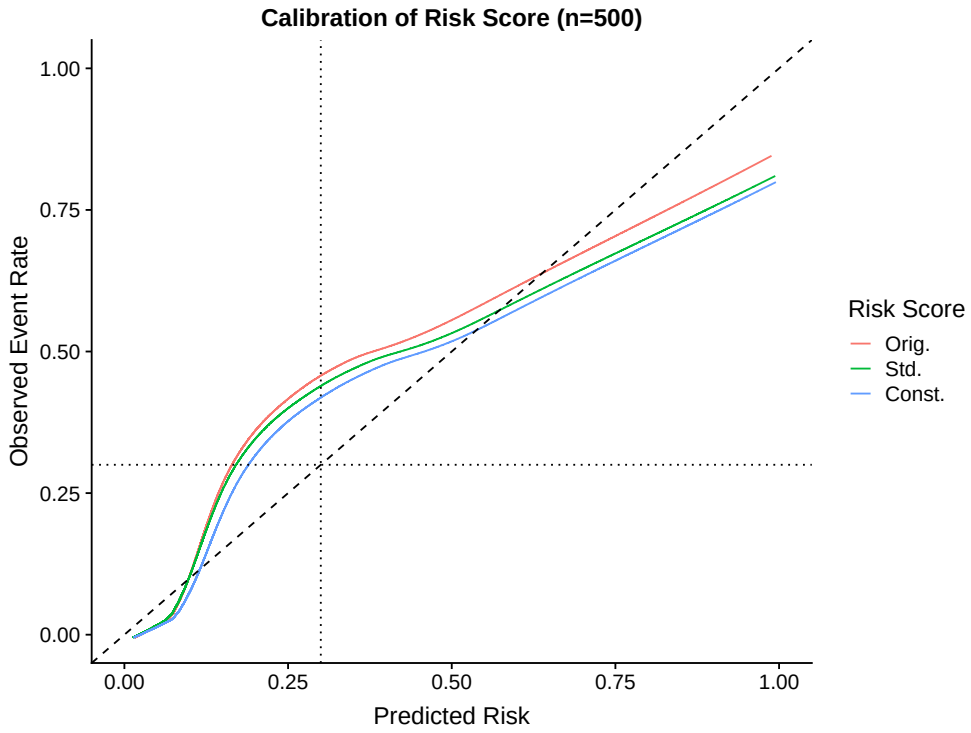


Figure B.22: Calibration curves for simulation example 3 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

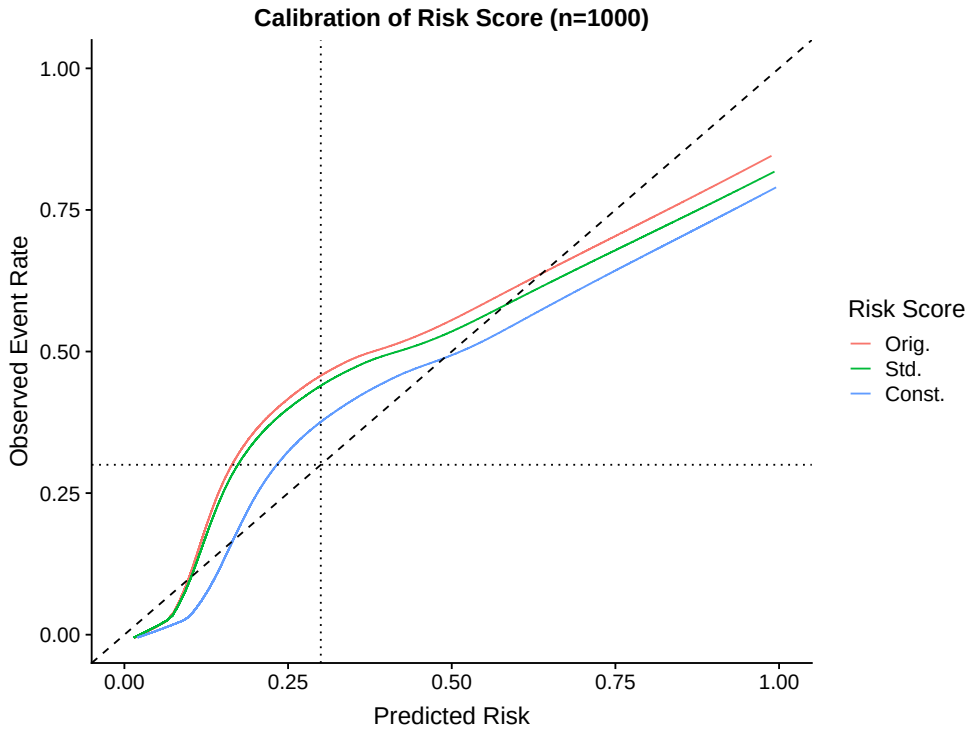


Figure B.23: Calibration curves for simulation example 3 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

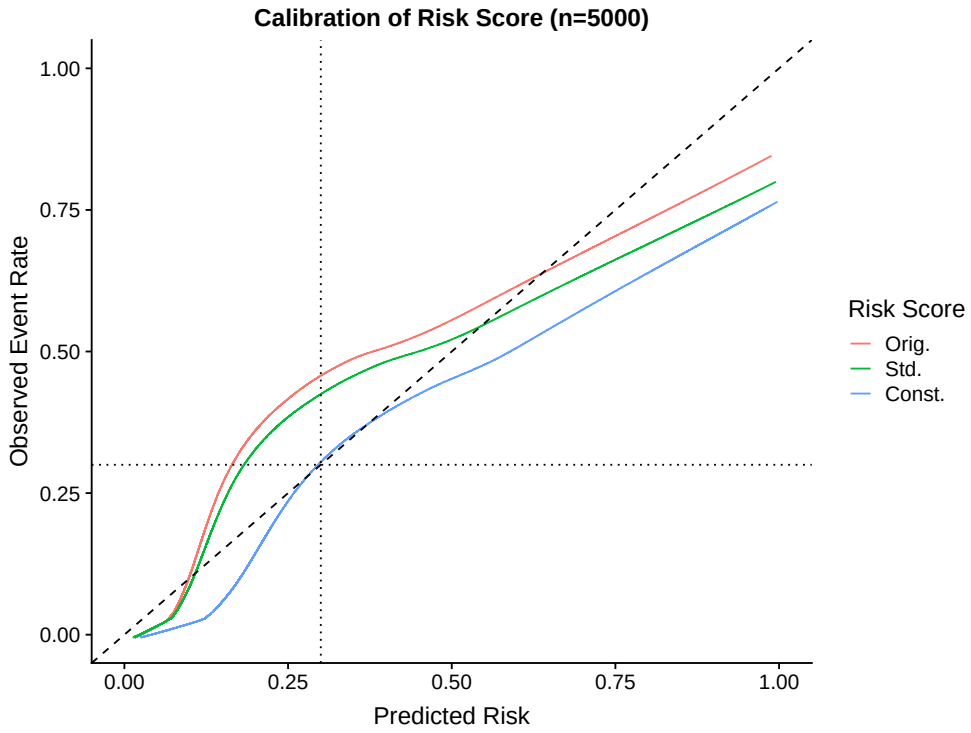


Figure B.24: Calibration curves for simulation example 3 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

B.2.4 Simulation Example 4

Table B.4 gives the parameters for Beta distributions used to simulate true risks used simulation example 4. Events are simulated from a $Bern(p_i)$ distributions. The miscalibrated risk score is obtained by applying piecewise polynomial $f_4(p)$. Figure B.25 shows the miscalibration pattern.

Table B.4: True Risk Parameters for Simulation Example 4

Subpopulation	π_m	$E[p_i]$	α_m	β_m
Subpop. 1	0.34	0.05	0.5	9.5
Subpop. 2	0.33	0.15	1	8.5
Subpop. 3	0.33	0.5	1	1
Overall prevalence = 0.23				

$$f_4(p_i) = \begin{cases} 0.3p^{0.8} & : p_i \in [0, 0.12) \\ 0.6p^{2.2} + 0.05 & : p_i \in [0.12, 0.46) \\ 1.7(p - 0.4)^{0.4} - 0.4 & : p_i \in [0.46, 1] \end{cases}$$

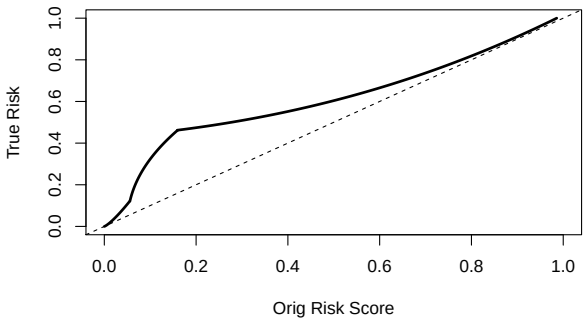


Figure B.25: Miscalibration setting for simulation example 4. All risk are underestimated, with moderate risks having worse miscalibration

