Commentary

Correction of Selection Bias in Survey Data: Is the Statistical Cure Worse Than the Bias?

James A. Hanley*

* Correspondence to Dr. James A. Hanley, McGill University, 1020, Pine Avenue West, Montréal, Québec, H3A 1A2, Canada
(e-mail: James.Hanley@McGill.ca).

Initially submitted June 20, 2016; accepted for publication December 20, 2016.

In previous articles in the *American Journal of Epidemiology* (*Am J Epidemiol.* 2013;177(5):431–442) and *American Journal of Public Health* (*Am J Public Health.* 2013;103(10):1895–1901), Masters et al. reported age-specific hazard ratios for the contrasts in mortality rates between obesity categories. They corrected the observed hazard ratios for selection bias caused by what they postulated was the nonrepresentativeness of the participants in the National Health Interview Study that increased with age, obesity, and ill health. However, it is possible that their regression approach to remove the alleged bias has not produced, and in general cannot produce, sensible hazard ratio estimates. First, one must consider how many nonparticipants there might have been in each category of obesity and of age at entry and how much higher the mortality rates would have to be in nonparticipants than in participants in these same categories. What plausible set of numerical values would convert the (“biased”) decreasing-with-age hazard ratios seen in the data into the (“unbiased”) increasing-with-age ratios that they computed? Can these values be encapsulated in (and can sensible values be recovered from) 1 additional internal variable in a regression model? Second, one must examine the age pattern of the hazard ratios that have been adjusted for selection. Without the correction, the hazard ratios are attenuated with increasing age. With it, the hazard ratios at older ages are considerably higher, but those at younger ages are well below 1. Third, one must test whether the regression approach suggested by Masters et al. would correct the nonrepresentativeness that increased with age and ill health that I introduced into real and hypothetical data sets. I found that the approach did not recover the hazard ratio patterns present in the unselected data sets: The corrections overshot the target at older ages and undershot it at lower ages.

age; effect modification; hazard ratios; sensitivity analysis

Abbreviations: *AJE*, *American Journal of Epidemiology*; *AJPH*, *American Journal of Public Health*; BMI, body mass index; NHIS, National Health Interview Study.

The Editors of the *American Journal of Public Health* (*AJPH*) and the *American Journal of Epidemiology* (*AJE*) asked me to comment on hazard ratio calculations in 2 articles by Masters et al. (1, 2). At issue are age-specific hazard ratios for the differences in mortality rates between obesity categories. The hazard ratio estimates in the *AJE* article (2) were derived using Cox regression models that allowed for different hazard ratios for different age intervals; those in the subsequent *AJPH* article (1) were obtained using smooth-in-age hazard ratio models. These estimates deserve scrutiny because they were derived from a statistical approach that

the authors claimed corrected the observed hazard ratios for selection bias (3), that is, a distortion caused by what they postulated was the nonrepresentativeness of the participants in the National Health Interview Study (NHIS) that increased with age, obesity, and ill health. In the *AJPH* article, the authors used these corrected hazard ratios to arrive at population attributable fractions (a secondary topic that is addressed in Web Appendices 1 and 2, available at <http://aje.oxfordjournals.org/>).

The way Masters et al. arrived at the age-specific hazard ratios was addressed in their response to a letter to the editor

Figure 1. Shown is an attempt to recover the true hazard ratio pattern after some data have been selectively removed. A) In a purely mathematical population, the true hazard ratios for exposed versus not exposed participants are approximately 1.1 at the lower end of the age category and 2.5 at the upper age (black line). From this population, 19 sample waves were simulated in which, increasingly with age at potential selection, exposed persons who did participate were less likely to die in the next 6 years than were their sampled exposed peers who did not participate. This attenuated the observed hazard ratio curve (red line). B) The selectivity that increases with age excludes some of the exposed persons who are so ill that they will die within 6 years. For example, the 6 red dots beginning at age 65 years (a potential age at entry) indicate what fractions of these frail 65-year-old individuals who die in the next 6 years were excluded from the survey wave. Applied to these selective data, the approach proposed by Masters et al. (A, blue line) was not able to recover the hazard pattern present in the unselected data.

in the *AJE* (4) and was also addressed in an unanswered letter to the editor in the *AJPH* (5). Web Appendix 3 and Web Figure 1 show my simple re-analyses of some of the NHIS data and the age-specific hazard ratios (which were very similar to those of Masters et al.) that I obtained using their correction for selection bias. However, it is possible that the way they removed the alleged bias has not produced, and in general cannot produce, sensible hazard ratio estimates.

Masters et al. postulated that the nonrepresentativeness of the NHIS participants, which increased with age and ill health, distorted the age-specific mortality gradients across the body mass index (BMI) categories: Participants who were older at entry and in the higher-risk BMI categories would have been healthier than their same-age counterparts who did not participate. Ideally, to directly and accurately adjust the observed age-specific hazard ratios for this discrepancy, one would 1) add appropriate (specific to age at entry) numbers of nonparticipants to each of the BMI categories, 2) assign them higher mortality rates than those seen in persons in the same category who did participate, and 3) calculate a new (higher) rate for each category.

