



Challenging estimation of tuning parameters in sPLSDR model to predict the incidence of rare diseases

Ian Meneghel Danilevicz ¹ EpiAgeing, Université Paris Cité, Inserm, Center for Research in Epidemiology and Statistics (CRESS)

Sam Vidil  EpiAgeing, Université Paris Cité, Inserm, Center for Research in Epidemiology and Statistics (CRESS)

Séverine Sabia  EpiAgeing, Université Paris Cité, Inserm, Center for Research in Epidemiology and Statistics (CRESS)

UCL Brain Sciences, Division of Psychiatry, University College London

Date published: 2025-10-07 Last modified: 2025-10-07

Abstract

We investigate some practical points of the sparse partial least squares deviance residuals (sPLSDR) model, a machine learning (ML) method with important application potential and a powerful alternative for Cox regression analysis when dimension reduction and variable selection are required. It can be used in a broad range of contexts, among which the health domain. This method requires to tune two hyper-parameters: eta (sparsity level) and D (number of components). Usually, the literature recommends cross-validation (CV) with a number of folds that depends on the sample size to estimate the hyperparameters. However, in case of dichotomous outcome for which the ratio of cases and non-cases is small this approach appears to be unstable. This is the case in rare disease such as incidence of major neurocognitive disorder (NCD). In that case, even large samples may perform as poorly as small samples. We specifically test this scenario of large samples with rare events using Monte Carlo simulations with different performance criteria (C-index and AIC). We observed that five folds are more conservative (biased) than one fold, and it presents more variability (less precise) in case of rare event outcomes. Previous authors suggested that Bootstrap or leave-one-out (LOO) could be a wise alternative. We checked Bootstrap with one fold in our real analysis application using the Whitehall II accelerometer sub-study (N total = 3,969, cases = 291 (7.3%)) as a training set and the UK Biobank as a validation set (N total = 54,833, cases = 750 (1.4%)), both considerably large sample size studies, with incidence of major neurocognitive disorders as an example of rare outcomes. This application shows that CV is sensitive to the number of folds and the initial seed. The counterpart Bootstrap was more robust and displayed two components with eta equal to 0.75 as a consistent optimum choice for the tuning parameters. However, this approach requires external validation, which is not always feasible. Further studies can better map the ideal number of folds for CV depending on the sample size and the ratio of cases/non-cases. As dichotomous outcomes and survival time are classical statistical subjects, the ML literature provide clear and precise recommendations for several available dawn methods.

Keywords: Machine learning, Survival analysis, Variable selection, Cross-validation, Bootstrap

1 Contents

2 1 Introduction

2

¹Corresponding author: ian.meneghel-danilevicz@inserm.fr

3	2 Method	3
4	2.1 The sPLS technique	3
5	2.2 The Cox proportional hazards model	4
6	2.3 Deviance residuals	4
7	2.4 The sPLSDR model	4
8	3 Monte Carlo study	5
9	3.1 Simulation plan	5
10	3.2 Simulation implementation	5
11	3.3 Simulation results	6
12	4 Real application	7
13	4.1 Real datasets	7
14	4.2 Real data results	8
15	5 Discussion	9
16	6 Conclusion	10
17	References	10
18	7 Declarations	12
19	7.1 Acknowledges:	12
20	7.2 Funding:	12
21	7.3 Data availability	12
22	7.4 Ethics approval and consent to participate:	12
23	7.5 Authors' contributions	12
24	7.6 Competing interests:	13
25	8 Finalize your submission	13
26	8.1 Handle R dependencies with <code>renv</code>	13
27	Session information	13

1 Introduction

Accelerometer devices objectively measure the wrist acceleration of human participants Migueles et al. (2019). The derived features calculated from the accelerometer three axial time series aggregated per minute for each participant during a certain period of time can be used to estimate rest-activity rhythm, daytime activity, sleep, and chronotype. These features can be used to predict several health outcomes such as mortality Chen et al. (2022), cardio-metabolic diseases Cassidy et al. (2018), and neurodegenerative diseases Nguyen et al. (2023). This represents a vast potential for medical research Chan et al. (2022) as these devices provide a more objective and detailed information than commonly used self-reported data.

In survival analysis, we model the duration from a time origin until an event of interest occurs, considering a set of predictors. In case of several candidate predictors, some selection and/or combination of predictors are often operated to allow for a parsimonious model, using machine learning (ML) methods such as the sparse partial least squares deviance residuals (sPLSDR) model Bastien et al. (2014).

ML methods often require tuning parameters to be fitted. These parameters are usually pre-estimated

using cross-validation (CV) techniques. The general recommendation consists of doing 3 to 10 folds (depending on the sample size) Lê Cao and Welham (2021), within a CV with all combinations of hyper-parameters. The general rule is $n/k > 5$, where n is the sample size and k the number of folds Lê Cao and Welham (2021)(page 82). The sPLSDR requires two hyper-parameters: First $\eta \in [0, 1]$, where 0 means a full dense model with all features, and values close to 1 a highly sparse model with only few features retained; second D components (each component is a combination of features as in PLS regression). One approach consists of determining a large maximum number of components D and a large grid for η , later trying a finer grid for η only on the best options of D Chen et al. (2022). To evaluate the hotspots which are considering the best hyperparameter options must be based on a criterion, eg, integrated AUC survival ROC (iAUC) Li (2006). Even if this approach is commonly used, there are situations when it lacks enough stability.

It happens that when the outcome of interest is rare, for example major neurocognitive disorder (NCD), the process of optimisation is unstable for different fold values, and even for the same fold value, it may vary according the initial seed, indicating a lack of reliability. This instability could be due to small sample size where 5 folds for CV are recommended, but we hypothesise that the low proportion of cases by non-cases could be problematic even for large and medium sample size studies.

