Targeted Recalibration Using Single-knot Splines to Improve the Clinical Utility of a Risk Model

Renee Russell

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

Master of Science

University of Washington

2020

Committee:

Kathleen Kerr

Noah Simon

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Targeted Recalibration Using Single-knot Splines
to Improve the Clinical Utility of a Risk Model

Renee Russell

Chair of the Supervisory Committee:
Kathleen Kerr
Department of Biostatistics

Risk prediction models are sometimes used to guide clinical decisions. Specifically, a patient may decide for or against an intervention based on whether his/her estimated risk falls above or below a critical threshold. When a risk model is used to make decisions in this way, it is especially important that the predicted risks are well calibrated. Miscalibrated risks are misleading and, moreover, reduce the clinical utility of risk-based decision-making to the patient population. Risk model miscalibration can occur due to overfitting or when a risk model is developed in one population and applied to another population. Recalibration methods can be used to address risk model miscalibration, but most methods do not account for how the risk model will be applied. We propose a new method of recalibration that is designed for risk models that will be used for risk-based decision-making. The method uses splines with a single, strategically placed knot to add flexibility and target good calibration where it is most important: at the critical threshold used for decision-making. We present simulation studies that compare our proposed recalibration method to existing methods and demonstrate our method's ability to improve the clinical utility of risk models in a variety of settings.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Risk prediction models that accurately estimate the probability of a binary health outcome or event are valuable in medical settings and sometimes guide clinical decisions. When deciding whether or not a specific intervention (e.g., a medical procedure, treatment, or screening) is advisable, the decision may be informed by the patient's predicted risk of a negative event absent the intervention. A clinician or patient may opt for the intervention when a risk model assigns a high predicted risk of the negative outcome. Conversely, a patient may forego a certain medical procedure or intervention when the predicted risk is low.

The magnitude of risk that merits intervention depends on the specific clinical context and the consequences of the intervention itself. Based on such knowledge, experts may establish a risk threshold across which different decisions are made. Specifically, patients with risks above the risk threshold are recommended to receive the intervention, while patients whose risks fall below the threshold are advised to forego it. A classic result from decision theory is that the risk threshold is directly connected to the relative misclassification costs [Pauker and Kassirer, 1975, 1980]. A risk threshold of 50%, for example, implies the consequences of either type of misclassification error - initiating an intervention when it is unnecessary to the patient or foregoing an intervention when it could benefit the patient - are equal. It would be unusual in a clinical context for the two types of errors to have the same severity.

As an example, the American College of Cardiology (ACC) defined a threshold of 7.5% for recommending statin therapy when considering 10-year atherosclerotic cardiovascular disease (ASCVD) risk [Goff et al., 2014]. The 7.5% risk threshold implies the benefit from giving statins to someone who would develop ASCVD is much greater

(specifically, $\frac{100-7.5}{7.5} = 12.3$ times greater) than the harms or burdens of giving statins to someone who will not have ASCVD. We provide further details of this relationship and calculation in Section 2.4.

When risk predictions are combined with a risk threshold to make clinical decisions, it is especially critical that the risks are well calibrated. Calibration means the predicted risks align with event rates, i.e., among the individuals with a predicted risk of $X\%$, $X\%$ will have the clinical event without intervention. Miscalibrated risk scores mislead clinicians and patients and lessen the value of risk-based decision-making [Baker et al., 2012; Mishra, 2019; Van Calster and Vickers, 2015; Van Calster et al., 2019]. For example, suppose a risk model for 10-year ASCVD risk systematically overestimates risks such that among all those assigned a risk score of 10%, only 6% would have an ASCVD event in 10 years even without any statin therapy. Presenting the score of 10% misleads these individuals about their risk, perhaps producing undue worry and resulting in excessive precautions. If following the ACC guidelines, these individuals would all receive statin therapy. However, assuming the 7.5% threshold accurately represents the benefits and burdens of statin therapy, treating these individuals is inappropriate. Making risk-based treatment decisions using the overestimated risks lessens the value of the risk-based treatment policy to the patient population.

One way miscalibration can occur is when a risk model is developed in one population and then applied to another population, such as an ethnic group not represented in the original population. In principle it may be ideal to develop a risk model especially for every population, but this is not always feasible and may be undesirable if the original risk model is well-established. Another option is risk model recalibration. Recalibration methods adjust the original (miscalibrated) risk scores to better reflect the new population. Cox's logistic recalibration [Cox, 1958] is the most common recalibration method. However, logistic recalibration and most other recalibration methods are agnostic about the way predicted risks will be used. When predicted risks will be compared to a fixed risk threshold to make clinical decisions, we care most about calibration at the

risk threshold. As far as we know, the only recalibration methods tailored specifically to this setting are weighted recalibration and constrained recalibration [Mishra, 2019]. In this work we provide another approach to risk model recalibration that incorporates knowledge of the risk threshold to target good calibration in this most critical area.

In Chapter 2, we give an overview of calibration and existing recalibration methods. Chapter 3 describes a new recalibration method for settings with a pre-established risk threshold. Chapter 4 describes simulations for examining performance of this method compared to other methods, the results of which are provided in Chapter 5. Finally, we provide discussion in Chapter 6.

Chapter 2

BACKGROUND

2.1 Notation and set-up

Let RS denote a risk model for a binary outcome Y (e.g., death or ASCVD) based on some predictor(s) X. Then $RS = \hat{P}(Y = 1|X)$ and we denote an individual's risk score as $RS_i = \hat{P}(Y_i = 1|X_i)$ for patient i. We use the terms "risk score" and "predicted risk" interchangeably when referring to these predictions or estimated risks.

We refer to individuals who experience the outcome or event of interest ($Y = 1$) as "cases" and those who do not ($Y = 0$) as "controls." It is also helpful to define $Z = \text{logit}(RS)$ here, since recalibration methods often operate using the logit-transformed risk scores rather than the risk scores on the probability scale.

For this paper, we assume there is some pre-existing risk model RS we wish to recalibrate. We also assume there is a dataset on a sample of patients who did not receive the intervention. The data contain patients' risk scores and whether or not they had the clinical event. Since our focus is recalibration, we are not concerned with the details underlying risk model RS or its mechanism of estimating risks. However, we do assume RS is "monotone" such that higher risk scores imply higher probability of the outcome: $RS_i > RS_j \implies P[Y_i = 1|RS_i] \geq P[Y_j = 1|RS_j]$.

2.2 Calibration

There are different notions of risk model calibration. In this paper, we say a risk model is calibrated if it has "calibration in the moderate sense" as defined by Van Calster et al. (2016). Calibration in the moderate sense means predicted risks agree with population event rates. Mathematically, if risk model RS is calibrated at risk r, then $P(Y = 1|RS =$

$r) = r$. Calibration at $r = 15\%$, for example, means among the subset of the population assigned a risk score of 15%, the event should occur for 15% of them. If a risk model is calibrated at all r, then we simply say that RS is calibrated.

If a clinical decision depends on patients' predicted risks, it is desirable that these predictions are calibrated, i.e., reliable and reflecting the observed rates of the outcome. Miscalibration might manifest as systematically overestimated or underestimated risks. One cause of miscalibration is heterogeneity between the setting used to develop a risk model and the setting where it will be applied. Differences in prevalence, demographics, changes over time, etc. might all lead to risk model miscalibration. For example, the ACC ASCVD risk model showed a strong tendency overestimate risks when applied to a Dutch (Rotterdam Study) cohort [Kavousi et al., 2014] and the MESA (Multi-Ethnic Study of Atherosclerosis) cohort [DeFilippis et al., 2015]. Alternatively, a risk model that is overfit may be miscalibrated in a different way, giving predicted risks that are too low for patients at low risk and too high for patients at high risk [Harrell, 2015; Pepe et al., 2015].

We can assess a risk model's degree and nature of miscalibration with a calibration curve that plots the observed event rate (vertical axis) across the predicted risk scores (horizontal axis). The observed binary events are 0/1-valued, but estimated event rates between 0 and 1 can be obtained by smoothing or averaging the outcomes by groups with similar predicted risks. A loess (locally weighted least squares regression) smoother is often chosen to obtain a calibration curve [Austin and Steyerberg, 2014; Harrell, 2015], and we use loess smoothers to make calibration curves for examples in this paper. We can compare a risk model's calibration curve to the diagonal line with slope 1 to assess miscalibration. Coincidence with the diagonal indicates calibration while deviations from the diagonal indicate miscalibration.

When poor calibration is observed or suspected for a risk model in a certain population, one approach is to develop or refit a new risk model. However, this may not be feasible without sufficient data and resources. It is possible to instead retain the original

risk model - which may already be well-established and familiar to clinical users - but use a method to correct or adjust the predicted risks to be better calibrated. This approach, called "recalibration," is described below.

2.3 Recalibration methods

A recalibration RS' of a risk model RS transforms the risk scores by a monotone non-decreasing function $f : [0, 1] \to [0, 1]$: $RS' = f(RS)$. Requiring this monotonicity of f guarantees the recalibrated risk scores maintain the same rank ordering as the original risk scores. We would consider a non-monotone transformation that re-orders predicted risks to result in a fundamentally different risk model, not a "recalibration" of RS.

The most prominent method of recalibration is Cox's logistic recalibration [Cox, 1958], where the mapping from original to recalibrated risk scores follows

$$f(Z) = \text{expit}(\alpha_0 + \alpha_1 Z).$$

We will refer to this method as "logistic recalibration" or sometimes "standard logistic recalibration" to differentiate it from extensions of this method. As the name suggests, the α coefficients are estimated from logistic regression of the outcome Y on the logit-transformed risk scores $Z = \text{logit}(RS)$. Thus, we obtain newly recalibrated scores by scaling and shifting the original (logit-transformed) scores Z by $\hat{\alpha}_1$ and $\hat{\alpha}_0$, respectively, then acquiring the recalibrated risk scores via the expit transformation (inverse of the logit). This mapping will be monotone non-decreasing - thus, a valid recalibration - when $\alpha_1 \geq 0$.

Another recalibration method is Beta recalibration, proposed by Kull et al. (2017). Whereas logistic recalibration has one location and one scale parameter, Beta recalibration has one location and two scale parameters. This parameterization allows for more flexibility beyond the familiar sigmoid shape of the logistic recalibration map; the Beta recalibration map produces both symmetric and assymmetric shapes. In particular, the Beta mapping from original to recalibrated risk scores follows

$$f(W_1, W_2) = \text{expit}(\alpha_0 + \alpha_1 W_1 + \alpha_2 W_2)$$

where $W_1 = \log(RS)$ and $W_2 = -\log(1-RS)$. The coefficients are estimated from logistic regression of the outcome Y on W_1 and W_2. The mapping is monotone non-decreasing when both $\alpha_1, \alpha_2 \geq 0$. Note that when $\alpha_1 = \alpha_2$, the Beta recalibration method reduces to standard logistic recalibration.

More flexible recalibration techniques exist, including non-parametric methods. Too much flexibility, however, runs the risk of overfitting. Moreover, more flexible methods may violate the monotonicity requirement. We are interested in parsimonious methods where we can also ensure monotonicity of the transformation, such as the above logistic and Beta methods. However, neither logistic nor Beta recalibration account for the application of the risk model to clinical decision-making.

