Adapting machine learning techniques to censored time-to-event health record data: a general-purpose approach using inverse probability of censoring weighting
Supplementary Material David M. Vock, Julian Wolfson, Sunayan Bandyopadhyay, Gediminas Adomavicius, Paul E. Johnson, Gabriela Vazquez-Benitez, Patrick J. O'Connor

1. Toy example

The table gives the data used in the toy example considered in Section 2.3: Intuition of IPCW and a toy example. Recalling the notation given in Section 2.1, X_i: Value of the binary ; V_i: Time until the adverse event or when the subject is lost-to-followup whichever comes first; δ_i: Indicator for whether or not an adverse event occurred during the follow-up period; E_i: Indicator for whether or not an adverse event occurred within 5 year. Subjects for whom E_i cannot be determined because they do not have $\tau = 5$ years of data and did not experience an adverse event during their follow-up are denoted by a question mark in the table. $\hat{G}\{\min(V_i, \tau)\}$: Estimated probability of remaining uncensored at V_i or $\tau = 5$ whichever comes first; ω_i: Calculated from Equation 1 of the main text.

Table S.1: Toy example to illustrate the use of IPCW.

X_i	V_i	δ_i	E_1	$\min(V_i, \tau)$	$\hat{G}\{\min(V_i, \tau)\}$	ω_i
1	0.0	1	1	0.0	1.00	1.00
0	0.2	1	1	0.2	1.00	1.00
0	0.4	0	?	0.4	0.98	0.00
0	0.7	0	?	0.7	0.96	0.00
1	0.8	0	?	0.8	0.94	0.00
1	0.8	1	1	0.8	0.94	1.07
0	0.9	1	1	0.9	0.94	1.07
1	0.9	1	1	0.9	0.94	1.07
0	1.0	1	1	1.0	0.94	1.07
0	1.0	1	1	1.0	0.94	1.07
1	1.3	0	?	1.3	0.89	0.00
0	1.3	0	?	1.3	0.89	0.00
1	1.4	1	1	1.4	0.89	1.12
0	1.6	1	1	1.6	0.89	1.12
1	1.6	1	1	1.6	0.89	1.12
0	2.1	0	?	2.1	0.87	0.00
0	2.3	0	?	2.3	0.84	0.00
1	2.3	1	1	2.3	0.84	1.19
1	2.4	0	?	2.4	0.81	0.00
0	2.5	1	1	2.5	0.81	1.23
1	2.6	1	1	2.6	0.81	1.23
0	2.8	0	?	2.8	0.79	0.00
1	3.2	0	?	3.2	0.73	0.00
0	3.2	1	1	3.2	0.73	1.37
1	3.2	0	?	3.2	0.73	0.00
0	3.3	0	?	3.3	0.70	0.00
0	3.4	0	?	3.4	0.67	0.00
1	3.4	1	1	3.4	0.67	1.49
0	3.5	1	1	3.5	0.67	1.49
0	3.7	1	1	3.7	0.67	1.49
0	3.7	1	1	3.7	0.67	1.49
0	3.8	1	1	3.8	0.67	1.49
0	3.9	0	?	3.9	0.63	0.00
1	4.2	1	1	4.2	0.63	1.58
1	4.3	1	1	4.3	0.63	1.58
0	4.9	1	1	4.9	0.63	1.58
1	5.3	0	0	5.0	0.63	1.58
1	5.7	1	0	5.0	0.63	1.58
0	5.8	0	0	5.0	0.63	1.58
0	6.1	0	0	5.0	0.63	1.58
1	6.4	1	0	5.0	0.63	1.58
0	6.5	1	0	5.0	0.63	1.58
0	6.6	1	0	5.0	0.63	1.58
0	6.6	1	0	5.0	0.63	1.58
1	6.6	1	0	5.0	0.63	1.58
0	6.8	0	0	5.0	0.63	1.58
1	6.8	0	0	5.0	0.63	1.58
1	7.6	1	0	5.0	0.63	1.58
1	7.8	1	0	5.0	0.63	1.58
0	8.7	0	0	5.0	0.63	1.58