This paper will discuss ways to overcome the challenge of tuning parameter estimation in cases of rare events. Our application is based on the identification of accelerometer data predicting major NCD, such as Alzheimer’s disease, using the Whitehall II ($n = 3969$) as a training set and the UK Biobank (UKB) as a validation set. Our proposition to get stable results for the tuning parameters is to keep all the data from the training set and bootstrap it, and then to test the validity of the model using in an external dataset (UKB). The optimization process for Whitehall II is unstable for different fold values, and even for the same fold value (with different initial seed), which indicates a lack of reliability. The instability is not due to a small sample size, as Whitehall II is a relatively large cohort. However, the outcome presents a small proportion of incident events (1.4%, the target disease is the major NCD, which includes Alzheimer’s disease). The problem was not fixed when considering the rules for small sample size (5 folds) which is far from the bottom level $n/k > 5$ suggest by Lê Cao and Welham (2021)(page 82).

A small simulation study was conducted to test the hypotheses of low stability in case of small ratio between cases and non-cases. We artificially created data sets with $n = 5000$ where the maximum observation time is 10 years of a disease for which incidence over time is small (resulting in few cases occurring during the observation period). In these simulations, we divided the data into five folds or keep the whole pool together, and we estimated the models using traditional performance criteria, such as the C-index and the Akaike information criterion (AIC).

2 Method

2.1 The sPLS technique

The sparse partial least squares regression (sPLS) was proposed by Chun and Keles (2010) to promote variable selection during PLS dimension reduction. The sPLS can select variables for the first direction vector and additional variables in the construction of the subsequent direction vectors Bastien et al. (2014).

A direct extension of PLS regression to sPLS regression could be provided by imposing L_1 constraint on PLS direction on n -vector w as

$$\max_w w^T M w \text{ subject to } w^T w = \|w\|_2 = 1, \|w\|_1 \leq c, \quad (1)$$

where $\|w\|_2 = w^T w$, $\|w\|_1 = \sum_{i=1}^n |w_i|$, $|u|$ is the absolute of u , $M = X^T Y Y^T X$, Y and X are matrices with the same number of rows, $c > 0$ is the L_1 constraint. If $Y = X$, then we have the special case sparse principal component analysis (sPCA) Jolliffe, Trendafilov, and Uddin (2003).

2.2 The Cox proportional hazards model

The model assumes the following hazard function for the occurrence of an event at time t in the presence of censoring as follows $h(t) = h_0(t) \exp(\beta^T X)$, where $h_0(t)$ is a baseline hazard function, β is a vector of regression coefficients and $X = (X_1, \dots, X_p)$ is a $n \times p$ matrix of features with $x_i = (x_{i1}, \dots, x_{ip})$ a row-vector for the i -th individual. The event could be any disease of interest. Based on the available data, Cox's partial likelihood can be written as

$$L(\beta) = \prod_{k \in \Omega} \frac{\exp(\beta^T x_k)}{\sum_{j \in \zeta_k} \exp(\beta^T x_j)},$$

where Ω is the set of indices of the events and ζ_k denotes the set of indices of the individuals at risk at time t_k . The goal is to find the coefficients which maximize the log partial likelihood function Bastien et al. (2014).

2.3 Deviance residuals

Here, we consider a Cox model with no time-dependent explanatory variables and, at most, one event per patient. Regardless, the Martingale residuals for the i -th subject with observation time t_i and event status δ_i , where $\delta_i = 0$ if t_i is a censored time, and $\delta_i = 1$ otherwise is as

$$m_i = \delta_i - \hat{\Delta}_0(t_i) \exp(\hat{\beta}^T x_i),$$

where $\hat{\Delta}_0(t_i)$ is the estimated cumulative hazard function at time t_i Bastien et al. (2014). However, the Martingale residuals used to be highly skewed Bastien et al. (2014). Although, the deviance residuals r_i are a normalized transform of the martingale residuals. For the Cox model, the deviance residuals Collett (1994) amount to a form as

$$r_i = \text{sign}(m_i) [2\{-r_i - \delta_i \log(\delta_i - m_i)\}]^{0.5}, \quad (2)$$

where $\text{sign}(u)$ is the sign of $u \in \mathbb{R}$.

2.4 The sPLSDR model

The sparse partial least squares deviance residual (sPLSDR) model was proposed by Bastien et al. (2014) as an upgrade concerning sPLS proposed by Chun and Keles (2010) and partial least squares Cox (PLS-Cox) Bastien and Tenenhaus (2001), Bastien (2008). The general idea is to apply the sPLS technique presented in Equation 1 on the deviance residual from Equation 2.

The algorithm was described in Bastien et al. (2014) as follows

1. Cox model without covariates to derive the null deviance residuals $r = (r_1, \dots, r_n)^T$ from Equation 2.
2. Computation of the sPLS components by using the sPLS regression from Equation 1 with the null deviance residuals as outcome.
3. Set $\hat{\beta}^{PLS} = 0$, $d = 1$, $Y = r$.
4. While $d \leq D$: 4.1. $w = (|z| - c/2) + \text{sign}(z)$, where $z = X^T Y / \|X^T Y\|_2$. 4.2. Update Ω as $\{d : \hat{w}_d \neq 0\} \cup \{d : \hat{\beta}_d^{PLS} \neq 0\}$. 4.3. Fit sPLS with X by using the d number of latent components. 4.4. Update $\hat{\beta}^{PLS}$ by using the new sPLS estimates of the direction vectors and update Y and d through $Y \leftarrow Y - X \hat{\beta}^{PLS}$ and $d \leftarrow d + 1$.
5. Cox model on the D retained sPLSDR components.

