Supplementary for

Frailty-aware extreme gradient boosting for high-dimensional clustered time-to-event outcomes

Saheed Kunle Ajibade, Oyebayo Ridwan Olaniran, Samuel Adewale Aderoju, Ayodeji Idowu, Nada MohammedSaeed Alharbi and Abdulhameed Bolakale

Contents

S1 Proofs **1**
S1.1 The boosting step 1
S1.2 The expectation step 2
S1.3 Concordance 3

S2 Additional simulation results **4**

S3 Tutorial: fitting MEXGB to clustered survival data **6**

S1 Proofs

We use the notation introduced in the main text. $\tau_1 < \cdots < \tau_K$ are the distinct event times, d_k the number of events at τ_k, $\mathcal{R}_k$ the risk set at τ_k, $S_k(\boldsymbol{\eta}) = \sum_{(i,j)\in\mathcal{R}_k} e^{\eta_{ij}}$, and $\mathcal{L}_{\mathrm{PL}}$ the negative log partial likelihood (3) of the main text. Numbered statements without the prefix S refer to the main manuscript.

S1.1 The boosting step

Lemma S1 (Gradient and Hessian of the partial likelihood)**.** *Let $\hat{\Lambda}_0(t) = \sum_{k:\tau_k\le t} d_k/S_k$ be the Breslow estimator at the current $\boldsymbol{\eta}$ and $\mu_{ij} = \hat{\Lambda}_0(Y_{ij})e^{\eta_{ij}}$. Then*

$$\frac{\partial\mathcal{L}_{\mathrm{PL}}}{\partial\eta_{ij}} = \mu_{ij} - \delta_{ij}, \qquad \frac{\partial^2\mathcal{L}_{\mathrm{PL}}}{\partial\eta_{ij}^2} = \mu_{ij} - \sum_{k:\tau_k\le Y_{ij}} d_k\frac{e^{2\eta_{ij}}}{S_k^2} \in (0, \mu_{ij}],$$

The lower bound is strict under the stated condition that every risk set containing (i, j) also contains another unit.

Proof. Unit (i, j) belongs to $\mathcal{R}_k$ if and only if $\tau_k \le Y_{ij}$, so $\partial S_k/\partial\eta_{ij} = e^{\eta_{ij}}\mathbb{1}\{\tau_k \le Y_{ij}\}$. Differentiating equation (3) gives

$$\frac{\partial\mathcal{L}_{\mathrm{PL}}}{\partial\eta_{ij}} = -\delta_{ij} + \sum_k d_k\frac{\partial S_k/\partial\eta_{ij}}{S_k} = -\delta_{ij} + e^{\eta_{ij}}\sum_{k:\tau_k\le Y_{ij}}\frac{d_k}{S_k} = \mu_{ij} - \delta_{ij}.$$

Differentiating the summand $d_k e^{\eta_{ij}}/S_k$ once more with respect to η_{ij} gives

$$d_k\Big\{\frac{e^{\eta_{ij}}}{S_k} - \frac{e^{2\eta_{ij}}}{S_k^2}\Big\} = d_k\frac{e^{\eta_{ij}}}{S_k}\Big(1 - \frac{e^{\eta_{ij}}}{S_k}\Big).$$

Since $0 < e^{\eta_{ij}} \le S_k$, each term lies in $[0, d_k e^{\eta_{ij}}/S_k]$, with strict positivity when $\mathcal{R}_k$ contains another unit; summing over k gives the stated range. □

The leaf-weight expression in equation (5) of the main text follows from the second-order objective of `xgboost`: a tree with leaf weights w_ℓ changes the loss by approximately $\sum_\ell\{G_\ell w_\ell + \frac{1}{2}(H_\ell + \lambda)w_\ell^2\}$ with $G_\ell = \sum_{I_\ell} g_{ij}$ and $H_\ell = \sum_{I_\ell} h_{ij}$, a sum of strictly convex quadratic functions, each minimised at $w_\ell^\star = -G_\ell/(H_\ell + \lambda)$.

