Hybrid Model for Recurrent Event Data

Ivo Sousa-Ferreira and Ana Maria Abreu

Abstract In the last four decades, there has been an increasing interest in developing survival models appropriate for multiple event data and, in particular, for recurrent event data. For these situations, several extensions of the Cox's regression model have been developed. Some of the most known models were suggested by: Prentice, Williams, and Peterson (PWP); Andersen and Gill (AG); Wei, Lin, and Weissfeld (WLW); and Lee, Wei, and Amato (LWA). These models can handle with situations where exist potentially correlated lifetimes of the same subject (due to the occurrence of more than one event for each subject) which is common in this type of data.

In this chapter we present a new model, which we call hybrid model, with the purpose of minimizing some limitations of PWP model. With this model we obtained an improvement in the precision of the parameters estimates and a better fit to the simulated data.

Keywords Correlated observations · Extensions of Cox model · Hybrid model · Recurrent events · Survival analysis

1 Introduction

A historical landmark that has revolutionized the survival analysis took place in 1972, when Sir David Cox [3] proposed a regression model capable of including factors that are assumed to affect the lifetime of the subjects (known as *prognostic*

I. Sousa-Ferreira (✉)
Universidade da Madeira, Funchal, Portugal
e-mail: ivo.ferreira@staff.uma.pt

A. M. Abreu
Universidade da Madeira, Funchal, Portugal

Centro de Investigação em Matemática e Aplicações (CIMA), Funchal, Portugal
e-mail: abreu@staff.uma.pt

S. E. Ahmed et al. (eds.), *Matrices, Statistics and Big Data*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-17519-1_2

or *risk factors*), which are represented by covariates. Based on this model, new extensions and approaches that seek to respond to the most varied problems have been developed.

Over the last few years there has been an increasing interest in studying the time until the observation of various events that may occur more than once for a given subject. The main feature of multiple events data is the observation of more than one lifetime for each subject, which makes the direct application of the Cox's regression model unfeasible. Therefore, several extensions of the Cox model have been suggested to analyze multiple events, in particular a single type of events that occurs more than once for the same subject. Such outcomes have been termed *recurrent events*. For these situations, the most applied extensions of the Cox model were suggested by: Prentice, Williams, and Peterson (PWP) [10]; Andersen and Gill (AG) [1]; Wei, Lin, and Weissfeld (WLW) [15]; and Lee, Wei, and Amato (LWA) [6].

According to Kelly and Lim [5], these models can be classified based on the dependency structure between events (of the same subject). The PWP and AG models have a conditional dependency structure, since subjects are not considered at risk for a given event unless the previous one has occurred. On the other hand, WLW and LWA models have a marginal dependency structure, by reason of it is considered that subjects are simultaneously at risk for the occurrence of any of events from the initial time, i.e., the occurrence of each event is not conditioned on the prior occurrence of any others. Therefore, the first two models are most appropriate to analyze recurrent events, since they allow to accommodate the orderly nature of such data.

One of the major problems in the application of these four models is related to the strong possibility of occurring within-subject correlation. In the Cox model and its extensions the estimation of regression parameters is made assuming that the observations are independent. In other words, we ignore the existence of within-subject correlation. For this reason, from the point of view of the estimation of the parameters, all of these extensions are also called *marginal models* [13]. Several authors [1, 7, 15] have proven that, under certain regularity conditions, the maximum likelihood estimator obtained thereby is still consistent and with the same asymptotic properties, even in the presence of correlated lifetimes.

Consequently, the estimate of variance for the regression parameters also treats each observation as independent. This means that, when the lifetimes are correlated, the usual estimate of the variance does not correctly evaluate the accuracy of estimated regression parameters. In order to offset this aspect, an adjustment in the estimation of variance should take place. Then, a robust estimator of covariance matrix—"sandwich" estimator—was developed to take that correlation into account [8, 15].

In this chapter we present a hybrid model that will focus on the two models that have a conditional dependency structure between events. The purpose of this hybrid model is an attempt to overcome two limitations pointed out by some authors [2, 13] about the PWP model: (1) the loss of heterogeneity throughout the study; and (2) the violation of the missing completely at random (MCAR) condition. Therefore, in the

next section we present the situations where the PWP and AG models are applied and, afterwards, we formalize the new hybrid model. In Sect. 3, the performance of the hybrid model is analyzed. For this purpose, the simulation of recurrent events was carried out through the R statistical software [11]. Finally, some considerations about the application of this model are discussed.

