Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Saheed Kunle Ajibade
Kwara State University
Oyebayo Ridwan Olaniran
ridwan.olaniran@kcl.ac.uk
King's College London
Samuel Adewale Aderoju
Kwara State University
Ayodeji Idowu
Babcock University
Nada MohammedSaeed Alharbi
Taibah University
Abdulhameed Bolakale
University of Ilorin

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11789696/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Saheed Kunle Ajibade[1,+], Oyebayo Ridwan Olaniran[1,2,3,*,+], Samuel Adewale Aderoju[1,+], Ayodeji Idowu[4], Nada MohammedSaeed Alharbi[5], and Abdulhameed Bolakale[2]

[1]Department of Statistics and Mathematical Sciences, Faculty of Pure and Applied Sciences, Kwara State University, PMB 1530, Malete, Kwara State, Nigeria.
[2]Department of Statistics, Faculty of Physical Sciences, University of Ilorin, PMB 1515, Ilorin, Kwara State, Nigeria.
[3]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
[4]Department of Information Resources Management, Babcock University, Ilishan-Remo, Ogun State, Nigeria
[5]Department of Mathematics, Taibah University, Faculty of Science, Al-Madinah, 42353, Al-Munawara, Saudi Arabia
[*]ridwan.olaniran@kcl.ac.uk
[+]These authors contributed equally to this work

ABSTRACT

Survival data arising from multiple centres or studies frequently exhibit cluster-level heterogeneity that is not accounted for by observed covariates. High-dimensional prognostic modelling therefore requires methods that can jointly accommodate such unobserved heterogeneity, nonlinear covariate effects, and principled variable selection. We propose mixed-effects extreme gradient boosting (MEXGB), a hybrid approach that integrates a relaxed-lasso Cox component with Newton-boosted decision trees and log-normal cluster-level frailties. Frailty parameters are estimated through a Poisson representation of the Cox model using cross-fitted risk predictions, and a calibration slope is subsequently applied to rescale the fixed-effect log-hazard contribution. Under explicitly stated assumptions, we derive theoretical guarantees that characterise the loss in predictive accuracy attributable to either ignoring clustering or enforcing a misspecified linear log-hazard structure. In simulation studies involving up to 2,000 predictors, MEXGB demonstrated improved predictive accuracy compared with the strongest competing method, particularly in settings where alternative approaches either neglected clustering structure or imposed an incorrect functional form. Performance was comparable to that of a correctly specified frailty model, and MEXGB attained the lowest integrated Brier score under a common procedure for constructing individual survival curves. In applications to a herd study and to multi-study breast- and ovarian-cancer cohorts, MEXGB achieved the highest mean predictive accuracy among the evaluated methods.

Introduction

Survival outcomes are frequently observed under a clustered sampling structure. Examples include patients enrolled in multi-centre clinical trials and treated across different hospitals, gene-expression cohorts aggregated from multiple studies, and livestock managed within distinct herds. Units within a given cluster typically share characteristics that are incompletely observed or unmeasured, such as surgical practice, clinical characteristics, patient selection mechanisms, assay platforms, or herd-management protocols. These shared factors can systematically shift the hazards of units within a cluster and induce dependence between their failure times. Estimation approaches that ignore clustering may therefore underestimate or otherwise misspecify inferential uncertainty and, in predictive settings, fail to use relevant information: outcomes among earlier patients treated at a given centre can be informative for subsequent patients within the same centre. The shared frailty model[1–4] represents this dependence by introducing a cluster-specific random effect that multiplicatively scales the hazards of all units in the cluster. Under a log-normal frailty distribution, estimation is commonly performed via penalised partial likelihood[5–7].

Modern applications often require frailty models to handle high-dimensional predictors and nonlinear effects. For example, one may wish to predict metastasis-free survival for a patient treated at one of several participating centres using several thousand gene-expression measurements. For the prediction task considered here, a useful model should provide four capabilities: (R1) accommodate centre-level heterogeneity via a cluster effect that is appropriately shrunk towards the population mean when the centre-specific sample size is small; (R2) permit nonlinear covariate effects and interactions, since thresholding, saturation, and pathway-level interactions are common in biological systems and a misspecified linear log-hazard can mis-rank patients even when frailties are known; (R3) remain estimable and enable variable selection in high-dimensional settings where the number of predictors exceeds the number of events; and (R4) yield centre-calibrated survival probabilities, rather than only a relative

risk ranking. Table 1 summarises the extent to which existing methods satisfy these requirements.

Table 1. Existing methods for clustered or correlated outcomes and the requirements of multi-centre survival prediction with many predictors: (R1) a shrunken cluster effect; (R2) nonlinear effects; (R3) estimation and variable selection in high-dimensional settings; (R4) survival probabilities for right-censored data. ✓: met; ○: partly met; –: not met; n.r.: not reported.

Method	Model	R1	R2	R3	R4	Main limitation for the present task
Frailty models with a linear predictor						
`coxme`[5,6]	log-normal frailty Cox	✓	–	–	✓	Linear log-hazard; not estimable when p approaches the number of events.
`PenCoxFrail`[8]	penalised frailty Cox	✓	–	○	✓	Selects predictors, but the log-hazard remains linear.
Flexible survival learners that ignore the cluster						
Cox boosting[9–11], XGBoost-Cox[12]	boosted Cox model	–	✓	✓	✓	No cluster effect: concordance and calibration limited when centres differ.
Random survival forests[13,14]	ensemble of survival trees	–	✓	✓	✓	As above; including cluster labels requires an appropriate encoding and fitting strategy.
DeepSurv[15], neural frailty machine[16]	neural Cox model	–	✓	○	✓	As above; the frailty of the latter is individual, not shared by a cluster.
Mixed-effects learners for uncensored outcomes						
MERF[17], LongituRF[18], GPBoost[19], MEGB[20], GMEXGBoost[21]	trees + random effects	✓	✓	✓	–	Gaussian or binary likelihood; no censoring and no baseline hazard.
Flexible survival models with a frailty						
SBART frailty[22]	soft BART, gamma frailty	✓	✓	–	✓	MCMC; interval-censored data; moderate number of covariates.
riAFT-BART[23]	BART, AFT, random intercept	✓	✓	–	✓	MCMC; accelerated failure time model; moderate number of covariates.
Bayesian survival tree[24]	single tree, frailty	✓	○	–	✓	A single tree approximates smooth effects coarsely; evaluated with few covariates.
RecForest[25]	forest, recurrent events	○	✓	✓	○	Cluster is the subject (recurrent events); no shared frailty across subjects.
Neural frailty Cox[26]	neural network, frailty	✓	✓	n.r.	✓	Needs large samples; no built-in variable selection.
MEXGB (this paper)	lasso + boosted trees, log-normal frailty	✓	✓	✓	✓	Assumes proportional hazards given the frailty; predictions for new clusters use the fixed effect only.

The methods in Table 1 address different combinations of these requirements. *Frailty models with a linear predictor* meet R1 and R4. The penalised frailty model of Groll et al.[8] also performs variable selection, but every model in this group fixes the form of the log-hazard, and we show in the theoretical section that under some certain regularity conditions, adjusting the coefficients cannot eliminate the loss of accuracy caused by a misspecified linear log-hazard.

Flexible survival learners such as Cox models boosted[9–12], random survival forests[13,14], and neural Cox models[15,16] meet R2 and R3 and have been extensively compared for high-dimensional survival prediction[14,27], but they treat units as independent. Under some assumptions, a score that ignores the cluster loses accuracy relative to the conditional log-hazard, regardless of its flexibility. Including cluster indicators is an alternative, although unpenalised cluster coefficients can be unstable when clusters contain few events. The behaviour of a cluster label in a tree model depends on how it is encoded and fitted.

Mixed effects tree learners[17–21,28] meet R1–R3 and have substantially improved prediction for clustered continuous and binary outcomes. MEGB, for example, reduced the mean squared error of mixed-effects random forests by 35–76%[20] but their expectation steps rely on a Gaussian or Bernoulli likelihood of fully observed responses and do not accommodate censoring or an unspecified baseline hazard.

Flexible frailty models such as Basak et al.[22] and Hu et al.[23] placed Bayesian additive regression trees in a frailty and an accelerated failure time model, Porndumnernsawat et al.[24] developed Bayesian multivariate survival trees under three frailty models, Murris et al.[25] adapted random survival forests to recurrent events, and Zhou et al.[26] replaced the linear risk function of a frailty Cox model by a neural network. These methods rely on Markov chain Monte Carlo, single trees or large samples; they have been developed and evaluated with tens of covariates (for example, 50 in the simulations of Porndumnernsawat et al.[24]) or for a different form of clustering; and their reported evaluations do not establish variable-selection performance with high-dimensional predictors.

A further difficulty arises whenever a flexible learner and random effects are estimated together. A flexible learner may capture part of the cluster-level variation through the predictors of its units. If the random effects are then estimated from the learner's in-sample residuals, the residuals may underrepresent cluster-level variation, leading to an underestimated frailty

variance. A boosted Cox model may also produce overly dispersed log-hazard scores, preserving their ranking while distorting the corresponding survival probabilities (R4).

Mixed-effects extreme gradient boosting (MEXGB) is designed to combine a shared frailty, a flexible and sparse log-hazard, scalability to high-dimensional predictors and cross-fitting to reduce overfitting in frailty estimation for right-censored data. Its log-hazard is the sum of a relaxed lasso Cox component, which captures sparse linear signal, and an ensemble of trees fitted by Newton steps of the Cox partial likelihood, which captures nonlinear effects and interactions; the cluster frailty enters as an offset. Frailty estimation uses cross-fitted fixed-effect predictions: each unit is predicted by a model fitted on its fold complement. A Poisson representation of the Cox model then allows mixed-model estimation to be applied to censored data. Cross-fitting is intended to reduce the absorption of cluster-level variation by the fixed-effect learner. The contributions of the paper are:

1. **Method.** A frailty extension of extreme gradient boosting for right-censored data, with a relaxed lasso Cox component, cross-fitted Poisson penalised quasi-likelihood (PQL) estimation of the frailties, and calibration of the boosted log-hazard. The relation to classical frailty estimation, including the additional restrictions required, is described in step 3. With the frailty variance fixed at zero, the method becomes a boosted Cox model with a relaxed lasso component.

2. **Theory.** An exact expression for the loss of accuracy of any risk score relative to the conditional log-hazard, at every follow-up horizon, from which the losses of cluster-blind scores (R1) and of linear frailty models (R2) follow; and a Brier-score decomposition that explains the role of calibration (R4).

3. **Evidence.** A simulation study with up to 2000 predictors, linear and nonlinear log-hazards and two censoring rates, and applications to a herd study and to multi-study breast- and ovarian-cancer gene-expression cohorts, assessed by concordance, integrated Brier scores, risk-group and cluster-specific survival curves and calibration.

4. **Software.** An implementation in the R package `MEXGB`.

Methods

Overview

MEXGB treats the cluster frailties as missing data and alternates between two estimation problems: estimating the fixed-effect log-hazard when the frailties are known, and estimating the frailties when the fixed effect is known (Fig. 1). The first problem is a boosted Cox regression with a known offset, which `xgboost`[12] and `glmnet`[29] solve directly. The second is a mixed-model problem for censored data, which becomes a standard generalised linear mixed model through the Poisson representation of the Cox model. Two procedures link the problems: cross-fitting, which reduces in-sample overfitting in the fixed-effect predictions used for frailty estimation, and a calibration slope, which corrects the scale of the boosted log-hazard.

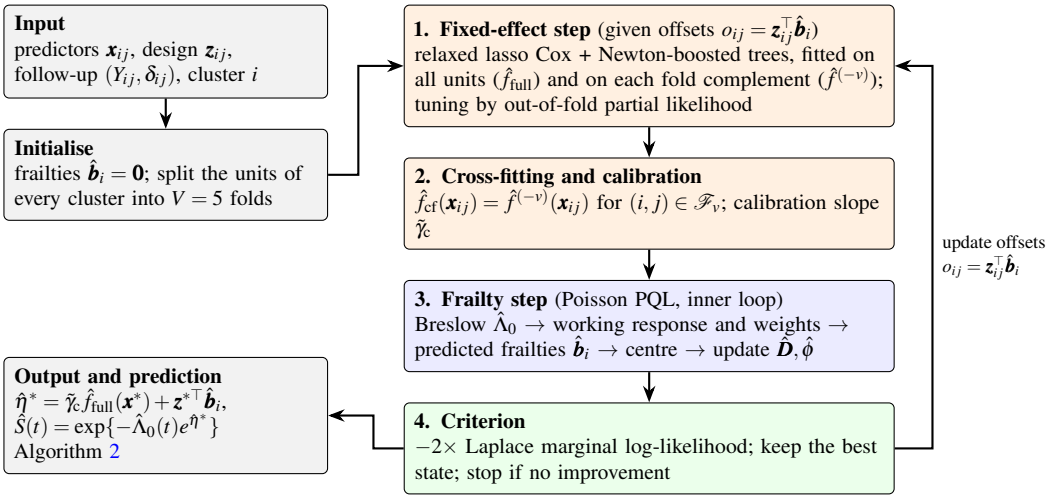

Figure 1. Structure of the MEXGB algorithm. Steps 1–4 are repeated (outer iterations) until the criterion stops improving; step 3 contains its own inner iterations. Orange: fixed-effect estimation; blue: frailty estimation; green: model selection.

