Biometrika (2016), **103**, 4, *pp.* 817–828
Printed in Great Britain

doi: 10.1093/biomet/asw049

Optimal stratification in outcome prediction using baseline information

By FLORENCE H. YONG

Department of Biostatistics, Harvard University, 655 Huntingdon Avenue,
Boston, Massachusetts 02115, U.S.A.
florenceyong@mail.harvard.edu

LU TIAN

Department of Biomedical Data Science, Stanford University,
150 Governor's Lane, Stanford, California 94305, U.S.A.
lutian@stanford.edu

SHENG YU

Center for Statistical Science, Tsinghua University, Beijing 100084, China
syu@tsinghua.edu.cn

TIANXI CAI and L. J. WEI

Department of Biostatistics, Harvard University, 655 Huntingdon Avenue,
Boston, Massachusetts 02115, U.S.A.
tcai.hsph@gmail.com wei@hsph.harvard.edu

Summary

A common practice in predictive medicine is to use current study data to construct a stratification procedure, which groups subjects according to baseline information and forms stratum-specific prevention or intervention strategies. A desirable stratification scheme would not only have small intra-stratum variation but also have a clinically meaningful discriminatory capability. We show how to obtain optimal stratification rules with such desirable properties from fitting a set of regression models relating the outcome to baseline covariates and creating scoring systems for predicting potential outcomes. We propose that all available optimal stratifications be evaluated with an independent dataset to select a final stratification. Lastly, we obtain inferential results for this selected stratification scheme with a holdout dataset. When only one study of moderate size is available, we combine the first two steps via crossvalidation. Extensive simulation studies are used to compare the proposed stratification strategy with alternatives. We illustrate the new proposal using an AIDS clinical trial for binary outcomes and a cardiovascular clinical study for censored event time outcomes.

Some key words: Cox regression model; Crossvalidation; Dynamic programming; Prediction score; Stratified medicine.

1. Introduction

One important aspect of stratified medicine development is the ability to predict the response probability for individual subjects. To construct a prediction procedure for a future subject's outcome via his or her baseline information, at the first step it is common practice to fit a parametric or semiparametric regression model in order to relate a subject's outcome to his or her covariates. If the model is a reasonable approximation to the true one, the resulting predicted value for an individual would be close to his or her outcome value. Such predicted values create a scoring system for all future subjects, which can have systematic bias when the regression model is misspecified. To eliminate the bias, one may further calibrate the scoring system nonparametrically (Tian et al., 2014). However, the calibrated prediction for subjects with a given score can be quite unstable due to sparseness of the observations. Consequently, the resulting fine-level prediction procedure may perform poorly. Common practice is to group the scores into several strata and use the average observed outcome in each stratum to predict outcomes of new subjects classified into that stratum. However, the choice of strata is often ad hoc in practice. In this paper, we present a stratification procedure based on baseline covariates which provides a clinically meaningful grouping strategy.

To illustrate current practice, we use data from a clinical study on treating HIV diseases (Hammer et al., 1997). This trial, a randomized, double-blind, placebo-controlled clinical study conducted by the AIDS Clinical Trials Group, demonstrated the overall efficacy of a combination of two nucleoside regimen with a protease inhibitor Indinavir for treating HIV-infected patients. The combination treatment concept has since been adopted for HIV patient management, but the combination therapy may be a poor choice for patients who do not have a reasonable chance of responding to the treatment. To identify these patients, we show a conventional, ad hoc procedure to construct a stratification scheme using baseline variables. For this study, there were 537 patients treated by the three-drug combination who had complete baseline information. One endpoint was a binary outcome Y, indicating whether or not the patient's HIV-RNA viral level was under an assay-detectable level, 500 copies/mL, at week 24. A nonresponder to the treatment was defined by the RNA level being above 500 copies/mL at week 24 or any dropout before week 24. The observed overall response rate was 45%. To build a predictive scoring system, consider an additive logistic regression model for Y with four baseline covariates: age; gender; CD4 count, denoted by CD4_0; and $\log_{10}$-transformed HIV-RNA values, denoted by $\log_{10}\text{RNA}_0$. For RNA values below 500, $\log_{10}\text{RNA}_0$ is replaced by $0{\cdot}5\log_{10}(500) = 1{\cdot}35$ in our analysis. We fit the model to the entire dataset, giving the individual predicted response rate

$$\psi(-0{\cdot}508 + 0{\cdot}044 \times \text{age} - 0{\cdot}493 \times \text{female} + 0{\cdot}004 \times \text{CD4}_0 - 0{\cdot}346 \times \log_{10}\text{RNA}_0), \quad (1)$$

where $\psi(s) = \{1 + \exp(-s)\}^{-1}$ is the expit function. The 537 predicted response rates range from 0·09 to 0·93. If the model is reasonably good, a future subject with a high score tends to respond to the treatment. A conventional way to group those patients is to stratify them into, for instance, four categories with roughly equal sizes by using the quartiles of the predicted scores, and this yields empirical average response rates of 31%, 42%, 38%, and 67%. Unfortunately this stratification scheme does not have good discriminatory capability across all the strata. The average response rates are not monotonically increasing over the ordered strata, perhaps due to the prediction model (1) being inadequate or an improper grouping of the prediction scores.