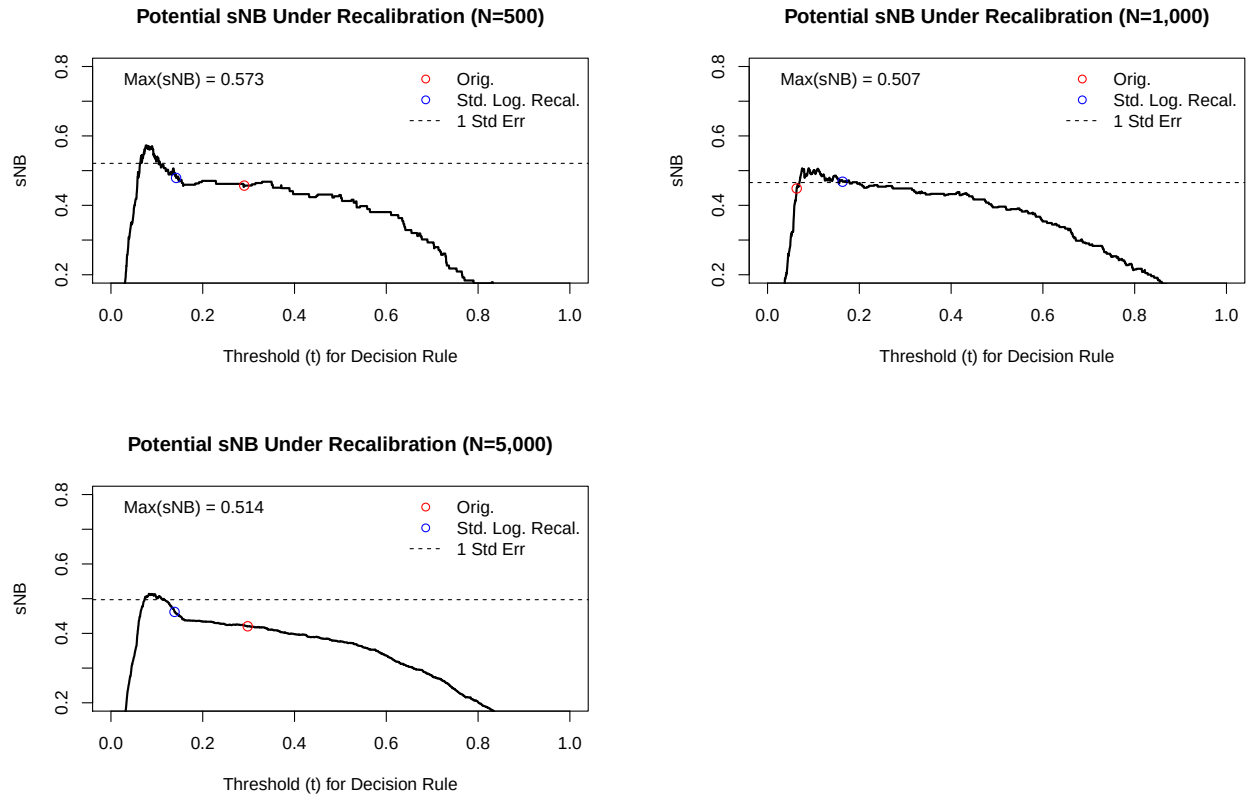


Figure B.26: Plot of potential sNB for simulation example 4, for differing training sample sizes. The dotted line shows a bandwidth of 1 standard error around the maximum attainable sNB. The estimated sNB for the standard recalibrated risk score is outside the standard error bandwidth, indicating that weighted recalibration could improve clinical utility beyond what is offered by standard logistic recalibration.

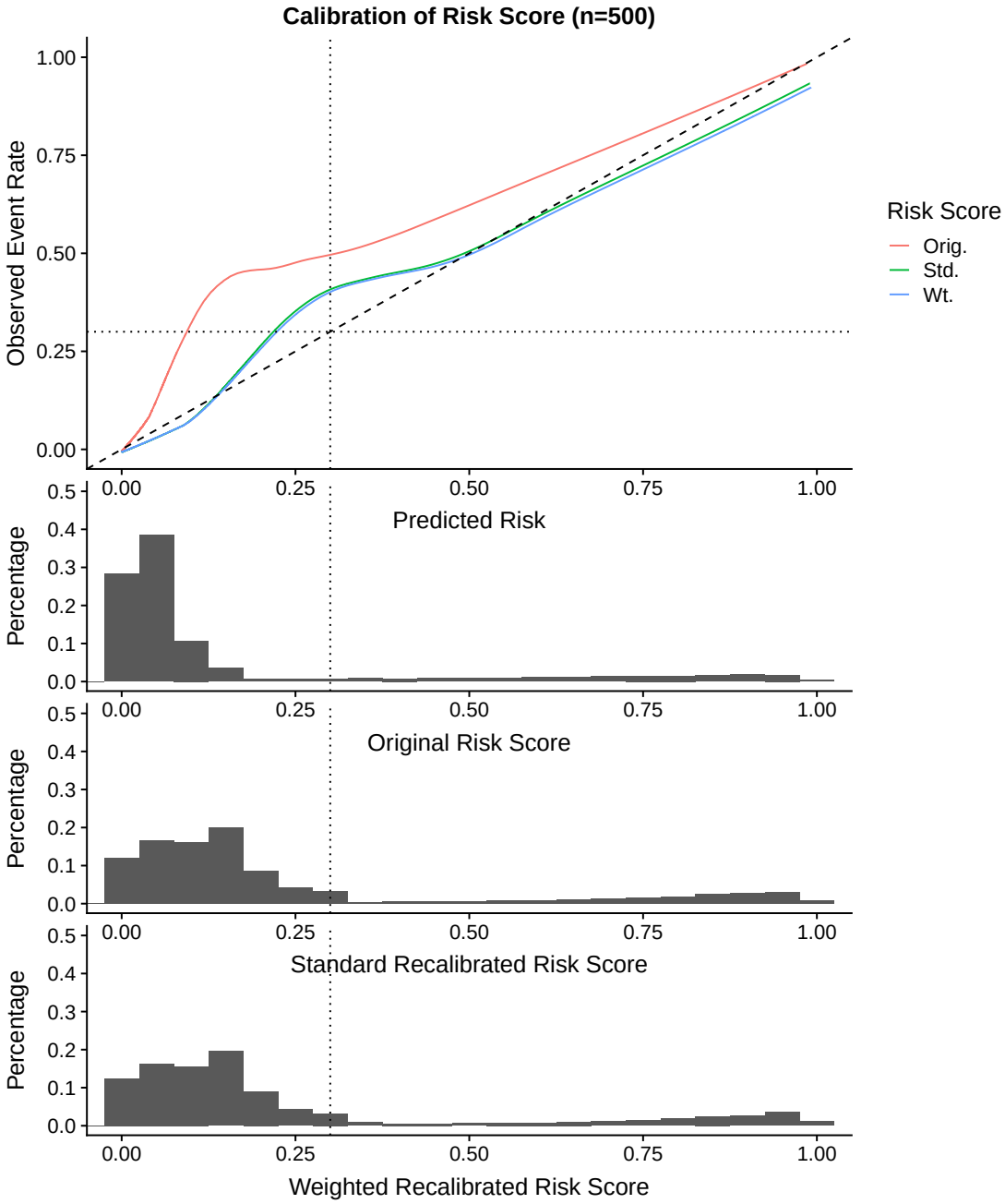


Figure B.27: Calibration curves for simulation example 4 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely.

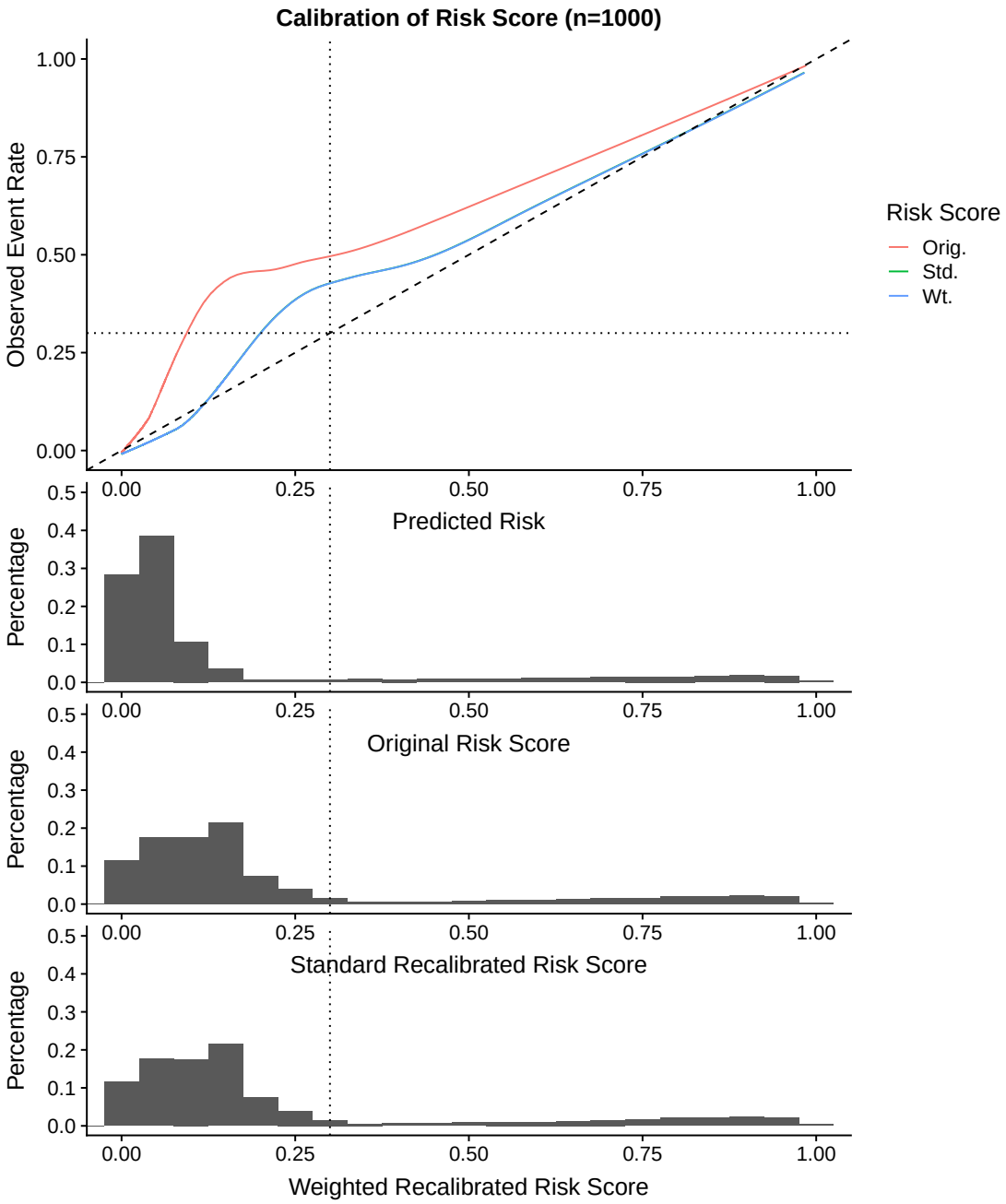


Figure B.28: Calibration curves for simulation example 4 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely.