To our knowledge, the only recalibration methods tailored to this application are Mishra's "constrained logistic" recalibration and "weighted logistic" recalibration [Mishra, 2019], extensions of standard logistic recalibration. The constrained logistic method uses the logistic recalibration map but estimates the intercept and slope parameters over a constrained parameter space restricted to only those which produce "high" sNB, standardized Net Benefit (a measure of clinical utility: see Section 2.4 below). If standard logistic recalibration produces a recalibrated model with high sNB, the standard and constrained logistic solutions can coincide. The following section provides more detail regarding the clinical utility of risk models for deciding for or against an intervention and how this relates to calibration.

2.4 Clinical Utility of Risk Models

Measures of clinical utility quantify how useful a risk model is to the relevant population for making clinical decisions, taking into account the decisions' costs and benefits. The measure of clinical utility we use in this paper is standardized Net Benefit (sNB), which incorporates the costs and benefits associated with a clinical decision as a weighted

average of true and false positive rates [Vickers and Elkin, 2006; Pepe et al., 2015; Kerr et al., 2016; Vickers et al., 2016].

Suppose for outcome Y, we use risk model RS with risk threshold R to decide for or against a clinical action (an intervention such as further screening, treatment, etc.). Standardized Net Benefit is defined as

$$sNB = TPR_R - \frac{R}{1-R}\frac{1-\pi}{\pi}FPR_R \tag{2.1}$$

where π is the prevalence of Y in the relevant population and TPR_R and FPR_R are the true and false positive rates, respectively, associated with using the risk model RS with risk threshold R.

As mentioned earlier, if the risk threshold R has been rationally selected, there is a relationship between R and the cost and benefit of the intervention [Pauker and Kassirer, 1975, 1980; Vickers and Elkin, 2006]. In particular, for a cost C of incorrectly intervening on a control and benefit B of correctly intervening on a case, the relation is $\frac{R}{1-R} = \frac{C}{B}$. Thus, we can rewrite (1) with,

$$sNB = TPR_R - \frac{C}{B}\frac{1-\pi}{\pi}FPR_R \tag{2.2}$$

As an example, a risk threshold R of 10% implies $\frac{C}{B} = \frac{1}{9}$, i.e., the expected benefit to a case of receiving the intervention is 9 times the expected cost or burden to a control of receiving the intervention. Note the key quantity is the ratio $\frac{C}{B}$ rather than the absolute values of C and B individually.

Notice that the highest possible value of sNB is 1 and would be obtained for a risk model that perfectly discriminates between cases and controls. Therefore, sNB is often considered on a percent scale. For example, a risk model with $sNB = 0.5 = 50\%$ means

that the model attains half the clinical utility of a perfect risk model. For reference, a treat-none policy (foregoing intervention for everyone, uniformly) has sNB of 0.

It has been shown both empirically and theoretically that the clinical utility of a risk model is closely related to its calibration. Simulations by Van Calster and Vickers (2015) illustrate that miscalibration leads to lower clinical utility. Baker et al. (2012) established theoretical results underlying the relationship between clinical utility and calibration. Mishra (2019) expresses the theoretical connection in the following corollary:

Corollary 1 (sNB of risk-based treatment policies and calibration of RS at R): *Let RS be a monotone risk model for binary outcome Y. Suppose RS is used to select individuals for an intervention based on $RS > R$, where R is a pre-specified risk threshold that represents the benefits and harms of the intervention. Then RS has maximum sNB among all recalibrated versions of RS if and only if RS is calibrated at R.*

Corollary 1 implies that it is important to have good calibration of the risk model RS at the decision threshold R if the goal is to optimize clinical utility. Or conversely, a method of recalibration that optimizes sNB can be expected to yield a recalibrated risk model that is calibrated at R.

In the next section, we propose a new approach to risk model recalibration that focuses on good calibration at the clinically important risk threshold. The simple idea is to add more flexibility compared to standard logistic recalibration and to target this additional flexibility at the clinically important risk threshold R.

Chapter 3

TARGETED SPLINE RECALIBRATION METHOD

We present a new method for risk model recalibration extending Cox's logistic recalibration to incorporate knowledge of the clinically important risk threshold R. Specifically, we propose fitting a linear spline with a single knot, where the known risk threshold informs the knot's placement. The overarching goal of the method is to improve the sNB of the risk model. The method leverages the relationship between clinical utility and calibration at the risk threshold (Corollary 1) to accomplish this.

We begin with an overview of the relevant details regarding splines.

3.1 Splines

In general, a spline is a function formed by polynomial segments of degree D (order $D+1$) pieced together at 'knots' such that the function is continuous and, for higher order splines, differentiable. Our proposed method will employ linear splines (degree $D=1$, order 2) which are composed of connected linear segments. In our method, a single knot 'k' is specified so that the spline function is composed of two linear segments meeting at 'k.'

We use the spline basis known as the "truncated polynomial basis." For a linear spline with predictor x, outcome y, and a single knot $k \in (0,1)$, this basis consists of $\{s_0, s_1, s_2\}$ where:

$$s_0 = 1$$

$$s_1 = x$$

$$s_2 = (x-k) \times 1_{\{x \geq k\}} = (x-k)_+$$

Using this basis, the resulting spline has the form:

$$y = \beta_0 + \beta_1 x + \beta_2 s_2$$

Note that β_1 gives the slope over the interval (0, k) and $\beta_1 + \beta_2$ gives the slope over the interval (k, 1).

3.2 Recalibration with Monotone Single-knot Linear Splines

The proposed method extends standard logistic recalibration, incorporating a single knot for added flexibility. In Section 3.3, we describe our method for choosing the knot's location, but for now assume the knot's location is known.

For a (logit-)linear spline with a single knot at k, we have the mapping

$$f(Z) = f(s_1(Z), s_2(Z)) = \text{expit}(\beta_0 + \beta_1 s_1(Z) + \beta_2 s_2(Z))$$

from original to recalibrated risk scores, where $s_1(Z) = Z = \text{logit}(RS)$ and $s_2(Z) = (Z - k) \times 1_{\{Z \geq k\}} = (Z - k)_+$. This is similar to logistic recalibration, but additional flexibility is afforded by the knot. Logistic regression of outcome Y on $s_1 = Z$ and s_2 is used to estimate the recalibration parameters (β_0, β_1, and β_2). Thus, the mapping f specifies that we obtain recalibrated scores by scaling the original (logit-transformed) scores Z by $\hat{\beta}_1$ or $\hat{\beta}_1 + \hat{\beta}_2$ and shifting by $\hat{\beta}_0$, followed by an expit transformation. (The scaling factor $\hat{\beta}_1$ applies to $Z < k$, i.e., Z left of the knot and $\hat{\beta}_1 + \hat{\beta}_2$ applies to $Z \geq k$, i.e., Z right of the knot.) This recalibration mapping is monotone when both $\hat{\beta}_1$ and $\hat{\beta}_1 + \hat{\beta}_2$ are ≥ 0.

3.3 Knot Placement

Recall that Corollary 1 implies that we care most about good calibration at the risk threshold R when we care about the sNB of the risk model. By definition, RS is calibrated at R if $P(Y = 1|RS = R) = R$. Since our goal is to achieve the best possible calibration at R, our method aims to place a single knot based on the location R^* such

that $P(Y = 1|RS = R^*) = R$. In other words, to place the knot, we need to identify the risk score of the miscalibrated risk model RS corresponding to event rate R. By virtue of RS being miscalibrated, we do not expect to place the knot at R, i.e., we expect $R^* \neq R$.

Figure 3.1 illustrates why we choose to place the knot according to R^*. Risk score RS is miscalibrated; in particular, at the critical risk threshold of 30% the event rate is nearly 60%. The figure shows that the group with $RS = 12\%$ has event rate 30%. Therefore, our method places a knot corresponding to $R^* = 12\%$.

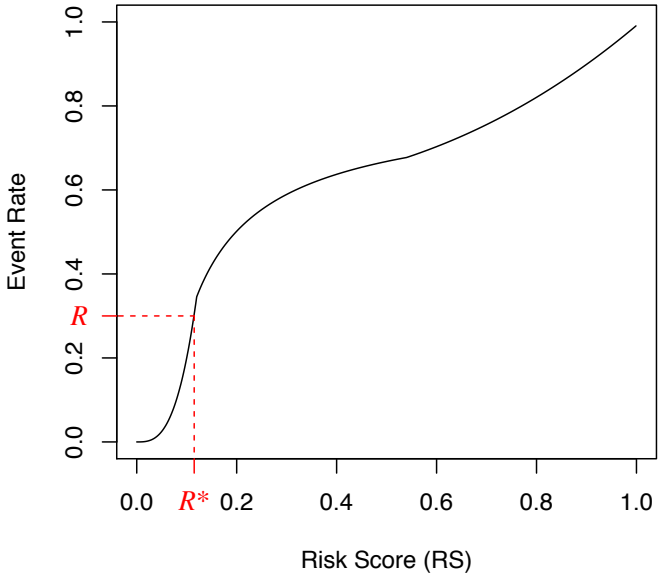


Figure 3.1: Calibration curve for risk score RS and choice of knot for targeted recalibration. The curve shows the rate of the clinical event by predicted risk RS. The group with risk score $RS = 12\%$ has event rate equal to the critical risk threshold $R = 30\%$. Targeted spline recalibration chooses a knot corresponding to $R^* = 12\%$.

We estimate R^* by considering a smoothing of the observed event rate in the dataset. This smoothing can be achieved via LOESS regression of the outcomes Y_i on the RS_i. In practice, we use the `loess()` function in **R** with the default smoothing span of 0.75 and polynomial degree option set to 1. We denote the smoothed event rate as $o(RS_i)$,

following the notation of Mishra (2019). Then, our spline recalibration method's choice of knot corresponds to the $\hat{R}^*$ that minimizes $|o(\hat{R}^*) - R|$. To compute $\hat{R}^*$, we first find the risk score RS_j in our data giving minimal $|o(RS_j) - R|$. We can then search a neighborhood of general values around this initial RS_j to potentially obtain a $\hat{R}^*$ that further minimizes $|o(\hat{R}^*) - R|$.

After estimating $\hat{R}^*$, we proceed to fit the linear spline as described in Section 3.2 with the knot $k = \text{logit}(\hat{R}^*)$, specifically. Since regression is on the logit-transformed risk scores, we similarly transform $\hat{R}^*$ to $\text{logit}(\hat{R}^*)$. Identifying $\hat{R}^*$ utilizes knowledge of the critical risk threshold R. By using this knowledge to place the knot, our single-knot spline recalibration affords extra flexibility of the recalibration function in the critical region.

Chapter 4

SIMULATION STUDIES

4.1 Comparator Recalibration Methods

We will present the performance of our targeted single-knot spline recalibration method compared with three existing parsimonious recalibration methods: standard logistic, constrained logistic, and Beta recalibration.

In addition, we compare our method to single-knot spline recalibrations with different criteria for placement of the knot. The purpose of showing these alternative single-knot recalibrations is to demonstrate that the advantages of our method do not simply result from using a more flexible recalibration function. In other words, we wish to demonstrate the importance of strategic knot placement that is informed by the goal of using the risk model for a clinical decision. The alternative knot placements are:

1. The median risk score.

2. The median risk score among cases only.

3. The risk score equal to R.

4. The risk score corresponding to event rate equal to the prevalence.

For (1), the median of all the original risks RS_i is taken as the knot location. Similarly, for (2), this is done only among cases (those with the event, $Y = 1$). For (3), we emphasize that this is where the original *miscalibrated* risk score equals R and not where the observed event rate is R (which is our proposed method). For (4), we estimate the prevalence of the outcome Y and use the same smoothing $o(RS_i)$ to estimate where the observed event rate equals the prevalence. Note that (3) is the only knot location

of the four that uses the clinically relevant risk threshold R, although it is not a very sensible use of R since the risk score is miscalibrated.

