Broadening the Time Horizon: Adaptive Risk Scores for Time-to-Event Prediction

Chirag Nagpal[1], **Artur Dubrawski**[1], **Berk Ustun**[2]
[1]**Carnegie Mellon** [2]**UC San Diego**

Abstract

Risk scores are simple models that allow users to make quick risk predictions by adding and subtracting a few numbers. These models are widely used to predict the risk that an event will take place within a given time horizon – e.g., to predict the risk that a patient will suffer a stroke within 24 hours or will survive following a heart failure for at least 3 years. In practice, the models are designed to predict a target obtained by thresholding a *time-to-event* outcome, casting a survival analysis task into a classification task. In this work, we present a new method to fit risk scores for direct time-to-event prediction called TSLIM – *Time-Adaptive Sparse Linear Integer Models*. Our method trains models by solving a mixed-integer non-linear program that minimizes the proportional hazard loss and enforces constraints to enforce sparsity and integrality. Our approach can customize models to obey a wide range of constraints and inform the customization process by returning a certificate of optimality. We evaluate our models on real-world clinical datasets where we build time-adaptive risk scores for disease staging and compare them to standard methods for survival analysis and classification.

1 Introduction

Simple integer risk scores are widely used in modern medical applications – be it to estimate the risk that a patient will experience a stroke in critical care [47], will be re-admitted to a hospital after discharge [17], or will survive after experiencing heart failure [43][1].

The widespread adoption of risk scores in such clinical decision support tasks stems from their *format*, which facilitates scrutiny from non-experts and clinical integration. Models that let individuals make predictions by adding and subtracting a few small numbers are easy to use and easy to understand [26, 42], which allows clinicians to scrutinize the prediction logic and to make an informed decision as to whether to use it. Since models allow users to make predictions by checking a set of conditions, they can be integrated into existing clinical workflow – without extensive training [38] and using a range of variety of technological solutions (see e.g., the MDCalc mobile app).

A considerable number of risk scores that are currently deployed in such settings predict the probability that an event will occur within a fixed time period – e.g., the risk that a patient in critical care will experience a stroke in the next 24 hours [47], that a patient will be re-admitted to a hospital *within 1 month* [17], or that a patient with a history of heart failure will survive for over 3 years [43]. In practice, these models predict a binary outcome by *thresholding* a *time-to-event outcome* [2].

The common practice of binarizing time-to-event outcomes treats what is an inherently a *survival* prediction task as a classification. This choice of building a classifier rather than a survival prediction model comes at a cost:

Censoring Data for real-world clinical risk prediction involves patients who are often lost to follow-up, a phenomenon known as *censoring* [33]. In the case of remission prediction, patients who may not have a time associated with remission could have either not experienced remission or been lost to follow-up. Binarizing these times would treat the patients lost to follow-up as patients who did not experience remission, underestimating risk of remission [see e.g., 53, 1, for a discussion on how miscalibrated risk estimates can lead to harmful decisions in medicine]

Time-Adaptivity A binary classifier may have predictive power at the time-to-event horizon it was trained on, but may lose this discriminative ability at other time horizons. The resulting models are inherently tied to the threshold that was used to define the outcome. This is a risk as we can no longer evaluate models across broader time horizons and models may lose their discriminative ability over time. This functionality may be valuable in settings where we predict an event of interest over a longer time of patients from different age groups horizons [e.g., mortality within X years as in 43] – since patients in younger age groups are inherently more interested in survival at longer time horizons.

Contributions The main contributions of this work are:

- We propose a new risk score model for tasks involving time-to-event outcomes. Our models provide a single score

[1] For a comprehensive list of clinical risk scores across a wide range of applications refer to www.mdcalc.com

[2] For example, Struck et al. [47] train a risk score to predict $\Pr(y_i = 1)$ where $y_i = 1[t_i \leq 24 \text{ hours}]$ indicates if a patient will have a seizure within the next 24 hours. Here, y_i is the target and t_i is the time-to-event outcome

function that end-users can use to scrutinize their predictions. In contrast to existing models, it provides benefits of survival regression e.g., the ability to obtain risk estimates at multiple time horizons accounted for censoring.

- We develop an efficient method to learn these models involving solving a mixed integer non-linear program using a cutting-plane algorithm. It can handle constraints that help improve the usability of the risk-scoring system and output a certificate of optimality.
- We demonstrate through several real-world examples that our approach is able to recover patient strata with high discriminative capability as well as calibration. Further, we demonstrate that our approach has better performance than existing approaches that are limited to binary classification.

Related Work

Risk Scores Our work is related to a stream of work on methods to learn sparse linear classifiers with small integer coefficients [see e.g., 12, 24, 50, 18, 8, 52, 31, 55, 58, 37, 36, 40]. We focus on models that are designed to output calibrated risk estimates [see e.g., 52, 8, 36] rather than yes-or-no predictions [see e.g., 12, 46, 50, 58, 37].

This body of work broadly seeks to develop modern approaches to build scoring systems, decision aids, and risk scores. In effect, the vast majority of risk scores that are used in practice are developed by panels of experts [34, 22] or by combining logistic regression with heuristics for rounding and feature selection. For example, the TIMI Risk Score of Antman et al. [2] which screens features via hypothesis tests, fits a logistic regression model on the remaining features, and then obtains integer coefficients by scaling and rounding.

We study a salient class of prediction tasks that can be cast as classification or survival analysis problems. In practice, the vast majority of models are designed to solve classification tasks by thresholding time-to-event outcomes [see e.g., 47, 17, 56]. Discussions surrounding model development ignore the potential effects of censoring or the possibility of more accurate predictions at other time horizons of interest. Decisions on thresholding stem from convention rather than validation. For example, the 30-day threshold that is commonly used to predict the risk of hospital readmissions [17] in the United States corresponds to a value chosen by the Hospital Readmissions Reduction Program enacted under the Patient Protection and Affordable Care Act [13].

Optimization We train our models by solving a mixed-integer non-linear programming formulation that fits a single model to assign calibrated risk estimates across multiple time horizons and address the bias from censoring. We solve these instances using a variant of the cutting-plane algorithms used by [7, 52] which relies on the use of a mixed-integer programming solver with callback functionality such as CPLEX [29], Gurobi [19], CBC [19].

Our work is broadly related to a stream of machine learning methods that solve discrete optimization problems (e.g. mixed-integer programs, and mixed-integer non-linear programs) with an optimization solver. Even as these are computationally intractable optimization problems in the worst case, we can now solve large instances using off-the-shelf solvers [see e.g., 23]. The viability of this approach stems from improvements in hardware, software, and research over the last three decades [see e.g., 9], prompting the development of practical methods for supervised learning [6, 12, 5, 44, 45, 4, 7, 27, 28].

Transparency & Fairness We propose an approach to train and evaluate a single model that individuals can use to obtain risk estimates over their desired time horizon. The resulting approach highlights an alternative strategy to personalize models [3], reflects some of the motivation for a stream of work on preference-based-fairness [57, 49, 32, 54, 16] and the need for participatory paradigms for prediction [30].

Our work highlights how routine decisions in problem specification can promote safety and transparency through model development and evaluation [41]. By training a model across multiple time horizons, we consider performance across time horizons that may be of interest to various subpopulations and use them to guide decisions such as feature selection [see 48]. This decision ensures that we resulting model can be evaluated across multiple time horizons to ensure that a model will not underperform at longer time horizons – and this effect in itself may be attenuated for minority subpopulations. In our setting, such effects can be evaluated through a disaggregated evaluation [see 11].

2 Problem Statement

In this section, we present **TSLIM** an integer scoring system for censored time-to-event outcomes.

2.1 Preliminaries

We consider a standard survival analysis task to predict the risk of an event at a time occurs from a set of features X.

We start with a dataset of n examples $\{(\boldsymbol{x}_i, t_i, \Delta_i)\}_{i=1}^n$. Each example i consists of features $\boldsymbol{x}_i = [x_{i,1}, \ldots, x_{i,d}] \in \mathbb{R}^d$, a *time-to-event* outcome $t_i \in \mathbb{R}_+$, and a censoring indicator $\Delta_i := 1[t_i \text{ is not censored}]$.