Model and notation

Let cluster $i = 1, \ldots, n$ contain n_i units, $N = \sum_i n_i$, and let unit (i, j) have predictors $\boldsymbol{x}_{ij} \in \mathbb{R}^p$, a random-effect design $\boldsymbol{z}_{ij} \in \mathbb{R}^q$ and a failure time T_{ij}. MEXGB specifies the conditional hazard

$$\lambda_{ij}(t \mid \boldsymbol{x}_{ij}, \boldsymbol{b}_i) = \lambda_0(t) \exp(\eta_{ij}), \qquad \eta_{ij} = f(\boldsymbol{x}_{ij}) + \boldsymbol{z}_{ij}^\top \boldsymbol{b}_i, \qquad \boldsymbol{b}_i \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{D}), \tag{1}$$

with an unspecified baseline hazard λ_0 and a fixed effect

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta} + T(\boldsymbol{x}), \qquad T(\boldsymbol{x}) = \sum_{m=1}^{M} \nu\, t_m(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{\beta}$ is a sparse coefficient vector, t_m are regression trees of depth at most d, and $\nu \in (0, 1]$ is the learning rate. The linear component efficiently represents sparse linear signals, while the trees capture nonlinear effects and interactions. Either component can be absent, and their complexity is selected from the data. With $\boldsymbol{z}_{ij} = 1$, model (1) is the shared log-normal frailty model with log-frailty b_i; a second column of $\boldsymbol{z}_{ij}$, such as the position of a unit within its cluster, allows a random slope and hence within-cluster variation in the frailty contribution. The failure times are conditionally independent given the frailties and are observed as $(Y_{ij}, \delta_{ij}) = (\min\{T_{ij}, C_{ij}\}, \mathbb{1}\{T_{ij} \le C_{ij}\})$, with censoring independent of T_{ij} given $(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)$ and non-informative. The conditional survival function is $S_{ij}(t) = \exp\{-\Lambda_0(t) e^{\eta_{ij}}\}$, where the cumulative hazard function is $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$.

Let $\tau_1 < \cdots < \tau_K$ be the distinct event times, d_k the number of events at τ_k, $\mathscr{R}_k$ the risk set and $S_k(\boldsymbol{\eta}) = \sum_{(i,j) \in \mathscr{R}_k} e^{\eta_{ij}}$. With Breslow's treatment of ties, the negative log partial likelihood[30] is

$$\mathscr{L}_{\mathrm{PL}}(\boldsymbol{\eta}) = -\sum_{i,j} \delta_{ij} \eta_{ij} + \sum_{k=1}^{K} d_k \log S_k(\boldsymbol{\eta}). \tag{3}$$

It does not involve λ_0, so the fixed-effect step does not require a specified baseline hazard. It is unchanged when the same constant is added to every η_{ij}, so the mean of the random intercepts is not identified; MEXGB fixes it at zero.

Step 1: fixed-effect step

Trees. The current frailties enter as known offsets $o_{ij} = \boldsymbol{z}_{ij}^\top \hat{\boldsymbol{b}}_i$, which `xgboost` accepts as base margins; the trees are fitted conditional on these estimated cluster effects. The Cox loss (3) is not quadratic, so its local curvature is used to construct an approximate update. `xgboost` instead takes a Newton step, using the gradient and the diagonal of the Hessian of (3):

$$g_{ij} = \hat{\mu}_{ij} - \delta_{ij}, \qquad h_{ij} = \hat{\mu}_{ij} - \sum_{k : \tau_k \le Y_{ij}} d_k \frac{e^{2\hat{\eta}_{ij}}}{S_k^2} \in (0, \hat{\mu}_{ij}], \tag{4}$$

where $\hat{\mu}_{ij} = \hat{\Lambda}_0(Y_{ij}) e^{\hat{\eta}_{ij}}$ is the fitted cumulative hazard of unit (i, j) and $\hat{\Lambda}_0(t) = \sum_{k : \tau_k \le t} d_k / S_k$ the Breslow estimator[31] at the current log-hazard. The optimal weight of a leaf containing the units I_ℓ is then available in closed form,

$$w_\ell^\star = \frac{\sum_{(i,j) \in I_\ell} (\delta_{ij} - \hat{\mu}_{ij})}{\sum_{(i,j) \in I_\ell} h_{ij} + \lambda}, \tag{5}$$

with λ the L_2 penalty on leaf weights. The numerator is the sum of martingale residuals observed minus expected events in the leaf. Because $\hat{\mu}_{ij}$ includes the frailty, a unit of a high-risk centre is expected to fail early, and the tree update responds to discrepancies between observed events and the fitted cumulative hazards, conditional on the estimated cluster effects.

Linear component. Before fitting the trees, MEXGB fits the lasso Cox model with the same offset, $\min_{\boldsymbol{\beta}} \{N^{-1} \mathscr{L}_{\mathrm{PL}}(\boldsymbol{o} + \boldsymbol{X}\boldsymbol{\beta}) + \lambda_{\mathrm{L}} \|\boldsymbol{\beta}\|_1\}$, by coordinate descent[32], and relaxes it[33]: the penalised coefficients are blended with an unpenalised Cox refit on the selected predictors, with weight $\gamma_{\mathrm{R}} \in \{0, \frac{1}{2}, 1\}$, which reduces the shrinkage bias of the lasso on strong predictors. The trees are then grown with $\boldsymbol{o} + \boldsymbol{X}\hat{\boldsymbol{\beta}}$ as base margin, so they model variation not captured by the linear component and estimated frailties.

Tuning. The lasso penalty λ_{L}, the blending weight γ_{R} and the number of trees M are chosen by the out-of-fold partial likelihood on the cross-fitting folds described below; for each validation fold the partial likelihood is computed over that fold's own risk sets. The number of trees is evaluated up to M_{max} and selected at the minimum out-of-fold loss, allowing the ensemble size to adapt to the signal. When p is large, the trees can be restricted to the $K \in \mathscr{K}$ predictors with the largest importance (tree gain plus absolute standardised lasso coefficient) in the first two outer iterations; K, or no restriction, is again chosen out of fold. These tuning choices are made within each model fit.

Step 2: cross-fitting and calibration

Estimating frailties from in-sample fixed-effect residuals can underestimate cluster-level variation if the trees have already absorbed part of that variation. MEXGB therefore estimates the frailties from *out-of-fold* predictions. The units of every cluster are split at random into V folds $\mathscr{F}_1, \ldots, \mathscr{F}_V$, so that each cluster appears in every fold. Step 1 is carried out on all training units, giving $\hat{f}_{\text{full}}$, and on each fold complement, giving $\hat{f}^{(-v)}$, and the cross-fitted fixed effect is $\hat{f}_{\text{cf}}(\boldsymbol{x}_{ij}) = \hat{f}^{(-v)}(\boldsymbol{x}_{ij})$ for $(i,j) \in \mathscr{F}_v$. The fixed-effect model used for each unit's cross-fitted prediction is fitted on that unit's fold complement. Because the partial likelihood identifies a log-hazard only up to a constant, each fold model is shifted so that its mean on its training units equals that of $\hat{f}_{\text{full}}$. The frailty step uses $\hat{f}_{\text{cf}}$; prediction uses $\hat{f}_{\text{full}}$.

Boosting the partial likelihood rewards extreme scores for early failures, so a boosted Cox model can rank units correctly while its log-hazard is too spread out; its survival curves are then too steep for high-risk and too flat for low-risk units. MEXGB multiplies the cross-fitted log-hazard by its calibration slope[34], the maximiser $\hat{\gamma}_{\text{c}}$ of the partial likelihood of the one-parameter model $\eta_{ij} = o_{ij} + \gamma \hat{f}_{\text{cf}}(\boldsymbol{x}_{ij})$. This likelihood is concave in γ and strictly concave under the non-degeneracy condition. Under a pure scale error and the stated regularity conditions, the unconstrained slope consistently estimates the scale correction.

Step 3: frailty step through the Poisson representation

The frailties cannot be estimated from the partial likelihood (3) with standard mixed-model procedures for two reasons. The partial likelihood is not a product over clusters, every risk set contains units of many clusters and most mixed-model procedures need a likelihood that factorises over clusters given the random effects; and the frailty variance $\boldsymbol{D}$ is a parameter of the distribution of $\boldsymbol{b}_i$, which requires integrating the likelihood over $\boldsymbol{b}_i$. The Poisson representation provides a practical route to estimation[35]. If the baseline hazard is a step function with a jump at each event time, the full likelihood of model (1) is, up to a factor that does not involve $\boldsymbol{\eta}$, the likelihood of independent Poisson variables δ_{ij} with means

$$\mu_{ij} = \Lambda_0(Y_{ij})\, e^{\eta_{ij}}, \qquad \text{that is,} \qquad \log \mu_{ij} = \log \Lambda_0(Y_{ij}) + f(\boldsymbol{x}_{ij}) + \boldsymbol{z}_{ij}^{\top} \boldsymbol{b}_i, \tag{6}$$

a Poisson generalised linear mixed model with log link and offset $\log \Lambda_0(Y_{ij})$. Maximising over the jumps of Λ_0 gives the Breslow estimator, and profiling the full survival likelihood over these jumps recovers the Cox partial-likelihood objective. The representation is an identity between likelihoods; it does not assume that event counts are Poisson distributed. Given $\hat{\Lambda}_0$, the Poisson log-likelihood is a sum over clusters, and the frailties can be estimated by penalised quasi-likelihood (PQL)[36], the method that underlies the penalised partial likelihood frailty model of Ripatti and Palmgren[5].

Each inner iteration of step 3 performs the following operations, with the fixed effect held at $\tilde{\gamma}_{\text{c}} \hat{f}_{\text{cf}}$.

1. Compute the Breslow estimator $\hat{\Lambda}_0$ at $\hat{\eta}_{ij} = \tilde{\gamma}_{\text{c}} \hat{f}_{\text{cf}}(\boldsymbol{x}_{ij}) + \boldsymbol{z}_{ij}^{\top} \hat{\boldsymbol{b}}_i$ and the fitted means $\hat{\mu}_{ij} = \hat{\Lambda}_0(Y_{ij}) e^{\hat{\eta}_{ij}}$.

2. Linearise the Poisson log-likelihood at $\hat{\eta}$. For a log link, the PQL working response and weight are

$$\tilde{y}_{ij} = \hat{\eta}_{ij} + \frac{\delta_{ij} - \hat{\mu}_{ij}}{\hat{\mu}_{ij}}, \qquad w_{ij} = \hat{\mu}_{ij}, \tag{7}$$

   and, to first order, $\tilde{y}_{ij} = \eta_{ij} + \xi_{ij}$ with $\mathbb{E}\xi_{ij} = 0$ and $\operatorname{Var} \xi_{ij} = 1/w_{ij}$. This yields the working linear mixed model for cluster i, $\tilde{\boldsymbol{y}}_i - \tilde{\gamma}_{\text{c}} \hat{f}_{\text{cf},i} = \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{r}_i$ with $\operatorname{Var}(\boldsymbol{r}_i) = \phi \boldsymbol{W}_i^{-1}$, $\boldsymbol{W}_i = \operatorname{diag}(w_{ij})$.

3. Predict the frailties by their best linear unbiased predictors in this working model,

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{D}} \boldsymbol{Z}_i^{\top} \left(\boldsymbol{Z}_i \hat{\boldsymbol{D}} \boldsymbol{Z}_i^{\top} + \hat{\phi} \boldsymbol{W}_i^{-1} \right)^{-1} \left(\tilde{\boldsymbol{y}}_i - \tilde{\gamma}_{\text{c}} \hat{f}_{\text{cf},i} \right), \qquad \boldsymbol{C}_i = \left(\hat{\boldsymbol{D}}^{-1} + \hat{\phi}^{-1} \boldsymbol{Z}_i^{\top} \boldsymbol{W}_i \boldsymbol{Z}_i \right)^{-1}, \tag{8}$$

   where $\boldsymbol{C}_i$ is the conditional covariance of $\boldsymbol{b}_i$. A cluster with few events has small weights w_{ij}, and its frailty is shrunk strongly towards zero.

