

RESEARCH ARTICLE

Regularized parametric survival modeling to improve risk prediction models

J. Hoogland^{1,2}  | T. P. A. Debray^{1,3}  | M. J. Crowther⁴  | R. D. Riley⁵  |
J. IntHout⁶  | J. B. Reitsma^{1,3}  | A. H. Zwinderman² 

¹Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

²Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Amsterdam, The Netherlands

³Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

⁴Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

⁵School for Medicine, Keele University, Keele, Staffordshire, UK

⁶Radboud Institute for Health Sciences (RIHS), Radboud University Medical Center, Nijmegen, The Netherlands

Correspondence

Jeroen Hoogland, Department of Epidemiology and Data Science, Amsterdam University Medical Centers, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands.

Email: j.hoogland@amsterdamumc.nl

Funding information

ZonMw, Grant/Award Numbers: 91215058, 91617050



This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

We propose to combine the benefits of flexible parametric survival modeling and regularization to improve risk prediction modeling in the context of time-to-event data. Thereto, we introduce ridge, lasso, elastic net, and group lasso penalties for both log hazard and log cumulative hazard models. The log (cumulative) hazard in these models is represented by a flexible function of time that may depend on the covariates (i.e., covariate effects may be time-varying). We show that the optimization problem for the proposed models can be formulated as a convex optimization problem and provide a user-friendly R implementation for model fitting and penalty parameter selection based on cross-validation. Simulation study results show the advantage of regularization in terms of increased out-of-sample prediction accuracy and improved calibration and discrimination of predicted survival probabilities, especially when sample size was relatively small with respect to model complexity. An applied example illustrates the proposed methods. In summary, our work provides both a foundation for and an easily accessible implementation of regularized parametric survival modeling and suggests that it improves out-of-sample prediction performance.

KEYWORDS

convex optimization, penalized maximum likelihood, prediction, regularization, survival analysis

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

1 | INTRODUCTION

The estimation of individualized survival probabilities is often of key interest in medical and biostatistical research (Harrell, 2015; Steyerberg, 2009). A suitable prediction model for this task describes survival probabilities as a function of time and the covariates of interest. In this context, fully parametric models provide a very direct and possibly parsimonious means to obtain predicted survival curves over time (Crowther & Lambert, 2014; Royston and Parmar, 2002). With respect to the ubiquitous semiparametric Cox model (Cox, 1975, 1972), note that the way in which it elegantly avoids estimation of the baseline hazard is not a feature in this context: the baseline hazard is of key interest to obtain predicted survival probabilities.¹

A particularly flexible class of parametric survival models uses splines to model time and was introduced by Royston and Parmar (2002), considerably increasing flexibility beyond classical parametric families (e.g., Weibull models), while still providing smooth survival estimates over time. Software implementations for Royston–Parmar model implementation are readily available (e.g., `stpm2` (Lambert, 2010) in Stata (StataCorp, 2021) and `rstpm2` (Liu et al., 2018) in R (R Core Team, 2021)). Nonetheless, none of these implementations provides the means for regularization (such as ridge regression (Hastie, 2020; Hoerl & Kennard, 1970), lasso regression (Tibshirani, 1996), or elastic net regression (Friedman et al., 2010)), while this has proven to be an important tool in prediction modeling to improve out-of-sample prediction accuracy (Hastie et al., 2017, 2015).

In this paper, we introduce such regularization methods for flexible parametric survival models with possibly time-varying covariate effects. The main aim is to improve out-of-sample accuracy of predicted survival probabilities over time in settings where sample size is limited with respect to model complexity. More specifically, we focus on models that are multiplicative on the hazard scale (like the Cox model) or cumulative hazard scale (like the typical Royston–Parmar model). The use of regularization methods with such models is nontrivial due to the presence of constrained functions of time described by splines (e.g., the hazard and cumulative hazard function) and possible interactions with this function (i.e., TV covariate effects). We provide a unified regularization approach for both log hazard and log cumulative hazard models. A software implementation is made available in R package `regsurv`.

2 | LOG (CUMULATIVE) HAZARD MODELS

Let $h(t|\mathbf{Z})$ be the hazard at time t , conditional on an $n \times p$ covariate matrix \mathbf{Z} (for n subjects and p covariates) with the corresponding coefficient vector $\boldsymbol{\beta}$ expressing the log hazard ratios. Furthermore, let $H(t|\mathbf{Z})$ denote the corresponding cumulative hazard. A *proportional* hazards (PH) model for either $\ln h(t|\mathbf{Z})$ or $\ln H(t|\mathbf{Z})$ can be written as

$$g(t) + \mathbf{Z}\boldsymbol{\beta}, \quad (1)$$

where $g(t)$ is a function of time describing the baseline (cumulative) hazard, and $\mathbf{Z}\boldsymbol{\beta}$ captures the proportional (time-constant) covariate effects. In the remainder, and in line with Royston and Parmar (2002), we use restricted cubic splines of log time based on truncated power bases to model $g(t)$. The outer knots are taken to be the minimum and maximum of the observed event times, and a total of $m - 2$ inner knots are set to ordered quantiles of the distribution of event times. This leads to m restricted cubic spline basis functions v_j for $j \in \{1, \dots, m\}$ (details are provided in the online supporting material, Part A). For $u = \ln(t)$, some set of knots \mathbf{k} , and coefficients $\boldsymbol{\alpha}$, the transformation can be written as

$$s(u|\boldsymbol{\alpha}, \mathbf{k}) = \alpha_0 + \alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_m v_m. \quad (2)$$

Due to the nature of the truncated power bases, v_1 is always equal to u , and the subsequent basis functions are all nonlinear.

¹ While either the Breslow estimate (of the cumulative baseline hazard) (Lin, 2007) or the Kalbfleisch Prentice estimate (of baseline survival) allow for survival predictions, both of these estimates involve a large number of parameters and are computationally intensive when sample size is large and/or in the presence of time-dependent effects.

2.1 | TV covariate effects

TV (or nonproportional) covariate effects $\beta(t)$ can be included as covariate interactions with time. That is, either the $\ln h(t|\mathbf{Z})$ or $\ln H(t|\mathbf{Z})$ is modeled as

$$s(u|\boldsymbol{\alpha}, \mathbf{k}) + \mathbf{Z}\boldsymbol{\beta} + s(u, \mathbf{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa}), \quad (3)$$

where $s(u|\boldsymbol{\alpha}, \mathbf{k})$ and $\mathbf{Z}\boldsymbol{\beta}$ are defined as in Equations (1) and (2), and $s(u, \mathbf{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ denotes the interaction of restricted cubic spline basis functions of u with covariate matrix \mathbf{Z}_I , where I is the subset of covariates for which a time-dependent effect is incorporated, $\boldsymbol{\kappa}$ denotes the knots for the spline of time, and $\boldsymbol{\gamma}$ denotes the corresponding coefficients. For example, when two continuous covariates each interact with a restricted cubic spline representation of time with one interior knot, $s(u, \mathbf{Z}_I|\boldsymbol{\gamma}, \boldsymbol{\kappa})$ can be written as

$$s(u, \mathbf{Z}_{1,2}|\boldsymbol{\gamma}, \boldsymbol{\kappa}) = \gamma_{1,1}v_1Z_1 + \gamma_{1,2}v_2Z_1 + \gamma_{2,1}v_1Z_2 + \gamma_{2,2}v_2Z_2,$$

where the γ subscripts index the covariates and spline basis functions, respectively. Note that $\boldsymbol{\kappa}$ (the set of knots for interactions with time) may differ from \mathbf{k} (the set of knots for the log cumulative baseline hazard) to allow for interactions with time that are less (or more) granular than the model for the baseline hazard.² For ease of reference, note that the parameters in Equation (3) are grouped in baseline (cumulative) hazard parameters $\boldsymbol{\alpha}$, main (proportional) effect parameters $\boldsymbol{\beta}$, and parameters relating to TV (nonproportional) effects $\boldsymbol{\gamma}$. Note that while the proposed model in Equation (3) can be used to model either the log hazard or the log cumulative hazard scale, the interpretation of the model coefficients of course strongly depends on the chosen scale.