Given the dataset, our goal is to estimate a *hazard rate* function for the event at time t conditioned on the covariates, $\boldsymbol{x}$. Here, the hazard rate is a function that describes the instantaneous rate at which an event occurs for an individual at a certain time.. More formally the hazard rate can be described as:

$$\boldsymbol{\lambda}(t) := \lim_{\Delta t \to 0} \frac{\Pr(t < T \le t + \Delta t \mid T > t)}{\Delta t}.$$

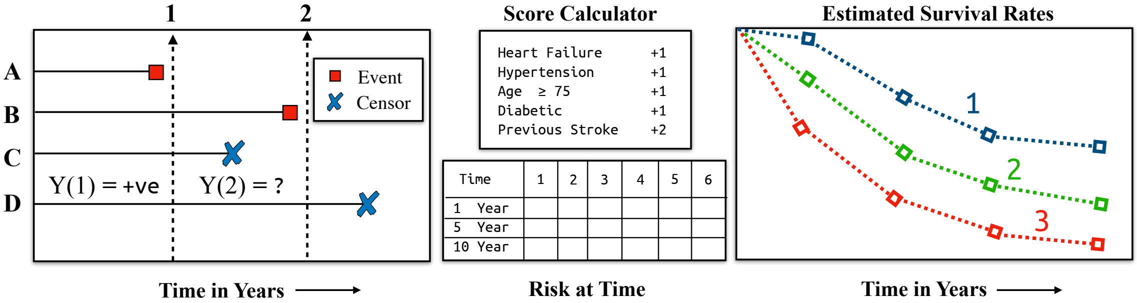
Reasoning in terms of the hazard rate, $\boldsymbol{\lambda}(t)$ is natural in survival analysis and reliability engineering. The hazard function can be used to estimate other quantities of interest in time-to-event prediction such as the survival rates.[3]

2.2 Model Form

We assume that the conditional hazard rate function follows a *proportional hazards* (PH) model:

$$\boldsymbol{\lambda}(t \mid X = \boldsymbol{x}) := \boldsymbol{\lambda}_0(t) \exp \frac{\boldsymbol{w}^\top \boldsymbol{x}}{c} \tag{1}$$

[3] The survival rate is the negative exponent of the cumulative hazard, ie. $\boldsymbol{S}(t) = \exp\big(-\int_0^t \boldsymbol{\lambda}(t)\big)$.



Heart Failure	+1
Hypertension	+1
Age ≥ 75	+1
Diabetic	+1
Previous Stroke	+2

Time	1	2	3	4	5	6
1 Year						
5 Year						
10 Year						

Figure 1: **a) Censoring and Time-to-Event Predictions**. **Patient C** was lost to follow-up after the first year. When estimating 1-year risk, we treat **C** as a positive (survived) sample. The survival status for **C** at 2 years was is ambigous, posing a challenge to traditional risk scoring models, a phenomenon known as *censoring*. **b) Illustrative example of a scoring system produced by TSLIM:** Our proposed approach allows estimation of risk and the corresponding survival curves of the patient **across time**, helping clinicians have a better estimate of the patient risk profile.

This model assumes that the ratio between the hazard rate for a point with features $\boldsymbol{x}$ at time t changes with respect to the *baseline hazard rate* $\boldsymbol{\lambda}_0(t)$[4] and that the rate of of change is determined by the score function $\frac{\boldsymbol{w}^\top \boldsymbol{x}}{c}$. In practice our goal is to estimate $\boldsymbol{w}$, a sparse set of integer coefficients that determine the estimated score (or risk level) for a patient with a certain set of features $\boldsymbol{x}$ through a affine function.

$\boldsymbol{c} > 0$ is a constant scaling factor that is set by a practitioner based on the problem statement and domain expertise. It determines the difference in terms of hazard ratio between the various levels of patients stratified by our risk score model. Smaller values of $\boldsymbol{c}$ would lead to the model recovering granular stratification while larger values would lead to fewer recovered risk levels.

2.3 Estimating the Conditional Survival Function

Once the parameters of the model in $\boldsymbol{w}$ are learned using the cutting plane formulation of MINLP formulation (Equation 7), we estimate individualized survival at a time horizon t, $\widehat{\mathbb{P}}(T > t | X = \boldsymbol{x}, \boldsymbol{w})$ using a non-parametric maximum likelihood estimation procedure, more commonly known as the Breslow's Estimator [10, 35].

$$\widehat{\mathbb{P}}(T > t | X = \boldsymbol{x}, \boldsymbol{w}) = \exp\big(-\widehat{\boldsymbol{\Lambda}}_0(t)\big)^{\exp(\frac{\boldsymbol{w}^\top \boldsymbol{x}}{c})}$$

$$\text{and,} \quad \widehat{\boldsymbol{\Lambda}}_0(t) = \sum_{t_i < t} \frac{1}{\sum_{j \in \mathcal{R}(t_i)} \exp\big(\frac{\boldsymbol{w}^\top \boldsymbol{x}_j}{c}\big)} \tag{2}$$

Here, $\widehat{\boldsymbol{\Lambda}}_0(t)$ is the estimated baseline cumulative hazard.

2.4 Evaluating Performance

Our goal is to train a risk score that performs well in terms of the following measures:

Brier Score: : The Brier Score is the Mean Squared Error of the binary forecast of survival at a certain time horizon of interest. As a proper scoring rule, the Brier Score gives a sense of both discriminative performance and calibration.

$$\mathrm{BS}(t) = \mathbb{E}_{\mathcal{D}}\big[\big(\mathbb{1}\{T > t\} - \widehat{\mathbb{P}}(T > t | X)\big)^2\big]$$

Expected ℓ_1 Calibration Error (ECE): The ECE measures the average absolute difference between the observed and expected (according to the risk score) event rates, conditional on the estimated risk score. At time t, let the predicted risk score be $s(t) = \widehat{\mathbb{P}}(T > t | X)$. Then, the ECE is:

$$\mathrm{ECE}(t) = \mathbb{E}\big[\big|\mathbb{P}(T > t | s(t)) - s(t)\big|\big]$$

by conditioning on the set of all possible estimated integer risk scores $\{\boldsymbol{w}^\top \boldsymbol{x}_i \, : \, i \in [n]\}$.

Area under ROC Curve (AUC): Treating the survival analysis problem as binary classification at different horizons of event times and computing the corresponding area under the curve.

The metrics described above are adjusted for censoring by standard Thompson-Horvitz style Inverse Propensity of Censoring Weights (IPCW) estimates learnt with a Kaplan-Meier estimator over the censoring times (Appendix A).

3 Methodology

We determine the values of the coefficients by solving the following mixed integer nonlinear program (MINLP):

Definition 1 (Hazard Scoring Problem)**.** *The hazard scoring problem is a discrete optimization problem of the form:*

$$\min_{\boldsymbol{w}} \;\; \mathcal{PL}(\mathcal{D}; \boldsymbol{w}, c) \;\; \text{s.t.} \;\; \boldsymbol{w} \in \mathcal{W} \;\; \textit{and} \;\; \|\boldsymbol{w}\|_0 \leq R^{\max} \tag{3}$$

- Here, the is the partial likelihood, $\mathcal{PL}(\mathcal{D}; \boldsymbol{w}, c)$ $= \sum_{i=1}^{n} \mathbf{1}_{\Delta_i \neq 0} \Big(\frac{\boldsymbol{w}^\top \boldsymbol{x}_i}{c} - \log \sum_{j \in \mathcal{R}(t_i)} \exp\big(\frac{\boldsymbol{w}^\top \boldsymbol{x}_j}{c}\big) \Big)$
- $\|\boldsymbol{w}\|_0 = \sum_{j=1}^{d} \mathbf{1}\{w_j \neq 0\}$ is the ℓ_0-seminorm;
- $\mathcal{W} \subset \mathbb{Z}^{d+1}$ is a user-specified coefficient set, e.g. $\mathcal{W} = \{-5, 5\}^{d+1}$;
- $R^{\max} \in \mathbb{Z}_+$ is a user-specified limit on model size.

[4] The base hazard rate $\boldsymbol{\lambda}$ is an infinite dimensional functional parameter of the model, which is estimated non-parametrically.

This problem captures what we desire in a scoring system. The objective minimizes the *partial likelihood* over the event rate to ensure models that are well-calibrated and have good discriminative performance. Further it penalizes the ℓ_0-seminorm (the count of non-zero coefficients) for sparsity. The constraints restrict coefficients to a set of small integers such as $\mathcal{W} := \{-5, \ldots, 5\}^{d+1}$, and may be customized to encode other model requirements such as those in Table 1.

