Gac Sanit. 2017;31(3):227–234

Original article

Analyzing recurrent events when the history of previous episodes is unknown or not taken into account: proceed with caution

Albert Navarro[a,*], Georgina Casanovas[a], Sergio Alvarado[b,c], David Moriña[a,d]

[a] GRAAL-Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
[b] Programa de Salud Ambiental, Escuela de Salud Pública, Facultad de Medicina, Universidad de Chile, Chile
[c] Facultad de Ciencias de la Salud, Universidad de Tarapacá, Arica, Chile
[d] Unit of Infections and Cancer (UNIC), Cancer Epidemiology Research Program (CERP), Catalan Institute of Oncology (ICO)-IDIBELL, Barcelona, Spain

ARTICLE INFO

Article history:
Received 23 June 2016
Accepted 8 September 2016
Available online 15 November 2016

Keywords:
Recurrence
Cohort studies
Risk assessment
Survival analysis
Bias

ABSTRACT

Objective: Researchers in public health are often interested in examining the effect of several exposures on the incidence of a recurrent event. The aim of the present study is to assess how well the common-baseline hazard models perform to estimate the effect of multiple exposures on the hazard of presenting an episode of a recurrent event, in presence of event dependence and when the history of prior-episodes is unknown or is not taken into account.
Methods: Through a comprehensive simulation study, using specific-baseline hazard models as the reference, we evaluate the performance of common-baseline hazard models by means of several criteria: bias, mean squared error, coverage, confidence intervals mean length and compliance with the assumption of proportional hazards.
Results: Results indicate that the bias worsen as event dependence increases, leading to a considerable overestimation of the exposure effect; coverage levels and compliance with the proportional hazards assumption are low or extremely low, worsening with increasing event dependence, effects to be estimated, and sample sizes.
Conclusions: Common-baseline hazard models cannot be recommended when we analyse recurrent events in the presence of event dependence. It is important to have access to the history of prior-episodes per subject, it can permit to obtain better estimations of the effects of the exposures

Análisis de eventos recurrentes cuando la historia de episodios previos es desconocida o no se tiene en cuenta: proceder con cautela

Palabras clave:
Recurrencia
Estudios de cohortes
Medición del riesgo
Análisis de supervivencia
Sesgo

RESUMEN

Objetivo: A menudo los investigadores en salud pública están interesados en examinar el efecto de varias exposiciones en la incidencia de un evento recurrente. El objetivo de este estudio es evaluar el funcionamiento de los modelos de riesgo basal común al estimar el efecto de múltiples exposiciones sobre el riesgo de presentar un episodio de un evento recurrente, cuando existe dependencia del evento y los antecedentes de los episodios por sujeto son desconocidos o bien no se tienen en cuenta.
Métodos: Mediante un estudio exhaustivo de simulación, utilizando modelos de riesgo basal específico como referencia, se evalúa el rendimiento de los modelos de riesgo basal común a través de diversos criterios: sesgo, error cuadrático medio, cobertura, longitud de los intervalos de confianza y compatibilidad con el supuesto de riesgos proporcionales.
Resultados: El sesgo empeora a medida que aumenta la dependencia del evento, llevando a una sobreestimación considerable del efecto de la exposición; los niveles de cobertura y el cumplimiento del supuesto de riesgos proporcionales son bajos o muy bajos, lo que empeora con el aumento de la dependencia del evento, el efecto a estimar y el tamaño muestral.
Conclusiones: El uso de modelos de riesgo basal común no puede recomendarse cuando analizamos eventos recurrentes en presencia de dependencia del evento. Es importante tener acceso a los antecedentes de episodios previos por sujeto, ya que ello puede permitir obtener mejores estimaciones de los efectos de las exposiciones.

Introduction

The outcome of interest in a biomedical or epidemiological study is often an event that can occur more than once in a subject.

* Corresponding author.
E-mail address: albert.navarro@uab.cat (A. Navarro).

http://dx.doi.org/10.1016/j.gaceta.2016.09.004

Therefore, identifying a statistical method suitable for studying recurrent events is of great interest to the field.

From a statistical point of view, recurrent event analysis presents two major challenges. The first is individual heterogeneity, i.e. the unmeasured effects produced by between-subject variability, presumably due to unobserved exposures. For instance, imagine that a study measuring the number of respiratory crises is not asking for smoking status. It is likely that smokers will have a different pattern from non-smokers, resulting in heterogeneity across the subjects that can't be attributed to any known factor as smoking status was not recorded. This issue is usually tackled using frailty models, which incorporate random effect terms to account for this "extra" variability. The second problem is within-subject correlations attributable to a single subject suffering multiple episodes of the event. These correlations are especially problematic in situations complicated by event dependence, in other words, when the risk of having a new episode depends on the number of previous episodes. This is the case of the number of sick leaves suffered by workers: A history of sick leaves increases the risk of a subsequent episode. Reis et al.[1] quantified the extent of this increase. If we fail to account for event dependence, our resulting estimators will be inefficient and potentially biased. As discussed in Box-Steffensmeier et al.,[2] common-baseline hazard models average the effects across all events not taking strata into account, being this averages biased in a predictable direction. In cohort studies, event dependence can be controlled by using survival models with specific-baseline hazards for each episode that the subject faces.[3]

Amorim and Cai[4] provide an excellent review of approaches to recurrent event analysis. The article describes the applicable statistical methods for epidemiological studies of recurrent events, working off of the assumption that researchers have access to all of the information required by each model. In practice, however, much of this data is typically unavailable. Specific-baseline hazard models assume that the exact number of previous episodes suffered by each subject is known, but in reality it is typically impractical to obtain an exhaustive history for each patient. This leaves us without a method to directly address event dependence. The usual practice in such cases is to fit models with a common-baseline hazard.