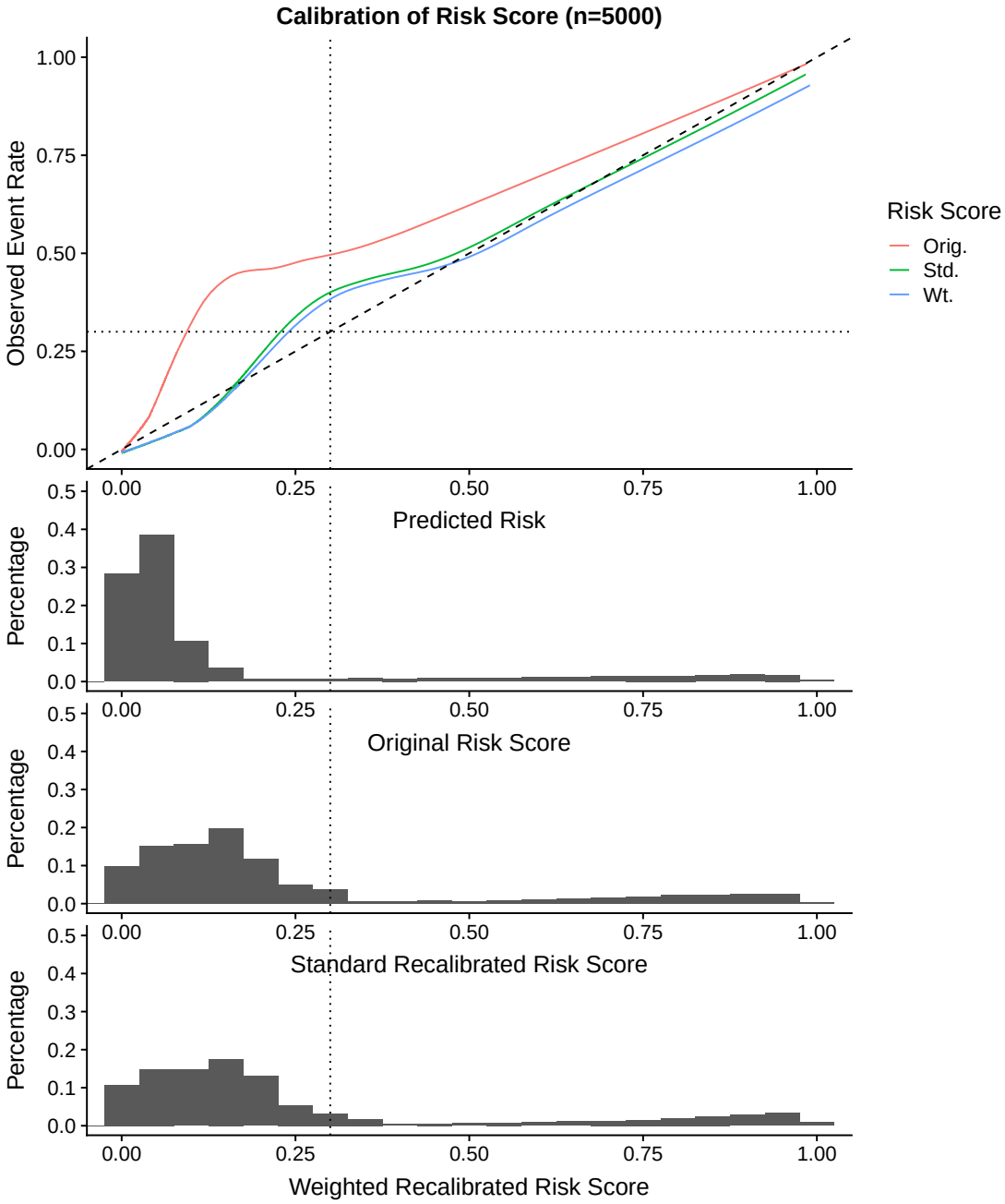


Figure B.29: Calibration curves for simulation example 4 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **weighted logistic recalibration**.

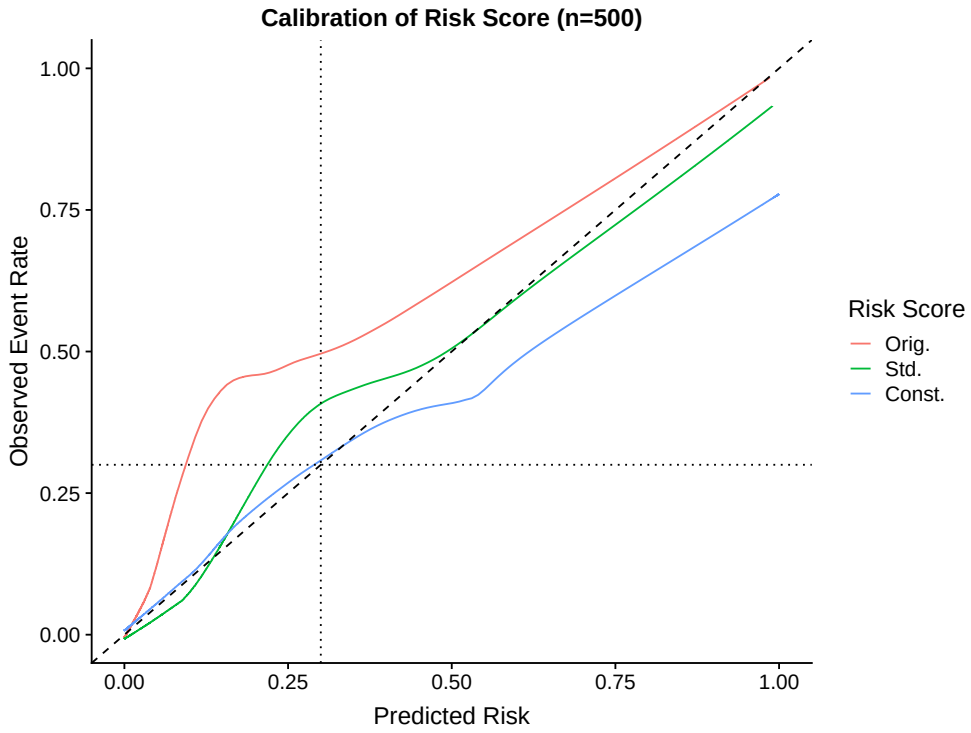


Figure B.30: Calibration curves for simulation example 4 with training sample size $N = 500$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

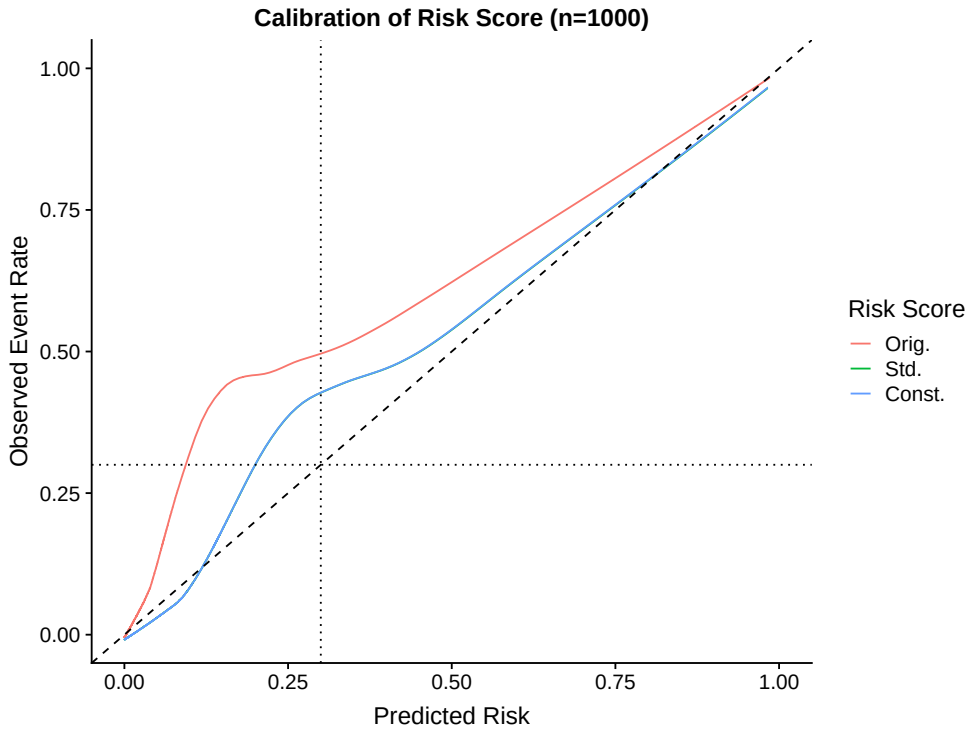


Figure B.31: Calibration curves for simulation example 4 with training sample size $N = 1000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**. The calibration curves for standard logistic recalibration (green) and weighted logistic regression (blue) overlap completely.

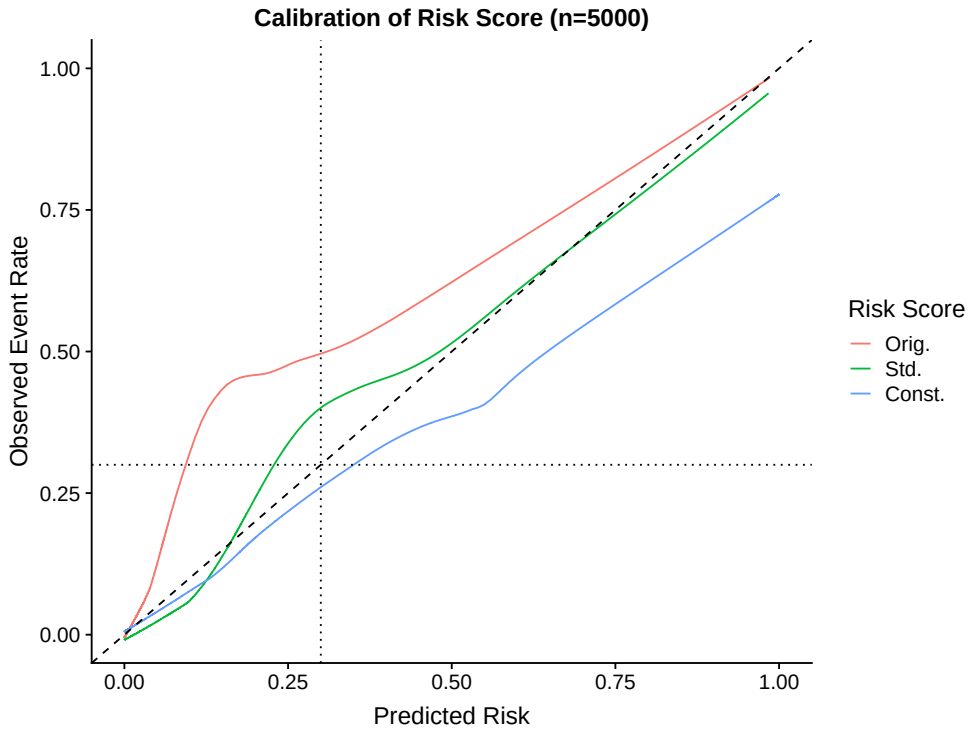


Figure B.32: Calibration curves for simulation example 4 with training sample size $N = 5000$. Comparison of calibration curves for the original risk model, risk model estimated from standard logistic recalibration, and risk model estimated from **constrained logistic recalibration**.