3 Monte Carlo study

3.1 Simulation plan

We used an exponential survival model, where the hazard function is

$$h_i(t) = \lambda \exp(X_i \beta), \quad (3)$$

and the cumulative hazard is

$$H_i(t) = \lambda t \exp(X_i \beta),$$

such as implemented on the **simsurv** R Package by Brilleman et al. (2021). We create artificial data with $n = 5000$ number of individuals, $\lambda = 0.05, 0.005$ rate parameter, $\beta^T = (\beta_1^T, \beta_2^T, \beta_3^T)$ is the coefficient parameters, composed by $\beta_1^T = (0.15, 0.1, 0.05)$, $\beta_2^T = (-0.15, -0.1, -0.05)$ and $\beta_3^T = (0, 0, 0, 0, 0)$; the covariate matrix is $X_{i,(1:3)} \sim N(\mu_1, R)$ a multivariate normal with $\mu_1 = (0.15, 0.1, 0.05)^T$ and R a matrix with variance on the diagonal of ones and covariance on the off-diagonals of 0.5; $X_{i,(4:6)} \sim N(\mu_2, R)$ a multivariate normal with $\mu_2 = (-0.15, -0.1, -0.05)^T$ and same covariance; $X_{i,(7:12)} \sim N(\mu_3, I)$ a multivariate normal with $\mu_3 = (0, 0, 0, 0, 0)^T$ and I an identity matrix. The maximum observed time is $m = 10$.

So this simple model may be well estimated by a sPLSDR model with two dimensions. We proceed with a Monte Carlo simulation, replicating $M = 1000$ times each λ and estimating the model changing $\eta = 0, 0.1, 0.2, \dots, 0.9$, using $k = 1, 5$ folds. We save the C-index and the AIC of each iteration of the Monte Carlo. We set all models with two dimensions ($D = 2$). Finally, we count how many times each value of η was considered the best in the row; this results in a frequency per value of η .

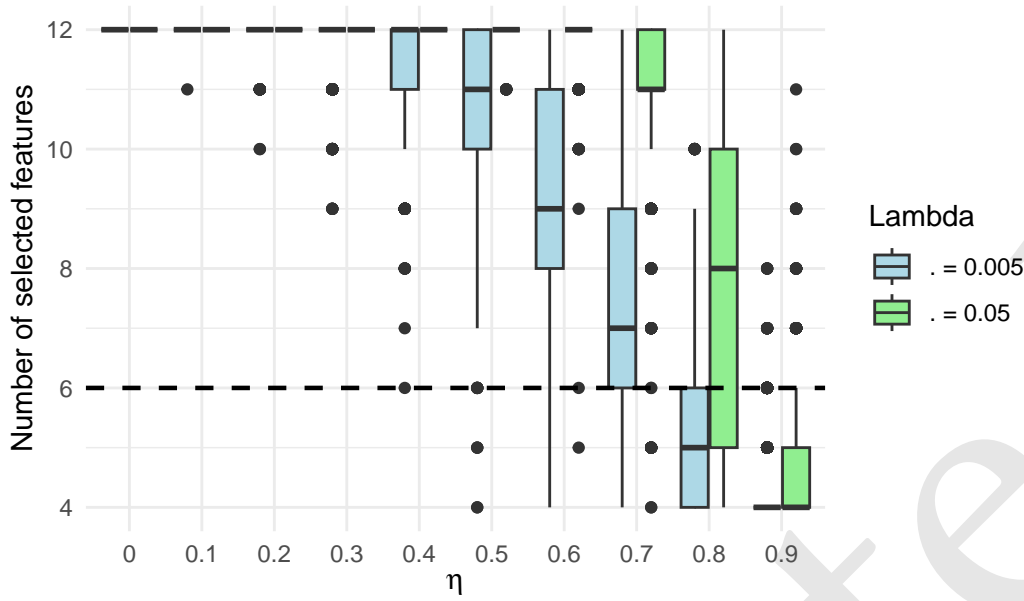
3.2 Simulation implementation

Before the results, we implemented some functions available in the R folder. Also, some packages are necessary, such as **pplsRcox** (the original package to estimate sPLSDR Bastien et al. (2014)), **Coxmos** (another package to estimate sPLSDR but with a convenient way to extract the W^* matrix of weights, this matrix is similar to a loading matrix in PCA Salguero García (2025)), **caret** (to split the dataset in folds Kuhn and Max (2008)).

Then we run the simulations using our loaded functions.