4. Centre the random intercepts, $\hat{b}_{i0} \leftarrow \hat{b}_{i0} - \bar{b}_0$. By the invariance of (3), this changes no fitted hazard (the Breslow estimator absorbs the shift) but stops the intercept mean from drifting and inflating $\hat{D}_{00}$.

5. Update the covariance and the dispersion,

$$\hat{\boldsymbol{D}} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^{\top} + \boldsymbol{C}_i \right), \qquad \hat{\phi} = \frac{1}{N} \sum_{i=1}^{n} \left\{ \left\| \boldsymbol{W}_i^{1/2} (\tilde{\boldsymbol{y}}_i - \tilde{\gamma}_{\text{c}} \hat{f}_{\text{cf},i} - \boldsymbol{Z}_i \hat{\boldsymbol{b}}_i) \right\|^2 + \operatorname{tr}\left(\boldsymbol{W}_i \boldsymbol{Z}_i \boldsymbol{C}_i \boldsymbol{Z}_i^{\top} \right) \right\}. \tag{9}$$

The inner iterations stop when the relative change in $(\hat{\boldsymbol{D}}, \hat{\phi})$ is below tolerance δ, or after t iterations. The update of $\hat{\boldsymbol{D}}$ is the expectation–maximisation update of the working model: the term $\boldsymbol{C}_i$ accounts for the uncertainty of $\hat{\boldsymbol{b}}_i$, without which $\hat{\boldsymbol{D}}$ would

be biased downwards. The single global dispersion ϕ is an additional modification of the classical working model. In the classical model $\phi = 1$, but here the working residual also contains the out-of-fold error of the boosted fixed effect; estimating ϕ provides a way to accommodate residual variation beyond the unit-dispersion working model.

Relation to existing estimators. With the trees removed, no penalty on $\boldsymbol{\beta}$, no cross-fitting and $\phi = 1$, step 3 is the penalised partial likelihood algorithm of Ripatti and Palmgren[5] for the log-normal frailty Cox model. The `coxme` package[6] uses the same penalised partial likelihood for the frailties but estimates $\boldsymbol{D}$ by maximising a Laplace approximation of the integrated partial likelihood, so the two estimators coincide in their frailty predictions given $\boldsymbol{D}$ and differ slightly in the estimate of $\boldsymbol{D}$. With $\boldsymbol{D}$ fixed at zero, MEXGB reduces to a boosted Cox model with a relaxed lasso component.

Step 4: model selection and stopping

Because the frailty step uses a cross-fitted fixed effect, the outer iterations do not maximise one objective, and their progress must be monitored by an external criterion. MEXGB evaluates, after each outer iteration, the Laplace approximation[37] of the marginal Poisson log-likelihood,

$$\ell_{\mathrm{La}} = \sum_{i=1}^{n} \Big\{ l_i(\hat{\boldsymbol{b}}_i) - \tfrac{1}{2}\hat{\boldsymbol{b}}_i^{\top}\hat{\boldsymbol{D}}^{-1}\hat{\boldsymbol{b}}_i - \tfrac{1}{2}\log\big|\boldsymbol{I}_q + \hat{\boldsymbol{D}}\boldsymbol{Z}_i^{\top}\boldsymbol{W}_i\boldsymbol{Z}_i\big| \Big\}, \qquad l_i(\boldsymbol{b}_i) = \sum_{j=1}^{n_i} \big\{ \delta_{ij}\log\mu_{ij} - \mu_{ij} \big\}, \tag{10}$$

where l_i is the conditional Poisson log-likelihood (6) of cluster i at the Breslow estimator and the cross-fitted log-hazard, so that the criterion uses cross-fitted fixed-effect predictions. The first term rewards fit, the second penalises large frailties and the third penalises frailty variance that the data do not support. MEXGB keeps the state with the smallest $-2\ell_{\mathrm{La}}$ and stops when its relative change is below δ, when it has not improved for V outer iterations, or after t iterations. The frailties start at zero, so the first outer iteration fits an ordinary boosted Cox model. Algorithm 1 summarises the fitting procedure.

Algorithm 1 MEXGB: fitting

Require: Training units $(\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, Y_{ij}, \delta_{ij})$, $i = 1,\dots,n$; learning rate ν; depth d; minimum Hessian weight per leaf w; leaf penalty λ; at most M_{max} trees; V folds; at most t_{inner} inner and t_{outer} outer iterations; tolerance δ.

Ensure: Linear coefficients $\hat{\boldsymbol{\beta}}$, trees, slope $\tilde{\gamma}_{\mathrm{c}}$, frailties $\hat{\boldsymbol{b}}_i$, $\hat{\boldsymbol{D}}$, $\hat{\phi}$, Breslow estimator $\hat{\Lambda}_0$.

$\hat{\boldsymbol{b}}_i \leftarrow \boldsymbol{0}$, $\hat{\phi} \leftarrow 1$; split the units of every cluster at random into folds $\mathscr{F}_1,\dots,\mathscr{F}_V$.
for $t = 1, 2, \dots$ **do**
  Offsets $o_{ij} \leftarrow \boldsymbol{z}_{ij}^{\top}\hat{\boldsymbol{b}}_i$.
  Step 1. On all units and on each fold complement $\mathscr{F}_v^{c}$: fit the relaxed lasso Cox model with offset $\boldsymbol{o}$, then Newton-boosted trees (5) with base margin $\boldsymbol{o} + \boldsymbol{X}\hat{\boldsymbol{\beta}}$; choose λ_{L}, γ_{R}, M (and, for $t \le 2$, the screened predictors) by the out-of-fold partial likelihood.
  Step 2. Align each fold model with $\hat{f}_{\mathrm{full}}$; assemble $\hat{f}_{\mathrm{cf}}$; compute $\tilde{\gamma}_{\mathrm{c}}$.
  repeat
    Step 3. Breslow $\hat{\Lambda}_0$ at $\tilde{\gamma}_{\mathrm{c}}\hat{f}_{\mathrm{cf}} + \boldsymbol{o}$; working response and weights (7); $\hat{\boldsymbol{b}}_i$ by (8); centre the random intercepts; $(\hat{\boldsymbol{D}}, \hat{\phi})$ by (9); $o_{ij} \leftarrow \boldsymbol{z}_{ij}^{\top}\hat{\boldsymbol{b}}_i$.
  until relative change of $(\hat{\boldsymbol{D}}, \hat{\phi}) < \delta$ or $t = t_{inner}$ iterations
  Step 4. Evaluate $-2\ell_{\mathrm{La}}$ (10); store the state if it is the best so far.
  stop if the relative change of $-2\ell_{\mathrm{La}}$ is below δ, there has been no improvement for V iterations, or $t = t_{outer}$.
end for
Restore the best state; compute $\hat{\Lambda}_0$ by the Breslow estimator at $\tilde{\gamma}_{\mathrm{c}}\hat{f}_{\mathrm{full}}(\boldsymbol{x}_{ij}) + \boldsymbol{z}_{ij}^{\top}\hat{\boldsymbol{b}}_i$ on the training units.

Prediction

For a new unit with predictors $\boldsymbol{x}^*$ and design $\boldsymbol{z}^*$, MEXGB returns a risk score and a survival curve (Algorithm 2). If the unit belongs to a cluster i seen in training, the risk score adds the cluster's predicted frailty to the calibrated fixed effect,

$$\hat{\eta}^* = \tilde{\gamma}_{\mathrm{c}}\hat{f}_{\mathrm{full}}(\boldsymbol{x}^*) + \boldsymbol{z}^{*\top}\hat{\boldsymbol{b}}_i, \qquad \hat{S}(t \mid \boldsymbol{x}^*, \boldsymbol{z}^*, i) = \exp\big\{-\hat{\Lambda}_0(t)\, e^{\hat{\eta}^*}\big\}. \tag{11}$$

If the cluster is new, its log-frailty contribution is unknown and is set to zero, its model-based mean. This gives a conditional survival curve at the median multiplicative frailty. The population-averaged curve, which integrates over the frailty distribution, can be obtained by averaging $\hat{S}$ over draws $\boldsymbol{b} \sim \mathscr{N}_q(\boldsymbol{0}, \hat{\boldsymbol{D}})$.

Algorithm 2 MEXGB: prediction

Require: Fitted model from Algorithm 1; new units $(\boldsymbol{x}_r^*, \boldsymbol{z}_r^*, c_r)$, $r = 1, \ldots, R$, with cluster labels c_r; time grid $t_1 < \cdots < t_G$.
Ensure: Risk scores $\hat{\eta}_r^*$ and survival probabilities $\hat{S}_r(t_g)$.

for $r = 1, \ldots, R$ **do**
  Fixed effect: $\hat{f}_r \leftarrow \boldsymbol{x}_r^{*\top}\hat{\boldsymbol{\beta}} + \sum_{m=1}^{\hat{M}} \nu\, t_m(\boldsymbol{x}_r^*)$ from $\hat{f}_{\text{full}}$; calibrated: $\hat{f}_r \leftarrow \tilde{\gamma}_c \hat{f}_r$.
  if cluster c_r was seen in training **then**
    $\hat{\eta}_r^* \leftarrow \hat{f}_r + \boldsymbol{z}_r^{*\top}\hat{\boldsymbol{b}}_{c_r}$ (conditional prediction)
  else
    $\hat{\eta}_r^* \leftarrow \hat{f}_r$ (median-frailty prediction)
  end if
  $\hat{S}_r(t_g) \leftarrow \exp\{-\hat{\Lambda}_0(t_g)\exp(\hat{\eta}_r^*)\}$ for $g = 1, \ldots, G$, where $\hat{\Lambda}_0$ is the step function of Algorithm 1.
end for
Optional, for a new cluster: population-averaged curve $\bar{S}_r(t_g) = B^{-1}\sum_{b=1}^{B} \exp\{-\hat{\Lambda}_0(t_g)\exp(\hat{f}_r + \boldsymbol{z}_r^{*\top}\boldsymbol{b}^{(b)})\}$ with $\boldsymbol{b}^{(b)} \sim \mathcal{N}_q(\boldsymbol{0}, \hat{\boldsymbol{D}})$.

Computational cost

One outer iteration fits $V+1$ lasso paths and tree ensembles, so its cost is about $V+1$ times that of a single boosted Cox model with the same number of trees; the $V+1$ fits are independent and can be run in parallel. The frailty step costs $O(N)$ per inner iteration for $q \le 2$, because applying the Woodbury identity reduces the required inversions to $q \times q$ matrices, and is negligible relative to boosting. In the simulations the criterion stopped improving after two to four outer iterations (Fig. 3).

Theoretical properties

The following results explain why both the frailty and a flexible fixed effect are needed for discrimination, and why calibration is needed for survival probabilities; proofs are given in Supplementary Section S1. Let (k, l) be two units drawn independently from the units of clusters seen in training, with conditional hazards $e^{\eta_k}\lambda_0$ and $e^{\eta_l}\lambda_0$, and let s be a risk score (larger values meaning higher risk). For a horizon $\tau \in (0, \infty]$ the truncated concordance probability is

$$C_\tau(s) = \mathbb{P}\big(s_k > s_l \mid T_k < T_l,\ T_k < \tau\big),$$

with tied scores receiving half credit. It is estimated by Harrell's C-index[38] when $\tau = \infty$ and there is no censoring, and by the inverse-probability-weighted index of Uno et al.[39] under independent censoring.

Lemma 1 *If T_k and T_l are conditionally independent with hazards $h_k\lambda_0$ and $h_l\lambda_0$, then $q_{kl} := \mathbb{P}(T_k < T_l, T_k < \tau \mid h_k, h_l) = \dfrac{h_k}{h_k + h_l}\big\{1 - e^{-(h_k+h_l)\Lambda_0(\tau)}\big\}$, and $q_{kl} - q_{lk}$ has the sign of $h_k - h_l$.*

Theorem 1 (Concordance optimality of the conditional log-hazard) *Let s be any risk score that is a function of the information available for prediction, and suppose T_k and T_l are conditionally independent given (η_k, η_l) and that information. Write $\mathbb{1}^{\mathrm{d}}_{kl}(s) = \mathbb{1}\{(s_k - s_l)(\eta_k - \eta_l) < 0\}$ for a discordant pair and $\mathbb{1}^{\mathrm{t}}_{kl}(s) = \mathbb{1}\{s_k = s_l,\ \eta_k \ne \eta_l\}$ for a pair tied by s only. Then, at every horizon τ,*

$$C_\tau(\eta) - C_\tau(s) = \frac{\mathbb{E}\big[\,|q_{kl} - q_{lk}|\,\{\mathbb{1}^{\mathrm{d}}_{kl}(s) + \tfrac{1}{2}\mathbb{1}^{\mathrm{t}}_{kl}(s)\}\big]}{2\,\mathbb{E}\,q_{kl}} \ \ge\ 0. \tag{12}$$

For $\tau = \infty$, $q_{kl} - q_{lk} = (e^{\eta_k} - e^{\eta_l})/(e^{\eta_k} + e^{\eta_l})$ and $\mathbb{E}q_{kl} = \tfrac{1}{2}$.

