Temporal Quilting for Survival Analysis

Changhee Lee
UCLA

William R. Zame
UCLA

Ahmed M. Alaa
UCLA

Mihaela van der Schaar
University of Cambridge
UCLA
Alan Turing Institute

Abstract

The importance of survival analysis in many disciplines (especially in medicine) has led to the development of a variety of approaches to modeling the survival function. Models constructed via various approaches offer different strengths and weaknesses in terms of discriminative performance and calibration, but no one model is best across all datasets or even across all time horizons within a single dataset. Because we require both good calibration and good discriminative performance over different time horizons, conventional model selection and ensemble approaches are not applicable. This paper develops a novel approach that combines the collective intelligence of different underlying survival models to produce a valid survival function that is well-calibrated and offers superior discriminative performance at different time horizons. Empirical results show that our approach provides significant gains over the benchmarks on a variety of real-world datasets.

1 Introduction

Survival analysis (time-to-event analysis) plays an important role in many disciplines and especially in medicine, which is the focus of the paper. The importance of survival analysis has prompted the development of a variety of approaches to model the survival function (the probability of surviving past a given time as a function of the covariates). Parametric and semi-parametric approaches construct models that rely on specific assumptions about the true underlying distribution; non-parametric approaches take a more agnostic point of view to construct models that rely on (variants of) familiar machine learning methods. The models produced by these various approaches offer different strengths and weaknesses in terms of both discriminative performance and calibration, and their relative performance varies across different datasets and at different time horizons within a single dataset. In particular, no single model is best across all datasets, and frequently no single model is best across all time horizons within a single dataset. This presents a challenge to familiar methods of model selection or ensemble creation. An additional challenge is that survival analysis needs to yield good performance at different time horizons while providing a valid and well-calibrated survival function; this makes the conventional model selection or ensemble methods actually inapplicable.

The usefulness of a survival model should be assessed both by how well the model *discriminates* among predicted risks and by how well the model is *calibrated*. The necessity of keeping both criteria in mind is illustrated by the case of heart transplantation, which is the treatment of last resort for patients with end-stage heart failure. Successful transplantation can mean many additional years of life for such patients, but there are many more patients in need of transplants than there are available donor hearts. So, it is important to correctly discriminate/prioritize recipients on the basis of risk. However, if the risk predictions of a given model are not well calibrated to the truth – i.e. if there is poor agreement between predicted and observed outcomes – then the model will have little prognostic value for clinicians.

This paper offers a novel approach that addresses these challenges. Our approach combines the collective intelligence of different underlying survival models to produce a valid survival function that is both discriminative and well-calibrated. Because we piece together these underlying models according to (endogenously determined) weights that vary over time, we refer to our construction as *temporal quilting*, and to the resultant model as a *Survival Quilt*. An illustration of

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

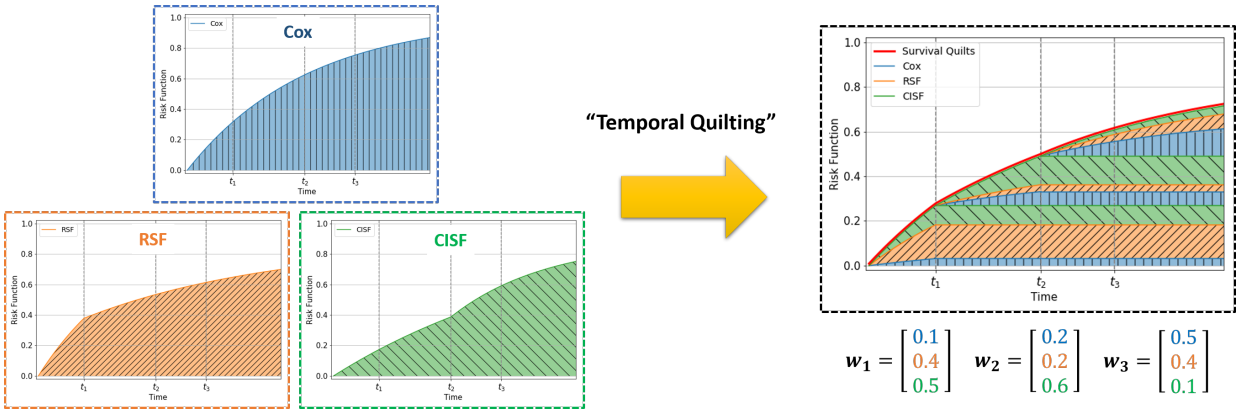


Figure 1: A toy example of temporal quilting with prescribed weights for survival models in $\mathcal{M} = \{\text{Cox}, \text{RSF}, \text{CISF}\}$ at t_1, t_2, and t_3. A risk function is constructed by stitching together the weighted increment functions of each survival model between two adjacent time horizons.

temporal quilting (for given weights) is provided in Figure 1. The core part of our method is an algorithm for configuring the weights sequentially over a (perhaps very fine) grid of time intervals. To render the problem tractable, we apply constrained Bayesian Optimization (BO) [1], which models the discrimination and calibration performance metrics as black-box functions, whose input is an array of weights (over different time horizons) and whose output is the corresponding performance achieved. Based on the constructed array of weights, our method makes a single predictive model – a Survival Quilt.

Our empirical results demonstrate that Survival Quilts provide significant performance gains over the underlying models (which we take as benchmarks) on a variety of real-world survival datasets. Because our approach automatically finds (an approximation to) the best temporal quilting of the underlying survival models, it provides a way to free clinicians from the concern of choosing one particular survival model for each dataset and for each time horizon of interest.

