Journal of Clinical Epidemiology 83 (2017) 75−84

Journal of Clinical Epidemiology

The number of primary events per variable affects estimation of the subdistribution hazard competing risks model

Peter C. Austin[a,b,c,*], Arthur Allignol[d], Jason P. Fine[e,f]

[a]Institute for Clinical Evaluative Sciences, G106, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada
[b]Institute of Health Management, Policy and Evaluation, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M6, Canada
[c]Schulich Heart Research Program, Sunnybrook Research Institute, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada
[d]Institute of Statistics, Ulm University, Helmholtzstr. 20, Ulm 89081, Germany
[e]Department of Biostatistics, University of North Carolina, 135 Dauer Drive, 3101 McGavran-Greenberg Hall, CB #7420 Chapel Hill, NC 27599-7420, USA
[f]Department of Statistics & Operations Research, University of North Carolina, 318 Hanes Hall, CB# 3260, Chapel Hill, NC 27599-3260, USA

Accepted 10 November 2016; Published online 12 January 2017

Abstract

Objectives: To examine the effect of the number of events per variable (EPV) on the accuracy of estimated regression coefficients, standard errors, empirical coverage rates of estimated confidence intervals, and empirical estimates of statistical power when using the Fine−Gray subdistribution hazard regression model to assess the effect of covariates on the incidence of events that occur over time in the presence of competing risks.

Study Design and Setting: Monte Carlo simulations were used. We considered two different definitions of the number of EPV. One included events of any type that occurred (both primary events and competing events), whereas the other included only the number of primary events that occurred.

Results: The definition of EPV that included only the number of primary events was preferable to the alternative definition, as the number of competing events had minimal impact on estimation. In general, 40−50 EPV were necessary to ensure accurate estimation of regression coefficients and associated quantities. However, if all of the covariates are continuous or are binary with moderate prevalence, then 10 EPV are sufficient to ensure accurate estimation.

Conclusion: Analysts must base the number of EPV on the number of primary events that occurred.

Keywords: Survival analysis; Competing risks; Subdistribution hazard model; Sample size; Fine−Gray regression model; Events per variable

1. Introduction

Quantifying the occurrence of an adverse event or outcome over time is an important issue in clinical medicine and public health research. There is an increasing interest in the incidence of nonfatal events (e.g., incidence of heart disease, occurrence of an infection) or the incidence of cause-specific mortality (e.g., incidence of death due to cardiovascular disease or death due to cancer). In such settings, the presence of competing risks must be taken into account when assessing the effect of prognostic factors on the incidence of an outcome over time. A competing risk is an event whose occurrence precludes the occurrence of the event of interest [1−6]. For instance, when evaluating the effect of risk factors on the incidence of death due to cardiovascular disease, death to noncardiovascular causes serves as a competing risk because subjects who die of a noncardiovascular cause

Funding: This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin was supported by Career Investigator awards from the Heart and Stroke Foundation. The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study were funded by a CIHR Team Grant in Cardiovascular Outcomes Research (CTP 79847 and CRT43823). These data sets were linked using unique, encoded identifiers, and analyzed at the Institute for Clinical Evaluative Sciences (ICES).

Conflict of interest: None.

* Corresponding author. Tel.: (416)-480-6131; fax: (416)-480-6048.
E-mail address: peter.austin@ices.on.ca (P.C. Austin).

http://dx.doi.org/10.1016/j.jclinepi.2016.11.017
0895-4356/

What is new?

Key findings

- The number of type 1 events (primary events) was more important than the number of events of any type for assessing the number of events per variable (EPV) when estimating a subdistribution hazard model.
- Forty to 50 EPV are necessary to ensure accurate estimation of regression coefficients and associated quantities.
- If all of the covariates are continuous or are binary with moderate prevalence, then 10 EPV are sufficient to ensure accurate estimation.

What this adds to what was known?

- Previous research has examined the number of EPV necessary to fit logistic regression models (binary outcomes) or a Cox proportional hazards models (survival outcomes in the absence of competing risks). The current research extends the findings of these earlier studies to the setting in which competing risks are present.

What is the implication and what should change now?

- Authors and analysts need to be aware that the number of type 1 events (or the primary event of interest) is the key number when determining the number of EPV and whether there are an adequate number of events for fitting the desired subdistribution hazard model.

(e.g., cancer) are no longer at risk of death due to cardiovascular disease.

Naïve use of the conventional Cox proportional hazards model that censors the competing event leads to biased estimates of the effect of covariates on incidence in the presence of competing risks [2,3,7]. In response, Fine and Gray [8] developed the subdistribution hazard model which allows one to model the effects of covariates on the cumulative incidence function in the presence of competing risks. It is increasingly being acknowledged that the subdistribution hazard model should be used when evaluating incidence of an outcome over time in the competing risks setting [9].

Peduzzi et al. published an influential series of articles examining the effect of the number of events per variable (EPV) on the accuracy of estimation of regression coefficients for the logistic regression model and for the Cox proportional hazards model in the absence of competing risks [10–12]. For a logistic regression model for use with binary outcomes, the number of events was defined to be the smaller of the number of events and the number of nonevents (or the smaller of the number of successes and the number of failures). For a Cox proportional hazards regression model, the number of events was defined as the number of subjects for whom an event was observed to occur (i.e., the number of noncensored subjects). Their studies used simulations based on 673 patients enrolled in a trial comparing medical and surgical management of coronary artery disease. Based on these simulations, they recommended that at least 10 EPV be observed to enable accurate estimation of the regression coefficients. These papers have been very influential, with the article on logistic regression being cited 1,610 times and the article on the Cox regression model being cited 527 times (source: Science Citation Index; Date accessed: June 16, 2016).

When analyzing survival data in which competing risks are present, there are multiple types of events: the primary event of interest (e.g., death due to cardiovascular causes) and the competing events (e.g., death due to noncardiovascular causes). The effects of the number of the different types of events on the accuracy of estimation of the coefficients of a subdistribution hazard model have not been explored. The results in Peduzzi et al. are not applicable owing in part to competing events and in part to a nonstandard weighting technique in the partial likelihood estimation procedure which addresses independent censoring [8]. Given the increasing use of the subdistribution hazard model for estimating the effect of covariates on incidence in the presence of competing events, the objective of the current paper is to examine the effect of the number of EPV on the accuracy of estimation of the coefficients of a subdistribution hazard model. The paper is structured as follows: in Section 2, we describe the design of a series of Monte Carlo simulations to examine the impact of the number of EPV on the accuracy of estimation of regression coefficients for a subdistribution hazard model. In Section 3, we report the results of these simulations. In Section 4, we summarize our findings, which differ somewhat from those in Peduzzi et al. and discuss them in the context of the existing literature.

2. Monte Carlo simulations—methods

In this section, we describe the design of a series of Monte Carlo simulations to examine the effects of the number of EPV on the accuracy of estimation of the coefficients of a subdistribution hazard model. In Section 2.1, we describe data on patients hospitalized with acute myocardial infarction (AMI or heart attack). In Section 2.2, we describe analyses that were conducted using these data to determine parameters for the data-generating process in the subsequent Monte Carlo simulations. In Section 2.3, we describe the data-generating process that was used to