The theorem expresses any loss of concordance as the probability-weighted share of pairs that the score orders differently from the conditional log-hazard. Two sources of misordering follow; they correspond to requirements R1 and R2 presented earlier.

Corollary 1 (Scores that ignore the cluster) *If $\boldsymbol{D} \ne \boldsymbol{0}$, $\boldsymbol{z}^\top\boldsymbol{b}$ has a positive density on $\mathbb{R}$, and $s = s(\boldsymbol{x})$ does not use the cluster label, then $C_\tau(s) < C_\tau(\eta)$ for every τ with $\Lambda_0(\tau) > 0$.*

Corollary 2 (Frailty models with a linear log-hazard) *Let $s = \boldsymbol{x}^\top\boldsymbol{\beta} + \boldsymbol{z}^\top\boldsymbol{b}$ be the risk score of a linear frailty model with the true frailties and any $\boldsymbol{\beta}$. If $f(\boldsymbol{x}) - \boldsymbol{x}^\top\boldsymbol{\beta}$ is not almost surely constant, the predictors of two units are independent and identically distributed, and the frailty difference has a positive density and is independent of the predictors, then $C_\tau(s) < C_\tau(\eta)$.*

MEXGB addresses the first loss with its frailty term and the second with its boosted fixed effect, whose boosting criterion targets f: given the true frailties, the expected Poisson log-likelihood (6) is strictly concave in η with maximiser $f(\boldsymbol{x})+\boldsymbol{z}^\top\boldsymbol{b}$. Concordance depends only on the ordering of the scores, so it cannot detect a wrong scale; the Brier score can.

Proposition 1 (Brier score) *Let $S(t)$ be the conditional survival function given the information used for prediction and $\hat{S}(t)$ any prediction that is a function of that information. Then $\mathbb{E}\{\mathbb{1}(T>t)-\hat{S}(t)\}^2=\mathbb{E}[S(t)\{1-S(t)\}]+\mathbb{E}\{S(t)-\hat{S}(t)\}^2$. If the frailties and Λ_0 are known and $\hat{\eta}=sf(\boldsymbol{x})+\boldsymbol{z}^\top\boldsymbol{b}$ with $s>0$, the excess Brier score $\mathbb{E}[\exp\{-\Lambda_0(t)e^{\eta}\}-\exp\{-\Lambda_0(t)e^{\hat{\eta}}\}]^2$ is zero if and only if $s=1$ (for f not almost surely zero), whereas without a frailty the concordance does not depend on s.*

The first term of the decomposition is irreducible; the second is the mean squared error in the predicted survival probabilities. The slope in step 2 targets one source of this error: an incorrect scale of the fixed-effect score. With frailties, a scale error also changes the relative weight of the fixed effect and the frailty in the risk score, and hence the ordering of units of different clusters, so calibration can improve concordance as well as the Brier score.

Simulation design

Data generation

The simulation design extends those of the published MEGB and LongituRF studies[18,20] with a larger sample. 50 clusters contain 20 units each ($N=1000$), with a unit-level position $t_{ij}\in\{0,1\}$. The $p\in\{6,170,2000\}$ predictors are drawn from $\mathcal{N}_p(\mathbf{0},\boldsymbol{\Sigma})$ with $\Sigma_{kk'}=0.6^{|k-k'|}$, and six of them are active through

$$f_{\text{lin}}(\boldsymbol{x})=2x_1-2x_2+1.5x_3-1.5x_4+x_5-x_6,$$
$$f_{\text{non}}(\boldsymbol{x})=2\sin(\pi x_1x_2)+4(x_3-0.5)^2+2x_4+e^{x_5}+x_6x_1.$$

The nonlinear function combines a sinusoidal interaction, a quadratic term, a linear term, an exponential term and a product, so that a linear fixed-effect specification is misspecified. The frailty design is $\boldsymbol{z}_{ij}=(1,t_{ij})^\top$ with $\boldsymbol{b}_i\sim\mathcal{N}_2(\mathbf{0},\boldsymbol{D})$, $D_{00}=D_{11}=0.5$ and correlation 0.6. Failure times follow the Weibull proportional hazards model

$$\lambda_{ij}(t)=\nu_{\text{W}}t^{\nu_{\text{W}}-1}\exp(\eta_{ij}),\qquad \eta_{ij}=\kappa f^*(\boldsymbol{x}_{ij})+b_{i0}+b_{i1}t_{ij},\qquad \nu_{\text{W}}=1.5,$$

generated by inversion, $T=\{-\log U/e^{\eta}\}^{1/\nu_{\text{W}}}$ with U uniform on $(0,1)$, where f^* is f standardised on a pilot sample. The multiplier κ is calibrated so that the population C-index of the fixed effect alone is 0.75 (Lemma 1 with $\tau=\infty$), which gives $\kappa=1.54$ for the linear and $\kappa=1.96$ for the nonlinear log-hazard and aligns the two scenarios in terms of target fixed-effect concordance. Censoring times are independent exponential variables with a rate chosen to give 30% censoring; a further scenario has 50% censoring (nonlinear, $p=170$). The design yields seven scenarios (Table 2).

Table 2. Simulation scenarios. Every scenario has 50 clusters of 20 units, a random intercept and slope with variances 0.5 and correlation 0.6, Weibull baseline hazard with shape 1.5, and 100 train/test splits.

Scenario	Log-hazard	p	Active predictors	Censoring
S1–S3	linear, f_{lin}	6, 170, 2000	6	30%
S4–S6	nonlinear, f_{non}	6, 170, 2000	6	30%
S7	nonlinear, f_{non}	170	6	50%

Competing methods

MEXGB is compared with six methods, each fitted with its own published R package. *CoxFrailty* is the log-normal frailty Cox model with a linear predictor and the same random intercept and slope (`coxme`[6]); it is the correctly specified model in the linear $p=6$ scenario. *XGBCox* is `xgboost` with the `survival:cox` objective[12], the cluster-blind base learner of MEXGB. *GBM* is boosting of the Cox partial likelihood (`gbm`[9]); *RSF* is the random survival forest (`randomForestSRC`[13]) with trees, *mtry* candidate predictors per split and node size selected via 5-fold cross-validation; *mboost* is componentwise linear boosting of the Cox model (`glmboost`[11]); and *CoxLasso* is the lasso Cox model (`glmnet`[32]) with the penalty chosen by 5-fold cross-validation with clusters kept together. For all comparator tree-boosting methods, the number of trees was determined via 5-fold cross-validation under the prespecified benchmark configuration, using a common maximum depth of 2 and a learning rate of 0.05. MEXGB employed the same depth constraint and learning rate, and selected the number of trees out-of-fold within each model fit (Methods), up to a maximum of 1000 trees. MEXGB and CoxFrailty use the conditional risk score $f(\boldsymbol{x})+\boldsymbol{z}^\top\hat{\boldsymbol{b}}_i$; the remaining methods use their fitted log relative hazard, and RSF the logarithm of its ensemble mortality, without a cluster effect. Table 3 summarises the configuration.

Table 3. Methods compared and their configuration.

Method	Settings	Frailty
MEXGB	Algorithm 1; $\nu = 0.05$, depth 2, $\lambda = 1$, number of trees out of fold, relaxed lasso Cox, Poisson PQL	intercept + slope
CoxFrailty	`coxme`, linear predictor	intercept + slope
XGBCox	`xgboost`, `survival:cox`, common budget	none
GBM	`gbm`, Cox partial likelihood, common budget	none
RSF	`randomForestSRC`, trees, node size and *mtry* via 5-fold	none
mboost	`glmboost`, `CoxPH`, iterations selected via 5-fold cross-validation, step 0.05	none
CoxLasso	`glmnet`, penalty by 5-fold cross-validation	none

Evaluation

For each scenario, we generate a dataset and perform repeated random splitting (10×10 repetitions), resulting in a total of 100 train/test partitions. In every split 30% of the units within every cluster (6 of 20) form the test set. All clusters are therefore represented in training, allowing the frailty models to use estimated cluster effects when predicting held-out units. This is the prediction task of a multi-centre study in which new patients are treated in participating centres. We report:

- *Discrimination accuracy*: Harrell's C-index[38] of the test risk scores.
- *Overall accuracy*: the integrated Brier score (IBS)[40,41], with inverse-probability-of-censoring weights, over 50 grid points up to the 75th percentile of the test follow-up. Every method's survival curves are $\exp\{-\hat{\Lambda}_0(t)e^{\hat{\eta}}\}$ with the Breslow estimator computed from its own training risk scores. This applies the same baseline-estimation procedure to all risk scores, although the estimated baseline differs between methods. The comparison therefore evaluates the complete score-to-survival construction for each method. For RSF the construction treats the log ensemble mortality as a log relative hazard, which is an approximation; its native survival estimator may give a different integrated Brier score.
- *Paired comparisons*: the difference between MEXGB and a competitor on the same split, the percentage of splits on which MEXGB is better, and the two-sided Wilcoxon signed-rank test. Split-level standard errors and tests are reported as descriptive summaries; the main interpretation concerns the magnitude and consistency of the performance differences.
- *Variable selection*: the proportion of active predictors selected (true positive rate, TPR) and the proportion of inactive predictors selected (false positive rate, FPR), using the six top-ranked predictors according to each method's importance measure.
- *Computation*: the elapsed time of one fit on one core.

Simulation results

Discrimination

Table 4 and Fig. 2 report the test C-index, and Table 5 the paired comparisons of MEXGB with the best competitor and with its base learner.

The main findings are as follows.

1. *Linear log-hazard, low dimension.* The correctly specified CoxFrailty model and MEXGB have very similar C-indices (0.817 and 0.816; paired difference -0.0003, MEXGB better on 43% of the splits). The absolute difference is small, consistent with recovery of the linear signal by the relaxed lasso component and little additional contribution from the trees.
2. *Linear log-hazard, high dimension.* At $p = 170$ and $p = 2000$, where CoxFrailty was not fitted under the benchmark configuration due to computational difficulty, MEXGB (0.818, 0.791) improves on the best competitor, CoxLasso (0.748, 0.730), by mean differences of 0.070 and 0.061, respectively, with an advantage on every split. The sparse linear Cox models have the right form of the log-hazard but no frailty (Corollary 1); the tree ensembles omit frailty and approximate the linear signal using trees, achieving 0.58–0.73.

Table 4. Simulation: test C-index, mean (standard error) over 100 train/test splits, 30% censoring. Higher is better; the best value in each column is in bold. CoxFrailty is fitted for $p \leq 50$ only.

	Linear log-hazard			Nonlinear log-hazard		
Method	$p = 6$	$p = 170$	$p = 2000$	$p = 6$	$p = 170$	$p = 2000$
MEXGB	0.816 (0.001)	**0.818 (0.001)**	**0.791 (0.002)**	**0.806 (0.001)**	**0.797 (0.002)**	**0.798 (0.002)**
CoxFrailty	**0.817 (0.001)**	–	–	0.691 (0.002)	–	–
XGBCox	0.725 (0.002)	0.683 (0.002)	0.647 (0.002)	0.749 (0.002)	0.732 (0.003)	0.725 (0.002)
GBM	0.727 (0.002)	0.691 (0.002)	0.655 (0.002)	0.750 (0.002)	0.737 (0.002)	0.727 (0.002)
RSF	0.711 (0.002)	0.636 (0.002)	0.580 (0.002)	0.748 (0.002)	0.710 (0.002)	0.669 (0.003)
mboost	0.751 (0.002)	0.745 (0.002)	0.730 (0.002)	0.641 (0.002)	0.624 (0.003)	0.605 (0.002)
CoxLasso	0.758 (0.002)	0.748 (0.002)	0.730 (0.002)	0.641 (0.002)	0.626 (0.003)	0.621 (0.003)