The aim of the present study is to assess how well these common-baseline hazard models perform when they are used to estimate the effect of multiple exposures on the hazard of presenting an episode of a recurrent event when the previous history is not taken into account.

Methods

Simulations

Example

We illustrate this work by reproducing a study from the literature[5] to analyze long-term sickness absence (SA) frequency in a cohort of Dutch workers. We will use the same baseline hazard as in the Dutch study, 0.0021 per worker-week. The between-episodes hazard ratios (HR) do not correspond exactly to those of any specific study, although Reis et al.[1] provide values for a wide range of SA-related diagnoses. SA is a commonly-used outcome in occupational health studies because it is considered a major economic and public health issue,[6–8] resulting in a growing interest in identifying the best method to quantitatively and efficiently analyze this phenomenon.[5,9–13]

Generation of populations

Six different populations of 250 000 workers, each with 20 years of history, were generated using the survsim[14,15] package in R 2.15.3 (R Foundation for Statistical Computing, Vienna, Austria).

Table 1
Characteristics of the simulated populations.

	Baseline hazard		HR	υ_i
	Worker-days	Worker-weeks		
Population 1				
$\beta_{01}=8.109$	0.000301	0.002106	1	None
$\beta_{02}=7.927$	0.000361	0.002526	1.20	
$\beta_{03}=7.745$	0.000433	0.003030	1.44	
Population 2				
$\beta_{01}=8.109$	0.000301	0.002106	1	Gamma (1,0.1)
$\beta_{02}=7.927$	0.000361	0.002526	1.20	
$\beta_{03}=7.745$	0.000433	0.003030	1.44	
Population 3				
$\beta_{01}=8.109$	0.000301	0.002106	1	None
$\beta_{02}=7.703$	0.000451	0.003160	1.50	
$\beta_{03}=7.298$	0.000677	0.004738	2.25	
Population 4				
$\beta_{01}=8.109$	0.000301	0.002106	1	Gamma (1,0.1)
$\beta_{02}=7.703$	0.000451	0.003160	1.50	
$\beta_{03}=7.298$	0.000677	0.004738	2.25	
Population 5				
$\beta_{01}=8.109$	0.000301	0.002106	1	None
$\beta_{02}=7.193$	0.000752	0.005263	2.50	
$\beta_{03}=6.276$	0.001881	0.013166	6.25	
Population 6				
$\beta_{01}=8.109$	0.000301	0.002106	1	Gamma (1,0.1)
$\beta_{02}=7.193$	0.000752	0.005263	2.50	
$\beta_{03}=6.276$	0.001881	0.013166	6.25	

β_{03} refers to β_0 for the third and subsequent episodes.
HR: hazard ratio.

For each subject i, the hazard of the next episode k was simulated through an exponential distribution:

$$h_{ik}(t) = \nu_i \cdot e^{\left(-\beta_{0k}+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3\right)} \quad (1)$$

where $h_{0k}(t) = e^{-\beta_{0k}}$, i.e. the baseline hazard for subjects exposed to episode k. The maximum number of SA episodes that a worker may suffer was not fixed, although the baseline hazard was considered constant when $k \geq 3$. X_1, X_2, and X_3 are the three covariates that represent the exposure, with $X_i \sim Bernoulli(0.5)$. β_1, β_2, and β_3 are the parameters of the three covariates that represent the effect, set independently of the episode k to which the worker is exposed, as: $\beta_1 = 0.25$, $\beta_2 = 0.5$, and $\beta_3 = 0.75$ in order to represent effects of different magnitudes. ν_i is a random effect.

Event dependence

Event dependence was addressed by using various values of $h_{0k}(t)$, specifying different β_{0k}. Table 1 presents the specifications for the generated populations in terms of the baseline hazards by SA episode and random effects used. Table 1 also presents the HR resulting from the comparison of the baseline hazard with that of the first episode, which gives us the event dependence for the phenomenon. Note that for populations 1 and 2, the HR = 1.20 and 1.44, respectively, for the second SA episode, as well as for the third and subsequent SA episodes with respect to the first. This means that between the second and third SA episodes, the baseline hazard was also increased by a factor of 1.20. The HR = 1.50 between episodes two and three for populations 3 and 4, and 2.50 for populations 5 and 6. We chose to simulate phenomena with increasing event dependence, given that Reis et al.[1] demonstrated that the hazard always increases in the presence of previous SA.

Individual heterogeneity

Individual heterogeneity was addressed by introducing ν_i, the random effect. This effect was held constant over the various episodes for a given subject but varied between subjects.