2.2 | Log-likelihood

Writing $\boldsymbol{\theta} = (\boldsymbol{\alpha} : \boldsymbol{\beta} : \boldsymbol{\gamma})$, the general form of the log-likelihood is

$$l(\boldsymbol{\theta}) = \boldsymbol{\delta} \ln h(t|\mathbf{Z}) - H(t|\mathbf{Z}) \quad (4)$$

with $\boldsymbol{\delta}$ the vector of event indicators taking value 0 for (right) censored cases and 1 for events. For log cumulative hazard models, $l(\boldsymbol{\theta})$ is available in closed form. For log hazard models, numerical integration is needed to approximate $H(t|\mathbf{Z})$, which was performed by means of Gauss–Legendre quadrature (Bower et al., 2016; Crowther & Lambert, 2014; Novomestky, 2013). Details on the log-likelihood contributions for both types of models are available in the online supporting material, Part B.

3 | REGULARIZATION

Regularization can be implemented by means of penalized maximum likelihood. We have implemented both an elastic net-type penalty and a group lasso penalty. Since the implemented penalties act on the size of the model coefficients, data are standardized to mean zero and standard deviation one by default.

3.1 | Elastic net

The elastic net penalty can be written as

$$P_{net}(\boldsymbol{\omega}, \boldsymbol{\theta}) = \lambda \sum_{d=1}^D \omega_d \phi_d |\theta_d| + \frac{1}{2} (1 - \omega_d) \phi_d \theta_d^2 \quad (5)$$

² In fact, $\boldsymbol{\kappa}$ could also be a matrix \mathbf{K}_I with different sets of knots per time-dependent covariate effect.

with global penalty scaling parameter λ scaling the weighted sum of regression coefficient-specific contributions to the penalty. The regression coefficient vector θ has elements $d \in 1, \dots, D$, corresponding parameter-specific penalty scaling factors $\phi_d \in [0, \infty)$, and mixing factors $\omega_d \in [0, 1]$ with extremes $\omega_d = 1$ being a lasso penalty and $\omega_d = 0$ a ridge penalty. Note that in contrast to the well-known and widely applied elastic net penalty in generalized linear models (Friedman et al., 2010), we allow for parameter-specific specification of the mixing factor (as opposed to a global choice). This allows the user to combine penalties that only shrink and penalties that may also remove coefficients from the model. This is especially relevant to the survival setting. For example, it allows one to choose ridge regression for baseline (cumulative) hazard parameters (to avoid selection of individual basis functions) and a penalty that also provides parameter selection for the remaining parts of the model. With respect to the penalty scale factors ϕ_d , note that $\phi_d = 0$ equals unpenalized θ_d and that $\phi_d = \infty$ leads to $\theta_d = 0$. Setting some elements of ϕ to zero could, for instance, be used to avoid penalization of the baseline hazard.

3.2 | Group lasso

We implemented a group lasso penalty that can be written as

$$P_{GL}(\omega, \theta) = \lambda \sum_{g=1}^G \omega_g \phi_g \|\theta_g\|_2 + \frac{1}{2} (1 - \omega_g) \phi_g \|\theta_g\|_2^2 \quad (6)$$

for partitions $g \in 1, \dots, G$ of θ . Note that in this case, mixing factors ω_g and penalty scaling factors ϕ_g relate to the norms of G partitions of θ denoted by θ_g . In addition to the usual group lasso formulation (e.g., Meier et al., 2008), and analogous to the elastic net penalty, the group lasso penalty in Equation (6) allows for group-specific ω_g , thus allowing some groups to follow a group lasso penalty and others to follow a ridge penalty. As for the elastic net case, this allows users to only shrink a subset of parameter-groups (ensuring that they stay in the model), while potentially also selecting among other groups of parameters (group lasso). Note that in the group lasso case, we restrict ω_g to take a value in $\{0, 1\}$, but this could be extended to the entire range $[0, 1]$.

3.3 | Survival-specific nuances

In ridge, lasso, and elastic net implementations for generalized linear models, it is standard practice to avoid penalization of the intercept by centering of both outcome y and the columns of design matrix X , and to allow penalization of the remaining parameters (Friedman et al., 2010; Hastie et al., 2017; Tibshirani, 1996). However, this strategy is not directly applicable in the case of parametric survival analysis. First, centering of the outcome is not possible, and the intercept therefore remains in the model and should be estimated. Our implementation treats the intercept as an unpenalized parameter (i.e., the scaling factor for the intercept penalty (ϕ_1) is always set to 0). Second, a log cumulative hazard model needs at least an intercept and a slope parameter to provide a sensible model. It is convenient that the first basis function of the implemented restricted cubic splines provides this slope in the form of a linear contribution of log time. Nonetheless, it may still be desirable to penalize the slope estimate. Thereto, log cumulative hazard models are estimated with a log time offset (i.e., slope equal to 1), effectively shrinking the slope parameter toward unity instead of zero. Consequently, the simplest model has an unpenalized intercept α_0 and a log time offset, which can be recognized as an exponential survival distribution with rate parameter e^{α_0} .

3.4 | Tuning parameter selection

The choice of tuning parameters ω and λ can be informed by a grid search using resampling such as cross-validation or bootstrapping. The log-likelihood or deviance can be used as a measure of out-of-sample performance (Meier et al., 2008).³ In the software (also refer Section 7), we have implemented k -fold cross-validation over a grid of λ for fixed ω . In short,

³ Note that the monotonicity constraints for log cumulative hazard models are only enforced for the covariate domain as reflected by the development data. In presence of many TV effects, extrapolation beyond this domain may lead to invalid (nonnegative) hazard estimates. Therefore, in the context of

the training data are split into k parts, and each part in turn serves as a holdout set where the performance of models fitted on the data not in part k are evaluated. For each holdout fold k , all combinations on the grid of λ for a given ω are evaluated and averaged in the end. This provides an estimate of the optimal λ value conditional on ω for the training data. In principle, if tuning of ω is also desired, repeated cross-validation runs for different choices of ω may provide further information. However, this is a computationally demanding task that comes with more uncertainty and may not lead to a unique optimal value (van Nee et al., 2023). In practice, it is our experience that the degree of sparsity is hard to estimate from the data. Preferably, the choice of ω is informed by some content knowledge about the expected degree of sparsity or by desired model characteristics with respect to sparsity (e.g., use a lasso penalty when sparsity is expected, such as when exploring many interactions, and use a ridge penalty when only shrinkage is expected to be required).

