Towards filling the gaps around recurrent events in high dimensional framework: a systematic literature review and application

Juliette Murris, Anais Charles-Nelson, Abir Tadmouri Sellier, Audrey Lavenu, Sandrine Katsahian

Towards Filling the Gaps around Recurrent Events in High Dimensional Framework: A Systematic Literature Review and Application

J. Murris[1,2,3]*, A. Charles-Nelson[4,5], A. Tadmouri[3], A. Lavenu[6,7,8]†, S. Katsahian[1,2,4,5,9]†

[1] Inserm, Centre de Recherche des Cordeliers, Université Paris Cité, Sorbonne Université, Paris, France;

[2] HeKA, Inria, PariSanté Campus, Paris, France;

[3] RWE & Data, Pierre Fabre, Boulogne-Billancourt, France;

[4] Unité de Recherche Clinique, Hôpital Européen Georges-Pompidou, Assistance Publique – Hôpitaux de Paris (AP-HP), APHP, Centre, Paris, France;

[5] Centre d'Investigation Clinique 1418 (CIC1418) Épidémiologie Clinique, Inserm, Paris, France;

[6] Faculté de Médecine, Rennes, Université de Rennes 1, France;

[7] Institut de Recherche Mathématique de Rennes (IRMAR), Rennes, France;

[8] Centre de Investigation Clinique (CIC) CIC 1414, Inserm, Université de Rennes 1, Rennes, France;

[9] Service d'Informatique Médicale, Biostatistiques et Santé Publique, Hôpital Européen Georges Pompidou, Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, France

*J. M. is the corresponding author, address: 33 Av. Emile Zola, 92100 Boulogne-Billancourt, France, phone: +33 6 73 52 36 39; e-mail: juliette.murris@pierre-fabre.com

†A.L. and S.K. contributed equally.

ORCID

Juliette Murris, https://orcid.org/0000-0002-7017-9865

Anaïs Charles-Nelson, https://orcid.org/0000-0001-6437-7059

Audrey Lavenu, https://orcid.org/0000-0002-0049-2397

Sandrine Katsahian, https://orcid.org/0000-0002-7261-0671

Word count: 4,172 words

Preprint available at https://arxiv.org/abs/2203.15694

Towards Filling the Gaps around Recurrent Events in High-Dimensional Framework: A Systematic Literature Review and Application

Individuals may experience repeated events over time. However, there is no consensus about learning approaches to use in a high-dimensional framework for survival data (when the number of variables exceeds the number of individuals, i.e., $p > n$). The aim of this study was to identify learning algorithms for analyzing/predicting recurrent events, and to compare them to standard statistical models on simulated data. A systematic literature review was conducted to provide state-of-the-art methodology. Data was then simulated according to the number of variables, the proportion of active variables, and the number of events. The performance of the models was assessed using Harrell's concordance index, Kim's C-index, and error rate for active variables. Seven publications were identified, of which four were methodological studies, one an application paper and two were reviews. On simulated data, the standard models failed when $p > n$. Penalized Andersen-Gill and frailty models outperformed, whereas RankDeepSurv gave poorer performances. With no current guidelines on a specific approach to use, this study deepens understanding of the mechanisms and limits of investigated methods in this context.

Keywords: recurrent events; survival analysis; high-dimensional data; machine learning; simulated data

1. Introduction

Individuals may experience repeated events over time, such as hospitalizations or cancer relapses. In either clinical trials or real-world settings, survival analysis usually focuses on modeling the time to the first event. However, variables may have a different effect on the first event and on subsequent occurrences. Thus, modeling recurrent events remains a big challenge (Figure 1).

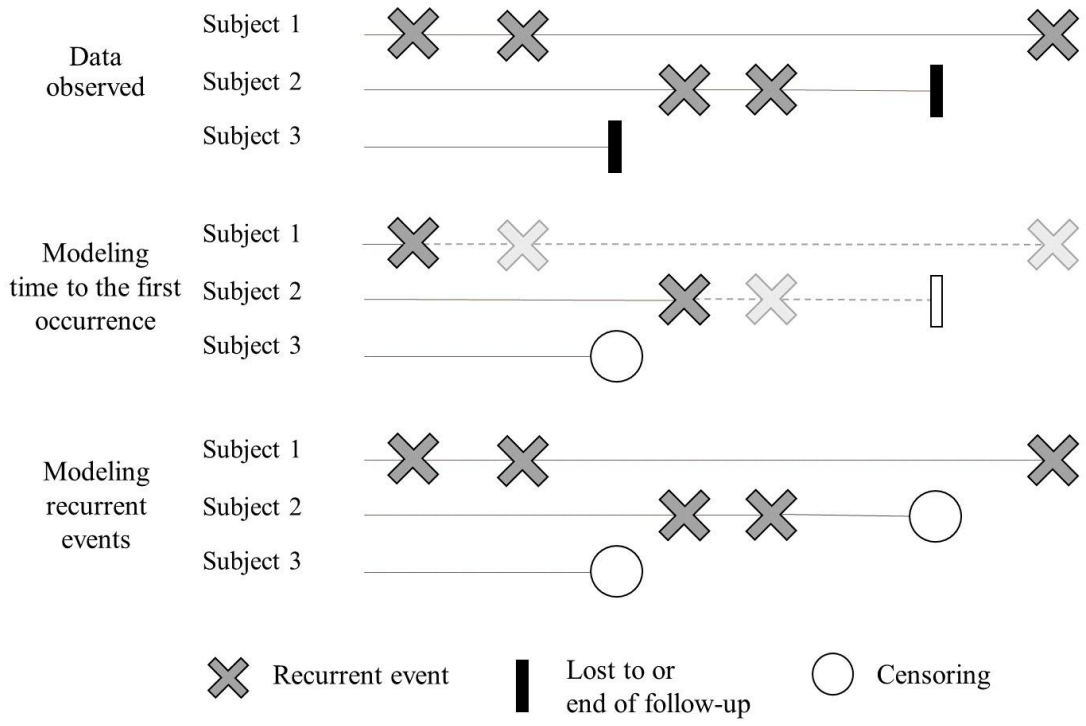

Figure 1. Recurrent Event Framework

Indeed, two main problems arise when analyzing recurrent events. Firstly, interindividual heterogeneity emerges as some subjects may be more likely than others to experience the event. Secondly, events experienced by an individual are not independent, leading to intraindividual heterogeneity. Various methods have been developed to deal with these two issues and can be classified into marginal and conditional models. Marginal models

involve implicitly averaging over the history of previous recurrent events. Conditional models can condition on event past history.

Furthermore, modern technologies enable data to be generated on thousands of variables or observations, as per genomics, medico-administrative databases, disease monitoring by intelligent medical devices, etc. While massive data describes large numbers of observations, high-dimensional data is defined as data with a number of variables of interest p greater than the number of individuals n. In this context, standard statistical models may no longer be applied, as they tend to face convergence problems and non-clinically relevant significance of the variables can arise. Machine learning methods have been developed to handle these problems.

Literature reviews were previously conducted on recurrent events, but none dealt with a high dimensional framework [1-3]. The aim of this article was to review innovative methodology available to analyze and predict high-dimensional recurrent events data. Simulations were performed to study properties of identified methods compared to standard methods, according to the number of variables at the modeling stage.

Section 2 hereafter describes the methodological setting regarding both the literature review and the statistical approaches for modeling and evaluating the models as well as the data simulation scheme. Then, section 3 provides the findings of the review that enabled the identification of the adequate methods. Next, the application results on simulated data are reported. Section 4 finally relates the discussion and gives a contextual perspective based on related work and theoretical considerations.

2. Materials and method

2.1. Systematic literature review

A systematic literature review (SLR) was performed to identify high-dimensional survival methods for analyzing recurrent events.

2.1.1. Data sources and search strategy

A first search in PubMed, as recommended by Cochrane [4,5], was performed in October 2022 to identify published articles with the following concepts and index terms (MeSh terms) in the title or abstract:

- Survival analysis,
- Recurrent event,
- High-dimensionality,
- Machine Learning.

Appendix 1 details the complete PubMed search strategy.

Secondly, hand searches were also carried out via research engines (Google, Google Scholar, Science direct, Web of Science) and conferences (International Society for Clinical Biostatistics, Association for Computing Machinery, Machine Learning Conference, Journées de Statistique, Medical and Health Informatics) to seek unpublished work such as conference abstracts, papers, and reviews [6]. The hand search strategy included the distinct concepts above that were combined using the following key terms: ‘survival’ or ‘survival analysis’ or ‘time to event’, ‘recurrence’ ‘recurrent’ or ‘repeated events’ or ‘relapse’ or ‘hospitalization’, ‘high-dimension’ or ‘machine learning’.