For each method, we find the maximum likelihood estimates of the recalibration parameters over a constrained parameter space that guarantees the resulting recalibration is monotone non-decreasing. For example, for the spline recalibrations, the parameter constraint space is β_1, $\beta_1 + \beta_2 \geq 0$. For estimation, we use the BFGS (Broyden, Fletcher, Goldfarb, Shanno) algorithm for optimization along with a logarithmic barrier that enforces the constraints by penalizing values near the boundaries. This is implemented with the `constrOptim()` function in **R**.

4.2 Simulation procedure

We simulate data across eight different settings exhibiting varying miscalibration behavior and risk distributions. The data simulation procedure is adopted from Mishra (2019). We first specify a mixture Beta distribution (3 sub-distributions) to generate true risks p_i. Then we generate outcomes Y_i following Bernoulli(p_i). We apply a monotone piecewise polynomial function $g : [0, 1] \to [0, 1]$ to the true risks p_i to obtain miscalibrated predicted risks. Different "miscalibration functions" g produce different types of miscalibration.

4.3 Simulation settings

We present eight examples. Four settings (Examples 1-4) are used in conjunction with critical risk threshold $R = 0.30$. Each exhibits a different type of miscalibration. The first setting (Example 1) is characterized by underestimated risk scores near the risk threshold and overestimated risk scores elsewhere. Conversely, the third setting (Example 3) is characterized by overestimated risks near the risk threshold and underestimated risks elsewhere. The second setting (Example 2) illustrates overall underestimation of risk and the fourth (Example 4) setting illustrates overall overestimation of risk.

The remaining four examples use critical risk threshold $R = 0.075$. Examples 5 and

6 use the same miscalibration function and are characterized by overestimation of most risks including slight overestimation near the risk threshold. Examples 5 and 6 differ in the underlying Beta mixture distributions of the p_i. Examples 7 and 8 also share a miscalibration function and are characterized by underestimation at very low risks and overestimation elsewhere, starting near the risk threshold. Examples 7 and 8 differ in the underlying Beta mixture distributions for the p_i.

The appendix provides more detail, including each setting's underlying Beta distribution (mixing and hyperparameters) and the functions g used to induce miscalibration of the risk scores.

4.4 Evaluation

For each recalibration method, recalibration parameters (including knots when applicable) are estimated from a simulated "training" dataset of size 500 or 1000. Performance is always evaluated on an independent "test" dataset of size 10 million simulated from the same data-generating mechanism. We use a large independent test set to essentially eliminate small sample variability from our assessment of the methods' performances, as we are interested in the performance of the recalibrated risk model in the population.

Each recalibration function estimated from a training dataset is applied to the test dataset's miscalibrated predicted risks to produce recalibrated risks and calculate a corresponding sNB. Note that we calculate sNB based on a risk threshold of $R = 0.30$ for Examples 1-4 and $R = 0.075$ for Examples 5-8. We also present illustrative calibration curves (based on a single simulation for each example) to graphically compare calibration of the original and recalibrated risk models.

Chapter 5

RESULTS

We compare the clinical utility (sNB) of recalibrated risk models produced by our targeted spline method and by other recalibration methods. For each simulation setting and recalibration method, we report mean (SD) of sNB over 1000 independent simulations. The sNB values are reported in percent units (%).

For the spline recalibration methods, certain knot placements can result in the situation where all individuals with risk scores below the knot are controls ($Y = 0$). That is, there may be no cases to the left of the knot. This is most likely when the knot is close to 0. In this situation, it does not make sense to fit a spline and estimate a slope over the pre-knot region. If this happened in practice, we anticipate that one would choose not to use a knot, which means reverting to standard logistic recalibration (estimating a single slope parameter). Thus, whenever this situation arose, we substituted standard logistic recalibration for the more flexible spline recalibration. When reporting results, we note the number of instances out of 1000 simulations that a method reverted to standard logistic recalibration.

5.1 Simulation Results

Tables 5.1 and 5.2 summarize the sNB of the recalibrated risk models for Examples 1 through 4 using training datasets with sample size N = 500 and 1000, respectively, and $R = 0.3$.

For Example 1, the original miscalibrated risk scores are underestimated where the event rate is $R = 0.3$ and otherwise overestimated (see Figure A.1 for details). Figure 5.1 shows some illustrative calibration curves for a single simulation. The sNB of the original risk model before any recalibration is 43.3%. We see that every recalibration

method improves the sNB but targeted spline recalibration produces recalibrated risk scores with the highest sNB on average. For either training set size, the average sNB from targeted spline recalibration is about 51%, an absolute improvement of about 7%. Standard logistic and constrained logistic recalibration perform similarly to one another and only increase sNB by about 1-2% when N=500 and 2-3% when N=1000.

In Example 2, all original miscalibrated risks are underestimated (see Figure A.2 for details). Figure 5.2 shows calibration curves for a single simulation. The sNB of the original risk model before any recalibration is 48.7%. Every recalibration method improves the sNB but targeted spline recalibration produces the best result, giving an average sNB over 53% with either size training set.

For Example 3, the original miscalibrated risk scores are overestimated near the risk threshold $R = 0.3$ and underestimated elsewhere (see Figure A.3 for details). Figure 5.3 shows calibration curves for a single simulation. The sNB of the original risk model before any recalibration is 43.7%. Every recalibration method improves the sNB but targeted spline recalibration produces the highest sNB, averaging about 51% with either size training set.

In Example 4, all original miscalibrated risks are overestimated (see Figure A.4 for details). Figure 5.4 shows calibration curves for a single simulation. The sNB of the original risk model before any recalibration is 28.8%, a low starting value compared to Examples 1-3. All recalibration methods increase the sNB by 12% or more on average, including standard logistic recalibration. Targeted spline recalibration improves sNB the most, a couple percent beyond the improvement from using standard or constrained logistic recalibration.

Tables 5.3 and 5.4 summarize sNB for simulation Examples 5 through 8 using training datasets with sample size N = 500 and 1000, respectively, and $R = 0.075$.

For Example 5, all original miscalibrated risks are overestimated, including slight overestimation near $R = 0.075$ (see Figure A.5 for details). Figure 5.5 shows calibration curves for a single simulation. The sNB of the original risk model is 50.9%. Every

recalibration method improves the sNB. Standard and constrained logistic recalibration give very similar improvements with resulting sNB between 53% and 54% on average. Targeted spline recalibration gives sNB near 58% for either sample size. The method using a knot based on the prevalence slightly outperforms targeted spline recalibration with an average sNB between 58% and 59%, the best performance of all the recalibration methods for this example.

Example 6 has the same miscalibration behavior as in Example 5, where risks are overestimated including slight overestimation near $R = 0.075$ (see Figure A.5 for details). Figure 5.6 shows calibration curves for a single simulation. The sNB of the original risk model is 56.8%. Every recalibration method improves upon this sNB by at least 7%. Standard and constrained logistic recalibration both result in sNB of about 64% on average. Targeted spline recalibration gives even greater sNB around 66-67% for either sample size. The method using a knot based on the prevalence has the best performance of all the recalibration methods, slightly outperforming targeted spline recalibration but by less than half a percent sNB on average.

In Example 7, original miscalibrated risks are underestimated at very low values and overestimated elsewhere starting near $R = 0.075$ (see Figure A.6 for details). Figure 5.7 shows calibration curves for a single simulation. The sNB of the original risk model is 59%. In this setting, we see standard and constrained logistic recalibration provide virtually no improvement in sNB. Targeted spline recalibration improves the original sNB by about one percent, with slightly larger improvement with more training data (N=1000 versus 500). Several of the other recalibration methods produce results very similar to the targeted spline recalibration method (Beta, knot based on prevalence, knot at R).

Example 8 has the same miscalibration behavior as in Example 7 where very low risks are underestimated and higher risks are overestimated, including risks near $R = 0.075$ (see Figure A.6 for details). Figure 5.8 shows calibration curves for a single simulation. The sNB of the original risk model is 66.8%. In this setting, we see standard

and constrained logistic recalibration provide no improvement in sNB. Targeted spline recalibration improves the original sNB by over one percent on average. Targeted spline recalibration yields the highest sNB among the recalibration methods. Several of the other recalibration methods produce results very similar to the targeted spline recalibration method (Beta, knot based on prevalence, knot at R).

We also find that the sNB variability (SD) from our proposed method is usually comparable to that of the other recalibration methods. In fact, for Examples 1-4, the variability in sNB for targeted spline recalibration is always lower than for the other methods.

Table 5.1: Standardized Net Benefit of recalibrated risk scores for Examples 1-4 based on training dataset of size N=500 and 1000 simulations.

sNB mean (SD)				
	Miscalibration Setting			
Recalibration Method	Example 1	Example 2	Example 3	Example 4
Original (uncalibrated)	43.3	48.7	43.7	28.8
Standard Logistic	45.0 (1.7)	51.4 (0.9)	46.3 (2.2)	41.5 (1.6)
Constrained Logistic	45.4 (1.9)	51.4 (0.9)	47.1 (2.3)	41.5 (1.6)
Beta calibration	47.6 (1.2)	52.5 (1.0)	50.9 (0.7)	42.9 (1.3)
Single Knot Placement at:				
Event rate = 0.3 (proposed method)	50.9 (0.6)	53.7 (0.7)	51.3 (0.4)	43.3 (1.0)
RS = 0.3	49.4 (0.8)	52.8 (1.0)	51.0 (0.6)	42.5 (1.4)
Median *RS* (all)	[1] 47.3 (1.8)	[2] 52.2 (1.1)	50.1 (1.0)	41.2 (1.9)
Median *RS* (cases)	48.4 (1.4)	52.2 (1.2)	49.7 (2.1)	42.7 (1.4)
Event Rate = prevalence	50.2 (1.5)	53.2 (1.0)	51.2 (0.5)	42.5 (1.6)

Knot placement resulted in zero cases in the pre-knot interval and method reverted to standard logistic recalibration, out of 1000 simulations:

1) 14 times
2) 1 time

Table 5.2: Standardized Net Benefit of recalibrated risk scores for Examples 1-4 based on training dataset of size N=1000 and 1000 simulations.

sNB mean (SD)				
	Miscalibration Setting			
Recalibration Method	Example 1	Example 2	Example 3	Example 4
Original (uncalibrated)	43.3	48.7	43.7	28.8
Standard Logistic	45.0 (1.2)	51.3 (0.7)	46.2 (1.6)	41.6 (1.2)
Constrained Logistic	46.1 (1.9)	51.3 (0.8)	47.7 (2.0)	41.7 (1.2)
Beta calibration	47.7 (0.8)	52.6 (0.8)	51.0 (0.5)	43.1 (0.8)
Single Knot Placement at:				
Event rate = 0.3 (proposed method)	51.1 (0.4)	53.9 (0.5)	51.4 (0.2)	43.6 (0.5)
RS = 0.3	49.4 (0.6)	52.9 (0.8)	51.0 (0.4)	42.7 (0.9)
Median *RS* (all)	[1] 47.3 (1.3)	52.2 (0.9)	50.1 (0.7)	41.3 (1.4)
Median *RS* (cases)	48.5 (1.0)	52.1 (1.0)	50.4 (1.3)	43.1 (0.8)
Event Rate = prevalence	50.5 (1.0)	53.4 (0.7)	51.3 (0.3)	42.8 (1.0)
Knot placement resulted in zero cases in the pre-knot interval and method reverted to standard logistic recalibration, out of 1000 simulations: 1) 1 time				