4 | OPTIMIZATION

The general optimization problem can be formulated as

$$\begin{aligned} & \text{maximize} && l_{pen}(\theta) = l(\theta) - P(\theta) \\ & \text{subject to} && h(\mathbf{u}|\theta, \mathbf{Z}) > \mathbf{0}, \end{aligned} \quad (7)$$

where $l(\theta)$ is the appropriate form of the log-likelihood in Equation (4) for either a log hazard or a log cumulative hazard model, $P(\theta)$ is either the elastic net penalty (Equation 5) or the group lasso penalty (Equation 6), and $h(\mathbf{u}|\theta, \mathbf{Z})$ denotes the hazard contributions. Note that for the latter, strict positivity could be relaxed to positivity except at event times. The online supporting material (Part C) shows the necessary objective functions and constraints can be written in an equivalent but convex form, such that convex optimization procedures can be used to find the global optimal value and corresponding solution(s) θ^* (for fixed values of ω and ϕ) (Boyd & Vandenberghe, 2015). Subsequently, efficient software is available for the optimization (Boyd & Vandenberghe, 2015; Domahidi et al., 2013) and is easily accessible by means of R package CVXR (Fu et al., 2020). More specifically, CVXR provides a user-friendly interface that transforms the standard convex programming form of the problem into a second-order cone program, that can subsequently be solved with interior-point solver ECOS (embedded conic solver) (Domahidi et al., 2013).

5 | SIMULATION STUDY

The main aim of the simulation study was to compare key survival modeling methods in settings that strike a balance between model complexity and sample size. The design and reporting of the simulation study adhere to the guidelines by Morris et al. (2019).

5.1 | Data-generating mechanism

We followed a proposal by Crowther and Lambert and simulated from a two-component Weibull mixture (Crowther & Lambert, 2013). The main motivation was to generate survival data that are sufficiently complex to resemble real data, and at the same time avoid that any of the models under evaluation contain the exact data-generating mechanism. Specifically, we sampled from a Weibull mixture distribution that was additive on the survival scale. Details on the derivation are available elsewhere (Crowther & Lambert, 2013), so we only restate the general form of the baseline hazard function

$$h_0(t) = \frac{\lambda_1 \gamma_1 t^{\gamma_1 - 1} p_{mix} e^{-\lambda_1 t^{\gamma_1}} + \lambda_2 \gamma_2 t^{\gamma_2 - 1} (1 - p_{mix}) e^{-\lambda_2 t^{\gamma_2}}}{p_{mix} e^{-\lambda_1 t^{\gamma_1}} + (1 - p_{mix}) e^{-\lambda_2 t^{\gamma_2}}}. \quad (8)$$

log cumulative hazard models, the default cross-validation implementation optimizes the objective function $l_{pen}(\theta)$ in the selected cases, while enforcing the nonnegative hazards constraint in the whole sample.

The Weibull mixture parameters were setup such that they describe a nonmonotone hazard function that first increases and subsequently decreases, such as might occur after an intervention with hazardous early side effects (e.g., difficult surgery) or in cancer studies (Crowther & Lambert, 2013). This nonmonotone pattern was chosen since it requires a fairly flexible model, and hence does not clearly favor simple parametric representations. The particular choice of parameters for the Weibull mixture was $\lambda_1 = 0.21$, $\lambda_2 = 0.05$, $\gamma_1 = 1.1$, $\gamma_2 = 1.4$, and $p_{mix} = 0.4$, and cases were censored administratively at time $t = 30$. For ease of reference, we will further refer to time units in months. In terms of covariates, the main aim was to simulate covariates with varying effect sizes and time-dependencies, as might be encountered in practice. Due to the complexity associated with the modeling of TV effects, a setting was chosen with the number of covariates considerably smaller than the number of events. Thereto, 11 covariates were simulated from a multivariate standard normal distribution with pairwise correlations set to 0.25. The particular choice of parameters for these 11 main effects was 0, 0, 0.5, -0.5, 0.25, -0.25, 0.125, -0.125, 0.0625, -0.0625, and 0.5, respectively. The effects of the first three covariates varied with time according to 0.9^t , with coefficients -1, 0.75, and -0.5. Combining the time-constant and TV effects, the log hazard ratio of the first and second covariates started at -1 and 0.75, respectively, and diminished over time, and the log hazard ratio for the third covariate started at 0 and its effect increased over time to 0.5. The online supporting material (Part D) visualizes the baseline hazard and TV effects corresponding to the data-generating mechanism. A population of 110,000 cases was generated from this data-generating mechanism. A fixed set of 10,000 was set aside as an independent validation cohort. The data for model development, also known as the training or discovery data, were sampled from the remaining 100,000 cases.

In addition to the main simulation setting, two additional settings were evaluated. First, to evaluate performance under settings with more covariates, 20 standard normal noise variables were added to the above described data-generating mechanism. Second, to evaluate performance under increased censoring, exponential censoring was added to the above described data-generating mechanism to arrive at 50% censoring.

5.2 | Simulation settings

For the main simulation settings, a total of 1000 simulation runs was performed for four development sample size settings: 100, 250, 500, and 1000. For the additional simulation settings with more covariates and with higher censoring, the $N = 100$ sample size was omitted since it was too small to evaluate all methods. In each simulation run, all survival models were fitted on the development sample and evaluated in the independent validation sample. To emulate realistic settings, where not all covariates that are relevant to the problem at hand are known and/or measured, the 11th covariate was considered to be unmeasured for all modeling methods (and thus not included in the models). The main motivation was to avoid a comparison of methods under near-perfect model specification that could not be expected in real data.

5.3 | Survival modeling methods

Ten different modeling techniques were compared:

1. Regularized log hazard model including time-varying effects (RegHazTV): regularized log hazard models with the log baseline modeled with a restricted cubic spline with 5 degrees of freedom (df), 10 linear main effects (i.e., for each measured covariate), and including interactions with log time by means of a 2 degrees of freedom restricted cubic spline for all 10 covariates. The log baseline hazard and main-effects parameters were penalized with a ridge penalty, and the TV effects with a group lasso penalty with separate groups for each covariate. With respect to the TV effects (i.e., interactions with spline basis functions), the group lasso penalty ensures that coefficients belonging to the same spline transformation are simultaneously zero or nonzero.
2. Regularized log cumulative hazard model including time-varying effects (RegCumHazTV): same as (1), but on the log cumulative hazard scale. NB: the interactions with time are therefore also on the log cumulative hazard scale and hence differ from the specification in (1).
3. Cox proportional hazards model (CoxPH): a CoxPH model with 10 linear main effects. Predicted survival was derived based on the Cox model and the corresponding Breslow estimate of the cumulative baseline hazard.
4. Cox model including time-varying effects (CoxTV): same as (3), but allowing for TV effects as a function of a 2 degrees of freedom restricted cubic spline of log time. To encode these TV effects, the data set was transformed into start-

stop format with splits at all percentiles of the observed event times and subsequent derivation of the covariate-time interaction columns (Therneau & Grambsch, 2011).

5. Cox proportional hazards lasso model (CoxPHlasso): same as (3), but with a lasso penalty on all parameters.
6. Cox time-varying effects ridge model (CoxTVridge): same as (4), but with a ridge penalty on all parameters. Note that a regular lasso penalty is not directly applicable due to the presence of spline components.
7. Royston–Parmar proportional hazards model (RPrcsPH): a proportional Royston–Parmar model (i.e., log cumulative hazard model) as implemented by (Liu et al., 2018), with a 5 degrees of freedom natural cubic spline for the log cumulative baseline hazard and 10 linear main covariate effects.
8. Royston–Parmar time-varying effects model (RPrcsTV): same as (7), but including interactions with log time by means of a 2 degrees of freedom restricted cubic spline for all 10 covariates.
9. Royston–Parmar proportional hazards model (RPssPH): a proportional Royston–Parmar model (i.e., log cumulative hazard model) as implemented by Liu et al. (2018), with a smoothing spline for the log cumulative baseline hazard and 10 linear main covariate effects.
10. Royston–Parmar time-varying effects model (RPssTV): same as (9), but including interactions with log time by means of a smoothing spline for all 10 covariates.