2.1.2. Eligibility criteria

Only published articles in English or French were included. Inclusion criteria were systematic or observational studies that analyzed any recurrent outcome(s), as well as reviews and/or surveys. Exclusion criteria were any Bayesian approach and clinical trial design. The rationale behind this strategy was to ensure consistency with frequentist approaches and real-world applications. Unstructured data such as textual or imaging data were not considered; in our opinion such data are disparate from structured data. Finally, no restrictions regarding the field of healthcare, medical indication or treatment were applied.

2.1.3. Study selection

Two reviewers assessed the eligibility of publications independently and any discrepancies were subsequently discussed. Forward and backward citation tracking was conducted to avoid missing any relevant literature. Eligible hits were subjected to title and abstracts screening after duplicate removal. The findings of this selection led to the next step in the systematic review process which was the full-text review.

2.1.4. Study characteristics

Study characteristics such as general study setting, location, sample size if applicable, research design, statistical/machine learning approaches, outcomes measured, metrics for evaluation, code availability / reproducibility and application of data (sample description if applicable) were extracted for each included study and summarized. Heterogeneity across studies was not assessed as it was deemed irrelevant to the objective.

2.2. Statistical analysis for application

2.2.1. Notations

Let $\boldsymbol{X}_i$ be a *p*-dimensional vector of covariates, β the associated regression coefficients, $\lambda_0(t)$ the baseline hazard function, $Y_i(t)$ an indicator of whether subject *i* is at risk at time *t*, $\delta_i = 1$ when the subject experienced the event (else 0). Let E_i and C_i be the time to event or censoring, $T_i = E_i \wedge C_i$ for the patient *i*, with $a \wedge b = \min(a, b)$. $N_i^*(t)$ denoted the number of events over the interval [0, *t*]. Of note, *i = 1, ..., n*, with *n* the number of subjects and $\mathbf{X} \in \mathbb{R}^{\boldsymbol{n*p}}$ denoted the covariates matrix for all subjects.

2.2.2. Standard statistical models for modeling recurrent events

Andersen-Gill (AG) [7], Prentice, William and Peterson (PWP) [8], Wei-Lin-Weissfeld (WLW) [9] and the frailty models [10] were developed as extensions of the Cox model [11]. These methodologies commonly use models to handle recurrent event data. Their characteristics are summarized in Table 1. Further details on time scales and how models accounted for subject at risk can be found in Appendix 2. While other statistical approaches exist to model recurrent events, we focused on risk outputs to be able to compare methodologies to one another containing identical metrics. However, this statistical model can handle low-dimensional data only, i.e., when the number of individuals is lower than the number of variables.

Table 1. Standard Statistical Models for Recurrent Events Analyses

Model	Components and specificities
AG	Conditional model, accounts for the counting process as a time scale and unrestricted set for subjects at risk Recurrent events within individuals are independent and share a common baseline hazard function Intensity of the model: $\lambda_i(t) = Y_i(t) \times \lambda_0(t) \times \exp(\beta^t X_i)$
PWP	Conditional model, counting process as time scale and restricted set for subjects at risk Stratified AG, stratum *k* collects all the k^{th} events of the individuals Hazard function for each event Hazard function: $\lambda_{ik}(t) = Y_i(t) \times \lambda_{0k}(t) \times \exp(\beta_k^t X_i)$
WLW	Marginal model, also stratified, calendar time scale and semi-restricted set for subjects at risk Intra-subject dependence Hazard function: $\lambda_{ik}(t) = Y_i(t) \times \lambda_{0k}(t) \times \exp(\beta_k^t X_i)$
Frailty	Extension of AG model Random term z_i for each individual to account for unobservable or unmeasured characteristics Hazard function: $\lambda_i(t) = Y_i(t) \times \lambda_0(t) \times z_i \times \exp(\beta^t X_i)$

AG, Andersen-Gill; PWP, Prentice, William and Peterson; WLW, Wei-Lin-Weissfeld.

2.2.3. Evaluation criteria to measure performance

Few methods are currently available to evaluate a model adjusted for recurrent events. This leads to a lack for model discrimination, i.e., the model cannot differentiate between high- and low-risk individuals who may be subject to the events. We selected the following three criteria to answer the study objectives:

Harrell's Concordance index. Harrell's C-index is a common evaluation criterion in survival analysis [12]. This measure is the proportion of pairs of individuals for which the order of survival times are concordant with the order of the predicted risk. In the presence of censoring, the denominator is the number of pairs of individuals with an event. The C-index is estimated as follows

$$\hat{\mathbb{C}} = \frac{\sum_{i \neq j} I\{\eta_i < \eta_j\} \times I\{T_i > T_j\} \times \delta_j}{\sum_{i \neq j} I\{T_i > T_j\} \times \delta_j} \quad (1)$$

With η_i the risk of occurrence of the event. Of note, when two individuals are censored, we cannot know which of the two has the event first. This pair is not included in the calculation. In the same way, if one of the individuals is censored and its censoring time is lower than the event time of another individual, we cannot know which one has the event first. This pair is also not included in the C-index calculation. If the C-index is equal to 1, it means a perfect prediction, and if the C-index ≤ 0.5, it implies that the model behaves similarly or worse than random. Models with a C-index close to 1 are preferred. Harrell's C-index was computed at each event.

Kim's C-index. Kim et al. [13] proposed a measure of concordance between observed and predicted event counts over a time interval of shared observations. It is the proportion of pairs of individuals for whom the risk prediction and the number of observed events is concordant:

$$\hat{\mathbb{C}}_{rec} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\right\} \times I\{\beta^t X_i > \beta^t X_j\}}{\sum_{i=1}^{n}\sum_{j=1}^{n} I\left\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\right\}} \quad (2)$$

This extension of the C-index implies:

- Two individuals are comparable up to the minimum time of follow-up;
- A pair contributes to the denominator if the two event counts are not equal.

As per Harrell's C-index in Equation (1), a score closes to 1 indicates a better performance of the model. As opposed to Harrell's C-index, Kim's C-index was computed once across all the events.

Error rate for active variables. When simulating the datasets, the active status of each variable is known. Methods report the significant variables with a p-value < 0.05 (except deep neural networks). Significant variables are considered as positive tests for their active status. Some active variables likely have a false negative test (FN), and some passive variables have a false

positive test (FP). The error rate (err) is the proportion of misclassified variables after prediction:

$$err = \frac{FP+FN}{p} \quad (3)$$

2.2.4. Simulation scheme

The following assumptions were made:

- Active variables were continuous, and have the same (non-zero) effect;
- The variables do not vary over time;
- Individuals were at risk continuously until end of follow-up;
- Censoring is not informative.

The generation of the covariate matrix, $\mathbf{X} \sim \mathcal{N}_m(\mu, \Sigma(\rho))$. $\mu = \left(\mu_1 \dots \mu_p\right) = (a \dots a)$ and $\Sigma(\rho)$ was the covariance matrix with an autoregressive correlation structure and $\rho \in (0,1)$. The coefficients $\beta = \left(\beta_1 \dots \beta_p\right) = (b, \dots, b, 0, \dots, 0)$ were associated with the p covariates. m coefficients were equal to a constant $b \in \mathbb{R}$ (the value of the active coefficients) and $p - m$ coefficients were equal to zero. The sparse rate was described by $\frac{m}{p}$. The baseline hazard function followed a Weibull distribution with scale $\alpha > 0$ and shape $\gamma > 0$, and $\lambda_0(t) = \alpha\gamma t^{\gamma-1}$. The cumulated baseline hazard function could be expressed as $\Lambda_0(t) = \int_0^t \lambda_0(s)\mathrm{d}s = \alpha t^{\gamma}$. Hence the cumulative hazard function could be expressed as $\Lambda(t) = \Lambda_0(t) \exp(\beta^t X_i)$. Conditional baseline hazard function was then defined as $\widetilde{\Lambda}_t(u) \coloneqq \widetilde{\Lambda}^i(u|T_{i-1} = t) = \Lambda(u + t) - \Lambda(t)$. A frailty term z_i i.i.d. was incorporated to account for heterogeneity.