Table 5.3: Standardized Net Benefit of recalibrated risk scores for Examples 5-8 based on training dataset of size N=500 and 1000 simulations.

sNB mean (SD)				
	Miscalibration Setting			
Recalibration Method	Example 5	Example 6	Example 7	Example 8
Original (uncalibrated)	50.9	56.8	59.0	66.8
Standard Logistic	53.4 (3.0)	63.9 (1.1)	59.0 (0.5)	66.7 (0.5)
Constrained Logistic	53.3 (3.2)	63.9 (1.2)	59.1 (0.5)	66.8 (0.6)
Beta calibration	55.2 (2.4)	65.1 (1.1)	60.0 (0.4)	67.9 (0.5)
Single Knot Placement at:				
Event rate = 0.075 (proposed method)	[1] 57.9 (2.1)	[5] 66.3 (1.8)	[8] 59.9 (0.8)	[12] 68.0 (0.6)
RS = 0.075	[2] 54.4 (3.6)	[6] 64.8 (2.2)	[9] 59.9 (0.4)	[13] 68.0 (0.4)
Median *RS* (all)	[3] 54.9 (3.1)	[7] 64.8 (2.1)	[10] 59.1 (0.7)	[14] 66.9 (0.5)
Median *RS* (cases)	55.4 (2.8)	64.2 (1.2)	59.1 (0.9)	66.8 (0.5)
Event Rate = prevalence	[4] 58.4 (1.9)	[5] 66.7 (1.5)	[11] 59.9 (0.7)	[5] 67.9 (0.6)

Knot placement resulted in zero cases in the pre-knot interval and method reverted to standard logistic recalibration, out of 1000 simulations:

1) 9 times
2) 258 times
3) 175 times
4) 7 times
5) 1 time
6) 189 times
7) 199 times
8) 10 times
9) 39 times
10) 142 times
11) 11 times
12) 3 times
13) 16 times
14) 198 times

Table 5.4: Standardized Net Benefit of recalibrated risk scores for Examples 5-8 based on training dataset of size N=1000 and 1000 simulations.

sNB mean (SD)				
	Miscalibration Setting			
Recalibration Method	Example 5	Example 6	Example 7	Example 8
Original (uncalibrated)	50.9	56.8	59.0	66.8
Standard Logistic	53.6 (2.0)	64.0 (0.8)	58.9 (0.3)	66.7 (0.4)
Constrained Logistic	53.8 (2.4)	64.0 (0.8)	59.0 (0.4)	66.8 (0.5)
Beta calibration	55.5 (1.6)	65.2 (0.8)	60.1 (0.3)	68.0 (0.4)
Single Knot Placement at:				
Event rate = 0.075 (proposed method)	58.4 (1.7)	67.0 (1.5)	60.1 (0.4)	68.2 (0.3)
RS = 0.075	[1] 56.1 (3.1)	[3] 65.5 (1.6)	60.0 (0.3)	68.0 (0.3)
Median *RS* (all)	[2] 56.2 (2.5)	[4] 65.5 (1.6)	[5] 59.1 (0.4)	[6] 66.8 (0.4)
Median *RS* (cases)	55.8 (1.9)	64.2 (0.8)	59.1 (0.3)	66.7 (0.3)
Event Rate = prevalence	58.9 (1.4)	67.2 (1.1)	60.1 (0.4)	68.2 (0.4)

Knot placement resulted in zero cases in the pre-knot interval and method reverted to standard logistic recalibration, out of 1000 simulations:

1) 91 times
2) 30 times
3) 39 times
4) 42 times
5) 29 times
6) 44 times

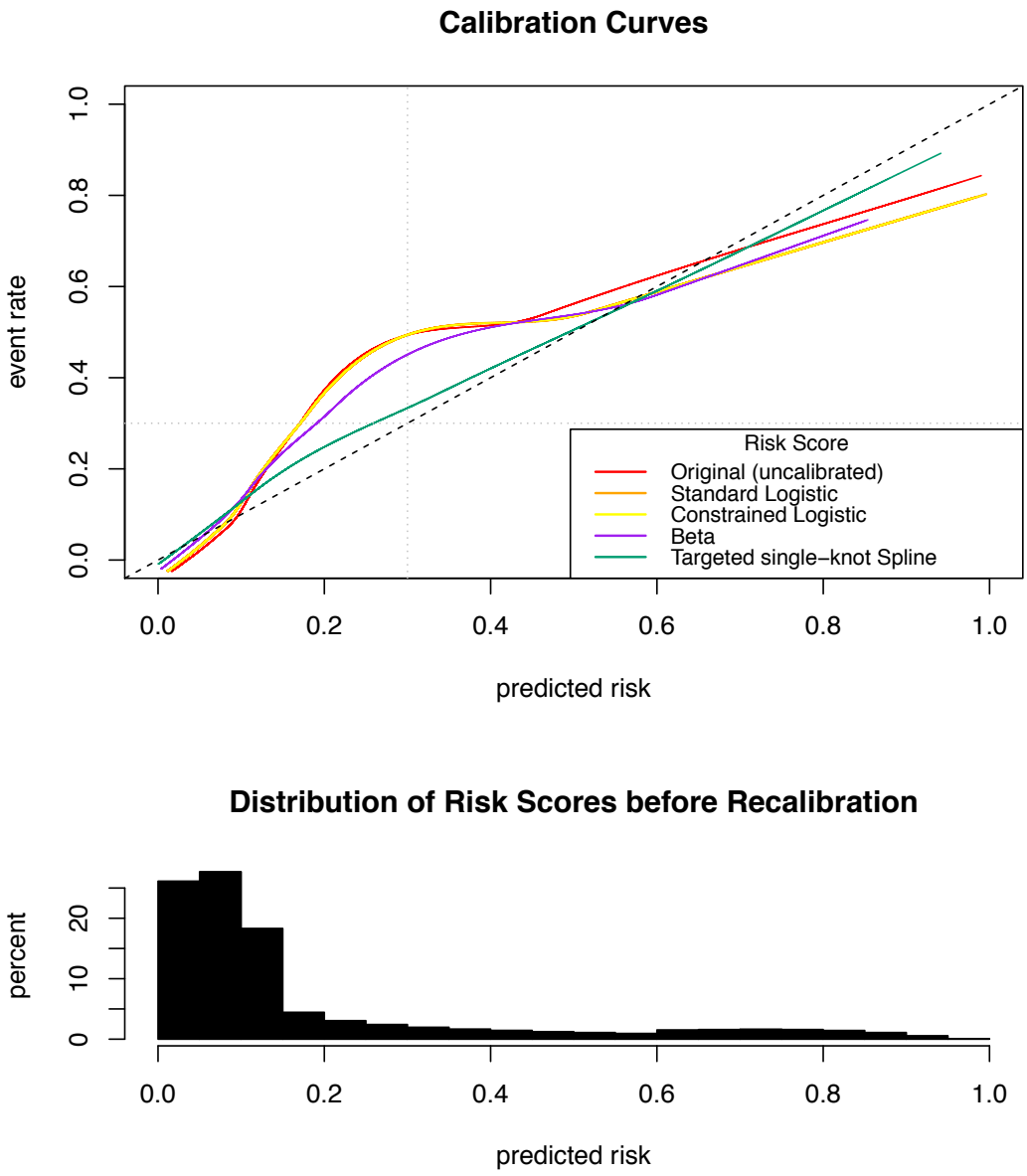


Figure 5.1: **Example 1** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.30.

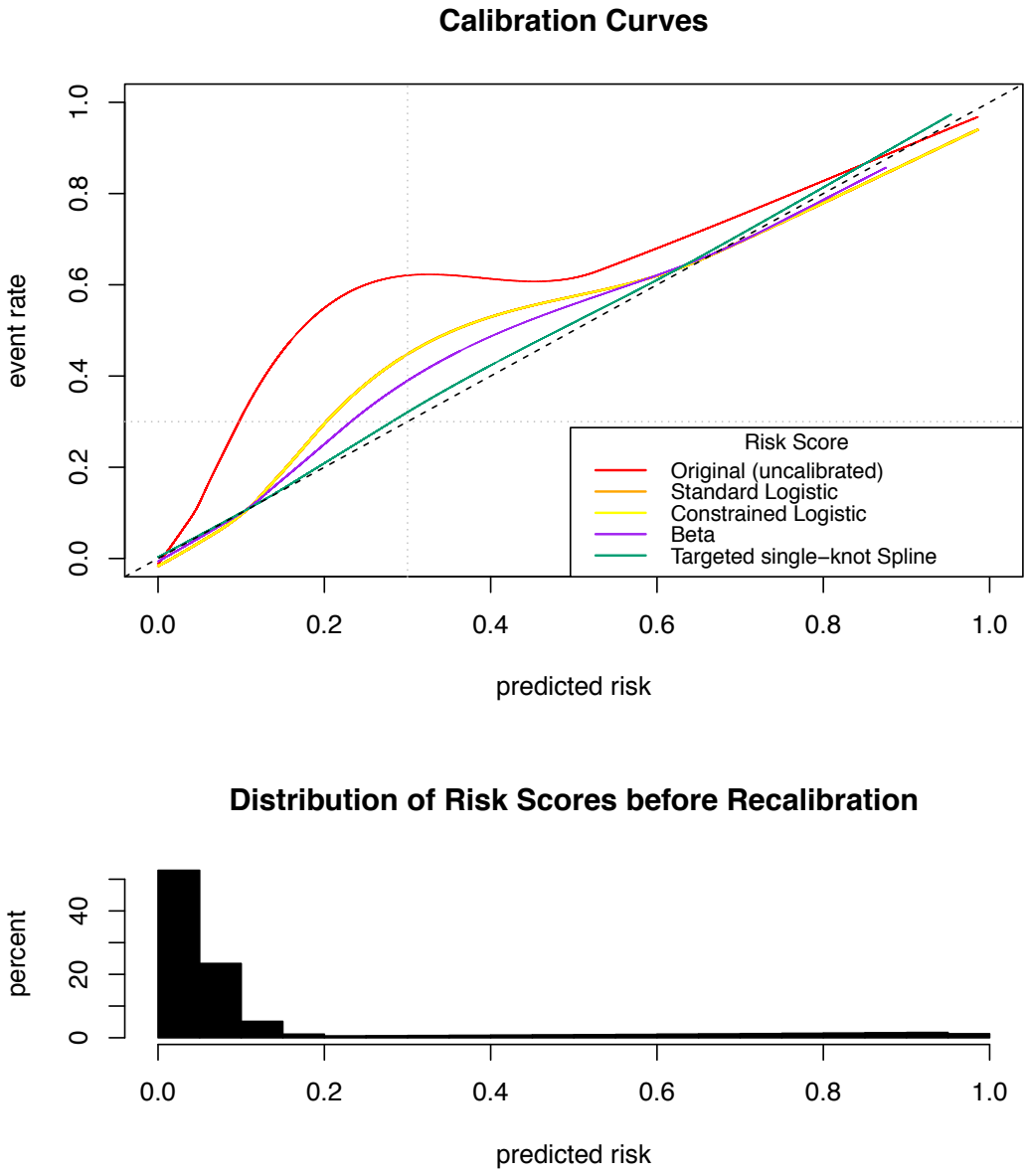


Figure 5.2: **Example 2** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.30.