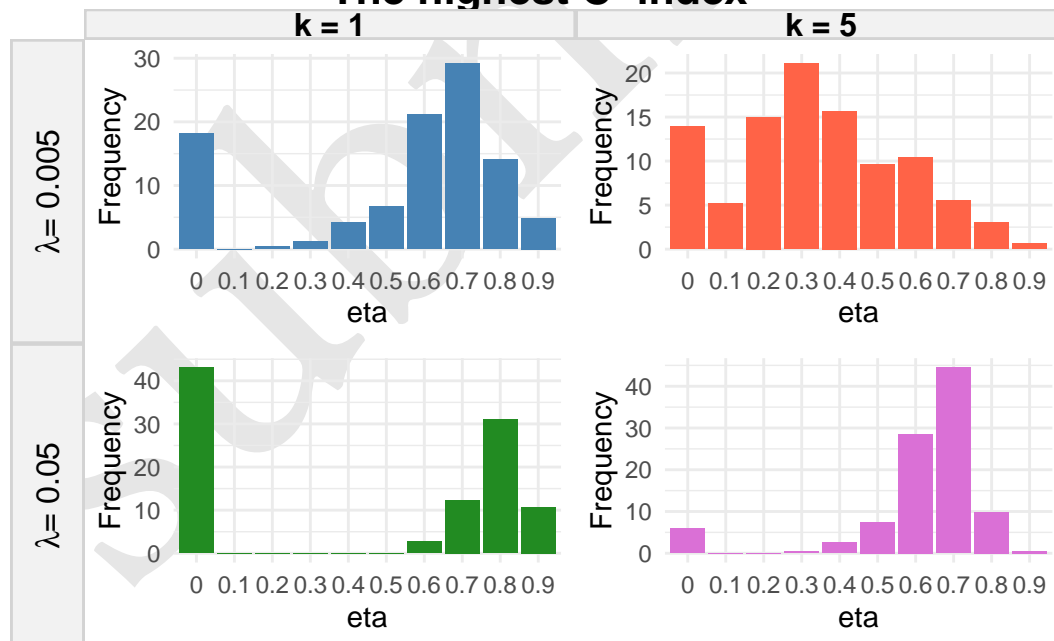
3.3 Simulation results

Selected features by eta and lambda



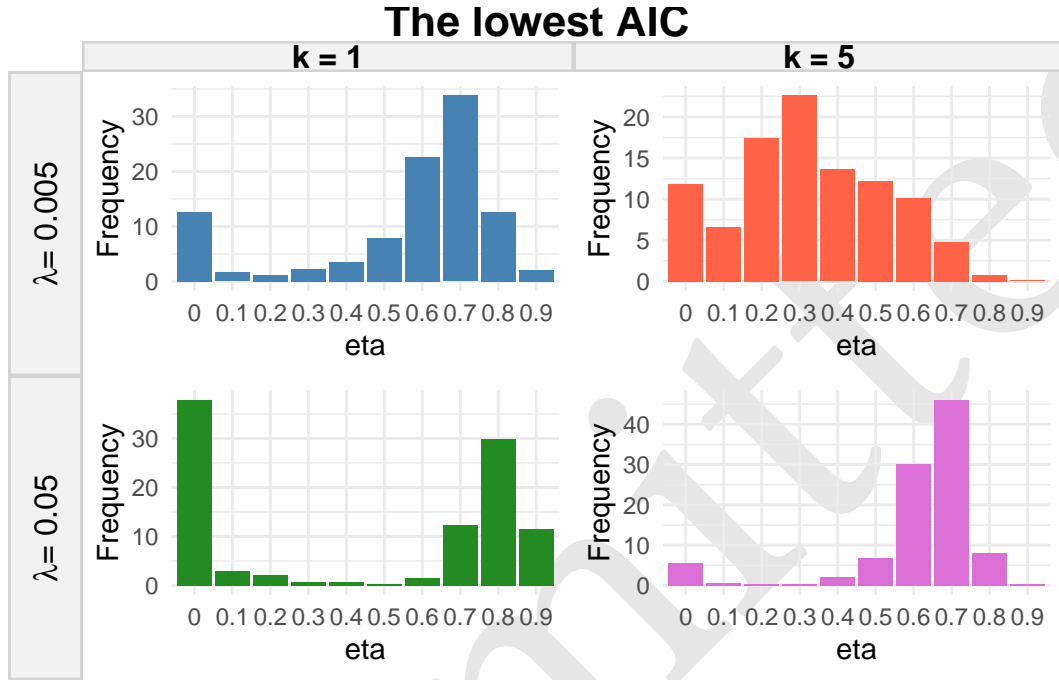
The above table shows the results for the number of selected features depending on the value of η . The number of selected features is decreasing as eta increases. Remember that we set a simulation within the coefficient parameter vector that has a length of 12, but 6 values equal to zero and six different from zero. So, a fair number of selected features must be equal or smaller than six, since there is a moderate correlation between the non-zero coefficients. Finally, the value of eta, which corresponds to a fair number of features, is $\eta \in [0.8, 0.9]$ for $\lambda = 0.005$ or $\eta \in [0.9]$ for $\lambda = 0.05$. These regions for eta are the targets that we will search for with C-index and AIC in the following figures.

The highest C-index



In the above figure, we see the results for the C-index. The figure shows the frequency decision taken by the highest C-index, ie at each simulation, we count the η , which corresponds to the highest C-index. On top, $\lambda = 0.005$, which corresponds to a scenario where there are few incident cases (eg,

rare disease), and on the bottom, $\lambda = 0.005$ is a scenario with several incident cases (eg, common disease). On the left, the estimation is done by a $k = 1$ fold, and on the right, a $k = 5$ folds does it. For $\lambda = 0.05$ the most frequent decision is $\eta = 0.7$ when using $k = 5$ which is in line with the general recommendation and with the complexity of the model, but using $k = 1$ the most frequent decision is $\eta = 0$ (too conservative decision). However, when $\lambda = 0.005$ the decision is the reverse: using $k = 1$ returns $\eta = 0.7$ as the most frequent decision and $k = 5$ returns $\eta = 0.3$ as the most frequent decision (which is conservative). Furthermore, we can see that the plot at the top right ($k = 5$) presents larger frequency variability than $k = 1$. Remember that the target regions are $\eta \in [0.8, 0.9]$ for $\lambda = 0.005$ or $\eta \in [0.9]$ for $\lambda = 0.05$. Therefore, the frequency distribution is more concentrated in the target region if $k = 1$ for $\lambda = 0.005$, and $k = 5$ for $\lambda = 0.05$.



In the above figure, we see the results for the AIC. These findings for AIC are very similar to those observed using C-index as the performance criteria. This reinforces that the findings are not due to the criteria chosen but reflect the need to select the number of folds k adequate to the problem based on sample size, ratio of events, number of dimensions, number of non-zero features, correlation between the features.

4 Real application

4.1 Real datasets

The Whitehall II study is an ongoing prospective cohort study established in 1985-1988 among 10,308 British civil servants aged 35 to 55 years with clinical examinations every four-five years since inception Marmot and Brunner (2005). An accelerometer measure was added to the 2012-2013 wave of data collection, where 4,880 members from the cohort were invited to take part in the sub-study. Participants were followed for incident NCD up to 2023.

The UK Biobank (UKB) is a prospective population-based cohort study on over 500,000 men and women aged between 40 and 69 years established in 2006-2010 Littlejohns et al. (2020). The accelerometer sub-study was proposed to 236,519 individuals between 2013-2015 Doherty et al. (2017). Participants were followed for incident NCD up to 2022.

The accelerometer signal from each participant in both cohorts were extracted by **GGIR** R package Migueles et al. (2019). A brief introduction to the 40 accelerometer derived features can be found in Chen et al. (2022) and Danilevicz, van Hees, et al. (2024). They represent features related to rest-activity rhythm, daytime activity, sleep, and chronotype.

When there are two datasets, usually the bigger is used as a training set and the smaller as a validation set Lê Cao and Welham (2021, 81). However, we intend to test our hypotheses in separate subpopulations, so we choose Whitehall II ($n = 3969$; 291 incident cases (7.3%)) as a training set in order to validate findings in UKB ($n = 54833$; 750 incident cases (1.4%)) and subgroups of this population. This should not be a problematic because Whitehall II is not a small cohort. Participants were included in the analyses if they had data from at least five valid full day windows (defined as wear times $\geq 2/3$ of both waking and following sleep periods) Danilevicz, Vidil, et al. (2024).