To maintain censoring rates, censored individuals were randomly drawn (censoring is not informative), as per Jahn-Eimermacher et al. [14]. The algorithm of Jahn-Eimermacher et

al. [15] was applied to simulate event times k for each subject *i*:

$$t_{i,1} = \Lambda^{-1}(t)(-\log(\varepsilon_1)) \quad (4)$$

$$\text{and } t_{i,k+1} = t_{i,k} + \widetilde{\Lambda}_{i,t_k}{}^{-1}(-\log(\varepsilon_{k+1})) \text{ with } \varepsilon_k \sim U[0,1]$$

Train-test split was employed with a 70-30% distribution. Datasets were generated with:

- N = 100 subjects (low sample size)
- Censoring rate of 20%
- $\rho = 0.7$
- b = 0.15
- $\alpha = 1$ and $\gamma = 2$
- $z \sim \mathcal{G}amma(0.25)$

Scenarios include variations of the number of covariates p = 25, 50, 100, 150, and 200 and the sparse rate = 0%, 25%, and 50%. For each one of the 15 scenarios, 100 datasets were generated to account for variability.

3. Results

3.1. Systematic literature review

The search strategy is summarized in Figure 2. Extraction led to the identification of 192 hits through electronic research on the PubMed database. Forty-one studies proceeded to the full-text review step, while the other 151 remaining papers were excluded for further consideration. Overall, after confirming the outcome of interest dealt with recurrence, the primary reason for exclusion was the non-consideration of recurrent events as time-to-event

for each occurrence. Recurrence was considered as a classifier (19/192), as a recurrence-free survival outcome (26/192), or as a time-to-first event (34/192). In this way, the challenge of recurrent data was avoided. The probability of the event was estimated without considering all available information (subsequent event occurrences were omitted in such cases), and were hence biased.

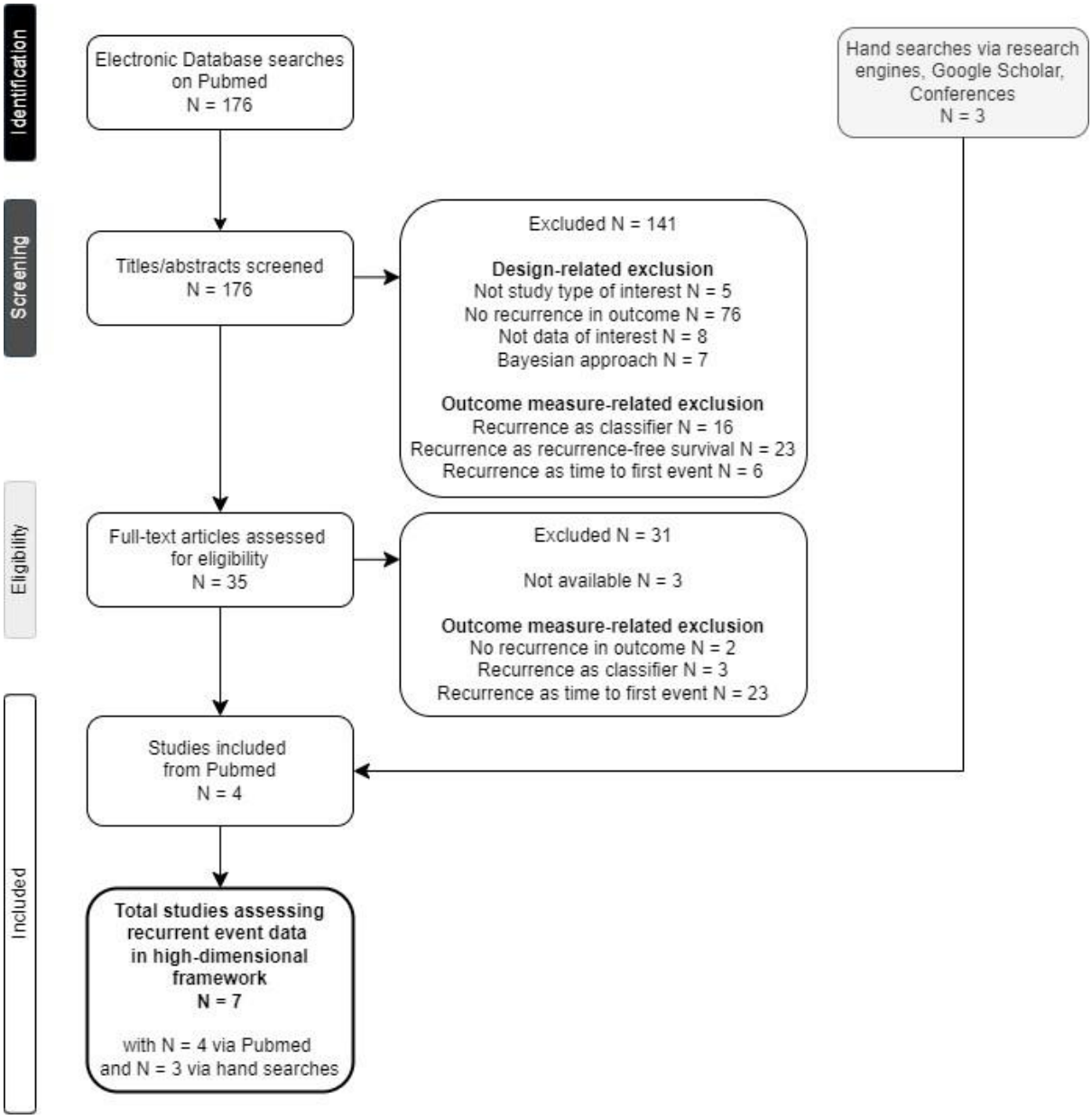

Figure 2. Flowchart of Included Publications via PubMed

This may be the illustration of authors' caution when dealing with recurrent events in high (framework?) dimensions, as no published guidelines or recommendations are available as far as we know. In the field of medicine, the most frequent disease application was cancer (77/192) probably due to the high level of sustained interest in this condition. Among cancers, however, no type stood out (colon/colorectal cancer 34/192, breast cancer 20/192, lung cancer 15/192, other cancer 8/192). In addition, four full-text articles could not be reviewed as they were not

available. After title, abstract and full-text thorough review, four publications were included from the electronic database search. Three additional papers from the hand searches were identified. Subsequently, a total of seven relevant publications were selected (Table 2).

The two first publications are literature reviews. Work from Wang et al. [16] and Bull et al. [17] are recent comprehensive surveys in which classical and contemporary methods of survival analysis are reported. They both underline the development over the last decade of more complex approaches to dealing with longitudinal data to predict survival outcomes, e.g., joint models and deep neural networks. Recurrent events may be seen as longitudinal outcomes but were not addressed per se and were only mentioned as specific data structure.

Four methodological articles were also selected presenting a significant variation in methodology. Two articles describing learning algorithms for variable selection strategies were identified. Firstly, Wu [18] focused on accelerating coefficient estimation with a coordinated descent algorithm and penalizing partial likelihood, followed by Zhao et al. [19] work which provided an extension of Ridge penalization for estimating and selecting variables simultaneously. Finally, the other two selected articles are from Gupta et al. [20] and Jing et al. [21], and developed deep neural networks extensions for the analysis of recurrence.

An additional paper was selected, which aimed at estimating time between two breast cancer recurrences using a Weibull Time To Event Recurrent Neural Network [22]. However, the methodology used was an extension of a recurrent neural network and was not published in any peer-reviewed journal [23].

Findings from the present review highlight the current gap in the literature and vast differences in the context and methods of interest. In particular, not all developed models were based on simulated data, as Jing et al. [21]. Additionally, none of the included

publications compared their performance to others. For instance, variable selection approaches were compared to standard statistical model only, while neural networks were compared to other neural networks or random forests. No head-to-head comparison across standard methods, learning algorithms and deep neural network seem to have been performed.