Certain groups of methods can be distinguished within these 10 methods. For instance, we will refer to methods 1, 2, 4, 6, 8, and 10 as methods that allow for TV effects, and to the complementary set of methods 3, 5, 7, and 9 as PH methods. In addition, the methods can be grouped into methods that incorporate regularization on the size of the model parameters (methods 1, 2, 5, and 6) and methods that do not (methods 3, 4, and 7–10).⁴ For all regularized models, for a fair comparison, the optimal value of the penalty parameter λ was estimated by means of 10-fold cross-validation minimizing the deviance. In practice, note that while k -fold cross-validation already uses all the available data, the uncertainty of the cross-validation estimates may be further reduced by repeated k -fold cross-validation.

In the additional simulation settings with added noise variables, these noise variables were added as main effects. In the additional simulation settings with increased censoring, all methods models were specified exactly as for the main simulation settings.

5.4 | Performance measures

In the simulation study setting, the *true* survival probabilities are known for all individuals across all time points. For further reference, we denote the true survival probability at time t for individual i as $p_i(t)$, and its estimate as $\hat{p}_i(t)$ (omitting dependence on covariate vector \mathbf{x}_i from the notation). Root mean squared prediction error (rMSPE) was evaluated as a measure of prediction accuracy and was defined as

$$\text{rMSPE}^s = \sqrt{\frac{1}{n_{sim}} \sum_i^{n_{sim}} (p_i(t) - \hat{p}_i(t))^2}, \quad (9)$$

where rMSPE^s is the rMSPE for simulation run s and n_{sim} is the number of cases in the validation data set. To emphasize performance at particular time points, rMSPE was evaluated with t set to each of the time points 2.5, 5, 7.5, 10, 20, and 30 months for all individuals. As a measure of overall prediction performance, rMSPE was evaluated with t_i set to the observed event times for the cases in the validation data set.

Likewise, both fixed time-point and time-averaged discriminative prediction performance were evaluated against the *true* survival probabilities by means of Harrell's C-statistic (Harrell, 2015; Harrell et al., 1996), which for our purposes can be defined as

$$\text{C-statistic}^s = \frac{\sum_i \sum_{j \neq i} \left[I(\hat{p}_i(t) < \hat{p}_j(t)) I(p(t)_i < p(t)_j) + \frac{1}{2} I(\hat{p}_i(t) = \hat{p}_j(t)) I(p(t)_i < p(t)_j) \right]}{\sum_i \sum_{j \neq i} [I(p(t)_i < p(t)_j)]}, \quad (10)$$

⁴ Note that while the smoothing splines in methods 9 and 10 do use penalization, they effectively penalize nonlinear covariate contributions toward linearity, as opposed to penalizing coefficient size.

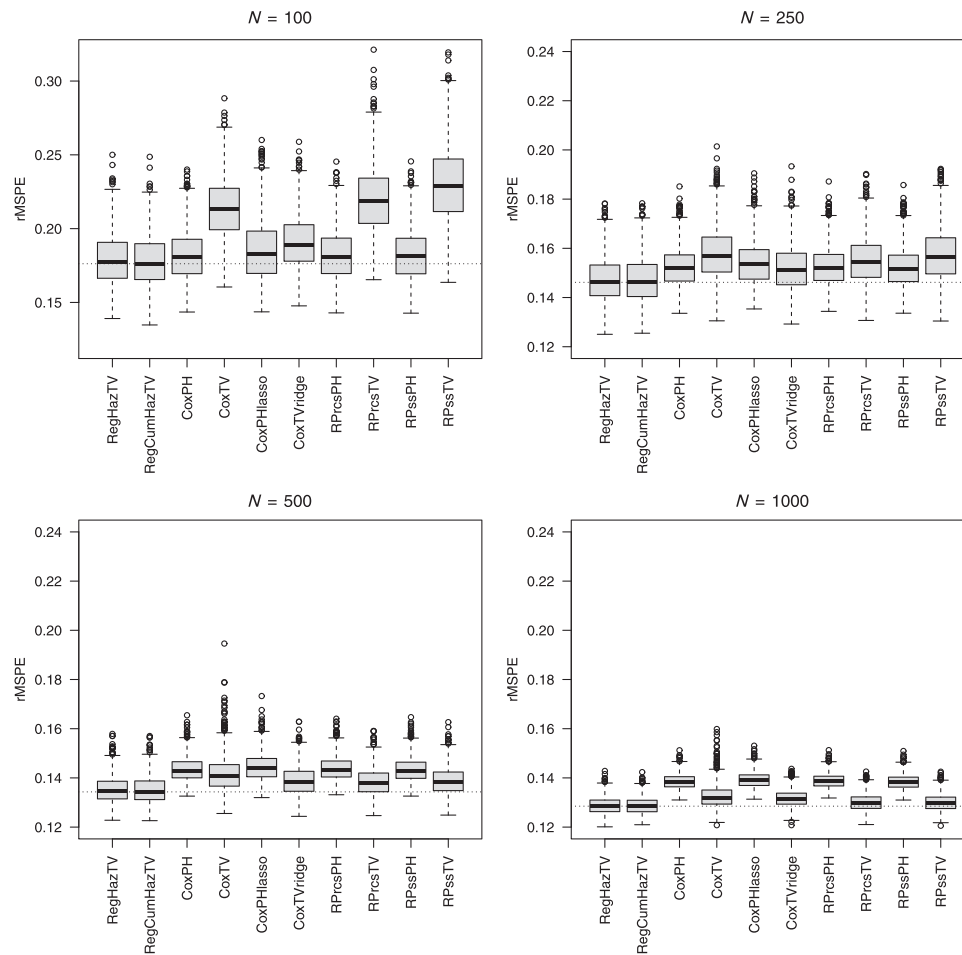


FIGURE 1 Boxplots of root mean squared prediction error (rMSPE) for each of the evaluated methods. Boxes cover the interquartile range and have a solid bar showing the median; whiskers extend to 1.5 times the interquartile range. The horizontal dotted line crosses the median rMSPE for the best performing method in a specific sample size setting. NB: The scale of the y-axis differs between the subfigure for $N = 100$ and the other subfigures to enhance visual clarity of the differences within scenarios.

where C-statistic^s is the C-statistic for simulation run s , i and j index the individuals $1, \dots, n_{sim}$ in the independent validation data, and $I(\cdot)$ is the indicator function. As for rMSPE, performance was evaluated at fixed time points by setting t to each of the time points 2.5, 5, 7.5, 10, 20, and 30 months, and as an overall average by setting t to the subject-specific observed event times in the validation data set for each individual.

Lastly, calibration performance was assessed using graphical calibration curves as proposed by Austin et al. (2020) at the same fixed time points. In short, calibration aims to evaluate whether the predicted survival probabilities for a particular time points are close to the true probabilities as evaluated over the whole range of predicted probabilities. The typical graph shows predicted versus “observed” probabilities, with perfect calibration corresponding to a straight line through the origin with slope equal to 1. Since survival probabilities cannot be directly observed in practice, the right-censored time-to-event data in the validation set are modeled as a flexible function of the complementary log–log of the predicted cumulative incidence for each validation case, at a fixed time t , using hazard regression. Details can be found in appendix B of Austin et al. (2020) and an implementation is available in R function `calibrate()` in our supplementary material.