In this article, we present an optimal stratification strategy incorporating model selection from a collection of candidates that satisfy certain criteria. Specifically, we show how to obtain an optimal grouping scheme for each candidate scoring system created from a regression model. For example, with the predicted response rates (1), we consider all possible discretization schemes with stratum sizes at least 10% of the study sample size and any monotonically increasing stratum-specific average response rates with an incremental value of at least 0·2 between consecutive strata. We then choose the final stratification to be the one that minimizes a certain overall prediction error among all such stratification schemes. Dynamic programming techniques are used to solve this optimization problem (Taha, 2003). With the data from the HIV example and model (1), our proposal results in three categories with stratum-specific average response rates of 11%, 42% and 69% and stratum sizes of 65, 343 and 129, respectively. If model (1) is appropriate, future patients classified to the first stratum may not benefit much from this rather costly three-drug combination therapy. We also describe how to evaluate competing models to select the final stratification scheme. Finally, we generalize the new proposal to handle censored event time outcomes and illustrate the procedure using data from a cardiovascular study (Braunwald et al., 2004).

2. Optimal stratification for a specific scoring system

Let Y be the outcome variable and V a vector of baseline covariates. Assume that the conditional mean $\mu(V) = E(Y \mid V)$ is the parameter of interest for future prediction. To estimate $\mu(V)$ when V is not a scalar, we generally use a working model which relates Y to Z, a vector transformation of V including components such as quadratic and interaction terms. Let $\mu(V) = g(\beta^{\mathrm{T}} Z)$, where $g(\cdot)$ is a given smooth monotone function. Let the data consist of n independent copies $\{(Y_i, V_i, Z_i), i = 1, \ldots, n\}$ of (Y, V, Z). An estimate $\hat{\beta}$ for β can be obtained via a regularized estimation procedure such as the lasso (Tibshirani, 1996), especially when the dimension of Z is large. If the regression model is a reasonable approximation to the true one, the resulting estimator $\hat{\mu}(V) = g(\hat{\beta}^{\mathrm{T}} Z)$ would be close to $\mu(V)$, and a large $\hat{\mu}(\cdot)$ indicates that the subject is expected to have a large outcome value Y. As an example, for the binary outcome Y in the HIV study discussed in § 1, one may use a logistic model with lasso or ridge regularization to obtain $\hat{\mu}(\cdot)$. The score (1) in § 1 gives the individual predicted response rate based on a simple additive logistic regression with four baseline covariates.

Suppose that we group the n subjects into K consecutive strata $S_1, \ldots, S_K$ based on the score $\hat{\mu}(\cdot)$. With a slight abuse of notation, we use K to denote the number of strata, a random quantity dictated by the working model and data. Let $\bar{Y}_k$ be the empirical mean outcomes in the kth stratum $(k = 1, \ldots, K)$. For a future subject who is classified to the kth stratum, we predict the individual outcome by the corresponding stratum-specific mean $\bar{Y}_k$. The performance of this stratification can be evaluated via a loss function

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in S_k} |Y_i - \bar{Y}_k|, \tag{2}$$

which quantifies the average within-stratum variation. When all the scores are distinct, an optimal stratification that minimizes (2) would result in n strata each with only one member and zero observed prediction error. However, the prediction error for future observations with such a sparse stratification would be unacceptably high. To increase prediction precision while ensuring stable subgroups, one may group the subjects under the condition of a minimal stratum size of at least a certain fraction p_0 of n. To ensure that the stratification scheme yields meaningful differences in group-specific average outcomes between subgroups, we further impose a constraint on all candidate stratification schemes such that

$$\bar{Y}_k - \bar{Y}_{k-1} \geqslant d_0 \quad (k = 2, \ldots, K),$$

where $d_0 > 0$ represents a minimum clinically meaningful increment.

Minimizing the loss function (2) over all qualifying stratification schemes is a challenging problem. In the Supplementary Material, we show how to identify the boundary values $\{\hat{c}_0, \ldots, \hat{c}_{\hat{K}}\}$ of the strata. We use (2) to evaluate the prediction error. The asymptotic optimality of the resulting stratification scheme is guaranteed by Lemma 1, whose justification is given in the Supplementary Material. This large-sample property of the optimal stratification scheme is also essential to ensure its stability under the crossvalidation setting discussed in § 3.