2 Methodology

The PWP and AG models will be formalized to subsequently construct the PWP-AG hybrid model. The characteristics of each model will be examined in order to understand in which situations the application of each of them is more appropriate. In the first instance, it is necessary to introduce some notation which will enable the construction of these survival models.

2.1 Notation

Suppose that there are n subjects in study and each subject can experience a maximum of S failures. Let $T_{is} = \min\left\{X_{is}, C_{is}\right\}$ be the observation time, where X_{is} and C_{is} represent the true lifetime and the censoring time of the sth event ($s = 1, \ldots, S$) in the ith subject ($i = 1, \ldots, n$), respectively. Define $\delta_{is} = I(X_{is} \leq C_{is})$ as being the indicator censoring variable, where $I(E) = 1$ when the event E holds, and $I(E) = 0$ otherwise. It is assumed that censoring is non-informative. Let $\mathbf{z}_{is}(t) = \left(z_{is1}(t), \ldots, z_{isp}(t)\right)'$ represent the p-vector of time-dependent covariates for the ith subject with respect to the sth event and $\mathbf{z}_i(t) = \left(\mathbf{z}'_{i1}(t), \ldots, \mathbf{z}'_{iS}(t)\right)$ denote his overall covariate vector. The true lifetime vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{iS})'$ and the censoring time vector $\mathbf{C}_i = (C_{i1}, \ldots, C_{iS})'$ are assumed to be independent conditional on the overall covariate vector $\mathbf{z}_i(t)$. If X_{is} or $\mathbf{z}_{is}$ is missing, we set $C_{is} = 0$, which ensures that $T_{is} = 0$ and $\delta_{is} = 0$. We require that such cases are MCAR.

2.2 Prentice, Williams, and Peterson (PWP) Model

In 1981, Prentice et al. [10] suggested one of the earliest extensions of the Cox model for the analysis of multiple events and it is often labeled as the *PWP model*. This model applies to the situations in which events occur in an orderly way, where it is considered that a subject cannot be at risk for the sth event until he has experienced the $s - 1$ order event (Fig. 1). Therefore, it means that the risk set is restrictive.

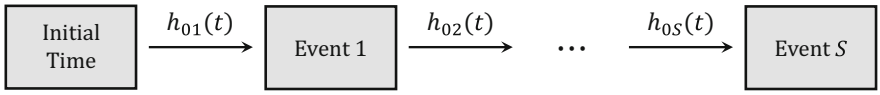

Fig. 1 Schematic representation of the PWP model

Furthermore, it is assumed that the risk of occurrence of the following event is modified by the occurrence of the previous one. This means that it is necessary to stratify the subjects according to the order in which events occur. Thus, if it has been observed s events, then there will be s ordered strata, wherein each of them will be associated a different baseline hazard function $h_{0s}(t)$, $t \geq 0$ and $s = 1, \ldots, S$.

The authors of PWP model have suggested two possible time scales to construct the risk intervals: counting process or gap time formulation. We will only consider the first formulation. Then the hazard function of the ith subject for the sth event is defined as

$$h\big(t; \mathbf{z}_{is}(t)\big) = h_{0s}(t) \exp\big(\boldsymbol{\beta}' \mathbf{z}_{is}(t)\big), \qquad t \geq 0, \tag{1}$$

where $h_{0s}(t) \geq 0$ is the event-specific baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ overall vector of unknown regression parameters.

The regression parameters are estimated through the partial likelihood function, where we admit that the observations within the same subject are independent. For a model with stratification, this function is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s=1}^{S} \left[\frac{\exp\big(\boldsymbol{\beta}' \mathbf{z}_{is}(t_{is})\big)}{\sum_{j=1}^{n} Y_{js}(t_{is}) \exp\big(\boldsymbol{\beta}' \mathbf{z}_{js}(t_{is})\big)} \right]^{\delta_{is}}, \tag{2}$$

where $Y_{is}(t) = I\big(t_{i(s-1)} < t \leq t_{is}\big)$ is the risk set indicator which represents the counting process formulation and t_{is} is the observation time of the ith subject with respect to the sth event.