B.2.5 Validity of Cross-Validation Variance Estimator

Gelman et al. (1992) give a variance estimator of convergence diagnostic statistic used when Markov Chain Monte Carlo with multiple chains are performed. The variance estimator accounts for both the variability of the statistic 'within-in" a single chain, and the variance of the statistic across chains. Analogously, we can use this framework to estimate the "within" repetition variance (i.e. variation in sNB from a single round of K-fold cross-validation) and the "between" repetition variance. We denote the 'within" repetition variance as W and the "between" repetition variance as B. We augment this formula slightly from that given in Gelman et al. (1992) to account for the fact that as the number of cross-validation repetitions

increases, the between repetition variability should decrease. The proposed estimator is

$$\widehat{\sigma}(cvSNB(\lambda)) = \sqrt{\frac{K-1}{K}W(\lambda) + \frac{1}{M}B(\lambda)}, \tag{B.19}$$

where

$$W(\lambda) = \frac{1}{M}\sum_{m=1}^{M}\sigma_m^2 = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{K-1}\sum_{i=1}^{K}(cvSNB_{i,m}(\lambda) - cvSNB_m(\lambda))^2,$$

$$B(\lambda) = \frac{K}{M-1}\sum_{m=1}^{M}\left[cvSNB_m(\lambda) - cvSNB(\lambda)\right]^2.$$

We present a small simulation study to assess if the estimator in B.19 is a reliable estimate of $\sigma(cvSNB)$. Data are simulated following the simulation setting for scenario 1. We implement the cross-validation procedure and estimate the $\sigma(cvSNB(RAW))$ as described in Section 3.3.3, for RAW tuning parameters 0.1,0.2,...,0.9. We compare these estimates to a empirical estimates of $cvSNB(\lambda)$ obtained from 500 Monte Carlo simulations. Figure B.33 shows the proposed estimator of $\sigma(cvSNB)$ compared to the Monte Carlo estimates, for sample sizes of $n = 500$ and $n = 5000$. Note there is increased variance in $cvSNB(RAW)$ when RAW is small for small sample sizes. When the sample size is small, and heavy down-weighting is applied (i.e. low RAW parameter) the effective sample size is small. This leads to increased variation in the estimated recalibration parameter, $\vec{\alpha}$ which is propogated to estimates of sNB. The proposed variance estimate captures this trend.

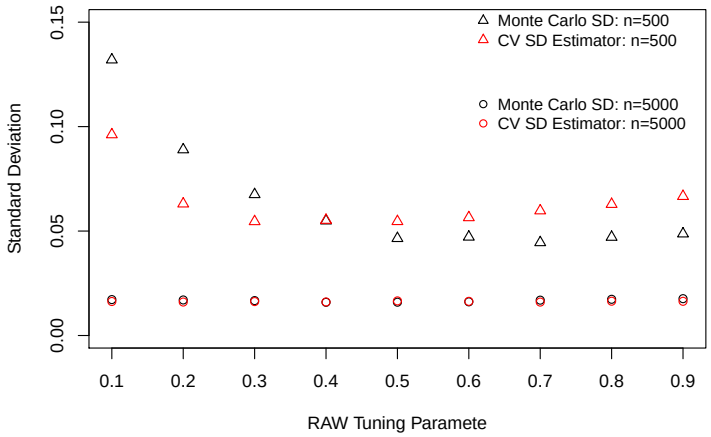


Figure B.33: Comparison of proposed estimator of cross-validation standard deviation and Monte Carlo estimates of standard deviation for simulation scenario 1. The black points give the empirical estimates of $\sigma(cvSNB)$ obtained from 500 Monte Carlo simulations. The red points show the proposed estimator of $\widehat{\sigma}(cvSNB)$. The different shapes denote different sample sizes.

B.3 Additional MESA Application Results

B.3.1 Calibration curves and plots of potential sNB under recalibration for race and gender cohorts in MESA

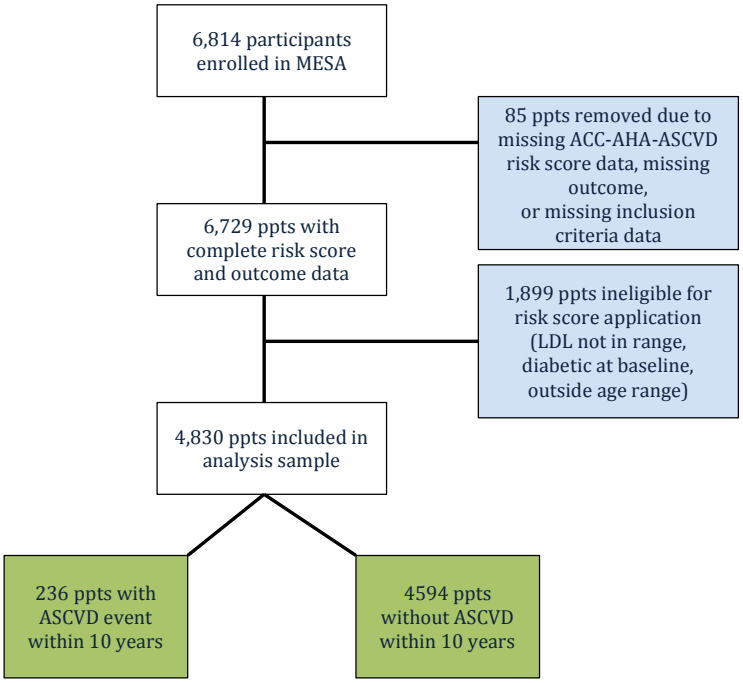


Figure B.34: Sample exclusions prior to data analysis. We excluded participants with missing ASCVD outcome data, data for the outcome, ACC-AHA-ASCVD risk score, or variables used to assess inclusion for the ACC-AHA-ASCVD risk score.

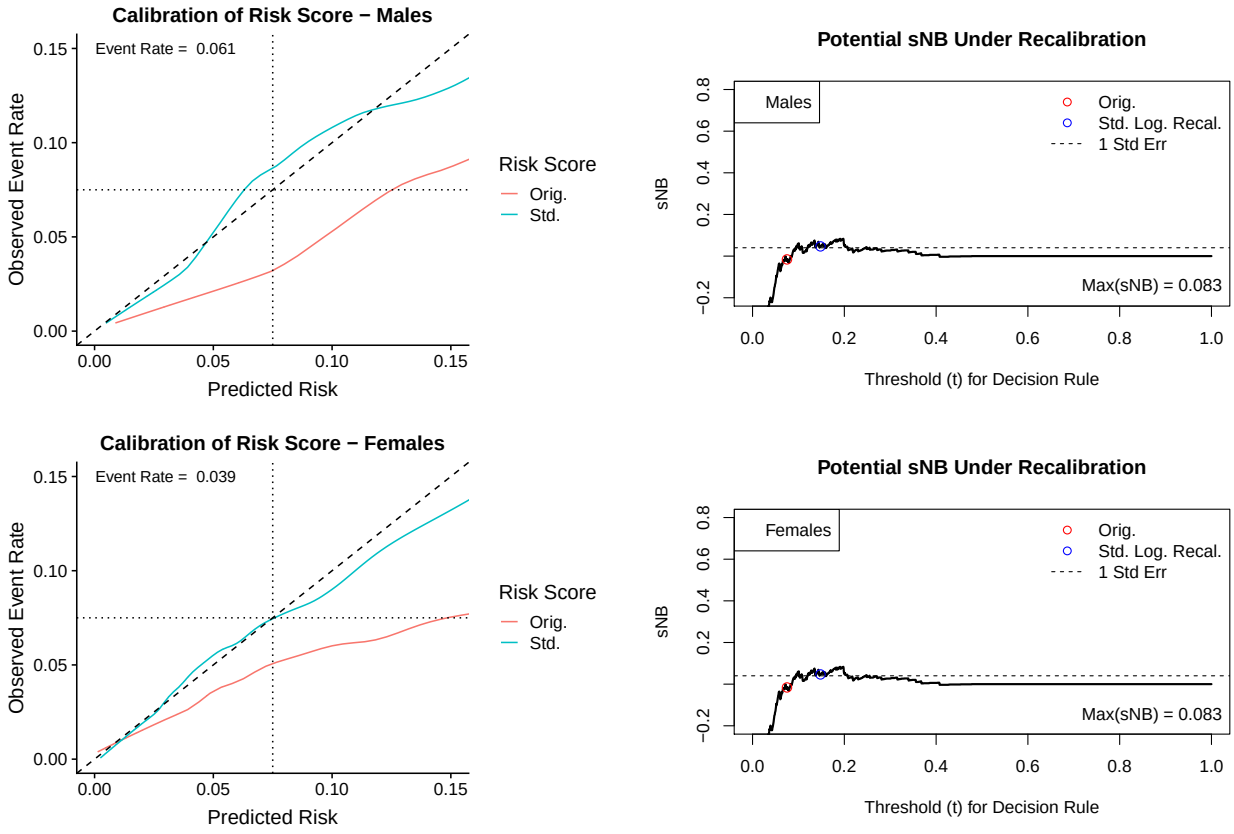


Figure B.35: Potential sNB curves and calibration curves by gender for ACC-AHA-ASCVD risk score in MESA analysis sample.