Model Requirement	**Example**
Feature Selection	Choose between **5** to **10** total features
Group Sparsity	Include **either** `male` **or** `female` but **not** both
Optimal Thresholding	Use at most 3 thresholds for a set of variables: $\sum_{k=1}^{100} 1\{\texttt{age} \leq k\} \leq 3$
Logical Structure	If `male` is **in** model, include `glucose` **or** `bmi` ≥ 30
Side Information	`stage`$\geq$ 5 **if** `male=+ve` **&** `diabetes=+ve`

Table 1: Model requirements that can be addressed by adding operational constraints

As is standard in survival analysis, we train the model to optimize the likelihood function:

$$\mathcal{L}(\mathcal{D}) \propto \prod_{i=1}^{n} \boldsymbol{\lambda}(t|X = x_i)^{\Delta_i} \boldsymbol{S}(t|X = x_i)$$

Here, $\boldsymbol{S}(t)$ is the *survival function*, Note that under the assumptions of PH in Equation 1, in practice, the model is learned by minimizing the *partial likelihood* function

$$\mathcal{PL}(\mathcal{D}; \boldsymbol{w}, c) := \sum_{i:\Delta_i \neq 0} \frac{\boldsymbol{w}^\top \boldsymbol{x}_i}{c} - \log \sum_{j \in \mathcal{R}(t_i)} \exp\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}_j}{c}\right)$$

which is independent of the baseline hazard rate, $\boldsymbol{\lambda}(\cdot)$. Here, $\mathcal{R}(t)$ is the *'risk set'* $\{i \in [n] : t_i > t\}$ – i.e., the indices of all points that have survived until time t.

3.1 Cutting-Plane Algorithm

We recover an optimal solution to the MINLP in Equation 5 with the lattice cutting-plane algorithm [52]. The cutting-plane algorithm solves a surrogate problem that replaces the loss function with a linear approximation composed of *cutting-planes*. This resulting problem can be expressed as a *mixed integer linear program* (MILP) rather than an MINLP (Equation 5), and has the following form:

$$\begin{aligned}
\min_{\boldsymbol{w}} \quad & L + \epsilon R \\
\text{s.t.} \quad & L \geq \mathcal{PL}(\boldsymbol{w}^t) + \nabla\mathcal{PL}(\boldsymbol{w}^k)(\boldsymbol{w} - \boldsymbol{w}^k) \quad k \in [K] && \textit{(loss cuts)} \\
& R = \sum_{j \in [d]} \alpha_j && \textit{(model size)} \\
& W_j^{\max} \alpha_j \geq w_j \quad j \in [d] && (w_j > 0 \implies \alpha_j = 1) \\
& -W_j^{\min} \alpha_j \geq -w_j \quad j \in [d] && (w_j < 0 \implies \alpha_j = 1) \\
& L \in [0, \ldots, L^{\max}] && \textit{(loss)} \\
& R \in \{0, \ldots, R^{\max}\} && \textit{(model size)} \\
& w_j \in \{W_j^{\min} \ldots, W_j^{\max}\} \quad j \in [d] && \textit{(coef for variable } j\textit{)} \\
& \alpha_j \in \{0, 1\} \quad j \in [d] && (\alpha_j := 1[w_j \neq 0]) \quad (4)
\end{aligned}$$

- L and R are '*auxiliary*' variables that represent the overall loss and the model size, respectively. In theory, these are redundant in that they could be replaced by a single quantity. In practice, we include them explicitly as they allow us to set bounds on feasible models via variable definitions.
- The parameter ϵ trade-offs between these competing objectives, and represents the maximum log-likelihood sacrificed to remove a feature from the optimal model.
- The formulation accounts for model size using the indicator variables $\alpha_j := 1[w_j \neq 0]$. These variables are set to 1 whenever $w_j \neq 0$ through the constraints on $\boldsymbol{\alpha}$.
- The coefficient for each variable is constrained to small integer values in the constraints. These constraints restrict each $\boldsymbol{w}_j$ to integers from $W_j^{\min}$ to $W_j^{\max}$. By default, we set these values to $W_j^{\min} = -5$ and $W_j^{\max} = +5$.

The main difference between the MILP formulation in (7g) and the MINLP formulation in (5) is that we compute the loss using a cutting-plane approximation of the loss function. The cutting-plane approximation is captured through K cuts. Each cut is a supporting hyperplane to the loss function at a specific point $\boldsymbol{w}^k$ – where the values of $\boldsymbol{w}^k$ represent integer-feasible solutions. Since we work with the partial likelihood (a convex loss function), the cutting-plane approximation is an under-approximation. For a more comprehensive discussion refer to Appendix C.

Algorithm 1 Cutting Plane Algorithm for TSLIM

```
Input : training data $(\boldsymbol{x}_i, t_i, \Delta_i)_{i=1}^n$; coefficient set $\mathcal{W}$
        model size $R^{\max}$;
$k \leftarrow 0$                                                   (iteration counter)
$\hat{l}^0(\boldsymbol{w}) \leftarrow \{0\}$                       (approximate loss with an empty set)
$V^{\min} \leftarrow \min_{\boldsymbol{w} \in \mathbb{R}^d} \mathcal{PL}(\boldsymbol{w}, \mathcal{D})$   (set lower bound by solving for reals)
$\varepsilon \leftarrow \infty$                                     (initialize optimality gap)
$\boldsymbol{w} \leftarrow \{0\}$                                   (initialize solution set)
while $\varepsilon > \varepsilon^{stop}$ do
    $(L^k, \boldsymbol{w}^k) \leftarrow$ provably optimal solution to TSLIM
    compute cut parameters $l(\boldsymbol{w}^k)$ and $\nabla l(\boldsymbol{w}^k)$
    $\hat{l}^{k+1}(\boldsymbol{w}) \leftarrow \max\{\hat{l}^k(\boldsymbol{w}), l(\boldsymbol{w}^k) + \langle \nabla l(\boldsymbol{w}^k), \boldsymbol{w} - \boldsymbol{w}^k \rangle\}$
                                                                    (update approximate loss)
    $V^{\min} \leftarrow L^k + C_0 \left\|\boldsymbol{w}^k\right\|_0$   (optimal value of TSLIM is lower bound)
    if $V(\boldsymbol{w}^k) < V^{max}$ then
        $V^{\max} \leftarrow V(\boldsymbol{w}^k)$                   (update upper bound)
        $\boldsymbol{w}^{best} \leftarrow \boldsymbol{w}^k$          (update incumbent)
    end
    $\varepsilon \leftarrow 1 - V^{\min}/V^{\max}$                  (update optimality gap)
    $k \leftarrow k + 1$                                            (increment counter)
end
Return: $\boldsymbol{w}^{\text{best}}$
```

Algorithm 1 presents the cutting plane approach to recover a provably optimal solution to the MINLP in Equation 5. TSLIM is paired with an optimality gap, $\boldsymbol{\varepsilon}$. In practice, a small optimality gap suggests that we have trained the best possible risk score that satisfies a specific set of constraints. If a risk score with a small optimality gap doesn't generalize, then one can attribute the performance deficit of the model to overly restrictive constraints and improve performance by relaxing them.