Conventionally, the overall maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ is obtained by adjusting a single vector of covariates, that in this case is the overall covariate vector $\mathbf{z}_i(t)$. However, since there is stratification, it is also possible to obtain the event-specific vector of unknown regression parameters $\boldsymbol{\beta}_s = (\beta_{s1}, \ldots, \beta_{sp})'$, one for each s stratum [7]. For this purpose, it is required to adjust the event-specific covariate vector of each stratum, in such a way that $\mathbf{z}_i(t) = \big(\mathbf{0}, \ldots, \mathbf{0}, \mathbf{z}_{is}(t), \mathbf{0}, \ldots, \mathbf{0}\big)'$, towards $s = 1, \ldots, S$. Thus, we obtain the event-specific estimates $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2 \ldots, \widehat{\boldsymbol{\beta}}_S$.

In the PWP model, the set of subjects at risk is restricted in the sense that subjects who have not experienced the sth event may not be included in the analysis of the $s + 1$ order event. In this way, the risk set will gradually decrease over the study, revealing increasingly less heterogeneous. Consequently, the event-specific parameters estimates will become unreliable. Therneau and Grambsch [13] presented two options to solve this limitation: (1) truncate the data set exactly in the event where

the number of subjects at risk is considered too small; or (2) agglomerate the final strata, starting from that one is considered to have a small number of subjects at risk. The latter option is more attractive, because it has the advantage of not wasting information that may be critical for the analysis.

In addition to the loss of heterogeneity, in a restrictive risk set the choice of subjects that will be at risk for a given event does not occur randomly, because it is determined by observation of the previous event. This leads to another limitation—the violation of the MCAR assumption [2].

2.3 Andersen and Gill (AG) Model

In 1982, Andersen and Gill [1] proposed a simple model for the analysis of recurrent events, usually referred to as *AG model*. This model was suggested in the same line of reasoning of the previous model but has stronger assumptions. The main assumption concerns with the independence of times between events within a subject.

In this model, the events follow a given order, but it is assumed that the events have equal risk of occurring (Fig. 2). Thus, there will be a common hazard function, $h_0(t)$, $t \geq 0$, to all events.

The AG model was conceived for the case where the occurrence of each event does not depend on time elapsed since the last observation, nor the number of events observed previously. This means that although the occurrence of each event is conditioned to the occurrence of previous events, it is considered that the times between the events are independent.

The authors of this model only considered the counting process formulation to construct the risk intervals. The hazard function for the ith subject with respect to the sth event is defined as

$$h\big(t; \mathbf{z}_{is}(t)\big) = h_0(t) \exp\big(\boldsymbol{\beta}' \mathbf{z}_{is}(t)\big), \qquad t \geq 0, \tag{3}$$

where $h_0(t) \geq 0$ is the common baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ overall vector of unknown regression parameters.

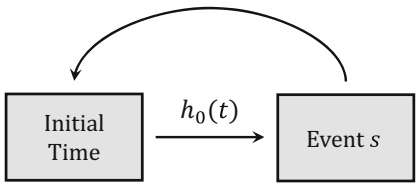

Fig. 2 Schematic representation of the AG model

As in this case there is no stratification, the parameters are estimated through the following partial likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s=1}^{S} \left[\frac{\exp\left(\boldsymbol{\beta}' \mathbf{z}_{is}(t_{is})\right)}{\sum_{j=1}^{n} \sum_{l=1}^{S} Y_{jl}(t_{is}) \exp\left(\boldsymbol{\beta}' \mathbf{z}_{jl}(t_{is})\right)} \right]^{\delta_{is}}, \tag{4}$$

where $Y_{is}(t) = I\left(t_{i(s-1)} < t \leq t_{is}\right)$ is the risk set indicator and t_{is} is the observation time of the ith subject with respect to the sth event. Since this model has a common hazard function to all events, it is only possible to obtain the overall maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$.

