

RESEARCH

Open Access



Left-censored recurrent event analysis in epidemiological studies: a proposal for when the number of previous episodes is unknown

Gilma Hernández-Herrera^{1,2}, David Moriña^{3,4,5*} and Albert Navarro^{6,7}

Abstract

Background: When dealing with recurrent events in observational studies it is common to include subjects who became at risk before follow-up. This phenomenon is known as left censoring, and simply ignoring these prior episodes can lead to biased and inefficient estimates. We aimed to propose a statistical method that performs well in this setting.

Methods: Our proposal was based on the use of models with specific baseline hazards. In this, the number of prior episodes were imputed when unknown and stratified according to whether the subject had been at risk of presenting the event before $t = 0$. A frailty term was also used. Two formulations were used for this “Specific Hazard Frailty Model Imputed” based on the “counting process” and “gap time.” Performance was then examined in different scenarios through a comprehensive simulation study.

Results: The proposed method performed well even when the percentage of subjects at risk before follow-up was very high. Biases were often below 10% and coverages were around 95%, being somewhat conservative. The gap time approach performed better with constant baseline hazards, whereas the counting process performed better with non-constant baseline hazards.

Conclusions: The use of common baseline methods is not advised when knowledge of prior episodes experienced by a participant is lacking. The approach in this study performed acceptably in most scenarios in which it was evaluated and should be considered an alternative in this context. It has been made freely available to interested researchers as R package *miRecSurv*.

Background

In epidemiological cohort studies, participants may have been at risk of the event of interest before entering follow-up, especially in observational designs. This can lead to unawareness of their history, specifically the time at risk, and whether they have experienced the event (including

how often) at study inception, which is problematic when the baseline hazard of the event is time dependent. There are two relevant situations to consider. First, if an event can only occur once and has already happened, the result is already determined and will require specific statistical techniques for analysis. Second, if the outcome of interest is a recurrent event that can occur more than once, the number of prior episodes will be unknown despite the potential for new events. This represents left censoring where the censored variable is of the discrete type and for which we can define different baseline hazards.

*Correspondence: dmorina@ub.edu

⁵ Facultat d'Economia i Empresa, Universitat de Barcelona (UB), Avinguda Diagonal, 690-694, 08034 Barcelona, Spain

Full list of author information is available at the end of the article



There is a need to clarify how data are handled when the prior history is unknown in a cohort and the outcome of interest is a recurrent event with event dependence. Specifically, this concerns situations where we know the moment from when individuals become at risk, but we do not know the number of prior episodes. For example, concerning the risk of sick leave in a work force, we will likely know the start date for employment; however, especially for people with ample trajectory, we may not know how many sick leaves they had prior to this employment. Another example can be seen in cohorts where the outcome is the incidence of infection with the human papilloma virus in adult women. It would be relatively simple to know how long they have been at risk (first sexual intercourse), but we will not know the true number of infections because they are typically asymptomatic.

This study has two objectives. The first is to describe an analysis proposal for recurrent phenomena in the presence of event dependence when the prior history is unknown for some or all subjects. The second is to compare the performance of our proposed model with another that ignores event dependence.

Methods

Event dependence

When analyzing a recurrent event, we often observe a phenomenon called event dependency in which the baseline hazard of an episode depends on the number of episodes that have already occurred. To date, the phenomenon has been estimated for falls [1], sickness absence [2, 3], hospitalizations in heart failure [4], and cardiovascular readmission after percutaneous coronary intervention [5]. All of these have shown that the baseline hazard increased most significantly with the number of prior episodes.

Recurrent event analyses in epidemiological studies are well described [6–9], typically as extensions of Cox’s classical model of proportional hazards [10, 11]. Specifically, baseline hazard models, also known as conditional models or Prentice, Williams, and Peterson (PWP) models [12], are used to consider the existence of event dependence. These models assume that the baseline hazard of an episode differs as a function of episodes that have already occurred, stratifying by how many there have been. This allows general or specific effects to be calculated for each episode, with all at-risk individuals included in the first strata, but only those with an episode in the previous strata subsequently considered at risk.

PWP models can be formulated in two ways depending on the risk interval used (i.e., how time is considered) [6]. In the first, called “counting process,” time is handled typically for the survival analysis, always referenced from the start of follow-up such that the beginning of the k_{th}

episode is always posterior at the end of the $k-1_{st}$. The hazard function is shown below (Eq. 1). In the second format, called “gap time,” time is always considered in relation to the prior episode such that the start of each new episode for a given individual is set at zero (Eq. 2)

$$\lambda_{ik}(t) = \lambda_{0k}(t)e^{X_i\beta}h_{ik}(t) = h_{0k}(t)e^{X_i\beta} \tag{1}$$

$$h_{ik}(t) = h_{0k}(t - t_{k-1})e^{X_i\beta} \tag{2}$$

where X_i represents the vector of covariables, β represents the coefficients of regression, k is the k_{th} event episode for individual i , and $h_{0k}(t)$ is the function of the baseline hazard dependent on k .

If the phenomenon under study lacks event dependence, these models can be simplified to a function of the common baseline hazard. These are known as Andersen–Gill models [13] and assign the same baseline hazard independent of the episodes that have already been experienced (Eq. 3).

$$h_i(t) = h_0(t)e^{X_i\beta} \tag{3}$$

where $h_0(t)$ is the common baseline hazard.