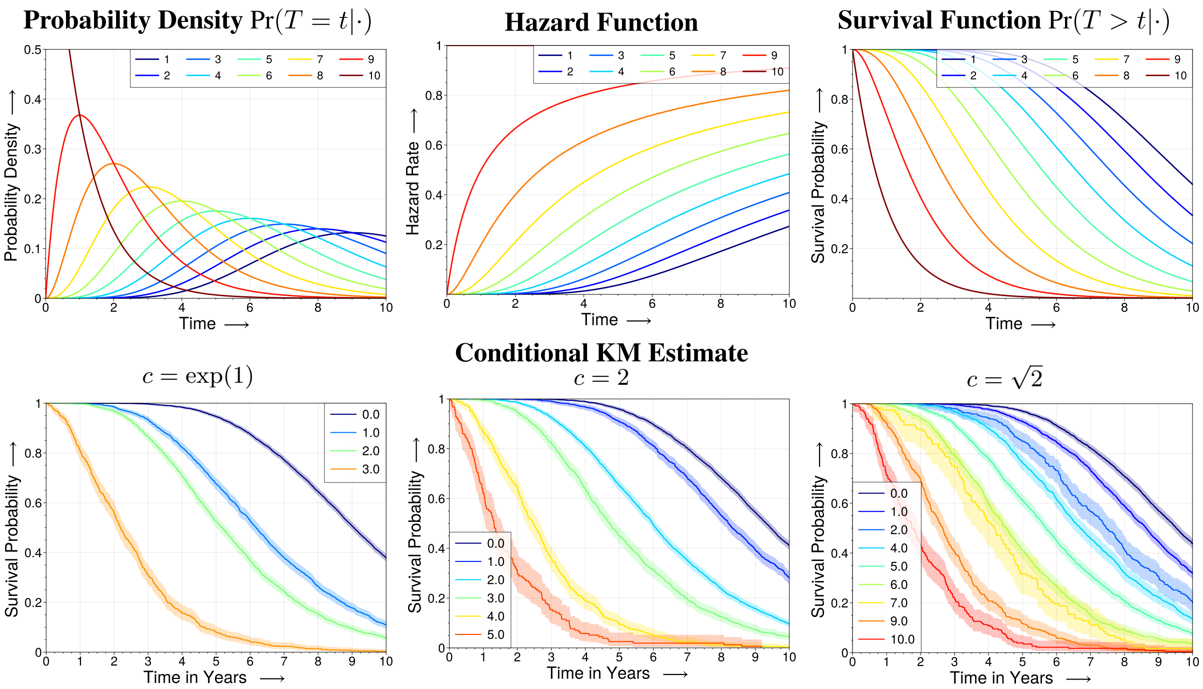


Figure 2: **Top**: The probability density, hazard and survival functions for the **synthetic** data stratified by the true risk score. **Bottom**: Kaplan-Meier survival curves recovered by **TSLIM** on the **synthetic** data for scaling factors $\boldsymbol{c} \in \{\sqrt{2}, 2, e\}$.

synthetic

	Brier Score			**Area Under ROC Curve**			**ECE**			**OPT**
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.010	0.186	0.637	0.980	0.841	0.699	0.010	0.229	0.751	-
CR	0.012	0.118	0.214	0.983	0.856	0.705	0.014	0.103	0.154	-
LR-L1-6	0.010	0.187	0.637	0.981	0.840	0.698	0.009	0.229	0.752	-
CR-L1-6	0.012	0.120	0.212	0.983	0.855	0.704	0.014	0.109	0.153	-
RiskSLIM	0.010	0.186	0.637	0.977	0.835	0.696	0.009	0.229	0.751	0.0%
TSLIM	0.012	0.118	0.214	0.982	0.856	0.701	0.014	0.101	0.159	0.0%

Table 2: Discriminative performance and Calibration of TSLIM vs. RiskSLIM on the **synthetic** data.

Dataset	Description	n	d	5-Yr Survival
flchain	effect of light chains on survival	6,524	39	86.54%
support	survival post ICU discharge	9,105	90	24.55%
seer-lymphoma	lymphoma/leukemia survival	60,486	55	54.16%

Table 3: Datasets used in Section 4. n and d denote the number of examples and features in each dataset, respectively. All datasets are publicly available.

4 Experiments

Synthetic Data We compare the performance of TSLIM on a synthetic dataset we generate to simualte a censored time-to-event study. The generative process captures what we seek in a real world clinical risk prediction task. We sample binary covariates $\boldsymbol{x}_{1:4}$ that determine the true Time-to-Event t_i^* along with noisy covariates $\boldsymbol{x}_{4:8}$ that are exogenous to the time-to-event outcome. Further, we randomly censor 75% of the population with a censoring time that is drawn uniformly between 0 and t_i^*. The final dataset consists of the observed time-to-event t_i, the covariates $\boldsymbol{x}_i$ and the censoring indicators Δ_i. Our generative process for the data is

$$\begin{aligned} \boldsymbol{x}_{1:8} &\sim \text{Bernoulli}(1/4), & s &:= \boldsymbol{x}_{1:4}^\top [1\ 1\ 3\ 4] \\ t^* &\sim \text{Gamma}(1/s), & c_i &\sim \text{Uniform}(0, t_i) \\ \Delta &\sim \text{Bernoulli}(3/4), & t &= \Delta \cdot c + (1 - \Delta) \cdot t^* \end{aligned}$$

Figure 2 presents the true event distributions for the synthetic data.

flchain

	Brier Score			Area Under ROC Curve			ECE			OPT
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.033	0.111	0.219	0.833	0.817	0.807	0.016	0.097	0.222	-
LR-L1-6	0.034	0.117	0.228	0.823	0.806	0.804	0.021	0.095	0.222	-
Cox	0.033	0.090	0.129	0.836	0.822	0.839	0.020	0.027	0.047	-
Cox-L1-6	0.034	0.098	0.142	0.817	0.811	0.824	0.030	0.068	0.097	-
RiskSLIM	0.034	0.113	0.223	0.815	0.790	0.778	0.010	0.096	0.223	2.6%
TSLIM	0.034	0.091	0.131	0.830	0.810	0.829	0.009	0.014	0.027	0.0%

support

	Brier Score			Area Under ROC Curve			ECE			OPT
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.224	0.227	0.253	0.684	0.696	0.721	0.046	0.091	0.292	-
LR-L1-6	0.241	0.244	0.257	0.637	0.663	0.698	0.088	0.121	0.287	-
Cox	0.226	0.219	0.170	0.676	0.696	0.730	0.040	0.035	0.064	-
Cox-L1-6	0.242	0.235	0.183	0.617	0.649	0.696	0.055	0.073	0.069	-
RiskSLIM	0.229	0.236	0.257	0.647	0.644	0.670	0.019	0.088	0.290	14.3%
TSLIM	0.232	0.225	0.173	0.643	0.668	0.707	0.019	0.016	0.028	0.8%

seer-lymphoma

	Brier Score			Area Under ROC Curve			ECE			OPT
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.138	0.213	0.262	0.808	0.798	0.778	0.022	0.168	0.249	-
LR-L1-6	0.167	0.253	0.294	0.707	0.731	0.753	0.076	0.194	0.261	-
Cox	0.143	0.174	0.177	0.787	0.808	0.813	0.026	0.028	0.023	-
Cox-L1-6	0.170	0.220	0.221	0.678	0.711	0.745	0.067	0.104	0.106	-
RiskSLIM	0.148	0.226	0.273	0.760	0.749	0.728	0.017	0.166	0.248	0.0%
TSLIM	0.151	0.190	0.194	0.746	0.768	0.772	0.034	0.023	0.022	0.0%

Table 4: Discriminative performance and Calibration of TSLIM vs. RiskSLIM on the **flchain**, **support** and **seer-lymphoma** data. All metrics are reported on the held-out test set and adjusted for censoring using Inverse Propensity of Censoring.

Real-World Clinical Studies We compare and evaluate the on real-world clinical prediction tasks. Each dataset includes demographic information such as sex, age, as well as clinical variables specific to the study or the results of a medical procedure. Table 3 presents summary statistics of the real-world datasets employed in the experiments.

flchain (Assay of Serum Free Light Chain): This is a public dataset introduced by [15] aiming to study the relationship between serum free light chain and mortality. It includes covariates like age, gender, serum creatinine and presence of monoclonal gammapothy. We removed all the individuals with missing covariates and experiment with the remaining subset of 6,524 individuals.

support The support dataset is derived from a study of the survival risk of critically-ill patients who were discharged from the ICU conducted by Connors et al. [14]. Here, we have records of 9,105 patients. The outcome variable indicates that a patient has died within six months of discharge. The features cover chronic health conditions (e.g., diabetic status, number of comorbidities), vital signs (e.g., mean blood pressure), and results of lab tests (e.g., white blood cell count). The dataset is publicly available[5].

seer-lymphoma (Surveillance, Epidemiology and End Results Study)[6] : We consider a cohort of 60,486 patients who were diagnosed with lymphoma or leukemia cancer between 2000-2004 and monitored as part of the National Cancer Institute SEER study [39]. Here, the outcome variable indicates if a patient dies within five years from any cause, and 45.83% of patients die within the first five years from diagnosis. The cohort includes patients from New Jersey, Greater California, Kentucky, Lousisiana and Georgia. The features reflect the morphology and histology of the tumor (e.g., size, metastasis, stage, node count and location, number and location of notes) as well as interventions that were administered at the time of diagnosis (e.g., surgery, chemo, radiology).