Table 2. Papers Identified from the Literature Review

#	Year	Author	Title	Type	Description	Data used for application	Evaluation measure	Code availability / reproducibility
#1	2013	Wu et al.	Lasso penalized semiparametric regression on high-dimensional recurrent event data via coordinate descent	Variable selection	Regularization method with penalization Use of the coordinated descent algorithm, which computes in a bi-directional way (forward and backward) the deviations of the optimization problem and updates the parameter value iteratively	Chronic septic granulomatosis	Number of selected predictor variables and regression coefficients	No
#2	2018	Zhao et al.	Variable selection for recurrent event data with broken adaptive ridge regression	Variable selection	Extension of the broken adaptive ridge method to recurrent events, involves repetition and reweighting of penalized L_2 models Simultaneous variable selection and parameter estimation, accounts for clustering effects	Chronic septic granulomatosis	MSE Number of predictor variables selected correctly, and number of predictor variables selected incorrectly	Yes
#3	2019	Wang et al.	Machine Learning for Survival Analysis: A Survey	Literature review	Introduction to survival analysis, overview of classical methods and overview of learning methods Recurrent events are mentioned, but ML methods are not developed	/	/	/
#4	2019	Gupta et al.	CRESA: A Deep Learning Approach to Competing Risks, Recurrent Event Survival Analysis.	Deep neural networks	LSTM neural networks with the introduction of the cumulative incidence curve to take into account competitive and/or recurrent events	MIMIC III Machine failure data	Harrell's C-index MAE	No

#5	2019	Jing et al.	A deep survival analysis method based on ranking	Deep neural networks	Extension of the DeepSurv model (neural networks for competitive events) with the use of ranking in the loss function on the differences between observed and predicted values	Myocardial infarction Breast cancer omic data	Harrell's C-index	Yes
#6	2020	Bull et al.	Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods	Literature review	Summary of existing methodology to provide clinical prediction depending on the nature on input data Both statistical and learning approaches are described, but no ML methods for recurrent events highlighted	/	/	/
#7	2021	Kim et al.	Deep Learning-Based Prediction Model for Breast Cancer Recurrence Using Adjuvant Breast Cancer Cohort in Tertiary Cancer Center Registry	Deep neural networks	Use of Weibull Time To Event Recurrent Neural Network, an extension of recurrent neural network to sequentially estimate time to next event	Breast cancer registry in Korea	Harrell's C-index MAE	Yes

LSTM, long short-term memory; MAE, mean absolute error. Articles were sorted by publication year. #1 to #3 were identified via hand searches and #4 to #7 via PubMed.

3.2. Application to simulated data

We proposed testing the identified methods on simulated data. Only two methods had open-sourced code: Variable selection from Zhao et al. [19] and the deep neural network RankDeepSurv from Jing et al. [21].

3.2.1. Methods selected

Learning algorithms for variable selection

A common approach to addressing high-dimension challenge is variable selection. Penalizing models helps to reduce the space of parameter coefficients, called shrinkage. Widely used for regression and classification problems, Lasso penalization accepts null coefficients to select variables [24] and Ridge helps to deal with multicollinearity in the data [25]. Both penalization approaches have been extended to Cox models in standard survival analysis framework [26,27]. The purpose is to solve a constrained optimization problem of the partial log-likelihood of the Cox model, which is written

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} \delta_i \beta^t \boldsymbol{X}_i - \sum_{i=1}^{n} \delta_i \times \log \sum_{j \in \mathcal{R}(\tau_i)}^{n} \exp(\beta^t \boldsymbol{X}_i) \quad (5)$$

With $\mathcal{R}(t)$ the set of individuals who are 'at risk' for the event at time t. For CoxLasso, regularization is performed using an L_1 norm penalty and $\hat{\beta} = \text{argmax}_{\beta}\, \mathcal{L}(\beta), ||\beta||_1 \leq s$ and for CoxRidge an L_2 norm penalty and $\hat{\beta} = \text{argmax}_{\beta}\, \mathcal{L}(\beta), ||\beta||_2 \leq s$, with $s \geq 0$. The lower the value of s, the stronger the penalization. Hyperparameters, named penalty coefficients, are used to determine its value, and enable the control of the impact of the penalty.

Zhao et al. [19] proposed an extension of these methods to recurrent events by developing the broken adaptive ridge (BAR) regression. The first iteration consists of a penalized L_2 model

$$\hat{\beta}^{(0)} = \text{argmin}_{\beta}\left(-2\,\mathcal{L}_{mod}(\beta) + \xi_n \sum_{j=1}^{p} \beta_j^2\right), \xi_n \geq 0 \quad (6)$$

If penalization hyperparameter $\xi_n > 0$, this is a Ridge penalty, and if $\xi_n = 0$ then $\hat{\beta}^{(0)}$ is not penalized. We update for each iteration ω:

$$\hat{\beta}^{(\omega)} = \text{argmin}_{\beta}\left(-2\,\mathcal{L}_{mod}(\beta) + \vartheta_n \sum_{j=1}^{p} \frac{\beta_j^2}{\left(\hat{\beta}_j^{(\omega-1)}\right)^2}\right), \omega \geq 1 \quad (7)$$

BAR estimates are defined by $\hat{\beta} = \lim_{k\to\infty} \hat{\beta}^{(\omega)}$. The estimator benefits from the oracle properties of both penalties for model covariate selection and estimation. Cross-validation is recommended to optimize values of hyperparameters ξ_n and ϑ_n. According to Kawaguchi et al. [28], estimates are not sensitive to variations of ξ_n and optimization can be performed only on ϑ_n. In the absence of a consensual single measure on cross-validation under recurrent events, two values for ϑ_n were studied in this paper, thereby covering two separate models. Such penalty was applied to models presented in the previous subsection.

Deep neural network

RankDeepSurv is a deep neural network proposed by Jing et al. [21] with fully connected layers (all neurons in one layer are connected to all neurons in another layer). The specificity of the RankDeepSurv neural network lies in the loss function adapted to survival, which results from the sum of two terms: one to constrain the survival model using an extension of the mean square error and the other to evaluate the rank error between observed and predicted values for two individuals. The loss function is written as

$$L_{loss}(\theta) = \alpha_1\, L_1(\theta) + \alpha_2\, L_2(\theta) + \mu\, ||\theta||_2^2 \quad (8)$$

with

- $\alpha_1, \alpha_2 > 0$ constant values, θ the weights of the network, μ the regularization parameter for L_2 ;
- $L_1 = \frac{1}{n}\sum_{i=1,I(i)=1}^{n}\left(y_{j,pred} - y_{j,obs}\right)^2$, $I(i) = 1$ if i is censored or if the predicted time to event is before observed time, else 0;
- $L_2 = \frac{1}{n}\sum_{I(i,j)=1}^{n}\left[\left(y_{j,obs} - y_{i,obs}\right) - \left(y_{j,pred} - y_{i,pred}\right)\right]^2$, $I(i,j) = 1$ if $y_{j,obs} - y_{i,obs} > y_{j,pred} - y_{i,pred}$, else 0.

Gradient descent is utilized for solving the minimization of L_{loss}.

3.2.2. Results of the application

Simulated datasets had identical characteristics in terms of number of individuals, structure of covariates, but differed across scenarios in terms of number of covariates and sparse rate. In the variance-covariance matrix, the covariates were highly correlated when they were close, then decreasingly correlated when they were further apart. Appendix Figure B captured this relationship across covariates with five datasets, regardless of the number of covariates. Figure 3 provides a visual representation of the history of nine individuals and their events over the follow-up period (and helps to understand Figure 1).

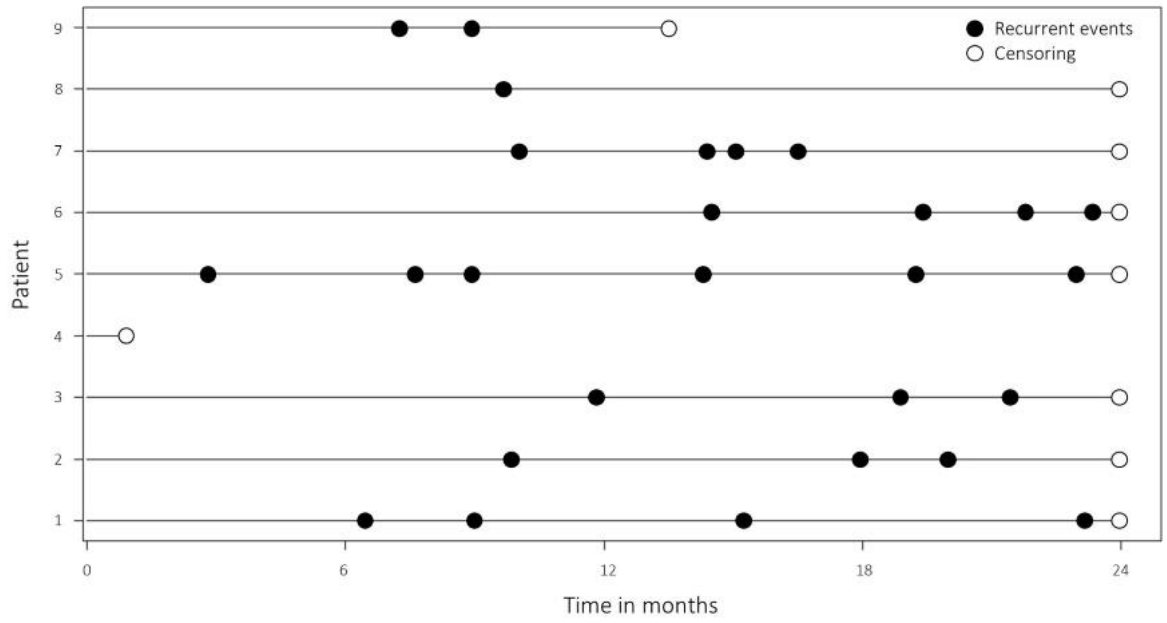

Figure 3. History of the 9 First Simulated Individuals (for a given training set). A row referred to a patient with their event history. The time on the x-axis was the follow-up period. Solid circles corresponded to events and empty circles represented censoring.