Proposition S2 (Profile Poisson likelihood)**.** *Let λ_0 be a step function with jumps $\lambda_k \geq 0$ at the event times and $\Lambda_0(t) = \sum_{\tau_k \leq t} \lambda_k$. The full log-likelihood given the linear predictors, $\ell(\boldsymbol{\eta}, \boldsymbol{\lambda}) = \sum_{ij} \{\delta_{ij} \log(\lambda_{k(ij)} e^{\eta_{ij}}) - \Lambda_0(Y_{ij}) e^{\eta_{ij}}\}$, where $k(ij)$ is the index of the event time Y_{ij} of a failed unit, is maximised over $\boldsymbol{\lambda}$ by the Breslow jumps $\hat{\lambda}_k = d_k / S_k(\boldsymbol{\eta})$, and the profile log-likelihood equals $-\mathcal{L}_{\mathrm{PL}}(\boldsymbol{\eta})$ plus a constant. As a function of $\boldsymbol{\eta}$, the full log-likelihood is the log-likelihood of independent Poisson variables δ_{ij} with means $\mu_{ij} = \Lambda_0(Y_{ij}) e^{\eta_{ij}}$, up to a term that does not depend on $\boldsymbol{\eta}$.*

Proof. Reversing the order of summation gives $\sum_{ij} \Lambda_0(Y_{ij}) e^{\eta_{ij}} = \sum_{ij} \sum_{k : \tau_k \leq Y_{ij}} \lambda_k e^{\eta_{ij}} = \sum_k \lambda_k S_k$. Hence

$$\ell = \sum_{ij} \delta_{ij} \eta_{ij} + \sum_k d_k \log \lambda_k - \sum_k \lambda_k S_k,$$

which is strictly concave in each λ_k with derivative $d_k / \lambda_k - S_k$, vanishing at $\hat{\lambda}_k = d_k / S_k$. Substituting, $\ell(\boldsymbol{\eta}, \hat{\boldsymbol{\lambda}}) = \sum \delta_{ij} \eta_{ij} - \sum_k d_k \log S_k + \sum_k (d_k \log d_k - d_k) = -\mathcal{L}_{\mathrm{PL}}(\boldsymbol{\eta}) + \text{const}$. To establish the second claim, write each summand as $\delta \log(\lambda_k e^{\eta}) - \Lambda_0(Y) e^{\eta} = \{\delta \log \mu - \mu\} + \delta \log\{\lambda_k / \Lambda_0(Y)\}$ with $\mu = \Lambda_0(Y) e^{\eta}$; the first bracketed term is the Poisson log-likelihood of δ with mean μ (up to $-\log \delta! = 0$), and the last term does not involve η [1]. □

Corollary S3 (Fisher consistency of the boosting step)**.** *Suppose model (1) holds, Λ_0 is known, and the offsets are the true frailty contributions $o = \boldsymbol{z}^\top \boldsymbol{b}$. Then, over measurable g, the expected Poisson log-likelihood $\mathbb{E}[\delta \{g(\boldsymbol{x}) + o\} - \Lambda_0(Y) e^{g(\boldsymbol{x}) + o}]$ is maximised by $g = f$, uniquely up to null sets.*

Proof. Condition on $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{b})$. Under (1) the counting process $N(t) = \mathbb{1}\{Y \leq t, \delta = 1\}$ has compensator $\int_0^t \mathbb{1}\{Y \geq u\} \lambda_0(u) e^{f(\boldsymbol{x}) + o} du = \Lambda_0(Y \wedge t) e^{f(\boldsymbol{x}) + o}$, so $\mathbb{E}[\delta \mid \cdot] = e^{f(\boldsymbol{x}) + o} \, \mathbb{E}[\Lambda_0(Y) \mid \cdot]$. Writing $A = \mathbb{E}[\Lambda_0(Y) \mid \cdot] > 0$, the conditional expectation of the log-likelihood at η is $\psi(\eta) = A e^{f + o} \eta - A e^{\eta}$, strictly concave with derivative $A(e^{f + o} - e^{\eta})$ vanishing only at $\eta = f(\boldsymbol{x}) + o$. Taking the expectation over the conditioning variables establishes the result. □

Corollary S4 (Cox PQL expectation step)**.** *Applying penalised quasi-likelihood to the Poisson representation with log link gives the working response and weight (7) of the main text, $\tilde{y}_{ij} = \hat{\eta}_{ij} + (\delta_{ij} - \hat{\mu}_{ij}) / \hat{\mu}_{ij}$ and $w_{ij} = \hat{\mu}_{ij}$, with $\hat{\mu}_{ij} = \hat{\Lambda}_0(Y_{ij}) e^{\hat{\eta}_{ij}}$. To first order, $\tilde{y}_{ij} = \eta_{ij} + \xi_{ij}$ with $\mathbb{E}\xi_{ij} = 0$ and $\operatorname{Var} \xi_{ij} = 1 / w_{ij}$.*