Methods We use each dataset to train a risk score with 70% of the data as the training set and test the performance of the learnt scoring system on the remaining 30% held out set. The set of possible model coefficients $\mathcal{W}$ are restricted to be between $\{-5, ..., 5\}$ and we fix the maximum size of the

[5] https://hbiostat.org/data/

[6] https://seer.cancer.gov/

flchain

Condition	Points
STUDY_GROUP=10	2
AGE<56.0	-2
AGE<63.5	-2
AGE<73.0	-2
AGE>80.0	3
CREATININE>1.4	2
SCORE	...

support

Condition	Points
CREATININE<2.2	-1
AGE>60	1
COMORBIDITIES>=1	1
CANCER	1
CANCER (METASTIZED)	1
COMATOSE	3
SCORE	...

seer-lymphoma

Condition	Points
PRIMARY SITE: BONE MARROW	2
HISTOLOGY: ACUTE MYELOID LEUKEMIA	3
HISTOLOGY: B-CELL LEUKEMIA	-3
MALIGNANCY COUNT<=1	2
AGE<55	-2
AGE<70	-1
SCORE	...

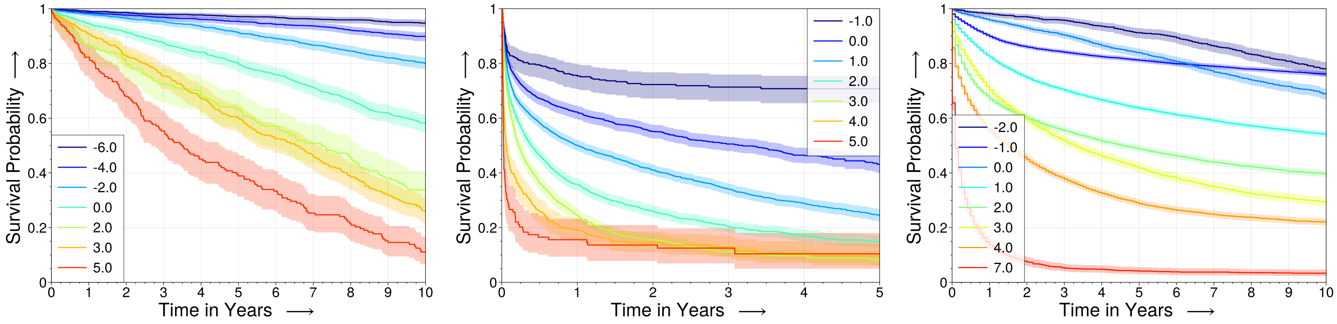


Figure 3: **TSLIM** risk scores on all clinical risk prediction tasks. along with the Kaplan-Meier curves that show risk estimates across time. As shown, TSLIM recovers adaptive risk scores that stratify patients reliably across multiple horizons.

model R^{max} to 6 and scaling factor $\boldsymbol{c}$ to $\sqrt{2}$. We compare the performance of our proposed TSLIM to RiskSLIM [51, 52] involving learning of integer scoring systems with binary outcomes. For completeness we also report performance of an unconstrained Logistic Regression (LR) and a Cox Regression (Cox) on the entire dataset as well as L1 regularized Logistic (LR-L1-6) and Cox Regression (Cox-L1-6) that select at most R^{max} features using the `glmnet` [20] package.[7]

For RiskSLIM, our horizon of a positive outcome is determined using best judgment for each dataset. For **flchain** and **seer-lymphoma** we consider an outcome to be positive if a patient survived the first 1 year from entry into the study. For **support** we consider survival post the first 6-months from discharge as a positive outcome. We found that for all three datasets we were able to solve the TSLIM problem close optimality ($< 1\%$) within a running time of < 10 minutes with 8 parallel threads using the IBM CPLEX solver, while RiskSLIM took longer to converge.

Further we demonstrate the superiority of TSLIM in the presence of higher amounts of censoring we also experiment by synthetically augmenting the amount of censoring in the above datasets by randomly sampling a certain percentage of the uncensored individuals and censoring their event times drawn from a uniform distribution in Appendix B.

Results In this section we describe the results of the proposed TSLIM approach vs RiskSLIM in terms of both Calibratation and Discriminative performance. Table 2 summarizes performance of TSLIM on the **synthetic** dataset. TSLIM had better discriminative performance than RiskSLIM at all horizons of time and was well calibrated at all time horizons.

TSLIM consistently had better discrimination performance as evidenced from the higher area under ROC scores at different horizons of time (Table 4). Further While RiskSLIM was calibrated at the horizon it was trained on, calibration deteriorated significantly at longer time horizons. For completeness, we also present the sparse integer scoring system outputs from TSLIM and RiskSLIM in Figure 3 as well as the corresponding survival curves stratified by the estimated risk score by TSLIM.

In order to better present the qualitative differences between various different bases for the risk scoring models, we also experiment with different values of the scaling base $\boldsymbol{c}$ coefficient (Figure 2) and find that smaller values lead to better stratification with more granular scoring stages, however this comes at a cost of calibration.

5 Concluding Remarks

We proposed **TSLIM** an integer risk scoring method that allows learning highly interpretable scoring systems involving censored time-to-event outcomes in a data-driven manner. Our formulation involves a mixed integer program and allows for the specification of several operational constraints helping improve the utility of the learnt scoring systems. We benchmark the performance of TSLIM to existing solutions and found that across multiple real world risk estimation studies, TSLIM recovered highly calibrated risk scoring systems with improved discriminative power.

[7] Note that these models recover coefficients that are reals and thus involve a much more complex hypothesis class than recovered from RiskSLIM and TSLIM.

References

[1] Alba, A. C.; Agoritsas, T.; Walsh, M.; Hanna, S.; Iorio, A.; Devereaux, P.; McGinn, T.; and Guyatt, G. 2017. Discrimination and calibration of clinical prediction models: Users' guides to the medical literature. *Journal of the American Medical Association*, 318(14): 1377–1384.

[2] Antman, E. M.; Cohen, M.; Bernink, P. J.; McCabe, C. H.; Horacek, T.; Papuchis, G.; Mautner, B.; Corbalan, R.; Radley, D.; and Braunwald, E. 2000. The TIMI risk score for unstable angina/non–ST elevation MI. *The Journal of the American Medical Association*, 284(7): 835–842.

[3] Awad, N. F.; and Krishnan, M. S. 2006. The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS quarterly*, 13–28.

[4] Bertsimas, D.; Dunn, J.; Gibson, E.; and Orfanoudaki, A. 2022. Optimal survival trees. *Machine Learning*, 111(8): 2951–3023.

[5] Bertsimas, D.; and King, A. 2017. Logistic regression: From art to science. *Statistical Science*, 367–384.

[6] Bertsimas, D.; King, A.; and Mazumder, R. 2015. Best subset selection via a modern optimization lens. *arXiv preprint arXiv:1507.03133*.

[7] Bertsimas, D.; King, A.; Mazumder, R.; et al. 2016. Best subset selection via a modern optimization lens. *Annals of statistics*, 44(2): 813–852.

[8] Billiet, L.; Van Huffel, S.; and Van Belle, V. 2018. Interval Coded Scoring: a toolbox for interpretable scoring systems. *PeerJ Computer Science*, 4: e150.

[9] Bixby, R.; and Rothberg, E. 2007. Progress in computational mixed integer programming: a look back from the other side of the tipping point. *Annals of Operations Research*, 149(1): 37–41.

[10] Breslow, N. E. 1972. Contribution to discussion of paper by DR Cox. *J. Roy. Statist. Soc., Ser. B*, 34: 216–217.

[11] Bynum, L.; Loftus, J.; and Stoyanovich, J. 2021. Disaggregated interventions to reduce inequality. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–13.

[12] Carrizosa, E.; Nogales-Gómez, A.; and Morales, D. R. 2016. Strongly agree or strongly disagree?: Rating features in support vector machines. *Information Sciences*, 329: 256–273.

[13] Centers for Medicare & Medicaid Services, U.; et al. 2019. Hospital readmissions reduction program (HRRP).

