HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

Epidemiology. 2018 July ; 29(4): 521–524. doi:10.1097/EDE.0000000000000849.

Survival-related selection bias in studies of racial health disparities: the importance of the target population and study design

Chanelle J. Howe[a] and **Whitney R. Robinson**[b,c]

[a]Centers for Epidemiology and Environmental Health, Department of Epidemiology, Brown University School of Public Health, Providence, Rhode Island [b]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina [c]Carolina Population Center, University of North Carolina, Chapel Hill, North Carolina

Abstract

The impact of survival-related selection bias has not always been discussed in relevant studies of racial health disparities. Moreover, the analytic approaches most frequently employed in the epidemiologic literature to minimize selection bias are difficult to implement appropriately in racial disparities research. This difficulty stems from the fact that frequently employed analytic techniques require that common causes of survival and the outcome are accurately measured. Unfortunately, such common causes are often unmeasured or poorly measured in racial health disparities studies. In the absence of accurate measures of the aforementioned common causes, redefining the target population or changing the study design represent useful approaches for reducing the extent of survival-related selection bias. To help researchers recognize and minimize survival-related selection bias in racial health disparities studies, we illustrate the aforementioned selection bias as well as how redefining the target population or changing the study design can be useful.

Keywords

Selection bias; survivor bias; racial disparities; study design; target population

INTRODUCTION

Survival-related selection bias can occur when studying the relationship between an exposure in early life and a health outcome later in life (1). For example, this selection bias can arise under the scenario where the early life exposure influences whether an individual

Corresponding author: Chanelle J. Howe, Centers for Epidemiology and Environmental Health, Department of Epidemiology, Brown University School of Public Health, 121 South Main Street, Providence, RI 02912 (Phone: 401-863-7406, chanelle_howe@brown.edu).

The authors report no conflicts of interest related to this research.

All data used to produce the results reported in this paper were generated via simulations. The simulation code is included in the eAppendix.

lives long enough to be a study participant and a common cause of survival and the outcome exists. Although race (defined at conception or birth) can be considered to be an early life exposure (2, 3), the aforementioned selection bias has not always been explicitly discussed in relevant applied studies of racial health disparities (4–7). This lack of discussion may in part be attributed to race being an ill-defined exposure (8) and less frequently conceived of as an early-life exposure.

Even if this survival-related selection bias is recognized in a study of racial health disparities, analytic approaches most often employed in the epidemiologic literature to minimize selection bias (e.g., inverse probability weighting (8–10)) require that the aforementioned common cause be accurately measured. However, common causes that might be most relevant to studying racial health disparities later in life may be unmeasured or poorly measured because they may also be early-life factors or are difficult to measure (e.g., childhood socioeconomic position) (11, 12). Even if such common causes are accurately measured, the appropriateness of using techniques such as inverse probability weighting to minimize selection bias related to death has been debated in the literature (8, 10, 13, 14) because death can be considered to be a competing risk/event. Given the vulnerability of analytic approaches such as inverse probability weighting to unmeasured or poorly measured common causes as well as debates surrounding their use in this setting, careful consideration of the target population and study design emerge as useful strategies for minimizing survival-related selection bias. To help researchers recognize and minimize survival-related selection bias in racial health disparities studies, we provide an example of the abovementioned selection bias and how redefining the target population or changing the study design can help.

EXAMPLE

Let the target population be defined as the population to whom inference is to be made. Furthermore, the study population is defined as a subset of the target population that is obtained by sampling from the target population and used to make inference about the target population. Now suppose that a researcher aims to study the effect of race (defined at birth) on infection with virus Z in the target population: Black and White residents of City X born in 1936. In 2016 the researcher conducts a cross-sectional study that enrolls all living Black and White City X residents born in 1936. The researcher assesses the Z status of enrollees. To focus discussion, we assume that migration and Z infections prior to or at birth in City X are non-existent/negligible.

The Figure shows a simplified causal diagram for the target and study population depicting the relationships among race, Z infection, and selection into the researcher's study. Specifically, the variable S denotes whether an individual is alive to be included in the researcher's study and is influenced by race given documented racial/ethnic differences in mortality (15, 16). Furthermore, the variable U represents a common cause of S and Z infection that was unmeasured in the study [e.g., neighborhood characteristics in childhood and adulthood (17–19)].