Proof. For a Poisson mean with log link, $\mu(\eta) = e^{\eta} \Lambda_0(Y)$, $\partial \mu / \partial \eta = \mu$ and $\operatorname{Var}(\delta) = \mu$. A first-order expansion at $\hat{\eta}$ gives $\mu = \hat{\mu} + \hat{\mu}(\eta - \hat{\eta}) + O\{(\eta - \hat{\eta})^2\}$. Writing $\delta = \mu + e$ with $\mathbb{E}(e \mid \boldsymbol{b}) = 0$ and $\operatorname{Var}(e \mid \boldsymbol{b}) = \mu$, $\tilde{y} = \hat{\eta} + (\mu - \hat{\mu} + e) / \hat{\mu} = \eta + e / \hat{\mu} + O\{(\eta - \hat{\eta})^2\}$, so $\xi = e / \hat{\mu}$ has mean zero and, to first order, variance $1 / \hat{\mu}$. Profiling Λ_0 at every linearisation (Proposition S2) makes the working likelihood that of the penalised partial likelihood [2]. □

S1.2 The expectation step

Proposition S5 (Non-identifiability and centring of the frailty mean)**.** *Let $\boldsymbol{z}_{ij}$ contain an intercept. Replacing every random intercept b_{i0} by $b_{i0} + c$ leaves $\mathcal{L}_{\mathrm{PL}}$ and the Breslow-profiled Poisson means $\hat{\mu}_{ij}$ unchanged and multiplies the Breslow estimator by e^{-c}. Consequently, the partial likelihood does not identify the random-intercept mean. The covariance update in equation (9) of the main text satisfies*

$$\hat{D}_{00} = \frac{1}{n} \sum_i (\hat{b}_{i0} - \bar{b}_0)^2 + \bar{b}_0^2 + \frac{1}{n} \sum_i C_{i,00},$$

Thus, drift in $\bar{b}_0$ inflates $\hat{D}_{00}$. Centring, $\hat{b}_{i0} \leftarrow \hat{b}_{i0} - \bar{b}_0$, enforces $\mathbb{E} b_{i0} = 0$ without changing any fitted hazard.

Proof. A common shift adds c to every η_{ij}. In (3), $-\sum \delta_{ij} \eta_{ij}$ changes by $-c \sum_k d_k$ and $\sum_k d_k \log S_k$ by $+c \sum_k d_k$, because every S_k is multiplied by e^c; the partial likelihood is invariant. The Breslow jumps d_k / S_k are multiplied by e^{-c} and $\hat{\mu}_{ij} = \hat{\Lambda}_0(Y_{ij}) e^{\eta_{ij}}$ is unchanged, as are the fitted hazards $\hat{\lambda}_0(t) e^{\eta_{ij}}$. The decomposition of $\hat{D}_{00}$ is the identity $n^{-1} \sum_i b_i^2 = n^{-1} \sum_i (b_i - \bar{b})^2 + \bar{b}^2$ applied to the first diagonal element of (8). □

Proposition S6 (The dispersion absorbs the cross-fitting error)**.** *In the working model of the expectation step, suppose* $\tilde{\gamma}_{\mathrm{c}}\hat{f}_{\mathrm{cf}}(\boldsymbol{x}_{ij}) = f(\boldsymbol{x}_{ij})+c+e_{ij}$ *for a constant* c *and errors* e_{ij} *with mean zero and variance* τ^2*, independent of the frailties and of the working errors. Then the scaled working residual* $w_{ij}^{1/2}(\tilde{y}_{ij}-\tilde{\gamma}_{\mathrm{c}}\hat{f}_{\mathrm{cf},ij}-\boldsymbol{z}_{ij}^{\top}\boldsymbol{b}_i)$ *has variance* $1+w_{ij}\tau^2$ *to first order. A working model with* $\phi\equiv 1$ *attributes the excess* $w_{ij}\tau^2$ *to the frailties and inflates* $\hat{\boldsymbol{D}}$*, whereas the dispersion estimate converges to the weighted average of* $1+w_{ij}\tau^2$*; when the weights are constant within clusters the frailty covariance is unaffected.*

Proof. By Corollary S4, $\tilde{y}_{ij} = \eta_{ij} + \xi_{ij}$ with $\operatorname{Var}\xi_{ij} = 1/w_{ij}$ to first order. The residual is $\xi_{ij} - e_{ij} - c$, whose variance is $1/w_{ij} + \tau^2$; multiplying by $w_{ij}^{1/2}$ gives $1 + w_{ij}\tau^2$. If $w_{ij} \equiv w_i$ within cluster i, the scaled residuals of the cluster have covariance $w_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^{\top} + (1+w_i\tau^2)\boldsymbol{I}$, which is the working model with dispersion $\phi_i = 1+w_i\tau^2$; the frailty part of the covariance is unchanged, and the residual-variance update (9) estimates the dispersion. The constant c is removed by the centring of Proposition S5. □

