Journal of Clinical Epidemiology 67 (2014) 1111−1120

Journal of Clinical Epidemiology

Missing information caused by death leads to bias in relative risk estimates

Nadine Binder[a,b,*], Martin Schumacher[b]

[a]Freiburg Center for Data Analysis and Modeling, University of Freiburg, Eckerstr. 1, 79104 Freiburg, Germany
[b]Center for Medical Biometry and Medical Informatics, Institute for Medical Biometry and Statistics, Medical Center − University of Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany

Accepted 12 May 2014; Published online 30 June 2014

Abstract

Objectives: In most clinical and epidemiologic studies, information on disease status is usually collected at regular follow-up visits. Often, this information can only be retrieved in individuals who are alive at follow-up, and studies frequently right censor individuals with missing information because of death in the analysis. Such ad hoc analyses can lead to seriously biased hazard ratio estimates of potential risk factors. We systematically investigate this bias.

Study Design and Setting: We illustrate under which conditions the bias can occur. Considering three numerical studies, we characterize the bias, its magnitude, and direction as well as its real-world relevance.

Results: Depending on the situation studied, the bias can be substantial and in both directions. It is mainly caused by differential mortality: if deaths without occurrence of the disease are more pronounced, the risk factor effect is overestimated. However, if the risk for dying after being diseased is prevailing, the effect is mostly underestimated and might even change signs.

Conclusion: The bias is a result of both, a too coarse follow-up and an ad hoc Cox analysis in which the data sample is restricted to the observed and known event history. This is especially relevant for studies in which a considerable number of death cases are expected.

Keywords: Bias (epidemiology); Cohort studies; Death-induced bias; Illness−death models; Risk factors; Survival analysis

1. Introduction

In most clinical and epidemiologic studies, information on disease status is usually collected at regular follow-up times. Often, this information can only be retrieved in individuals who are alive at follow-up but will be missing for those who died before, resulting in incompletely observed study data with missing information because of death. It has been emphasized that such problems are especially prominent when studying longitudinal health-related variables in elderly populations [1−3] and in end-of-life research [4] leading to the general advice that missingness because of death and missingness because of other reasons, for example, nonresponse should be distinguished in the statistical analysis [5].

When investigating risk factors on the progression of a specific disease, a Cox proportional hazards model on the hazard of interest is often used, in which individuals who died are right censored at the last follow-up visit. In this article, we investigate the following question: How reliable are hazard ratio estimates when performing such an ad hoc Cox analysis, in which the data set analyzed is restricted to individuals with observed and known event history.

The potential bias resulting from applying ad hoc methods is already known for a long time. In their textbook on clinical epidemiology, Fletcher and Fletcher [6] use the term ‘‘false cohort’’ to indicate that the cohort of survivors is treated as if it were the original ‘‘true cohort.’’ Saracci [7] illustrates on three cohort study examples survival-related biases due to specific selection of study participants. However, the situation we address here is not that the cohort is already set up incorrectly but in which the analysis is restricted to the surviving individuals *afterward*. We emphasize that in a comprehensive analysis, the death cases are as important as the illness cases and should be adequately accounted for by considering an illness-death multistate model [8]. Joly et al. [9] seem to be among the first who

Conflict of interest: None of the authors have conflict of interest.

Funding: This research was funded by the German Research Foundation (DFG SCHU 756/12-1). The funding source had no involvement in this manuscript.

* Corresponding author. Tel.: +49-761-2036662; fax: +49-761-2036680.

E-mail address: nadine@imbi.uni-freiburg.de (N. Binder).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2014.05.010

What is new?

- An ad hoc analysis of data with missing disease status information because of death that simply treats the death cases as right-censored observations can yield biased hazard ratio estimates of potential risk factors.
- The bias can be substantial and can be in both directions, depending on the follow-up design and the situation studied.
- Potential bias resulting from applying ad hoc methods is not sufficiently recognized. The considered bias arises in the situation in which the analysis is restricted to the surviving individuals *afterward*. It was so far only investigated in incidence estimates but not in relative risk estimates.
- Study investigators should treat observed death cases in cohort studies that primarily investigate a specific disease of interest with care.

systematically investigated the problem of missingness due to death in the framework of an illness-death model. However, the aim of Joly et al. was to correctly estimate the *incidence* of dementia (and not relative risk estimates of potential risk factors) in a cohort study based on regular follow-up times. Onset of dementia was interval censored and missing if an individual died during two scheduled follow-up times. Harezlak et al. [10] developed a similar approach, also considering a longitudinal dementia study.

This article aims to provide a broad characterization of the potential bias in hazard ratio estimates resulting from using ad hoc complete-case Cox analyses by using established multistate techniques in different settings and many examples. We start by describing the circumstances under which the bias can occur. We then perform three different numerical studies to characterize the nature of the bias: First, we systematically explore the amount and direction of the bias for different settings in a simulation study and compare the results to bias values obtained from an extension of an approximate formula by Joly et al. [9]. Second, we illustrate the real-world relevance of the bias by a simulation using the design and variables from the Indianapolis Dementia Study [10] and, third, a study on nephropathy in diabetic patients [11, p. 30f] with complete follow-up information and artificially induce missing information due to death. We conclude with a discussion.

1.1. A study example

We briefly illustrate the problem of missingness because of death in an investigation of the association of traffic-related air pollution and incident type 2 diabetes within the Study on the Influence of Air Pollution on Lung, Inflammation and Aging (SALIA) cohort [12]. The SALIA cohort study is based on consecutive cross-sectional surveys that were performed between 1985 and 1994 in a specific area of Germany. Of all 54- to 55-year-old women living in this study region, a total of 4,874 women agreed to participate. In 2006, the study participants received a follow-up questionnaire that included additional questions on diabetes. However, by that time, 585 participants (12%) were already deceased. The date of death could be retrieved from death registers, but there was no information available whether they had developed type 2 diabetes before. To assess the potential impact of air pollution on diabetes incidence, the authors decided to exclude individuals with missing information because of death from their main analysis (which is the same as right censoring at study start) and dealt with this problem in sensitivity analyses only.

2. Description of the bias

Whenever one deals with potential illness–death data as for example outlined in the aforementioned study example, it is most appropriate to consider it within the framework of illness–death modeling. In doing so, it is still not possible to retrieve the death cases, but it allows us to assess the potential existence or magnitude of the bias that might result from an ad hoc analysis. Details are provided in the following.

2.1. Illness–death data and statistical techniques

We consider studies in which individuals are observed over time with focus on a specific disease of interest and in which death cases can possibly occur. Multistate models have been successfully implemented to describe and analyze the development and/or progression of a specific disease, thereby accounting for death [8]. A simple but prominent example is the so-called three-state illness–death model with an initial state 0 (healthy and alive), an illness state 1, and a death state 2 (Fig. 1).

The arrows between the boxes represent the possible transitions between the states. They are directly associated with the hazard rates for each transition $\alpha_{01}(t)$, $\alpha_{02}(t)$, and $\alpha_{012}(t)$, that is, like usual hazard rates (ie, instantaneous risks) but marked with the specific state transition $0 \rightarrow 1$, $0 \rightarrow 2$, or $1 \rightarrow 2$. Assuming a nonhomogeneous Markov model, the three hazard rates completely define the event history process, that is, the progression of the disease and death over time.

Particular statistical methodology is available to adequately analyze event history data [11]. For instance, nonparametric statistics based on the hazard rates gain insight into the dynamics of the processes. Prominent examples that have gained much attention in literature are