Individual heterogeneity

If a model is perfectly specified, with all possible covariates accounted for, then the baseline hazard reflects the randomness of the event time given the value of the covariates. In practice, however, it is rarely possible to account for all relevant covariates [14]. This requires another aspect to be considered: the unmeasured effects produced by between-subject variability, presumably due to unobserved exposures. This phenomenon is called individual heterogeneity and in practice is analyzed by adding a frailty to the model, v_i , making an individual random effect to account for the “extra” variability. Given that v_i has a multiplicative effect, we can imagine that it represents the cumulative effect of one or more omitted covariates [15]. The most commonly adopted frailty terms are $E[v_i]=1$ and $V[v_i]=\theta$ [16–18]. Thus, the models specified in Eq. 1, Eq. 2, and Eq. 3 change to those specified in Eq. 4 for “Conditional Frailty Model - Counting Process,” Eq. 5 for “Conditional Frailty Model - Gap Time,” and Eq. 6 for “Shared Frailty Model,” respectively.

$$h_{ik}(t) = v_i h_{0k}(t)e^{X_i\beta} \tag{4}$$

$$h_{ik}(t) = v_i h_{0k}(t - t_{k-1})e^{X_i\beta} \tag{5}$$

$$h_i(t) = v_i h_0(t)e^{X_i\beta} \tag{6}$$

The problem of being unaware of the previous history for some individuals

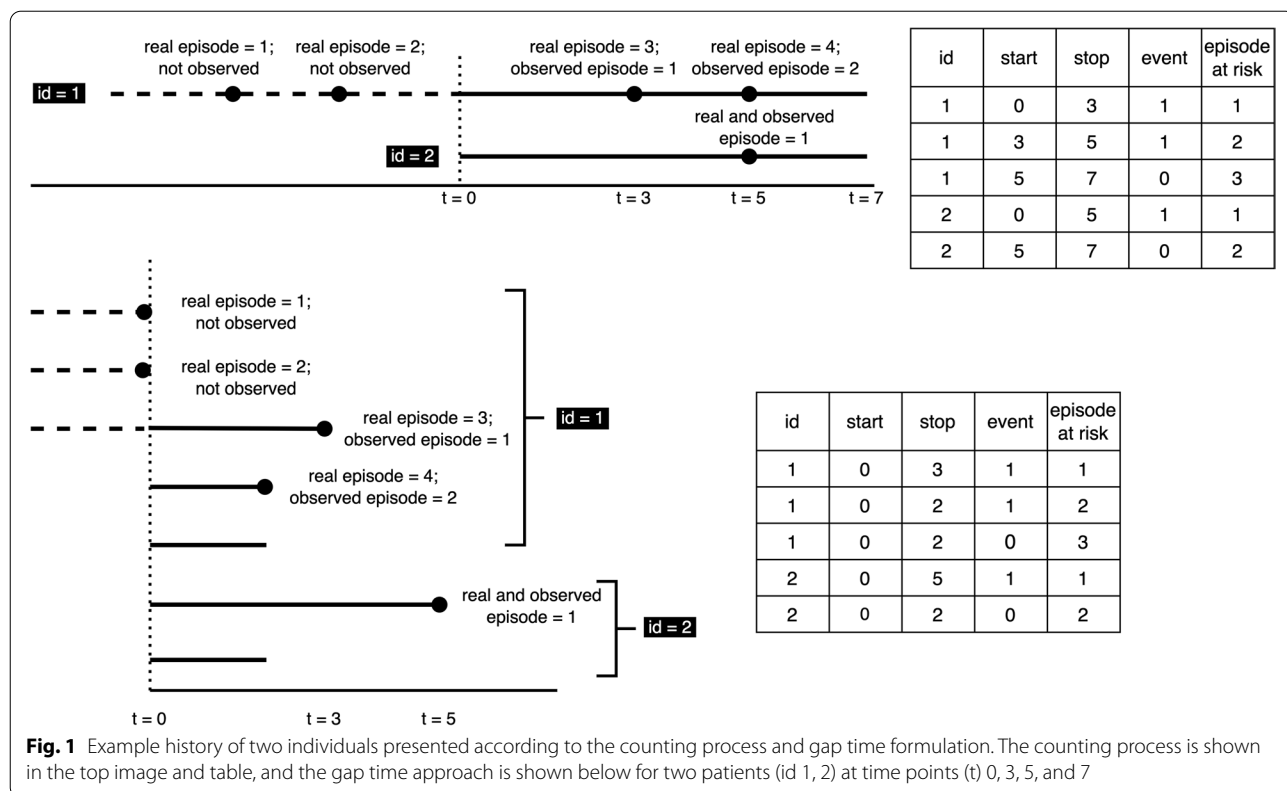
Specific baseline hazard methods, either with or without frailties, can be applied when all required information is known, particularly the number of prior episodes for each individual. In practice, however, this information is not always available, preventing the inclusion of basic information needed to consider event dependence.