Proposition S7 (Calibration slope)**.** *(i) The log partial likelihood of the one-parameter model* $\eta_{ij} = o_{ij} + \gamma u_{ij}$*,* $u_{ij} = \hat{f}_{\mathrm{cf}}(\boldsymbol{x}_{ij})$*, is concave in* γ*, and is strictly concave unless* u *is constant on every risk set. Under strict concavity, the slope estimate is unique whenever a finite maximiser exists. (ii) If* $u_{ij} = sf(\boldsymbol{x}_{ij})$ *for some* $s > 0$ *and the offsets are the true frailty contributions,* $\hat{\gamma}_{\mathrm{c}} \to 1/s$ *in probability, and the calibrated log-hazard recovers* f*. (iii) In general,* $\hat{\gamma}_{\mathrm{c}}$ *converges to the slope that minimises the Kullback–Leibler divergence between the true conditional law and the one-parameter working model.*

Proof. (i) The second derivative of $-\mathcal{L}_{\mathrm{PL}}$ with respect to γ is

$$-\sum_k d_k\Big\{\frac{\sum_{\mathcal{R}_k} u^2 e^{\eta}}{S_k} - \Big(\frac{\sum_{\mathcal{R}_k} u e^{\eta}}{S_k}\Big)^2\Big\},$$

which is the negative of a sum of weighted variances of u over the risk sets, which is non-positive and negative unless u is constant on every risk set with an event. (ii) Under the stated conditions, the one-parameter model is correctly specified at $\gamma = 1/s$, and the maximiser of a strictly concave, correctly specified log partial likelihood is consistent [3]. (iii) The normalised log partial likelihood converges uniformly on compacts to a strictly concave limit whose maximiser is the Kullback–Leibler projection; the argmax of a sequence of concave functions converging pointwise to a function with a unique maximiser converges to it. □

S1.3 Concordance

Proof of Lemma 1.

$$q_{kl} = \int_0^{\tau} h_k\lambda_0(t)e^{-h_k\Lambda_0(t)}e^{-h_l\Lambda_0(t)}dt = \frac{h_k}{h_k+h_l}\int_0^{\tau}(h_k+h_l)\lambda_0(t)e^{-(h_k+h_l)\Lambda_0(t)}dt = \frac{h_k}{h_k+h_l}\big\{1-e^{-(h_k+h_l)\Lambda_0(\tau)}\big\},$$

because the last integrand is the density of a failure time with hazard $(h_k+h_l)\lambda_0$. The factor in braces is symmetric in (k,l), so $q_{kl}-q_{lk} = \{(h_k-h_l)/(h_k+h_l)\}\{1-e^{-(h_k+h_l)\Lambda_0(\tau)}\}$ has the sign of h_k-h_l. For $\tau=\infty$, $q_{kl} = h_k/(h_k+h_l)$ irrespective of the baseline hazard under the stated survival-model assumptions. □

Proof of Theorem 1. By conditional independence, $\mathbb{P}(T_k < T_l,\, T_k < \tau \mid \eta_k, \eta_l, \text{information}) = q_{kl}$ with $h = e^{\eta}$. The denominator of $C_\tau(s)$ is $\mathbb{P}(T_k < T_l,\, T_k < \tau) = \mathbb{E}q_{kl}$, which does not depend on s, and the numerator is

$$\mathcal{N}(s) = \mathbb{E}\big[\mathbb{1}\{s_k > s_l\}q_{kl}\big] + \tfrac{1}{2}\mathbb{E}\big[\mathbb{1}\{s_k = s_l\}q_{kl}\big].$$

Because (k,l) is an exchangeable pair, $\mathcal{N}(s)$ also equals $\mathbb{E}[\mathbb{1}\{s_l > s_k\}q_{lk}] + \frac{1}{2}\mathbb{E}[\mathbb{1}\{s_k = s_l\}q_{lk}]$. Averaging the two expressions,

$$\mathcal{N}(s) = \tfrac{1}{2}\mathbb{E}\big[\mathbb{1}\{s_k > s_l\}q_{kl} + \mathbb{1}\{s_l > s_k\}q_{lk} + \tfrac{1}{2}\mathbb{1}\{s_k = s_l\}(q_{kl}+q_{lk})\big].$$