It is difficult to see how estimates of these additional (age-at-entry-specific sets of) quantities implicated in the selection

bias can be extracted from the NHIS data simply by using a regression model. Masters et al. claim to have removed the bias merely by adding a centered version of age at entry as an effect modifier in their regression model and then (after fitting) setting it to that central value (see Web Appendix 3 for details). Epidemiologists often set a known confounder to a typical central value and use a statistical model to standardize the comparison of interest. In the present case, Masters et al. relied on a piece of information that was recorded only for the participants to somehow fill in (impute) some never-estimated number of participants who were selectively missing. This new statistical cure for selection bias appears to be too simple to be true. In this commentary, I focus on 3 concerns.

The first arises when one considers the plausible values to be used in the ideal adjustment described above; this sensitivity analysis would provide bounds for the correct age-specific hazard ratio estimates. For example, might one plausibly double the hazard ratios for those missing from the higher-risk BMI categories? It is unlikely that the majority of the 10%–15% who chose not to participate in the 19 NHIS waves (6) were in these older and higher-risk categories. Moreover, not all of the missing minority are necessarily at higher risk than those in the same category who did participate. With all of

these constraints, can an uncorrected hazard ratio of 1.2 at, for example, 80 years of age realistically become a corrected hazard ratio of 2.4?

A second and very concrete concern is the age pattern of the “adjusted-for-selection” hazard ratios shown in Figure 2 of the article by Masters et al. in the *AJE* (2) and Figure 2 of the article by Masters et al. in the *AJPH* (1). Without their correction, the hazard ratios are attenuated with age. With it, the hazard ratios at older ages are considerably higher but those at younger ages are well below 1. The media coverage of these articles missed an important implication for persons younger than 50 or even 55 years of age. All other things being equal, should life insurance premiums be lower for those in an obese BMI category than for those in the normal-weight category?

The third, and equally worrying, concern stems from 2 simulations in which the regression approach of Masters et al. did not fix the nonrepresentativeness that increased with age and ill health that I deliberately introduced. I introduced this selection bias into samples from known populations. In 1 scenario, the correct hazard ratios were higher at older ages, whereas in the other, they were lower at older ages. The simulations and results are described in detail in Web Appendix 4, and the R code is provided in Web Appendix 5. The simulation for hazard ratios that are higher at older ages, shown in black in Figure 1, is based on a purely mathematical population in which the true hazard ratios for exposed versus not exposed are approximately 1.1 at the lower end of the age category and 2.5 at the upper end. From it, I simulated 19 sample waves in which (increasingly with age at selection) exposed persons who did participate were less likely to die in the next 6 years than were their sampled exposed peers who did not participate, thereby attenuating the observed hazard ratio curve. Applied to these selective data, the approach proposed by Masters et al. was not able to recover the hazard ratio pattern present in the unselected data. The corrected age-specific hazard ratios overshot the target at older ages and undershot it at lower ages.

The second simulation was based on hazard ratio patterns seen in actual populations, in which the hazard ratio for cohort-type mortality rates in men versus women is typically greater than 2 at age 60 years, 1.5 at age 80 years, and near 1 at age 100 years (see Web Figures 2 and 3 and Web Appendix 4). Using the published cohort mortality rates for Finland, I simulated 19 sample waves in which, increasingly with age at selection, men who did participate were less likely to die in the next 5 years than were their sampled male peers who did not participate. Again, applied to these selective data, the approach of Masters et al. was not able to recover the “lower-at-older-ages” hazard ratio pattern present

in the full data (Web Figure 3 and Appendix 4). The greater the selectivity I introduced, the more the corrected age-specific male-to-female hazard ratios overshot the target at older ages and undershot it at lower ages.

The Web Appendices 1, 2, and 4 contains several implicit pleas to authors. I end with 2 explicit ones: 1) that new ways to directly correct heretofore uncorrectable biases first be tested on simulated data generated by known parameter values and 2) that reported model-based corrections not be so drastic that (as with the life insurance premiums) they seemingly correct the problem at one end of the age scale by creating one at the other end. I also make a plea to editors: Instead of asking authors to report what software—and what version—they used to prepare the data and derive the reported results, might they ensure instead that the computer code used is publicly available?

ACKNOWLEDGMENTS

Author affiliation: Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada (James A. Hanley).

This article was jointly published in the *American Journal of Epidemiology* (*Am J Epidemiol.* 2016;185(00):0000–0000) and the *American Journal of Public Health* (*Am J Pub Health.* 2016;107(4):503–505).

Conflict of interest: none declared.

REFERENCES

1. Masters RK, Reither EN, Powers DA, et al. The impact of obesity on US mortality levels: the importance of age and cohort factors in population estimates. *Am J Public Health.* 2013;103(10):1895–1901.
2. Masters RK, Powers DA, Link BG. Obesity and US mortality risk over the adult life course. *Am J Epidemiol.* 2013;177(5):431–442.
3. Chapter 11: Selection bias. In: Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, CA: Lifetime Learning Publications; 1982:194–219.
4. Masters RK, Powers DA, Link BG. The authors reply [letter]. *Am J Epidemiol.* 2014;179(4):530–532.
5. Wang Z, Liu M. Obesity-mortality association with age: wrong conclusion based on calculation error [letter]. *Am J Public Health.* 2014;104(7):e3–e4.
6. Galea S, Tracy M. Participation rates in epidemiologic studies. *Ann Epidemiol.* 2007;17(9):643–653.