Figure 1 shows representative data for two subjects with their respective risk intervals. These are shown according to the counting process on the top and according to gap time on the bottom. Notice that the difference between these formulations is that the gap time “restarts” time to risk at $t=0$ after each episode. When starting our study, the first subject ($id=1$) had two prior episodes and had been at risk for a considerable amount of time, making this a left-censored observation. However, we are only aware of the two episodes that occur from the start of follow-up. By contrast, for the second subject ($id=2$), exposure starts at $t=0$, they have an episode at $t=5$, and follow-up stops at $t=7$. The data tables show the resulting analysis: note that the prior history for Subject 1 “disappears” in both approaches such that the risk only appears at the same instant as it does for Subject 2.

Thus, two problems exist. First, if there is event dependence, and we fail to stratify by number of

episodes, we will mix individuals that have different baseline hazards at that instant (e.g., risk starts for Subject 2 at baseline, when Subject 1 is already at risk of a third). Second, we mix temporal scales: on one side there will be subjects (e.g., Subject 2) whose follow-up time corresponds with their time at risk, but on the other side there will be subjects (e.g., Subject 1) whose follow-up time does not correspond with their time at risk. Consequently, two subpopulations with different baseline hazards are mixed, which goes against the assumptions of the Cox model and its extensions for recurrent phenomena.

When there is missing data for prior history and specific baseline hazard models are unsuitable, an alternative approach is to ignore that history and fit the models based on common baseline hazards (i.e., eqs. 3 and 6). However, these assign the same baseline hazard to all episodes, and as such, they consider neither the possible effect of the number of prior episodes nor the effect of comparing two subjects at the same instant who have times to risk that can be radically different. The use of such models with common baseline hazards for analyzing recurrent phenomena with event dependence has been shown to be highly inefficient, generating high levels of bias in parameter estimates and confidence intervals. Results can be unsound, even if when



the event dependence is of small intensity [19]. Thus, alternative analytic methods are necessary for this context.

Proposal

Our proposal starts with the assumption that, even though the history of a given subject may be unknown, we do know who was at risk before starting follow-up. This is based on three fundamental considerations: 1) imputing k , the number of previous episodes for those subjects at risk before follow-up starts; 2) treating the subpopulation “previously at risk” separately from that “not previously at risk” and 3) using a frailty term to capture the impact of unobserved effects, including uncertainty related to the imputation process. This gives two formulations that we call “Specific Hazard Frailty Model Imputed,” as shown in Eq. 7 for the counting process (SPECIFIC.CP) and Eq. 8 for gap time (SPECIFIC.GT).

$$h_{ikr}(t) = v_i h_{0kr}(t) e^{X_i \beta} \tag{7}$$

$$h_{ikr}(t) = v_i h_{0kr}(t - t_{k-1}) e^{X_i \beta} \tag{8}$$

where k is the number of previous episodes in individual i when they are known or the imputed value when they are not known; r indicates the “previously at risk” or “not previously at risk” subpopulation to which the individual belongs.

In practice, this proposal means that we stratify by the interaction between risk prior to the start of follow-up and the number of prior episodes. Therefore, the use of the individual random error term v_i intends to capture the error that will be made when imputing, as well as the effect of any variable with a non-nil effect not being considered in the analysis. Stratifying by the number of prior episodes intends to safeguard event dependence: doing it as an interaction with the fact that it is an individual previously at risk, or not, separates the two subpopulations to avoid mixing times that are not comparable on the same scale.

To impute the prior history, a generalized linear model (GLM) based on the Conway–Maxwell Poisson distribution (COMPOisson) [20] is fitted to the observed number of events. Imputed values are randomly sampled from the corresponding distribution with the parameters obtained in the previous step, including random noise generated from a normal distribution. To produce proper estimation of uncertainty, the described methodology is used within a multiple imputation framework. The data in the Results section reflect the combination of 5 imputed data sets (following the classic advice to use a low number of imputed data sets and considering that there is little to gain from using more [21–23]), according to Rubin’s

rules [23] and based on the following steps in a Bayesian context:

1. Fit the corresponding COMPOisson model and find the posterior mean $\hat{\beta}$ and variance $V(\hat{\beta})$ of model parameters β .
2. Draw new parameters β^* from $N(\hat{\beta}, V(\hat{\beta}))$.
3. Draw imputations from the COMPOisson distribution using the scores obtained in the previous step as parameters.

A detailed description of the whole imputation process can be reported by Hernández et al. (2020) [24].

Simulation study

Six populations were simulated through the R package, *survsim* [25], corresponding to previously described cohorts of workers [19, 26]. The first three simulated worker history depending on the occurrence of long-term sick leave with several intensities of event dependence. Given that the simulations are carried out through Weibull distributions with ancillary parameters equal to 1 (that is, exponential distributions) the hazard was deemed constant within each episode, but different between episodes (in other words, there was event dependence). For the fourth, fifth, and sixth populations, the outcome was sick leave according to diagnosis (i.e., respiratory system, musculoskeletal system, and mental and behavioral disorders). In this case, the hazard functions were not constant within the episode.

Table 1 shows the characteristics of each episode in each population. The maximum number of episodes that a subject may suffer was not fixed, but the baseline hazard was considered constant when $k \geq 3$. The exposure is represented by covariates X_1 , X_2 and X_3 , with $X_j \approx \text{Bernoulli}(0.5)$ and $\beta_1 = 0.25$, $\beta_2 = 0.5$, and $\beta_3 = 0.75$ representing effects of different magnitudes set independently of the episode k to which the worker is exposed.