Let $M_{kl} = \max(q_{kl}, q_{lk})$. By Lemma 1 the larger of q_{kl} and q_{lk} corresponds to the unit with the larger η failing first, and the smaller equals $M_{kl} - |q_{kl}-q_{lk}|$. The bracket therefore equals M_{kl} when s orders the pair as η does, $M_{kl}-|q_{kl}-q_{lk}|$ when the pair is discordant, and $M_{kl}-\frac{1}{2}|q_{kl}-q_{lk}|$ when s assigns equal scores to the pair; when $\eta_k = \eta_l$ the three values coincide because $q_{kl} = q_{lk}$. Hence the bracket equals $M_{kl} - |q_{kl}-q_{lk}|\{\mathbb{1}^{\mathrm{d}}_{kl}(s) + \frac{1}{2}\mathbb{1}^{\mathrm{t}}_{kl}(s)\}$ for every score and M_{kl} for $s=\eta$. Subtracting, $\mathcal{N}(\eta)-\mathcal{N}(s) = \frac{1}{2}\mathbb{E}[|q_{kl}-q_{lk}|\{\mathbb{1}^{\mathrm{d}}_{kl}(s)+\frac{1}{2}\mathbb{1}^{\mathrm{t}}_{kl}(s)\}]$, and dividing by $\mathbb{E}q_{kl}$ gives (12). For $\tau=\infty$, $q_{kl} = e^{\eta_k}/(e^{\eta_k}+e^{\eta_l})$, $q_{lk} = 1-q_{kl}$, and $\mathbb{E}q_{kl} = \frac{1}{2}$ by exchangeability. As a check, for $s=-\eta$ and $\tau=\infty$ every pair with $\eta_k \neq \eta_l$ is discordant and (12) gives $C(\eta)-C(-\eta) = \mathbb{E}|2q_{kl}-1| = 2C(\eta)-1$, the identity $C(-\eta) = 1-C(\eta)$. □

Proof of Corollary 1. Consider two units of different clusters and suppose $s_k > s_l$. Write $\eta_k - \eta_l = \{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_l)\} + w$ with $w = \boldsymbol{z}_k^\top \boldsymbol{b}_k - \boldsymbol{z}_l^\top \boldsymbol{b}_l$. Given the predictors and designs, w has a positive density on $\mathbb{R}$ because $\boldsymbol{D} \neq \boldsymbol{0}$ and the frailties of different clusters are independent, so $\mathbb{P}(\eta_k < \eta_l \mid \boldsymbol{x}_k, \boldsymbol{x}_l, \boldsymbol{z}_k, \boldsymbol{z}_l) > 0$; on this event the pair is discordant and $|q_{kl} - q_{lk}| > 0$ by Lemma 1, because $\Lambda_0(\tau) > 0$. The same holds for $s_k < s_l$ with the roles exchanged, and when $s_k = s_l$ the pair is tied by s with $\eta_k \neq \eta_l$ almost surely. Pairs of units of different clusters have positive probability, so the numerator of (12) is positive. □

Proof of Corollary 2. Write $w = \boldsymbol{z}_k^\top \boldsymbol{b}_k - \boldsymbol{z}_l^\top \boldsymbol{b}_l$, $A = (\boldsymbol{x}_k - \boldsymbol{x}_l)^\top \boldsymbol{\beta}$ and $B = f(\boldsymbol{x}_k) - f(\boldsymbol{x}_l)$, so that $s_k - s_l = A + w$ and $\eta_k - \eta_l = B + w$. Since $u(\boldsymbol{x}) = f(\boldsymbol{x}) - \boldsymbol{x}^\top \boldsymbol{\beta}$ is not almost surely constant and $\boldsymbol{x}_k, \boldsymbol{x}_l$ are independent copies, $B - A = u(\boldsymbol{x}_k) - u(\boldsymbol{x}_l) \neq 0$ with positive probability. On that event, the open interval between $-A$ and $-B$ has positive length, and w, which is independent of the predictors and has a positive density, falls in it with positive conditional probability. Then $A + w$ and $B + w$ have opposite signs: the pair is discordant and $\eta_k \neq \eta_l$, so $|q_{kl} - q_{lk}| > 0$. The numerator of (12) is therefore positive. □