The investigator compares Z infection by racial group (i.e., Black versus White). Because the study population is a selected sample of the target population comprised of persons who are alive to enroll, a box around S denoting conditioning on the selected sample appears in the Figure. This conditioning means that the researcher's racial comparison will yield an estimate that may be subject to selection bias because S is a collider (8, 20). Thus, a racial difference in Z infection may be observed in the study although race does not influence Z infection in the target population. The potential for selection bias for the effect of race on Z infection was confirmed via simulations that are included in the Table.

Two approaches could be used to minimize the survival-related selection bias: (a) change the target population or (b) change the study design. Regarding the first approach, to choose a target population that successfully minimizes survival-related selection bias, the researcher should redefine the target population in ways that minimizes (a) the possibility of exclusions due to death or (b) the extent to which race or unmeasured determinants of the outcome influence whether someone is a member of the study or target population (2). For example, the researcher could redefine the target population to be: Black and White City X residents born in 1976. As confirmed by the simulations in the Table, using a younger target population lowers the likelihood that an eligible resident is excluded from the cross-sectional study population due to dying before enrollment and minimizes potential selection bias.

In contrast, if the target population were simply redefined to be the study population (i.e., Black and White City X residents who were born in 1936 and are living at enrollment in 2016), being a member of the study and target population now both require being alive when study enrollment occurs and inclusion in the study and target population will be influenced by race and U. Thus, an association between race and Z infection in the target and study population may occur even though race does not influence Z infection. Therefore, selection bias will not be minimized. If the researcher were to redefine the target population to be Black and White City X residents born in 1936 who would be alive at enrollment regardless of their race, then in this target population race would not influence survival to enrollment. An estimate obtained for this target population would be equivalent to the survivor average causal effect (21) and not be subject to selection bias. However, even when the target population is appropriately redefined, inferences based on this first approach may pertain to fewer people.

A second approach to minimize selection bias would be to change the study design so that the study population includes a more representative and in turn less selected sample of the target population. For example, the researcher could change the study design to a cohort study. Specifically, the researcher could use (a) birth registry data to include all Black and White City X residents born in 1936 in a cohort study population regardless of vital status in 2016 and (b) surveillance data on Z infections to capture diagnoses of Z infection that occur subsequent to birth by 2016 among cohort members.

However, changing the design may have limitations. For instance, the use of surveillance data would miss infections that occur among individuals who were never tested for the Z virus. Missed infections would potentially result in measurement bias. Furthermore, a cohort design would likely require analyzing the resulting data using a time-to-event framework so

that deaths that occur by 2016 and before Z infection are appropriately handled (14). A time-to-event analysis would in turn require estimating the time at which Z infection occurred, which may be difficult using surveillance data, as well as ascertaining death dates. However, as shown via simulations in the Table, if the researcher accurately estimated times to Z infection and death and applied a time-to-event framework that included simply censoring follow up at death and fitting a Cox model, bias is reduced likely because Z infections that occurred before death were no longer excluded from the study.

DISCUSSION

Studies of racial health disparities may be subject to selection bias due to the exclusion of persons who die subsequent to conception or birth before the outcome of interest occurs or can be assessed in the study. Gauging the potential for this survival-related selection bias requires careful consideration of the target population and study design. Like in the Z infection example, careful consideration can include using a causal diagram to reflect the target population and study design and in turn identify potential sources of selection bias. Such consideration can be implemented in the setting of many commonly employed epidemiologic study designs (e.g., cross-sectional, cohort, case–control) informed by prior work (8, 9, 20, 22). If the potential for selection bias exists, changing the target population or study design may lessen the possibility of such selection bias. However, findings from the new target population or study design may apply to fewer people or may be more subject to other biases (e.g., measurement).

Therefore, the advantages of altering the target population or study design should be weighed against the disadvantages. If researchers deem that the disadvantages outweigh the advantages (e.g., selection bias expected to be less impactful than measurement bias), they should at least explicitly acknowledge the potential for survival-related selection bias when studying racial health disparities. Part of this acknowledgement includes articulating the expected magnitude and direction of such selection bias when reporting results, if possible (2, 23). Performing simulations as done here (see eAppendix for code) and previously (23) can help predict the magnitude and direction of bias and inform whether to change the target population or study design.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sources of funding: Dr. Whitney Robinson was supported by the National Cancer Institute grant number K01 CA172717 and is grateful to the Carolina Population Center and its NIH Center grant (P2C HD050924) for general support.

The authors thank Drs. Tyler VanderWeele, Stephen Cole, Miguel Hernán, and Lauren Cain for helpful feedback on earlier drafts of this manuscript.