Lemma 1. *Suppose that the following hold.*

(i) *The regularized estimator $\hat\beta$ of the working regression model converges to a constant vector β_0 within a compact parameter space in probability.*
(ii) *The score estimate $\hat\mu(v)$ converges uniformly to a deterministic function $\tilde\mu(v)$ in probability on the support of V.*
(iii) *The score $\tilde\mu(V)$ is a continuous random variable with bounded support and the joint density function of the continuous components of the random vector $(Y, V^{\mathrm{T}})^{\mathrm{T}}$ is continuously differentiable.*
(iv) *The outcome Y has bounded support.*

Then $\mathrm{pr}\left\{L(\hat c) \leqslant L(c_0) + \epsilon\right\} \rightarrow 0$ *as* $n \rightarrow \infty$ *for any given* $\epsilon > 0$, *where* $\hat c = (\hat c_1, \ldots, \hat c_{\hat K})^{\mathrm{T}}$ *represents the estimated cut-off values by solving the constrained optimization problem based on observed data, and* $c_0 = (c_1, \ldots, c_K)^{\mathrm{T}}$ *is the minimizer of*

$$L(c) = E\left\{\left|Y - f(V \mid c)\right|\right\} \quad \textit{subject to} \quad \begin{cases} \mathrm{pr}\left\{c_{k-1} < \tilde\mu(V) \leqslant c_k\right\} \geqslant p_0, \\ \mu_k - \mu_{k-1} \geqslant d_0, \end{cases}$$

with $f(V \mid c) = \sum_{k=1}^{K} I\{c_{k-1} < \tilde\mu(V) \leqslant c_k\}\mu_k$, $\mu_k = E\{Y \mid c_{k-1} < \tilde\mu(V) \leqslant c_k\}$ *and* $I(\cdot)$ *being the indicator function.*

3. Selecting an optimal stratification scheme

One can obtain the optimal stratified prediction procedure for a prediction score system created by a given working regression model as presented in § 2. To make inferences about the resulting stratified prediction procedure, one may use an independent dataset. When only a single study is available, we may split its data into two separate parts, say, I and II. Using Part I data, we obtain the optimal stratification scheme. Then we use Part II data for inference. Moreover, if there is a collection of competing stratified score systems considered as potential candidates, we may further split the Part I data into two parts, say, Ia and Ib. The data from Part Ia are used for obtaining the optimal stratification schemes for each candidate scoring system as in § 2, whereas Part Ib data are used to evaluate all candidates and to select the best stratification scheme. This multi-stage procedure is similar to that proposed in Li et al. (2016).

Suppose that there are several optimal stratification schemes available with the data from Part Ia. In this section, we show how to evaluate them and choose the best one. Let the data from Part Ib be denoted by n^* independent identically distributed observations $\{(Y_i^*, V_i^*, Z_i^*), i = 1, \ldots, n^*\}$, where a generic variable A^* is defined as A in § 2. For each candidate scoring system, we obtain its optimal stratified counterpart from the data of Part Ia with boundary points $\{\hat c_0, \ldots, \hat c_{\hat K}\}$ and the stratum-specific predictions $\{\bar Y_k, k = 1, \ldots, \hat K\}$. To evaluate the predictive performance of such a scheme with the data from Part Ib, we consider the loss function

$$\mathcal{L}^* = \frac{1}{n^*} \sum_{k=1}^{\hat K} \sum_{\hat\mu(V_i^*) \in (\hat c_{k-1}, \hat c_k]} |Y_i^* - \bar Y_k| \tag{3}$$

where $\hat{\mu}(V_i^*) = g(\hat{\beta}^{\mathrm{T}} Z_i^*)$. An optimal stratification scheme minimizes (3) among all candidates under consideration. On the other hand, we bear in mind that a parsimonious model may be appealing in practice if the associated $\mathcal{L}^*$ is comparable to the minimum value derived from a more complex model.

Since Parts Ia and Ib may be small, one may use Monte Carlo crossvalidation (Xu & Liang, 2001; Yong et al., 2013) to obtain a more stable (3). Specifically, we randomly split the Part I dataset into Ia and Ib, say, N times. For the jth split, we repeat the above model building and evaluation procedure for each candidate model and obtain $\mathcal{L}_j^*$ from (3). We then compute the average, $\bar{\mathcal{L}}^* = N^{-1} \sum_{j=1}^{N} \mathcal{L}_j^*$. For each candidate model, we refit the entire Part I data and denote the final realized stratification rule by $\mathcal{M}^*$. The pair $(\bar{\mathcal{L}}^*, \mathcal{M}^*)$ reflects the magnitude of the estimated within-stratum variation and the complexity of each candidate model. The selection of an optimal stratification rule would be based on such pairs. With the data from Part II, we then construct confidence intervals for the stratum-specific mean values of the outcome variables for the selected final stratification scheme. Although crossvalidation may not choose the best stratification based on the Part I data, it identifies the procedure with the best average performance in constructing the stratification. Thus it is reasonable to expect that the stratification rule from the selected procedure will be highly competitive if not optimal.

We now use the data from the HIV study to illustrate our proposal. For this study, other than the four baseline variables discussed in § 1, there are seven additional baseline covariates and two short-term marker values at week 4, including CD4 count, denoted by $\mathrm{CD4}_4$, and $\log_{10}$ RNA, denoted by $\log_{10} \mathrm{RNA}_4$. The additional baseline covariates are race, injection-drug use, haemophilia, CD8 count, weight, Karnofsky performance score, and months of prior zidovudine therapy. Patients with missing covariate values account for 4% of the data. Any missing covariates are replaced by the corresponding sample averages.

