Annals of Epidemiology xxx (2018) 1–6

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Ghost-time bias from imperfect mortality ascertainment in aging cohorts

Eric J. Jacobs, PhD [a,*], Christina C. Newton, MSPH [a], Ying Wang, PhD [a], Peter T. Campbell, PhD [a], W. Dana Flanders, MD, DSc [b], Susan M. Gapstur, PhD [a]

[a] *Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA*
[b] *Rollins School of Public Health, Emory University, Atlanta, GA*

A R T I C L E I N F O

Article history:
Received 20 September 2017
Accepted 11 June 2018
Available online xxx

Keywords:
Cohort studies
Bias
Mortality
Epidemiologic methods
Aging

A B S T R A C T

Purpose: Many cohort studies in the United States link with the National Death Index to detect deaths. Although linkage with National Death Index is relatively sensitive, some participant deaths will be missed. These participants continue to contribute person-time to the data set after their death, resulting in bias, which we refer to as ghost-time bias. We sought to evaluate the influence of ghost-time bias on mortality relative risk (RR) estimates.
Methods: Simulations were performed to determine the magnitude of ghost-time bias under a variety of plausible conditions.
Results: Our simulations demonstrate that ghost-time bias can be substantial, particularly among the elderly, where it can reverse the direction of the RR. For example, we conducted a simulation of a cohort of men beginning follow-up at age of 70 years, assuming 5% missed deaths and a true RR of 2.0. In this simulation, observed RRs were 1.89 during the year the cohort was aged 85 years, 1.60 during the year the cohort was aged 90 years, and 0.61 during the year the cohort was aged 95 years. We also provide results from actual cohort data that are consistent with ghost-time bias.
Conclusions: Ghost-time bias may meaningfully affect mortality RR estimates under conditions that can plausibly occur in aging cohorts.

Introduction

Many large cohort studies in the United States, for example, the Cancer Prevention Study II [1], the Multiethnic Cohort [2,3], and the National Health and Nutrition Examination Survey [4] have used computerized linkage with the National Death Index (NDI) as the primary method of follow-up for mortality outcomes. Linkage with NDI allows cohorts to follow up large numbers of study participants for mortality outcomes without the need for personal contact, potentially for several decades.

Linkage with NDI is generally considered to have good sensitivity for detecting deaths, although sensitivity is substantially higher when information on social security number (SSN) is available [5,6]. The largest analysis to date to examine the sensitivity of NDI included over 5000 known deaths in Cancer Prevention Study II participants [1]. In that study, sensitivity was estimated at 97% among participants who had provided a complete SSN and 87% among participants who had not. Reasons that contributed to missed deaths included incomplete information on SSN or date of birth, disagreement between the SSN provided by the participant and that listed on a death certificate, disagreement on birth month or year, the use of informal first names, and misspelled names [1]. In other studies, sensitivities have ranged from 93% to 97% among individuals who provided an SSN and from 88% to 96% among participants who did not [5,6].

From the perspective of a researcher analyzing cohort data, study participants whose deaths are missed by linkage with NDI are "immortal" within the data set. These immortal participants inappropriately contribute person-time accrued after their actual death date. In this report, we refer to person-time inappropriately accrued after a participant's death as "ghost-time" and we refer to bias in mortality relative risk (RR) estimates resulting from ghost-time as "ghost-time bias."

The authors have no conflicts of interest to disclose.

* Corresponding author. Behavioral and Epidemiology Research Group, American Cancer Society, National Home Office, 250 Williams St, NW, Atlanta, GA 30303. Tel.: +1-404-329-7916; fax: +1-404-327-6450.
E-mail address: Eric.Jacobs@cancer.org (E.J. Jacobs).

https://doi.org/10.1016/j.annepidem.2018.06.002

To our knowledge, the influence of ghost-time bias on mortality RR estimates has not been assessed, although one report documented that ghost-time resulted in overestimated longevity in an elderly study population in Ohio [7]. Ghost-time bias is likely to be negligible during the early years of follow-up of a cohort of young or middle-aged people because nearly all participants are truly alive and therefore only a very small proportion of observed person-time is actually ghost-time. However, ghost-time will inevitably account for a steadily increasing proportion of observed person-time as follow-up continues, and the proportion of participants who are truly alive steadily declines. Ghost-time bias may therefore be an increasingly important concern as cohorts of predominantly middle-aged adults established during the 1980s and 1990s age into their 80s and beyond.

Understanding how and when ghost-time bias can influence results of cohort studies may be useful for researchers who work with data accrued from older study participants. In this report, we describe how ghost-time bias occurs, simulate the magnitude of bias in mortality RR estimates under various conditions, provide an example of actual results consistent with ghost-time bias, and discuss strategies to reduce ghost-time bias.