Four different situations, which were combinations between two possible follow-up times (2 and 5 years) and two maximum times at risk prior to the beginning of the cohort (2 and 10 years), were simulated for each population. For example, for 10 years’ maximum time at risk and 5 years’ follow-up, we generated dynamic populations with 15 years’ history and selected subjects who were either present at 10 years’ follow-up or incorporated after that date. Follow-up time was then re-scaled, setting $t = 0$ at 10 years for subjects already present in the population, and setting $t = 0$ as the beginning of the follow-up period for those incorporated later. This procedure allowed us to obtain a cohort in which some subjects had worked up to 10 years and

Table 1 Characteristics of the simulated populations

| Episode | Distribution | β_0 | Ancillary | HR |
|----------|--------------|-----------|-----------|-------|
| 1 | Weibull | 8.109 | 1 | 1 |
| 2 | Weibull | 7.927 | 1 | 1.20 |
| ≥ 3 | Weibull | 7.745 | 1 | 1.44 |
| 1 | Weibull | 8.109 | 1 | 1 |
| 2 | Weibull | 7.703 | 1 | 1.50 |
| ≥ 3 | Weibull | 7.298 | 1 | 2.25 |
| 1 | Weibull | 8.109 | 1 | 1 |
| 2 | Weibull | 7.193 | 1 | 2.50 |
| ≥ 3 | Weibull | 6.276 | 1 | 6.25 |
| 1 | Log-normal | 7.195 | 1.498 | 1 |
| 2 | Log-logistic | 6.583 | 0.924 | 1.77 |
| ≥ 3 | Weibull | 6.678 | 0.923 | 2.53 |
| 1 | Log-logistic | 7.974 | 0.836 | 1 |
| 2 | Weibull | 7.109 | 0.758 | 3.81 |
| ≥ 3 | Log-normal | 5.853 | 1.989 | 7.19 |
| 1 | Log-normal | 8.924 | 1.545 | 1 |
| 2 | Log-normal | 6.650 | 2.399 | 10.13 |
| ≥ 3 | Log-normal | 6.696 | 2.246 | 11.19 |

Weibull distribution: $f(t) = \lambda p t^{p-1} e^{-\lambda t^p}, \lambda = e^{-p\beta_0}$
 Lognormal distribution: $f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\frac{1}{\sigma}\{\log(t)-\mu\})^2}{2}}$, $\mu = \beta_0$
 Logistic distribution: $f(t) = \frac{\lambda^{1/\gamma} t^{1/\gamma-1}}{\gamma\{1+(\lambda t)^{1/\gamma}\}^2}, \lambda = e^{-\beta_0}$

were treated as unknown with 5 years of effective follow-up (observed between 10 and 15 years in the original simulated follow-up). In each case, $m = 100$ samples of sizes 250, 500, and 1000 were simulated with different proportions of subjects at risk prior to entering the cohort (0.1, 0.3, 0.5, 0.75 and 1).

Our proposal was compared to a model with frailty and common baseline hazard in terms of the number of previous events, but with different risk per subpopulation. We named this the ‘‘Common Hazard Frailty Model with stratification by subpopulation’’ (COMMON). Importantly, this model does not take event dependence into account, as expressed in Eq. 6, but separates individuals by prior risk:

$$\lambda_{ir}(t) = v_i \lambda_{0r}(t) e^{X_i \beta} \tag{9}$$

Additionally, a second simulation study was conducted with the same technical settings but with $m = 1000$ simulation instances to avoid spurious fluctuations and with only one covariate with no effect ($\beta = 0$), to assess type I error rates for rejecting the null hypothesis of no exposure effect. Table 2 presents the performance measures used for evaluating the different methods.