2. Technical details of IPCW

Formal proofs for IPCW are given in Robins and Finkelstein [1], Bang and Tsiatis [2], Tsiatis [3], but we aim here to give some more technical details of the method. If there were no censoring, class probability estimates (across all the machine learning scenarios considered in this manuscript) would be obtained using some form of maximum likelihood estimator (e.g., non-parametric, penalized, etc.). For example, in logistic regression, we parameterize the log odds (i.e., $\log[\pi_i(\mathbf{x})/\{1-\pi_i(\mathbf{x})\}]$) in terms of a linear combination of $\mathbf{x}$ and estimate the coefficients using maximum likelihood. In generalized additive logistic models, the log odds are related to a linear combination of $\mathbf{z}$, a basis expansion of $\mathbf{x}$, and regression coefficients are estimated using penalized maximum likelihood. In Bayesian networks, the continuous components of are modeled via maximum likelihood using a (mixture of) Gaussian densities (see section below in Supplementary Material), and the discrete components are modeled the non-parametric maximum likelihood estimator (i.e., conditional sample average). Once we identify the terminal nodes in binary trees or the neighbors in k-nearest neighbors, we compute probability estimates by taking sample averages within nodes; these averages are non-parametric maximum likelihood estimators. It has been well established that, as the sample size tends to infinity and other regularity conditions hold (including the number of observations in terminal nodes and number of neighbors also increase), maximum likelihood estimators consistently estimate the population parameters and, therefore, the risk probability [4, 5, 6].

Inverse probability of censoring weighted maximum likelihood estimators, "work" because the weighted log likelihood approximates the log-likelihood that we would have obtained had there not been censoring. Consider the general likelihood $\ell(\beta;\mathbf{X},\mathbf{E}) = \frac{1}{n}\sum_{i=1}^{n}\ell_i(\beta;\mathbf{X}_i,E_i)$ for the parameter β had no observations been right-censored. By the weak law of large numbers,

$$\tfrac{1}{n}\sum_{i=1}^{n}\ell_i(\beta;E_i,\mathbf{Z_i}) \xrightarrow{p} \mathcal{E}\{\ell_i(\beta;E_i,\mathbf{Z_i})\}$$

where $\mathcal{E}(\cdot)$ is the expectation and $\xrightarrow{p}$ denotes convergence in probability.

In the case of IPCW estimators, we maximize the likelihood $\frac{1}{n}\sum_{i=1}^{n}\omega_i\ell_i(\beta;E_i,\mathbf{Z_i})$. Note that we have

$$\omega_i = \mathbb{I}[\min(T_i,\tau) < C_i]/\hat{G}\{\min(T_i,\tau)\} \xrightarrow{p} \mathbb{I}[\min(T_i,\tau) < C_i]/G\{\min(T_i,\tau)\}$$

and hence it follows that

$$\begin{aligned}
\mathcal{E}\left\{\tfrac{1}{n}\sum_{i=1}^{n}\omega_i\ell_i(\beta;\mathbf{X}_i,E_i)\right\} &= \mathcal{E}\left\{\frac{\mathbb{I}\{\min(T_i,\tau)<C_i\}}{G\{\min(T_i,\tau)\}}\ell_i(\beta;\mathbf{X}_i,E_i)\right\}\\
&= \mathcal{E}\left(\mathcal{E}\left[\frac{\mathbb{I}\{\min(T_i,\tau)<C_i\}}{G\{\min(T_i,\tau)\}}\ell_i(\beta;E_i,\mathbf{Z_i})|\mathbf{X}_i,T_i\right]\right)\\
&= \mathcal{E}\left(\frac{\mathcal{E}[\mathbb{I}\{\min(T_i,\tau)<C_i\}|\mathbf{X}_i,T_i]}{G\{\min(T_i,\tau)\}}\ell_i(\beta;\mathbf{X}_i,E_i)\right)\\
&= \mathcal{E}\left(\frac{\mathcal{E}[\mathbb{I}\{\min(T_i,\tau)<C_i\}]}{G\{\min(T_i,\tau)\}}\ell_i(\beta;\mathbf{X}_i,E_i)\right)\\
&= \mathcal{E}\left(\frac{\mathcal{E}[G\{\min(T_i,\tau)\}]}{G\{\min(T_i,\tau)\}}\ell_i(\beta;\mathbf{X}_i,E_i)\right)\\
&= \mathcal{E}\left\{\ell_i(\beta;\mathbf{X}_i,E_i)\right\}
\end{aligned}$$

That is, for large samples the IPCW log-likelihood converges to the same quantity as the fully-observed (i.e., uncensored) likelihood. Because the difference in the IPCW and fully-observed likelihoods are (asymptotically) negligible, in large samples the IPCW approach inherits all the properties of machine learning estimators if we had full data. The above argument relies on the assumption that the censoring time C is independent of the event time T and all features $\mathbf{X}$. In our application, most patients are censored due to the end of the study or because they disenroll from the insurance plan due to a change in employment, reasons unrelated to their health status (i.e., $\mathbf{X}$ and T), so this independence assumption is reasonable.

