A Theoretical Statements and Proofs

The results in Lemma 1 and Corollary 1 assume concave ROC curve, which follows from the monotonicity of the risk models we consider in this article. We note that an alternative statement and proof of Lemma 1 by Baker et al. (2012)\(^1\) (equation 6).

**Lemma 1.** *(risk model calibration and ROC slope.) Let \(S\) be a risk model monotonically increasing with event rate for binary outcome \(Y\) and let \(r \in (0, 1)\). Then*

\[
P[Y = 1 | RS = r] = \frac{1}{1 + \text{ROCSLOPE} - \frac{1 - \pi}{\pi}},
\]

*(1)*

*where ROCSLOPE is the slope of the ROC curve of risk score \(S\) at cutpoint \(r\).*

**Proof.** Let \(S\) be a risk score. Let \(f_{S|Y=1}(rs)\) and \(f_{S|Y=0}(rs)\) represent the densities of \(S\) for cases and controls, respectively. Let \(r\) \(\epsilon\) be an \(\epsilon\)-width interval around cutpoint \(r\), where \(r \in (0, 1)\). Here we relate the height of the calibration curve at \(r\) and the slope of the ROC curve for cutpoint \(r\).

\[
P(Y = 1 | S = r) = \lim_{\epsilon \to 0} P(Y = 1 | S \in r_{\epsilon}) = \lim_{\epsilon \to 0} \frac{P(S \in r_{\epsilon} | Y = 1)P(Y = 1)}{P(S \in r_{\epsilon})}
\]

\[
= \lim_{\epsilon \to 0} \frac{1}{1 + \frac{P(S \in r_{\epsilon} | Y = 0)P(Y = 0)}{P(S \in r_{\epsilon} | Y = 1)P(Y = 1)}}
\]

\[
= \lim_{\epsilon \to 0} \frac{1}{1 + \frac{f_{S|Y=0}(rs)}{f_{S|Y=1}(rs)} \frac{1 - \pi}{\pi}}
\]

\[
= \frac{1}{1 + \text{ROCSLOPE}(r)^{-1 \frac{1 - \pi}{\pi}}}
\]

where ROCSLOPE\((r)\) is the slope of the ROC curve at the point corresponding to cutpoint \(r\). The last equality uses an established relationship (Result 4.3 in Pepe (2003)\(^2\)).

Using the results of the above Lemma, we arrive at the following corollary. We note that this corollary also follows directly from optimization requirement shown in Baker et al. (2009).\(^3\)

**Corollary 1.** *(sNB of risk-based treatment policies and calibration of \(S\) at \(R\).) Let \(S\) be a risk score for binary outcome \(Y\), that is monotonically increasing with event rate. Suppose \(S\) is used to select individuals for an intervention based on \(S > R\), where \(R\) is a pre-specified risk threshold that represents the benefits and harms of the intervention. Then \(S\) has maximum sNB among all recalibrated versions of \(S\) if and only if \(S\) is calibrated at \(R\).*
\textbf{Proof.} For a risk score, $S$ with differentiable and concave ROC curve, the cutpoint $r$ that maximizes $sNB$ has
\[
\text{ROCSLOPE}(r) = \frac{1 - \pi C}{\pi B}.
\]
Let $R$ be the risk threshold corresponding to the harms and benefits of an intervention, i.e., \( \frac{R}{1-R} = \frac{C}{B} \). Risk score $S$ is calibrated at $R$ if and only if $R$ maximizes $sNB$ among all possible cutpoints $r$. Suppose $S$ is calibrated at $R$, i.e., \( P(Y = 1|RS) = R \). Then
\[
R = \frac{1}{1 + \text{ROCSLOPE}(R) - \frac{1 - \pi}{\pi}}
\Rightarrow \text{ROCSLOPE}(R) = \frac{1 - \pi C}{\pi B}.
\]
Suppose $R$ maximizes $sNB$ among all possible cutpoints for risk score $S$. Then $\text{ROCSLOPE}(R)$ is
\[
\text{ROCSLOPE}(R) = \frac{1 - \pi C}{\pi B}
\Rightarrow P(Y = 1|RS = R) = \frac{1}{1 + \frac{\pi}{1 - \frac{C}{B} - \pi}} = R,
\]
and $S$ is calibrated at $R$. \hfill \square

\textbf{Proposition.} (variance of $\hat{sNB}(S)$ at cutpoint $r$.)