When the PWP and AG models have been suggested, they did not present any adjustment for the within-subject correlation. However, the authors of these models were conscious of this strong possibility and recommended attempt to capture this correlation including time-dependent covariates on the model. A few years later, it was realized that it was possible to take advantage of the fact that these two models are also classified as marginal models from the point of view of the parameters estimation. From this point, the robust estimator of the covariance matrix was applied to take the within-subject correlation into account [7, 13].

In contrast with the previous model, the AG model reveals neither loss of the heterogeneity nor the violation of the MCAR condition because the set of subject at risk is unrestrictive. This means that all risk intervals of all subjects may contribute to the risk set for any given event, regardless of the number of events observed previously for each subject [5].

2.4 PWP-AG Hybrid Model

In order to overcome the limitations pointed for the PWP model, we present a slightly different option from those presented in [13].

In fact, there may exist another reason to agglomerate the final strata. Suppose that, initially, the PWP model has a very heterogeneous risk set, whereby the differences between the hazard functions of the various subjects are due, in particular, to the effect of several covariates with quite different values for each of them. Suppose that the risk set related to the second event (which contains only the subjects that had suffered the first event) no longer contains the subjects who belong to a certain category of a covariate. This means that, in addition to subjects being less heterogeneous, this covariate is no longer important to the model. Thus, its effect on the survival shall be embedded in the baseline hazard functions, which necessarily have to be different from the baseline hazard function of the first event. With that in mind, assume that after a certain event, denoted by S^*, the subjects at risk will be more homogeneous, in such a way that the baseline hazard functions

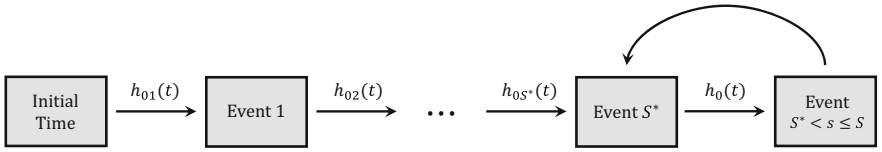

Fig. 3 Schematic representation of the PWP-AG hybrid model

for each event no longer have to be different. Then, instead of continuing to apply the PWP model, the AG model can be implemented after the observation of the S^* event, as illustrated in Fig. 3.

In general, the events $1, 2, \ldots, S^*$ are analyzed with the PWP model (1) and the events $S^* + 1, S^* + 2, \ldots, S$ are analyzed with the AG model (3). The proposal to agglomerate the final strata gives rise to the PWP-AG hybrid model that, as far as we know, has not yet been mathematically formalized in the available literature. Therefore, considering the counting process formulation, the hazard function for the ith subject regarding the sth event is defined as

$$h\big(t; \mathbf{z}_{is}(t)\big) = \begin{cases} h_{0s}(t) \exp\big(\boldsymbol{\beta}' \mathbf{z}_{is}(t)\big), & 0 < s \leq S^* \\ h_0(t) \exp\big(\boldsymbol{\beta}' \mathbf{z}_{is}(t)\big), & S^* < s \leq S \end{cases}, \quad t \geq 0.$$

where $h_{0s}(t) \geq 0$ is the event-specific baseline hazard function, $h_0(t) \geq 0$ is the common baseline hazard function, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ overall vector of unknown regression parameters.

Similarly it is necessary to adapt the partial likelihood functions of PWP (2) and AG (4) models to this situation. Admitting that the observations within the same subject are independent, the overall maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ is obtained through the following function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{s=1}^{S} \left[\frac{\exp\big(\boldsymbol{\beta}' \mathbf{z}_{is}(t_{is})\big)}{\mathbf{Q}_s(\boldsymbol{\beta}, t_{is})^{\Delta_s} \, \overline{\mathbf{Q}}(\boldsymbol{\beta}, t_{is})^{1-\Delta_s}} \right]^{\delta_{is}},$$