Proof of Proposition 1. Conditional on the information used for prediction, $\mathbb{1}(T > t)$ is Bernoulli with mean $S(t)$, and $\hat{S}(t)$ is fixed, so $\mathbb{E}[\{\mathbb{1}(T > t) - \hat{S}(t)\}^2 \mid \cdot] = S(t)\{1 - S(t)\} + \{S(t) - \hat{S}(t)\}^2$; taking expectations gives the decomposition, and the Brier score is minimised by $\hat{S} = S$. With known frailties and baseline, $S(t) = \exp\{-\Lambda_0(t)e^{\eta}\}$ and $\hat{S}(t) = \exp\{-\Lambda_0(t)e^{\hat{\eta}}\}$; for $\Lambda_0(t) > 0$ the map $x \mapsto \exp\{-\Lambda_0(t)e^{x}\}$ is strictly decreasing, so the integrand of the excess Brier score vanishes if and only if $\hat{\eta} = \eta$, that is, $(s - 1)f(\boldsymbol{x}) = 0$, which holds almost surely only if $s = 1$ when f is not almost surely zero. Without a frailty, $\hat{\eta} = sf$ is a strictly increasing transformation of $\eta = f$, so the pairwise ordering, and hence concordance, does not depend on s. □

With frailties, $\hat{\eta}_k - \hat{\eta}_l = s\{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_l)\} + w$ and $\eta_k - \eta_l = \{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_l)\} + w$ can have opposite signs when $s \neq 1$, by the argument of Corollary 2; a scale error can therefore reduce concordance as well as affect probability calibration. The calibration slope is intended to correct that scale error under the stated assumptions.

S2 Additional simulation results

Table S1: Simulation, nonlinear log-hazard, $p = 170$: test C-index and integrated Brier score, mean (standard error) over 100 splits, by censoring rate. The best value in each column is shown in bold.

	C-index		Integrated Brier score	
Method	30%	50%	30%	50%
MEXGB	**0.797 (0.002)**	**0.821 (0.002)**	**0.132 (0.001)**	**0.117 (0.001)**
XGBCox	0.732 (0.003)	0.762 (0.002)	0.166 (0.001)	0.144 (0.001)
GBM	0.737 (0.002)	0.763 (0.002)	0.160 (0.001)	0.142 (0.001)
RSF	0.710 (0.002)	0.736 (0.003)	0.195 (0.001)	0.176 (0.001)
mboost	0.624 (0.003)	0.642 (0.003)	0.198 (0.001)	0.179 (0.001)
CoxLasso	0.626 (0.003)	0.644 (0.003)	0.198 (0.001)	0.179 (0.001)

Table S2: Simulation: true positive rate (%) of the six top-ranked predictors, mean (standard error). The best value in each column is shown in bold.

	Linear			Nonlinear		
Method	$p = 6$	$p = 170$	$p = 2000$	$p = 6$	$p = 170$	$p = 2000$
MEXGB	**100.0 (0.0)**	**100.0 (0.0)**	**91.0 (0.9)**	94.7 (0.8)	54.3 (0.8)	51.3 (0.5)
CoxFrailty	**100.0 (0.0)**	–	–	**100.0 (0.0)**	–	–
XGBCox	**100.0 (0.0)**	87.7 (0.8)	70.7 (1.1)	**100.0 (0.0)**	54.2 (0.8)	53.3 (0.8)
GBM	**100.0 (0.0)**	85.8 (0.9)	68.3 (1.0)	**100.0 (0.0)**	52.3 (0.6)	51.3 (0.6)
RSF	99.8 (0.2)	71.2 (0.9)	47.7 (1.1)	95.8 (0.8)	**80.7 (1.0)**	**67.5 (1.1)**
mboost	**100.0 (0.0)**	98.5 (0.5)	87.8 (0.8)	87.2 (1.1)	50.5 (0.6)	46.5 (0.8)
CoxLasso	**100.0 (0.0)**	**100.0 (0.0)**	89.5 (0.8)	79.7 (1.7)	51.2 (0.7)	44.3 (1.0)

Table S3: Simulation: false positive rate (%), mean (standard error), for $p \geq 170$ (at $p = 6$ all predictors are active). The best value in each column is shown in bold.

	Linear		Nonlinear	
Method	$p = 170$	$p = 2000$	$p = 170$	$p = 2000$
MEXGB	**0.00 (0.00)**	**0.02 (0.00)**	1.54 (0.04)	0.13 (0.00)
XGBCox	0.45 (0.03)	0.09 (0.00)	1.68 (0.03)	0.14 (0.00)
GBM	0.52 (0.03)	0.10 (0.00)	1.74 (0.02)	0.15 (0.00)
RSF	1.05 (0.03)	0.16 (0.00)	**0.71 (0.04)**	**0.10 (0.00)**
mboost	0.05 (0.02)	0.04 (0.00)	1.81 (0.02)	0.16 (0.00)
CoxLasso	**0.00 (0.00)**	0.03 (0.00)	1.74 (0.03)	0.16 (0.00)

Table S4: Simulation: elapsed time per fit (seconds, one core), mean (standard error). The best value in each column is shown in bold.