5.5 | Simulation study results

Figure 1 shows that the time-averaged rMSPE of the proposed regularized models (RegHazTV and RegCumHazTV) were among the best performing methods in all sample size settings. In the smallest sample size setting ($N = 100$), prediction

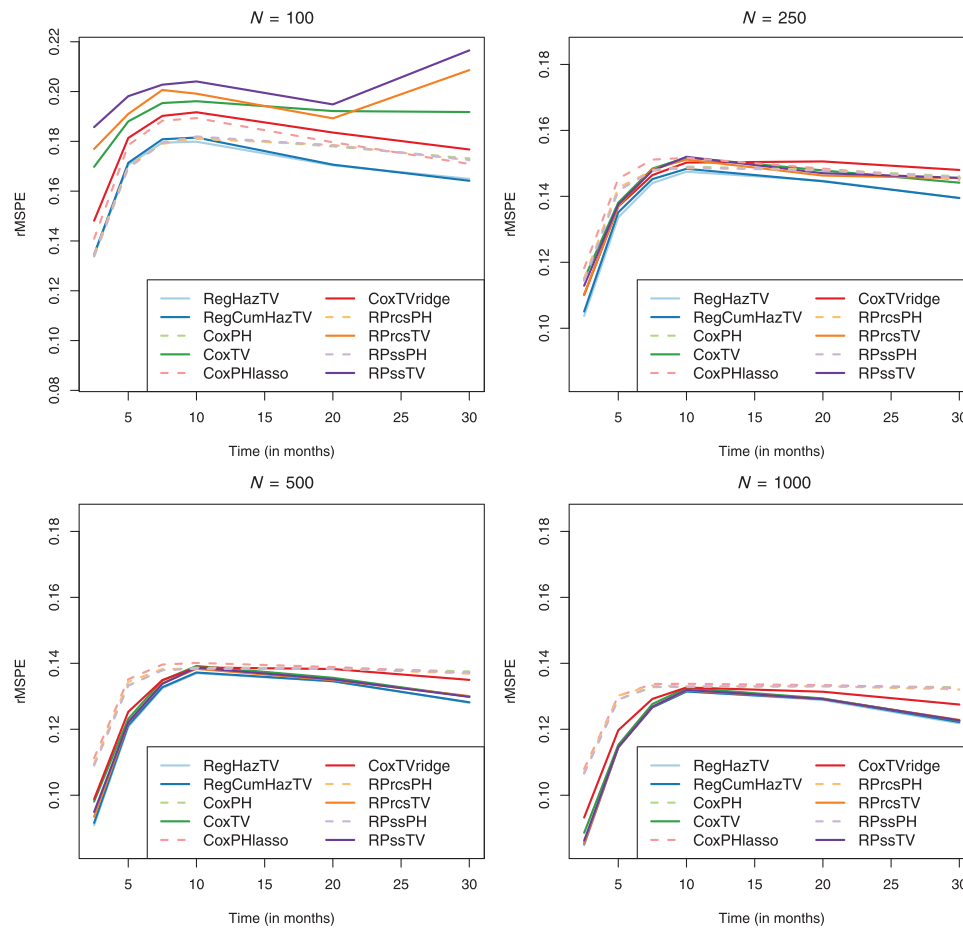


FIGURE 2 Root mean squared prediction error (rMSPE) over time each of the evaluated methods. Note that the line for Cox proportional hazards model (CoxPH) is hardly visible since its curve is almost identical to the curves of Royston–Parmar proportional hazards model (RPrCsPH) and RPssPH. Solid lines are for models allowing for time-varying coefficients; dashed lines for proportional hazards models. NB: The scales of the y-axis differs between the subfigures to enhance visual clarity of the differences within scenarios; rMSPE decreases with increasing sample size.

accuracy of RegHazTV and RegCumHazTV was comparable to modeling methods assuming PH, and clearly outperformed other TV effects methods. With increasing sample size ($N = 250$), the other TV effects methods start to catch up with the PH models. Further increase in sample size ($N = 500$) shows that the TV effects models start to more fully capture the data-generating mechanism generally and overcomes their tendency to overfit: all of the TV effects models outperform the PH methods. The final increase in sample size up to $N = 1000$ shows that the nonregularized time-dependent effects models (CoxTV, RPrCsTV, and RPssTV) start to catch up with the proposed regularized models.

While time-averaged rMSPE provides a summary measure of accuracy, in practical applications there is often interest in the prediction accuracy for particular clinically relevant time points (McLernon et al., 2023). Figure 2 shows rMSPE results at each of the evaluated fixed time points. In line with the time-average results, the proposed regularized parametric methods performed well across all sample size settings. Their benefit was most apparent for later prediction times. As was the case for the time-averaged results, the PH methods were at a clear disadvantage in large sample size settings (due to misspecification), but had the edge over nonregularized TV methods in the smallest sample size setting (due to decreased risk of overfitting). The benefit of modeling TV coefficients strongly depends on the available sample size due to the increase in model complexity. Furthermore, the gain in precision of course depends on the underlying data-generating process, where the amount of deviation of the TV effects from a time-constant approximation will determine the gain in precision when accounting for this. Importantly, this may depend on the time point of interest for prediction purposes. As an illustration, across all of the simulation settings, the agreement between PH and TV models is best for predictions around 10 months, which is where a time-constant approximation of the TV effects is close to the true value.

The online supporting material shows discriminative performance (Part E) and calibration performance (Part F) in the validation data for all particular fixed time points of interest. For discrimination, time-averaged results are also provided. Results were in line with the rMSPE results, with RegHazTV and RegCumHazTV consistently performing well in terms of both time-averaged discriminative performance and discriminative performance at each of the fixed time points. With respect to calibration, average calibration curves across simulations converged toward the diagonal with increasing sample size for all TV effects methods except for CoxTVridge. Average calibration curves for RegHazTV and RegCumHazTV did so faster than other TV effects methods. PH-method curves clearly reflected misspecification in the larger sample size settings, especially for early and late time points. The latter was according to expectation since the trends of the true TV effects over time were all monotone and thereby cross the time-constant (PH) approximation that captures the average across time, resulting in good approximations near the crossing point and bad approximations further away.

With respect to computation times, online supplementary Figure G.1 shows the computation times for each of the methods. Most methods were fast across all settings, providing results within seconds. The exceptions are RegCumHazTV, RPsTV, CoxTVRidge, and RegHazTV, needing up to a median of 2, 7, 10, and 28, min, respectively, for a single run (including tuning parameter selection) in the $N = 1000$ setting.

Simulation results with respect to the two additional groups of settings are provided in the online supplementary material (Part H). For settings with 20 added noise variables, the benefit of regularization was more apparent than in the main simulation settings, with the proposed methods performing better than the alternatives in terms of fixed time-point or time-averaged rMSPE, discrimination, and calibration. Patterns in the results were similar to the main results but exaggerated. In settings with increased censoring, results were very similar to the main simulation study results, with more marginal and sample size-dependent benefit of the proposed regularization methods. An interesting difference with the main results can be seen in the fixed time-point accuracy (rMSPE) results, where prediction error tends to increase with the prediction horizon for all methods. This relates to the decrease in information in the data over time due to increased censoring. In line, early prediction errors are very similar between methods, and differences become more apparent at later prediction times.