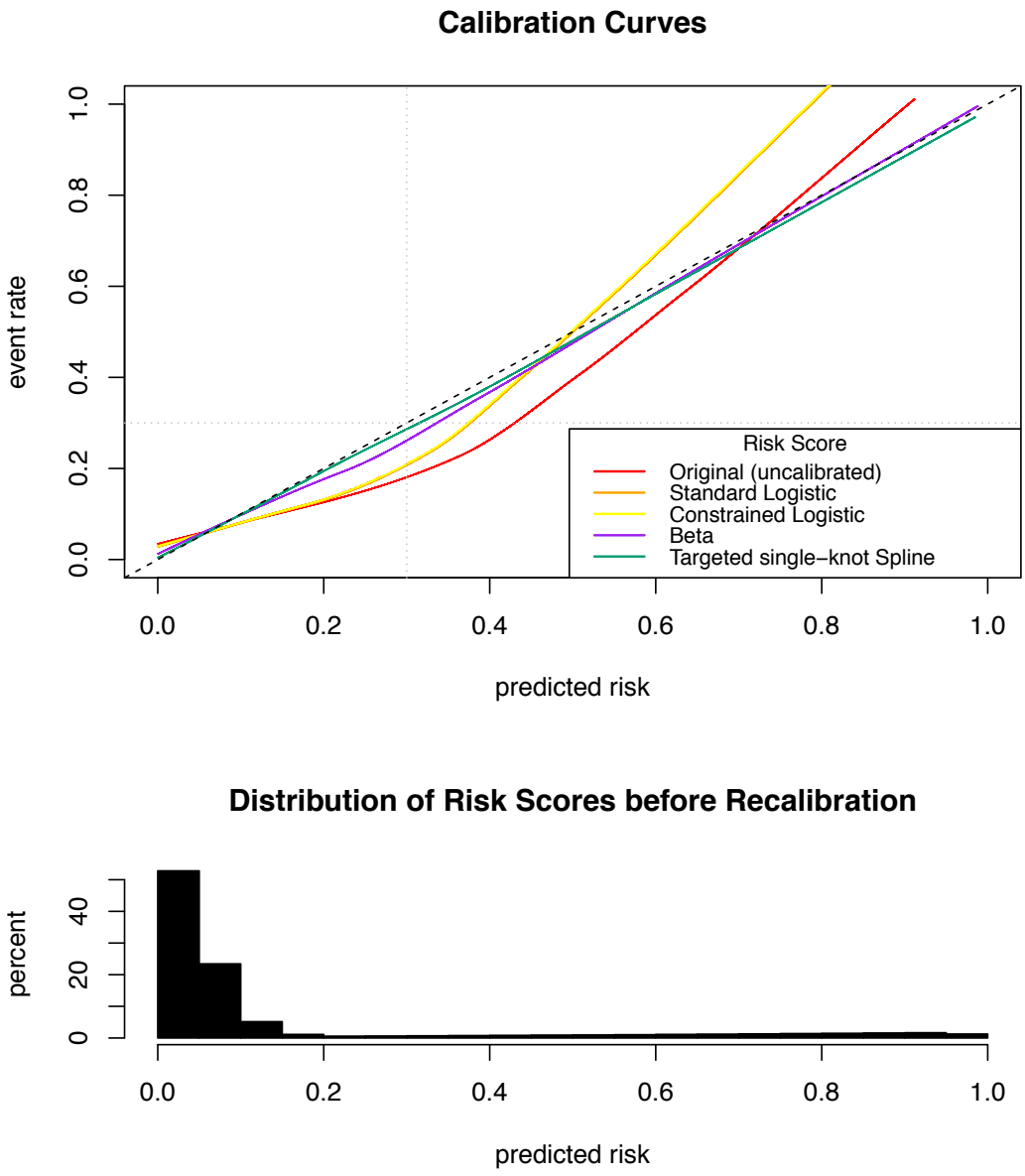


Figure 5.3: **Example 3** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.30.

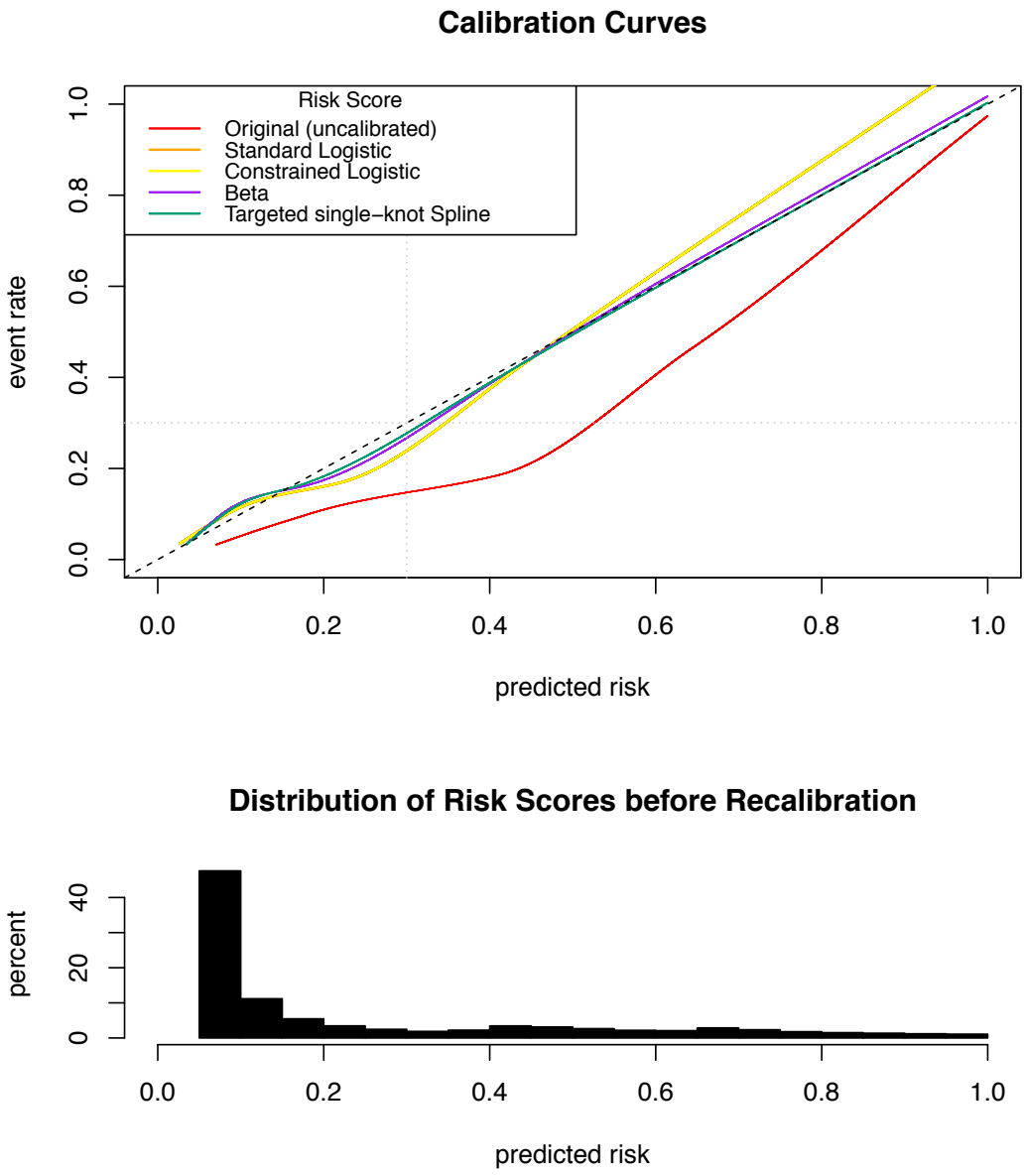


Figure 5.4: **Example 4** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.30.

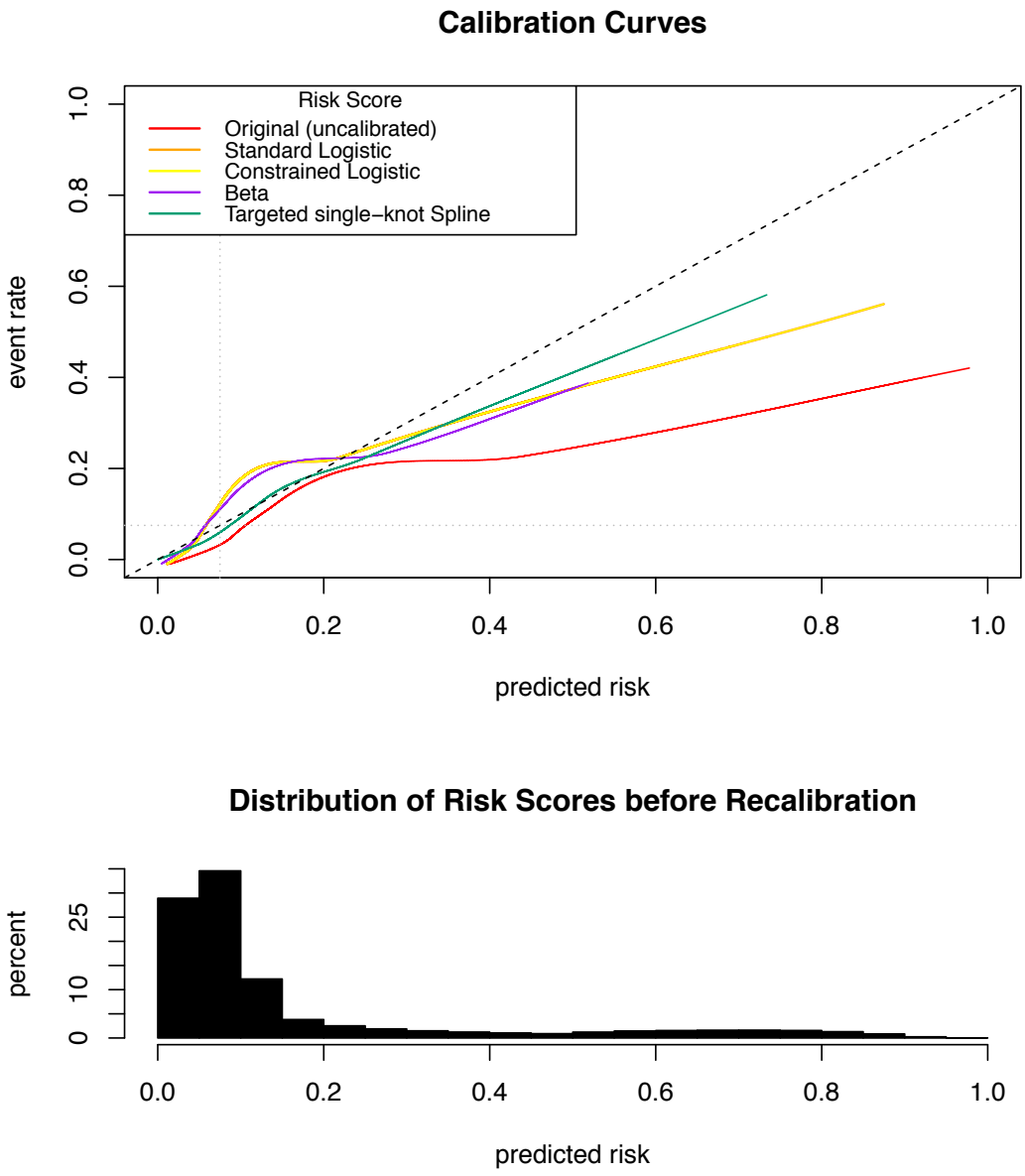


Figure 5.5: **Example 5** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.075.