	Linear			Nonlinear		
Method	$p = 6$	$p = 170$	$p = 2000$	$p = 6$	$p = 170$	$p = 2000$
MEXGB	10.5 (0.2)	38.3 (0.9)	585.7 (24.9)	13.9 (0.2)	37.3 (0.9)	798.8 (28.7)
CoxFrailty	3.7 (0.1)	–	–	3.8 (0.1)	–	–
XGBCox	0.2 (0.0)	2.5 (0.0)	27.0 (0.5)	0.2 (0.0)	2.2 (0.0)	25.6 (0.5)
GBM	**0.1 (0.0)**	1.3 (0.0)	17.1 (0.4)	**0.1 (0.0)**	1.2 (0.0)	16.5 (0.4)
RSF	3.2 (0.0)	24.4 (0.2)	144.8 (1.4)	3.5 (0.1)	23.6 (0.3)	142.1 (2.1)
mboost	2.0 (0.0)	1.8 (0.0)	**2.5 (0.1)**	1.9 (0.0)	1.7 (0.0)	**2.3 (0.0)**
CoxLasso	**0.1 (0.0)**	**0.4 (0.0)**	6.6 (0.2)	**0.1 (0.0)**	**0.4 (0.0)**	8.6 (0.3)

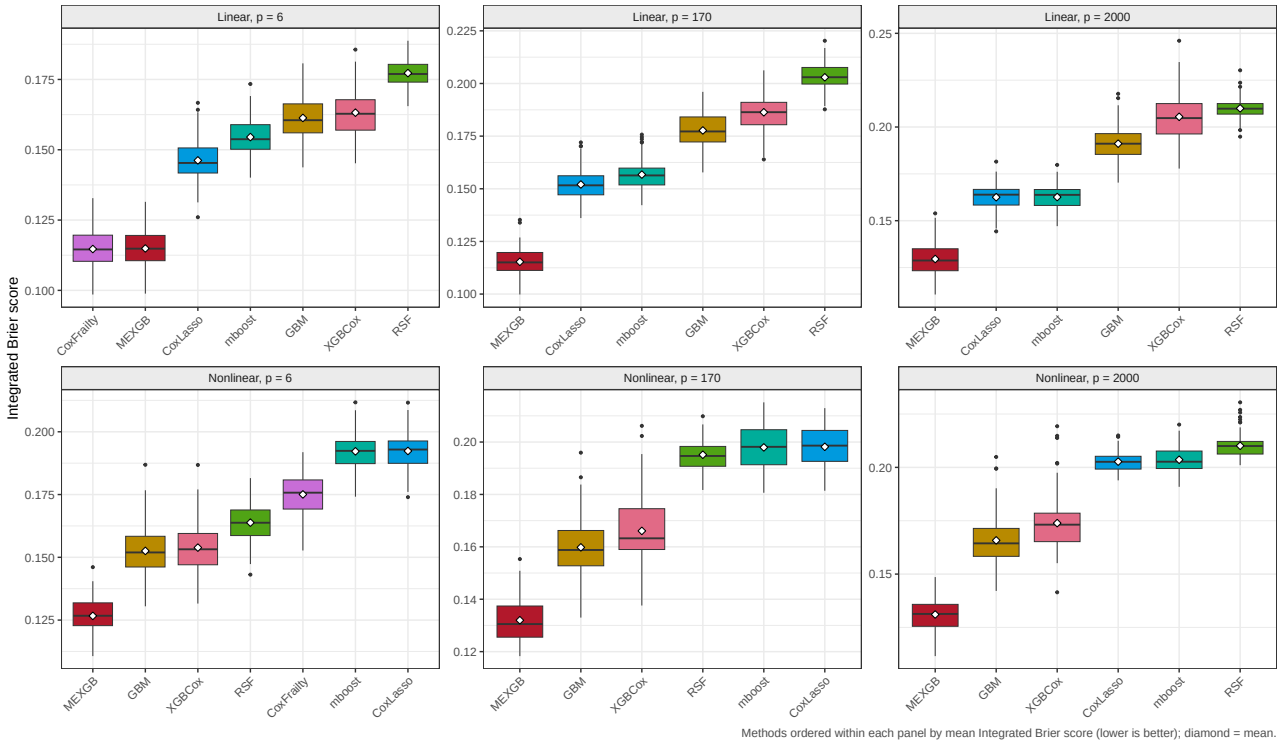

Figure S1: Simulation: test integrated Brier score over the 100 splits of each scenario (30% censoring), with methods ordered by their mean within each panel (lower is better).