4.2 Real data results

Our first model fit is similar to the simulated one in Equation 3, but now it has a real interpretation. It is as follows

$$h_i(t) = \lambda \exp(Z_i^{(D,\eta)} \beta),$$

where $h_i(t)$ is the time for the event for the i -th participant, $Z_i^{(D,\eta)}$ is a matrix of reduced accelerometer-derived data which depend of our choice of tuning parameters D and η , β is a D -vector of component coefficients.

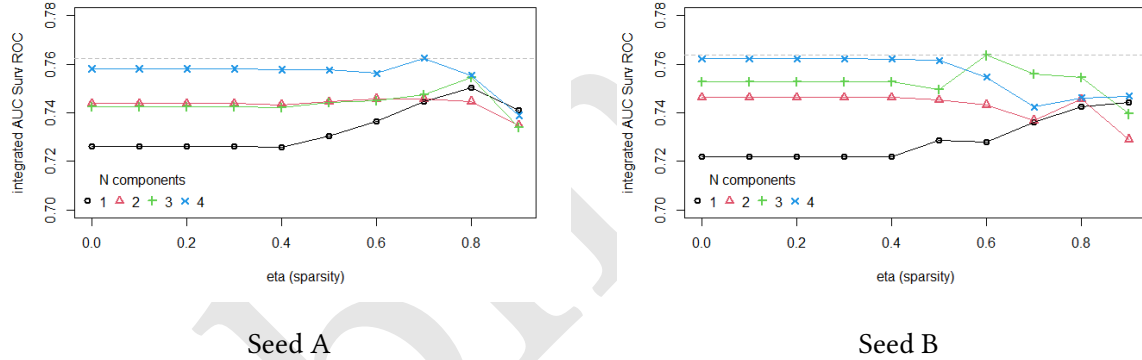


Figure 1: C-index plots for $\eta = 0, 0.05, \dots, 0.95$ under different seeds, with $k = 5$ (left: seed A, right: seed B).

The Figure 1 shows the iAUC for $\eta = 0, 0.1, \dots, 0.9$ for $D = 1, \dots, 4$ (components) from 5-fold CV. At the left, the maximum is achieved by $D = 4$ components at $\eta = 0.7$ and at the right by $D = 3$ at $\eta = 0.6$. The only difference is the seed to sampling the participants in each fold. This unstable and conflicting results were observed irrespective of the number of folds used for CV, ie, for $k = 3, \dots, 10$ folds.

We move from CV to bootstrap, with the results in the Figure 2 (left). First, we fit the same combinations for CV, ie, $D = 1, \dots, 4$ and $\eta = 0, 0.1, \dots, 0.9$. We can see on the left that the results are much more linear, and the number of components is monotonous in terms of four always being higher than three, three always being higher than two, and so on for the respective C-index.

Later, we fit a model that combines the previous model with known risk factors for the event, which is as follow

$$h_i(t) = \lambda \exp(Z_i^{(D,\eta)} \beta + X_i \alpha),$$

where X_i is the row-vector of known risk factors (sex, age, ethnicity, education, working status, cohabitation status, smoking, alcohol consumption, daily consumption of fruits and vegetables, body mass index (BMI), hypertension, diabetes, hyperlipidemia, multiple comorbidities, central nervous system (CNS) medication) for the i -th participant, α is a 15-vector of the risk factor coefficient. Then in order to identify the best tuning parameters that allow the selected model to improve the predictive value for NCD as compared to known risk factors, we compared the C-index of the model with only the known risk factors (dotted line in the Figure 2 (right)) with the C-index of models including these risk factors and the fitted accelerometer data for $\eta = 0, 0.05, \dots, 0.95$ for $D = 1, \dots, 4$. The C-index mean (points) and CI (bars) were estimated using bootstrap with 1 to 4 components derived from the accelerometer features and all the previously described known risk factors. The C-index for the covariates only was around 0.78. We found that two components are significantly higher than known risk factors only for values of η in the range $[0, 0.75]$. So the retained model was the one for $D = 2$ and $\eta = 0.75$, allowing to improve C-index while being as sparse as possible.

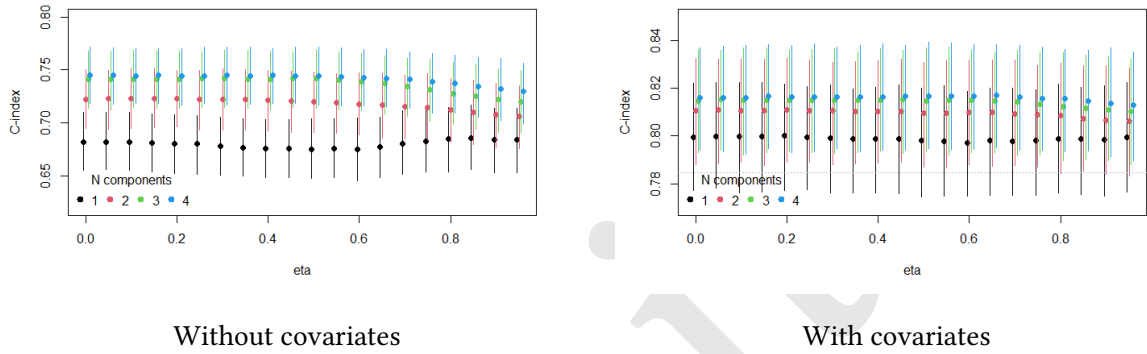


Figure 2: C-index (mean and CI) for $\eta = 0, 0.05, \dots, 0.95$ for $D = 1, \dots, 4$ components, $k = 1$, without (left) and with (right) covariates using Bootstrap with 2000 repetitions (dotted line represents the C-index of sociodemographic, behavioural and health-related covariates only).