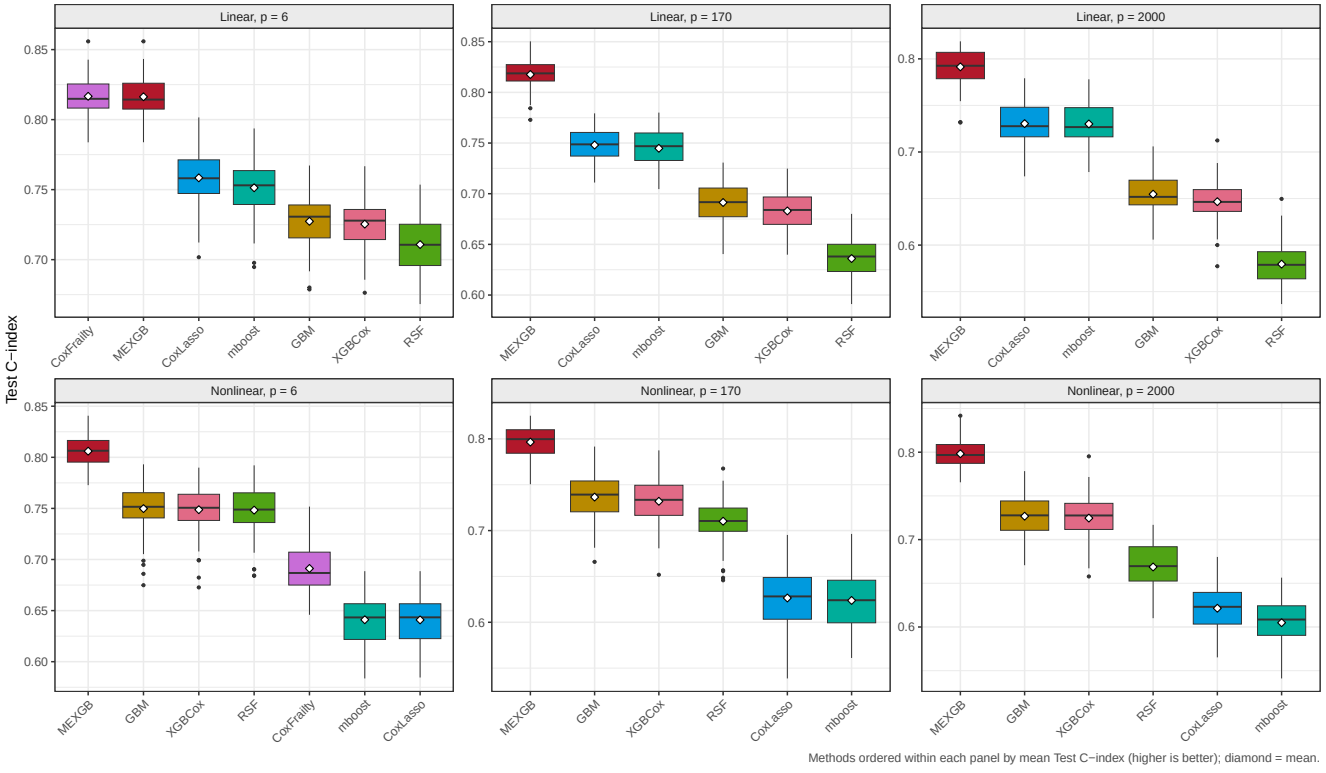

Figure 2. Simulation: test C-index over the 100 train/test splits of each scenario (30% censoring). Within each panel the methods are ordered by their mean C-index (diamonds), from the highest on the left.

Table 5. Simulation: paired comparison of the test C-index of MEXGB with the best competitor (by mean) and with its base learner XGBCox on identical splits. Diff.: mean of MEXGB minus competitor; Better: percentage of splits on which MEXGB has the higher C-index; p: two-sided Wilcoxon signed-rank test.

Scenario	Competitor	Splits	Diff.	Better (%)	p
Linear, $p = 6$	CoxFrailty	100	−0.0003	43	0.044
	XGBCox	100	0.0909	100	< 0.001
Linear, $p = 170$	CoxLasso	100	0.0696	100	< 0.001
	XGBCox	100	0.1346	100	< 0.001
Linear, $p = 2000$	CoxLasso	100	0.0609	100	< 0.001
	XGBCox	100	0.1448	100	< 0.001
Nonlinear, $p = 6$	GBM	100	0.0562	100	< 0.001
	XGBCox	100	0.0573	99	< 0.001
Nonlinear, $p = 170$	GBM	100	0.0600	100	< 0.001
	XGBCox	100	0.0646	100	< 0.001
Nonlinear, $p = 2000$	GBM	100	0.0715	100	< 0.001
	XGBCox	100	0.0737	100	< 0.001
Nonlinear, $p = 170$, 50% censored	GBM	100	0.0580	100	< 0.001
	XGBCox	100	0.0589	100	< 0.001

3. *Nonlinear log-hazard.* MEXGB (0.806, 0.797, 0.798) improves on the best competitor, GBM (0.750, 0.737, 0.727), by mean differences of 0.056–0.072, with an advantage on every split. The linear methods reach 0.60–0.69; CoxFrailty, with the true frailty structure but a linear log-hazard, reaches 0.691 at $p = 6$, consistent with the limitation described in Corollary 2.

4. *Gain over the base learner.* MEXGB improves on XGBCox by 0.091–0.145 in the linear and by 0.057–0.074 in the nonlinear scenarios. From $p = 170$ to $p = 2000$ the C-index of MEXGB falls by 0.027 with the linear log-hazard and remains essentially unchanged with the nonlinear one, whereas those of XGBCox, GBM and RSF fall by 0.036–0.056 and 0.007–0.041, respectively.

The C-index of MEXGB exceeds the population C-index of the fixed effect (0.75) because its risk score contains the predicted frailty, which provides additional information for predicting held-out failure times within an observed cluster.

Integrated Brier score and censoring

MEXGB has the lowest integrated Brier score in every scenario (0.115–0.132; Table 6, Supplementary Fig. S1), tied with CoxFrailty in the linear $p = 6$ scenario and ahead of the next best method by 0.033–0.037 (CoxLasso, linear $p \geq 170$) and by 0.026–0.035 (GBM, nonlinear). These Brier-score results complement the discrimination findings. Proposition 1 explains why probability calibration matters even when risk ranking is accurate. With 50% instead of 30% censoring (scenario S7), MEXGB retains its advantage (0.821 against 0.763 for GBM and 0.762 for XGBCox; paired difference 0.058, better on every split) and the lowest integrated Brier score (0.117 against 0.142; Supplementary Table S1). All reported C-indices increase in the higher-censoring scenario. Harrell's index depends on the censoring distribution, which changes the set of comparable pairs; these higher values should therefore not be interpreted as evidence that additional censoring improves prediction.

Variable selection and computational time

In the linear scenarios MEXGB places all six active predictors among its six most important predictors at $p = 6$ and $p = 170$ and, on average, 91% of them at $p = 2000$, the highest rate of all methods (CoxLasso 89.5%, mboost 87.8%, XGBCox 70.7%), with false positive rates of 0.00% and 0.02% (Supplementary Tables S2 and S3). In the nonlinear scenarios MEXGB, XGBCox and GBM recover 51–54% of the active predictors at $p \geq 170$ whereas the random survival forest recovers 68–81%. The predictors x_1, x_2 and x_6 act only through interactions, which may explain why the deeper forest trees rank them more successfully than the depth-two boosted trees. MEXGB takes 10–14 s at $p = 6$, about 38 s at $p = 170$ and 10–13 min at $p = 2000$ on one core, 22–31 times as long as XGBCox at $p = 2000$ (Supplementary Table S4); its five fold models can be fitted in parallel. Figure 3 shows the criterion (10) across outer iterations: it decreases over the first two to four iterations, after which the stopping rule ends the alternation; the selected criterion values are similar across the minimum leaf sizes examined.

Table 6. Simulation: test integrated Brier score, mean (standard error) over 100 train/test splits, 30% censoring. Lower is better; the best value in each column is in bold.

	Linear log-hazard			Nonlinear log-hazard		
Method	$p = 6$	$p = 170$	$p = 2000$	$p = 6$	$p = 170$	$p = 2000$
MEXGB	**0.115 (0.001)**	**0.115 (0.001)**	**0.130 (0.001)**	**0.127 (0.001)**	**0.132 (0.001)**	**0.131 (0.001)**
CoxFrailty	**0.115 (0.001)**	–	–	0.175 (0.001)	–	–
XGBCox	0.163 (0.001)	0.186 (0.001)	0.205 (0.001)	0.154 (0.001)	0.166 (0.001)	0.174 (0.001)
GBM	0.161 (0.001)	0.178 (0.001)	0.191 (0.001)	0.153 (0.001)	0.160 (0.001)	0.166 (0.001)
RSF	0.177 (0.000)	0.203 (0.001)	0.210 (0.001)	0.164 (0.001)	0.195 (0.001)	0.210 (0.001)
mboost	0.155 (0.001)	0.157 (0.001)	0.163 (0.001)	0.192 (0.001)	0.198 (0.001)	0.204 (0.001)
CoxLasso	0.146 (0.001)	0.152 (0.001)	0.163 (0.001)	0.192 (0.001)	0.198 (0.001)	0.203 (0.000)

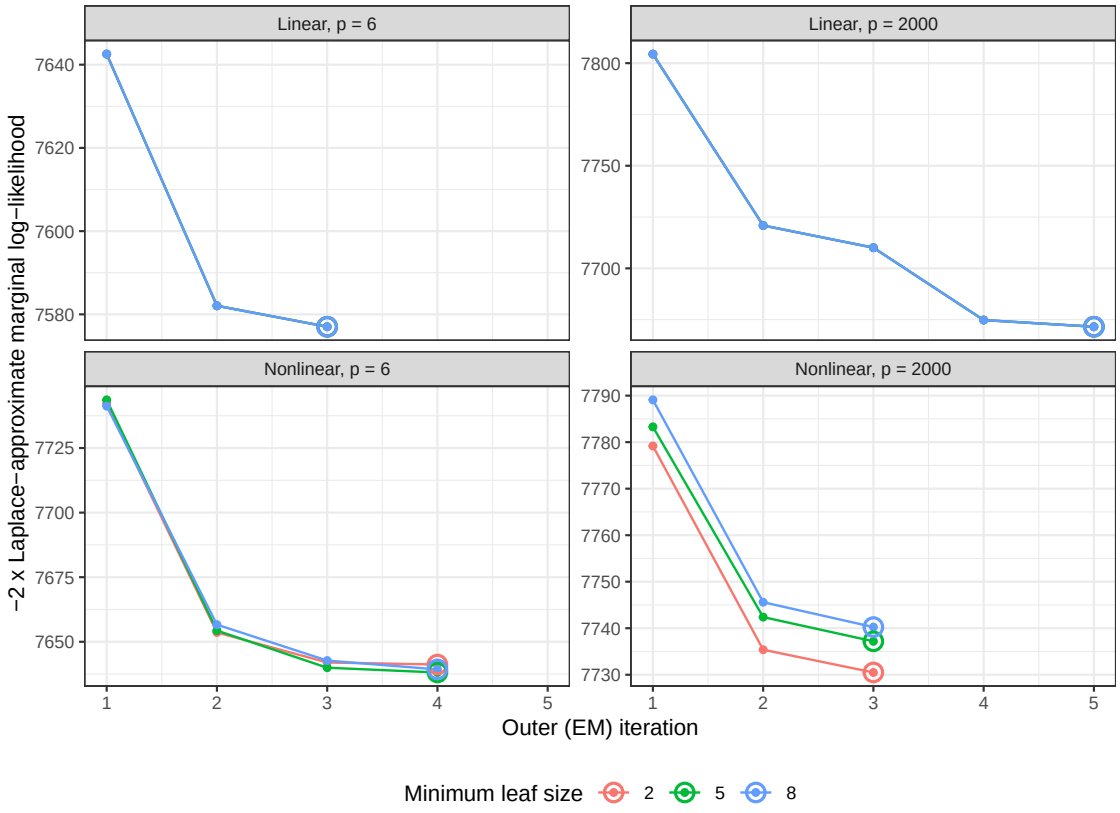

Figure 3. MEXGB in the simulation: $-2\times$ the Laplace approximation of the marginal log-likelihood (10) at the cross-fitted log-hazard along the outer iterations, for one dataset in each of four scenarios and minimum leaf sizes 2, 5 and 8. Open circles mark the iteration returned.

Applications to three clustered survival datasets

Data

We analysed one low-dimensional and two high-dimensional clustered datasets (Table 7).

Insemination[4], distributed with the R package `parfm`, records the time from calving to first insemination of dairy cows clustered in herds. We retained the 174 herds with at least ten cows ($N = 10476$ cows, 11 to 174 per herd, median 57); 9902 cows (94.5%) were inseminated during follow-up. The $p = 4$ predictors are milk urea and protein concentrations, parity and an indicator of first parity, and the random effect is the herd frailty. Herd management can influence insemination timing, so herd membership may provide substantial predictive information.

BreastMulti combines four breast-cancer cohorts profiled on the Affymetrix HG-U133A array: VDX[42] (344 patients), TRANSBIG[43] (198), MAINZ[44] (200) and UNT[45] (125), $N = 867$ in total. The outcome is distant-metastasis-free survival censored at 10 years (235 events, 27.1%).

OvarianMulti combines studies available in the Bioconductor package `curatedOvarianData`[46] with overall-survival information. We retained tumour samples, excluded samples flagged by the package as duplicates or outliers, and included studies with at least 40 patients and 15 deaths. The resulting dataset contains 15 studies and $N = 2269$ patients (42 to 510 per study), with 1232 deaths (54.3%) over a follow-up censored at 10 years.