[14] Connors, A. F.; Dawson, N. V.; Desbiens, N. A.; Fulkerson, W. J.; Goldman, L.; Knaus, W. A.; Lynn, J.; Oye, R. K.; Bergner, M.; Damiano, A.; et al. 1995. A controlled trial to improve care for seriously iII hospitalized patients: The study to understand prognoses and preferences for outcomes and risks of treatments (SUPPORT). *Jama*, 274(20): 1591–1598.

[15] Dispenzieri, A.; Katzmann, J. A.; Kyle, R. A.; Larson, D. R.; Therneau, T. M.; Colby, C. L.; Clark, R. J.; Mead, G. P.; Kumar, S.; Melton III, L. J.; et al. 2012. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, volume 87, 517–523. Elsevier.

[16] Do, V.; Corbett-Davies, S.; Atif, J.; and Usunier, N. 2021. Online certification of preference-based fairness for personalized recommender systems. *arXiv preprint arXiv:2104.14527*.

[17] Donzé, J.; Aujesky, D.; Williams, D.; and Schnipper, J. L. 2013. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA internal medicine*, 173(8): 632–638.

[18] Ertekin, Ş.; and Rudin, C. 2015. A Bayesian approach to learning scoring systems. *Big Data*, 3(4): 267–276.

[19] Forrest, J. F.; and Ralphs, T. 2017. COIN Branch and Cut. https://projects.coin-or.org/Cbc.

[20] Friedman, J. 2009. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4).

[21] Gerds, T. A.; and Schumacher, M. 2006. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6): 1029–1040.

[22] Gillespie, B. M.; and Marshall, A. 2015. Implementation of safety checklists in surgery: a realist synthesis of evidence. *Implementation Science*, 10(1): 137.

[23] Gleixner, A.; Hendel, G.; Gamrath, G.; Achterberg, T.; Bastubbe, M.; Berthold, T.; Christophel, P.; Jarck, K.; Koch, T.; Linderoth, J.; et al. 2021. MIPLIB 2017: data-driven compilation of the 6th mixed-integer programming library. *Mathematical Programming Computation*, 13(3): 443–490.

[24] Goel, S.; Rao, J. M.; Shroff, R.; et al. 2016. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, 10(1): 365–394.

[25] Graf, E.; Schmoor, C.; Sauerbrei, W.; and Schumacher, M. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18): 2529–2545.

[26] Hales, B.; Terblanche, M.; Fowler, R.; and Sibbald, W. 2008. Development of medical checklists for improved quality of patient care. *International Journal for Quality in Health Care*, 20(1): 22–30.

[27] Hazimeh, H.; and Mazumder, R. 2020. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5): 1517–1537.

[28] Hazimeh, H.; Mazumder, R.; and Saab, A. 2022. Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Mathematical Programming*, 196(1): 347–388.

[29] ILOG, I. 2022. CPLEX Optimizer 22.1. https://www.ibm.com/products/ilog-cplex-optimization-studio.

[30] James, H.; Nagpal, C.; Heller, K. A.; and Ustun, B. 2023. Participatory Personalization in Classification. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

[31] Jung, J.; Concannon, C.; Shroff, R.; Goel, S.; and Goldstein, D. G. 2020. Simple rules to guide expert classifications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3): 771–800.

[32] Kim, M. P.; Korolova, A.; Rothblum, G. N.; and Yona, G. 2019. Preference-informed fairness. *arXiv preprint arXiv:1904.01793*.

[33] Klein, J. P.; Moeschberger, M. L.; et al. 2003. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer.

[34] Kramer, H. S.; and Drews, F. A. 2017. Checking the lists: A systematic review of electronic checklist use in health care. *Journal of biomedical informatics*, 71: S6–S12.

[35] Lin, D. 2007. On the Breslow estimator. *Lifetime data analysis*, 13: 471–480.

[36] Liu, J.; Zhong, C.; Li, B.; Seltzer, M.; and Rudin, C. 2022. FasterRisk: fast and accurate interpretable risk scores. *Advances in Neural Information Processing Systems*, 35: 17760–17773.

[37] Makhija, Y.; De Brouwer, E.; and Krishnan, R. G. 2022. Learning predictive checklists from continuous medical data. *arXiv preprint arXiv:2211.07076*.

[38] Morse, K. E.; Bagley, S. C.; and Shah, N. H. 2020. Estimate the hidden deployment cost of predictive models to improve patient care. *Nature medicine*, 26(1): 18–19.

[39] National Cancer Institute, S. R. P., DCCPS. 2019. Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1975-2016).

[40] Ning, Y.; Li, S.; Ong, M. E. H.; Xie, F.; Chakraborty, B.; Ting, D. S. W.; and Liu, N. 2022. A novel interpretable machine learning system to generate clinical risk scores: An application for predicting early mortality or unplanned readmission in a retrospective cohort study. *PLOS Digital Health*, 1(6): e0000062.

[41] Passi, S.; and Barocas, S. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, 39–48.

[42] Patel, J.; Ahmed, K.; Guru, K. A.; Khan, F.; Marsh, H.; Khan, M. S.; and Dasgupta, P. 2014. An overview of the use and implementation of checklists in surgical specialities–a systematic review. *International Journal of Surgery*, 12(12): 1317–1323.

[43] Pocock, S. J.; Ariti, C. A.; McMurray, J. J.; Maggioni, A.; Køber, L.; Squire, I. B.; Swedberg, K.; Dobson, J.; Poppe, K. K.; Whalley, G. A.; et al. 2013. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *European heart journal*, 34(19): 1404–1413.

[44] Sato, T.; Takano, Y.; and Miyashiro, R. 2015. Piecewise-Linear Approximation for Feature Subset Selection in a Sequential Logit Model. *arXiv preprint arXiv:1510.05417*.

[45] Sato, T.; Takano, Y.; Miyashiro, R.; and Yoshise, A. 2016. Feature subset selection for logistic regression via mixed integer optimization. *Computational Optimization and Applications*, 1–16.

[46] Sokolovska, N.; Chevaleyre, Y.; and Zucker, J.-D. 2018. A provable algorithm for learning interpretable scoring systems. In *International Conference on Artificial Intelligence and Statistics*, 566–574. PMLR.

[47] Struck, A. F.; Ustun, B.; Ruiz, A. R.; Lee, J. W.; LaRoche, S. M.; Hirsch, L. J.; Gilmore, E. J.; Vlachy, J.; Haider, H. A.; Rudin, C.; and Westover, M. B. 2017. Association of an Electroencephalography-Based Risk Score With Seizure Probability in Hospitalized Patients. *JAMA Neurology*, 74(12): 1419–1424.

[48] Suriyakumar, V. M.; Ghassemi, M.; and Ustun, B. 2023. When Personalization Harms Performance: Reconsidering the Use of Group Attributes in Prediction. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 33209–33228. PMLR.

[49] Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, 6373–6382.

[50] Ustun, B.; and Rudin, C. 2016. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3): 349–391.

[51] Ustun, B.; and Rudin, C. 2017. Optimized risk scores. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1125–1134.

[52] Ustun, B.; and Rudin, C. 2019. Learning Optimized Risk Scores. *Journal of Machine Learning Research*, 20(150): 1–75.

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

[54] Viviano, D.; and Bradic, J. 2020. Fair Policy Targeting. *arXiv preprint arXiv:2005.12395*.

[55] Wu, M.; Hughes, M.; Parbhoo, S.; Zazzi, M.; Roth, V.; and Doshi-Velez, F. 2018. Beyond sparsity: Tree regularization of deep models for interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[56] Yamga, E.; Mantena, S.; Rosen, D.; Bucholz, E. M.; Yeh, R. W.; Celi, L. A.; Ustun, B.; and Butala, N. M. 2023. Optimized Risk Score to Predict Mortality in Patients With Cardiogenic Shock in the Cardiac Intensive Care Unit. *Journal of the American Heart Association*, e029232.

[57] Zafar, M. B.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, 228–238.

[58] Zhang, H.; Morris, Q.; Ustun, B.; and Ghassemi, M. 2021. Learning Optimal Predictive Checklists. *Advances in Neural Information Processing Systems*, 34.