6 | VETERANS' ADMINISTRATION LUNG CANCER (VALC) STUDY

The VALC study is a randomized trial of two chemotherapy treatments in males with advanced inoperable lung cancer (Kalbfleisch & Prentice, 2002). The primary endpoint was time to death and 128 of 137 patients died during follow-up. Data on a selection of variables are available in Kalbfleisch and Prentice (2002) and include time-to-event, event status, and data on treatment assignment (standard vs. new chemotherapy), age (in years), prior therapy (yes/no), histological type (squamous, small cell, adeno, and large cell), performance status (Karnofsky rating from 0 to 100, with higher scores relating to better status), and time between diagnosis and randomization (in months). A CoxPH model including all of these measures as main effects shows signs of nonproportionality based on the Grambsch and Therneau test on Schoenfeld residuals (Therneau & Grambsch, 2011) ($p = 3.2e^{-5}$), with individual contributions of cell type ($\chi^2_3 = 15.2$, $p = 0.0016$) and Karnofsky rating ($\chi^2_3 = 12.9$, $p = 0.0003$). This provides us with an interesting setting to illustrate the methods as applied in the simulation study in suitable variations for this applied example (note that model specification was not informed by the ad hoc nonproportionality test):

1. RegHazTV, with the log baseline modeled with a restricted cubic spline with 4 degrees of freedom, all main effects, and including linear interactions with log time. The log baseline hazard parameters were penalized with a ridge penalty, and the remaining parameters with a lasso penalty (group lasso in case of cell type which had three groups).
2. RegCumHazTV: same as (1), but on the log cumulative hazard scale.
3. A CoxPH model with all main effects. Predicted survival was derived based on the Breslow estimate of the cumulative baseline hazard.
4. A Cox model similar to (3), but including time-varying effects (CoxTV) as a linear function of log time.
5. Cox proportional hazards ridge model (CoxPHridge): same as (3), but with a ridge penalty on all parameters. Ridge was preferred over lasso due to presence of a categorical variable with three groups (cell type).
6. CoxTVridge: same as (4), but with a ridge penalty on all parameters.
7. RPrCsPH: a proportional log cumulative hazard model with a 4 degrees of freedom natural cubic spline for the log cumulative baseline hazard and all main covariate effects.
8. RPrCsTV: same as (7), but including interactions with log time for all 10 covariates.

TABLE 1 Mean and standard error of the time-average C^{td} (time-dependent C-statistic) and the C^{td} up to 60-day follow-up (censoring event times $> t = 60$ days) are shown as derived based on 1000 out-of-bag estimates for the Veterans' Administration Lung Cancer study.

	Average C^{td} (se)	60-day C^{td} (se)
RegHazTV	0.698 (0.044)	0.742 (0.043)
RegCumHazTV	0.694 (0.047)	0.741 (0.043)
CoxPH	0.705 (0.035)	0.740 (0.044)
CoxTV	0.683 (0.057)	0.736 (0.045)
CoxPHridge	0.709 (0.034)	0.744 (0.043)
RPrCsPH	0.706 (0.035)	0.740 (0.044)
RPrCsTV	0.693 (0.043)	0.736 (0.046)
RPssPH	0.705 (0.035)	0.739 (0.045)
RPssTV	0.677 (0.060)	0.732 (0.046)
RegHazPH	0.713 (0.035)	0.752 (0.042)
RegCumHazPH	0.713 (0.035)	0.751 (0.042)

9. RPssPH: a proportional log cumulative hazard model with a smoothing spline for the log cumulative baseline hazard and all main covariate effects.
10. RPssTV: same as (9), but including interactions with log time for all 10 covariates.
11. Regularized log hazard model (RegHazPH), same as (1), but without the time-varying effects.
12. Regularized log cumulative hazard model (RegCumHazPH): same as (2), but without the TV effects.

Due to the limited sample size in the VALC data, note that, compared to the simulation study, 1 df less was spent on the baseline hazard for parametric models, and that TV effects were modeled linearly instead of using splines for all methods where applicable. For the same reason, the last two models were added as simplifications of the first two in light of the simulation results.

A bootstrapping approach was used to evaluate model performance (1000 repetitions). All penalty parameters were selected based on 10-fold cross-validation as nested in the bootstrap procedure. Out-of-bag performance was measured in terms of time-dependent C-statistic (Antolini et al., 2005) and graphical calibrations curves (Austin et al., 2020). Graphical calibration curves were derived at $t = 60$, which was close to the median event time ($t = 62$), and at $t = 120$ and $t = 180$. The time-dependent C-statistic as described by Antolini et al. (2005) was adapted to match Harrell's C-statistic for censored data (Harrell et al., 1996) in case of PH by counting tied predictions for discordant outcomes as 0.5 instead of 0. It is defined as

$$C^{td} = \frac{\sum_i \sum_{j \neq i} \text{conc}_{ij}}{\sum_i \sum_{j \neq i} \text{comp}_{ij}} \tag{11}$$

with

$$\text{comp}_{ij} = I(t_i < t_j \& d_i = 1) + I(t_i = t_j \& d_i = 1 \& d_j = 0) \tag{12}$$

and

$$\text{conc}_{ij} = I[\hat{p}_i(t_i) < \hat{p}_j(t_i)] \cdot \text{comp}_{ij} + \frac{1}{2} I[\hat{p}_i(t_i) = \hat{p}_j(t_i)] \cdot \text{comp}_{ij}. \tag{13}$$

Equation (12) defines comparability of pairs i, j , with case i being comparable to case j if its event indicator equals 1 ($d_i = 1$) and j has a later event time or equal event time and censored status ($d_j = 0$). Equation (13) defines concordance of the predicted survival probabilities at time t_i and adds the 0.5 for tied predictions of comparable pairs. Hence, C^{td} estimates the concordance probability among comparable pairs.

Results are shown in Table 1, Figure 3, and supplementary Figures I.1 and I.2 for all methods except CoxTVridge, which did not converge regardless of the choice of penalty. Even though the differences were small, RegHazTV, RegCumHazTV, RegHazPH, RegCumHazPH, and RPrCsTV performed somewhat better than the remaining methods in terms of average