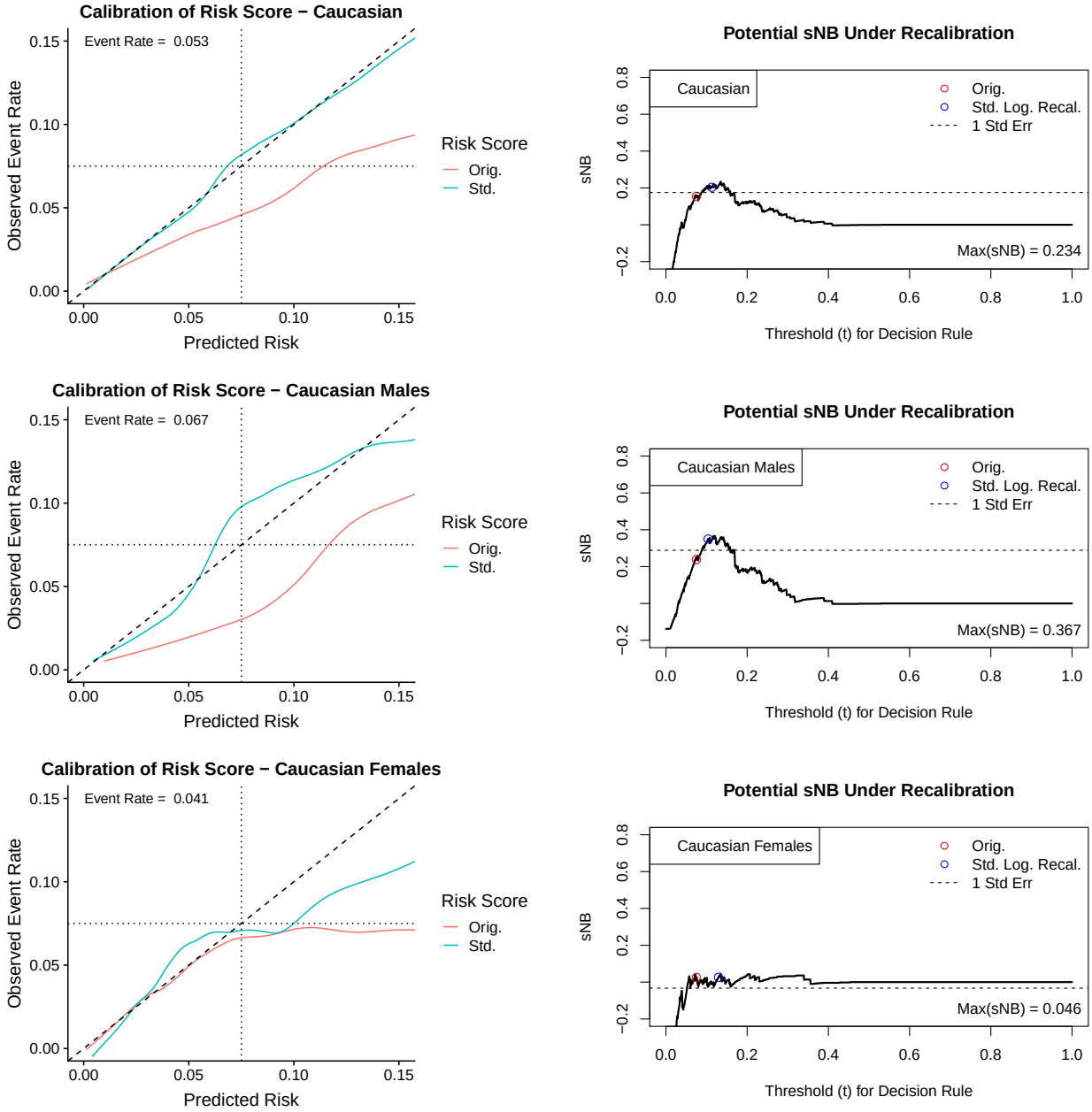


Figure B.36: Potential sNB curves and calibration curves by gender for ACC-AHA-ASCVD risk score in **Caucasian** MESA analysis sample.

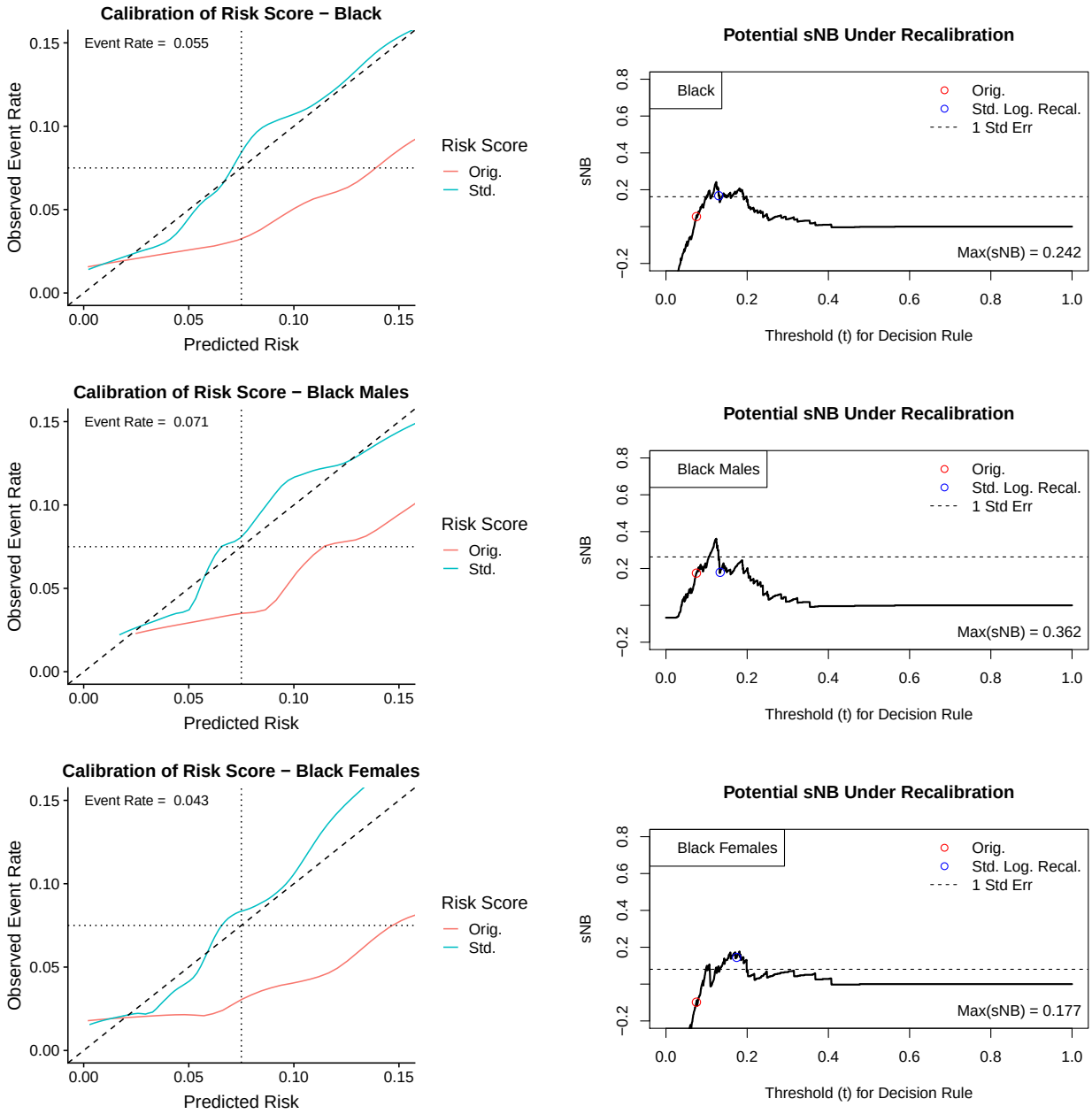


Figure B.37: Potential sNB curves and calibration curves by gender for ACC-AHA-ASCVD risk score in **Black** MESA analysis sample.

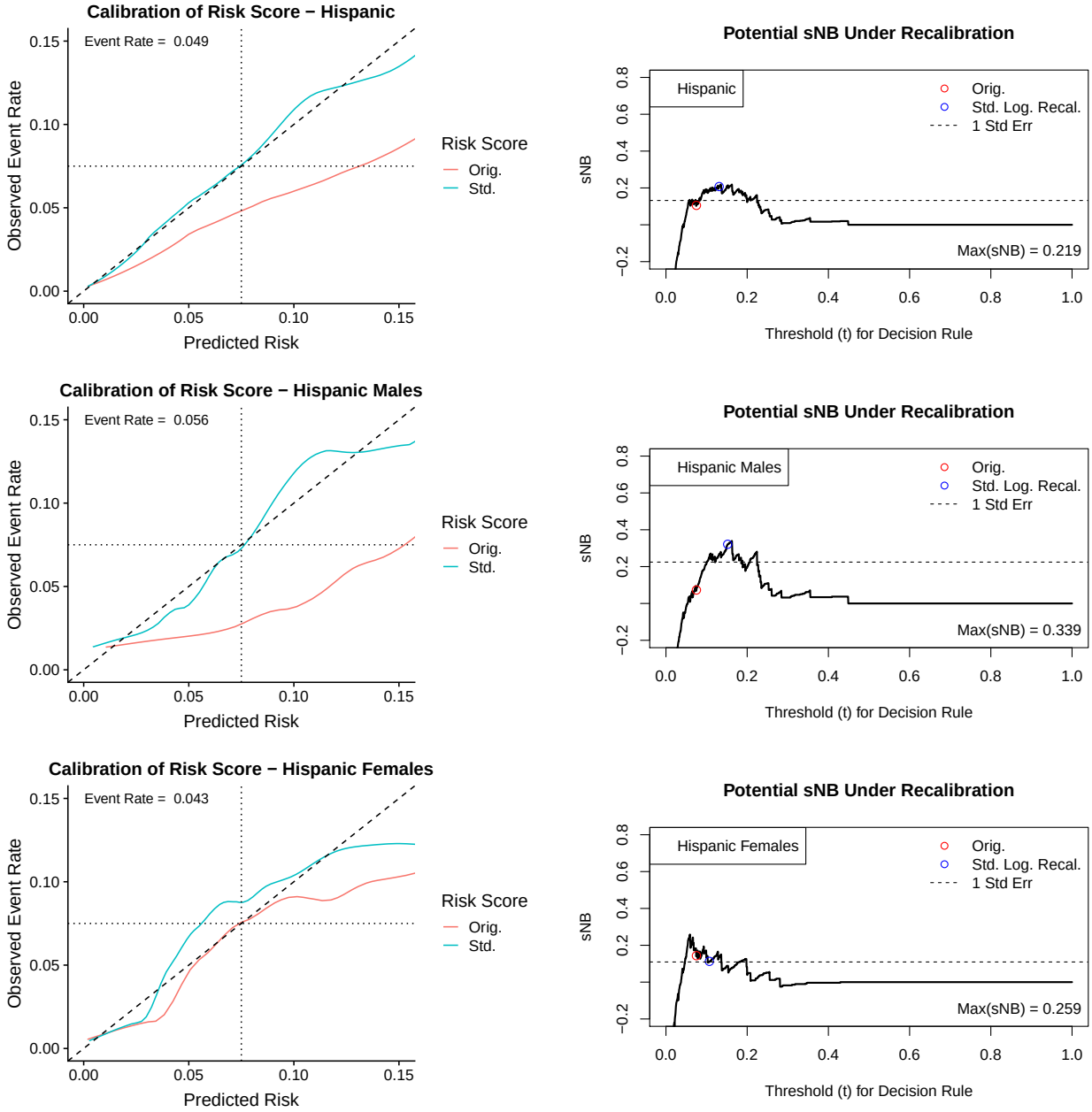


Figure B.38: Potential sNB curves and calibration curves by gender for ACC-AHA-ASCVD risk score in **Hispanic** MESA analysis sample.