Supplementary Materials

A Adjusting performance metrics in the presence of censoring

Area under ROC Curve (AUC): The ROC curve is defined as a plot between the True Positive Rate/Sensitivity (TPR) and the False Positive Rate (FPR) for all thresholds at which a classifier can be deployed. Note that the FPR is equal to $1-$Specificity. We employ the technique proposed by [59, 60] to adjust the Sensitivity using IPCW estimates of the censoring distribution. The Specificity is computed on the uncensored instances.

$$\widehat{\text{Se}}(c,t) = \frac{\sum\limits_{i=1}^{n} \omega_i \cdot \mathbb{1}\{\pi_i(t) > c, T_i \leq t\}\cdot}{\sum\limits_{i=1}^{n} \omega_i \cdot \mathbb{1}\{T_i < t\}\cdot}; \quad \omega_i = \frac{\delta_i}{n \cdot \hat{G}(T_i)}; \quad \widehat{\text{Sp}}(c,t) = \frac{\sum\limits_{i=1}^{n} \mathbb{1}\{\pi_i(t) \leq c, T_i > t\}\cdot}{\sum\limits_{i=1}^{n} \mathbb{1}\{T_i > t\}\cdot}$$

$\widehat{\text{Se}}(c,t)$ and $\widehat{\text{Sp}}(c,t)$ refer to the estimated sensitivity and specificity at classification threshold c and time horizon t respectively. $\hat{G}(t)$ is a Kaplan-Meier estimator of the censoring distribution and $\pi_i(t)$ is the estimated survival probability, $\widehat{\mathbb{P}}(T > t|X =_i)$ by the classifier. This curve is plotted for all thresholds $c \in [0,1]$ and the area under the curve is used to AUC. For a larger discussion around comparisons of various strategies to compute ROC curves in the presence of censoring refer to [61].

Expected ℓ_1 Calibration Error (ECE): The ECE measures the average absolute difference between the observed and expected (according to the risk score) event rates, conditional on the estimated risk score. At time t, let the predicted risk score be $R(t) = \widehat{\mathbb{P}}(T > t|X)$. Then, the ECE approximates

$$\text{ECE}(t) = \mathbb{E}\big[\big|\mathbb{P}(T > t|R(t)) - R(t)\big|\big]$$

by partitioning the risk scores R into q quantiles $\{[r_j, r_{j+1})\}_{j=1}^{q}$. and computing the Kaplan-Meier estimate of the event rate $\text{KM}_j(t) \approx P(T > t|R \in [r_j, r_{j+1}))$, and the average risk score $\overline{R}_j = \frac{q}{n}\sum_{i:R_i \in [r_j, r_{j+1})} R_i$ in each bin. Altogether, the estimated ECE is

$$\widehat{\text{ECE}}(t) = \frac{1}{q}\sum_{j=1}^{q} |\text{KM}_j(t) - \overline{R}_j(t)|.$$

In practice, we fix the number of quantiles to be the minimum of 20 or the total number of discovered risk levels by the scoring system for our experiments.

Brier Score (BS): The Brier Score involves computing the Mean Squared Error around the binary forecast of survival at a certain event quantile of interest. Brier Score is a proper scoring rule and can be decomposed into components that measure both discriminative performance and calibration.

$$\begin{aligned}
\text{BS}(t) &= \mathbb{E}_{\mathcal{D}}\big[\big(\mathbb{1}\{T_i > t\} - \widehat{\mathbb{P}}(T > t|X)\big)^2\big] \\
\widehat{\text{BS}}_{\text{IPCW}}(t) &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\pi_i(t)^2 \mathbb{1}\{T \leq t, \delta_i = 1\}}{\hat{G}_i(T_i)} + \frac{\big(1 - \pi_i(t)\big)^2 \mathbb{1}\{T > t\}}{\hat{G}_i(t)}\right]; \\
\text{where, } \pi_i(t) &= \widehat{\mathbb{P}}(T > t|X_i)
\end{aligned}$$

The adjusted Brier Score adjusted for Censoring using IPCW is given by $\widehat{\text{BS}}_{\text{IPCW}}(t)$ as proposed in [25, 21] Here, $\hat{G}(.)$ is the Kaplan Meier estimate of the Censoring Distribution. When the Censoring distribution is independent of the Event distribution, the above quantity is an unbiased estimate of the Brier Score.

B Additional Results

Tables 5, 6 and 7 present the discriminative performance and Calibration of TSLIM in the presence of induced censoring.

`flchain`

`25%+ Censoring`

	Brier Score			**Area Under ROC Curve**			**ECE**			**OPT**
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.033	0.112	0.222	0.820	0.803	0.787	0.018	0.104	0.231	-
LR-L1-6	0.034	0.119	0.232	0.828	0.810	0.816	0.023	0.104	0.231	-
Cox	0.033	0.090	0.132	0.829	0.813	0.832	0.016	0.036	0.059	-
Cox-L1-6	0.034	0.099	0.145	0.815	0.809	0.822	0.027	0.064	0.098	-
RiskSLIM	0.034	0.114	0.223	0.821	0.800	0.789	0.011	0.104	0.229	2.46%
TSLIM	0.033	0.090	0.133	0.830	0.809	0.829	0.011	0.024	0.046	0.0%

`flchain`

`50%+ Censoring`

	Brier Score			**Area Under ROC Curve**			**ECE**			**OPT**
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.033	0.115	0.226	0.810	0.775	0.759	0.021	0.113	0.240	-
LR-L1-6	0.035	0.121	0.234	0.829	0.802	0.796	0.027	0.114	0.241	-
Cox	0.033	0.092	0.139	0.826	0.809	0.829	0.019	0.057	0.096	-
Cox-L1-6	0.034	0.102	0.156	0.817	0.810	0.822	0.026	0.069	0.116	-
RiskSLIM	0.033	0.114	0.225	0.817	0.807	0.799	0.015	0.112	0.238	2.58%
TSLIM	0.033	0.093	0.143	0.828	0.807	0.827	0.018	0.055	0.095	0.0%

Table 5: Discriminative performance and Calibration of TSLIM vs. RiskSLIM on the `flchain` data with enhanced censoring. All metrics are reported on the held-out test set and adjusted for censoring using Inverse Propensity of Censoring.

`support`

`25%+ Censoring`

	Brier Score			**Area Under ROC Curve**			**ECE**			**OPT**
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.225	0.235	0.279	0.681	0.692	0.713	0.066	0.137	0.337	-
LR-L1-6	0.240	0.251	0.283	0.636	0.662	0.693	0.096	0.147	0.338	-
Cox	0.229	0.224	0.184	0.673	0.693	0.723	0.083	0.082	0.082	-
Cox-L1-6	0.242	0.239	0.194	0.622	0.655	0.695	0.087	0.094	0.086	-
RiskSLIM	0.229	0.242	0.282	0.647	0.651	0.681	0.062	0.140	0.341	13.5%
TSLIM	0.233	0.231	0.189	0.650	0.666	0.697	0.082	0.081	0.086	0.67%

`support`

`50%+ Censoring`

	Brier Score			**Area Under ROC Curve**			**ECE**			**OPT**
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.232	0.251	0.319	0.672	0.685	0.701	0.129	0.206	0.404	-
LR-L1-6	0.245	0.266	0.325	0.636	0.661	0.687	0.141	0.219	0.417	-
Cox	0.243	0.244	0.208	0.661	0.685	0.719	0.171	0.181	0.185	-
Cox-L1-6	0.256	0.258	0.219	0.620	0.653	0.693	0.182	0.194	0.195	-
RiskSLIM	0.240	0.260	0.329	0.645	0.654	0.674	0.131	0.208	0.408	12.71%
TSLIM	0.247	0.248	0.212	0.642	0.669	0.693	0.172	0.184	0.191	0.74%

Table 6: Discriminative performance and Calibration of TSLIM vs. RiskSLIM on the `support` data with enhanced censoring. All metrics are reported on the held-out test set and adjusted for censoring using Inverse Propensity of Censoring.