Impact of the number of covariates on the average C-index

Average C-indices were computed across all 15 scenarios (Figure 4). As expected, the standard models failed as soon as $p > n$. Whereas the C-indices were also expected to be around 0.5 when the sparse rate was zero, they increased as the sparse rate increased. The best performance was obtained using the frailty model. Other models showed similar trends, except for the WLW and RankDeepSurv models. The C-indices of these two models remained around the value of 0.5 (and even below) regardless of the scenario. The Kim's C-index was more stable across the different number of covariates and sparse rates, although it tended to decrease as the number of covariates increased with sparse rate = 50%. Small differences across penalty values were noticed as 0.05 penalized models and 0.1 penalized models followed similar trends.

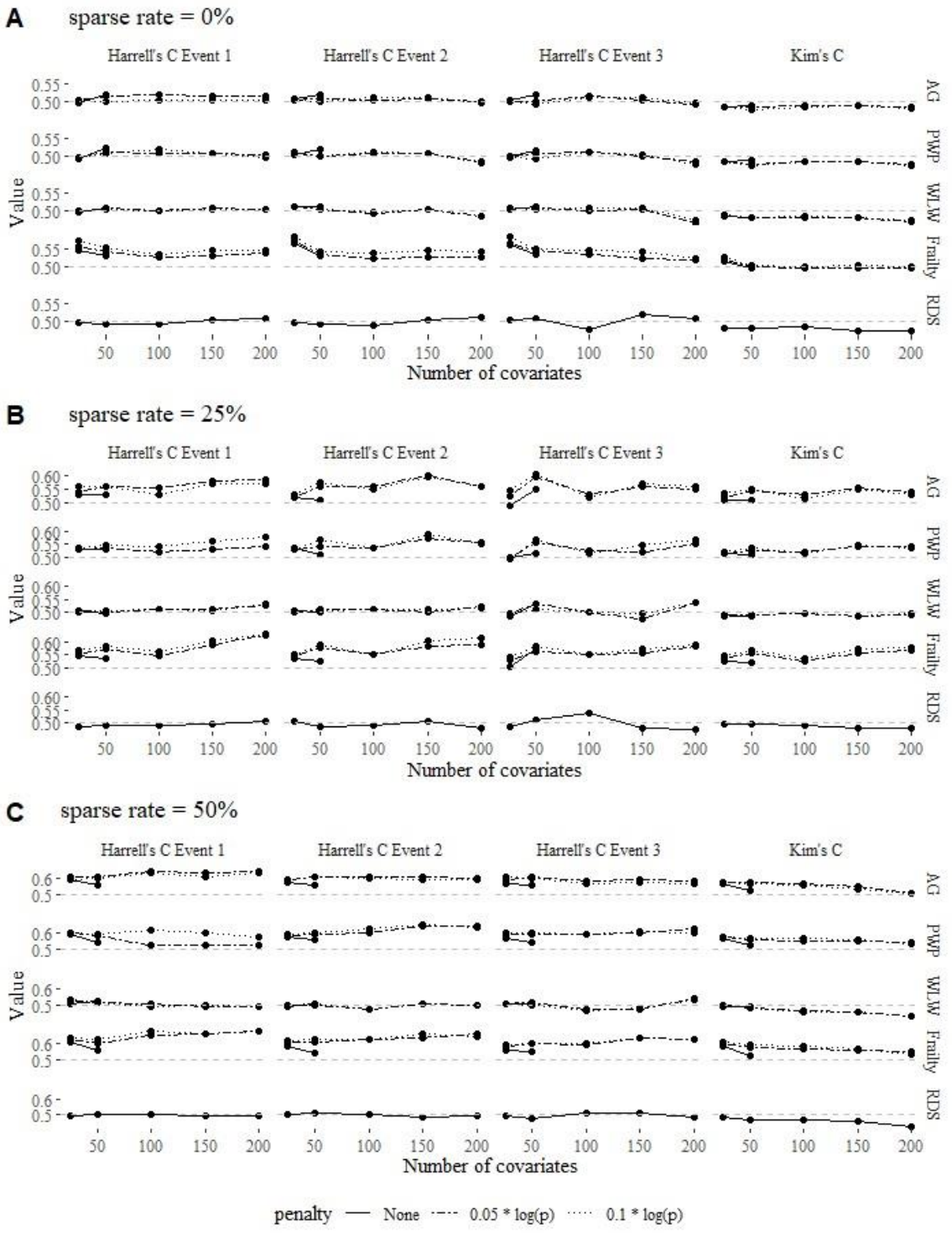

Figure 4. Impact of the number of covariates on average C-indices with Sparse Rate Equal to 0% (A), 25% (B) and 50% (C). p the number of covariates. For each sparse rate, model and penalty, the average C-indices of the 100 simulated datasets were displayed over the number of covariates. The penalties were equal to 0 (unpenalized), $0.05 \times \log(p)$ and $0.1 \times \log(p)$, respectively. Penalties > 0 were applicable only for standard statistical models, RankDeepSurv deep neural network was hence not penalized. Unpenalized standard statistical models did not converge as soon as $p > n$, performance was therefore not available. AG: Andersen-Gill; PWP: Prentice, RDS: RankDeepSurv, William, and Peterson; WLW: Wei-Lin-Weissfeld.

Focus on the variability of C-indices for two extreme scenarios

Two extreme scenarios were thoroughly studied: one with no active variable and only 25 variables (A), and another one in which models overall reported greater performance with a sparse rate of 25% and over 150 variables (B). Similar trends in variability were observed across these two scenarios and for each C-index (Appendix Figure C). Kim's C-index was the less volatile across models and their penalties, with values ranging between 0.39 and 0.63 and 0.28 and 0.76 for (A) and (B), respectively. Harrell's C-index was increasingly variable in the first event (A: min = 0.26 and max = 0.74; B: min = 0.24 and max = 0.81), second event (A: min = 0.30 and max = 0.75; B: min = 0.17 and max = 0.85), and third event (A: min = 1 and max = 0).

Error rate for active variables

Results regarding average error rates are displayed in Appendix Figure D. For scenarios where the sparse rate was equal to 0%, all models reported average error rates below 0.5, except penalized WLW with error rates around 0.75. Average error rates appeared similar when no penalty was applied. The AG model had the lowest average error rate for each p, with a minimum value of 0.018 for the penalized model at $0.1 \times \log(p)$ and $p = 200$. Average error rate decreased when the penalty increased when $p > n$. For scenarios with a sparse rate equal to 25%, the unpenalized frailty model had the best performance, while the other models provided higher values. Similarly, penalties decreased the average error rate. Penalized AG models reported average error rates lower than 0.3. Finally, when the sparse rate was equal to 50%, almost constant average error rates around 0.5 were observed for each model regardless of p.

4. Discussion

The present systematic literature review enabled the identification of emerging approaches. A total of seven publications were included, highlighting the limitation of available resources in this area. Methods herein identified were based on existing model extensions to the recurrent survival framework, which included variable selection approaches and neural networks. As with the standard models presented, they have both strengths and drawbacks. It is therefore necessary to tailor them to the clinical setting in order to meet the stated goals in a meaningful way.

At the same time, these approaches had not tested against one another. This may lead to erratic behavior and confusion when researchers aim to conduct robust and reliable analyses in this context. The present study thus proposed to evaluate some of the available open-sourced innovative learning algorithms developed to solve the high-dimensional framework when considering recurrent events. The investigation of the beforementioned 15 scenarios on simulated data highlighted specificities of both the methodology and measures used for the evaluation of their performance.

Firstly, unpenalized standard approaches failed as soon as $p > n$ as expected, while penalized approaches helped to improve their performance when $p < n$. This was typically expected as standard statistical models were not designed for $p > n$ cases. AG and PWP models reported equivalent performance, while the frailty model consistently had the best performance. This was due to the construction of the frailty term from the simulation scheme. The WLW model performed in an inferior manner, regardless of penalization or not. This finding was consistent with results in the literature, suggesting WLW models to be more appropriate with events of different types rather than recurrent events [29,30]. Nevertheless, these models, each with their own specificities, can respond to differing needs, especially

related to the research questions [1,3,31]. Secondly, variable selection with penalties did not significantly increase performance, and few variables were even selected when the sparse rate was zero. Since only two values for the hyperparameter were explored, it seemed quite unlikely these would maximize model performance. The deep neural network reported poorer performance; one reason could be that the format of the data was not suitable for the code. In this case, average error rates increased with the sparse rate. When the number of active variables was higher, models tended to select the wrong variables. It appears as if the models have a harder time learning and selecting the true active variables in the advent of a high sparse rate, however they managed to report better C-indices in this situation. This was related to the variance-covariance structure chosen for data simulation. With regards to evaluation metrics, Kim's C-index has shown higher stability and robustness compared to other metrics and stood for a criterion evaluating the entire set of events. Harrell's C, on the other hand, was measured at each event, making it difficult to be interpreted in terms of global performance.