Results

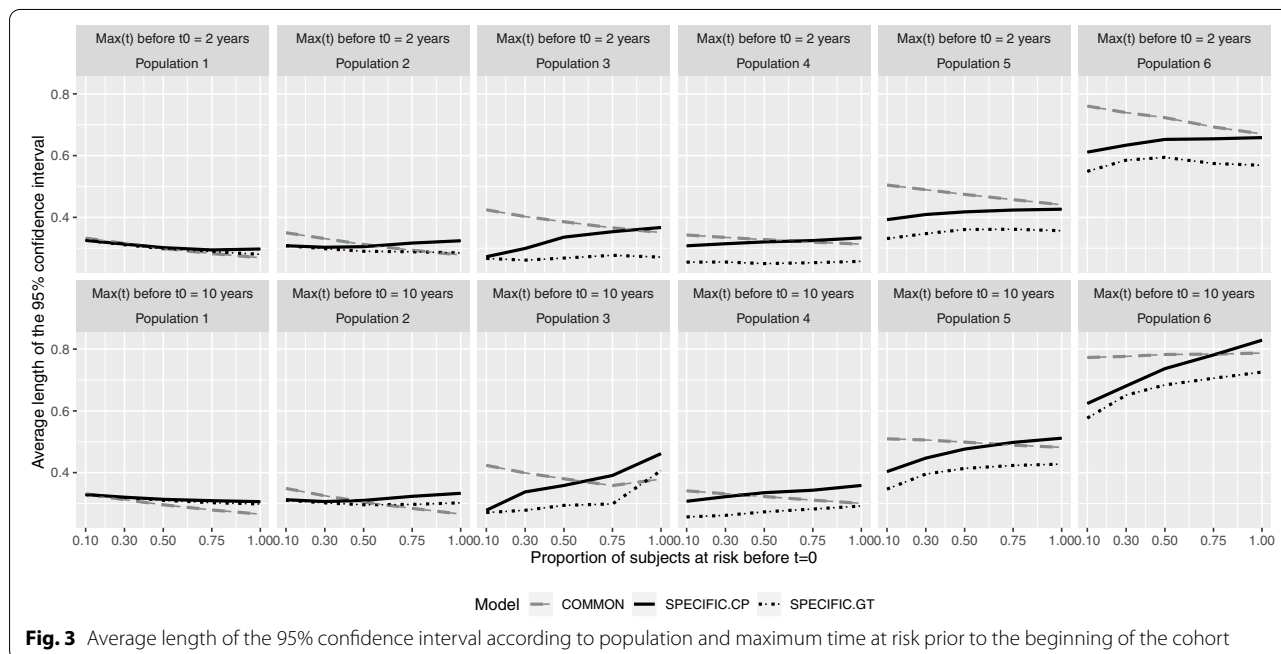
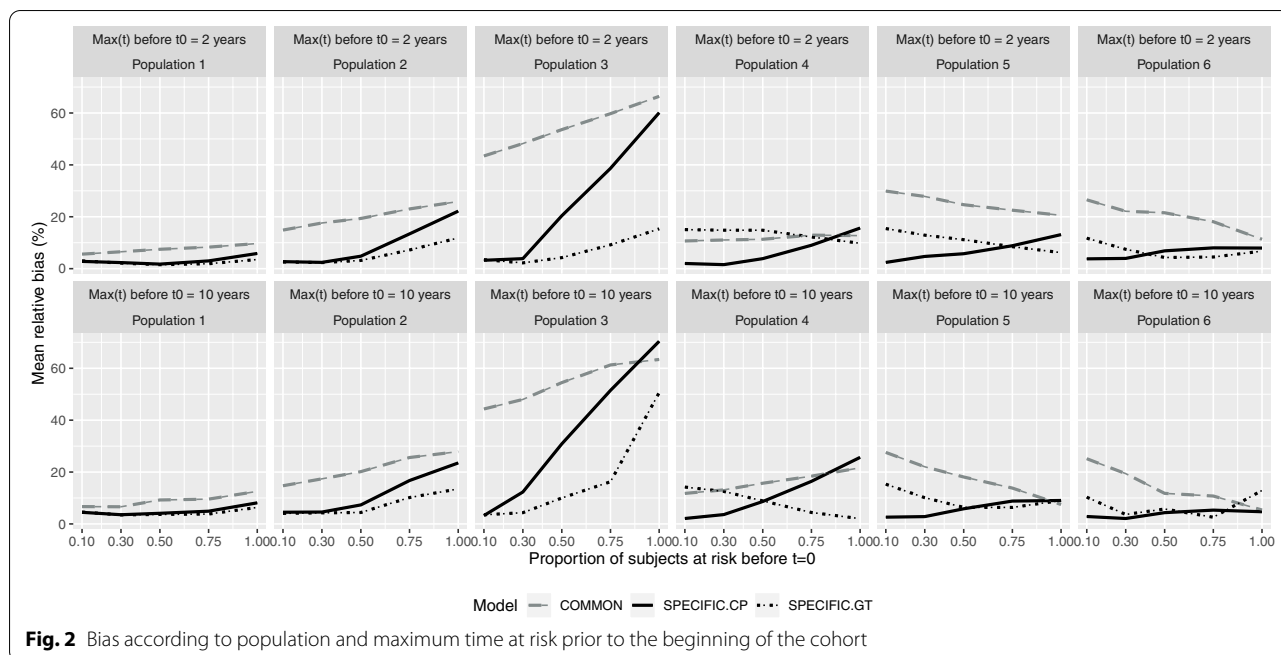
The presented results refer to $n = 1000$ and 5 years’ follow-up because the differences observed for $n = 250$, $n = 500$, and 2 years’ follow-up were not considered relevant (see [Supplementary Material](#)). Regarding bias, Fig. 2 highlights that the COMMON model only obtained values under 10% in population 1 and in some cases when 100% of individuals were at prior risk. Models SPECIFIC.CP and SPECIFIC.GT generally obtain biases < 10% in most situations, except for SPECIFIC.GT in populations 4 and 5, which is slightly above this threshold, and for SPECIFIC.CP, especially in population 3 when 100% of individuals had prior risk at $t = 0$, which seemed more sensitive. For the first three populations, SPECIFIC.GT shows equal or less bias than SPECIFIC.CP, whereas the converse is true for the second three populations, provided that risk starts during follow-up for at least 50% of the cohort.

The average 95% CIs are largest in the COMMON model, excluding population 1, when up to 30% of individuals were at previous risk, always overcoming the accuracy of the SPECIFIC.CP model in populations 1, 2, and 3, with the SPECIFIC.GT model most accurate in populations 4, 5, and 6 (Fig. 3).