Let $S$ be a risk model predicting binary random variable $Y$. Assume that the harm to benefit ratio of the intervention associated with the outcome is $\frac{R}{1-R}$. For a random sample of $n$ individuals from the relevant population, the estimated standardized net benefit of the policy that treats individuals with $S > r$ is
\[
\hat{sNB}(S) = \hat{TPR}_{Y}(S) - \frac{R}{1 - R \hat{P}(Y = 1)} \hat{FPR}_{Y}(S)
\]
Define the sample proportions:
\[
\hat{p}_{11} = \frac{1}{n} 1[S \geq r, Y = 1] \quad \hat{p}_{01} = \frac{1}{n} 1[S < r, Y = 1], \quad \hat{p}_{10} = \frac{1}{n} 1[S \geq r, Y = 0] \quad \hat{p}_{00} = \frac{1}{n} 1[S < r, Y = 0].
\]
The estimator in (2) can be expressed as a function of $\hat{p} = \{\hat{p}_{11}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{00}\}$, when $\hat{p}_{11} + \hat{p}_{01} > 0$ and $\hat{p}_{10} + \hat{p}_{00} > 0$ (since we can never divide by 0),
\[
\hat{sNB}(\hat{p}) = \frac{\hat{p}_{11}}{\hat{p}_{11} + \hat{p}_{01}} - \frac{C}{B} \frac{\hat{p}_{10} + \hat{p}_{00}}{\hat{p}_{11} + \hat{p}_{01}} \frac{\hat{p}_{10}}{\hat{p}_{10} + \hat{p}_{00}}.
\]
Then, the variance of $\hat{sNB}(\hat{p})$ given harm to benefit ratio, $\frac{C}{B}$, is
\[
V\left[\hat{sNB}(\hat{p})\right] = \left(\frac{1}{\hat{p}_{11} + \hat{p}_{01}}\right)^2 \left[ \left(\frac{R}{1-R}\right)^2 \hat{p}_{10} + \frac{p_{11}(\hat{p}_{01} + \frac{R}{1-R} \hat{p}_{10})^2}{(\hat{p}_{11} + \hat{p}_{01})^2} + \frac{p_{01}(\frac{C}{B} \hat{p}_{10} - \hat{p}_{11})^2}{(\hat{p}_{11} + \hat{p}_{01})^2} \right],
\]
where
\[
p_{11} = P(S \geq r, Y = 1) \quad p_{01} = P(S < r, Y = 1), \quad p_{10} = P(S \geq r, Y = 0) \quad p_{00} = P(S < r, Y = 0),
\]
and $p_{11} + p_{01} > 0$ and $p_{10} + p_{00} > 0$.
Proof. For $i = 1, \ldots, n$ define

$$
N_{11} = \sum_{i=1}^{n} 1[S_i \geq r, Y_i = 1] \quad N_{01} = \sum_{i=1}^{n} 1[S_i < r, Y_i = 1] \\
N_{10} = \sum_{i=1}^{n} 1[S_i \geq r, Y_i = 0] \quad N_{00} = \sum_{i=1}^{n} 1[S_i < r, Y_i = 0]
$$

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,

$$
\begin{align*}
p_{11} &= P(S \geq r, Y = 1) \\
p_{01} &= P(S < r, Y = 1), \quad \text{(8)} \\
p_{10} &= P(S \geq r, Y = 0) \\
p_{00} &= P(S < r, Y = 0), \quad \text{(9)}
\end{align*}
$$

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

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

where

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

The estimand of (2) can be expressed as a function of $\bar{p} = \{p_{11}, p_{10}, p_{01}, p_{00}\}$,

$$
\begin{align*}
sNB(\bar{p}) &= \frac{p_{11}}{p_{11} + p_{01}} - R \left( \frac{p_{10} + p_{00}}{p_{11} + p_{01}} \right) \frac{p_{10}}{p_{11} + p_{00}} \\
\quad &= \frac{p_{11}}{p_{11} + p_{01}} - \frac{R}{1 - R} \frac{p_{11} + p_{01}}{p_{11} + p_{01}} \frac{p_{10}}{p_{11} + p_{00}} \\
\quad &= \frac{1}{p_{11} + p_{01}} \left[ p_{11} - \frac{R}{1 - R} p_{10} \right], \quad \text{(10)}
\end{align*}
$$

The gradient of (11) is

$$
\nabla sNB(\bar{p}) = \begin{pmatrix} \frac{p_{11} + R}{p_{11} + p_{01}} \\ \frac{p_{10} + R}{(p_{11} + p_{01})^2} \\ -\frac{p_{11} + R}{(p_{11} + p_{01})^2} \end{pmatrix}.
$$

Then by the invariance property of the maximum likelihood estimator and invoking the $\delta$-method,

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

where

$$
\Sigma_2 = \nabla sNB(\bar{p})^T \Sigma \nabla sNB(\bar{p})
= \left( \frac{1}{p_{11} + p_{01}} \right)^2 \left[ \left( \frac{R}{1 - R} \right)^2 \frac{p_{11}(p_{11} + R p_{10})^2}{(p_{11} + p_{01})^2} + \frac{p_{01}(R p_{10} - p_{11})^2}{(p_{11} + p_{01})^2} \right].
$$
B Calculating Weights for Weighted Logistic Recalibration

As described in section 2.3, the weight function used in weighted logistic recalibration has flexible forms. The choice of weight function depends on the context in which the risk score will be used and where good calibration of the risk score is most important. In the following section we provide more detail on construction of the weight functions and guidance for selecting associated tuning parameters, including a cross-validation procedure.

B.1 Weight components and weight shape

The weight function is comprised of five components: the risk threshold $R$, the smoothed observed event rate $o(RS_i)$, the critical risk interval $[R_l, R_u]$ containing $R$, and tuning parameters $\lambda$ and $\delta$. The smoothed observed event rate is obtained from a LOESS regression of $Y_i$ on $RS_i$, and is used to identify observations with event rate close to $R$. The critical risk interval defines a range of risk scores around the risk threshold where calibration has heightened clinical importance. Finally, $\delta$ sets a constant weight outside the critical risk interval and $\lambda$ sets the rate at which observations further from $R$ are down-weighted.

