Computational Statistics and Data Analysis 63 (2013) 16–30

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Performance of parametric survival models under non-random interval censoring: A simulation study

Nikos Pantazis [a,*,1], Michael G. Kenward [b,1], Giota Touloumi [a,1]

[a] Department of Hygiene, Epidemiology and Medical Statistics, Athens University Medical School, Greece

[b] Medical Statistics Department, London School of Hygiene and Tropical Medicine, United Kingdom

ARTICLE INFO

Article history:
Received 29 May 2012
Received in revised form 21 December 2012
Accepted 8 January 2013
Available online 29 January 2013

Keywords:
Interval censoring
Parametric survival model
Generalised Gamma
Informative censoring
HIV
Virologic response
Antiretroviral treatment

ABSTRACT

In many medical studies, individuals are seen periodically, at a set of pre-scheduled clinical visits. In such cases, when the outcome of interest is the occurrence of an event, the corresponding times are only known to fall within an interval, formed by the times of two consecutive visits. Such data are called interval censored. Most methods for the analysis of interval-censored event times are based on a simplified likelihood function which relies on the assumption that the only information provided by the censoring intervals is that they contain the actual event time (i.e. non-informative censoring). In this simulation study, the performance of parametric models for interval-censored data when individuals miss some of the pre-scheduled visits completely at random (MCAR), at random (MAR) or not at random (MNAR) was assessed comparing also with a simpler approach that is often used in practice. A sample of HIV-RNA measurements and baseline covariates of HIV-1 infected individuals from the CASCADE study is used for illustration in an analysis of the time between the initiation of antiretroviral treatment and viral load suppression to undetectable levels. Results suggest that parametric models based on flexible distributions (e.g. generalised Gamma) can fit such data reasonably well and are robust to irregular visit times caused by an MCAR or MAR mechanism. Violating the non-informative censoring assumption though, leads to biased estimators with the direction and the magnitude of the bias depending on the direction and the strength of the association between the probability of missing visits and the actual time-to-event. Finally, simplifying the data in order to use standard survival analysis techniques, can yield misleading results even when the censoring intervals depend only on a baseline covariate.

1. Introduction

In conventional time-to-event analyses it is assumed that the time between the study's origin and the onset of the event of interest is either exactly known or right censored (i.e. greater than the available follow up time). An array of methods is available for the analysis of such data such as the Kaplan–Meier estimator of the survival function, log-rank tests for comparisons between groups or various types of semi-parametric (e.g. proportional hazards Cox model) or parametric regression methods for modelling the hazard of the event or the survival time in terms of a set of covariates. However, in many cases the onset of the event cannot be immediately observed and the analyst knows only the interval of time within which the event occurred. For example, in many medical studies patients are seen at a set of pre-scheduled visits at the

* Correspondence to: Department of Hygiene, Epidemiology and Medical Statistics. University of Athens, Medical School, M. Asias 75, 115 27 Athens, Greece. Tel.: +30 210 7462088; fax: +30 210 7462205.
E-mail address: npantaz@med.uoa.gr (N. Pantazis).

[1] On behalf of CASCADE Collaboration in EuroCoord.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2013.01.014

clinic where the physician can determine if a condition is present or not. The last visit at which the condition was absent and the first at which the condition was present can be used to form a time interval within which the event of interest must have occurred, giving rise to the so-called interval-censored data.

More formally, if T_i is the random variable representing the time to the event of interest of the i-th ($i = 1, 2, \ldots, n$) individual and data are interval censored, instead of observing T_i, the observables are intervals $(L_i, R_i]$ such that $T_i \in (L_i, R_i]$. This definition allows also for exactly observed, right-censored and left-censored data for which $L_i = R_i$, $R_i = \infty$ and $L_i = 0$, respectively.