We created a score with $D = 2$ and $\eta = 0.75$ for the UKB using the same features and weight matrix W^* identified in the Whitehall II cohort. This score significantly improved the predictive value of a model including known risk factors in the external cohort.

Note: Due to the inclusion of sensitive patient information, the datasets used in this section are not publicly available in order to comply with ethical and data privacy regulations. However, access to the data may be granted upon reasonable request to the respective institutions responsible for their maintenance, as referenced in the data availability disclaimer.

5 Discussion

To evaluate the stability of hyperparameter selection in the context of rare events, we conducted a Monte Carlo simulation comparing one-fold and five-fold cross-validation. The results show that using one fold is more stable than five folds when cases are rare. Specifically, $\lambda = 0.005$ associated with $k = 5$ display more conservative decisions ($\eta < 0.5$) and larger variability than $\lambda = 0.005$ associated with $k = 1$. In case of rare events, identifying a stable decision of hyper-parameters may be a challenge. Our findings suggest that it may be more reliable to use one fold, opening to two issues: 1) if there is no repetition, we cannot estimate a confidence interval (CI); 2) we cannot validate our findings as there is no remaining data to test. This raises the importance of having an independent datasets as validation sample. The necessity of external validation is a limitation of this approach, as it is not always available.

Our real data application shows that CV is unstable for $k = 5$ in a large sample size as Whitehall II ($n = 3969$). Changing only the initial seed, CV for $k = 5$ returns contradictory results such as different optimum values for η and D . The bootstrap result shows a very stable shape for η . The conclusion was that $D = 2$ with $\eta = 0.75$ significantly improved (higher C-index) compared with a model including known risk factors only.

Applied works commonly used 5 folds Chen et al. (2022) or 10 folds Kohavi (1995), Lê Cao, Boitard, and Besse (2011). A few studies give general advice for CV for similar machine learning techniques; they can recommend between 2 and 10 folds Breiman and Spector (1992) or 3 and 10 folds Lê Cao and Welham (2021) [page 82] depending on the sample size. However, they do not consider the ratio of cases by non-cases in dichotomous outcomes. Lê Cao and Welham suggest leave-one-out (LOO) for tiny samples and that Bootstrap might be an interesting new approach Lê Cao and Welham (2021, 82). Our simulation and real application agree that $k = 1$ is a promising approach for specific situations with few positive outcomes (eg, rare diseases).

Further studies can be developed in several directions. First, this was not an exhaustive simulation study and other parameter may be chosen such as n , λ , D , β , R . Second, other researchers can develop their own simulation to answer their specific goals in terms of sample size, the ratio of events, the number of dimensions, the number of non-zero features, and the correlation between the features (which is simple as this paper displays all the codes for reproducibility). Third, instead of bootstrapping, an alternative and interesting approach is to use U-statistics. Kang et al. (2015) proved that U-statistics can be used to achieve asymptotic distribution of the C-index (confidence intervals may be easily estimated), which saves a lot of time and financial resources if we do cloud processing data.

6 Conclusion

The general advice for CV proceedings takes into account the sample size. However, from our study, the number of cases (in dichotomous outcomes) also plays an important role and should not be neglected. Furthermore, dividing the data by folds to CV does not seem a stable strategy if the ratio of cases/noncases is too extreme (which is the case for rare diseases). In fact, a bootstrap on the entire sample appears to be a good solution and we hypothesize that U-statistics could also be useful to derive C-index confidence intervals. However, these approaches require external validation, which is not always available. Further studies can confirm other scenarios where this situation occurs.

References

- Bastien, P. 2008. "Deviance residuals based PLS regression for censored data in high dimensional setting." *Chemometr Intell Lab Syst* 91 (1): 78–86. <https://doi.org/10.1016/j.chemolab.2007.09.009>.
- Bastien, P., F. Bertrand, N. Meyer, and Myriam Maumy-Bertrand. 2014. "Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data." *Bioinformatics* 31 (3): 397–404. <https://doi.org/10.1093/bioinformatics/btu660>.
- Bastien, P., and M. Tenenhaus. 2001. "PLS generalized linear regression. Application to the analysis of life time data." In *Proceedings of the PLS'01 International Symposium Anacapri*, 131–40. Italy.
- Breiman, L., and P. Spector. 1992. "Submodel Selection and Evaluation in Regression." *Int Stat Rev* 60 (3): 291–319. <https://doi.org/10.2307/1403680>.
- Brilleman, S. L., R. Wolfe, M. Moreno-Betancur, and Michael J. Crowther. 2021. "Simulating Survival Data Using the Simsurv r Package." *J Stat Soft* 97 (3): 1–27. <https://doi.org/10.18637/jss.v097.i03>.
- Cassidy, S., H. Fuller, J. Chau, Michael Catt, Adrian Bauman, and Michael I Trenell. 2018. "Accelerometer-Derived Physical Activity in Those with Cardio-Metabolic Disease Compared to