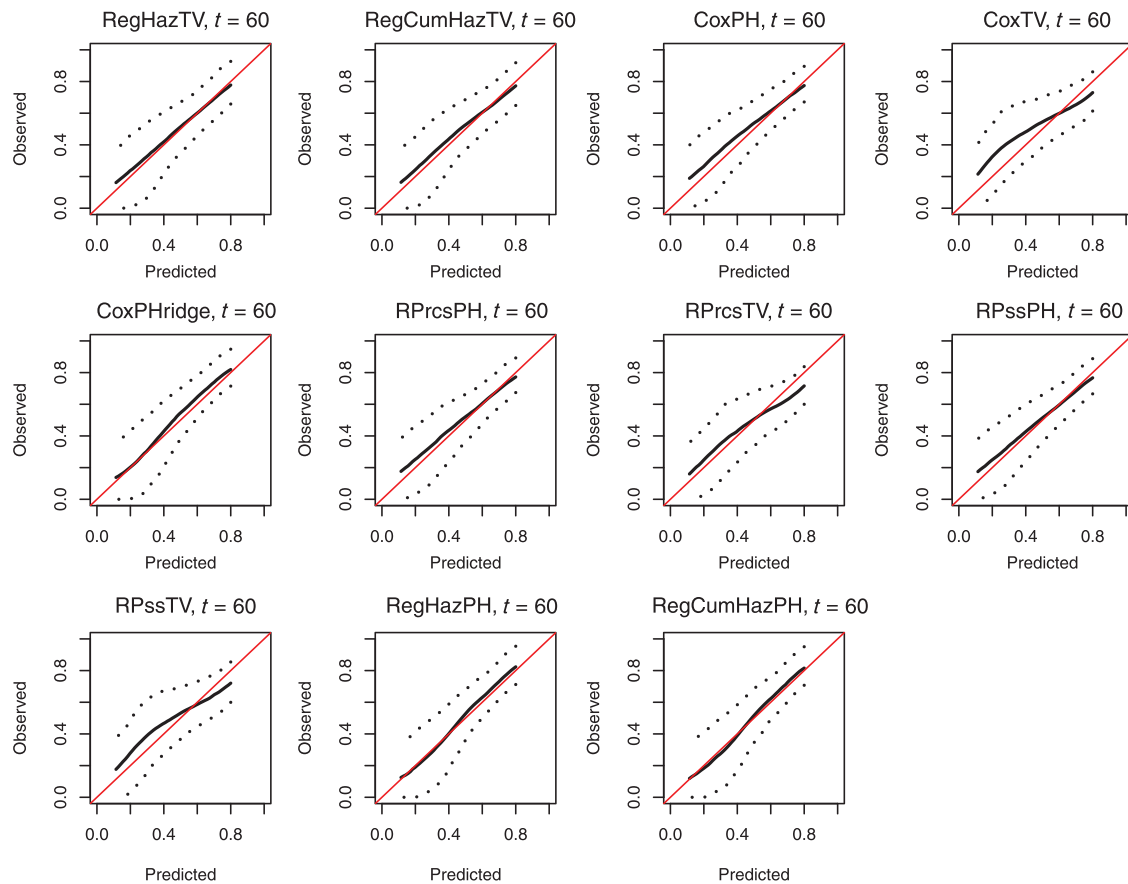


FIGURE 3 Calibration curves for out-of-bag predictions at 60 days follow-up in the Veterans' Administration Lung Cancer (VALC) data. Solid lines represent the average calibration curve over 1000 out-of-bag estimates; dotted lines are for the 10th and 90th percentiles.

time-dependent C-statistic (C^{td}). The fixed time-horizon 60-day C^{td} statistics (censoring individuals with a time-to-event > 60 days) had a similar rank-ordering but were even more similar between methods. In terms of calibration, the differences are more apparent, with RegHazPH and RegCumHazPH being closest to the diagonal for $t = 60$ days (Figure 3), $t = 120$ days, and $t = 180$ days (supplementary Figures I.1 and I.2, respectively). Calibration curves for the other regularized models also look reasonable, especially for the earlier time points. Summarizing, differences were small, but regularization was the preferred option on all performance measures. Even though allowing for TV effects is quite a stretch given the limited sample size, the regularized time-dependent effects models performed reasonably well.

7 | SOFTWARE

Regularized log hazard and log cumulative hazard modeling has been implemented in R (R Core Team, 2021) package **regsurv**, which is available on GitHub (<https://github.com/jeroenhoogland/regsurv>) and provides functions for model optimization, penalty parameter tuning, prediction, and plot methods for loss and coefficients paths across a penalty parameter grid. Royston–Parmar modeling software is readily available (e.g., `stpm2` (Lambert, 2010) in Stata (Stata-Corp, 2021) and `rstpm2` (Liu et al., 2018) in R), and the same holds for standard cox modeling (e.g., the **survival** package (Therneau, 2022) in R) and regularized cox modeling (e.g., the **glmnet** package (Friedman et al., 2010, 2021, Simon et al., 2011) and the **penalized** package (Goeman, 2010) in R). R script for replication of the simulation study and applied example is available as supplementary material.

Comparing **regsurv** against the popular **glmnet** and **penalized** in the context of survival modeling, the key difference is that **regsurv** was developed for fully parametric models, whereas **glmnet** and the **penalized** package were developed for Cox models. Both **glmnet** and **penalized** offer very fast implementations for lasso, ridge, and elastic net penalties (and fused lasso in **penalized**) on the linear predictor parameters in a CoxPH model. These estimates are combined with

a separate estimate of the baseline cumulative hazard (e.g., a Breslow estimate) for prediction purposes. While **glmnet** also allows for TV coefficients, this implementation did not perform well in the simulation study. Also, prediction of survival probabilities or cumulative hazards from Cox models with time-dependent coefficients is not implemented. Briefly, our **regsurv** package provides (i) smooth (cumulative) baseline hazard estimates; (ii) straightforward incorporation of TV coefficients; (iii) simultaneous optimization and regularization of all model parameters based on the full likelihood, (iv) group lasso penalization, (v) easy prediction from TV effects model, and (vi) the choice between log hazard or log cumulative hazard modeling. This comes at the cost of computational complexity and the risk of misspecification that comes with any fully parametric model.

8 | DISCUSSION

We have introduced regularization methods for parametric survival models with a flexible baseline hazard or cumulative hazard. This opens an important toolbox that constrains the risk of overfitting and increases prediction accuracy for a flexible class of models. Importantly, these models explicitly model the baseline (cumulative) hazard, which is of interest for the prediction of survival probabilities over time.

The introduced penalty functions include the elastic net penalty (and hence ridge and lasso penalty) and the group lasso penalty. From a theoretical perspective, the corresponding optimization problems were shown to be convex, enabling a unified optimization approach and providing guarantees with respect to global optimality of the solution(s). From an applied perspective, a freely accessible software implementation was written in R package **regsurv** with high level functions for model fitting, cross-validation, and prediction to make the methods easily accessible. Simulation results showed that the proposed methods performed well in comparison to alternative methods including Cox regression, regularized Cox regression, and Royston–Parmar models of various types. Importantly, regularization was beneficial even in large sample size settings. In line with the simulation results, the applied example in the VALC study showed that the proposed methods were competitive in terms of discrimination and had a slight edge in terms of calibration.

Regarding practical applications, it should be noted that while regularization helps to balance model complexity against limited sample size, it is not an alternative to larger sample sizes (Riley et al., 2020; Van Calster et al., 2020). Especially when sample sizes are small, it is difficult to estimate the regularization parameter(s). Hence, choices with respect to model complexity and regularization should be made judiciously, based on available content knowledge and sample size considerations (Riley et al., 2019). As a second practical consideration, some discussion on the choice between popular semiparametric options (e.g., a Cox model combined with a Breslow estimate of the cumulative baseline hazard) and fully parametric models is warranted. Advantages of the latter include smoothness of (cumulative) hazards over time, ease of model sharing, the relative ease with which TV effects can be incorporated, and the unified regularization approach that enables optimization of the whole model in one go. A disadvantage is the risk of misspecification with respect to the baseline hazard, which is mitigated by the use of splines. Benefits of the semiparametric approach include its familiarity, well-known properties, easy access to software across platforms, and very flexible baseline cumulative hazard. In our experience, the proposed regularized parametric approach is primarily beneficial when modeling TV effects, as also illustrated in the simulation study. As a third practical consideration, the choice between log hazard and log cumulative hazard modeling deserves some further attention. While performance was very comparable for both model types across the investigated settings, log hazard models may be preferable in case of many TV effects. This is because they naturally enforce nonnegative hazards even when extrapolating beyond the development data. Both our implementation of the log cumulative hazards model and the version implemented by Liu et al. (2018) enforce this constraint only within the domain of the development data, since the monotonicity of the time-covariate interactions essentially depend on the covariate distributions.