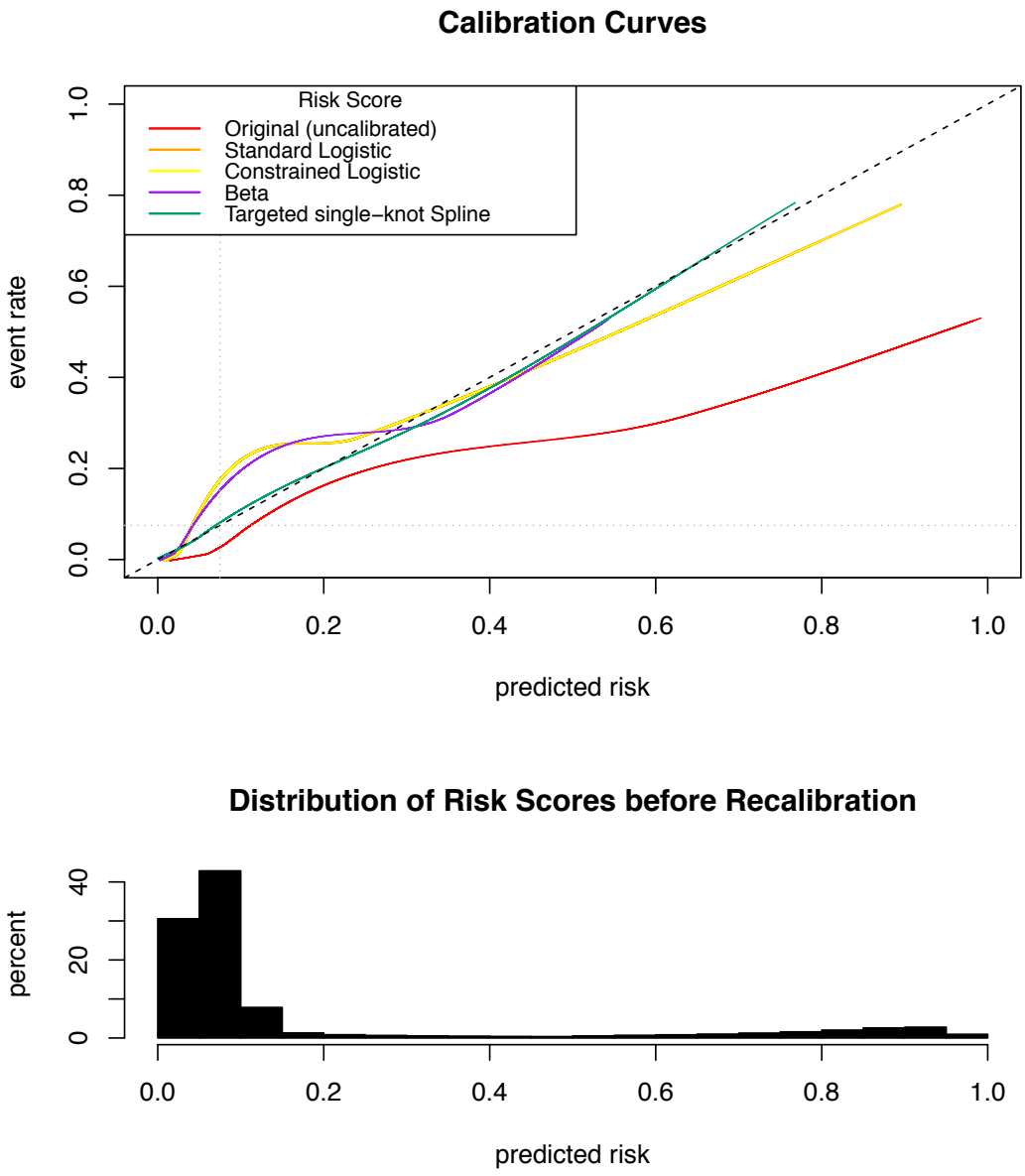


Figure 5.6: **Example 6** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.075.

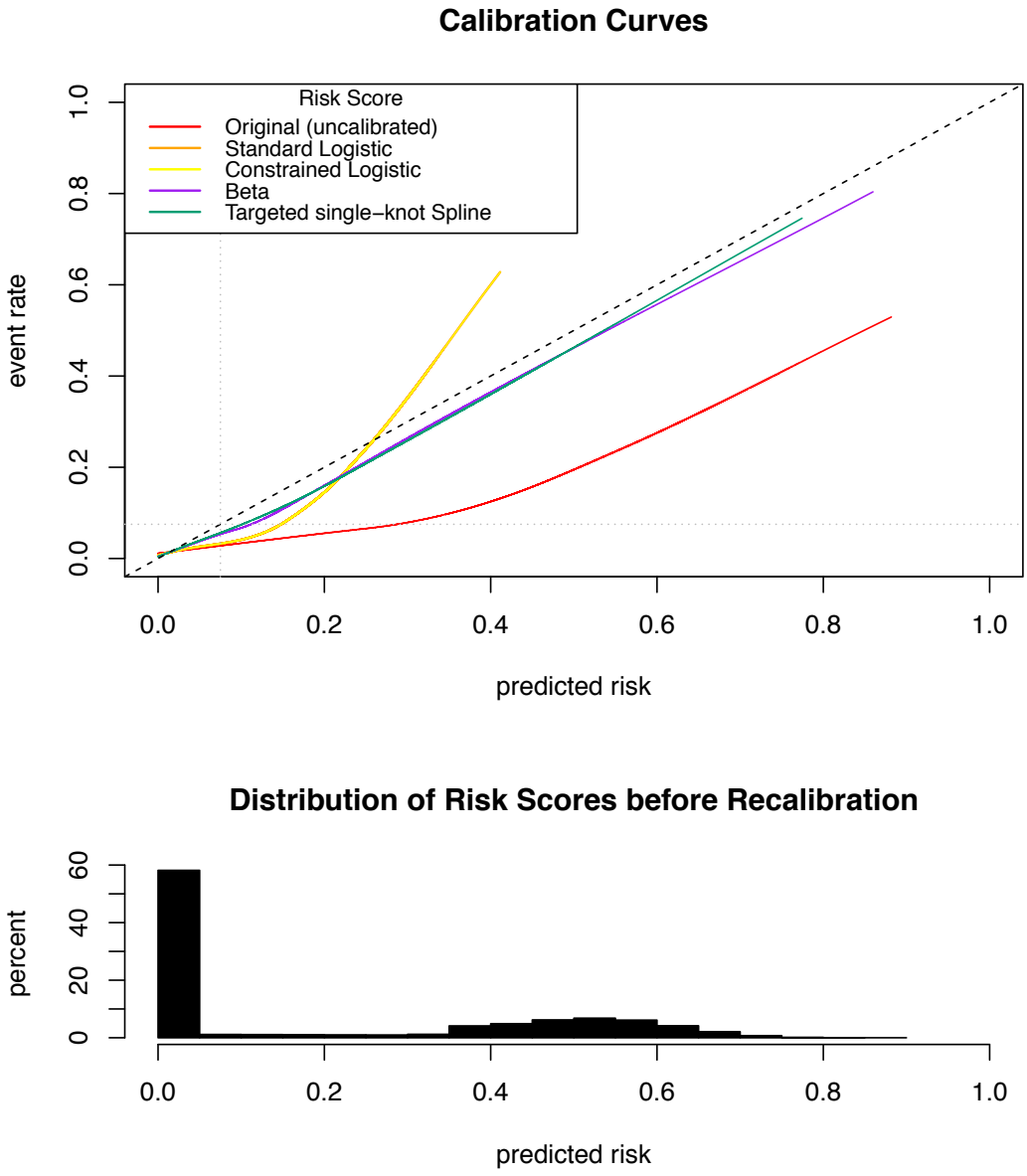


Figure 5.7: **Example 7** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.075.

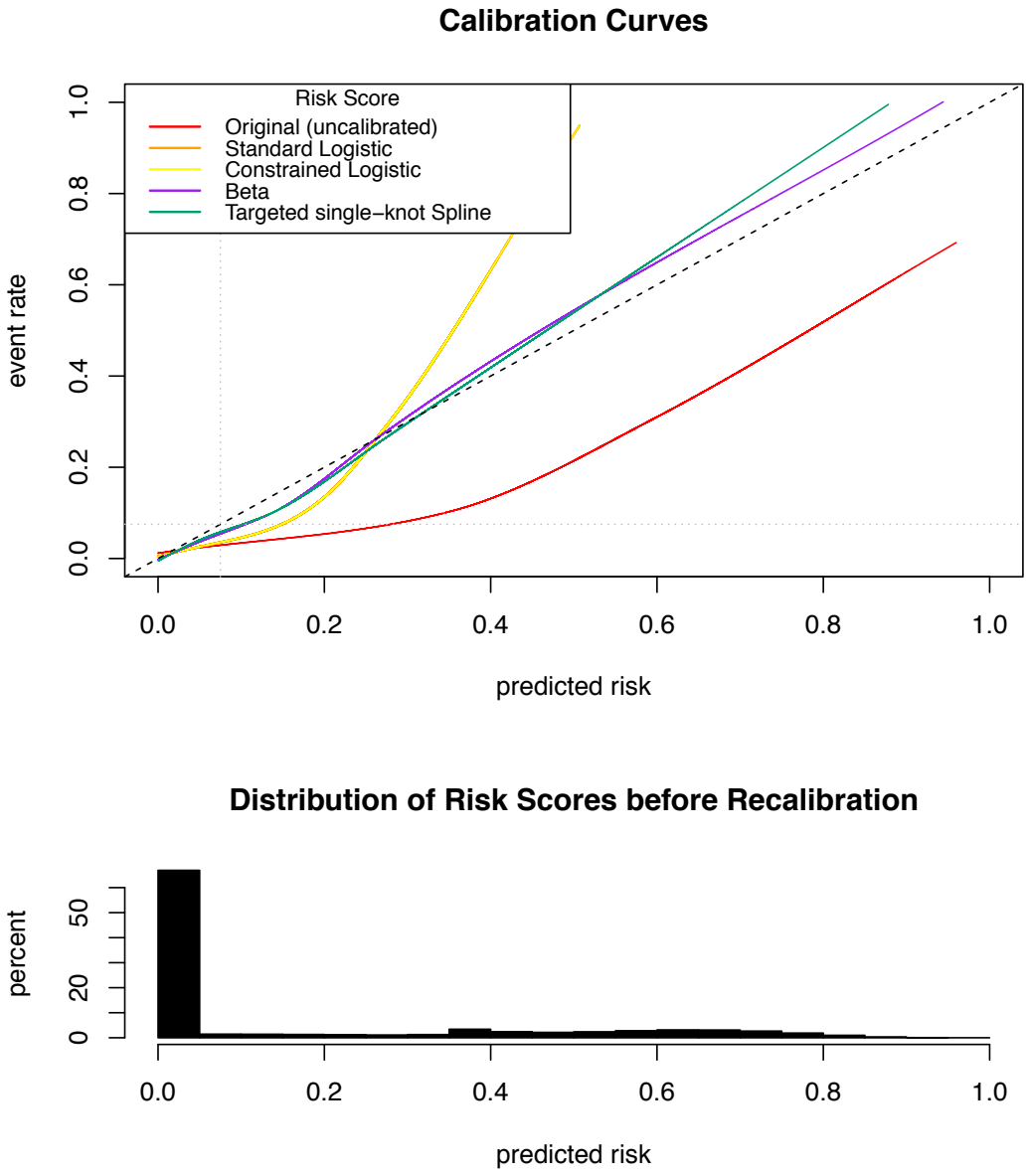


Figure 5.8: **Example 8** Calibration Curves and Original Risk Score Distribution from one simulation (N=1000 observations used for estimating recalibration functions and 1 million independent observations used for constructing calibration curves). Calibration curves are shown for original (uncalibrated) risk scores and recalibrated risk scores from standard logistic, constrained logistic, Beta, and targeted single-knot spline recalibration methods. The diagonal identity line represents perfect calibration and the gray lines mark the critical risk threshold, 0.075.