- Healthy Adults.” *Acta Diabetol.* 55 (9): 975–79. <https://doi.org/10.1007/s00592-018-1161-8>.
- Chan, A., D. Chan, H. Lee, C. C. Ng, and Yeo A. H. L. 2022. “Reporting adherence, validity and physical activity measures of wearable activity trackers in medical research: A systematic review.” *Int J Med Inform.* 160: 104696. <https://doi.org/10.1016/j.ijmedinf.2022.104696>.
- Chen, M., B. Landré, P. Marques-Vidal, Vincent T van Hees, April C E van Gennip, Mikaela Bloomberg, Manasa S Yerramalla, Mohamed Amine Benadjaoud, and Séverine Sabia. 2022. “Identification of physical activity and sedentary behaviour dimensions that predict mortality risk in older adults.” *EClinicalMedicine* 55: 101773. <https://doi.org/10.1016/j.eclinm.2022.101773>.
- Chun, H., and S. Keles. 2010. “Sparse partial least squares regression for simultaneous dimension reduction and variable selection.” *J. R. Stat. Soc. B* 72: 3–25. <https://doi.org/10.1111/j.1467-9868.2009.00723.x>.
- Collett, D. 1994. *Modelling Survival Data in Medical Research*. Chapman; Hall/CRC. <https://doi.org/10.1201/b18041>.
- Danilevicz, I. M., V. T. van Hees, Frank C. T. van der Heide, Louis Jacob, Benjamin Landré, Mohamed Amine Benadjaoud, and Séverine Sabia. 2024. “Measures of fragmentation of rest activity patterns: mathematical properties and interpretability based on accelerometer real life data.” *BMC Med. Res. Methodol* 24 (1, 132). <https://doi.org/10.1186/s12874-024-02255-w>.
- Danilevicz, I. M., Sam Vidil, Benjamin Landré, Aline Dugravot, Vincent Theodor van Hees, and Séverine Sabia. 2024. “Reliable measures of rest-activity rhythm fragmentation: how many days are needed?” *Eur Rev Aging Phys Act* 21 (29). <https://doi.org/10.1186/s11556-024-00364-5>.
- Doherty, A., D. Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H. Granat, Tom White, et al. 2017. “Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers.” *PloS One* 12 (2): e0169649. <https://doi.org/10.1371/journal.pone.0169649>.
- Jolliffe, I, N. T. Trendafilov, and M. Uddin. 2003. “A modified principal component technique based on the LASSO.” *J. Comput. Graph. Stat.* 12: 531–47. <https://doi.org/10.1198/1061860032148>.
- Kang, L., W. Chen, Nicholas A Petrick, and Brandon D Gallas. 2015. “Comparing two correlated C indices with right-censored survival outcome.” *Stat Med.* 34 (4): 685–703. <https://doi.org/10.1002/sim.6370>.
- Kohavi, R. 1995. “A study of cross-validation and bootstrap for accuracy estimation and model selection.” In *14 Int. Joint Conf. AI*, 1137–43. Montreal. <https://doi.org/10.5555/1643031.1643047>.
- Kuhn, and Max. 2008. “Building Predictive Models in r Using the Caret Package.” *Journal of Statistical Software* 28 (5): 1–26. <https://doi.org/10.18637/jss.v028.i05>.
- Lê Cao, K.-A., S. Boitard, and P. Besse. 2011. “Sparse PLS discriminant analysis.” *BMC Bioinformatics* 12 (253). <https://doi.org/10.1186/1471-2105-12-253>.
- Lê Cao, K.-A., and Z. M. Welham. 2021. *Multivariate Data Integration Using R*. Chapman; Hall/CRC. <https://doi.org/10.1201/9781003026860>.
- Li, L. 2006. “Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information.” *Bioinformatics* 22 (4): 466–71. <https://doi.org/10.1093/bioinformatics/bti824>.
- Littlejohns, T. J., J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, and et al. 2020. “The UK Biobank imaging enhancement of 100,000 participants.” *Nat Commun.* 11 (1): 2624. <https://doi.org/10.1038/s41467-020-15948-9>.
- Marmot, M., and E. Brunner. 2005. “Cohort profile: the Whitehall II study.” *Int J Epidemiol.* 34 (2): 251–56. <https://doi.org/10.1093/ije/dyh372>.
- Miguelles, J. H., A. V. Rowlands, F. Huber, Séverine Sabia, and Vincent T. van Hees. 2019. “GGIR: a research community-driven open source R package for generating physical activity and sleep outcomes from multi-day raw accelerometer data.” *J. Meas. Phys. Behav* 2 (3): 188–96. <https://doi.org/10.1123/jmpb.2018-0063>.
- Nguyen, S., A. Z. LaCroix, K. M. Hayden, C. Di, P. Palta, M. L. Stefanick, J. E. Manson, S. R. Rapp, M. J. LaMonte, and J. & Bellettiere. 2023. “Accelerometer-measured physical activity and sitting

with incident mild cognitive impairment or probable dementia among older women.” *Alzheimer Dement.* 19 (7): 3041–54. <https://doi.org/10.1002/alz.12908>.
Salguero García, Pedro. 2025. *Coxmos: Cox MultiBlock Survival*. <https://github.com/BiostatOmics/Coxmos>.

7 Declarations

7.1 Acknowledges:

We thank all of the participating civil service departments and their welfare, personnel, and establishment officers; the British Occupational Health and Safety Agency; the British Council of Civil Service Unions; all participating civil servants in the Whitehall II study; and all members of the Whitehall II study team. The Whitehall II Study team comprises research scientists, statisticians, study coordinators, nurses, data managers, administrative assistants and data entry staff, who make the study possible.

7.2 Funding:

The Whitehall II study is and was supported by grants from the National Institute on Aging, NIH (R01AG056477, R01AG062553); UK Medical Research Council (R024227, S011676), and the Wellcome Trust (221854/Z/20/Z). SS is funded by the European Union (ERC grant number 101043884).

The funding agencies had no role in the study design, data collection, analyses, and interpretation of the data or writing of the manuscript. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the funding agencies. Neither the European Union nor the granting authority can be held responsible for them.

7.3 Data availability

Data cannot be made publicly available because of ethics and Institutional Review Board (IRB) restrictions. However, a data sharing portal allows access to data to undertake analyses within the secure portal in [Whitehall II study](#). The UK Biobank data are available through a procedure described at [UK Biobank data](#).

7.4 Ethics approval and consent to participate:

Concerning the Whitehall II data, written informed consent for participation was obtained at each contact. Research ethics approval was obtained from the University College London ethics committee (latest reference number 85/0938).

Concerning the UK Biobank data, approval was received from the National Information Governance Board for Health and Social Care and the National Health Service North West Centre for Research Ethics Committee (reference number 11/NW/0382). All participants gave written consent for participation. Our study was conducted using the UKB Resource under application number 96856.

7.5 Authors' contributions

IMD and SS developed the hypothesis and study design. IMD conducted data analysis. IMD wrote the first and successive drafts of the manuscript. IMD, SV and SS were involved in the conception, and design of the study, analysis and interpretation of the data, and drafting or critically revising the manuscript for important intellectual content, or, in addition, acquired data. All authors have read and approved the final manuscript.