With respect to limitations, it should be noted that the simulation study and applied example were intended as a proof-of-concept for the introduced methodology, and future research may inform on a wider range of settings. Our particular simulation setting reflected the tension between model complexity and sample size in the context of TV effects. Other settings include higher dimensional main-effects settings (e.g., with $p \gg n$), models exploring many interaction effects, and more extensive negative controls (where many parameters are actually zero). Also, simulations that are closely inspired by real data may sometimes be very helpful to better understand method characteristics for a particular application. As a second limitation, the computation time for regularization paths can be considerable for the log hazard models due to the required numerical integration. Nonetheless, computation times for the **glmnet** implementation of Cox regression with TV effects may be even longer without coarsening the grid of event times used to represent TV effects. Lastly, we have

implemented restricted cubic splines based on truncated power series since they allow the user to regularize either or both of linear and nonlinear contributions. Different types of splines may offer different benefits for particular situations (Perperoglou et al., 2019). For instance, in the context of parametric survival modeling without regularization, smoothing splines have been implemented for log-cumulative hazard models (Liu et al., 2018) and log hazard models (Fauvernier et al., 2019), avoiding the need for knot specification. As such, the interplay between different types of splines and regularization is an interesting topic for further research, and has already been explored in the context of generalized linear models (Chouldechova & Hastie, 2015).

Summarizing, parametric log hazard and log cumulative hazard models provide a flexible tool for survival analysis, and the addition of regularization enhances control on overfitting in settings with limited sample size in light of model complexity. This is of particular interest for the development of prediction models with the aim to predict survival probabilities over time.

ACKNOWLEDGMENTS

J. Hoogland, T. P. A. Debray, J. IntHout, and J. B. Reitsma acknowledge financial support from the Netherlands Organisation for Health Research and Development (Grant 91215058). T. P. A. Debray also acknowledges financial support from the Netherlands Organisation for Health Research and Development (Grant 91617050).


CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data for the Veterans' Administration Lung Cancer study are publicly available in the **survival** package (Therneau, 2022) in R and in Kalbfleisch and Prentice (2002). R scripts are made available to replicate all results, including both the applied example and the simulation study. In addition, software for general purpose use of the proposed models is available in R package **regsurv** (<https://github.com/jeroenhoogland/regsurv>).

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

J. Hoogland  <https://orcid.org/0000-0002-2397-6052>

T. P. A. Debray  <https://orcid.org/0000-0002-1790-2719>

M. J. Crowther  <https://orcid.org/0000-0001-8378-8259>

R. D. Riley  <https://orcid.org/0000-0001-8699-0735>

J. IntHout  <https://orcid.org/0000-0002-6127-0747>

J. B. Reitsma  <https://orcid.org/0000-0003-4026-4345>

A. H. Zwinderman  <https://orcid.org/0000-0003-0361-3139>

REFERENCES

- Antolini, L., Boracchi, P., & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24, 3927–3944.
- Austin, P. C., Harrell, F. E., & Klaveren, D. (2020). Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Statistics in Medicine*, 39, 2714–2742.
- Bower, H., Crowther, M. J., & Lambert, P. C. (2016). Strcs: A command for fitting flexible parametric survival models on the log-hazard scale. 16, 989–1012.
- Boyd, S. P., & Vandenberghe, L. (2015). *Convex optimization* (print ed.). Cambridge University Press.
- Chouldechova, A., & Hastie, T. (2015). Generalized additive model selection. *arXiv preprint arXiv:1506.03850*.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2), 187–220.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.
- Crowther, M. J., & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32, 4118–4134.

- Crowther, M. J., & Lambert, P. C. (2014). A general framework for parametric survival analysis. *Statistics in Medicine*, 33, 5280–5297.
- Domahidi, A., Chu, E., & Boyd, S. (2013). ECOS: An SOCP solver for embedded systems. In *2013 European Control Conference (ECC)*, Zurich (pp. 3071–3076). IEEE.
- Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., Remontet, L., & the Challenges in the Estimation of Net Survival Working Survival Group. (2019). Multi-dimensional penalized hazard model with continuous covariates: Applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68, 1233–1257.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
- Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Yang, J. (2021). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.
- Fu, A., Narasimhan, B., & Boyd, S. (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14), 1–34.
- Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52(1), 70–84.
- Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer Series in Statistics. Springer.
- Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361–387.
- Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, 62, 426–433.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer Series in Statistics. Springer.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. No. 143 in Monographs on Statistics and Applied Probability. CRC Press; Taylor & Francis Group.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.
- Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Wiley Series in Probability and Statistics. Wiley.
- Lambert, P. (2010). *STPM2: Stata module to estimate flexible parametric survival models*. Statistical Software Components, Boston College Department of Economics.
- Lin, D. Y. (2007). On the Breslow estimator. *Lifetime Data Analysis*, 13, 471–480.
- Liu, X.-R., Pawitan, Y., & Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27, 1531–1546.
- McLernon, D. J., Giardiello, D., Van Calster, B., Wynants, L., van Geloven, N., van Smeden, M., Therneau, T., Steyerberg, E. W., McLernon, D. J., Giardiello, D., Van Calster, B., Wynants, L., van Geloven, N., van Smeden, M., Therneau, T., Steyerberg, E. W., Bossuyt, P., Boyles, T., Taylor, J., ... topic groups 6 and 8 of the STRATOS Initiative. (2023). Assessing performance and clinical usefulness in prediction models with survival outcomes: Practical guidance for Cox proportional hazards models. *Annals of Internal Medicine*, 176, 105–114.
- Meier, L., van de Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B*, 70(1), 53–71.
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods: Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.
- Novomestky, F. (2013). gaussquad: Collection of functions for Gaussian quadrature. R Package Version 1.0-2.
- Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, 19, 46.
- Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part I—Continuous outcomes. *Statistics in Medicine*, 38, 1262–1275.
- Riley, R. D., Snell, K. I., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2020). Penalisation and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, 132, 88–96.
- Royston, P., & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21, 2175–2197.
- R Core Team. (2021). R: A Language and Environment for Statistical Computing. <https://www.R-project.org/>
- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5), 1–13.
- StataCorp. (2021). Stata Statistical Software. <https://www.stata.com/>
- Steyerberg, E. (2009). *Clinical prediction models*. Statistics for Biology and Health. Springer New York.
- Therneau, T. M. (2000). A Package for Survival Analysis in R. <https://CRAN.R-project.org/package=survival>
- Therneau, T. M., & Grambsch, P. M. (2011). *Modeling survival data: Extending the Cox model*. Springer.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
- Van Calster, B., van Smeden, M., De Cock, B., & Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, 29, 3166–3178.
- van Nee, M. M., van de Brug, T., & van de Wiel, M. A. (2023). Fast marginal likelihood estimation of penalties for group-adaptive elastic net. *Journal of Computational and Graphical Statistics*, 32, 950–960.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hoogland, J., Debray, T. P. A., Crowther, M. J., Riley, R. D., IntHout, J., Reitsma, J. B., & Zwinderman, A. H. (2024). Regularized parametric survival modeling to improve risk prediction models. *Biometrical Journal*, 66, 2200319. <https://doi.org/10.1002/bimj.202200319>