where $\mathbf{Q}_s(\boldsymbol{\beta}, t) = \sum_{j=1}^{n} Y_{js}(t) \exp\big(\boldsymbol{\beta}' \mathbf{z}_{js}(t)\big)$, $\overline{\mathbf{Q}}(\boldsymbol{\beta}, t) = \sum_{s=1}^{S} \mathbf{Q}_s(\boldsymbol{\beta}, t)$ and $\Delta_s = I(s \leq S^*)$ denote the indicator model variable which takes the value $\Delta_s = 1$ when the PWP model is considered ($0 < s \leq S^*$) and $\Delta_s = 0$ when the AG model is considered ($S^* < s \leq S$). It should be noted that Δ_s does not depend on the i index, which means that for all subjects we define that the AG model is applied from the S^* event. Furthermore, it is noteworthy that when $\Delta_s = 1$ we can also obtain the event-specific estimators $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_{S^*}$. The estimation of the event-specific regression parameters is performed by the same procedure described in Sect. 2.2.

3 PWP-AG Hybrid Model with Simulated Data

The application of the marginal models can be easily accomplished by `R`, S-Plus, or SAS statistical software. In this contribution we used the `R` statistical software [11], version 3.4.0, where for the analysis we used the `survival` package [12].

In order to evaluate the performance of the PWP-AG hybrid model, we proceeded with the simulation of a recurrent event data, bearing in mind the characteristics of the situations where this model is applied. The data set was simulated with `survsim` package [9], where we considered that the time to right censoring and to events follows a Weibull distribution. The values of the covariates were simulated from Bernoulli distribution (with probability of success $p = 0.5$), uniform distribution (that takes values in the range [0, 1]), and standard gaussian distribution. Let them be denoted by `x`, `x.1` and `x.2`, respectively. The procedure used for the simulation of this data set was recently presented by Ferreira [4].

Before applying any model, we decided to analyze the evolution of the risk for each event over time. The cumulative hazard functions from Kaplan-Meier estimates on the left side of Fig. 4 show that the first four events have different risks of occurring, but after that the risk is more similar. This was the main reason why we considered the PWP-AG hybrid model with $S^* = 4$ (right side of Fig. 4). Also, in Table 1 it can be seen the number of subjects at risk and observed events in each event number, where the decreasing over the strata becomes obvious. This means that if we want to calculate the event-specific estimates for the PWP model, these will be unreliable or even missing.

The implementation code of the PWP-AG hybrid model is very similar to the code of the PWP model [13, 14], the only difference lies in the way that we define the stratification variable. For the hybrid model it is necessary to define a new

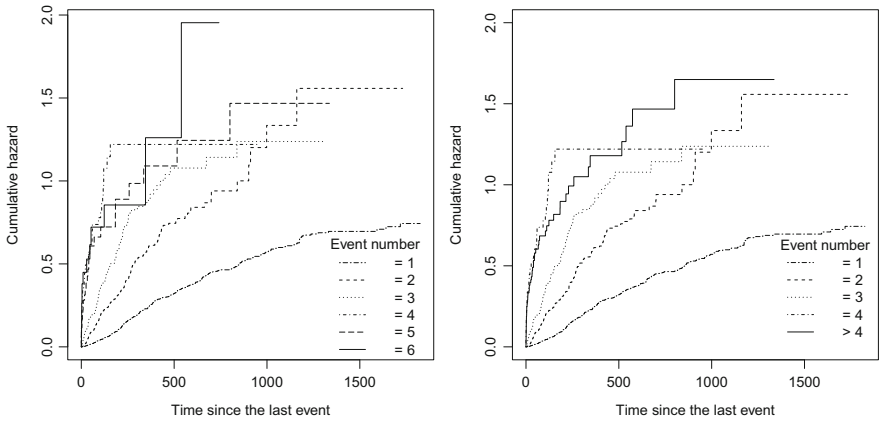

Fig. 4 Cumulative hazards from Kaplan-Meier estimates of the first 6 events (*left*) and of the first 4 events with the following ones agglomerated in the last stratum (*right*)

Table 1 Number of subjects at risk and observed events

	Event number									
	1	2	3	4	5	6	7	8	9	10
Subjects at risk	1000	365	162	87	48	27	15	10	6	3
Observed events	365	162	87	48	27	15	10	6	3	0