2 Related Work

Different approaches, ranging from statistical methods to machine learning based methods, have been proposed for survival analysis. One approach employs (semi-)parametric models that are constructed on the basis of assumptions on the true underlying distribution. This includes i) survival models based on the Cox proportional hazard (Cox-PH) assumption [2], and a variety of extensions [3, 4, 5] and ii) the accelerated failure time (AFT) model based on the Weibull distribution, and extensions [6, 7]. Other approaches employ nonparametric models, including i) ensembles of survival trees constructed via bagging [8, 9] or boosting [10], and ii) deep learning methods [11]. In general, nonparametric models provide better survival predictions than do (semi-)parametric models when the true underlying distribution is unknown or is mis-specified. However, nonparametric models often yield inaccurate predictions at time horizons for which the number of subjects in the dataset who are "at risk" is small [12].

As we have noted in the Introduction, methods based on model selection and ensemble creation that are familiar for classification problems (including the Auto-ML framework [13, 14]) do not extend to the survival setting because we need to construct a valid survival function that provides good discriminative performance at different time horizons and is also well-calibrated. Our work is most closely related to a model based on stacking [15], which estimates an optimally weighted combination of different survival models on the basis of calibration performance. However, in order to produce a valid survival function, that model requires the weights to be independent of time. By contrast, our approach exploits weights that depend on time to provide a valid survival function that is well-calibrated and achieves superior discriminative performance at different time horizons. To the best of our knowledge, this paper is the first that combines different survival models in a time-dependent manner to provide both discriminative and prognostic power.

3 Problem Formulation

For convenience, we couch our description in the medical setting, although our approach is entirely applicable to any time-to-event problem. In our setting, some patients experience the event of interest (e.g. death) and some are censored (lost to follow-up). The data for an individual patient i therefore consists of a vector

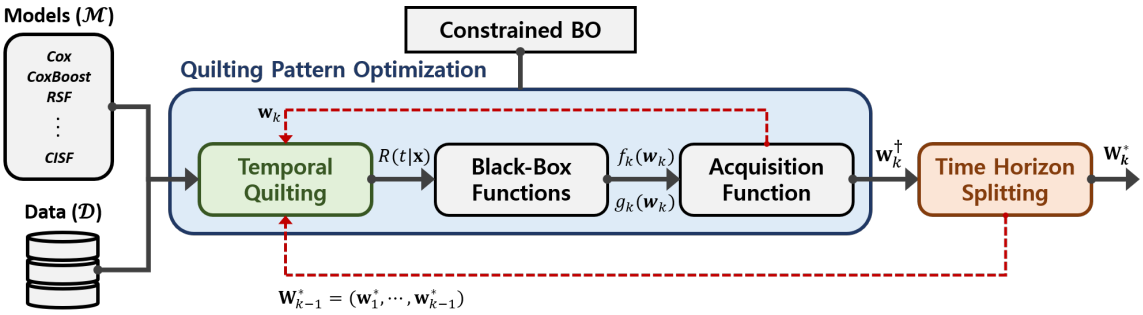


Figure 2: A schematic depiction of Survival Quilts and its pattern optimization at step k. Survival Quilts provide risk functions that are constructed on the basis of the final quilting pattern $\mathbf{W}_K^*$. Here, colored boxes are the three main components of our method and dotted lines imply feedback loops for sequential computations.

of covariates $\mathbf{x}_i \in \mathcal{X}$ ($\mathcal{X}$ is the space of all covariates), either a time-to-event, $T_i \in \mathbb{R}^+$, or a time-to-censoring, $C_i \in \mathbb{R}^+$ (both from the initial moment of observation), and an indicator $\Delta_i = I(T_i < C_i)$; $\Delta = 1$ if the patient experienced the event of interest and $\Delta = 0$ if the patient was right-censored. Note that censoring provides the information that the patient had *not* experienced the event (e.g. was alive) up to time C_i. We are given data for N patients so the entire time-to-event dataset is $\mathcal{D} = \{(\mathbf{x}_i, \Delta_i T_i + (1-\Delta_i)C_i, \Delta_i)\}_{i=1}^N$.

Our goal is to estimate the *risk function* $R : \mathcal{X} \times \mathbb{R}^+ \to [0, 1]$

$$R(t|\mathbf{x}) = \mathbb{P}(T \leq t|\mathbf{x}), \tag{1}$$

which is the probability of the event occurring at or before time t given the covariates $\mathbf{x}$. (Equivalently, we could estimate the *survival function* $S : \mathcal{X} \times \mathbb{R}^+ \to [0, 1]$; $S(t|\mathbf{x}) = \mathbb{P}(T > t|\mathbf{x}) = 1 - R(t|\mathbf{x})$ is the probability of the event occurring after time t, given covariates $\mathbf{x}$.)

Since we aim at finding the best predictive model among the set of all models that provide well-calibrated risk functions, it is natural to formulate the optimization problem as maximizing discriminative performance subject to a constraint on calibration. If we write $\mathcal{R}$ for the set of all risk functions, $f(\cdot)$ for a metric of discriminative performance, and $g(\cdot)$ as a metric of calibration, then our problem is to find the risk function $R^* \in \mathcal{R}$ that solves the following maximization problem:

$$\begin{aligned} \max_{R \in \mathcal{R}} \quad & f(R) \\ \text{s.t.} \quad & g(R) \leq c, \end{aligned} \tag{2}$$

where $c > 0$ is some prescribed tolerance of predictive error. (In the experiments reported below, we take f and g to be the time-dependent C-index and Brier Score, respectively, but other metrics could be used.)

4 Survival Quilts

As noted in the introduction, the existing survival models may fail to capture the true survival behavior in different settings and over different time horizons. (See also the discussion in Section 5.) Survival Quilts address both these failings by forming *time-varying* ensembles of different survival models.