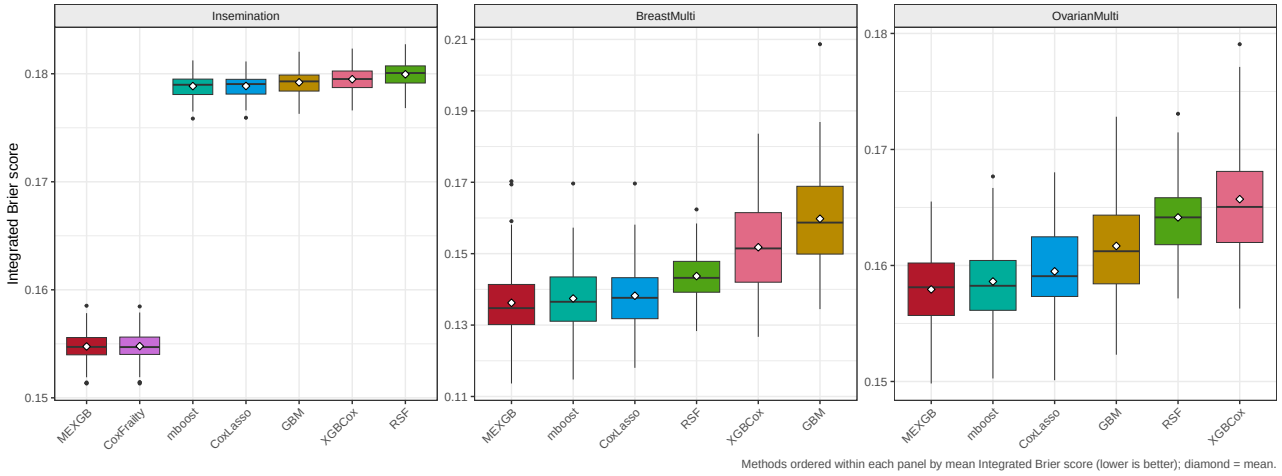

Figure S2: Real data: test integrated Brier score over the 100 within-cluster splits, with methods ordered by their mean within each panel (lower is better).

S3 Tutorial: fitting MEXGB to clustered survival data

To fit MEXGB to clustered survival data, supply a two-column matrix containing follow-up times and event indicators and set `family = "cox"`:

```
devtools::install_github("rid4stat/MEXGB")
install.packages("survival")
library(MEXGB); library(survival)
## simulated clustered failure times: 50 clusters x 20 units, p = 170
set.seed(1)
cal <- calibrateSim("cox", target = 0.75, linear = FALSE, rel_p = 6,
                    time_points = 20, random_sd_slope = sqrt(0.5))
d <- simLong(n = 50, p = 170, rel_p = 6, time_points = 20, linear = FALSE,
             family = "cox", random_sd_slope = sqrt(0.5), censoring = 0.3,
             calibration = cal)
X <- as.matrix(d[, grep("^Var", names(d))])
Z <- as.matrix(d[, c("RandomIntercept", "RandomSlope")])
te <- unlist(tapply(seq_len(nrow(d)), d$id,
                    function(i) sample(i, round(0.3 * length(i))))) # 30% of every cluster
tr <- setdiff(seq_len(nrow(d)), te)
fit <- MEXGB(X[tr, ], cbind(d$Y[tr], d$status[tr]), id = d$id[tr], Z = Z[tr, ],
           time = d$time[tr], family = "cox", interaction.depth = 2)
eta <- predict(fit, X[te, ], Z[te, ], id = d$id[te], time = d$time[te], type = "link")
concordance(Surv(d$Y[te], d$status[te]) ~ eta, reverse = TRUE)$concordance
fit$var_random_effects # estimated frailty covariance D
summary(fit, top = 10) # predictor importance (tree gain + |lasso coefficient|)
```

Survival curves for held-out units are obtained from equation (11) of the main text, $\hat{S}(t) = \exp\{-\hat{\Lambda}_0(t)e^{\hat{\eta}}\}$, with the Breslow estimator stored in the fitted object (`fit$baseline`). The example was run using `MEXGB` version 1.0.0. For an individual from a new cluster, `predict` sets the log-frailty contribution to zero.

References

[1] John Whitehead. Fitting Cox's regression model to survival data using GLIM. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):268–275, 1980.

[2] Samuli Ripatti and Juni Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022, 2000.

[3] Per K. Andersen and Richard D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.