Nevertheless, some limitations should be noted. The literature review presented several drawbacks. Publications whose objective was variable selection without explicit dimension reduction, such as Tong et al. [32] and Chen et al. [33] could not be captured because of the elaborated research strategy. In addition, it is not always simple to assess how the outcome was considered, especially for neural networks that make little mention of the expected structure to process the data. Furthermore, as mentioned above, the lack of hyperparameter optimization for variable selection made BAR approach inconclusive. Lastly, a cross-validation would have highlighted the robustness of the results [34].

Other evaluation measures have been used in the literature, e.g., the mean square error, the mean absolute error, the log-likelihood [18,19,30]. An additional approach to investigating active variables would be to assess the importance of the variables by permutation [35]. When

choosing a performance measure beforehand, this consists in permuting k times for the order of the covariates and calculating k times for the performance of the model. We note however that the simulations scheme itself presented several drawbacks. Covariates were not time-dependent and shared the same effect on the outcome, which may seem implausible in real life and made the interpretation of the results difficult to generalize. Also, although the simulation of the data maintained censoring rates, it was not based on a distribution of censoring time, while one should be able to genuinely control [36,37].

5. Conclusion

As far as we know, this is the first study to compare standard methods, variable selection algorithms, and a deep neural network in modeling recurrent events in a high-dimensional framework, and to specifically measure the impact of the number of covariates.

Progress in medical care is leading to the use of embedded artificial intelligence (AI) technologies, evidenced by the booming market for AI medical devices. In this context, these systems are typically designed to prevent the occurrence of events at the hospital, elderly care home or outpatient setting, for example. Where these events are likely to occur repeatedly, and all available data/knowledge is captured, then thorough, robust and appropriate analysis of recurrent events is crucial [38].

Overall, this work raises many concerns for recurrent event data analysis in high-dimensional settings. In addition, it highlights the current need for developing further approaches in order to assess their performance in a relevant manner.

Acknowledgments

The authors would like to thank the reviewers for their constructive comments that led to improvements of the manuscript.

Funding

This work was partially funded by a grant from the Association Nationale de la Recherche et de la Technologie, with Pierre Fabre, Convention industrielle de formation par la recherche number 2020/1701.

Declaration of interest statement

The authors declare no conflicts of interest.

Data availability statement

All data were simulated in line with simulation scheme detailed in the article.

References

1. Rogers JK, Pocock SJ, McMurray JJV, et al. Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to CHARM-Preserved. Eur J Heart Fail. 2014;16(1):33–40.

2. Twisk J, Smidt N, de Vente W. Applied analysis of recurrent events: a practical overview. J Epidemiol Community Health. 2005;59(8):706–10.

3. Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44(1):324–33.

4. Tacconelli E. Systematic reviews: CRD's guidance for undertaking reviews in health care. Lancet Infect Dis. 2010;10(4):226.

5. Higgins JPT, Thomas J, et al. Cochrane Handbook for Systematic Reviews of Interventions. John Wiley & Sons; 2019. 726 p.

6. Guide to the methods of technology appraisal 2013. 2013;94.

7. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. Ann Stat. 1982;10(4):1100–20.

8. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. Biometrika. 1981;68(2):373–9.

9. Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. J Am Stat Assoc. 1989;84(408):1065–73.

10. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography. 1979;16(3):439–54.

11. Cox DR. Regression Models and Life-Tables. J R Stat Soc Ser B Methodol. 1972;34(2):187–202.

12. Harrell FE Jr, Califf RM, Pryor DB, et al. Evaluating the Yield of Medical Tests. JAMA. 1982;247(18):2543–6.

13. Kim S, Schaubel DE, McCullough KP. A C-index for recurrent event data: Application to hospitalizations among dialysis patients. Biometrics. 2018;74(2):734–43.

14. Jahn-Eimermacher A, Ingel K, Ozga AK, et al. Simulating recurrent event data with hazard functions defined on a total time scale. BMC Med Res Methodol. 2015;15(1):16.

15. Jahn-Eimermacher A. Comparison of the Andersen-Gill model with Poisson and negative binomial regression on recurrent event data. Comput Stat Data Anal. 2008;52(11):4989–97.

16. Wang P, Li Y, Reddy CK. Machine Learning for Survival Analysis: A Survey. ACM Comput Surv. 2019;51(6):110:1–110:36.

17. Bull LM, Lunt M, Martin GP, et al. Harnessing repeated measurements of predictor variables for clinical risk prediction: a review of existing methods. Diagn Progn Res. 2020;4:9.

18. Wu TT. Lasso penalized semiparametric regression on high-dimensional recurrent event data via coordinate descent. J Stat Comput Simul. 2013;83(6):1145–55.

19. Zhao H, Sun D, Li G, Sun J. Variable selection for recurrent event data with broken adaptive ridge regression. Can J Stat. 2018;46(3):416–28.

20. Gupta G, Sunder V, Prasad R, Shroff G. CRESA: A Deep Learning Approach to Competing Risks, Recurrent Event Survival Analysis. In: Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing; 2019. p. 108–22.

21. Jing B, Zhang T, Wang Z, et al. A deep survival analysis method based on ranking. Artif Intell Med. 2019;98:1–9.

22. Kim JY, Lee YS, Yu J, et al. Deep Learning-Based Prediction Model for Breast Cancer Recurrence Using Adjuvant Breast Cancer Cohort in Tertiary Cancer Center Registry. Front Oncol. 2021 May 4;11:596364.

23. Martinsson E. A model for sequential prediction of time-to-event in the case of discrete or continuous censored data, recurrent events or time-varying covariates. 2017.

24. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.

25. Hilt DE, Seegrist DW, States U, Northeastern Forest Experiment Station (Radnor P). Ridge, a computer program for calculating ridge regression estimates [Internet]. Upper Darby, Pa: Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station; 1977. 10 p.

26. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385–95.

27. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw [Internet]. 2011 [cited 2022 Feb 16];39(5). Available from: http://www.jstatsoft.org/v39/i05/

28. Kawaguchi ES, Suchard MA, Liu Z, Li G. A surrogate ℓ0 sparse Cox's regression with applications to sparse high-dimensional massive sample size time-to-event data. Stat Med. 2020;39(6):675–86.

29. Ozga AK, Kieser M, Rauch G. A systematic comparison of recurrent event models for application to composite endpoints. BMC Med Res Methodol. 2018;18(1):2.

30. Ullah S, Gabbett TJ, Finch CF. Statistical modelling for recurrent events: an application to sports injuries. Br J Sports Med. 2014;48(17):1287–93.

31. Charles-Nelson A, Katsahian S, Schramm C. How to analyze and interpret recurrent events data in the presence of a terminal event: An application on readmission after colorectal cancer surgery. Stat Med. 2019;sim.8168.

32. Tong X, Zhu L, Sun J. Variable selection for recurrent event data via nonconcave penalized estimating function. Lifetime Data Anal. 2009;15(2):197–215.

33. Chen X, Wang Q. Variable selection in the additive rate model for recurrent event data. Comput Stat Data Anal. 2013;57(1):491–503.

34. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49.

35. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. :81.

36. Wan F. Simulating survival data with predefined censoring rates for proportional hazards models: Simulating censored survival data. Stat Med. 2017;36(5):838–54.

37. Pénichoux J, Moreau T, Latouche A. Simulating recurrent events that mimic actual data: a review of the literature with emphasis on event-dependence. ArXiv150305798 Stat [Internet]. 2015; Available from: http://arxiv.org/abs/1503.05798

38. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. NPJ Digit Med. 2021;4(1):153.

Appendix 1 – PubMed search strategy

Table A. Search Strategies in PubMed Database

Concept	Research strategy keyword	Research	# Results
Survival analysis	"survival analysis"[MeSH Terms] OR "survival analysis"[Text Word] OR "time-to-event"[All Fields]	# 1	338,066
Recurrence	("recurren*"[All Fields] OR ("relapse"[All Fields] OR ("repeated"[All Fields] OR ("multiple"[All Fields] AND ("event"[All Fields]))))	# 2	930,483
High-dimension	"high dimension*"[All Fields]	#3	8,435
Machine learning	"machine learning"[MeSH Terms] OR "machine learning"[Text Word]	#4	69,520
Survival analysis for recurrent events	#1 and #2	#5	75,826
High-dimension or machine learning	#3 or #4	#6	76,554
Total	#5 and #6	#7	192

Appendix 2 – Data components for modeling recurrent events when using standard statistical approaches

Set of individuals at risk

Standard statistical models described do not encounter for individuals at risk in the same way. This induces prior data management for appropriate application.