References

1. Rothman, KJ., Greenland, S., Lash, TL. Modern Epidemiology. 3. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
2. VanderWeele TJ, Robinson WR. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology. 2014; 25(4):473–484. doi:410.1097/EDE.0000000000000105. [PubMed: 24887159]
3. Howe CJ, Dulin-Keita A, Cole SR, et al. Evaluating the Population Impact on Racial/Ethnic Disparities in HIV in Adulthood of Intervening on Specific Targets: A Conceptual and Methodological Framework. Am J Epidemiol. 2018 Feb 1; 187(2):316–325. doi: 10.1093/aje/kwx247. [PubMed: 28992096]
4. Chatterjee R, Brancati FL, Shafi T, et al. Non-traditional risk factors are important contributors to the racial disparity in diabetes risk: the atherosclerosis risk in communities study. J Gen Intern Med. 2014; 29(2):290–297. [PubMed: 23943422]
5. Simoni JM, Huh D, Wilson IB, et al. Racial/Ethnic disparities in ART adherence in the United States: findings from the MACH14 study. J Acquir Immune Defic Syndr. 2012; 60(5):466–472. [PubMed: 22595873]
6. Anastos K, Schneider MF, Gange SJ, et al. The association of race, sociodemographic, and behavioral characteristics with response to highly active antiretroviral therapy in women. J Acquir Immune Defic Syndr. 2005; 39(5):537–544. [PubMed: 16044004]
7. Glasser SP, Judd S, Basile J, et al. Prehypertension, racial prevalence and its association with risk factors: Analysis of the REasons for Geographic And Racial Differences in Stroke (REGARDS) study. American journal of hypertension. 2011; 24(2):194–199. [PubMed: 20864944]
8. Hernán, MA., Robins, J. Causal Inference. Boca Raton: Chapman & Hall/CRC; 2018. Forthcoming
9. Howe CJ, Cole SR, Lau B, et al. Selection Bias Due to Loss to Follow Up in Cohort Studies. Epidemiology. 2016; 27(1):91–97. [PubMed: 26484424]
10. Tchetgen Tchetgen EJ, Glymour M, Shpitser I, et al. To weight or not to weight? On the relation between inverse-probability weighting and principal stratification for truncation by death. Epidemiology. 2012; 23(4):644–646. [PubMed: 22659551]
11. Howe CJ. Reducing HIV Racial/Ethnic Disparities: What's Good Data Got to Do with It? Epidemiology. 2017; 28(2):221–223. [PubMed: 27779496]
12. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research: concepts, methodologies, and guidelines. Annu Rev Public Health. 1997; 18:341–378. [PubMed: 9143723]
13. Chaix B, Evans D, Merlo J, et al. Commentary: Weighing up the dead and missing: reflections on inverse-probability weighting and principal stratification to address truncation by death. Epidemiology. 2012; 23(1):129–131. discussion 132–127. [PubMed: 22157307]
14. Cole SR, Lau B, Eron JJ, et al. Estimation of the standardized risk difference and ratio in a competing risks framework: application to injection drug use and progression to AIDS after initiation of antiretroviral therapy. Am J Epidemiol. 2015; 181(4):238–245. [PubMed: 24966220]
15. Harper S, Rushani D, Kaufman JS. Trends in the black-white life expectancy gap, 2003–2008. Jama. 2012; 307(21):2257–2259. [PubMed: 22706828]
16. Shiels MS, Chernyavskiy P, Anderson WF, et al. Trends in premature mortality in the USA by sex, race, and ethnicity from 1999 to 2014: an analysis of death certificate data. Lancet. 2017
17. Galea S, Tracy M, Hoggatt KJ, et al. Estimated deaths attributable to social factors in the United States. American journal of public health. 2011; 101(8):1456–1465. [PubMed: 21680937]
18. Latkin CA, German D, Vlahov D, et al. Neighborhoods and HIV: a social ecological approach to prevention and care. The American psychologist. 2013; 68(4):210–224. [PubMed: 23688089]
19. Warner TD, Giordano PC, Manning WD, et al. Everybody's Doin' It (Right?): Neighborhood Norms and Sexual Activity in Adolescence. Soc Sci Res. 2011; 40(6):1676–1690. [PubMed: 22427712]
20. Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15(5):615–625. [PubMed: 15308962]

21. Vanderweele TJ. Principal stratification--uses and limitations. The international journal of biostatistics. 2011; 7(1) pii: Article 28. doi: 10.2202/1557-4679.1329. Epub 2011 Jul 11.