The two multi-study datasets were preprocessed in the same way. Among the genes measured in every study, genes were ranked by their standard deviation within each study on the original expression scale, the ranks were averaged over studies, and the 2000 genes with the best average rank were kept. Each retained gene was then standardised to mean zero and unit variance within each study, and missing values were set to zero, the study mean. These steps do not use the outcomes and were carried out once, before the data were split. Within-study standardisation adjusts gene-wise means and variances, but does not remove all possible platform or laboratory effects. Differences in baseline survival may reflect patient selection, treatment era and follow-up practice; the study frailty models residual survival heterogeneity between studies.

Table 7. Real clustered survival datasets. Test units per split: 30% of the units within every cluster.

Data set	N	Clusters	Per cluster, median [range]	p	Events (%)	Median follow-up	Test units
Insemination	10476	174 herds	57 [11–174]	4	9902 (94.5)	2.99 months	3144
BreastMulti	867	4 cohorts	199 [125–344]	2000	235 (27.1)	7.75 years	260
OvarianMulti	2269	15 studies	122 [42–510]	2000	1232 (54.3)	2.58 years	680

Predictive performance

Each dataset was split 100 times (ten repetitions of ten random splits), with 30% of the units within every cluster in the test set and the methods and settings of Table 3; CoxFrailty was fitted to the Insemination data only. Table 8 and Fig. 4 report the results, and Table 9 the paired comparisons.

Table 8. Real data: test C-index and integrated Brier score (IBS), mean (standard error) over 100 within-cluster splits. Best value per column in bold.

	Insemination		BreastMulti		OvarianMulti	
Method	C-index	IBS	C-index	IBS	C-index	IBS
MEXGB	**0.6617 (0.0004)**	**0.1548 (0.0001)**	**0.6892 (0.0026)**	**0.1362 (0.0011)**	**0.6296 (0.0014)**	**0.1579 (0.0004)**
CoxFrailty	0.6616 (0.0004)	**0.1548 (0.0001)**	–	–	–	–
XGBCox	0.5104 (0.0005)	0.1795 (0.0001)	0.6696 (0.0028)	0.1518 (0.0013)	0.6098 (0.0013)	0.1657 (0.0005)
GBM	0.5104 (0.0005)	0.1792 (0.0001)	0.6554 (0.0028)	0.1598 (0.0014)	0.6100 (0.0013)	0.1617 (0.0004)
RSF	0.5057 (0.0005)	0.1800 (0.0001)	0.6707 (0.0030)	0.1437 (0.0007)	0.5982 (0.0014)	0.1641 (0.0003)
mboost	0.5081 (0.0005)	0.1789 (0.0001)	0.6835 (0.0027)	0.1374 (0.0010)	0.6182 (0.0014)	0.1586 (0.0004)
CoxLasso	0.5037 (0.0004)	0.1789 (0.0001)	0.6825 (0.0029)	0.1382 (0.0009)	0.6143 (0.0015)	0.1595 (0.0004)

Insemination. MEXGB and CoxFrailty are practically tied (C-index 0.6617 and 0.6616, integrated Brier score 0.1548 for both; MEXGB ahead on 70% of the splits by a mean of 0.0001), while every cluster-blind method has a C-index of 0.50–0.51 and an integrated Brier score of about 0.179. In this analysis, the four cow-level predictors provide little discrimination, while estimated herd effects substantially improve prediction. This pattern illustrates the value of accounting for clustering and is consistent with Corollary 1. MEXGB and the log-normal frailty Cox model perform similarly.

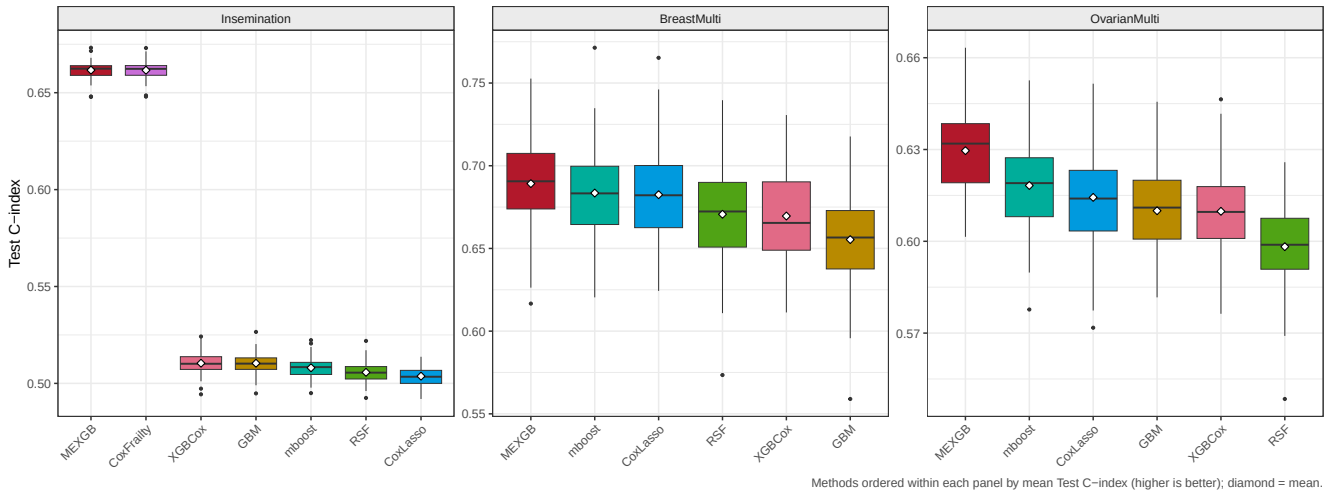

Figure 4. Real data: test C-index over the 100 within-cluster splits, ordered by mean (diamonds) within each panel.

Table 9. Real data: paired comparison of the test C-index of MEXGB with the best competitor (by mean) and with XGBCox on identical splits. Columns as in Table 5.

Data set	Competitor	Splits	Diff.	Better (%)	p
Insemination	CoxFrailty	100	0.0001	70	< 0.001
	XGBCox	100	0.1512	100	< 0.001
BreastMulti	mboost	100	0.0057	63	0.002
	XGBCox	100	0.0196	79	< 0.001
OvarianMulti	mboost	100	0.0114	88	< 0.001
	XGBCox	100	0.0198	89	< 0.001

BreastMulti. MEXGB has the highest C-index (0.689) and the lowest integrated Brier score (0.136), ahead of mboost (0.684, 0.137) and CoxLasso (0.683, 0.138); the paired improvement over mboost is 0.006 (63% of the splits, $p = 0.002$) and over XGBCox 0.020. The tree ensembles reach 0.655–0.671.

OvarianMulti. MEXGB again has the highest C-index (0.630) and the lowest integrated Brier score (0.158), ahead of mboost (0.618) and CoxLasso (0.614); the paired improvement over mboost is 0.011 (88% of the splits, $p < 0.001$) and over XGBCox 0.020. With fifteen studies, the study frailty captures differences in baseline survival between studies, and the strong performance of penalised linear methods suggests that the linear component of MEXGB may capture an important part of the gene-expression signal.

Survival curves

To illustrate the corresponding survival predictions, we took the first of the 100 splits of each dataset (the same split, seeds and settings as in the benchmark), refitted every method, and produced risk-group curves, cluster-specific curves and calibration plots for the test units (Figs. 5–7). The survival curve of every method is $\exp\{-\hat{\Lambda}_0(t)e^{\hat{\eta}}\}$ with the Breslow estimator from its own training risk scores, so all scores are evaluated using the same survival-curve construction. On this split the C-indices of MEXGB were 0.660 (Insemination), 0.702 (BreastMulti) and 0.637 (OvarianMulti), close to the benchmark means of Table 8.

Risk groups. Risk groups were formed separately for each method: the test units were ranked by the method's predicted risk score and split at the tertiles of the test-set scores into approximately equal-sized low-, middle- and high-risk groups for illustration (Fig. 5). Because the groups are formed from predictions only, their separation measures how well each score discriminates. On the herd data the groups of MEXGB and of the frailty model separate clearly (hazard ratio of the high- versus the low-risk group 3.86, 95% confidence interval 3.50–4.26, for MEXGB and 3.89 for CoxFrailty; log-rank $p < 10^{-16}$ for both), whereas those of XGBCox coincide (hazard ratio 1.01, log-rank $p = 0.45$): without the herd, the cow-level predictors do not separate the cows. On BreastMulti the high-risk group of MEXGB has 4.69 times the hazard of its low-risk group (2.46–8.94; log-rank $p = 1.1 \times 10^{-6}$), against 4.09 for mboost ($p = 3.6 \times 10^{-6}$) and 2.32 for XGBCox ($p = 0.002$); 51% of the high-risk patients of MEXGB were metastasis-free at ten years, against 83% of its low-risk patients. On OvarianMulti the groups of MEXGB and mboost separate to a similar extent (hazard ratios 2.41 and 2.50; log-rank $p = 4.1 \times 10^{-11}$ and 3.9×10^{-13}), and more widely than those of XGBCox (2.29; $p = 2.0 \times 10^{-9}$). On this split mboost separated the extreme groups slightly better, while MEXGB achieved higher overall pairwise concordance (C-index 0.637 against 0.626).

Cluster-specific curves. Figure 6 compares, within single herds and studies, the Kaplan–Meier estimate for the test units with their mean predicted survival curve. These comparisons illustrate the contribution of the estimated cluster effects. In the herds with slow insemination (herds 62 and 137) the cluster-blind curve of XGBCox predicts that almost all cows are inseminated within five months, whereas MEXGB follows the observed curve; in the fast herds (166 and 8) XGBCox predicts too slow a decline. The mean absolute distance between the predicted and the observed curve was 0.02–0.21 for MEXGB and 0.22–0.61 for XGBCox in these four herds, and the frailty model gave the same curves as MEXGB. In the breast-cancer cohorts the distance was 0.011–0.027 for MEXGB, against 0.036–0.070 for mboost and 0.035–0.063 for XGBCox, because the cohort frailty shifts the predicted curves of the VDX cohort down and those of the UNT and MAINZ cohorts up. In the ovarian-cancer studies MEXGB was closest to the observed curve in three of the four studies; in the largest study (TCGA, 153 test patients), whose survival is close to the average of all studies, mboost was slightly closer (0.031 against 0.039).

Calibration. Figure 7 plots the observed against the predicted survival probability in quintiles of predicted survival probability at a selected horizon. The mean absolute calibration error of MEXGB was 0.021 on the herd data (CoxFrailty 0.021, XGBCox 0.036) and 0.036 on BreastMulti at five years (mboost 0.065, XGBCox 0.093). On OvarianMulti at three years MEXGB (0.053) and mboost (0.048) were similarly calibrated and both better than XGBCox (0.077). The cluster-blind XGBCox spreads its predicted probabilities too widely on BreastMulti and OvarianMulti: its low-risk quintile is predicted to fare better, and its high-risk quintile worse, than observed, a pattern consistent with an inflated log-hazard, which the MEXGB calibration step is designed to address (Proposition 1).