As shown in Figure 2 in the main text, the weight function can take an exponential or step function form. Depending on the desired shape of the weight function some parameters may be considered ancillary. Under a fully exponential decay weight shape, as pictured in Figure 2(A) of the main text, recalibration is focused at the risk threshold. In this setting there may not be an already defined clinically relevant critical risk intervals, and one can set $[R_l, R_u] = [0, 1]$. Then, $\delta$ is not a relevant parameter for the weight function; only $\lambda$ requires specification. Under the step function weight form, as pictured in Figure 2(B) of the main text, the critical risk interval and $\delta$ should be specified, and $\lambda$ can be arbitrarily large (e.g $\lambda \geq 10$) to achieve a “flat” weight function within the critical risk interval. Variations between these weight forms can be easily achieved by varying $\lambda$, $\delta$, or $[R_l, R_u]$.

Before selecting a tuning parameter, we recommend that researchers first select the general shape of the weight (exponential, step, or some intermediary). Specifying the shape will identify which tuning parameters are necessary to specify, and which tuning parameters are ancillary.

B.2 Tuning parameter selection and relative average weights (RAW)

Once the general shape is chosen the next step is to select tuning parameters. However, neither $\lambda$ nor $\delta$ has a clinical interpretation, which contributes to the challenge of selecting these tuning parameters.

We define a more interpretable measure, the relative average weight (RAW), that can be used to elicit or select $\lambda$ or $\delta$. RAW compares the average weight of observations within $[R_l, R_u]$ to the average weight of observations outside the interval. For the step function weight form, the critical risk interval is already defined. For exponential decay weights, such as shown in Figure 2(A), $[R_l, R_u]$ is selected and used exclusively for defining RAW and is not otherwise used in the weighting scheme. We define the relative average weight

$$ RAW = \frac{n_R \sum_{i=1}^{n} w_i 1(o(S_i) \notin (R_l, R_u))}{n_R \sum_{i=1}^{n} w_i 1(o(S_i) \in (R_l, R_u))}, $$

(12)

where $n_R$ ($n_R$) is the number of observations within (outside) the interval $[R_l, R_u]$.

A useful interpretation of RAW is a summary metric capturing how much down-weighting is applied. RAW is bounded by $(0, 1)$. A RAW value close to 1 implies that there is little differential weighting and $\lambda$ or $\delta$ are large. 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 $\lambda$ or $\delta$.

If there is a prior notion of how much observations outside the interval should be down-weighted, $\lambda$ or $\delta$ can be identified by fixing one parameter (depending on desired weight shape) and numerically solving for the other using (12). Otherwise, a range of RAW values can be used to obtain a grid of $\lambda$ (or $\delta$ for step weight function). We propose a cross-validation procedure to select the tuning parameter based on maximizing $sNB$. 

4
B.3 Cross-validation for tuning parameter selection

In this section we present a repeated $K$-fold cross-validation procedure for selecting tuning parameters. We designed the cross-validation procedure to select the tuning parameter that yields highest $sNB$ when weighted recalibration is employed. An important note is that in order to implement this cross-validation procedure, at least either $\delta$ or $\lambda$ must be pre-specified, as well as $[R_l, R_u]$. As discussed in section B.2, the choice of which parameter is pre-specified depends on the desired weight function form. For convenience, we describe the procedure when an exponential decay type weight function is desired (as in Figure 2A of main text), and therefore $\delta$ is fixed $a$ priori and tuning is conducted to select $\lambda$. The procedure is equivalent if a step weight function is desired (as in Figure 2B of main text), where $\lambda$ is fixed and a cross-validation is used to select $\delta$.

I: Obtain grid of tuning parameters
As described in section 2.3, tuning parameters $\delta$ and $\lambda$ do not have clear clinical interpretation. Instead, a range of suitable RAW values can be used to obtain a grid of $\lambda$, (e.g. $\text{RAW} \in [0.1, 0.2, \ldots, 0.9]$). Given values of $[R_l, R_u]$, $\delta$, and RAW, solve for $\lambda$ using equation (7), using uniroot function in R. Repeat for all values of RAW values to obtain a grid of $\lambda$ values.

II: Implement single round of $K$-fold cross-validation
Once a range of tuning parameters is obtained, randomly partition data into $K$ equal folds. Calculate the estimated $sNB$ of $\lambda$ ($\delta$) within fold $k$, $sNB_{k,m}(\lambda)$. Repeat for all $K$ folds, and define

$$cv.sNB_m(\lambda) = \frac{1}{K} \sum_{i=k}^{K} cv.sNB_{k,m}(\lambda)$$

to be the cross-validated estimate of $sNB$ from one round of $K$-fold cross-validation for tuning parameter $\lambda$. This process is repeated for the grid of $\lambda$ values obtained in step I.