Chapter 6

DISCUSSION

The evaluation of risk models often focuses on discrimination, but calibration is an important and distinct aspect of risk model performance. For example, the commonly used performance measure AUC is a measure of discrimination but *not* a measure of calibration or clinical utility; it is agnostic to the benefits and harms of true and false positives [Vickers and Elkin, 2006]. Regardless of whether a risk model discriminates well between cases and controls, if it is miscalibrated then it is misleading to clinicians and patients [Van Calster et al., 2019]. Moreover, if the risk model is miscalibrated at a risk threshold used for decision-making, then the clinical utility of the model is undermined.

We found that our targeted spline approach for risk model recalibration resulted in improved sNB in a variety of situations. We provided evidence that these improvements were not simply the result of a more flexible recalibration function, as our method performed similarly to or better than other splines approaches with a single knot. We do not claim that our method of placing the knot based on R is optimal in every application. In fact, we discovered situations where other choice of knot offered small improvements in sNB compared to our method. However, we claim that our method's knot placement is strategic. It makes sense to give the recalibration function maximal flexibility in the area where we care about calibration the most.

We used linear splines in our presentation of targeted spline recalibration, but we could have used higher order splines, e.g., quadratic or cubic splines. We chose to use linear splines because our goal is to attain more flexibility than logistic recalibration while still fitting a parsimonious function. We are primarily concerned with situations where there is insufficient data to build a new risk model, so there would also be limited data to fit a highly flexible recalibration function. Limited data mean that overfitting is

a concern, so parsimony is important.

Our method involves smoothing the event rate across the risk scores $RS_i \in [0, 1]$ to place the knot. We also explored smoothing the observed event rate on the logit-transformed risk scores ($Z_i = \text{logit}(RS_i)$) to place the knot. Smoothing on the logit scale and on the original, probability scale produced similar sNB results, although using the probability scale tended to produce very slightly higher sNB of recalibrated risk models in our examples.

We propose our targeted spline recalibration method when a risk model shows evidence of miscalibration and the intended application is decision-making based on a clinically established risk threshold. Our method extends standard logistic recalibration and our simulation studies demonstrate it is a viable recalibration approach. Other methods of recalibration performed similarly to our method in some settings, or better by very slim margins in some settings. However, no method performed well as consistently as our targeted method. We conclude that recalibration methods with a similar degree of flexibility can outperform our method, but this is somewhat by chance. In contrast, the targeted method uses what is known theoretically, that calibration at R imparts clinical utility, and we have shown empirically it can indeed improve sNB in a variety of settings.

BIBLIOGRAPHY

[Austin and Steyerberg, 2014] Austin, P. C. and Steyerberg, E. W. (2014). Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 33:517–535.

[Baker et al., 2012] Baker, S., 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.

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

[DeFilippis et al., 2015] 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.

[Goff et al., 2014] Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., Greenland, P., Lackland, D. T., Levy, D., O'Donnell, C. J., Robinson, J. G., Schwartz, J. S., Shero, S. T., Smith, S. C., Sorlie, P., Stone, N. J., and Wilson, P. W. F. (2014). 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation*, 129(25_suppl_2):S49–S73.

[Harrell, 2015] Harrell, F. E. (2015). *Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, 2nd edition.

[Kavousi et al., 2014] Kavousi, M., Leening, M. J. G., Nanchen, D., Greenland, P., Graham, I. M., Steyerberg, E. W., Ikram, M. A., Stricker, B. H., Hofman, A., and Franco, O. H. (2014). Comparison of Application of the ACC/AHA Guidelines, Adult Treatment Panel III Guidelines, and European Society of Cardiology Guidelines for Cardiovascular Disease Prevention in a European Cohort. *JAMA*, 311(14):1416–1423.

[Kerr et al., 2016] Kerr, K. F., Brown, M. D., Zhu, K., and Janes, H. (2016). Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *Journal of Clinical Oncology*, 34(21):2534–2540.

[Kull et al., 2017] Kull, M., Silva Filho, T., and Flach, P. (2017). Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631, Fort Lauderdale, FL, USA.

[Mishra, 2019] Mishra, A. (2019). *Methods for Risk Markers that Incorporate Clinical Utility*. PhD thesis, University of Washington.

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

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

[Pepe et al., 2015] 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.

[Van Calster et al., 2019] Van Calster, B., McLemon, D. J., van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive medicine. *BMC Medicine*, 17(230).

[Van Calster et al., 2016] 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*, 74:167–176.

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

[Vickers and Elkin, 2006] 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.

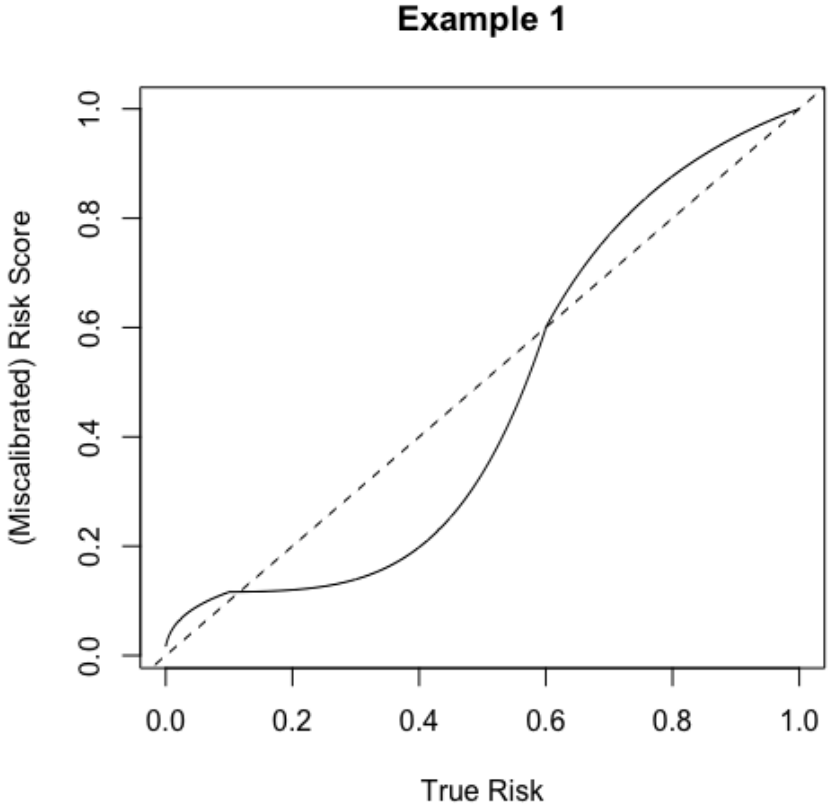
[Vickers et al., 2016] Vickers, A. J., Van Calster, B., and Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352(i6).

Appendix A

SIMULATION SETTING DETAILS

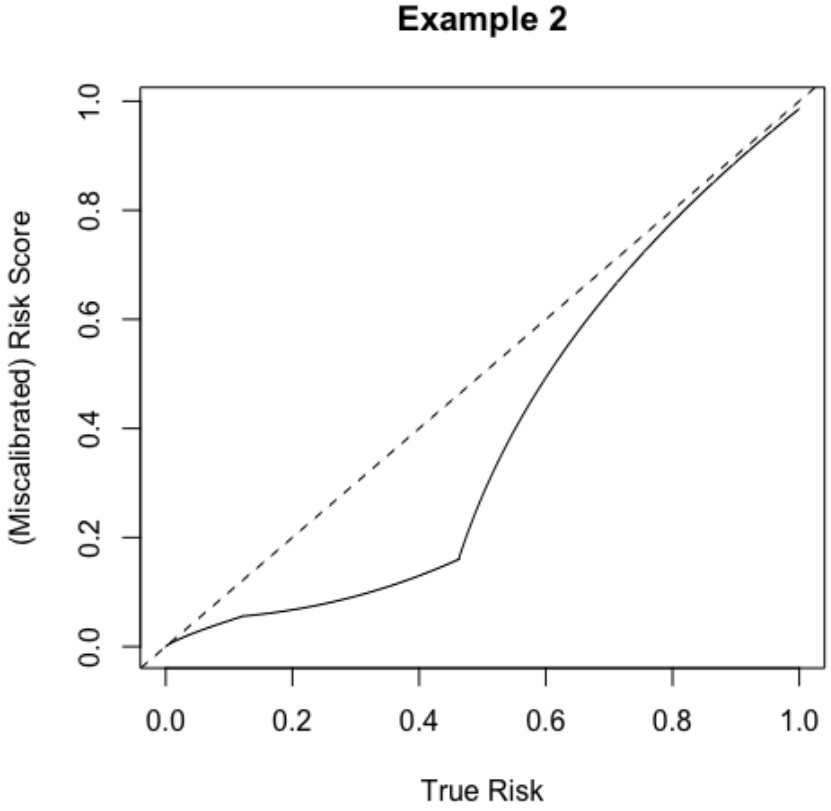
The specific parameters and functions used to simulate true risks and miscalibrated risk scores for each setting, Examples 1 - 8, are reported here. Table A.1 gives details of the Beta mixture distributions. Figures A.1-A.6 show the miscalibration functions.