Table 2 Overall estimates of the various parameters associated with each model

Covariate/Model	$\widehat{\beta}_j$	$\exp(\widehat{\beta}_j)$	$se(\widehat{\beta}_j)$	$se_r(\widehat{\beta}_j)$	p-value
x					
PWP	0.52734	1.69442	0.08076	0.08624	9.66e−10
PWP-AG	0.52234	1.68598	0.07998	0.08615	1.33e−09
x.2					
PWP	0.67673	1.96742	0.04759	0.05003	<2e−16
PWP-AG	0.67702	1.96800	0.04694	0.04943	<2e−16

stratification variable, where we specified that after the $S^* = 4$ event the last strata are agglomerate in the $S^* + 1 = 5$ event number. So, in the stratification variable of the PWP-AG hybrid model the first four strata remain unchanged and the last strata are agglomerated in the stratum number five. Consequently, there will be an event-specific baseline hazard function for the first four events and a common baseline hazard function for the subsequent events.

The analysis revealed that the covariate `x.1` was not significant in both models (p-value=0.553 and p-value=0.403 in PWP and PWP-AG hybrid models, respectively). Therefore, in Table 2 we present the results of the models with the remaining two covariates. The parameters estimates were similar but the standard errors were slightly smaller in PWP-AG hybrid model, thus improving the accuracy of the estimates. For both models, the robust standard errors were inflated compared to the usual ones. This observed inflation suggests that there is less variation within-subjects than between-subjects [5].

In addition, the value of concordance for both models is 0.722. However, the value of R^2 is better for PWP-AG hybrid model ($R^2 = 0.131$ vs $R^2 = 0.129$).

4 Conclusions and Future Work

The proposed PWP-AG hybrid model revealed to be an alternative to the PWP model. The decision of gathering the last events was mainly based on the similarity of the cumulative hazard functions and not only on the dimension of the risk set. The fact that subjects become more homogeneous does not ensure that the hazard functions corresponding to the subsequent events are the same because the

mechanism that triggers such events may cause differences in these functions. This is the reason why it is important to represent the cumulative hazard function of each event.

On the other hand, the simulation showed that the parameters estimates become more accurate. Moreover, in this case the PWP-AG hybrid model has resulted in a better fit to the simulated data.

Although the PWP-AG hybrid model may not completely overcome the limitations of the PWP model (the loss of heterogeneity and the violation of the MCAR assumption), nevertheless these limitations are reduced. Therefore, the PWP-AG hybrid model is a compromise between PWP and AG models, which allows to compile the features of each of them.

Further work is required with this model, namely a simulation study, which can clarify when this model is more appropriate than the other models for recurrent events.

Acknowledgements This research was partially supported by FCT—Fundação para a Ciência e a Tecnologia with Portuguese Funds, Project UID/MAT/04674/2013.

References

1. Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics, 10*(4), 1100–1120.
2. Cai, J., & Schaubel, D. E. (2004). Analysis of recurrent event data. In N. Balakrishnan, & C. R. Rao (Eds.), *Handbook of statistics*. Advances in survival analysis (vol. 23, pp. 603–623). North Holland: Elsevier.
3. Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B. 34*(2), 187–220.
4. Ferreira, I. M. S. *Modelos para acontecimentos múltiplos*. Master's dissertation, University of Madeira, Funchal (2016). http://hdl.handle.net/10400.13/1326.
5. Kelly, P. J., & Lim, L. L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine, 19*(1), 457–481.
6. Lee, E. W., Wei, L. J., & Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In J. P. Klein, & P. K. Goel (Eds.), *Survival analysis: State of the art* (pp. 237–247). Dordrecht: Kluwer Academic Publisher.
7. Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine, 13*(21), 2233–2247.
8. Lin, D. Y., & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association, 22*(4), 1074–1078.
9. Moriña, D., & Navarro, A. (2015). *Survsim: Simulation of simple and complex survival data*. R package version 1.1.4. http://CRAN.R-project.org/package=survsim.
10. Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika, 68*(2), 373–379.
11. R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. http://www.R-project.org/.
12. Therneau, T. M. (2017). *Survival: A package for survival analysis in S*. R package version 2.41-3. http://CRAN.R-project.org/package=survival.

13. Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. New York: Springer.
14. Therneau, T. M., & Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine, 16*(18), 2029–2047.
15. Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association, 84*(408), 1065–1073.