III: Repeat $K$-fold cross-validation M times
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. Repeat step II $M$ times, with a new random partitioning each time. Define

$$cv.sNB(\lambda) = \frac{1}{M} \sum_{m=1}^{M} cv.sNB_m(\lambda)$$

to be the repeated cross-validated estimate of $sNB$ for tuning parameter $\lambda$. We used $cv.sNB(\lambda)$ to denote the average cross-validated estimate of $sNB$ for tuning parameter $\lambda$. We have found that 25 replications of 5-fold cross-validation leads to a suitable reduction in noise.

IV: Compute cross-validated standard error and apply “one-standard error rule”
Following Friedman et al.\textsuperscript{4} we implement a “one-standard error” rule for tuning parameter selection. 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. Under the “one-standard error” rule the selected tuning parameter

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

where $\lambda^*$ is the tuning parameter with largest $cv.sNB(\lambda)$ and $\hat{\sigma}(cv.sNB(\lambda^*))$ is the corresponding standard error of $\lambda^*$. That is, we propose to select the largest tuning parameter (and therefore apply less down-weighting) such that the cross-validated $sNB$ of that tuning parameter is still within one standard error of the tuning parameter with highest observed $cv.sNB(\lambda)$. 

5
To employ the one-standard error rule, we must estimate the standard error of $cv.sNB(\lambda)$, which we will refer to as $\sigma(cv.sNB(\lambda))$. The calculation of $\sigma(cv.sNB(\lambda))$ requires accounting for the dependence between estimates of $cv.sNB_m(\lambda)$. We can therefore think of the variance of $cv.sNB(\lambda)$ as being composed of two parts: (1) the variation in $sNB$ from a single round of $K$-fold cross-validation (i.e. $\text{var}(cv.sNB_m(\lambda))$), which comprises the “within” repetition variance, and (2) the variation between each of the $M$ estimates of $sNB$ (i.e. $\text{var}(cv.sNB_1(\lambda), cv.sNB_2(\lambda), \ldots, cv.sNB_M(\lambda))$), which we refer to as the “between” repetition variation.

Gelman (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” a single chain, and the variance of the statistic across, or “between”, chains. An analogously, we can use this framework to estimate the “within” repetition variance and the “between” repetition variance. We adapt the procedure presented by Gelman (1992) to obtain a variance estimate of $\sigma(cv.sNB)$. We denote the “within” repetition variance as $W$ and the “between” repetition variance as $B$.

The “within” repetition is the mean of $\text{var}(cv.sNB_m(\lambda))$, across the $M$ repetitions,

$$W(\lambda) = \frac{1}{M} \sum_{m=1}^{M} \text{var}(cv.sNB_m(\lambda))$$

$$= \frac{1}{M} \sum_{m=1}^{M} \frac{1}{K-1} \sum_{k=1}^{K} (cv.sNB_{k,m}(\lambda) - cv.sNB_m(\lambda))^2,$$

Following Gelman, we define the “between” repetition variance to be

$$B(\lambda) = K \text{var}[cv.sNB_1(\lambda), cv.sNB_2(\lambda), \ldots, cv.sNB_M(\lambda)]$$

$$= \frac{K}{M-1} \sum_{m=1}^{M} [cv.sNB_m(\lambda) - cv.sNB(\lambda)]^2,$$

where the $K$ multiplier accounts for dependence between each $cv.sNB_m(\lambda)$ and fold size. Note that if $K$ is small and there will be less variability in the estimates of $cv.sNB_m$ across replications, since the each round of cross-validation can produce relatively coarse partitions. Comparatively, if $K$ is large the variability may increase due increased likelihood of observing unique partitions across the $M$ replications.

Combining these two piece give a heuristic estimate of the cross-validation variance. We augment the formula slightly from that given in Gelman (1992) to account for the fact that as the number of cross-validation repetitions increases, the between-repetition variability should decrease, since the likelihood of sampling the same fold partition will increase as repetitions increase. The proposed estimator cross-validation standard error is

$$\hat{\sigma}(cv.sNB(\lambda)) = \sqrt{\frac{K-1}{K}W(\lambda) + \frac{1}{M}B(\lambda)}.$$
C Additional Simulation Results

Table 1 shows the hyper-parameters used to simulate risks for each scenario considered. Figures 1 - 4 show the piece-wise polynomial functions used to induce miscalibration for each simulation example. Figures 5-8 give additional information about each example, including the distribution of predicted risks and calibration curves. Tables 2-5 give detailed numerical results. Finally, Table 6 summarizes results for each example across 500 repeated simulations, showing the proposed methods produce higher average sNB with comparable variability compared to standard logistic recalibration.