We first randomly split the entire dataset of 537 patients into Part I and Part II, with sample sizes 268 and 269. In the Monte Carlo crossvalidation, we take N=200 and the sizes of Part Ia and Ib to be equal for each crossvalidation. For illustration, we consider four working models: three logistic regression models with lasso regularization methods and tuning parameters selected via a 20-fold crossvalidation procedure (Friedman et al., 2010), and a null model using the overall mean response proportion in Part Ia to predict future outcomes. Table 1 summarizes the composition of each model. We also present $\bar{\mathcal{L}}^*$, obtained by averaging the 200 $\mathcal{L}_j^*$ values; and for the corresponding $\mathcal{M}^*$, we report the numbers of informative baseline covariates needed to compute the score $\hat{\mu}(V)$ and nonzero regression coefficients in $\hat{\beta}$ estimated based on the entire Part I data to summarize its complexity. In constructing all candidate stratification schemes, we use $d_0 = 0{\cdot}2$ and $p_0 = 0{\cdot}1$.

From Table 1, Models 1 and 3 have almost the same $\bar{\mathcal{L}}^*$ values, but the $\mathcal{M}^*$ of Model 1 has fewer baseline covariates. The resulting predicted response rate is

$$\psi(-0{\cdot}23 - 0{\cdot}075 \times \log_{10} \mathrm{RNA}_0 - 0{\cdot}46 \times \log_{10} \mathrm{RNA}_4 + 0{\cdot}00036 \times \mathrm{CD4}_0 \\ + 0{\cdot}0028 \times \mathrm{CD4}_4 + 0{\cdot}029 \times \mathrm{age}).$$

This final selected stratification scheme $\mathcal{M}^*$ has three strata with cut-offs $\hat{c}_1 = 0{\cdot}25$ and $\hat{c}_2 = 0{\cdot}45$. The stratum-specific means are 0·06, 0·36, and 0·62; and the corresponding stratum sizes are 51, 107, and 110, based on the Part I data. These stratum-outcome-average estimates may be biased due to the extensive model building, evaluation and selection. To obtain valid inferences for this final stratification rule, we use the boundary values $\hat{c}_1$ and $\hat{c}_2$ to group subjects from Part II data. The resulting point estimates and 0·95 confidence intervals for the three stratum-average

Table 1. *Regression model candidates for HIV data, prediction accuracy $\bar{\mathcal{L}}^*$, complexity of $\mathcal{M}^*$, and $\|\hat{\beta}\|_0$, the number of nonzero components of $\hat{\beta}$*

Model	Candidate independent variables	dim(Z)	$\bar{\mathcal{L}}^*$	$\mathcal{M}^*$	
				# covariates	$\|\hat{\beta}\|_0$
1	age, sex, CD4 count and $\log_{10}$RNA at baseline and week 4	6	0·42	5	5
2	all baseline covariates plus their first-order interaction terms	78	0·47	12	14
3	all baseline covariates and CD4 and $\log_{10}$RNA at week 4 plus their first-order interaction terms	105	0·41	13	20
4	none	0	0·48	0	0

Table 2. *Response probabilities* (%) *of nine subgroups in simulation settings 3 and 4*

		Setting 3			Setting 4		
		Baseline RNA ($\times 10^5$)			Baseline RNA ($\times 10^5$)		
		$[0, 0{\cdot}56]$	$(0{\cdot}56, 2{\cdot}04]$	$(2{\cdot}04, \infty)$	$[0, 0{\cdot}56]$	$(0{\cdot}56, 2{\cdot}04]$	$(2{\cdot}04, \infty)$
Week 4	$[0, 0{\cdot}22]$	61	64	64	74	78	78
RNA ($\times 10^3$)	$(0{\cdot}22, 3{\cdot}02]$	47	48	44	47	48	40
	$(3{\cdot}02, \infty)$	15	19	27	3	6	14

response rates are 0·17 (0·06, 0·28), 0·41 (0·31, 0·51) and 0·65 (0·57, 0·73), with stratum sizes 47, 91, and 131; see the Supplementary Material.

4. A SIMULATION STUDY

To compare our stratification method with alternatives, we conducted a simulation study where data were generated mimicking the AIDS Clinical Trials Group trial described in § 1. Throughout the simulation, we used the observed baseline RNA and week 4 RNA levels as the covariates. In the first setting we simulated the binary response according to a logistic regression model with $\log_{10}$-transformed baseline RNA, $\log_{10}$ RNA$_0$, and week 4 RNA, $\log_{10}$ RNA$_4$, and their product as the independent variables. The regression coefficients of the logistic regression model were set as the maximum likelihood estimators based on the original data. In the second setting, we doubled all the regression coefficients in the logistic regression model, and adjusted the intercept term to maintain the same overall prevalence rate. In the third setting, the binary response was simulated with different probabilities in nine subgroups summarized in Table 2, where the cut-off values $(0{\cdot}56, 2{\cdot}04) \times 10^5$ and $(0{\cdot}22, 3{\cdot}02) \times 10^3$ are tertiles of baseline and week 4 RNA levels, respectively. The response probability was chosen to be the observed response rate of the corresponding subgroup in the AIDS Clinical Trials Group 320 trial. In the last setting, the response probabilities of the aforementioned nine subgroups were modified to increase the between-group differences, as shown in Table 2.