The 95% CI coverage of the COMMON model was clearly under 95% in most cases, decreasing as the event dependence increases, and for the first four populations, the proportion at risk prior to $t = 0$. The coverages for models SPECIFIC.CP and SPECIFIC.GT were generally acceptable, even when excessively conservative in some cases (> 95%). The SPECIFIC.CP model failed in

Table 2 Performance evaluation criteria

| Simulation study 1 | Evaluation criteria (covariate $j = 1, 2, 3$ and sample $m = 1, \dots, 100$) |
|-------------------------|---|
| Mean relative bias | $\frac{\sum_{j=1}^3 \sum_{m=1}^{100} \frac{ \hat{\beta}_{jm} - \beta_j }{\beta_j}}{3 \times 100} \times 100$ |
| Average length of 95%CI | $\frac{\sum_{j=1}^3 \sum_{m=1}^{100} 2Z_{1-\alpha/2} SE(\hat{\beta}_{jm})}{3 \times 100}$ |
| Coverage | Percentage of times the 95% confidence interval $\hat{\beta}_{jm} \pm Z_{1-\alpha/2} SE(\hat{\beta}_{jm})$ includes β_j |
| Simulation study 2 | Evaluation criterium |
| Type I error | Percentage of times the $H_0: \beta = 0$ is rejected, for $m = 1, \dots, 1000$ |

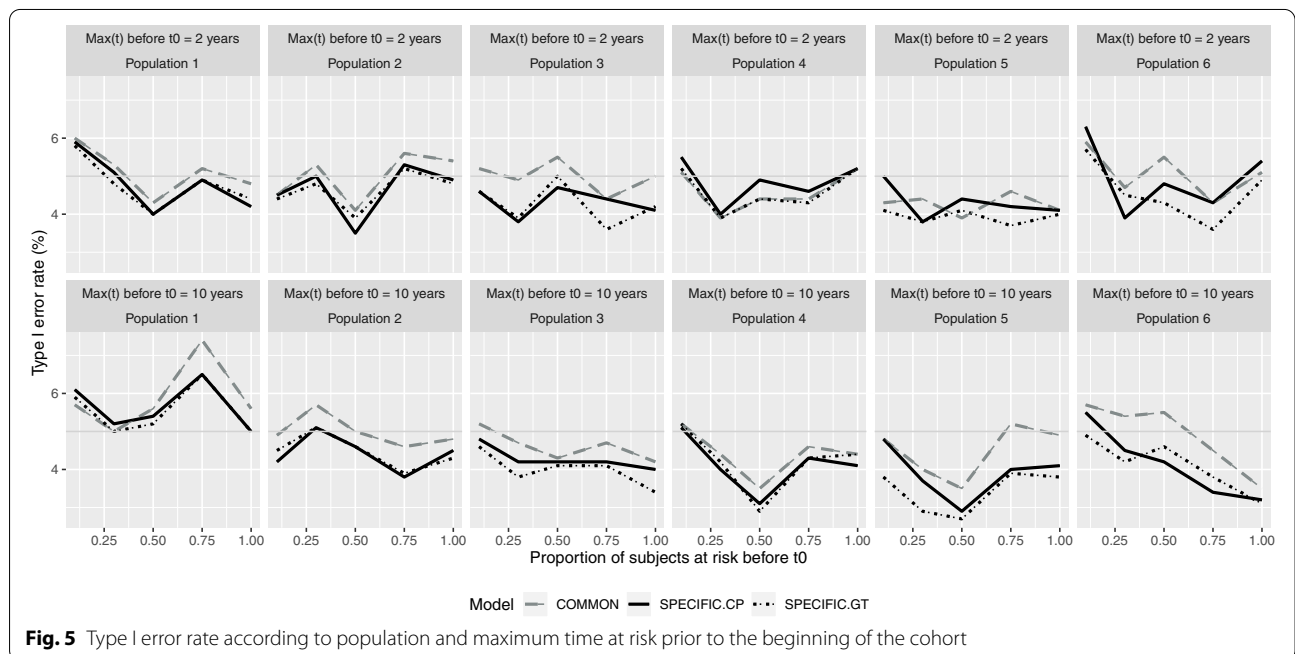
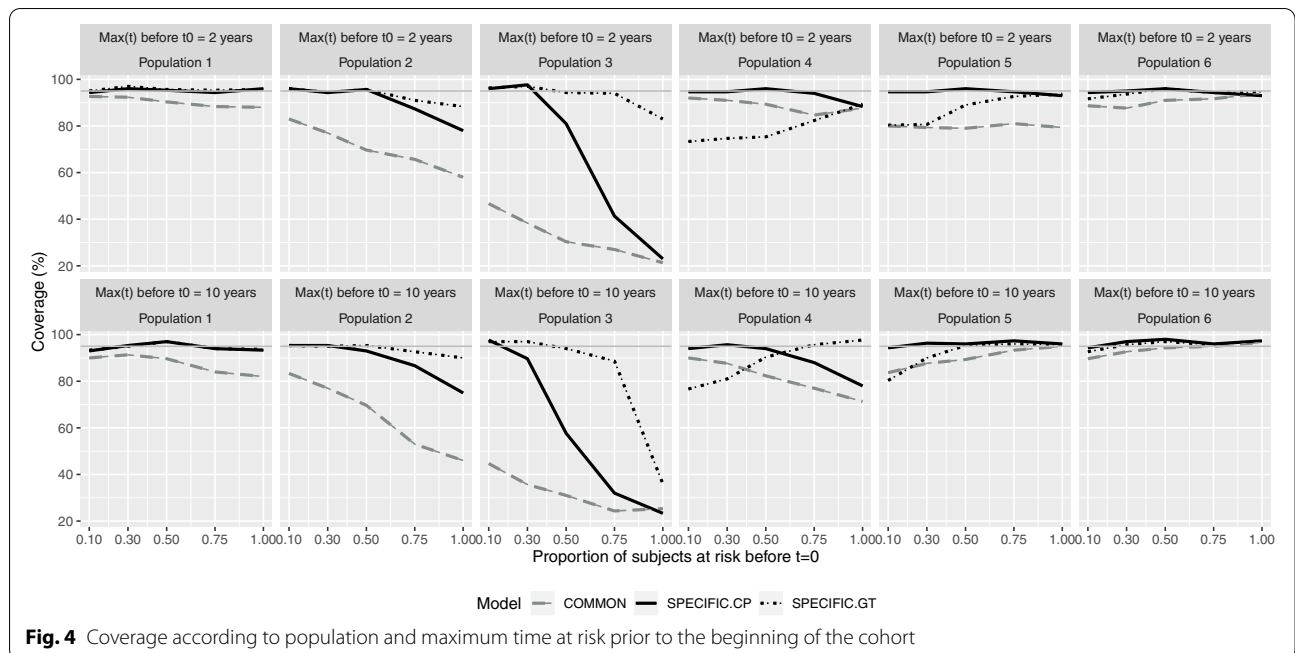