Table 1: Hyperparameters for Beta distribution used for generating true risks ($p_i$) in simulation examples

| Subpopulation | $b_m$ | $E[p_i]$ | $\alpha_m$ | $\beta_m$ |
|---------------|-------|---------|-----------|----------|
| Example 1     |       |         |           |          |
| 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        |
| Example 2     |       |         |           |          |
| 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        |
| Example 3     |       |         |           |          |
| 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        |
| Example 4     |       |         |           |          |
| Subpop. 1     | 0.4   | 0.05    | 1         | 19       |
| Subpop. 2     | 0.4   | 0.1     | 2         | 18       |
| Subpop. 3     | 0.2   | 0.5     | 1         | 1        |

Figure 1: Miscalibration setting for simulation example 1. Moderate risks are underestimated, while high and low risks are overestimated.
\[
\begin{align*}
\text{Scenario 2} \\
\text{Original Risk Score} \\
\text{True Risk} \\
\end{align*}
\]

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

\[
\begin{align*}
\text{Scenario 3} \\
\text{Original Risk Score} \\
\text{True Risk} \\
\end{align*}
\]

Figure 3: Miscalibration setting for simulation example 3. Moderate risk scores are overestimated, while low and high risk scores tend towards underestimation.

\[
\begin{align*}
\text{Scenario 4} \\
\text{Original Risk Score} \\
\text{True Risk} \\
\end{align*}
\]

Figure 4: Miscalibration setting for simulation example 4. All risk are overestimated, with moderate risks having worse miscalibration.
Figure 5: Calibration curves and risk score distribution for simulation examples with sample size $N = 10000$. Calibration curves for the original, standard recalibrated, weighted recalibrated, and constrained recalibrated risk models are shown for each simulation example setting. The histogram of the original simulated risk score is also presented. Dotted lines indicate the clinically relevant risk threshold, $R = 0.3$. The dashed is the identity line.
Figure 6: Calibration curves and risk score distribution for simulation examples with sample size $N = 5000$. Calibration curves for the original, standard recalibrated, weighted recalibrated, and constrained recalibrated risk models are shown for each simulation example setting. The histogram of the original simulated risk score is also presented. Dotted lines indicate the clinically relevant risk threshold, $R = 0.3$. The dashed is the identity line.
Figure 7: Calibration curves and risk score distribution for simulation examples with sample size \( N = 1000 \). Calibration curves for the original, standard recalibrated, weighted recalibrated, and constrained recalibrated risk models are shown for each simulation example setting. The histogram of the original simulated risk score is also presented. Dotted lines indicate the clinically relevant risk threshold, \( R = 0.3 \). The dashed is the identity line.
Figure 8: Calibration curves and risk score distribution for simulation examples with sample size $N = 500$. Calibration curves for the original, standard recalibrated, weighted recalibrated, and constrained recalibrated risk models are show for each simulation example setting. The histogram of the original simulated risk score is also presented. Dotted lines indicate the clinically relevant risk threshold, $R = 0.3$. The dashed is the identity line.
Table 2: Performance of standard, weighted (using exponential form and step form weights), and constrained logistic recalibration in Example 1.

| Measure          | Orig. (No Recal.) | Std. Recal. | Wt. Recal. Wt. Recal. | Const. Recal. |
|------------------|-------------------|-------------|------------------------|---------------|
|                  |                   |             | Exponential Wt.         |               |
|                  |                   |             | Step Wt.                |               |
| sNB              | 0.430             | 0.457       | 0.499                  | 0.499         |
| TPR              | 0.581             | 0.654       | 0.745                  | 0.744         |
| FPR              | 0.083             | 0.097       | 0.135                  | 0.135         |
|                  |                   |             | (0.89, 1.13)           |               |
|                  |                   |             | (0.96, 1.15)           |               |
| \(\hat{\alpha}_0, \hat{\alpha}_1\) | -                | (0.40, 1.16) | (1.60, 1.59)           |               |
| CV-Selected RAW (\(\lambda\)) | -                | -           | 0.24 (\(\lambda = 0.034\)) | 0.06 (\(\delta = 0.06\)) |
| Effective Sample Proportion (%) | 100              | 100         | 26.1                   | 9.39          |
| Sample Size n=10,000 |                  |             |                       |               |
| sNB              | 0.430             | 0.455       | 0.503                  | 0.500         |
| TPR              | 0.581             | 0.629       | 0.759                  | 0.750         |
| FPR              | 0.083             | 0.096       | 0.140                  | 0.137         |
|                  |                   |             | (0.86, 1.09)           |               |
|                  |                   |             | (0.96, 1.12)           |               |
| \(\hat{\alpha}_0, \hat{\alpha}_1\) | -                | (0.34, 1.12) | (1.46, 1.45)           |               |
| CV-Selected RAW (\(\lambda\)) | -                | -           | 0.18 (\(\lambda = 0.024\)) | 0.8 (\(\delta = 0.8\)) |
| Effective Sample Proportion (%) | 100              | 100         | 21.0                   | 12.4          |
| Sample Size n=5,000 |                  |             |                       |               |
| sNB              | 0.430             | 0.445       | 0.445                  | 0.445         |
| TPR              | 0.581             | 0.608       | 0.609                  | 0.609         |
| FPR              | 0.083             | 0.090       | 0.090                  | 0.090         |
|                  |                   |             | (0.57, 1.08)           |               |
| \(\hat{\alpha}_0, \hat{\alpha}_1\) | -                | (0.19, 1.07) | (0.19, 1.07)           |               |
| CV-Selected RAW (\(\lambda\)) | -                | -           | 0.99 (\(\lambda = 6.893\)) | 0.99 (\(\delta = 0.99\)) |
| Effective Sample Proportion (%) | 100              | 100         | 99.9                   | 99.0          |
| Sample Size n=1,000 |                  |             |                       |               |
| sNB              | 0.430             | 0.446       | 0.446                  | 0.446         |
| TPR              | 0.581             | 0.610       | 0.610                  | 0.610         |
| FPR              | 0.083             | 0.090       | 0.090                  | 0.090         |
|                  |                   |             | (0.36, 1.11)           |               |
| \(\hat{\alpha}_0, \hat{\alpha}_1\) | -                | (0.22, 1.10) | (0.22, 1.10)           |               |
| CV-Selected RAW (\(\lambda\)) | -                | -           | 0.99 (\(\lambda = 6.442\)) | 0.99 (\(\delta = 0.99\)) |
| Effective Sample Proportion (%) | 100              | 100         | 99.0                   | 99.0          |
| Sample Size n=500  |                  |             |                       |               |
| sNB              | 0.430             | 0.446       | 0.446                  | 0.446         |
| TPR              | 0.581             | 0.610       | 0.610                  | 0.610         |
| FPR              | 0.083             | 0.090       | 0.090                  | 0.090         |
|                  |                   |             | (0.36, 1.11)           |               |
| \(\hat{\alpha}_0, \hat{\alpha}_1\) | -                | (0.22, 1.10) | (0.22, 1.10)           |               |
| CV-Selected RAW (\(\lambda\)) | -                | -           | 0.99 (\(\lambda = 6.442\)) | 0.99 (\(\delta = 0.99\)) |
| Effective Sample Proportion (%) | 100              | 100         | 99.0                   | 99.0          |
Table 3: Performance of standard, weighted (using exponential form and step form weights), and constrained logistic recalibration in Example 2.