For each simulated training dataset of 537 observations, a misspecified simple logistic model

$$\text{logit}\{\text{pr}(Y = 1 \mid \text{RNA}_0, \text{RNA}_4)\} = \beta_0 + \beta_1 \log_{10} \text{RNA}_0 + \beta_2 \log_{10} \text{RNA}_4$$

was used to generate a scoring system. Then various stratification methods were used to form strata. In addition to our method, we also selected the cut-off values simply as the median, tertiles and quartiles. In our proposal, we let $p_0 = 0{\cdot}1$ and $d_0 = 0{\cdot}2$. We also employed classification and regression trees to build the stratification system (Breiman et al., 1984), with minimum node size $0{\cdot}1n$ and the tree pruned back based on crossvalidation. Lastly, we also calibrated the continuous score by the nonparametrically estimated conditional expectation of Y given the score.

For each setting, the mean square errors for predicting future observations were estimated for various stratification rules and the calibrated continuous score. As references, the mean square errors were also computed for the null and true models, using the mean response rate in the training set and the true individual response rate as the risk prediction, respectively. For each stratification, the number of strata and the differences between neighbouring strata were recorded.

The results based on 1000 simulations are reported in Table 3. The mean square errors across different methods lie between those of the null and true models, as expected. Compared with other stratification methods, our method usually yields the smallest prediction error at a level comparable to that of the continuous score in all settings, in spite of the fact that the latter is asymptotically optimal. The stratification from classification and regression trees is competitive and tends to yield fewer strata. The strata constructed by the new proposal guarantee that a high proportion of between-strata differences is greater than the prespecified threshold and the violation is mild when that occurred. Not all the differences satisfy the constraint, due to the randomness of $\bar{Y}_k - \bar{Y}_{k-1}$ in the training set.

5. Event time as the outcome variable

If the outcome T is the time to a specific event, it may be censored and its mean or median may be badly estimated. A common summary parameter is the event rate at a specific time-point τ, but this does not include information about the event occurrence profile. The restricted mean survival time has been argued to be a clinically meaningful summary for such a distribution (Royston, 2006; Royston & Parmar, 2011; Zhao et al., 2013). Specifically, let $Y = TI(T \leqslant \tau) + \tau I(T > \tau)$ and $\mu(V) = E(Y \mid V) = \int_0^\tau S(t \mid V)\,\mathrm{d}t$ as defined in § 2, where $S(t \mid V) = \mathrm{pr}(T > t \mid V)$. Here, $\mu(V)$ is the average event-free time for all subjects with covariate V, which would be followed up to time-point τ. The definition of restricted mean survival time and subsequent results depend on the choice of τ, which is often set to be close to the maximum study follow-up time. In practice, T may be right-censored by an independent random variable C. However, one can always observe (X, V, Δ), where $X = \min(T, C)$ and $\Delta = I(T \leqslant C)$. Therefore, the observed data consist of n independent copies $\{(X_i, V_i, \Delta_i), i = 1, \ldots, n\}$ of (X, V, Δ). When $\Delta_i = 1$ or $X_i \geqslant \tau$, $Y_i = \min(T_i, \tau)$ is observed.

Inferences about $\mu(V)$ have been extensively studied (Zucker, 1998; Tian et al., 2014). For example, the area under the Kaplan–Meier curve is a consistent estimator of the restricted mean survival time for a single group. To create a scoring system for $\mu(V)$, one may use the Cox (1972) regression model

$$\log\{-\log S(t \mid V)\} = \log\{-\log S_0(t)\} + \beta^{\mathrm{T}} Z,$$

where $S_0(\cdot)$ is an unknown baseline survival function and β is the regression coefficient vector. A regularized estimate $\hat{\beta}$ of β can be obtained by minimizing a penalized log partial likelihood function. The baseline survival function $S_0(t)$ can then be estimated by $\exp\{-\hat{\Lambda}_0(t)\}$, where $\hat{\Lambda}_0(t)$ is the cumulative hazard function estimator of Breslow (1972). It follows that the restricted mean survival time for subjects with covariate V can be estimated as

$$\hat{\mu}(V) = \int_0^\tau \exp\left\{-\hat{\Lambda}_0(t) e^{\hat{\beta}^{\mathrm{T}} Z}\right\} \mathrm{d}t.$$

Table 3. *Mean square prediction errors, MSE, and proportions of between-strata differences* (%) *satisfying the constraint for the different methods based on* 1000 *simulations. The percentage within the parenthesis reflects the empirical distribution of K*