population 3 when 100% had prior risk, due to the high bias, whereas the SPECIFIC.GT model obtained coverage < 80% even in population 4 when < 100% of individuals were at prior risk (Fig. 4).

Tables S1 and S2 (Supplementary Material) summarize the results using arbitrary relative bias (< 10%) and

coverage (92.5–97.5%) criteria. Considering all situations, the COMMON model rarely met these criteria, while the SPECIFIC.CP model showed the highest level of compliance.

Figure 5 shows the type I error distribution for each method, with acceptable rates of around 5% observed in all cases.



Discussion

Our results indicate that not knowing the number of prior episodes an individual has experienced should not justify the use of models with a common baseline hazard, which in most situations show higher bias and less coverage than specific baseline hazard models. The first step, which is indispensable for choosing the correct method,

should be to describe the baseline hazard form of the phenomenon under study in each episode. The SPECIFIC.GT model should be selected if this is a function of constant risk, while the SPECIFIC.CP model would also be acceptable when information about the number of prior episodes is lacking in up to 30–50%; however, the SPECIFIC.CP model should be selected for phenomenon

ruled non-constant when there is a lack of prior information in up to 50% of cases. Neither model is sufficiently robust if the number of previously experienced episodes is unknown for all individuals, even though the SPECIFIC.GT model could be adequate in some cases.

Some readers may be surprised by the finding that bias for the COMMON models in populations 5 and 6 decreased when increasing the proportion of workers at risk before $t=0$. This may be because these populations present high levels of recurrence: thus, more subjects are at risk prior to $t=0$ meaning that more subjects will have already had ≥ 2 episodes (and will be imputed with values ≥ 2). Given that our simulation set had the same baseline hazard from the second episode onwards (a realistic convergence phenomenon), many subjects will have had the same baseline hazard and the COMMON model will have performed better.

Other unexpected results are seen with the partially non-monotonous bias and CI coverage functions associated with increasing the proportion of missing data (e.g., the SPECIFIC.CP and SPECIFIC.GT models in population 1). These situations were in the minority, and the reader should keep in mind that the actual differences were very low despite appearing relevant (the Y-axis scales in some figures are very accurate). The results for $n=250$ and $n=500$ (Supplementary Material) did not follow the same pattern, which indicates that these differences are probably spurious.

Finally, it is important to note that the effectiveness of imputing prior missing episodes to account for left-censored data depends strongly on the assumption of ‘missing at random’ (i.e., that measured covariates actually predict the episode history of an individual). Therefore, a correctly specified imputation (and similarly, an appropriate final survival model specification) is essential to obtain unbiased and efficient estimates.

Conclusions

This simulation study reveals that simply ignoring prior episodes of the outcome of interest and using common baseline hazard models (the most common approach in epidemiological cohort studies) can lead to extremely biased and inefficient estimates. In turn, this will generate inaccurate conclusions, especially when the amount of missing information is high. However, the performance of these estimates can be improved by using the proposed methodology, which is already available on standard software and ready to be used by any interested researcher as R package *miRecSurv* [27].

Abbreviations

COMMON: Common hazard frailty model with stratification by subpopulation; CI: Confidence interval; COMPoisson: Conway-Maxwell Poisson; CP: Counting

process; GLM: Generalized linear model; GT: Gap time; HR: Hazard ratio; PWP: Prentice, Williams, and Peterson model; SPECIFIC.CP: Specific hazard frailty model imputed (Counting process formulation); SPECIFIC.GT: Specific hazard frailty model imputed (Gap Time formulation).

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12874-022-01503-1>.

Additional file 1.

Additional file 2.

Acknowledgements

Not applicable.