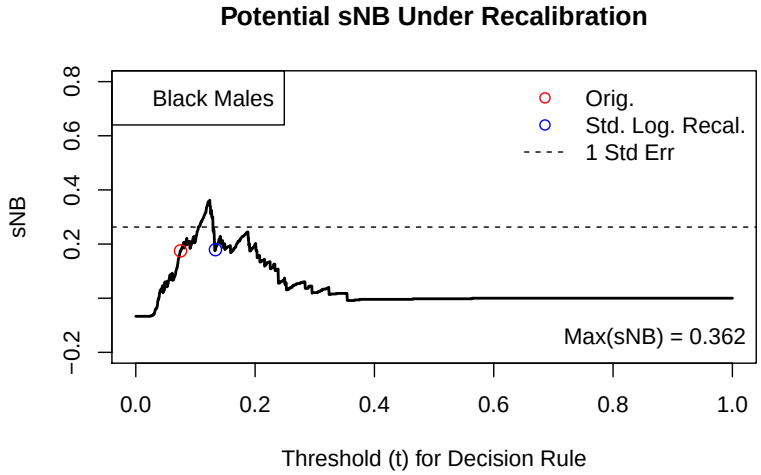


Figure B.39: Plot of sNB for different risk thresholds illustrating potential gains in sNB under recalibration (right) for black, male, MESA cohort. The dotted lines on the right panel show pointwise 95% confidence interval. The red dot shows the estimated sNB for the ACC-AHA-ASCVD risk score. The blue dot shows the estimated sNB for the recalibrated risk score, using standard logistic recalibration.

B.3.2 Results for application of methods to Black, male MESA cohort

In this section, we apply standard, weighted, and constrained logistic recalibration methods to recalibration the ACC-AHA-ASCVD risk score in the Black, male MESA cohort. Among the 538 black men in the analysis sample, the event rate of cardiovascular events is 0.071. The average 10-year risk of cardiovascular disease as estimated by the ACC-AHA-ASCVD risk score 12.49%. Figure B.39 shows the potential sNB that could be achieved under recalibration, which indicates specialized methods of recalibration, such as weighted or constrained logistic recalibration, could offer improvement in sNB and calibration at the risk threshold beyond what is achieved by standard logistic recalibration.

Table B.5 shows the estimated recalibration parameters $\hat{\alpha}$, standardized net benefit (and its components), event rate in the risk interval, and proportion treated. Under weighted

logistic recalibration the estimated sNB of the risk score 12.3% higher than under standard logistic recalibration, after optimization correction (on absolute scale). The estimated maximum achievable sNB under recalibration for this sample was 0.362, with estimated standard error 0.102. Therefore, the lower bound used to to define the constrained parameter space was $\widehat{sNB}(\vec{\alpha}) = 0.260$. The estimated sNB under constrained logistic recalibration is 0.274. This is a 9.5% increase in $\widehat{sNB}$ compared to standard logistic recalibration (on absolute scale).

The green curve in Figure B.40 shows the predicted risks obtained from standard logistic recalibration. The blue curve shows the calibration curve for the risk score obtained from weighted logistic recalibration. Figure B.40 suggests that standard logistic recalibration results in risks at the risk threshold being underestimated, overcorrecting for the overestimation under the original risk score. Weighted logistic recalibration produces a risk score that shows better calibration at and near the risk threshold. Figure B.41 shows the calibration curves for the standard logistic recalibrated risk score and the constrained logistic recalibrated risk scores. Under the constrained approach, the calibration curve suggests a small improvement in recalibration at the risk threshold. Given the low number of events in the sample ($n_D = 38$), there may be few events around the risk threshold, making calibration curves difficult to interpret. However, as shown in Table B.5 the event rate near the risk threshold is slightly higher under the constrained approach compared to standard logistic recalibration.

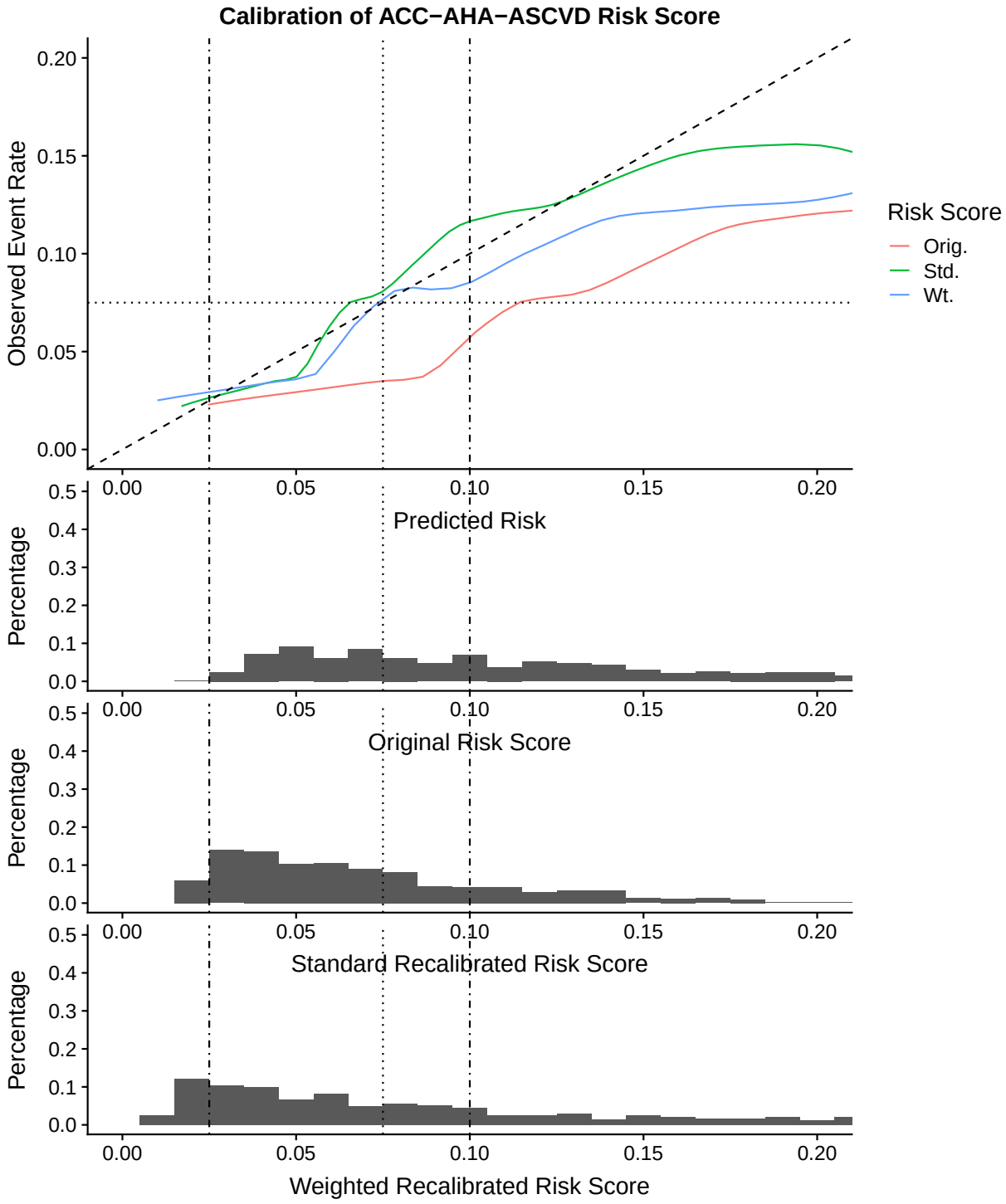


Figure B.40: Calibration curves of original, standard recalibrated, and weighted recalibrated ACC-AHA-ASCVD risk score for black males in MESA cohort. The dotted line indicates the risk threshold, $R = 0.075$. The dashed-dotted line indicates the relevant risk interval, $[0.025, 0.10]$. Compared to standard logistic recalibration, weighted logistic recalibration has better calibration in the key risk interval.