Setting	Stratification	MSE	Proportion of between-strata differences satisfying the constraint				
	method	(×100)	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
1	New	22·1	100	53·0	27·4		
			(13·1)	(81·3)	(5·6)		
	Median	22·5	100				
	Tertile	22·1		50·0			
	Quartile	22·2			0·0		
	CART	22·2	100	50	33·7	29·4	10
			(50·4)	(39·0)	(8·7)	(1·7)	(0·2)
	Continuous	22·1					
	Null model	24·7					
	True model	21·5					
2	New	16·6	100	99·8	54·2	28·1	
			(0·1)	(52·6)	(46·5)	(0·8)	
	Median	18·2	100				
	Tertile	17·2		100·0			
	Quartile	17·3			52·4		
	CART	16·8	100	98·8	50·3	46·7	
			(9·3)	(76·4)	(12·8)	(1·5)	
	Continuous	21·8					
	Null model	24·9					
	True model	16·0					
3	New	22·6	100	43·7	33·3		
			(33·4)	(66·3)	(0·3)		
	Median	22·9	100				
	Tertile	22·3		47·2			
	Quartile	22·6			0		
	CART	22·6	100	45·9	22·9	16·7	25
			(59·7)	(27·9)	(9·6)	(1·8)	(0·4)
	Continuous	22·6					
	Null model	24·7					
	True model	21·8					
4	New	17·6	100	99·9	63·6	50·0	
			(0·1)	(75·5)	(24·1)	(0·3)	
	Median	19·3	100				
	Tertile	17·7		100·0			
	Quartile	18·4			66·7		
	CART	17·6	100	99·8	66·7	50·0	
			(6·3)	(78·3)	(13·6)	(1·8)	
	Continuous	22·0					
	Null model	24·6					
	True model	17·0					

New, the proposed method; Median, median-based stratification; Tertile, Tertile-based stratification; Quartile, Quartile-based stratification; CART, classification and regression tree; Continuous, calibrated continuous score.

For any scoring system, we can then use the technique described in § 3 to obtain an optimal stratification. Specifically, in the limit, we are interested in minimizing $L(c)$. If the censoring time is independent of the survival time T and covariates V, the prediction error $L(c)$ can be estimated as

$$n^{-1}\sum_{k=1}^{K}\sum_{i\in S_k} w_i|Y_i-\bar{Y}_k|,$$

where $w_i=\{\Delta_i+(1-\Delta_i)I(X_i\geqslant\tau)\}/\hat{G}(Y_i)$ and $\hat{G}(\cdot)$ is the Kaplan–Meier estimator for the censoring distribution using the entire dataset. Here, $\bar{Y}_k$ is a consistent estimator for the kth stratum-specific restricted mean survival time, which is the weighted average $\sum_{i\in S_k} w_iY_i/\sum_{i\in S_k} w_i$. With the same constraints as those described in § 2, an optimal stratification can be obtained via the dynamic programming technique given in the Supplementary Material. If the obtained scoring system for restricted mean survival time converges to a deterministic limit as the sample size increases, the finite-sample stratified scheme would have the same asymptotic property as that for the noncensored case.

To select the best scoring model from the competing scoring systems, one can use the procedure in § 3 with the weighted version of (3) to evaluate the candidate stratification schemes.

6. Example with censored event time outcomes

We use the data from a cardiovascular clinical trial, Prevention of Events with Angiotensin Converting Enzyme Inhibition, to illustrate the proposal with an event time outcome variable. The trial is a double-blind, placebo-controlled study (Braunwald et al., 2004) of 8290 patients enrolled to investigate whether the addition to the conventional therapy of an angiotensin-converting-enzyme inhibitor would provide benefit with respect to, for example, the patient's specific cardiovascular event-free survival. The inhibitor is trandolpril at a target dose of 4 mg/day. The outcome is assumed to be the time to death, nonfatal myocardial infarction or coronary revascularization, whichever occurred first. There are 2110 patients who experienced this composite event, with a median follow-up time of 54 months. A 0·95 confidence interval for the hazard ratio is (0·86, 1·02). Since there was no statistically significant treatment effect, we combined the data from the two treatment groups. The overall observed event times in months range between 0·1 and 81·5 with quartiles 12·8 and 42·4. If we let $\tau=72$ months, the estimated restricted mean event time for the entire group is 60·4 months. This suggests that for future patients in this study population, one expects to have an average of 60·4 months event-free with a follow-up time of 72 months.

Based on the results of Solomon et al. (2006), we considered the following baseline covariates for prediction: the study treatment indicator, age, gender, left ventricular ejection fraction, history of myocardial infarction, history of hypertension, history of diabetes, and estimated glomerular filtration rate as a four-category discretized version represented by three indicator variables with cut-points of 45, 60, and 75. Left ventricular ejection fraction is missing for 412 patients and other covariate information is almost complete, with no more than 11 missing values each. We imputed the missing covariate values with the averages for continuous variables and the most frequently observed category for binary variables. We then randomly split the data into Parts I and II, each consisting of 4145 patients, and split Part I randomly for crossvalidation with $N=200$. Several candidate models are listed in Table 4. Model 2 is built upon the observation that there is potential treatment and estimated glomerular filtration rate interaction (Solomon et al., 2006).