7.6 Competing interests:

The authors declare that they have no competing interests.

8 Finalize your submission

8.1 Handle R dependencies with renv

To make your work reproducible, you need to fix the packages and environment used to run your analysis. For the R system, the `renv` package is one of the possible reliable method, supported by the community. You basically need a couple of commands to setup your environment on your local machine. First,

will initialize your repository. Then you just need to install the dependencies required to run your contribution, for instance,

Non-CRAN packages (e.g. Github packages) can be used. Once you are done, you can fix everything with the command

! Important

The only file that needs to be versioned by git is `renv.lock`. By default, the rest is ignored thanks to `.gitignore`.

Session information

R version 4.5.1 (2025-06-13 ucrt)

Platform: x86_64-w64-mingw32/x64

Running under: Windows 10 x64 (build 19045)

Matrix products: default

LAPACK version 3.12.1

locale:

[1] LC_COLLATE=French_France.utf8 LC_CTYPE=French_France.utf8

[3] LC_MONETARY=French_France.utf8 LC_NUMERIC=C

[5] LC_TIME=French_France.utf8

time zone: Europe/Paris

tzcode source: internal

attached base packages:

[1] grid stats graphics grDevices datasets utils methods

[8] base

other attached packages:

[1] cowplot_1.2.0 patchwork_1.3.2 tidyr_1.3.1 dplyr_1.1.4

[5] caret_7.0-1 simsurv_1.0.0 Coxmos_1.1.5 plsRcox_1.8.0

[9] mixOmics_6.32.0 ggplot2_3.5.2 lattice_0.22-7 MASS_7.3-65

[13] survcomp_1.58.0 prodlim_2025.04.28 survival_3.8-3

```

418 loaded via a namespace (and not attached):
419   [1] splines_4.5.1          polspline_1.1.25       fields_17.1
420   [4] tibble_3.3.0           hardhat_1.4.2          pROC_1.19.0.1
421   [7] rpart_4.1.24           lifecycle_1.0.4        Rdpack_2.6.4
422  [10] rstatix_0.7.2          globals_0.18.0         backports_1.5.0
423  [13] magrittr_2.0.3         Hmisc_5.2-4            rmarkdown_2.29
424  [16] spam_2.11-1            RColorBrewer_1.1-3     lubridate_1.9.4
425  [19] multcomp_1.4-28        maps_3.4.3             abind_1.4-8
426  [22] purrr_1.0.4            rmeta_3.0              nnet_7.3-20
427  [25] TH.data_1.1-4          sandwich_3.1-1         ipred_0.9-15
428  [28] lava_1.8.1             KMSurv_0.1-6           survivalROC_1.0.3.1
429  [31] ggrepel_0.9.6          listenv_0.9.1          ellipse_0.5.0
430  [34] vegan_2.7-1            MatrixModels_0.5-4     RSpectra_0.16-2
431  [37] parallelly_1.45.1      svglite_2.2.1          permute_0.9-8
432  [40] codetools_0.2-20       tidyselect_1.2.1       SuppDists_1.1-9.9
433  [43] shape_1.4.6.1          farver_2.1.2           matrixStats_1.5.0
434  [46] stats4_4.5.1           base64enc_0.1-3        risksetROC_1.0.4.1
435  [49] jsonlite_2.0.0         Formula_1.2-5          iterators_1.0.14
436  [52] systemfonts_1.3.1      foreach_1.5.2          tools_4.5.1
437  [55] progress_1.2.3         sna_2.8                Rcpp_1.0.14
438  [58] glue_1.8.0             rARPACK_0.11-0         gridExtra_2.3
439  [61] xfun_0.52              mgcv_1.9-3             withr_3.0.2
440  [64] fastmap_1.2.0          boot_1.3-31            SparseM_1.84-2
441  [67] digest_0.6.37          timechange_0.3.0       R6_2.6.1
442  [70] textshaping_1.0.3      colorspace_2.1-2       scattermore_1.2
443  [73] generics_0.1.4         pls_2.8-5              renv_1.1.5
444  [76] data.table_1.17.6      recipes_1.3.1          corpcor_1.6.10
445  [79] class_7.3-23           prettyunits_1.2.0      htmlwidgets_1.6.4
446  [82] ModelMetrics_1.2.2.2   pkgconfig_2.0.3        gtable_0.3.6
447  [85] timeDate_4041.110     bipartite_2.23         survMisc_0.5.6
448  [88] furrr_0.3.1           htmltools_0.5.8.1     carData_3.0-5
449  [91] dotCall64_1.2          scales_1.4.0           gower_1.0.2
450  [94] plsRglm_1.6.0          knitr_1.50             km.ci_0.5-6
451  [97] rstudioapi_0.17.1      reshape2_1.4.4         coda_0.19-4.1
452 [100] statnet.common_4.12.0  checkmate_2.3.3        nlme_3.1-168
453 [103] survAUC_1.4-0         zoo_1.8-14             stringr_1.5.1
454 [106] KernSmooth_2.23-26    parallel_4.5.1         foreign_0.8-90
455 [109] pillar_1.10.2         vctrs_0.6.5           ggpubr_0.6.1
456 [112] car_3.1-3             xtable_1.8-4           cluster_2.1.8.1
457 [115] htmlTable_2.4.3       evaluate_1.0.4         tinytex_0.57
458 [118] mvtnorm_1.3-3         cli_3.6.5              compiler_4.5.1
459 [121] rlang_1.1.6           crayon_1.5.3           rms_8.0-0
460 [124] future.apply_1.20.0    ggsignif_0.6.4         labeling_0.4.3
461 [127] survminer_0.5.1       plyr_1.8.9            stringi_1.8.7
462 [130] viridisLite_0.4.2     network_1.19.0         BiocParallel_1.42.2
463 [133] glmnet_4.1-10         quantreg_6.1           Matrix_1.7-3
464 [136] lars_1.3              hms_1.1.3             future_1.67.0
465 [139] kernlab_0.9-33        rbibutils_2.3          igraph_2.1.4
466 [142] broom_1.0.10          bootstrap_2019.6

```