Figure A.1: Example 1 Miscalibration Function



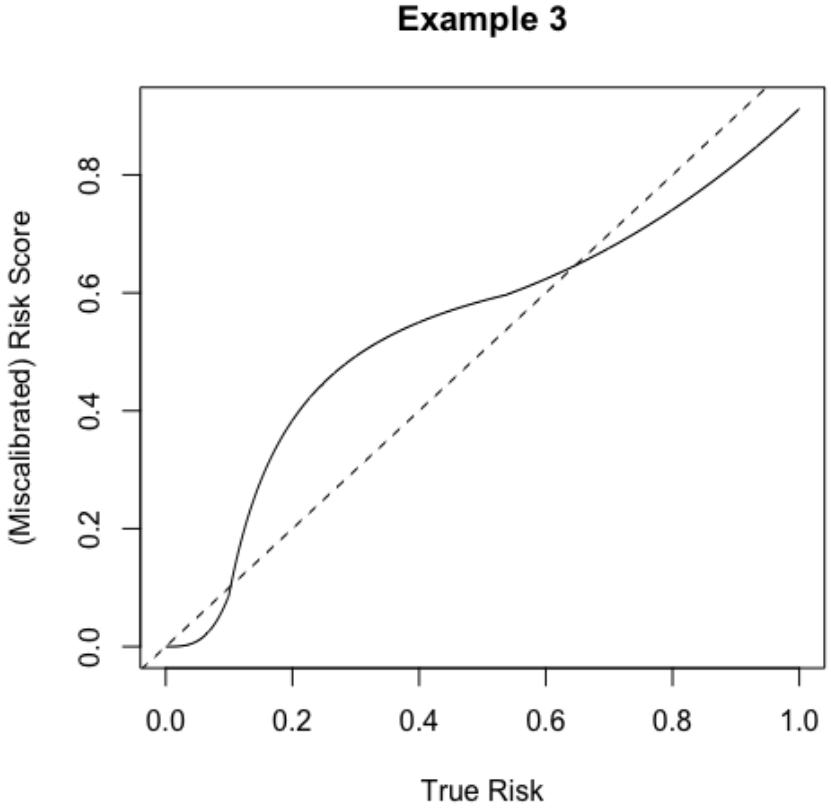
$$f_1(x) = \begin{cases} 0.3(x - 0.003)^{0.26} - 0.05 & x \in [0, 0.1) \\ 4.5x^{4.37} + 0.116 & x \in [0.1, 0.6) \\ -0.2x^{2.15} + 1.2 & x \in [0.6, 1] \end{cases}$$

Figure A.2: Example 2 Miscalibration Function



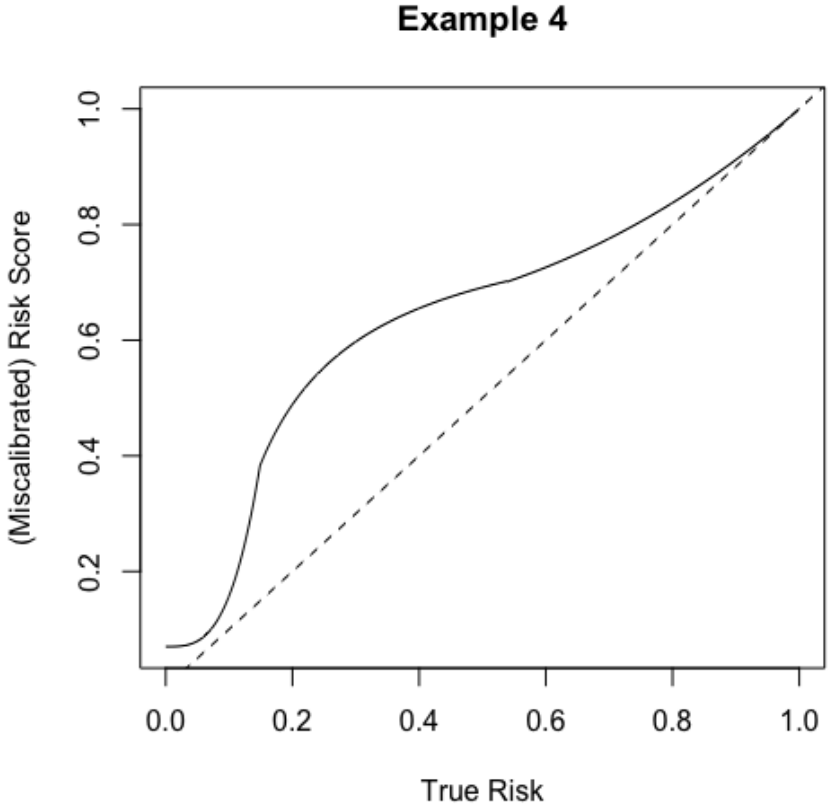
$$f_2(x) = \begin{cases} 0.3x^{0.8} - 0.05 & x \in [0, 0.12) \\ 0.6x^{2.2} + 0.05 & x \in [0.12, 0.464) \\ 1.7(x - 0.4)^{0.4} - 0.4 & x \in [0.464, 1] \end{cases}$$

Figure A.3: Example 3 Miscalibration Function



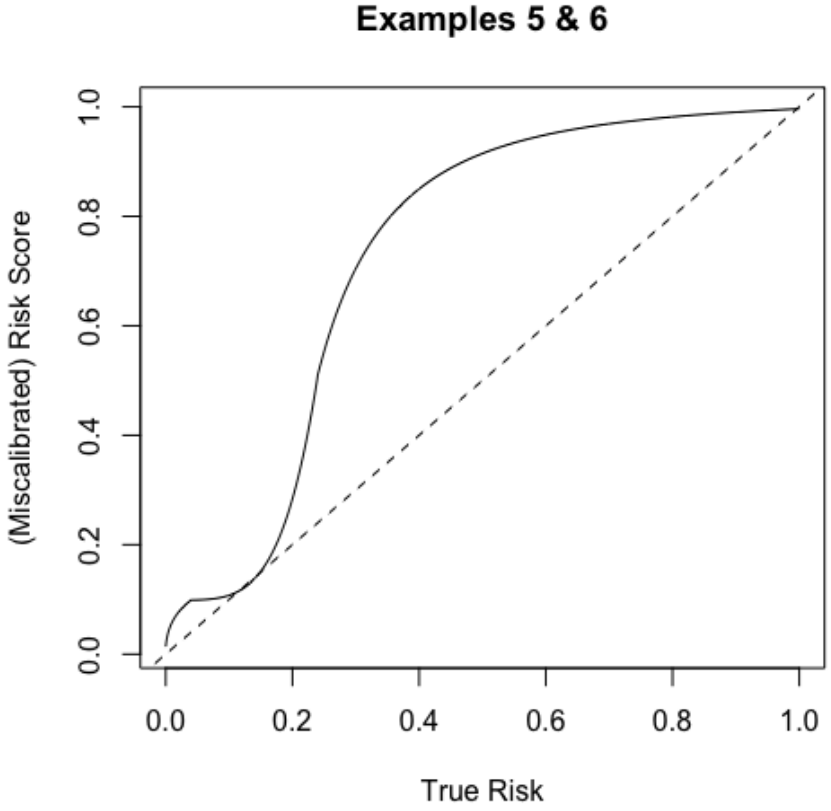
$$f_3(x) = \begin{cases} 139x^{3.2} & x \in [0, 0.1) \\ -0.1x^{-0.83} + 0.764 & x \in [0.1, 0.54) \\ 0.4x^{2.5} + 0.512 & x \in [0.54, 1] \end{cases}$$

Figure A.4: Example 4 Miscalibration Function



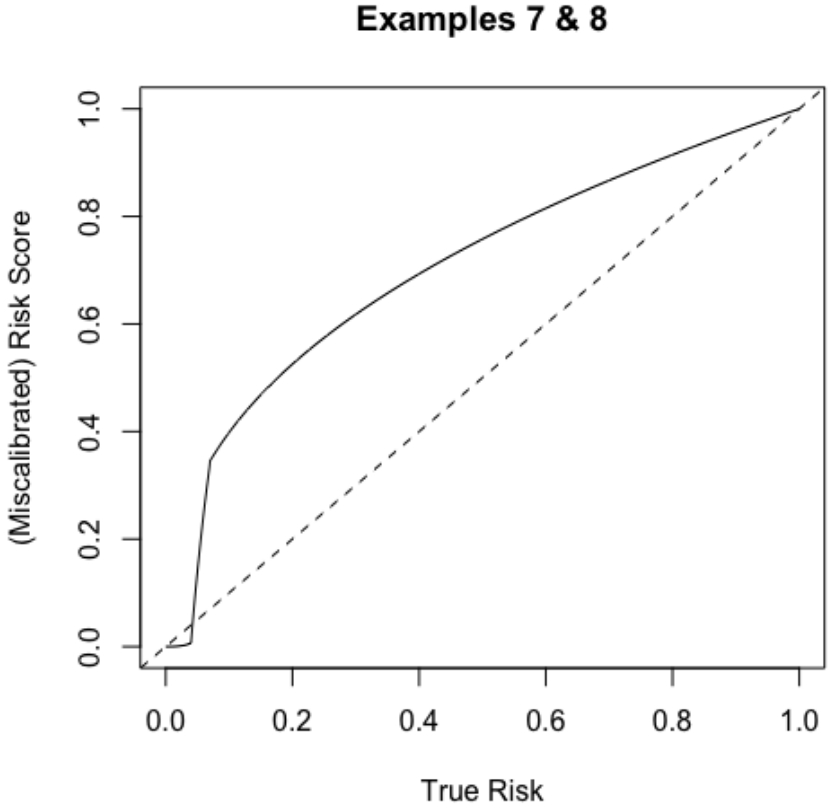
$$f_4(x) = \begin{cases} 139x^{3.2} + 0.07 & x \in [0, 0.15) \\ -0.1x^{-0.83} + 0.869 & x \in [0.1, 0.54) \\ 0.38x^{2.5} + 0.62 & x \in [0.54, 1] \end{cases}$$

Figure A.5: Examples 5-6 Miscalibration Function



$$f_{5-6}(x) = \begin{cases} 0.255(2.5x + 0.003)^{0.26} - 0.0425 & x \in [0, 0.04) \\ 3.825(2.5x)^{4.37} + 0.09775 & x \in [0.04, 0.24) \\ -0.17(2.5x)^{-2.15} + 1.02 & x \in [0.24, 1] \end{cases}$$

Figure A.6: Examples 7-8 Miscalibration Function



$$f_{7-8}(x) = \begin{cases} 100x^3 & x \in [0, 0.04) \\ 8x^{0.1} - 5.79 & x \in [0.04, 0.07) \\ x^{0.4} & x \in [0.07, 1] \end{cases}$$

Table A.1: Details of Beta mixture distributions for generating true risks.

Parameters of Beta mixture distributions used to generate true risks p_i and the corresponding overall event rates $E(p_i)$.			
	Mixing Proportion	α	β
Example 1			
Subpopulation 1	0.33	0.6	19.4
Subpopulation 2	0.34	0.5	9.5
Subpopulation 3	0.33	4.0	4.0
$E(p_i)$	19.2%		
Example 2			
Subpopulation 1	0.34	0.5	9.5
Subpopulation 2	0.33	1.0	8.5
Subpopulation 3	0.33	1.0	1.0
$E(p_i)$	21.7%		
Example 3			
Subpopulation 1	0.34	1.0	19.0
Subpopulation 2	0.33	1.5	8.5
Subpopulation 3	0.33	1.0	1.0
$E(p_i)$	23.2%		
Example 4			
Subpopulation 1	0.4	1.0	19.0
Subpopulation 2	0.4	2.0	18.0
Subpopulation 3	0.2	1.0	1.0
$E(p_i)$	16.0%		
Examples 5 & 7			
Subpopulation 1	0.33	0.6	40.0
Subpopulation 2	0.34	0.5	20.0
Subpopulation 3	0.33	4.0	15.0
$E(p_i)$	8.3%		
Examples 6 & 8			
Subpopulation 1	0.6	0.6	40.0
Subpopulation 2	0.2	1.0	20.0
Subpopulation 3	0.2	3.0	6.0
$E(p_i)$	11.3%		