Table 1: List of survival models used in Survival Quilts

Cox-PH model	AFT model	Survival Forest
Cox	Weibull	RSF
CoxRidge	LogNormal	CISF
CoxBoost	Exponential	

Given a time-to-event dataset $\mathcal{D}$ and a set of survival models $\mathcal{M}$ e.g., $\mathcal{M} = \{\text{Cox}, \text{Weibull}, \text{RSF}, \cdots\}$, (a full list of survival models used in this paper is provided in Table 1), our method outputs a predictive model – a *Survival Quilt* – that provides a valid risk function. A Survival Quilt is constructed endogenously from the data following three steps. The first step is *temporal quilting* which constructs valid risk functions for a *given* array of weights (a *quilting pattern*) for survival models in $\mathcal{M}$ over time horizons. The second step models the performance of these risk functions as black-box functions and applies constrained BO to (approximately) optimize the quilting pattern. The final step splits the time horizons in order to insure robustness of the (approximately optimized) quilting pattern. A schematic overview of our method is illustrated in Figure 2; details of each of these steps are described in the following subsections.

4.1 Temporal Quilting: Constructing a New Risk Function

Constructing a survival model entails learning a risk function that spans a continuum of time horizons. We

do not treat predictions at each time horizon as separate problems, but rather provide a natural construct for the entire risk function; risk predictions at past time horizons are carried forward to future time horizons to provide a consistent risk function. More specifically, given an increasing sequence of time horizons $\mathcal{T} = \{t_0 = 0, t_1, \cdots, t_K\}$, we first break down the risk functions provided by each survival model in $\mathcal{M}$ into *pieces* by focusing on the increment between two adjacent time horizons, t_{k-1} and t_k for $k = 1, \cdots, K$. We then assemble the pieces in a *quilting pattern* that, on each time interval, assigns weights to each of the increment functions of the underlying survival models and then sums the weighted combination of the increment functions over time.

We define the *increment function* of model $m \in \mathcal{M}$ on the interval $[a, b]$, given covariate $\mathbf{x}$, to be

$$i_m(a, b|\mathbf{x}) = R_m(b|\mathbf{x}) - R_m(a|\mathbf{x}), \tag{3}$$

where R_m is the risk function issued by model $m \in \mathcal{M}$. Because R_m is non-decreasing on the interval $[a, b]$, i_m is non-decreasing and non-negative on the interval $[a, b]$. Let $\mathbf{w}$ be a $|\mathcal{M}|$-dimensional *weight vector*, where $\mathbf{w}[m] \in [0, 1]$ indicates the weight for model m and $\sum_{m\in\mathcal{M}} \mathbf{w}[m] = 1$. Given $\mathbf{w}$, the *weighted increment function* on the interval $[a, b]$ is defined to be

$$I_{\mathbf{w}}(a, b|\mathbf{x}) = \sum_{m\in\mathcal{M}} \mathbf{w}[m] \cdot i_m(a, b|\mathbf{x}). \tag{4}$$

Then, given weights $\mathbf{w}_1, \ldots, \mathbf{w}_k$ and a time $t \in [t_{k-1}, t_k]$, we set

$$R_0(t|\mathbf{x}) = \sum_{\ell=1}^{k-1} I_{\mathbf{w}_\ell}(t_{\ell-1}, t_\ell|\mathbf{x}) + I_{\mathbf{w}_k}(t_{k-1}, t|\mathbf{x}), \tag{5}$$

where the first term is the aggregate risk up to time t_{k-1} and the second term is the *incremental* from time t_{k-1} to time $t \in [t_{k-1}, t_k]$. Now, we define the *risk function* at time t to be

$$R(t|\mathbf{x}) = \min\big\{1, R_0(t|\mathbf{x})\big\}. \tag{6}$$

A few words of explanation may be useful.

- Note that $R_m(0|\mathbf{x}) = 0$ (patients are alive at the beginning of the observation period) so that $i_m(0, t|\mathbf{x}) = R_m(t|\mathbf{x})$. Hence, if $t \in [0, t_1]$, then $R(t|\mathbf{x}) = \sum_{m\in\mathcal{M}} \mathbf{w}_1[m] \cdot R_m(t|\mathbf{x})$.
- R_0 might exceed 1, in which case it could not be a valid risk function. (The probability that the event has occurred cannot exceed 1.) Hence, we truncate by setting $R = \min\{1, R_0\}$.
- Because the weighted increment functions are non-decreasing and non-negative, the functions R_0, R are also non-decreasing – hence R is a valid risk function.

We frequently refer to the array of weights $\mathbf{W}_K = (\mathbf{w}_1, \cdots, \mathbf{w}_K)$ as a *quilting pattern*; we refer to the construction above as *temporal quilting*. Figure 1 illustrates a quilting pattern and the resulting risk function constructed via temporal quilting.

4.2 Quilting Pattern Optimization via BO

Let $\mathbf{W}_K = (\mathbf{w}_1, \cdots, \mathbf{w}_K)$ be a quilting patterns (configuration of weights); write $\mathbf{W}_k = (\mathbf{w}_1, \cdots, \mathbf{w}_k)$ for the configuration up to time t. Our approach is to find the best risk function R that can be formed as in (6). Because R is completely defined by the configuration of weights, this amounts to finding the best quilting pattern $\mathbf{W}_K^*$ – i.e., the quilting pattern that solves the following maximization problem:

$$\begin{aligned} \max_{\mathbf{W}_K} \quad & f(\mathbf{W}_K) \\ \text{s.t.} \quad & g(\mathbf{W}_K) \le c, \end{aligned} \tag{7}$$