For each regression candidate model, $d_0=3$ months and $p_0=0{\cdot}05$ in the Part Ia training data. Table 4 summarizes $\bar{\mathcal{L}}^*$ for the optimal stratification based on each regression working model and the numbers of informative baseline covariates used in computing the estimated scores and nonzero regression coefficients of $\hat{\beta}$ for $\mathcal{M}^*$. Model 2 has the smallest $\bar{\mathcal{L}}^*$ and yields three strata with $\hat{c}_1=56{\cdot}5$ and $\hat{c}_2=60{\cdot}5$ months. The range of the estimated restricted mean survival time in

Table 4. *Regression model candidates for cardiovascular data, prediction accuracy $\bar{\mathcal{L}}^*$, complexity of $\mathcal{M}^*$, and $\|\hat{\beta}\|_0$, the number of nonzero components of $\hat{\beta}$*

Model	Candidate independent variables	dim(Z)	$\bar{\mathcal{L}}^*$	$\mathcal{M}^*$	
				# covariates	$\|\hat{\beta}\|_0$
1	age, gender, left ventricular ejection fraction, history of myocardial infarction, history of hypertension, history of diabetes, estimated glomerular filtration rate, ACE inhibitor treatment	10	16·92	6	7
2	variables in Model 1 plus three treatment and estimated glomerular filtration rate interaction terms	13	16·90	6	9
3	variables in Model 1 plus their first-order interaction terms	55	16·97	6	9
4	none	0	18·65	0	0

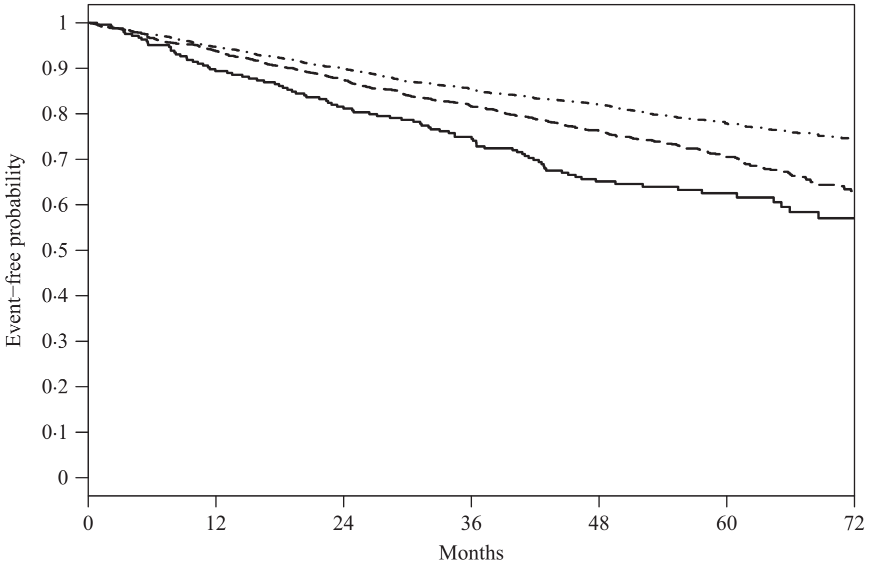

Fig. 1. Stratum-specific Kaplan–Meier estimates obtained from Part II data of the cardiovascular study with $\hat{c}_1 = 56{\cdot}5$ and $\hat{c}_2 = 60{\cdot}5$. The restricted mean survival time estimators and corresponding 95% confidence intervals for the 245 subjects in the first stratum (solid), 1350 subjects in the second stratum (dashes), and 2550 subjects in the third stratum (dot-dash) are 54 (51, 57) months, 59 (58, 60) months, and 62 (61, 63) months, respectively.

the Part I data is from 51·0 to 63·8 months. The corresponding estimated restricted mean survival times of the strata are 54·5, 58·7 and 62·3 months. To make inferences about the prediction of this selected final stratification scheme, we apply it to the Part II data. The corresponding Kaplan–Meier curves for three strata are given in Fig. 1. Based on the restricted area under the Kaplan–Meier curves derived from 1000 bootstrap samples, the point estimates and 0·95 confidence intervals for the stratum-specific restricted mean survival times are 54·3 (51·0, 57·2), 58·9 (57·7, 60·0) and 62·0 (61·2, 62·8) months for the three strata, with $n = 245, 1350$, and 2550 respectively.

Since this set of results depends on the choice of τ used in the restricted mean survival time, we also conducted extensive sensitivity analyses with τ from 60 to 80 months. The resulting stratification scheme is quite stable when $\tau \leqslant 77$ months. When $\tau \geqslant 78$ months, the stratification rule starts to deviate substantially, perhaps due to the presence of extreme weights at low values of the censoring survival estimator $\hat{G}(\cdot)$.