| Measure            | Orig. (No Recal.) | Std. Recal. | Wt. Recal. | Wt. Recal. | Const. Recal. |
|--------------------|-------------------|-------------|------------|------------|---------------|
|                    |                   |             | Exponential Wt. | Step Wt. |               |
| Sample Size n=10,000 |                  |             |         |             |               |
| $sNB$              | 0.475             | 0.511       | 0.521     | 0.518      | 0.525         |
| $TPR$              | 0.546             | 0.615       | 0.640     | 0.636      | 0.650         |
| $FPR$              | 0.052             | 0.077       | 0.088     | 0.086      | 0.092         |
| $(\hat{\alpha}_0, \hat{\alpha}_1)$ | -              | (0.75, 0.89) | (0.99, 0.96) | (0.826, 0.883) | (0.97, 0.90) |
| Prop. Assigned Intervention | 0.17        | 0.21        | 0.22      | 0.22       | 0.23          |
| CV-Selected RAW ($\lambda$) | -                | -           | 0.70 ($\lambda = 0.195$) | 0.29 ($\delta = 0.29$) | -             |
| Effective Sample Proportion (%) | 100          | 100         | 71.6      | 33.5       | 100           |
| Sample Size n=5,000 |                  |             |         |             |               |
| $sNB$              | 0.475             | 0.511       | 0.523     | 0.519      | 0.524         |
| $TPR$              | 0.546             | 0.619       | 0.644     | 0.637      | 0.647         |
| $FPR$              | 0.052             | 0.080       | 0.089     | 0.087      | 0.091         |
| $(\hat{\alpha}_0, \hat{\alpha}_1)$ | -              | (0.64, 0.82) | (0.86, 0.88) | (0.70, 0.81) | (0.75, 0.82) |
| Prop. Assigned Intervention | 0.17        | 0.21        | 0.22      | 0.22       | 0.22          |
| CV-Selected RAW ($\lambda$) | -                | -           | 0.74 ($\lambda = 0.214$) | 0.33 ($\delta = 0.33$) | -             |
| Effective Sample Proportion (%) | 100          | 100         | 75.2      | 36.7       | 100           |
| Sample Size n=1,000 |                  |             |         |             |               |
| $sNB$              | 0.475             | 0.491       | 0.491     | 0.491      | 0.491         |
| $TPR$              | 0.546             | 0.581       | 0.581     | 0.581      | 0.581         |
| $FPR$              | 0.052             | 0.066       | 0.066     | 0.066      | 0.066         |
| $(\hat{\alpha}_0, \hat{\alpha}_1)$ | -              | (0.51, 0.84) | (0.52, 0.84) | (0.52, 0.84) | (0.51, 0.83) |
| Prop. Assigned Intervention | 0.17        | 0.19        | 0.19      | 0.19       | 0.19          |
| CV-Selected RAW ($\lambda$) | -                | -           | 0.99 ($\lambda = 7.557$) | 0.99 ($\delta = 0.99$) | -             |
| Effective Sample Proportion (%) | 100          | 100         | 99.0      | 99.1       | 100           |
| Sample Size n=500  |                  |             |         |             |               |
| $sNB$              | 0.475             | 0.510       | 0.513     | 0.511      | 0.524         |
| $TPR$              | 0.546             | 0.616       | 0.624     | 0.617      | 0.647         |
| $FPR$              | 0.052             | 0.077       | 0.082     | 0.077      | 0.091         |
| $(\hat{\alpha}_0, \hat{\alpha}_1)$ | -              | (0.77, 0.90) | (0.86, 0.93) | (0.77, 0.90) | (0.99, 0.94) |
| Prop. Assigned Intervention | 0.17        | 0.21        | 0.21      | 0.21       | 0.22          |
| CV-Selected RAW ($\lambda$) | -                | -           | 0.90 ($\lambda = 0.775$) | 0.99 ($\delta = 0.99$) | -             |
| Effective Sample Proportion (%) | 100          | 100         | 90.5      | 99        | 100           |
Table 4: Performance of standard, weighted (using exponential form and step form weights), and constrained logistic recalibration in Example 3.