where $c > 0$ is the prescribed tolerance of predictive error. In (7), we take the function f to be the average of functions f_k that are the metric of time-dependent discriminative performance at t_k (see the definition below in (11)); similarly we take the function g to be the average of functions g_k that are the metric of time-dependent calibration performance at t_k (see the definition below in (12)). Formally, $f(\mathbf{W}_K) = \frac{1}{K}\sum_{k=1}^K f_k(\mathbf{W}_k)$ and $g(\mathbf{W}_K) = \frac{1}{K}\sum_{k=1}^K g_k(\mathbf{W}_k)$. Since the objective and constraint functions in (7) have no analytic form, we treat them as black-box functions $f, g : [0, 1]^{K\times|\mathcal{M}|} \to \mathbb{R}$. That is, given a quilting pattern $\mathbf{W}_K$, we can only evaluate the noisy versions of f and g which are given by $\frac{1}{J}\sum_{j=1}^J \mathcal{L}_f(\mathbf{W}_K; \mathcal{D}_{\text{tr}}^{(j)}, \mathcal{D}_{\text{va}}^{(j)})$ and $\frac{1}{J}\sum_{j=1}^J \mathcal{L}_g(\mathbf{W}_K; \mathcal{D}_{\text{tr}}^{(j)}, \mathcal{D}_{\text{va}}^{(j)})$, respectively. Here, $\mathcal{L}_f$ and $\mathcal{L}_g$ are the empirical values for the given performance metrics f and g, respectively, and $\mathcal{D}_{\text{tr}}^{(j)}$ and $\mathcal{D}_{\text{va}}^{(j)}$ denote training and validation splits of $\mathcal{D}$ in the j-th fold of J-fold cross-validation.

To search for the optimal quilting pattern $\mathbf{W}_K^*$, we use Bayesian optimization (BO) and solve a black-box optimization problem under a black-box constraint [1]. The BO algorithm specifies a Gaussian process (GP) prior on f and g as

$$\begin{aligned} f &\sim \mathcal{GP}(\mu_f(\mathbf{W}_K), \kappa_f(\mathbf{W}_K, \mathbf{W}'_K)) \\ g &\sim \mathcal{GP}(\mu_g(\mathbf{W}_K), \kappa_g(\mathbf{W}_K, \mathbf{W}'_K)) \end{aligned} \tag{8}$$

where $\mu_f(\mathbf{W}_K)$ and $\mu_g(\mathbf{W}_K)$ are the mean functions, encoding the expected performance of different quilting patterns, and $\kappa_f(\mathbf{W}_K, \mathbf{W}'_K)$ and $\kappa_g(\mathbf{W}_K, \mathbf{W}'_K)$ are the covariance kernels [16], encoding the similarity between different quilting patterns for f and g, respectively. We refer to the optimization problem in (7) as the Quilting Pattern Composition Problem (QPCP).

4.3 Sequential BO for QPCP

The functions f, g are defined over a space of dimension $D = K \times |\mathcal{M}|$. Note that D is large even for relatively small sets $\mathcal{M}$ of underlying survival models and a relatively coarse grid of time horizons; e.g. $D = 80$ if $|\mathcal{M}| = 8$ (as in our experiments) and $K = 10$. (In practice, it seems desirable to allow the grid of time horizons to be much finer than this; e.g. if the most distant horizon is 10 years we might want the grid to consist of 40 quarters or 120 months or perhaps even something finer. Moreover, although here we use only eight underlying models, it might well be desirable to use many more models – and one of the virtues of our approach is that this is possible.) This high-dimensionality renders standard GP-based BO infeasible because both the sample complexity of nonparametric estimation of the functions f, g and the computational complexity of maximizing the acquisition function are exponential in D [17, 14]. For these reasons, we propose instead a sequential greedy algorithm that incrementally selects a time horizon and performs constrained BO on that time horizon. (For a more detailed discussion of the computational complexity, see the Supplementary Material.)

Let $\mathbf{W}^*_{k-1} = (\mathbf{w}^*_1, \cdots, \mathbf{w}^*_{k-1})$ be the configuration of weights found through step $k-1$ (i.e., the time horizon t_{k-1}). Following the greedy approach, we find the weights at step k (i.e., the time horizon t_k) by solving the following BO:

$$\begin{aligned} \max_{\mathbf{w}_k} \quad & f_k(\mathbf{w}_k; \mathbf{W}^*_{k-1}) \\ \text{s.t.} \quad & g_k(\mathbf{w}_k; \mathbf{W}^*_{k-1}) \leq c, \end{aligned} \tag{9}$$

where we have written $\mathbf{w}_k; \mathbf{W}^*_{k-1}$ as shorthand for $(\mathbf{w}^*_1, \cdots, \mathbf{w}^*_{k-1}, \mathbf{w}_k)$. We have chosen this notation to emphasize that $\mathbf{W}^*_{k-1}$ is fixed so f_k, g_k depend only on $\mathbf{w}_k$. BO specifies GP priors on f_k and g_k as $f_k \sim \mathcal{GP}(\mu_{f_k}(\mathbf{w}_k; \mathbf{W}^*_{k-1}), \kappa_{f_k}(\mathbf{w}_k, \mathbf{w}'_k; \mathbf{W}^*_{k-1}))$ and $g_k \sim \mathcal{GP}(\mu_{g_k}(\mathbf{w}_k; \mathbf{W}^*_{k-1}), \kappa_{g_k}(\mathbf{w}_k, \mathbf{w}'_k; \mathbf{W}^*_{k-1}))$. From this point forward we simplify notation by omitting the dependence on $\mathbf{W}^*_{k-1}$.

4.3.1 Black-box Constrained BO

At step k, to solve the black-box constrained BO in (9), we approximate the problem by an augmented Lagrangian framework as proposed in [18]. In particular, (9) can be relaxed to minimizing the augmented Lagrangian problem given by

$$\begin{aligned} L(\mathbf{w}_k; \lambda, \rho) = & - f_k(\mathbf{w}_k) + \lambda \cdot (g_k(\mathbf{w}_k) - c) \\ & + \frac{1}{\rho} \max(0, g_k(\mathbf{w}_k) - c)^2, \end{aligned} \tag{10}$$

where $\rho > 0$ and $\lambda \geq 0$ indicate a penalty parameter and a Lagrange multiplier, respectively.