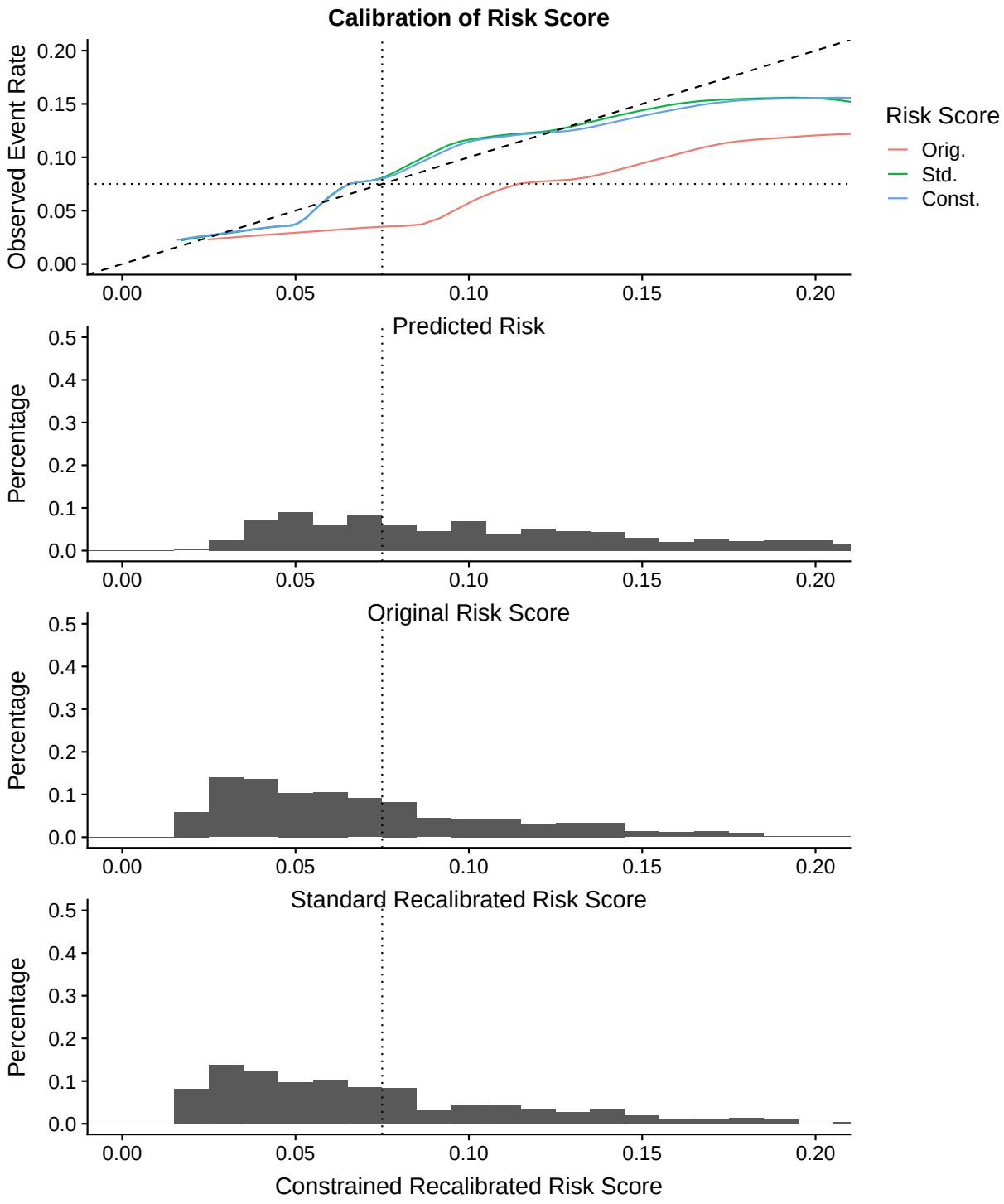


Figure B.41: Calibration curves comparing standard logistic recalibration and constrained logistic recalibration in black male participants in MESA study. The dotted line indicates the risk threshold of 0.075.

Table B.5: Comparison of recalibration methods in MESA black, male cohort for RAW = 0.01. An indicator weight is used for the weighted recalibration approach.

Measure	Orig	Std. Recal.	Wt. Recal.	Const. Recal
$(\hat{\alpha}_0, \hat{\alpha}_1)$	-	(-0.911, 0.856)	(0.192, 1.316)	(-0.836, 0.879)
Effective Sample Proportion %	-	100	70	100
$\widehat{sNB}$ (95% CI)[1]	0.175 (-0.082, 0.432)	0.179 (-0.069, 0.429)	0.304 (0.076, 0.599)	0.274 (0.130, 0.417)
$\widehat{sNB}$ (95% CI)[2]	-	0.034 (-0.111, 0.178)	0.147 (-0.070, 0.452)	0.130 (-0.013, 0.274)
$\widehat{TPR}^2$	0.868	0.423	0.663	0.632
$\widehat{FPR}^2$	0.650	0.414	0.413	0.399

[1]Delta-method derived std error used for original risk score 95% CI calculation.
Bootstrap used with 500 replications used for standard and weighted recalibration
Estimated corrected for optimism bias using bootstrap method with 500 replications.

Appendix C

APPENDIX C: SUPPLEMENTARY MATERIAL FOR CHAPTER 5

C.1 Additional Simulation Results

C.1.1 Standard bivariate normal setting

For this simulation setting we consider two risk markers that come from the following disease-specific distributions. Outcome data are drawn from $Y \sim Bern(0.10)$.

$$\mathbf{X}|Y=1 \sim N\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right) \quad \mathbf{X}|Y=0 \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right).$$

Table C.1: Mean and (standard deviation) of sNB, TPR, and FPR, for standard bivarate normal data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 1$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB}).

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$s\hat{N}B$		
5000	0.517 (0.01)	0.518 (0.01)	0.519 (0.01)	0.518 (0.01)	0.519 (0.01)
1000	0.509 (0.03)	0.505 (0.03)	0.517 (0.03)	0.504 (0.03)	0.517 (0.03)
800	0.505 (0.04)	0.499 (0.03)	0.516 (0.03)	0.498 (0.03)	0.515 (0.03)
500	0.502 (0.04)	0.488 (0.05)	0.515 (0.04)	0.487 (0.04)	0.514 (0.04)
			$T\hat{P}R$		
5000	0.753 (0.03)	0.740 (0.023)	0.759 (0.014)	0.740 (0.024)	0.759 (0.014)
1000	0.739 (0.07)	0.680 (0.046)	0.755 (0.029)	0.679 (0.048)	0.755 (0.030)
800	0.732 (0.07)	0.665 (0.049)	0.752 (0.031)	0.664 (0.051)	0.751 (0.031)
500	0.730 (0.08)	0.637 (0.067)	0.754 (0.041)	0.635 (0.070)	0.752 (0.043)
			$F\hat{P}R$		
5000	0.236 (0.032)	0.222 (0.021)	0.239 (0.013)	0.222 (0.021	0.239 (0.013)
1000	0.231 (0.062)	0.176 (0.033)	0.238 (0.028)	0.176 (0.034	0.238 (0.029)
800	0.227 (0.065)	0.166 (0.033)	0.236 (0.030)	0.166 (0.034	0.235 (0.030)
500	0.229 (0.072)	0.151 (0.041)	0.240 (0.040)	0.151 (0.043	0.239 (0.041)

Table C.2: Mean and (standard deviation) of sNB, TPR, and FPR, for standard bivarate normal data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 2$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB})

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$s\hat{N}B$		
5000	0.349 (0.014)	0.352 (0.013)	0.354 (0.013)	0.352 (0.013)	0.354 (0.013)
1000	0.341 (0.029)	0.346 (0.027)	0.352 (0.026)	0.346 (0.027)	0.352 (0.026)
800	0.337 (0.033)	0.341 (0.031)	0.350 (0.029)	0.341 (0.031)	0.350 (0.029)
500	0.328 (0.046)	0.332 (0.043)	0.347 (0.037)	0.331 (0.042)	0.347 (0.037)
			$T\hat{P}R$		
5000	0.573 (0.049)	0.576 (0.036)	0.585 (0.016)	0.575 (0.036)	0.585 (0.016)
1000	0.556 (0.078)	0.537 (0.052)	0.582 (0.034)	0.536 (0.052)	0.582 (0.034)
800	0.554 (0.083)	0.519 (0.057)	0.579 (0.037)	0.519 (0.058)	0.578 (0.037)
500	0.538 (0.108)	0.489(0.073)	0.576 (0.050)	0.488 (0.073)	0.575 (0.050)
			$F\hat{P}R$		
5000	0.111 (0.023)	0.112 (0.017)	0.115 (0.008)	0.112 (0.017)	0.115 (0.008)
1000	0.108 (0.037)	0.096 (0.021)	0.116 (0.016)	0.096 (0.022)	0.116 (0.017)
800	0.108 (0.038)	0.089 (0.022)	0.115 (0.018)	0.089 (0.022)	0.114 (0.018)
500	0.106 (0.048)	0.079 (0.025)	0.115 (0.023)	0.079 (0.025)	0.114 (0.024)

Table C.3: Mean and (standard deviation) of sNB, TPR, and FPR, for standard bivarate normal data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 0.5$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB})

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$\hat{sNB}$		
5000	0.673 (0.011)	0.674 (0.011)	0.676 (0.011)	0.674 (0.011)	0.676 (0.011)
1000	0.665 (0.026)	0.660 (0.030)	0.674 (0.023)	0.660 (0.031)	0.674 (0.023)
800	0.664 (0.028)	0.656 (0.032)	0.672 (0.025)	0.656 (0.032)	0.672 (0.025)
500	0.657 (0.037)	0.649 (0.047)	0.669 (0.033)	0.648 (0.048)	0.668 (0.033)
			$\hat{TPR}$		
5000	0.882 (0.028)	0.858 (0.018)	0.884 (0.011)	0.858 (0.018)	0.884 (0.011)
1000	0.877 (0.052)	0.817 (0.049)	0.880 (0.025)	0.817 (0.050)	0.879 (0.026)
800	0.879 (0.051)	0.812 (0.053)	0.877 (0.026)	0.811 (0.054)	0.876 (0.026)
500	0.872 (0.071)	0.807 (0.082)	0.875 (0.036)	0.806 (0.084)	0.873 (0.038)
			$\hat{FPR}$		
5000	0.416 (0.057)	0.368 (0.029)	0.414 (0.021)	0.368 (0.029)	0.414 (0.022)
1000	0.423 (0.104)	0.316 (0.067)	0.411 (0.048)	0.315 (0.068)	0.410 (0.050)
800	0.426 (0.101)	0.310 (0.073)	0.406 (0.049)	0.310 (0.074)	0.404 (0.050)
500	0.428 (0.133)	0.319 (0.114)	0.408 (0.068)	0.320 (0.117)	0.406 (0.071)

C.1.2 Standard bivariate normal setting

For this simulation setting we consider two risk markers that come from the following non-proportional disease-specific distributions. Outcome data are drawn from $Y \sim Bern(0.10)$.

$$\mathbf{X}|Y=1 \sim N\left(\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}9 & 12.6\\12.6 & 36\end{bmatrix}\right) \quad \mathbf{X}|Y=0 \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}7.29 & 3.78\\3.78 & 4\end{bmatrix}\right).$$

Table C.4: Mean and (standard deviation) of sNB, TPR, and FPR, for non-proportional covariance bivariate normal data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 2$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB})