Material and methods

Simulations

To evaluate ghost-time bias, we simulated observed RRs for mortality, over time, associated with an exposure with a given true RR. For ease of interpretation, we based our simulations on a hypothetical large cohort of people who enrolled at the same age and calendar year and were followed for mortality for many years through linkage with NDI. We assumed that being immortal at enrollment (not being linkable to NDI in the event of death) was unrelated to exposure status and that exposure status and the true RR did not change during cohort follow-up. We describe our calculations as "simulations" because they were based on hypothetical cohorts; we did not conduct Monte Carlo style simulations with repeated iterations.

The end result of our simulations was the observed RR in the cohort in each individual year of follow-up, meaning the RR that would be observed by a researcher conducting analyses without knowledge of which person-time in the data set was actually ghost-time. We refer to the observed RR in the ith year of follow-up as $RR^i_{(obs)}$.

We began each simulation by assigning values to the six "starting" variables shown in Table 1. The values assigned were chosen because they might plausibly occur in a contemporary aging cohort where mortality was ascertained through linkage with the NDI or through another linkage with similar sensitivity. We then used the assigned values from Table 1 to calculate the intermediate variables shown in Table 2. The first four variables shown in Table 2 designated as "N" denote the number of people in each of the four possible groups based on exposure status and "immortal" status (exposed immortals, unexposed immortals, exposed mortals, and unexposed mortals). The number of exposed immortals and unexposed immortals never changed during simulations because, being immortal, they never died in the data set. The number of mortals, however, declined due to deaths during each year of follow-up at rates determined by their exposure status, age, and sex.

The fifth variable in Table 2, E^i, the proportion of mortal participants who are exposed, can be calculated in a straightforward way from the four "N" variables. It should be noted that E^i is needed to "back calculate" the absolute risk of death among unexposed members of the hypothetical cohort (R^i_U) from published national mortality rates, as national mortality data are not available by exposure status. Once the mortality rate in the unexposed is calculated, the mortality rate in the exposed can be calculated by multiplying by the true RR (RR^T).

Next, we used the variables in Table 2 to calculate $RR^i_{(obs)}$, the end result of interest, using the steps shown below. First, we calculated the observed death rate in the exposed in year i (denoted as $R^i_{E\ (obs)}$) by dividing the observed deaths among the exposed ($D^i_{E\ (obs)}$) by the number of exposed people who were categorized as alive in the data set ($N^i_{E,M} + N_{E,I}$), as shown in Equation 1 below. It is important to note that the denominator in this equation ($N^i_{E,M} + N_{E,I}$) includes both people who are actually alive and people who are dead but whose death was missed by NDI linkage.

$$R^i_{E\ (obs)} = D^i_{E\ (obs)} \Big/ \left(N^i_{E,M} + N_{E,I}\right) \quad (1)$$

Then, we used a similar equation, Equation 2, to calculate the observed death rate in the unexposed (denoted as $R^i_{U\ (obs)}$).

$$R^i_{U\ (obs)} = D^i_{U\ (obs)} \Big/ \left(N^i_{U,M} + N_{U,I}\right) \quad (2)$$

Finally, using Equation 3, we calculated the observed RR ($RR^i_{(obs)}$) as the ratio of the observed death rates in the exposed and unexposed:

$$RR^i_{(obs)} = R^i_{E\ (obs)} \Big/ R^i_{U\ (obs)} \quad (3)$$

Analyses of data from Cancer Prevention Study II

In actual cohort data, in contrast to simulations, ghost-time bias cannot be precisely measured because the exact proportion of immortals at enrollment is unknown. However, we hypothesized that patterns of results consistent with ghost-time bias would be observed in actual data where participants reached advanced ages. We therefore analyzed 30 years of follow-up data (1982–2012) from men and women aged 60–74 years at enrollment into the Cancer Prevention Study II cohort (there were too few people aged 75 years and older at enrollment to analyze). As described in an earlier analysis of diabetes and all-cause and cause-specific mortality in this cohort [9], diabetes was self-reported at baseline in 1982. Exclusions and adjustment variables are the same as those used in the earlier analysis. However, this analysis includes an

Table 1
Assigned values of starting variables for ghost-time bias simulations

Variable	Definition	Assigned values
N	Number of participants in the cohort at study enrollment	100,000
A	Age at cohort enrollment	60, 70, or 80
E	Proportion exposed at enrollment	10%, 50%, or 90%
P	Proportion of participants whose death would not be identifiable through linkage with NDI due to insufficient or inaccurate information (i.e., immortals)	2.5%, 5%, or 10%
RR^T	True RR of exposure	0.5, 0.75, 1.0, 1.5 or 2.0
R^i	True mortality risk during the ith year of follow-up	Age- and sex-specific values from 2011 U.S. vital statistics [8]

Please cite this article in press as: Jacobs EJ, et al., Ghost-time bias from imperfect mortality ascertainment in aging cohorts, Annals of Epidemiology (2018), https://doi.org/10.1016/j.annepidem.2018.06.002