- The set of individuals at risk for the kth event comprised individuals who were at risk for the event. Different definition existed for the set of individuals at risk, mainly based on baseline hazard function:
- The unrestricted set, in which each subject could be at risk for any event regardless of the number of events presented, at all-time intervals;
- The restricted set contained only the time intervals for the kth event of subjects who had already presented $k-1$ events;
- The semi-restricted set contained for the kth event the subjects who had $k-1$ or fewer events.

Timescales

Timescales also embody key components to address at the data management stage. Three common timescales are:

- Calendar time, in which the times denotes the time since randomization/beginning of the study until an event occurs;
- Gap time, or waiting scale, resets the time to zero when an event occurs, i.e., it corresponds to the time elapsed since the last event previously observed;
- Counting process is constructed as per calendar time, although it enables late inclusions and/or censoring.

Illustrations for timescales were provided in Figure A.

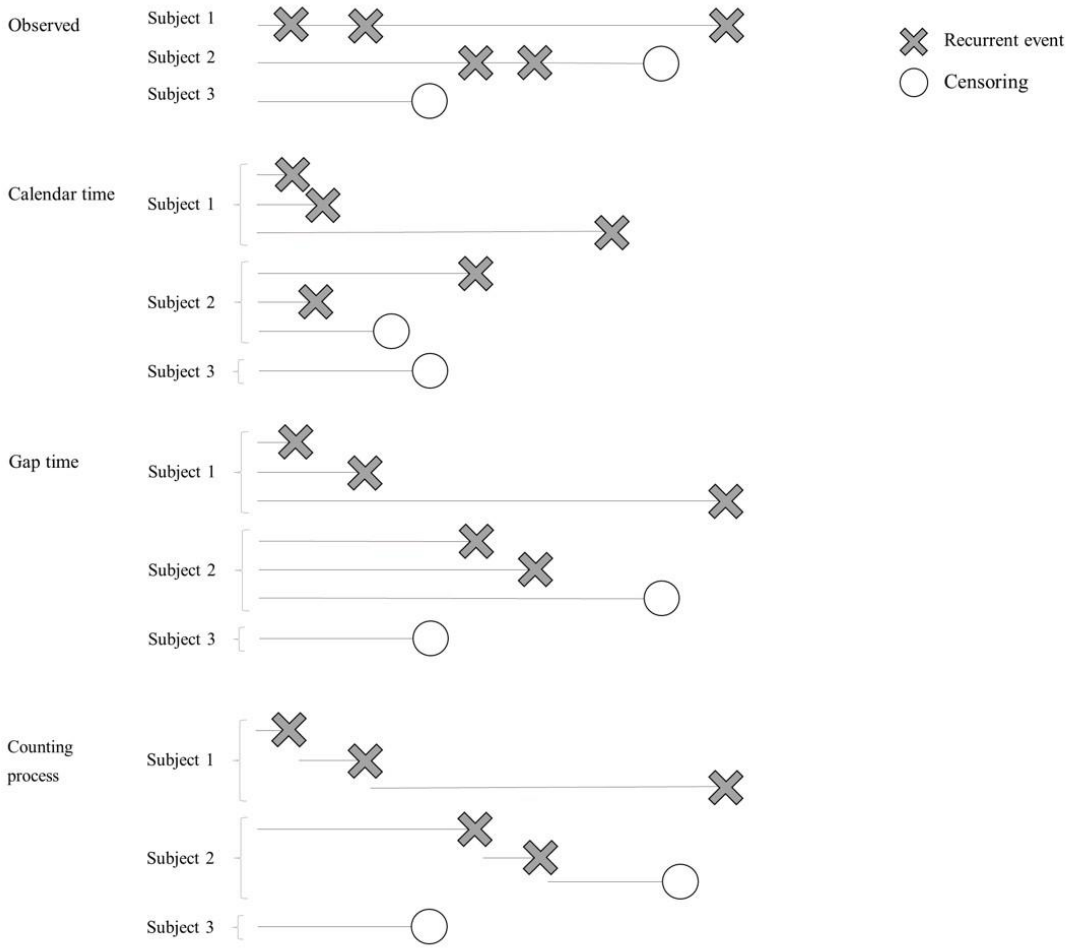

Figure A. Timescales in Recurrent Events Analysis

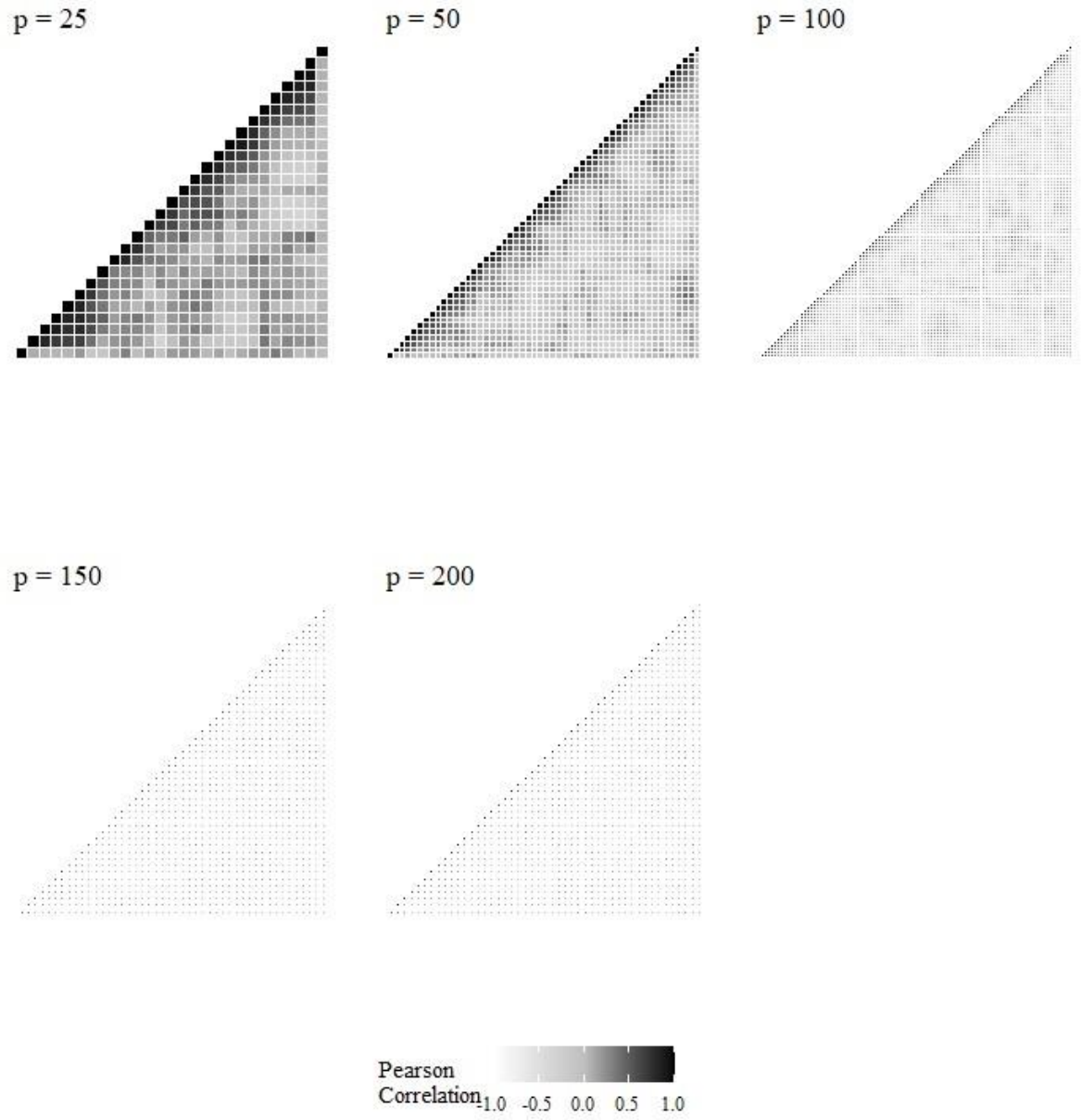

Figure B. Heatmaps of Correlation with Variations of the Number of Variables (25, 50, 100, 150, 200). Each square provided the Pearson correlation coefficient between the covariate on the x-axis and the one on the y-axis. All coefficients on the diagonal were equal to 1 as it was

the correlation coefficient between a covariate and itself.