Authors' contributions

G.H.H. and A.N. conceived the theoretical framework, and D.M. and G.H.H. designed the simulation study and participated in the analysis of the data. All authors have read and agreed to the published version of the manuscript.

Funding

Not applicable.

Availability of data and materials

R codes used in the described analyses are available as supplementary material and in the GitHub repository <https://github.com/dmorinya/miRecSurv>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

¹Instituto de Investigaciones Médicas, Facultad de Medicina, Universidad de Antioquia, Medellín, Colombia. ²Methodology of Biomedical Research and Public Health, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain. ³Department of Econometrics, Statistics and Applied Economics, Riskcenter-IREA, University of Barcelona (UB), Barcelona, Spain. ⁴Centre de Recerca Matemàtica (CRM), Cerdanyola del Vallès, Spain. ⁵Facultat d'Economia i Empresa, Universitat de Barcelona (UB), Avinguda Diagonal, 690-694, 08034 Barcelona, Spain. ⁶Psychosocial Risks, Organization of Work and Health (POWAH), Autonomous University of Barcelona (UAB), Cerdanyola del Vallès, Spain. ⁷Biostatistics Unit, Faculty of Medicine, Autonomous University of Barcelona (UAB), Cerdanyola del Vallès, Spain.

Received: 16 July 2021 Accepted: 3 January 2022

Published online: 16 January 2022

References

1. Navarro A, Ancizu I. Analyzing the occurrence of falls and its risk factors: some considerations. *Prev Med (Baltim)*. 2009;48:298–302.
2. Navarro A, Reis RJ, Martin M. Some alternatives in the statistical analysis of sickness absence. *Am J Ind Med*. 2009;52(10):811–6.
3. Reis RJ, Utzet M, La Rocca PF, et al. Previous sick leaves as predictor of subsequent ones. *Int Arch Occup Environ Health*. 2011;84:491–9.
4. Braga JR, Tu JV, Austin PC, et al. Recurrent events analysis for examination of hospitalizations in heart failure: insights from the enhanced feedback for effective cardiac treatment (EFFECT) trial. *Eur Hear J Qual Care Clin Outcomes*. 2018;4:18–26.

5. Vasudevan A, Choi JW, Feghali GA, et al. Event dependence in the analysis of cardiovascular readmissions postpercutaneous coronary intervention. *J Investig Med*. 2019;67:943–9.
6. Kelly PJ, Lim LL. Survival analysis for recurrent event data: an application to childhood infectious diseases. *Stat Med*. 2000;19:13–33.
7. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer; 2000.
8. Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. *Int J Epidemiol*. 2015;44:324–33.
9. Box-Steffensmeier JM, De Boef S. Repeated events survival models: the conditional frailty model. *Stat Med*. 2006;25:3518–33.
10. Cox DR. Partial likelihood. *Biometrika*. 1975;62:269–76.
11. Cox DR. Regression models and life table. *J R Stat Soc*. 1972;B34:187–220.
12. Prentice R, Williams B, Peterson A. On the regression analysis of multivariate failure time data. *Biometrika*. 1981;68:373–9.
13. Andersen PK, Gill RD. Cox's regression model counting process: a large sample study. *Ann Stat*. 1982;10:1100–20.
14. Balan TA, Putter H. A tutorial on frailty models. *Stat Methods Med Res*. 2020;29:3424–54.
15. O'Quigley J, Stare J. Proportional hazards models with frailties and random effects. *Stat Med*. 2002;21:3219–33.
16. Rondeau V, Commenges D, Joly P. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Anal*. 2003;9:139–53.
17. Balakrishnan N, Peng Y. Generalized gamma frailty model. *Stat Med*. 2006;25:2797–816.
18. Govindarajulu US, Lin H, Lunetta KL, et al. Frailty models: applications to biomedical and genetic studies. *Stat Med*. 2011;30:2754–64.
19. Navarro A, Casanovas G, Alvarado S, et al. Analyzing recurrent events when the history of previous episodes is unknown or not taken into account: proceed with caution. *Gac Sanit*. 2017;31:227–34.
20. Shmueli G, Minka TP, Kadane JB, et al. A useful distribution for fitting discrete data: revival of the COM-Poisson. *Appl Stat*. 2005;54:127–42.
21. Schafer J, Olsen M. Multiple imputation for multivariate missing-data problems: a data Analyst's perspective. *Multivariate Behav Res*. 1998;33:545–71.
22. Schafer J. Analysis of incomplete multivariate data. London: Chapman & Hall; 1997.
23. Rubin D. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.
24. Hernández-Herrera G, Navarro A, Moriña D. Regression-based imputation of explanatory discrete missing data. arXiv preprint; 2020. Available at: <https://arxiv.org/abs/2007.15031>
25. Moriña D, Navarro A. The R package survsim for the simulation of simple and complex survival data. *J Stat Softw*. 2014;59:1–20.
26. Navarro A, Moriña D, Reis R, et al. Hazard functions to describe patterns of new and recurrent sick leave episodes for different diagnoses. *Scand J Work Environ Health*. 2012;38:447–55.
27. Moriña D, Hernández-herrera G, Navarro A. miRecSurv Package: Prentice-Williams-Peterson models with multiple imputation of unknown number of previous episodes. *R J*. 2021;13(2). <https://journal.r-project.org/archive/2021/RJ-2021-082/RJ-2021-082.pdf>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