| Measure                | Orig. (No Recal.) | Std. Recal. | Wt. Recal. Exponential Wt. | Wt. Recal. Step Wt. | Const. Recal. |
|------------------------|-------------------|-------------|-----------------------------|---------------------|---------------|
| **Sample Size n=10,000** |                   |             |                             |                     |               |
| sNB                    | 0.440             | 0.449       | 0.502                       | 0.503               | 0.503         |
| TPR                    | 0.844             | 0.836       | 0.757                       | 0.751               | 0.751         |
| FPR                    | 0.282             | 0.271       | 0.173                       | 0.173               | 0.178         |
| (\(\hat{\alpha}_0, \hat{\alpha}_1\)) | -                 | (-0.31, 0.69) | (-0.69, 0.53) | (-0.663, 0.684) | (-0.66, 0.66) |
| Prop. Assigned Intervention | 0.41            | 0.40        | 0.32                        | 0.31                | 0.31          |
| CV-Selected RAW (\(\lambda/\delta\)) | - | - | 0.57 (\(\lambda = 0.096\)) | 0.11 (\(\delta = 0.11\)) | -            |
| Effective Sample Proportion (%) | 100             | 100         | 59.6                        | 17.4                | 100           |
| **Sample Size n=5,000** |                   |             |                             |                     |               |
| sNB                    | 0.440             | 0.458       | 0.484                       | 0.480               | 0.488         |
| TPR                    | 0.844             | 0.830       | 0.795                       | 0.801               | 0.789         |
| FPR                    | 0.282             | 0.260       | 0.217                       | 0.224               | 0.210         |
| (\(\hat{\alpha}_0, \hat{\alpha}_1\)) | -                 | (-0.30, 0.76) | (-0.54, 0.64) | (-0.45, 0.76) | (-0.49, 0.79) |
| Prop. Assigned Intervention | 0.41            | 0.39        | 0.35                        | 0.36                | 0.34          |
| CV-Selected RAW (\(\lambda/\delta\)) | - | - | 0.73 (\(\lambda = 0.200\)) | 0.22 (\(\delta = 0.22\)) | -            |
| Effective Sample Proportion (%) | 100             | 100         | 74.5                        | 27.3                | 100           |
| **Sample Size n=1,000** |                   |             |                             |                     |               |
| sNB                    | 0.440             | 0.440       | 0.457                       | 0.440               | 0.476         |
| TPR                    | 0.844             | 0.844       | 0.831                       | 0.844               | 0.808         |
| FPR                    | 0.282             | 0.282       | 0.202                       | 0.282               | 0.232         |
| (\(\hat{\alpha}_0, \hat{\alpha}_1\)) | -                 | (-0.36, 0.57) | (-0.46, 0.54) | (-0.36, 0.58) | (-0.52, 0.58) |
| Prop. Assigned Intervention | 0.41            | 0.41        | 0.39                        | 0.41                | 0.36          |
| CV-Selected RAW (\(\lambda/\delta\)) | - | - | 0.89 (\(\lambda = 0.526\)) | 0.99 (\(\delta = 0.99\)) | -            |
| Effective Sample Proportion (%) | 100             | 100         | 89.8                        | 99.1                | 100           |
| **Sample Size n=500**  |                   |             |                             |                     |               |
| sNB                    | 0.440             | 0.459       | 0.461                       | 0.460               | 0.475         |
| TPR                    | 0.844             | 0.828       | 0.827                       | 0.828               | 0.810         |
| FPR                    | 0.282             | 0.258       | 0.256                       | 0.257               | 0.232         |
| (\(\hat{\alpha}_0, \hat{\alpha}_1\)) | -                 | (-0.38, 0.66) | (-0.39, 0.65) | (-0.38, 0.66) | (-0.49, 0.63) |
| Prop. Assigned Intervention | 0.41            | 0.39        | 0.39                        | 0.39                | 0.36          |
| CV-Selected RAW (\(\lambda/\delta\)) | - | - | 0.99 (\(\lambda = 6.211\)) | 0.99 (\(\delta = 0.99\)) | -            |
| Effective Sample Proportion (%) | 100             | 100         | 99.1                        | 99.1                | 100           |
Table 5: Performance of standard, weighted (using exponential form and step form weights), and constrained logistic recalibration in Example 4.