How to handle this so-called "dependent censoring" is a current area of research in statistics [3], and to our knowledge very little of this work has been applied in the machine learning domain.

3. Technical details of Bayesian networks

A number of parametric and semi-parametric approaches to modeling the covariate distributions of continuous features are possible and have been described elsewhere [7, 8]; here describe a few possible techniques which we make use of in the data analysis. One common assumption is that the density of X_j given $\text{Pa}(X_j)$ and E is a normal density (or a mixture of normal densities). If the number of parents is small, then one could estimate a separate mean and variance parameter for each level of $\text{Pa}(X_j)$ and E. In this case the maximum likelihood estimators of the mean, $\mu_{j,\text{Pa}(x_j),e}$, and variance, $\Sigma_{j,\text{Pa}(x_j),e}$, given parents $\text{Pa}(x_j)$ and event status e are given by:

$$\begin{aligned}\mu_{j,\text{Pa}(x_j),e} &= \frac{\sum_{i=1}^n X_{ij}\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}}{\sum_{i=1}^n \mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}} \\ \Sigma_{j,\text{Pa}(x_j),e} &= \frac{\sum_{i=1}^n (X_{ij}-\mu_{j,\text{Pa}(x_j),e})(X_{ij}-\mu_{j,\text{Pa}(x_j),e})^T\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}}{\sum_{i=1}^n \mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}},\end{aligned} \tag{1}$$

If the number of parent nodes is large (or the nodes have several levels) there may be too few observations in some combinations of $\text{Pa}(X_j)$ to reliably estimate $\mu_{j,\text{Pa}(x_j),e}$ and $\Sigma_{j,\text{Pa}(x_j),e}$ in this way. Therefore, we may obtain more efficient estimators of the density of $X_j|\text{Pa}(X_j),E$ by using a regression model. In particular, we might assume that the mean of the conditional density of X_j given $\text{Pa}(X_j)$ and E is related to the levels of the parent nodes and event status through an additive model and that the conditional variance is constant across all levels $\text{Pa}(X_j)$ and E. For example, if the m_j parent nodes of X_j are denoted by $PX_{j1},\ldots PX_{j,m_j}$ and the k^{th} parent node has p_k levels denoted generically as $1,\ldots,p_k$ (recall that in our application the parent nodes are discrete), we could assume that $X_j|\text{Pa}(X_j),E$ has mean $\beta_{0j}+\sum_{k=1}^{m_j}\sum_{l=1}^{p_k-1}\beta_{jkl}\mathbb{I}(PX_k=l)$ and that the conditional variance σ_j^2 is constant across all levels of the parents. Then the log-likelihood takes the following form which can be solved to obtain the maximum likelihood estimators of $\beta_j=[\beta_{0j},\{\beta_{jkl}\}_{k=1,\ldots,m_j,l=1,\ldots,p_k-1}]$ and σ_j^2:

$$\ell(\beta_j,\sigma_j^2;\mathbf{X},\mathbf{E})=\sum_{i=1}^n\left(\frac{-1}{2\sigma_j^2}\left[X_{ij}-\{\beta_{0j}+\sum_{k=1}^{m_j}\sum_{l=1}^{p_k-1}\beta_{jkl}\mathbb{I}(PX_k=l)\}\right]^2-\frac{1}{2}\log(2\pi\sigma_j^2)\right), \tag{2}$$

The IPCW estimators of the mean and variance for continuous variables X_j can be obtained using a weighted maximum likelihood where the contribution of the i^{th} subject to the likelihood is weighted by ω_i. The formulas for the parameter estimates previously given in Eq. (1) become

$$\begin{aligned}\mu_{j,\text{Pa}(x_j),e} &= \frac{\sum_{i=1}^n \omega_i X_{ij}\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}}{\sum_{i=1}^n \omega_i\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}} \\ \Sigma_{j,\text{Pa}(x_j),e} &= \frac{\sum_{i=1}^n \omega_i(X_{ij}-\mu_{j,\text{Pa}(x_j),e})(X_{ij}-\mu_{j,\text{Pa}(x_j),e})^T\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}}{\sum_{i=1}^n \omega_i\mathbb{I}\{\text{Pa}(X_{ij})=\text{Pa}(x_{ij}),E_i=e\}}.\end{aligned} \tag{3}$$

Similarly, the contribution of the i^{th} subject to the log-likelihood for the regression parameters in (2) is weighted by ω_i so that the regression parameter estimates maximize the following weighted log-likelihood:

$$\ell^{\omega}(\beta_j,\sigma_j^2;\mathbf{X},\mathbf{E})=\sum_{i=1}^n\omega_i\left(-\frac{1}{2\sigma_j^2}\left[X_{ij}-\{\beta_{0j}+\sum_{k=1}^{m_j}\sum_{l=1}^{p_k-1}\beta_{jkl}\mathbb{I}(PX_{ik}=l)\}\right]^2-\frac{1}{2}\log(2\pi\sigma_j^2)\right), \tag{4}$$

4. Model results

Below are the coefficient estimates for the IPCW logistic regression model and the final constructed decision tree (after choosing the tuning parameters as discussed in the main text). Because the decision tree is relatively large we did not have space within the graphic to display the CV event probabilities in each node. The terminal nodes "TRUE" indicates a CV event and FALSE indicates no CV event.