Algorithm 1 Augmented Lagrangian optimization

Initialize: $\lambda^{(0)} \geq 0$, $\rho^{(0)} > 0$, and $\mathbf{w}^{(0)}_k$
for $n = 1, 2, \cdots, n_{\max}$ **do**
  Find $\mathbf{w}^{*(n)}_k$ that approximately solve (10)
  Update $\lambda^{(n)} \leftarrow \max\left(0, \lambda^{(n-1)} + \frac{1}{\rho^{(n-1)}}(g_k(\mathbf{w}^{*(n)}_k) - c)\right)$
  Update $\mathbf{w}^{\dagger}_k \leftarrow \mathbf{w}^{*(n)}_k$
  if $g_k(\mathbf{w}^{\dagger}_k) \leq c$ **then**
    Update $\rho^{(n)} \leftarrow \rho^{(n-1)}$
  else
    Update $\rho^{(n)} \leftarrow \frac{1}{2}\rho^{(n-1)}$
  end if
end for

An efficient algorithm in [19] transforms the original constrained problem into a sequence of subproblems: at the n-th subproblem, we find a weight vector at t_k, which is denoted as $\mathbf{w}^{(n)}_k$, by solving (10) given $\rho^{(n-1)}$ and $\lambda^{(n-1)}$. After finding a candidate solution $\mathbf{w}^{*(n)}_k$, the penalty parameter and approximate Lagrange multipliers are updated and the process repeats until termination conditions are satisfied. We denote the final output of the constrained BO at step k by $\mathbf{w}^{\dagger}_k$. (Throughout the experiments, we set the terminal condition to be satisfaction of the constraint by $\mathbf{w}^{\dagger}_k$ or n reaching the maximum number of subproblems $n_{\max}$.) Algorithm 1 gives the specific updates utilized in this paper.

4.3.2 Endogenous Time Horizon Splitting

In principle, we could always use $\mathbf{w}^{\dagger}_k$ to extend the sequence of weights. However, doing so would make the construction fragile because the optimal weights might become over-fitted. In order to make the construction more robust, we introduce a required margin of improvement $\delta > 0$; if using $\mathbf{w}^{\dagger}_k$ to extend the sequence of weights leads to an improvement in discriminative performance of at least δ, we set $\mathbf{w}^*_k = \mathbf{w}^{\dagger}_k$; otherwise we set $\mathbf{w}^*_k = \mathbf{w}^*_{k-1}$. In the former case, t_k represents an endogenously learned split in the time horizon – a time when the quilting pattern changes. The overall process of our method is illustrated in Algorithm 2. (The overall computational complexity of training our method (off-line) and predicting new risk functions (on-line) is provided in the Supplementary Material.)

5 Experiments

In this section, we present discriminative performance results in comparison to competitive baseline algorithms on six real-world time-to-event datasets. We set $K = 50$ and $\Delta t = \frac{T_{\max}}{K}$ where $T_{\max}$ indicates the maximum of time-to-event and time-to-censoring in each dataset. Throughout the evaluation, we report

Table 2: Descriptive statistics on the six real-world datasets. Mean (standard deviation) times in days are provided for the time-to-event/censoring.

Statistics	Datasets					
	MAGGIC	SUPPORT	METABRIC	UNOS-I	UNOS-II	BPD
No. Patients	5000	9105	1981	792	5000	2510
Events	1827 (36.5%)	6201 (68.1%)	888 (44.8%)	363 (45.8%)	2395 (47.9%)	1999 (79.6%)
Censored	3173 (63.5%)	2904 (31.9%)	1093 (55.2%)	429 (54.2%)	2605 (52.1%)	511 (20.4%)
Time-to-Event	885.3 (957.0)	206.0 (321.9)	2318.5 (1613.8)	141.0 (213.9)	2161.3 (2084.0)	613.5 (853.0)
Time-to-Censoring	927.7 (1032.4)	1060.3 (516.1)	3464.8 (1773.7)	327.6 (380.5)	2733.1 (2151.8)	1331.8 (1407.9)
No. Features	33	42	21	16	50	48

Algorithm 2 Sequential BO for QPCP

Initialize: $\mathbf{W}_0^* = \varnothing$, $\delta > 0$, and $\Delta t > 0$
for $k = 1, 2, \cdots, K$ **do**
  Set $t_k \leftarrow t_{k-1} + \Delta t$
  Obtain $\mathbf{w}_k^\dagger$ from **Algorithm 1** with $\mathbf{W}_{k-1}^*$ and t_k
  if $f_k(\mathbf{w}_k^\dagger) - f_k(\mathbf{w}_{k-1}^*) > \delta$ **then**
    Update $\mathbf{w}_k^* \leftarrow \mathbf{w}_k^\dagger$
  else
    Update $\mathbf{w}_k^* \leftarrow \mathbf{w}_{k-1}^*$
  end if
  Set $\mathbf{W}_k^* \leftarrow (\mathbf{w}_1^*, \mathbf{w}_2^*, \cdots, \mathbf{w}_k^*)$
end for

results using the average of 5 random 80/20 train/test splits.

5.1 Experimental Setup

5.1.1 Survival Models

The survival models that are used for constructing Survival Quilts and for the comparisons are listed below along with description of the implementations used to compute them: the standard Cox-PH model (**Cox**) [2] and the modification with ridge regression (taking $\alpha = 1$) (**CoxRidge**) are implemented with Python package `scikit-surv`; the survival regression models using the Weibull (**Weibull**), Log-normal (**LogNormal**) and Exponential (**Exponential**) distributions are implemented with R package `survival`; the Cox-PH model with the component-wise likelihood-based boosting algorithm [4] (**CoxBoost**) is implemented with R package `CoxBoost` with 500 iterations; the bagging-based Random Survival Forest [8] (**RSF**) is implemented with the R package `RandomForestSRC` with 1000 trees; and the Conditional Inference Survival Forest [9] (**CISF**) is implemented with the R package `pec` with 1000 trees.