seer-lymphoma										
25% Censoring										
	Brier Score			**Area Under ROC Curve**			**ECE**			**OPT**
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.136	0.220	0.273	0.807	0.797	0.775	0.026	0.200	0.281	-
LR-L1-6	0.165	0.261	0.308	0.706	0.729	0.750	0.081	0.228	0.300	-
Cox	0.142	0.177	0.182	0.785	0.806	0.811	0.042	0.066	0.065	-
Cox-L1-6	0.170	0.225	0.229	0.660	0.692	0.732	0.058	0.101	0.117	-
RiskSLIM	0.146	0.232	0.284	0.761	0.751	0.732	0.030	0.200	0.282	0.0%
TSLIM	0.149	0.191	0.200	0.750	0.769	0.767	0.049	0.074	0.074	0.0%

seer-lymphoma										
50% Censoring										
	Brier Score			**Area Under ROC Curve**			**ECE**			**OPT**
	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	1-Year	5-Year	10-Year	
LR	0.137	0.234	0.292	0.805	0.793	0.77	0.064	0.242	0.323	-
LR-L1-6	0.167	0.275	0.329	0.704	0.727	0.747	0.108	0.271	0.351	-
Cox	0.147	0.191	0.200	0.783	0.804	0.807	0.089	0.141	0.154	-
Cox-L1-6	0.172	0.237	0.248	0.674	0.708	0.743	0.121	0.194	0.203	-
RiskSLIM	0.146	0.245	0.302	0.760	0.747	0.725	0.069	0.245	0.326	0.0%
TSLIM	0.153	0.206	0.219	0.744	0.763	0.765	0.094	0.147	0.164	0.0%

Table 7: Discriminative performance and Calibration of TSLIM vs. RiskSLIM on the **seer-lymphoma** data with enhanced censoring. All metrics are reported on the held-out test set and adjusted for censoring using Inverse Propensity of Censoring.

C TSLIM as a Mixed Integer Non-Linear Program

Definition 2 (Hazard Scoring Problem)**.** *The hazard scoring problem is a discrete optimization problem of the form:*

$$\min_{\boldsymbol{w}} \quad \mathcal{PL}(\mathcal{D};\boldsymbol{w},c) \quad \text{s.t.} \quad \boldsymbol{w} \in \mathcal{W} \quad \textit{and} \quad \|\boldsymbol{w}\|_0 \leq R^{\max}, \tag{5}$$

where:

- $\mathcal{PL}(\mathcal{D};\boldsymbol{w},c) = \sum_{i=1}^{n} \mathbf{1}_{\delta_i \neq 0}\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}_i}{c} - \log \sum_{j \in \mathcal{R}(t_i)} \exp\left(\frac{\boldsymbol{w}^\top \boldsymbol{x}_j}{c}\right)\right)$ *is the partial likelihood;*
- $\|\boldsymbol{w}\|_0 = \sum_{j=1}^{d} \boldsymbol{1}\{w_j \neq 0\}$ *is the* ℓ_0*-seminorm;*
- $\mathcal{W} \subset \mathcal{Z}^{d+1}$ *is a set of feasible coefficient vectors, e.g.,* $\mathcal{W} = \{-5, 5\}^{d+1}$*;*
- $R^{\max} \in \mathcal{Z}$ *is a user-specified parameter to impose sparsity in the learnt coefficient set.*

This problem captures what we desire in a scoring system. The objective minimizes the *partial likelihood* over the event rate which inturn helps recovering models that are well calibrated with good discriminative performance. Further it penalizes the ℓ_0-seminorm (the count of non-zero coefficients) for sparsity. The trade-off parameter ϵ controls the balance between these competing objectives, and represents the maximum log-likelihood that is sacrificed to remove a feature from the optimal model. The constraints restrict coefficients to a set of small integers such as $\mathcal{W} := \{-5, \ldots, 5\}^{d+1}$, and may be customized to encode other model requirements such as those in Table 1.

We optimize the problem in Equation 5 by solving the following MINLP:

$$\min_{\boldsymbol{w}} \qquad L + \epsilon R$$

$$\text{s.t.} \qquad L = \sum_{i=1}^{n} \mathbf{1}_{\delta_i \neq 0}\left(\boldsymbol{w}^\top \boldsymbol{x}_i / c - \log \sum_{j \in \mathcal{R}(t_i)} \exp\left(\boldsymbol{w}^\top \boldsymbol{x}_j / c\right)\right) \qquad \textit{partial likelihood} \tag{6a}$$

$$R = \sum_{j \in [d]} \alpha_j \qquad \textit{model size} \tag{6b}$$

$$W_j^{\max} \alpha_j \geq w_j \qquad j \in [d] \qquad w_j > 0 \implies \alpha_j = 1 \tag{6c}$$

$$-W_j^{\text{min}}\alpha_j \geq -w_j \qquad j \in [d] \qquad w_j < 0 \implies \alpha_j = 1 \tag{6d}$$

$$\begin{aligned}
&L \in [0, \ldots, L^{\text{max}}] && && \textit{loss} && (6e)\\
&R \in \{0, \ldots, R^{\text{max}}\} && && \textit{model size} && (6f)\\
&w_j \in \{W_j^{\text{min}} \ldots, W_j^{\text{max}}\} && j \in [d] && \textit{coef for variable } j && (6g)\\
&\alpha_j \in \{0, 1\} && j \in [d] && \alpha_j := 1[w_j \neq 0] && (6h)
\end{aligned}$$

- L and R are "auxiliary" variables that represent the overall loss and the model size, respectively. In theory, these variables are redundant in that they could be replace by the quantities in (6a) and (6b). In practice, we include them because they allow us to set upper and lower bounds on feasible models via "variable definition constraints" in (6e) and (6f).
- The formulation accounts for model size using the indicator variables $\alpha_j := 1[w_j \neq 0]$. These variables are set to 1 whenever $w_j \neq 0$ through the constraints in (6c) and (6d).
- The coefficient for each variable is constrained to small integer values in Constraints . These constraints restrict each w_j to integers from W_j^{min} to W_j^{max}. By default, we set these values to $W_j^{\text{min}} = -5$ and $W_j^{\text{max}} = +5$.

Cutting-Plane Formulation We recover an optimal solution to the MINLP in (6) with the lattice cutting-plane algorithm in [52]. The cutting-plane algorithm solves a surrogate problem that replaces loss function with a linearized "cutting-plane" approximation. This problem is a MINLP (6) with the following form:

$$\begin{aligned}
\min_{\boldsymbol{w}} \quad & L + \epsilon R \\
\text{s.t.} \quad & L \geq \mathcal{PL}(\boldsymbol{w}^t) + \nabla\mathcal{PL}(\boldsymbol{w}^t)(\boldsymbol{w} - \boldsymbol{w}^t) && t \in [T] && \textit{loss cuts} && (7a)\\
& R = \sum_{j \in [d]} \alpha_j && && \textit{model size} && (7b)\\
& W_j^{\text{max}}\alpha_j \geq w_j && j \in [d] && w_j > 0 \implies \alpha_j = 1 && (7c)\\
& -W_j^{\text{min}}\alpha_j \geq -w_j && j \in [d] && w_j < 0 \implies \alpha_j = 1 && (7d)\\
& L \in [0, \ldots, L^{\text{max}}] && && \textit{loss} && (7e)\\
& R \in \{0, \ldots, R^{\text{max}}\} && && \textit{model size} && (7f)\\
& w_j \in \{W_j^{\text{min}} \ldots, W_j^{\text{max}}\} && j \in [d] && \textit{coef for variable } j && (7g)\\
& \alpha_j \in \{0, 1\} && j \in [d] && \alpha_j := 1[w_j \neq 0] && (7h)
\end{aligned}$$

The main difference with the MIP formulation in (7) and the MINLP formulation in (6) is that we now compute the loss using a cutting-plane approximation of the loss function. The cutting-plane approximation is captured through T cuts (7a). Each cut is a supporting hyperplane to the loss function at a specific point $\boldsymbol{w}^t$ – where the values of $\boldsymbol{w}^t$ represent integer-feasible solutions. Since we with the Cox partial likelihood (i.e, a convex loss function), the cutting-plane approximation is an under-approximation.

Supplementary References

[59] H. Uno, T. Cai, L. Tian, and L.-J. Wei, "Evaluating prediction rules for t-year survivors with censored regression models," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 527–537, 2007.

[60] H. Hung and C.-t. Chiang, "Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data," *Scandinavian journal of statistics*, vol. 37, no. 4, pp. 664–679, 2010.

[61] A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona, "Time-dependent roc curve analysis in medical research: current methods and applications," *BMC medical research methodology*, vol. 17, no. 1, pp. 1–19, 2017.