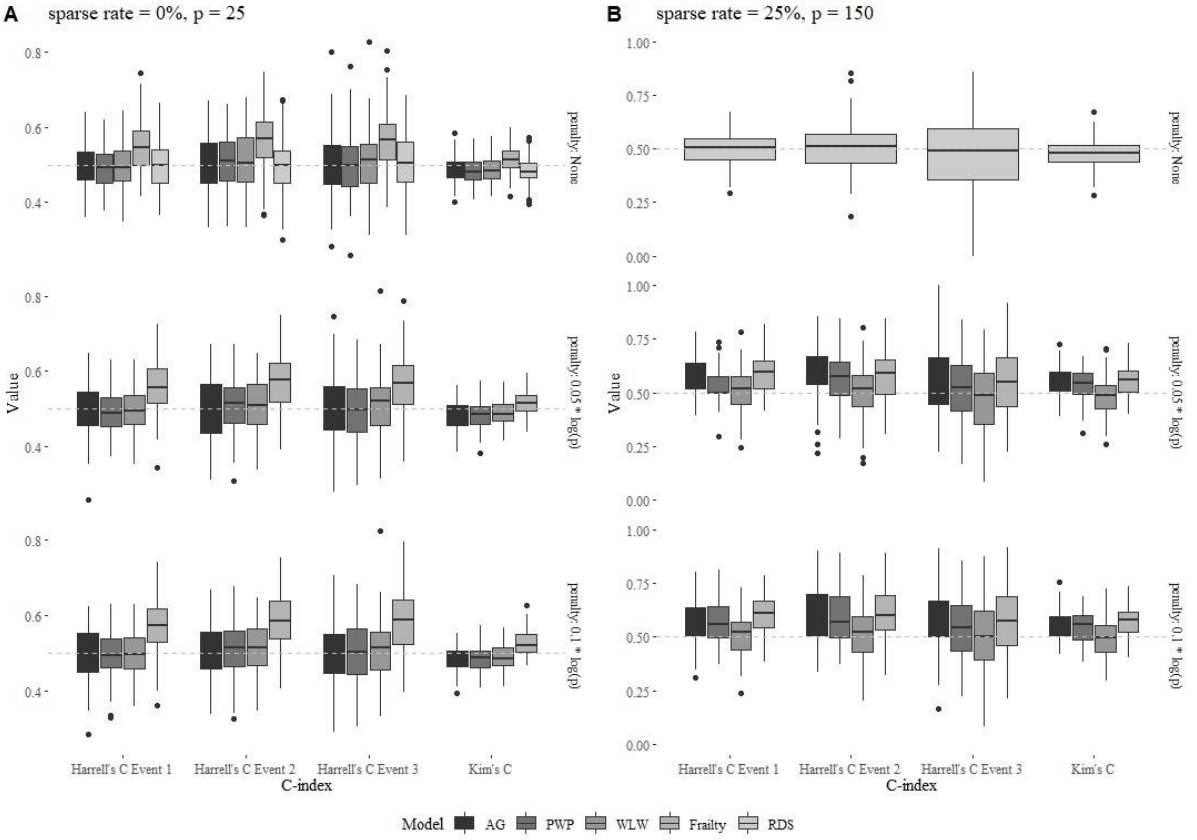

Figure C. Variability of C-indices for Two Extreme Scenarios: Sparse Rate = 0%, p = 25 (A) and Sparse Rate = 25%, p = 150 (B). p was the number of covariates. For each model and penalty, the C-indices of the 100 simulated datasets were summarized in a boxplot. The penalties were equal to 0 (unpenalized), $0.05 \times \log(p)$ and $0.1 \times \log(p)$, respectively. Penalties > 0 were applicable only for standard statistical models, RankDeepSurv deep neural network was hence not penalized. Unpenalized standard statistical models did not converge as soon as $p > n$, performance was therefore not available. C- AG: Andersen-Gill; PWP: Prentice, William, and Peterson; RDS: RankDeepSurv, WLW: Wei-Lin-Weissfeld

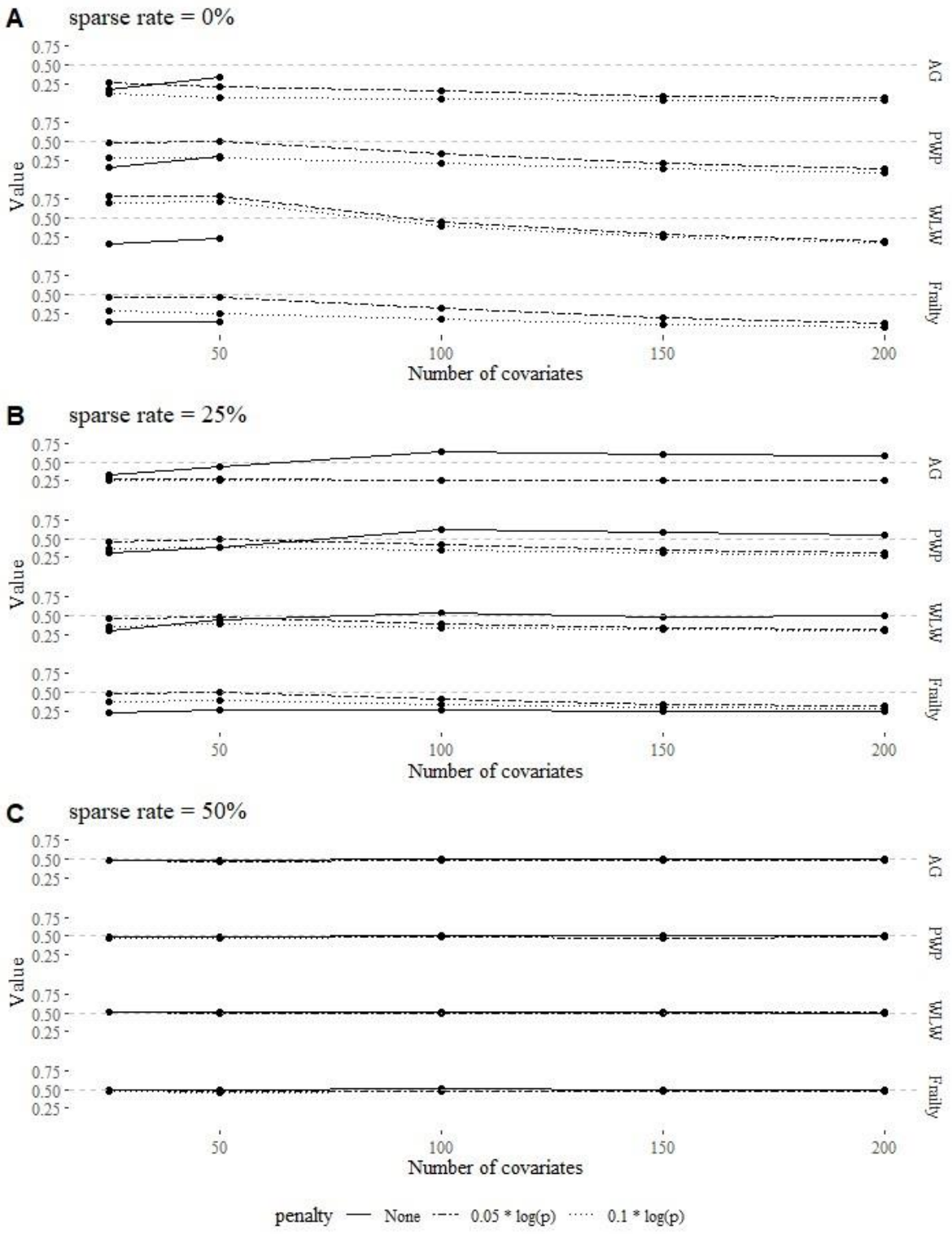

Figure D. Average Error Rates with Sparse Rate Equal to 0% (A), 25% (B) et 50% (C). p was the number of covariates. For each sparse rate, model and penalty, the average error rates of the 100 simulated datasets were displayed over the number of covariates. The penalties were equal to 0 (unpenalized), $0.05 \times \log(p)$ and $0.1 \times \log(p)$, respectively. Unpenalized standard statistical models did not converge as soon as $p > n$, error rate assessment was therefore not available. AG: Andersen-Gill; PWP: Prentice, William and Peterson; RDS: RankDeepSurv,; WLW: Wei-Lin-Weissfeld.