5.1.2 Performance Metrics

As discussed above, we assess the predictions of all the survival models with respect to how well the predictions discriminate among individual risks and how accurate the predictions are. As the metric of discriminative power, we use the time-dependent concordance index (C-index) [20], defined by

$$C(t) = \mathbb{P}(R(t|\mathbf{x}_i) > R(t|\mathbf{x}_j) | \Delta_i = 1, T_i \leq t, T_i < T_j). \tag{11}$$

As the metric of calibration, we use the Brier Score (BS) [21] which is the mean square error adjusted for the survival setting:

$$BS(t) = \mathbb{E}\left[(\mathbf{1}(T_i \leq t) - R(t|\mathbf{x}_i))^2\right]. \tag{12}$$

These metrics can be evaluated over different time horizons and are adjusted for censoring as defined in [20] and [21].

5.2 Datasets

We conducted experiments to investigate the performance of Survival Quilts on six real-world medical datasets from a variety of clinical settings: a preventive care database on chronic heart failure (**MAGGIC**) [22], a study to understand seriously ill hospitalized adults (**SUPPORT**) [23], a study on breast cancer subgroups (**METABRIC**) [24], databases on heart transplant management for patients (**UNOS-I**) wait-listed for transplantation and on patients who underwent a heart transplant (**UNOS-II**)[1], and preventative care records on bipolar disorder (**BPD**) [25]. In Table 2, we provide a summary of these time-to-event datasets. (The detailed descriptions on the datasets are provided in the Supplementary Material.)

5.3 Performance Evaluation

In Tables 3 - 5, we report the discriminative performance of the various survival models for the MAGGIC, SUPPORT, and METABRIC datasets at three different time horizons, representing the 25%, 50%, and 75%-quantiles of time-to-event. (In view of space constraints, we provide the discriminative performances for the other datasets in the Supplementary Material.) We emphasize that the time horizons used for testing are different from the time horizons that are used in the construction of Survival Quilts, so we are not prejudicing the evaluations in our favor.

[1] Available at `https://www.unos.org/data/`

Table 3: C-index (mean±std) for the MAGGIC dataset at different time horizons. Blue highlighting indicates that the Brier Score constraints are satisfied.

Models	Time-Horizons (quantiles)		
	25%	50%	75%
Best benchmark	RSF	RSF	RSF
Cox	0.709±0.01*	0.694±0.02*	0.679±0.01*
CoxRidge	0.711±0.01*	0.695±0.02*	0.679±0.01*
Weibull	0.710±0.01*	0.695±0.02*	0.679±0.01*
LogNormal	0.719±0.02*	0.699±0.01*	0.676±0.01*
Exponential	0.708±0.02*	0.695±0.02*	0.679±0.01*
CoxBoost	0.707±0.02*	0.689±0.02*	0.672±0.01*
RSF	0.755±0.02	0.725±0.01	0.692±0.01†
CISF	0.740±0.02	0.708±0.01*	0.683±0.01*
Survival Quilts			
exog. $K=1$	0.761±0.02	0.730±0.01	0.701±0.00
exog. $K=2$	0.759±0.02	0.731±0.01	0.702±0.00
exog. $K=3$	0.758±0.02	0.731±0.01	0.702±0.00
endogenous	**0.764±0.02**	**0.735±0.01**	**0.705±0.00**

∗ indicates p-value < 0.01
† indicates p-value < 0.05

Table 4: C-index (mean±std) for the SUPPORT dataset at different time horizons. Blue highlighting indicates that the Brier Score constraints are satisfied.

Models	Time-Horizons (quantiles)		
	25%	50%	75%
Best benchmark	RSF	CISF	CISF
Cox	0.786±0.01*	0.750±0.01*	0.726±0.01*
CoxRidge	0.786±0.01*	0.750±0.01*	0.727±0.01*
Weibull	0.778±0.01*	0.745±0.01*	0.724±0.01*
LogNormal	0.797±0.01*	0.759±0.01*	0.731±0.01†
Exponential	0.772±0.01*	0.742±0.01*	0.722±0.01*
CoxBoost	0.785±0.01*	0.745±0.01*	0.719±0.01*
RSF	0.849±0.02	0.784±0.01	0.740±0.01
CISF	0.847±0.02	0.787±0.01	0.741±0.01
Survival Quilts			
exog. $K=1$	0.842±0.02	0.782±0.01	0.743±0.01
exog. $K=2$	0.843±0.02	0.781±0.01	0.742±0.01
exog. $K=3$	0.846±0.01	0.784±0.01	0.743±0.01
endogenous	**0.851±0.02**	**0.789±0.01**	**0.750±0.01**

∗ indicates p-value < 0.01
† indicates p-value < 0.05

Table 5: C-index (mean±std) for the METABRIC dataset at different time horizons. Blue highlighting indicates that the Brier Score constraints are satisfied.