Table S.2: Coefficient estimates for the IPCW logistic regression model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.96	0.22	-41.60	<0.001
Female (ref: Male)	-0.34	0.0370	-9.32	<0.001
Age (per 10 year increase)	0.72	0.015	47.02	<0.001
SBP (per 10 mmHg increase)	0.045	0.014	3.21	<0.001
BMI (per 5 kg/m^2 increase)	0.11	0.015	7.11	<0.001
HDL (per 10 mg/dL increase)	-0.023	0.013	-1.72	0.086
TC (per 10 mg/dL increase)	-0.003	0.005	-0.69	0.49
Diabetes	0.31	0.051	6.20	< 0.001
Current smoker (ref: non-smoker)	0.77	0.051	15.25	< 0.001
Former smoker (ref: non-smoker)	0.032	0.055	0.58	0.56
Takeing SBP medication	1.22	0.052	23.65	<0.001

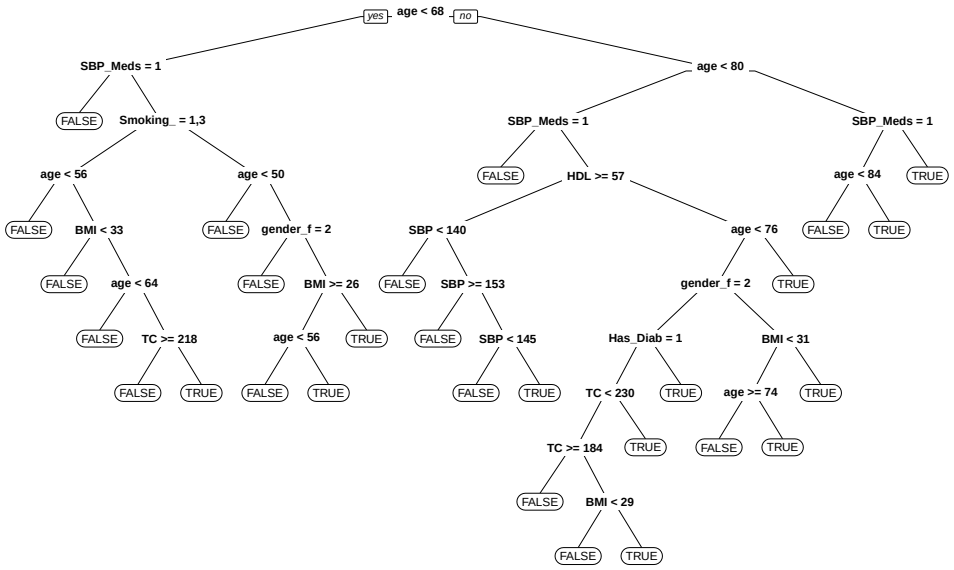

Figure S.1: Final decision tree using IPCW to account for right censoring. The terminal nodes "TRUE" indicates a CV event and "FALSE" indicates no CV event.

References

[1] J. M. Robins, D. M. Finkelstein, Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests, Biometrics 56 (3) (2000) 779–788.

[2] H. Bang, A. A. Tsiatis, Estimating medical costs with censored data, Biometrika 87 (2) (2000) 329–343.

[3] A. A. Tsiatis, Semiparametric Theory and Missing Data, Springer, New York, 2006.

[4] C. Sundberg, T. Aulin, N. Rydbeck, The rate of convergence of k-NN regression estimates and classification rules, IEEE Transactions on Information Theory 20 (1972) 429–435.

[5] E. Lehmann, G. Casella, Theory of Point Estimation, vol. 31, Springer, 1998.

[6] W.-Y. Loh, Classification and regression tree methods, Encyclopedia of Statistics in Quality and Reliability .

[7] P. Domingos, M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29 (1997) 103–130, URL `http://link.springer.com/article/10.1023/A:1007413511361`.

[8] G. H. John, P. Langley, Estimating Continuous Distributions in Bayesian Classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995.