The motivation for this investigation stems from epidemiological studies on HIV infected individuals and more specifically from those focusing on the so-called virologic response to treatment. In such studies, the time origin is usually the initiation of "combination antiretroviral treatment" (cART; simultaneous administration of at least three anti-HIV drugs) and the event of interest is the suppression of the HIV viral load to levels which are below the threshold of detection of modern assays. Inference focuses usually on the estimation of the survival or cumulative incidence functions and between groups comparisons, often summarised by the estimated median time-to-virologic response and hazard ratios, respectively. However, HIV viral load is only periodically measured and the exact time T_i of achieving undetectability for the first time after cART initiation, is only known to lie between the time of the last measurement at which viral load was detectable (L_i) and the first one at which it was undetectable (R_i). A common approach, often used in this type of studies, is to "simplify" the data by assuming that the time of virologic response coincides with the time of the first undetectable measurement (i.e. $T_i = R_i$; will be referred to as the "rightpoint" method) and then analyse the data using standard survival analysis techniques (Althoff et al., 2010; Pence et al., 2007). There are also other variations of these data simplification techniques (e.g. imputing T_i using the middle of the $(L_i, R_i]$ interval; which will be referred to as the "midpoint" method) but they are rarely used in HIV research (Geretti et al., 2009).

In general, simplifying interval-censored data by single imputation methods and then applying standard survival analysis methods can yield biased results (Law and Brookmeyer, 1992; Odell et al., 1992; Dorey et al., 1993) and should be avoided, especially given the availability of methods and software appropriate for the analysis of interval-censored data. For example, non-parametric maximum likelihood estimation (Turnbull, 1976) (NPMLE) of the survival function can be performed using the `Icens` (Gentleman and Vandal, 2010) package in R (R Development Core Team, 2010) and semi-parametric proportional hazards Cox-like models can be fitted using the `intcox` (Henschel et al., 2009) package also in R. Spline-based distributional models and fully parametric models for interval-censored survival data can be fitted within Stata (StataCorp, 2011) by using the `stpm` (Royston, 2007) and `intcens` (Griffin, 2005) programs. Gomez et al. (2009) and Lesaffre et al. (2005) give excellent tutorials on methods and software for interval-censored data.

In most methods which have been proposed for the analysis of interval-censored data (including those mentioned earlier), inference is based on a simplified version of the likelihood function. More specifically, let $F_{L,R,T}$ be the joint distribution of the random variable T and the observables (L, R) with range $\{(l, r, t) : 0 \leq l < t \leq r < \infty\}$, $D = \{(l_i, r_i], i = 1, 2, \ldots, n\}$ the observable data set and $S(t) = \Pr(T > t)$ the survival function. Then, the contribution to the likelihood of the i-th individual with observed interval $(l_i, r_i]$ is given by $dF_{L,R}(l_i, r_i) = \Pr(L \in dl_i, R \in dr_i, T \in (l_i, r_i])$ thus the overall likelihood is given by:

$$L_0(S(\cdot)|D) = \prod_{i=1}^{n} dF_{L,R}(l_i, r_i) = \prod_{i=1}^{n} \Pr(L \in dl_i, R \in dr_i, T \in (l_i, r_i]). \tag{1}$$

However, if the censorship and survival processes are independent, as for example in a longitudinal study with periodic follow-up where the monitoring times are not influenced by T, the censoring process is called non-informative and can be ignored. In general, non-informative conditions imply that the only information provided by the censoring interval $(l, r]$ about the survival time t, is that the interval contains t (Self and Grossman, Biometrics) or alternatively that the observables (l, r) are not influenced by the specific value of t in $(l, r]$ (Gomez et al., 2004). In such cases the likelihood function can be simplified to:

$$L(S(\cdot)|D) = \prod_{i=1}^{n} \Pr(T \in (l_i, r_i]) = \prod_{i=1}^{n} [S(l_i) - S(r_i)]. \tag{2}$$

In medical studies though, it is common for study participants to skip or delay their pre-scheduled or suggested visits to the clinics. In such cases, the censorship process may be associated with observed or unobserved variables. For example in our HIV example mentioned earlier, individuals infected through intravenous drug use (IDU) may have a higher probability of missing visits compared to those infected through sex between men (MSM). Moreover, individuals with poorer compliance to their treatment (a covariate which is often not available or is inaccurately recorded) and hence lower probabilities of virologic response, may also have a higher tendency to skip or delay clinical visits. In the latter case, the censoring intervals will be associated with the actual and unobserved time-to-event and the non-informative conditions will no longer hold.

Motivated by this example, our objective was to assess the performance of various methods for the analysis of interval-censored data in cases where individuals are periodically monitored for the onset of a condition but they may miss some visits with the probability of missing visits being common for all study participants, depending on a baseline covariate or depending on their time-to-event. Initially, non-parametric and fully parametric methods are applied to real data from