Models	Time-Horizons (quantiles)		
	25%	50%	75%
Best benchmark	CISF	RSF	CISF
Cox	0.663±0.02*	0.676±0.01*	0.669±0.01†
CoxRidge	0.674±0.03*	0.682±0.01*	0.674±0.01†
Weibull	0.660±0.02*	0.673±0.01*	0.668±0.01*
LogNormal	0.679±0.02*	0.686±0.01*	0.673±0.01†
Exponential	0.661±0.02*	0.674±0.01*	0.670±0.01†
CoxBoost	0.674±0.03*	0.676±0.01*	0.668±0.01*
RSF	0.757±0.04	0.741±0.03	0.694±0.02
CISF	0.758±0.02	0.739±0.01	0.698±0.01
Survival Quilts			
exog. $K=1$	0.753±0.03	0.739±0.02	0.698±0.02
exog. $K=2$	0.752±0.03	0.740±0.02	0.698±0.02
exog. $K=3$	0.752±0.04	0.739±0.02	0.693±0.02
endogenous	**0.761±0.03**	**0.744±0.02**	**0.701±0.02**

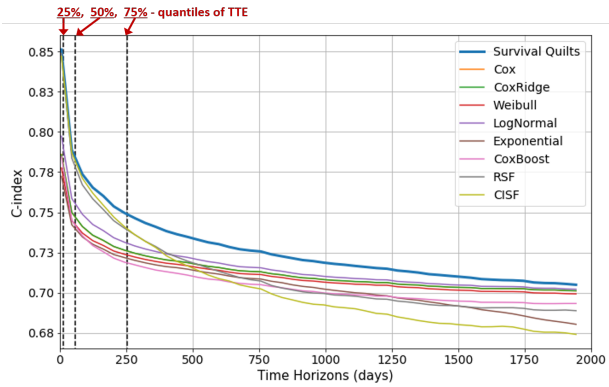
∗ indicates p-value < 0.01
† indicates p-value < 0.05

Overall, several things are important to note: i) the best performing benchmarks are *different* across the datasets and time horizons, ii) not all of the benchmarks satisfy the Brier Score constraints; i.e., they are not sufficiently well-calibrated, iii) in most cases the performance of Survival Quilts is better than that of the best benchmark, and the improvement is statistically significant over most of the benchmarks, and iv) in some cases (i.e., the UNOS-II and BPD datasets), the performance of Survival Quilts coincides with the best benchmark because it gives full weight to that benchmark.

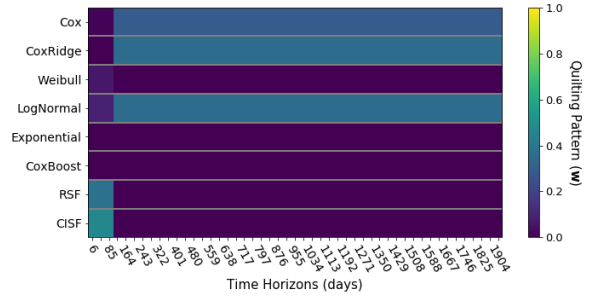
5.3.1 Endogenous Time-Horizon Splits

To illustrate the impact of choosing the quilting patterns endogenously, we call attention to Figure 3. The discriminative performance of RSF and CISF usually decreases at longer time horizons. In large part this is because RSF and CISF are nonparametric models and do less well over time horizons in which the number of patients at risk and the number of events are smaller. In contrast, the discriminative performances of the (semi-)parametric models decrease less over longer time horizons. Because our method constructs quilting patterns that change over time, it is able to give greater weight to models whose increments of risk predictions provide good discriminative performance in different time horizons. For example, in the SUPPORT dataset, the weights on RSF and CISF decrease and the weights on the Cox, CoxRidge and LogNormal increase at around $t = 100$ because the performance of RSF and CISF degrade earlier and more abruptly compared to that of Cox, CoxRidge, and LogNormal. (Corresponding figures for the other datasets are in the Supplementary Materials.)

Tables 3 - 5 compare the performance of Survival Quilts against the benchmarks at three time horizons. To highlight the gain achieved by our endogenous construction, we also provide the performance of Survival Quilts constructed using *exogenous* time horizons. When $K = 1$, we are using weights that do not vary with time as an alternative of the time-independent stacking [15]; for $K = 2, 3$, we have chosen exogenous time horizons with very coarse grids. As seen in the tables, the endogenous construction of Survival Quilts provides the best performance because it chooses the time intervals endogenously and allows for different weights in different time intervals. In the tables, we highlight in blue the results for models and time horizons in which the Brier Score constraints are satisfied; note that satisfaction of the constraints changes over different horizons. (The values for the Brier Scores are provided in the Supplementary Material.) Asterisks and daggers indicate that the performance improvements of Survival Quilts are statistically significant at the 0.01 and 0.05 levels, respectively.



(a) Discriminative performance



(b) Quilting Pattern

Figure 3: Discriminative performance and quilting patterns over time for the SUPPORT dataset. The dotted black lines depict the 25%, 50%, and 75%-quantiles of time-to-event.

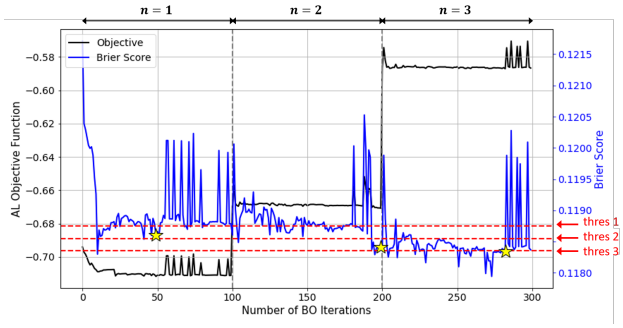


Figure 4: Illustration of change in the augmented Lagrangian objective (10) and Brier Score (g) with respect to the number of BO optimization iterations. The stars mark the minimal point of the objective for each subproblem. We set the maximum number of subproblems, $n_{\max}$, and the number of BO steps to 3 and 100, respectively. The constrained BO is solved at the time horizon t_1 for the MAGGIC dataset.

Table 6: Optimal weights, $\mathbf{w}_1^\dagger$ with varying Brier Score constraints in Figure 4.