| Measure                  | Orig. (No Recal.) | Std. Recal. | Wt. Recal. | Wt. Recal. | Const. Recal. |
|--------------------------|-------------------|-------------|------------|------------|--------------|
|                          |                   |             | Exponential Wt. | Step Wt.   |              |
| Sample Size n=10,000     |                   |             |             |            |              |
| $sNB$                    | 0.282             | 0.422       | 0.430      | 0.430      | 0.431        |
| $TPR$                    | 0.759             | 0.641       | 0.622      | 0.622      | 0.618        |
| $FPR$                    | 0.212             | 0.097       | 0.086      | 0.086      | 0.083        |
| ($\hat{\alpha}_0, \hat{\alpha}_1$) | -                 | (-0.90, 0.97) | (-0.98, 0.92) | (-0.989, 0.954) | (-1.01, 0.95) |
| Prop. Assigned Intervention | 0.30              | 0.18        | 0.17       | 0.17       | 0.17         |
| CV-Selected RAW ($\lambda$) | -                 | -           | 0.76 ($\lambda = 0.212$) | 0.38 ($\delta = 0.38$) | -         |
| Effective Sample Proportion (%) | 100               | 100         | 76.7       | 40.4       | 100          |
| Sample Size n=5,000      |                   |             |             |            |              |
| $sNB$                    | 0.282             | 0.413       | 0.421      | 0.425      | 0.435        |
| $TPR$                    | 0.759             | 0.652       | 0.641      | 0.638      | 0.605        |
| $FPR$                    | 0.212             | 0.106       | 0.098      | 0.095      | 0.076        |
| ($\hat{\alpha}_0, \hat{\alpha}_1$) | -                 | (-0.84, 0.99) | (-0.90, 0.95) | (-0.92, 0.97) | (-1.11, 1.11) |
| Prop. Assigned Intervention | 0.30              | 0.19        | 0.19       | 0.18       | 0.16         |
| CV-Selected RAW ($\lambda$) | -                 | -           | 0.86 ($\lambda = 0.413$) | 0.45 ($\delta = 0.45$) | -         |
| Effective Sample Proportion (%) | 100               | 100         | 86.4       | 47.1       | 100          |
| Sample Size n=1,000      |                   |             |             |            |              |
| $sNB$                    | 0.282             | 0.417       | 0.417      | 0.417      | 0.417        |
| $TPR$                    | 0.759             | 0.647       | 0.647      | 0.646      | 0.647        |
| $FPR$                    | 0.212             | 0.102       | 0.102      | 0.102      | 0.102        |
| ($\hat{\alpha}_0, \hat{\alpha}_1$) | -                 | (-0.87, 0.81) | (-0.87, 0.81) | (-0.87, 0.81) | (-0.87, 0.81) |
| Prop. Assigned Intervention | 0.30              | 0.19        | 0.19       | 0.19       | 0.19         |
| CV-Selected RAW ($\lambda$) | -                 | -           | 0.99 ($\lambda = 5.101$) | 0.99 ($\delta = 0.99$) | -         |
| Effective Sample Proportion (%) | 100               | 100         | 99.0       | 99.0       | 100          |
| Sample Size n=500        |                   |             |             |            |              |
| $sNB$                    | 0.282             | 0.417       | 0.418      | 0.418      | 0.417        |
| $TPR$                    | 0.759             | 0.646       | 0.645      | 0.646      | 0.646        |
| $FPR$                    | 0.212             | 0.102       | 0.101      | 0.102      | 0.102        |
| ($\hat{\alpha}_0, \hat{\alpha}_1$) | -                 | (-0.87, 0.97) | (-0.88, 0.97) | (-0.87, 0.97) | (-0.87, 0.97) |
| Prop. Assigned Intervention | 0.30              | 0.19        | 0.19       | 0.19       | 0.19         |
| CV-Selected RAW ($\lambda$) | -                 | -           | 0.99 ($\lambda = 6.414$) | 0.99 ($\delta = 0.99$) | -         |
| Effective Sample Proportion (%) | 100               | 100         | 99.0       | 99.0       | 100          |
Table 6: Summary of sNB obtained after recalibration across 500 repeated simulations. Mean (SD) of sNB of risk score after standard, weighted (with exponential and step weights), and constrained recalibration. Recalibration parameters are estimated in a training dataset with 1000 observations and sNB is evaluated in an independent dataset $10^6$ observations.

| Example   | Standard  | Weighted Exp. | Weighted Step | Constrained |
|-----------|-----------|---------------|---------------|-------------|
| Example 1 | 0.478 (0.009) | 0.479 (0.009) | 0.478 (0.009) | 0.484 (0.012) |
| Example 2 | 0.466 (0.016) | 0.476 (0.017) | 0.474 (0.018) | 0.484 (0.018) |
| Example 3 | 0.463 (0.015) | 0.472 (0.018) | 0.470 (0.018) | 0.482 (0.019) |
| Example 4 | 0.383 (0.021) | 0.392 (0.020) | 0.388 (0.021) | 0.401 (0.021) |
References

[1] Baker SG, Van Calster B, Steyerberg EW. Evaluating a new marker for risk prediction using the test tradeoff: an update. The International Journal of Biostatistics. 2012;8(1):1–37.

[2] Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press; 2003.

[3] Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2009;172(4):729–748.

[4] Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics New York, NY, USA; 2001.

[5] Gelman A, Rubin DB, et al. Inference from iterative simulation using multiple sequences. Statistical Science. 1992;7(4):457–472.