7. Discussion

A common practice in predictive medicine is to create an ordered category system to classify future subjects according to their baseline information. A desirable quantitative stratification procedure would have both a small overall prediction error and a reasonable discriminatory capability across the strata. In this article, we provide a systematic approach to constructing such a stratification rule. The user can determine important components including the loss function, the minimum stratum size and the desired clinically meaningful difference between strata. Our method uses a heuristically interpretable metric, a loss function based on the L_1-norm, for quantifying the prediction error, but it could be replaced by, for example, mean squared error, integrated Brier score and other metrics for prediction performance (Graf et al., 1999; Choodari-Oskooei et al., 2012a,b). As a possible modification, one may also introduce a weight function $w(\cdot)$ in the modified loss $n^{-1}\sum_k \sum_{i\in S_k} |Y_i - \bar{Y}_k| w(V_i)$ to encourage finer or coarser stratification for high- or low-risk patients. The proposed stratification requires a minimum stratum size to avoid unstable small strata. The choice of this size depends on the amount of information in the training set Part Ia, which is usually quantified by the sample size or the observed event rate. In choosing its value, one also needs to consider the objective of the analysis. For example, if one wants to identify a small subgroup of high-risk individuals, the minimum stratum size needs to be smaller than the anticipated proportion of the high-risk stratum. To enhance the discriminatory ability of the scheme, we set a minimum incremental value between two consecutive stratum-specific predicted values. The choice of this value depends on clinical inputs. For example, for the cardiovascular study, the range of the restricted mean survival time scores is from 51·0 to 63·8 months based on the Part I training data, which is relatively narrow. A choice of an incremental value of three months for illustration in § 6 seems appropriate.

An obvious extension of the new proposal is to construct an optimal stratification procedure for treatment selections based on data either from randomized clinical trials or from observational studies. In such a case, one needs to predict the treatment effect measured by the difference of potential clinical outcomes of the patient under different treatments rather than the outcome itself. Unfortunately, the L_1 loss function utilized in this article cannot be generalized to deal with this important problem. Further research on the choice of a clinically meaningful metric for quantifying the prediction error for treatment selections is warranted.

Acknowledgement

This research was supported by the U.S. National Institutes of Health and the National Cancer Institute. We thank the editor, associate editor and two referees for their valuable and constructive comments.

Supplementary material

Supplementary material available at *Biometrika* online includes a more detailed account of the dynamic programming algorithm and the asymptotic properties of the optimal stratification scheme.

References

Braunwald, E., Domanski, M. J., Fowler, S. E., Geller, N. L., Gersh, B. J., Hsia, J., Pfeffer, M. A., Rice, M. M., Rosenberg, Y. D. & Rouleau, J. L. (2004). Angiotensin-converting-enzyme inhibition in stable coronary artery disease. *New Engl. J. Med.* **351**, 2058–68.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth.

Breslow, N. E. (1972). Discussion of Professor Cox's paper. *J. R. Statist. Soc.* B **34**, 216–7.

Choodari-Oskooei, B., Royston, P. & Parmar, M. K. (2012a). A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statist. Med.* **31**, 2627–43.

Choodari-Oskooei, B., Royston, P. & Parmar, M. K. (2012b). A simulation study of predictive ability measures in a survival model II: Explained randomness and predictive accuracy. *Statist. Med.* **31**, 2644–59.

Cox, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc.* B **34**, 187–220.

Friedman, J. H., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.

Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statist. Med.* **18**, 2529–45.

Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., Eron Jr, J. J., Feinberg, J. E., Balfour Jr, H. H., Deyton, L. R. et al. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New Engl. J. Med.* **337**, 725–33.

Li, J., Zhao, L., Tian, L., Cai, T., Claggett, B., Callegaro, A., Dizier, B., Spiessens, B., Ulloa-Montoya, F. & Wei, L. J. (2016). A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. *Biometrics* **72**, 877–87.

Royston, P. (2006). Explained variation for survival models. *Stata J.* **6**, 83–96.

Royston, P. & Parmar, M. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statist. Med.* **30**, 2409–21.

Solomon, S. D., Rice, M. M., Jablonski, K. A., Jose, P., Domanski, M., Sabatine, M., Gersh, B. J., Rouleau, J., Pfeffer, M. A., Braunwald, E. & Prevention of Events with ACE Inhibition (PEACE) Investigators (2006). Renal function and effectiveness of angiotensin-converting enzyme inhibitor therapy in patients with chronic stable coronary disease in the Prevention of Events with ACE inhibition (PEACE) trial. *Circulation* **114**, 26–31.

Taha, A. H. (2003). *Operations Research: An Introduction*. Upper Saddle River, New Jersey: Pearson Education.

Tian, L., Zhao, L. & Wei, L. J. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* **15**, 222–33.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **68**, 267–88.

Xu, Q. S. & Liang, Y. Z. (2001). Monte Carlo cross validation. *Chemomet. Intel. Lab. Syst.* **56**, 1–11.

Yong, F., Cai, T., Tian, L. & Wei, L. J. (2013). Classical model selection. *Handbook of Survival Analysis*, J. P. Klein, H. C. van Houwelingen, J. G. Ibrahim, & T. H. Scheike, eds. New York: CRC Press, pp. 265–83.

Zhao, L., Tian, L., Cai, T., Claggett, B. & Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. *J. Am. Statist. Assoc.* **108**, 527–39.

Zucker, D. M. (1998). Restricted mean life with covariates: Modification and extension of a useful survival analysis method. *J. Am. Statist. Assoc.* **93**, 702–9.

[*Received March* 2015. *Revised August* 2016]