Models	Brier Score Constraint		
	thres 1	*thres* 2	*thres* 3
Cox	0.07	0	0
CoxRidge	0.06	0.04	0
Weibull	0	0.14	0.15
LogNormal	0	0.21	0.20
Exponential	0.19	0	0
CoxBoost	0.02	0	0
RSF	0.35	0.42	0.44
CISF	0.31	0.19	0.21

5.4 Effect of Constrained BO

In this subsection, we address the effect of using constrained BO and how the optimal weight vector, $\mathbf{w}_k^\dagger$, changes as the number of BO iterations increases. Figure 4 illustrates the change in augmented Lagrangian objective in (10) and the change in time-dependent Brier-Score, g_k, with setting $k = 1$ for the MAGGIC dataset. As seen in the figure, if a strict constraint c is chosen (e.g., $c = thres\ 3$ in the figure), the optimal weights for the first two subproblems of (10) do not satisfy the Brier Score constraint. Thus, our BO solves the next subproblem with updated λ and ρ, which in turn gives more weight to the calibration performance than in the previous subproblems. In this example, the optimal weight of the third subproblem satisfies the Brier Score constraint and, thus, is selected as $\mathbf{w}_1^\dagger$.

Table 6 shows the optimal weight vector $\mathbf{w}_k^\dagger$ that is chosen as we set a stricter constraint as illustrated in Figure 4. As seen in the table, Survival Quilts puts more weights on the Cox-PH based methods (Cox, CoxRidge, and CoxBoost), and Exponential when the constraint in (10) is less strict. However, with stricter constraints, our method reduces weights on the Cox-PH based methods and Exponential, and instead assigns higher weights on Weibull and LogNormal.

6 Conclusion

This paper offers a novel approach to survival analysis that creates time-varying ensembles of existing survival models that we call Survival Quilts. Survival Quilts exploit existing models by giving them greater weight in time intervals where these models provide better incremental performance and lesser weight in time intervals where these models provide less good incremental performance. The superiority of Survival Quilts over previous survival models is demonstrated over six real-world datasets. One of the virtues of our approach is that we can adapt to use other survival models as those become available and prove their value.

Acknowledgments

This research is supported by the Office of Naval Research (ONR) and the NSF (Grant number: ECCS1462245, ECCS1533983, and ECCS1407712).

References

[1] J. M. Hernández-Lobato, M. A. Gelbart, R. P. Adams, M. W. Hoffman, and Z. Ghahramani. A general framework for constrained bayesian optimization using information-based search. *Journal of Machine Learning Research*, 17:1–53, 2016.

[2] David R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.

[3] Greg Ridgeway. The state of boosting. *Computing Science and Statistics*, pages 172–181, 1999.

[4] Harald Binder and Martin Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9(1), 2008.

[5] Jared Katzman, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger. Deep survival: A deep cox proportional hazards network. *arXiv preprint arXiv:1606.00931*, 2016.

[6] T. Fernández, N. Rivera, and Y. W. Teh. Gaussian processes for survival analysis. *In Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*, 2016.

[7] Ahmed M. Alaa and Mihaela van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. *In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[8] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, September 2008.

[9] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

[10] Alexis Bellot and Mihaela van der Schaar. Boosted trees for risk prognosis. *In Proceedings of the 3st Machine Learning for Healthcare Conference (MLHC 2018)*, 2018.

[11] Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *In Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI 2018)*, 2018.

[12] Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *arXiv preprint arXiv:1708.04649*, 2017.

[13] Lars Kotthoff, Chris Thornton, Holger H. Hoos, Frank Hutter, and Kevin Leyton-Brown. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research*, 18(25):1–5, 2017.

[14] Ahmed M Alaa and Mihaela van der Schaar. Autoprognosis: Automated clinical prognostic modeling via bayesian optimization with structured kernel learning. *In Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

[15] Andrew Wey, John Connett, and Kyle Rudser. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, 16(3):537–549, February 2015.

[16] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[17] Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. High dimensional bayesian optimisation and bandits via additive models. *In Proceedings of the 32th International Conference on Machine Learning (ICML 2015)*, 2015.

[18] J. Nocedal and S. J. Wright. *Numerical Optimization, 2nd Edition*. Springer, 2006.

[19] R. B. Gramacy, G. A. Gray, S. Le Digabel, H. K.H. Lee, P. Ranjan, G. Wells, and S. M. Wild. Modeling an augmented lagrangian for blackbox constrained optimization. *Technometrics*, 58(1):1–11, 2016.

[20] Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, 2013.

[21] Ulla B. Mogensen, Hemant Ishwaran, and Thomas A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software*, 50(11), 2012.

[22] Chih M. Wong, Nathaniel M. Hawkins, Mark C. Petrie, Pardeep S. Jhund, Roy S. Gardner, Cono A. Ariti, Katrina K. Poppe, Nikki Earle, Gillian A. Whalley, Iain B. Squire, Robert N. Doughty, and John J.V. McMurray. Heart failure in younger patients: the meta-analysis global group in chronic heart failure (MAGGIC). *European Heart Journal*, 35(39):2714–2721, June 2014.

[23] William A. Knaus, Frank E. Harrell, Joanne Lynn, Lee Goldman, Russell S. Phillips, Alfred F. Connors, Neal V. Dawson, William J. Fulkerson, Robert M. Califf, Norman Desbiens, Peter Layde, Robert K. Oye, Paul E. Bellamy, Rosemarie B.

Hakim, and Douglas P. Wagner. The SUPPORT prognostic model. objective estimates of survival for seriously ill hospitalized adults. study to understand prognoses and preferences for outcomes and risks of treatments. *Annals of Internal Medicine*, 122(3):191–203, February 1995.

[24] Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346–352, June 2012.

[25] Joseph F. Hayes, Louise Marston, Kate Walters, John R. Geddes, Michael King, and David P. J. Osborn. Adverse renal, endocrine, hepatic, and metabolic events during maintenance mood stabilizer treatment for bipolar disorder: A population-based cohort study. *PLOS Medicine*, 13(8):e1002058, August 2016.