22. Hernán MA. Invited Commentary: Selection Bias Without Colliders. Am J Epidemiol. 2017; 185(11):1048–1050. [PubMed: 28535177]

23. Mayeda ER, Tchetgen Tchetgen EJ, Power MC, et al. A Simulation Platform for Quantifying Survival Bias: An Application to Research on Determinants of Cognitive Decline. Am J Epidemiol. 2016; 184(5):378–387. [PubMed: 27578690]

24. Arias E, Heron M, Xu JQ. United States life tables, 2012. National vital statistics reports. 2016 Nov; 65(8):1–65.

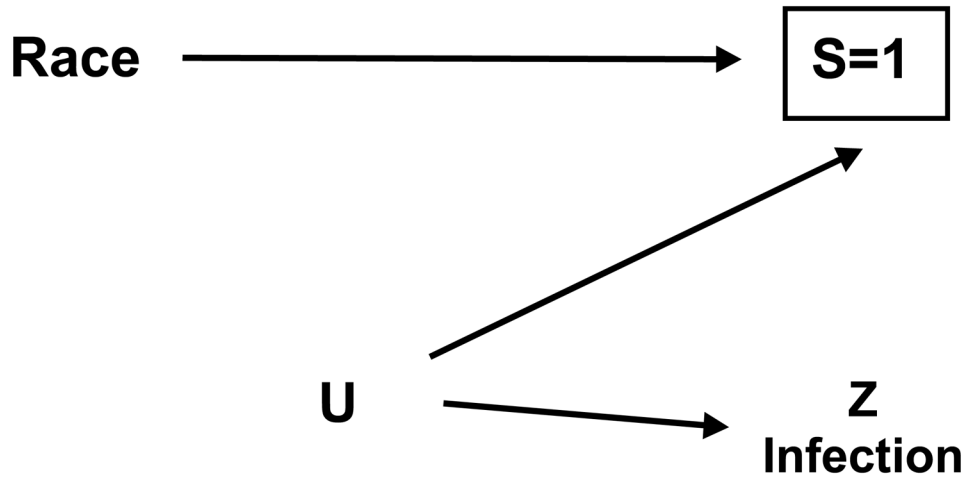

Figure.

Causal diagram for the target and study population in the researcher's original cross-sectional study that depicts the relationship between race and Z infection, where S=1 is an indicator of remaining alive in the target population to be included in the study population, U is an unmeasured factor, and a box denotes conditioning. In the researcher's original cross-sectional study, the target population is Black and White residents of City X born in 1936 and the study population is all Black and White City X residents who were born in 1936 and were living when study enrollment occurs in 2016. If the target population in the researcher's study were simply redefined (i.e., changed) to be the study population (i.e., Black and White City X residents who were born in 1936 and are living at enrollment in 2016), then S=1 is now an indicator of remaining alive in City X to be included in both the target population and study population.

Table

Simulation results summarizing the bias and mean squared error associated with various target and study populations when estimating the effect of race (defined at birth) on Z infection. Results based on 500 simulations of a target population of 10,000 individuals in the United States with specifications described in the eAppendix. The code that was used to generate these simulation results is included in the eAppendix.

Approach	Target population	Study population	Proportion of individuals in target population who are alive when researcher's original cross-sectional study would have begun enrollment[a]	True ln prevalence ratio or hazard ratio for effect of Black race on Z infection	Average ln prevalence ratio or ln hazard ratio observed in study population for effect of Black race on Z infection	Bias[b] observed in study population	Mean squared error[c] observed in study population
Researcher's original target population and cross-sectional study	Black and White residents of City X born in 1936	Black and White City X residents who were born in 1936 and living when study enrollment occurs in 2016	Approximately 0.19	ln prevalence ratio= 0	−0.06	−0.06	0.01
Change the target population without changing the cross-sectional study design	Black and White residents of City X born in 1976	Black and White City X residents who were born in 1976 and living when study enrollment occurs in 2016	Approximately 0.93	ln prevalence ratio= 0	−0.01	−0.01	0.00
Change the study design to a cohort study without changing the target population	Black and White residents of City X born in 1936	Black and White residents of City X born in 1936	Approximately 0.19	ln hazard ratio= 0	−0.01	−0.01	0.00

[a] Simulations conducted to be consistent with United States vital statistics (24) that indicate that between 18% and 23% of persons born in 1936 and between 93% and 95% of persons born in 1976 were alive at some point in 2016.

[b] Average of 500 differences between the estimate (e.g., ln prevalence ratio) and the true value

[c] Square of the bias plus the variance of the 500 estimates