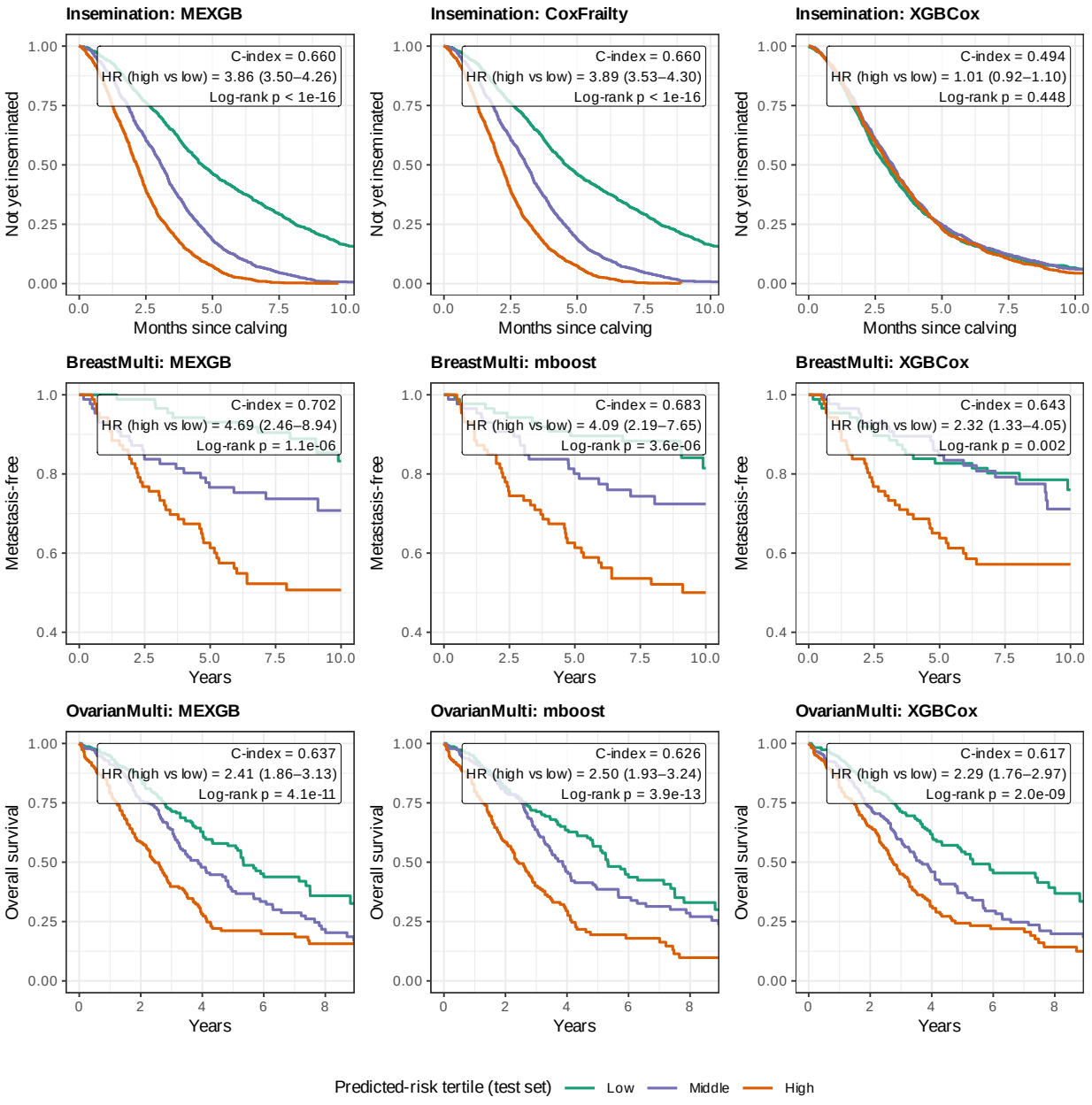

Figure 5. Kaplan–Meier curves of the test units in three risk groups, for MEXGB (left), the best competitor (middle) and the cluster-blind base learner XGBCox (right), on one within-cluster split of each dataset. For each method, the test units were ranked by the method's own predicted risk score and divided at the tertiles of the test-set scores into three approximately equal-sized groups (low, middle and high risk); no outcome information was used to form the groups. Labels give the test C-index, the hazard ratio (95% confidence interval) of the high- versus the low-risk group from a Cox model with the group as covariate, and the *p*-value of the log-rank test of equality of the three curves (two degrees of freedom).

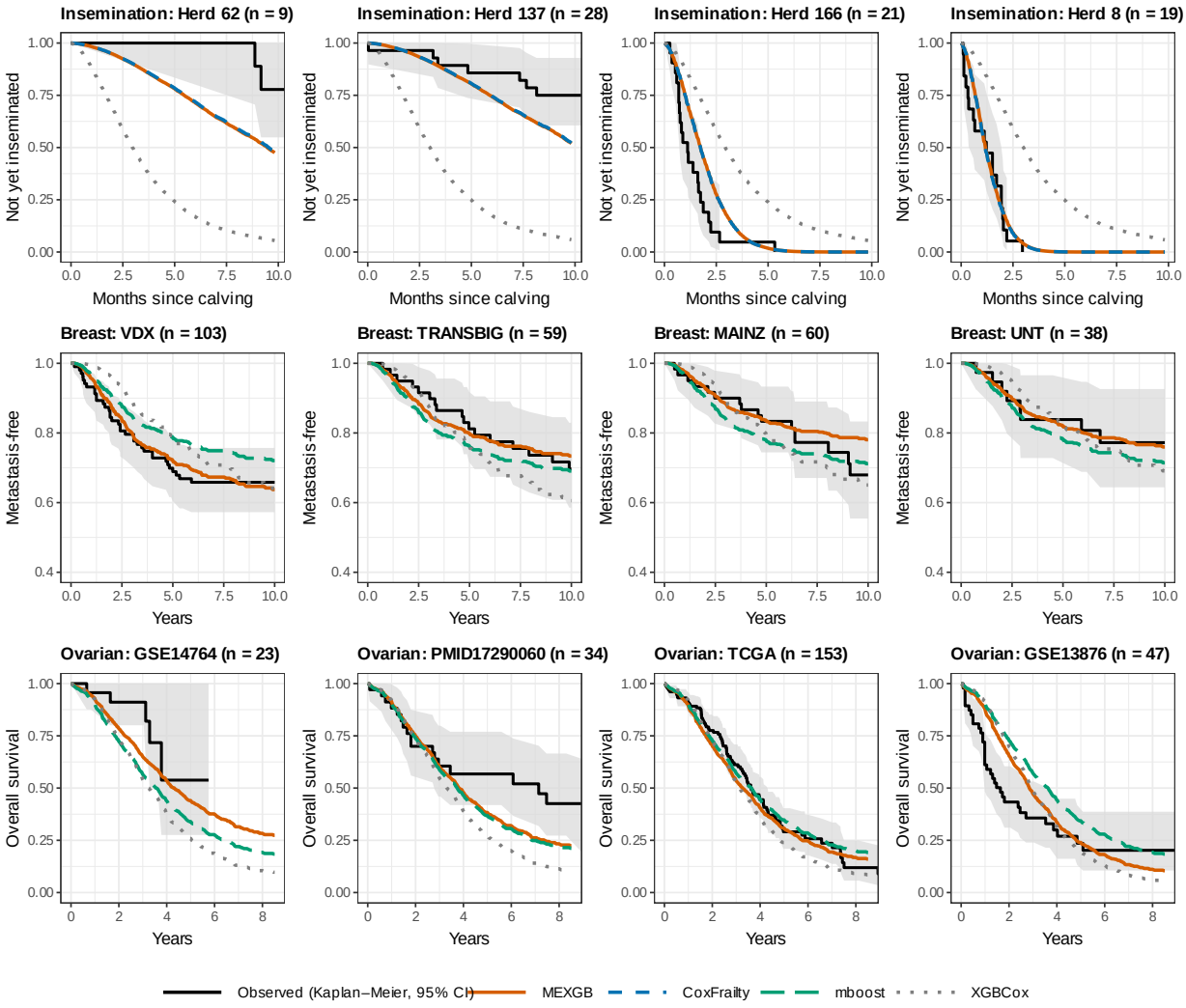

Figure 6. Observed and predicted survival in individual clusters of the test set (legend at the bottom). Black step function and grey band: Kaplan–Meier curve of the test units of the cluster with 95% confidence band. Coloured curves: mean predicted survival curve of the same units under MEXGB (orange, solid), the best competitor (CoxFrailty, blue dashed, top row; mboost, green long-dashed, middle and bottom rows) and XGBCox (grey, dotted). In the top row the MEXGB and CoxFrailty curves overlap. Top row: four herds of the Insemination data (the two herds with the lowest and the two with the highest observed test event rate among herds with at least eight test cows); middle row: the four breast-cancer cohorts; bottom row: four ovarian-cancer studies selected in the same way as the herds.

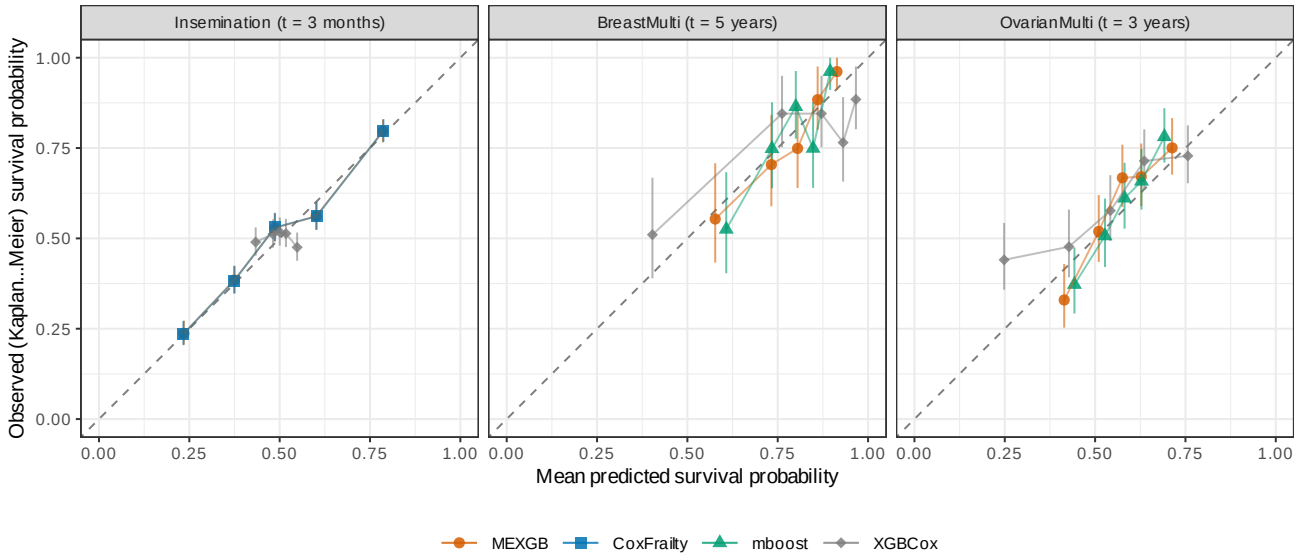

Figure 7. Calibration at a fixed horizon (three months, five years and three years): Kaplan–Meier survival estimate against the mean predicted survival probability, with 95% confidence intervals, in quintiles of the predicted probability of each method. The dashed line is perfect calibration. On the Insemination data the MEXGB and CoxFrailty points coincide.

Discussion

This study investigated methods for predicting survival outcomes for newly enrolled individuals within previously observed clusters under high-dimensional conditions, allowing for potentially nonlinear covariate effects. Three principal results emerged. First, in simulation studies, MEXGB achieved the highest mean C-index whenever competing approaches omitted either the clustering structure or the correct functional form. In these settings, MEXGB improved the C-index by 0.056–0.072 relative to the strongest competing method and by 0.057–0.145 relative to its own cluster-blind base learner. When a correctly specified frailty model was available, MEXGB exhibited performance that was effectively indistinguishable from that model. Second, MEXGB yielded the lowest integrated Brier score across all simulated scenarios, with one scenario showing a tie with the correctly specified frailty model. Third, across three empirical datasets, MEXGB attained the highest mean C-index and the lowest integrated Brier score; improvements were substantial compared with cluster-blind learners and modest compared with the best-performing linear methods in the gene-expression cohorts. Below, we examine these results in detail and situate them in the context of recent related work.

Cluster effect

The most pronounced impact of frailty was observed in the herd-level dataset: all cluster-blind approaches achieved C-indices of only 0.50–0.51, whereas both MEXGB and the frailty Cox model attained 0.66. Consistently, the risk-group survival curves indicated no meaningful separation among the three XGBCox groups (log-rank $p = 0.45$), while the corresponding MEXGB groups exhibited clear stratification. These results suggest that cow-level covariates contribute negligible information for predicting time to insemination, whereas herd membership (i.e., the clustering structure) contains substantial prognostic signal, motivating the use of these data as an illustrative example for frailty modelling[4]. Corollary 1 further indicates that this performance ceiling is not attributable to any specific learning algorithm: under the stated assumptions, omission of the cluster effect yields a strictly positive expected loss in concordance relative to the conditional log-hazard. The same mechanism accounts for the finding that MEXGB cluster-specific curves closely matched the empirical survival patterns of slow versus fast herds and high- versus low-risk breast-cancer cohorts, whereas cluster-blind learners produced curves that were systematically shrunk toward the marginal (population-average) survival.

These findings add to recent comparative evaluations of machine-learning methods for survival analysis. Spooner et al.[27] compared ten learning algorithms for high-dimensional dementia prediction, and Suliman et al.[14] evaluated oblique and standard random survival forests against Cox models across 18 simulated scenarios; in both studies, observations were assumed to be independent. In contrast, our simulation results indicate that under clustered data-generating mechanisms, explicit adjustment for cluster structure can materially improve performance even when highly flexible learners are employed: across all scenarios, the best-performing cluster-agnostic (cluster-blind) learners underperformed MEXGB by at least 0.056.

Analogous benefits of incorporating random effects into tree-based ensembles have been reported for uncensored outcomes; for example, MEGB reduced the mean squared error of mixed-effects random forests by 35–76%[20], and Gaussian process boosting outperformed boosting approaches that ignore dependence[19]. The present results extend this empirical pattern to the setting of right-censored survival times via MEXGB.