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$\hat{sNB}$		
5000	0.323 (0.008)	0.226 (0.026)	0.223 (0.024)	0.320 (0.007)	0.304 (0.008)
1000	0.312 (0.030)	0.214 (0.064)	0.211 (0.060)	0.312 (0.022)	0.298 (0.022)
800	0.307 (0.033)	0.216 (0.067)	0.213 (0.063)	0.308 (0.034)	0.295 (0.034)
500	0.291 (0.050)	0.198 (0.084)	0.199 (0.079)	0.297 (0.046)	0.287 (0.047)
			$\hat{TPR}$		
5000	0.373 (0.019)	0.327 (0.031)	0.368 (0.026)	0.374 (0.017)	0.425 (0.009)
1000	0.362 (0.044)	0.310 (0.063)	0.359 (0.064)	0.361 (0.029)	0.422 (0.023)
800	0.358 (0.049)	0.306 (0.068)	0.359 (0.068)	0.355 (0.039)	0.419 (0.033)
500	0.343 (0.065)	0.283 (0.087)	0.346 (0.085)	0.338 (0.050)	0.411 (0.047)
			$\hat{FPR}$		
5000	0.025 (0.010)	0.050 (0.015)	0.073 (0.013)	0.027 (0.008)	0.060 (0.008)
1000	0.026 (0.016)	0.049 (0.025)	0.075 (0.028)	0.025 (0.011)	0.063 (0.017)
800	0.026 (0.018)	0.046 (0.025)	0.075 (0.031)	0.024 (0.011)	0.063 (0.019)
500	0.026 (0.022)	0.043 (0.033)	0.075 (0.036)	0.021 (0.013)	0.063 (0.023)

Table C.5: Mean and (standard deviation) of sNB, TPR, and FPR, for non-proportional covariance bivariate normal data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 0.5$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB})

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$\hat{sNB}$		
5000	0.494 (0.038)	0.471 (0.040)	0.442 (0.029)	0.423 (0.021)	0.383 (0.012)
1000	0.489 (0.063)	0.485 (0.060)	0.448 (0.060)	0.467 (0.069)	0.386 (0.030)
800	0.478 (0.067)	0.483 (0.062)	0.441 (0.063)	0.469 (0.071)	0.384 (0.036)
500	0.475 (0.080)	0.481 (0.076)	0.446 (0.077)	0.474 (0.085)	0.379 (0.054)
			$\hat{TPR}$		
5000	0.916 (0.136)	0.830 (0.172)	0.808 (0.034)	0.530 (0.097)	0.635 (0.019)
1000	0.816 (0.181)	0.853 (0.193)	0.806 (0.070)	0.731 (0.261)	0.637 (0.045)
800	0.788 (0.186)	0.865 (0.190)	0.797 (0.071)	0.768 (0.264)	0.632 (0.046)
500	0.756 (0.205)	0.842 (0.209)	0.800 (0.089)	0.797 (0.268)	0.629 (0.065)
			$\hat{FPR}$		
5000	0.842 (0.233)	0.722 (0.284)	0.729 (0.041)	0.214 (0.163)	0.504 (0.044)
1000	0.664 (0.289)	0.747 (0.320)	0.716 (0.091)	0.554 (0.434)	0.503 (0.096)
800	0.626 (0.292)	0.770 (0.315)	0.708 (0.097)	0.617 (0.434)	0.494 (0.096)
500	0.574 (0.310)	0.735 (0.339)	0.702 (0.128)	0.674 (0.429)	0.499 (0.130)

C.1.3 Outlier Simulation Scenario

Next, we implement a simulation setting used in Meisner et al. (2017) and Fong et al. (2016), where outliers are present. Performance of standard logistic regression may suffer due to the presence of outliers, however robust logistic regression should be less sensitive to their impact. To induce a moderate case of outliers we first simulate a binary indicator $\Delta \sim Bern(0.1)$. Then we simulate risk marker $\mathbf{X}$

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = (1-\Delta) \times Z_1 + \Delta \times Z_2,$$

where

$$\mathbf{Z}_1 \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.2 & 0.18 \\ 0.18 & 0.2 \end{bmatrix} \right) \qquad \mathbf{Z}_2 \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \right)$$

Outcomes are drawn from $Y \sim Bern(p)$, where p are true risks generated from a logistic-linear model $p = \text{expit}(-1 + 4X_1 - 3X_2 - 0.8(X_1 - X_2)^3)$. Based on this mean model, the prevalence of disease is approximately 30%. Figure 5.1 shows the distribution of the two biomarkers by disease status. Note the slightly bimodal nature of the distribution of markers among cases.

Table C.6: MeanMean and (standard deviation) of sNB, TPR, and FPR, for outlier data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 2$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB})

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$s\hat{N}B$		
5000	0.014 (0.009)	0.014 (0.004)	0.014 (0.003)	0.014 (0.004)	0.014 (0.003)
1000	0.014 (0.017)	0.012 (0.008)	0.014 (0.006)	0.013 (0.011)	0.015 (0.009)
800	0.016 (0.021)	0.012 (0.010)	0.013 (0.008)	0.014 (0.018)	0.016 (0.015)
500	0.015 (0.022)	0.011 (0.011)	0.013 (0.009)	0.015 (0.019)	0.016 (0.017)
			$T\hat{P}R$		
5000	0.080 (0.051)	0.072 (0.034)	0.058 (0.015)	0.078 (0.036)	0.064 (0.017)
1000	0.089 (0.076)	0.085 (0.058)	0.068 (0.035)	0.096 (0.066)	0.076 (0.039)
800	0.096 (0.088)	0.082 (0.063)	0.066 (0.038)	0.095 (0.075)	0.078 (0.053)
500	0.103 (0.096)	0.089 (0.076)	0.070 (0.048)	0.103 (0.089)	0.084 (0.064)
			$F\hat{P}R$		
5000	0.034 (0.024)	0.030 (0.017)	0.023 (0.007)	0.033 (0.018)	0.026 (0.008)
1000	0.039 (0.036)	0.038 (0.031)	0.028 (0.017)	0.043 (0.035)	0.032 (0.019)
800	0.041 (0.040)	0.036 (0.032)	0.027 (0.018)	0.042 (0.036)	0.032 (0.023)
500	0.045 (0.046)	0.041 (0.040)	0.030 (0.024)	0.046 (0.044)	0.035 (0.029)

Table C.7: Mean and (standard deviation) of sNB, TPR, and FPR, for outlier simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 0.5$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB}).

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$s\hat{N}B$		
5000	0.507 (0.015)	0.507 (0.015)	0.506 (0.014)	0.506 (0.015)	0.505 (0.014)
1000	0.506 (0.033)	0.506 (0.034)	0.505 (0.033)	0.505 (0.033)	0.504 (0.032)
800	0.501 (0.038)	0.502 (0.038)	0.502 (0.037)	0.502 (0.038)	0.501 (0.036)
500	0.502 (0.045)	0.503 (0.047)	0.502 (0.047)	0.504 (0.047)	0.502 (0.046)
			$T\hat{P}R$		
5000	0.986 (0.020)	0.979 (0.021)	0.975 (0.007)	0.976 (0.023)	0.972 (0.008)
1000	0.964 (0.043)	0.968 (0.040)	0.969 (0.018)	0.962 (0.044)	0.964 (0.021)
800	0.956 (0.050)	0.960 (0.045)	0.968 (0.021)	0.952 (0.051)	0.962 (0.026)
500	0.949 (0.053)	0.961 (0.047)	0.965 (0.028)	0.955 (0.053)	0.958 (0.034)
			$F\hat{P}R$		
5000	0.976 (0.035)	0.964 (0.037)	0.958 (0.013)	0.959 (0.041)	0.953 (0.015)
1000	0.938 (0.081)	0.945 (0.069)	0.948 (0.034)	0.934 (0.077)	0.941 (0.038)
800	0.924 (0.092)	0.931 (0.077)	0.947 (0.038)	0.915 (0.094)	0.935 (0.052)
500	0.910 (0.104)	0.933 (0.082)	0.943 (0.051)	0.918 (0.100)	0.928 (0.067)

C.1.4 Skewed Data Scenario

Markers $\mathbf{X} = (X_1, X_2)$ are correlated and are informative of disease status.

$$\mathbf{X}|Y = 1 \sim LogNorm\left(\begin{bmatrix}1.1\\1.1\end{bmatrix}, \begin{bmatrix}0.04 & 0.09\\0.09 & 0.5\end{bmatrix}\right) \quad \mathbf{X}|Y = 0 \sim LogNorm\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}0.05 & 0.015\\0.015 & 0.05\end{bmatrix}\right).$$

The remaining risk marker, X_3 is uncorrelated to the previous two, and has the same distribution among cases and controls,

$$X_3|Y = 1 \equiv X_3|Y = 0 \sim LogNorm(1.65, 4.66)$$

Outcome data are simulated from $Y \sim Bern(0.5)$.

Table C.8: Mean and (standard deviation) of sNB, TPR, and FPR, for skewed data simulation setting with $\omega = \frac{C}{B}\frac{\hat{P}(D=0)}{\hat{P}(D=1)} \approx 0.5$. Results for direct maximization are compared to standard (Std.) and robust (Rob.) logistic regression. Measures based on 500 Monte Carlo samples. Logistic methods use a decision threshold based on optimizing sNB given the linear combination estimated from logistic regression (t_{opt}) or the harm-benefit ratio (t_{CB}).

	Direct Optimization	Std. Logistic Regression		Rob. Logistic Regression	
n		t_{opt}	t_{CB}	t_{opt}	t_{CB}
			$s\hat{N}B$		
5000	0.589 (0.010)	0.564 (0.014)	0.547 (0.012)	0.571 (0.014)	0.554 (0.012)
1000	0.585 (0.024)	0.562 (0.032)	0.546 (0.028)	0.568 (0.032)	0.552 (0.030)
800	0.583 (0.028)	0.560 (0.036)	0.545 (0.033)	0.566 (0.036)	0.551 (0.034)
500	0.579 (0.044)	0.558 (0.041)	0.545 (0.039)	0.564 (0.042)	0.551 (0.041)
			$T\hat{P}R$		
5000	0.941 (0.014)	0.940 (0.015)	0.986 (0.007)	0.941 (0.014)	0.986 (0.006)
1000	0.937 (0.024)	0.940 (0.024)	0.981 (0.016)	0.941 (0.022)	0.981 (0.015)
800	0.933 (0.027)	0.939 (0.026)	0.978 (0.020)	0.939 (0.024)	0.978 (0.018)
500	0.930 (0.052)	0.939 (0.029)	0.974 (0.023)	0.939 (0.028)	0.975 (0.021)
			$F\hat{P}R$		
5000	0.705 (0.026)	0.753 (0.032)	0.878 (0.026)	0.741 (0.028)	0.865 (0.024)
1000	0.705 (0.044)	0.758 (0.057)	0.874 (0.055)	0.747 (0.049)	0.861 (0.052)
800	0.700 (0.048)	0.758 (0.060)	0.867 (0.061)	0.746 (0.054)	0.856 (0.057)
500	0.705 (0.066)	0.766 (0.076)	0.863 (0.075)	0.754 (0.069)	0.852 (0.072)