Flexible fixed effect

Under the nonlinear log-hazard, the frailty Cox model exhibited a marked reduction in discrimination, with the C-index decreasing from 0.817 to 0.691, whereas MEXGB showed only a modest decline (from 0.816 to 0.806). Consistently, the frailty model's integrated Brier score exceeded that of MEXGB by 0.048. These findings align with Corollary 2. Under its assumptions, the same theoretical constraint extends to penalised frailty approaches such as `PenCoxFrail`[8], which perform variable selection while maintaining a linear log-hazard specification.

In contrast, when the log-hazard was linear, MEXGB achieved performance comparable to the correctly specified model, plausibly because its relaxed lasso component captured the linear effect structure. Among the cluster-blind methods, the random survival forest yielded the lowest accuracy in the linear log-hazard setting, consistent with Suliman et al.[14], who reported superior discrimination for Cox models relative to random survival forests in most scenarios. Notably, however, in nonlinear settings the forest most effectively recovered active predictors (68–81% versus 51–54% for the boosting methods at $p \geq 170$), suggesting that increased tree depth may facilitate the identification of interaction effects when the number of predictors is large.

Recent flexible frailty methods have been studied in different settings. Porndumnernsawat et al.[24] found that the accuracy of Bayesian frailty survival trees increased with the number and size of clusters and decreased as censoring increased, in simulations with 50 covariates. Murris et al.[25] reported C-indices of about 0.68–0.74 for random survival forests for recurrent events with 300 covariates, against 0.49–0.55 for the Ghosh–Lin model, which illustrates the cost of a rigid model in high dimensions in a different form of clustered data. Our simulations extend this evidence to shared frailties with up to 2000 predictors: the C-index of MEXGB changed by at most 0.027 between 170 and 2000 predictors, and its integrated Brier score did not deteriorate when censoring increased from 30% to 50%. The increase in Harrell's C-index with censoring reflects the changing set of comparable pairs rather than better prediction.

Gene-expression cohorts

Across the two multi-study cohorts, the performance improvements of MEXGB relative to sparse linear Cox models were modest, with increases of 0.006 and 0.011 in mean C-index, respectively; among the alternatives, penalised linear approaches constituted the most competitive baselines. This pattern is consistent with prior findings in comparable settings. Haibe-Kains et al.[47] reported that complex genome-wide predictors did not outperform a single proliferation gene for breast-cancer prognosis, with validation C-indices ranging from 0.56 to 0.72. Similarly, Waldron et al.[48] summarised C-indices of 0.54–0.60 for fourteen published ovarian-cancer signatures evaluated across ten datasets. In the present analysis, the observed C-indices of 0.69 (breast) and 0.63 (ovarian) are near the upper bounds of these reported ranges; however, they were obtained by predicting patients from studies that were represented in the training data, in contrast to the cross-study validation protocols used in the aforementioned work. This within-study prediction setting may yield optimistic performance due to the ability to leverage estimated study effects.

Mechanistically, when the underlying association between predictors and outcome is approximately linear, the linear component can explain a substantial fraction of the prognostic signal, whereas the frailty term can accommodate study-level heterogeneity. The out-of-fold selection strategy further constrains model complexity, favouring a small tree ensemble when incremental nonlinear structure is limited. Calibration performance was heterogeneous across tasks: MEXGB exhibited superior calibration relative to the two comparators shown for the breast-cancer cohorts, and calibration comparable to mboost for the ovarian-cancer cohorts.

Limitations

Several limitations should be considered when interpreting these results. First, all methods relied on a Breslow baseline hazard, thereby implicitly treating the logarithm of the random-forest mortality estimate as a log–relative hazard. This mapping is approximate and does not assess the random forest under its native survival function estimator. Second, MEXGB presupposes proportional hazards conditional on the frailty assumption, which is satisfied in the present simulation setting. In addition, at $p = 2000$, MEXGB required approximately 22–31 times the runtime of a single boosted Cox model; however, the fold-specific models can be trained in parallel.

Methodological extensions warranting further investigation include: (i) competing-risks analyses implemented through cause-specific hazard models; (ii) nested frailty formulations (e.g., patients nested within surgeons nested within hospitals); and (iii) the construction of uncertainty intervals for subject-specific survival curves that appropriately propagate uncertainty arising from frailty estimation.

Conclusion

MEXGB integrates a relaxed-lasso Cox submodel, Newton-boosted decision trees, and log-normal frailty terms to perform survival prediction in clustered, high-dimensional settings. Cross-fitted fixed-effect predictions are used to enable Poisson penalized quasi-likelihood (PQL) estimation of the frailty component, with an additional calibration slope applied to rescale these predictions. Under the stated assumptions, the accompanying theoretical analysis characterises the degradation in predictive accuracy arising from either omission of cluster-level effects or adoption of a misspecified linear log-hazard structure. In simulation studies, MEXGB increased the C-index by 0.056–0.072 relative to the strongest competing method when alternative approaches lacked either the frailty or the nonlinear boosting component; it exhibited comparable performance to the correctly specified frailty model and attained the smallest integrated Brier score under the common construction of survival curves. Across three empirical applications, MEXGB yielded the highest mean predictive accuracy among the methods evaluated and generated interpretable cluster-specific survival curves. Collectively, these results support the use of MEXGB for prediction in clustered survival (time-to-event) analyses. An implementation of MEXGB for survival outcomes is available in the R package `MEXGB`.

Data availability

The Insemination dataset is provided within the R package `parfm`. The breast cancer cohorts are available through the Bioconductor packages `breastCancerVDX`, `breastCancerTRANSBIG`, `breastCancerMAINZ`, and `breastCancerUNT`, whereas the ovarian cancer studies are distributed via the Bioconductor package `curatedOvarianData` (version 1.44.0). All datasets are publicly accessible and de-identified; accordingly, no additional data collection was undertaken for this study. The simulated datasets can be reproduced using the code available at https://github.com/rid4stat/MEXGB-validation.

Code availability

The `MEXGB` R package (version 1.0.0) is publicly accessible via GitHub at https://github.com/rid4stat/MEXGB. The full set of simulation scripts and data-analysis code necessary to reproduce the results reported in this study is available at https://github.com/rid4stat/MEXGB-validation.

References

1. Vaupel, J. W., Manton, K. G. & Stallard, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439–454 (1979).
2. Clayton, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141–151 (1978).
3. Hougaard, P. *Analysis of Multivariate Survival Data* (Springer, New York, 2000).
4. Duchateau, L. & Janssen, P. *The Frailty Model* (Springer, New York, 2008).
5. Ripatti, S. & Palmgren, J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**, 1016–1022 (2000).
6. Therneau, T. M., Grambsch, P. M. & Pankratz, V. S. Penalized survival models and frailty. *J. Comput. Graph. Stat.* **12**, 156–175 (2003).
7. Gorfine, M. & Zucker, D. M. Shared frailty methods for complex survival data: a review of recent advances. *Annu. Rev. Stat. Its Appl.* **10**, 51–73 (2023).
8. Groll, A., Hastie, T. & Tutz, G. Selection of effects in Cox frailty models by regularization methods. *Biometrics* **73**, 846–856 (2017).
9. Ridgeway, G. The state of boosting. *Comput. Sci. Stat.* **31**, 172–181 (1999).
10. Binder, H. & Schumacher, M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinforma.* **9**, 14 (2008).
11. Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. & Hofner, B. Model-based boosting 2.0. *J. Mach. Learn. Res.* **11**, 2109–2113 (2010).
12. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).

13. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals Appl. Stat.* **2**, 841–860 (2008).

14. Suliman, A. *et al.* Comparison of oblique random survival forest, random survival forest, and statistical models for time-to-event data using simulation study. *Sci. Reports* **15**, 44005 (2025).

15. Katzman, J. L. *et al.* DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).

16. Wu, R. *et al.* Neural frailty machine: beyond proportional hazard assumption in neural survival regressions. In *Advances in Neural Information Processing Systems*, vol. 36 (2023).

17. Hajjem, A., Bellavance, F. & Larocque, D. Mixed-effects random forest for clustered data. *J. Stat. Comput. Simul.* **84**, 1313–1328 (2014).

18. Capitaine, L., Genuer, R. & Thiébaut, R. Random forests for high-dimensional longitudinal data. *Stat. Methods Med. Res.* **30**, 166–184, DOI: 10.1177/0962280220946080 (2021).

19. Sigrist, F. Gaussian process boosting. *J. Mach. Learn. Res.* **23**, 1–46 (2022).

20. Olaniran, O. R., Olaniran, S. F., Allohibi, J., Alharbi, A. A. & Alharbi, N. M. Mixed effect gradient boosting for high-dimensional longitudinal data. *Sci. Reports* **15**, 30927 (2025).

21. Asadi, F., Homayounfar, R., Mehrali, Y., Masci, C. & Zayeri, F. Innovative statistical method for longitudinal and hierarchical data modeling: the gmexgboost method. *BMC Med. Res. Methodol.* **26**, 24 (2026).

22. Basak, P., Linero, A., Sinha, D. & Lipsitz, S. Semiparametric analysis of clustered interval-censored survival data using soft Bayesian additive regression trees (SBART). *Biometrics* **78**, 880–893 (2022).

23. Hu, L., Ji, J., Liu, H. & Ennis, R. A flexible approach for assessing heterogeneity of causal treatment effects on patient survival using large datasets with clustered observations. *Int. J. Environ. Res. Public Heal.* **19**, 14903 (2022).

24. Porndumnernsawat, P., Frank, T. D. & Ingsrisawang, L. On a Bayesian multivariate survival tree approach based on three frailty models. *Sci. Reports* **15**, 12017, DOI: 10.1038/s41598-025-96198-x (2025).

25. Murris, J., Bouaziz, O., Jakubczak, M., Katsahian, S. & Lavenu, A. Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event. *BMC Med. Res. Methodol.* **25**, 262, DOI: 10.1186/s12874-025-02678-z (2025).

26. Zhou, R., He, K., Wang, D. *et al.* Neural network for correlated survival outcomes using frailty model. *J. Data Sci.* **23**, 624–637 (2025).

27. Spooner, A. *et al.* A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Sci. Reports* **10**, 20410 (2020).

28. Prevett, M. L., Hui, F. K. C., Tho, Z. Y., Welsh, A. H. & Westveld, A. H. Gradient boosted mixed models: Flexible estimation of mean and variance components for clustered data. *arXiv preprint arXiv:2511.00217* (2025).

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

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

31. Breslow, N. E. Discussion of the paper by D. R. Cox. *J. Royal Stat. Soc. Ser. B* **34**, 216–217 (1972).

32. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).

33. Meinshausen, N. Relaxed lasso. *Comput. Stat. & Data Analysis* **52**, 374–393 (2007).

34. van Houwelingen, J. C. & le Cessie, S. Predictive value of statistical models. *Stat. Medicine* **9**, 1303–1325 (1990).

35. Whitehead, J. Fitting Cox's regression model to survival data using GLIM. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **29**, 268–275 (1980).

36. Breslow, N. E. & Clayton, D. G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **88**, 9–25 (1993).

37. Tierney, L. & Kadane, J. B. Accurate approximations for posterior moments and marginal densities. *J. Am. Stat. Assoc.* **81**, 82–86 (1986).

38. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *J. Am. Med. Assoc.* **247**, 2543–2546 (1982).

39. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L. J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Medicine* **30**, 1105–1117 (2011).

40. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Medicine* **18**, 2529–2545 (1999).

41. Gerds, T. A. & Schumacher, M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48**, 1029–1040 (2006).

42. Wang, Y., Klijn, J. G. M., Zhang, Y. *et al.* Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365**, 671–679 (2005).

43. Desmedt, C., Piette, F., Loi, S. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* **13**, 3207–3214 (2007).

44. Schmidt, M., Böhm, D., von Törne, C. *et al.* The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68**, 5405–5413 (2008).

45. Sotiriou, C., Wirapati, P., Loi, S. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262–272 (2006).

46. Ganzfried, B. F. *et al.* curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database* **2013**, bat013 (2013).

47. Haibe-Kains, B., Desmedt, C., Sotiriou, C. & Bontempi, G. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics* **24**, 2200–2208 (2008).

48. Waldron, L. *et al.* Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *J. Natl. Cancer Inst.* **106**, dju049 (2014).

Acknowledgements

Not applicable.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author contributions statement

ORO and SKA designed the simulation study. ORO and SKA conducted the simulation study, analysed the results, and drafted the manuscript. ORO, SKA, and SAA developed the `MEXGB` package. SAA, AB, AI and NMA contributed to the conceptualisation and study design, interpretation of results, and critical revision of the manuscript. ORO and SAA provided oversight of the project. All authors read and approved the final manuscript.

Additional information

Competing interests: The authors declare no competing interests.

Supplementary information accompanies this paper: proofs of the theoretical results, additional simulation tables and figures, and a tutorial for the `MEXGB` package.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Frailtyawareextremegradientboostingsupplementary18112026.pdf