Journal of the American Heart Association

VIEWPOINTS

On the Perils of Conducting Observational Research on Racial Health Disparities

Akshay Pendyal, MD, MHS

Racial disparities in cardiovascular health outcomes are well documented, as is the presence of unequal treatment of patients in racial and ethnic minority groups in the health care setting.[1,2] Frequently, these disparities are brought to light using large administrative claims data sets. Such data sets are powerful research tools, often containing clinical and demographic information on tens, if not hundreds of thousands, of individual patients. The use of administrative claims data sets in health services research is increasing.

An example of a project using one such data set (Optum Clinformatics, a large commercial insurance data set marketed for research purposes) was recently published in the *Journal of the American Heart Association* (*JAHA*), examining racial disparities in the treatment of non–ST-segment–elevation myocardial infarction.[3] The authors demonstrate that Black and Hispanic individuals are less likely to be treated according to an invasive strategy when compared with White individuals. These findings are noteworthy given the preponderance of evidence supporting an invasive strategy as the standard of care for patients presenting with non–ST-segment–elevation myocardial infarction,[4] as well as a steady increase in non–ST-segment–elevation myocardial infarction incidence over time.[5]

However, some caution is warranted when interpreting these results, as well as the results of any observational study attempting to document racial disparities using administrative claims data. The challenges associated with ascertaining race in large observational and administrative data sets are well-known.[6] In many cases, the methods used to derive race and ethnicity variables are unclear, and the variables themselves have been incompletely validated.[7] And such difficulties are arguably even more relevant in commercial (as opposed to publicly available, government-led) administrative claims data sets, given their proprietary nature.

Commercial administrative claims data sets often use automated algorithms to derive race and ethnicity data. Optum Clinformatics, for instance, previously used software known as E-Tech to construct its race and ethnicity variables, "incorporating racial and ethnic neighborhood composition from the US Census, residential zip code, and first and last name."[8] But such algorithms are prone to error. A study of the "diagnostic accuracy" of E-Tech in determining a subject's race noted that the software misclassified "52% of all [B]lack participants…as [W]hite," and, in doing so, systematically "underestimated [B]lack race."[9] Perhaps even more troubling, it appears that Optum may no longer be using E-Tech software, leaving researchers wanting to work with its data set with no published validation of the "proprietary algorithm" that Optum now uses to impute race and ethnicity data.[7]

Returning to the example I cite at the beginning of this piece, then, an algorithm that routinely underestimates Black race could significantly confound the relationship between race and appropriate treatment of non–ST-segment–elevation myocardial infarction. And with commercial administrative claims data sets, such as Optum Clinformatics, only increasing in popularity, when it comes to studying racial disparities, these

Key Words: administrative claims data ■ ethnicity ■ health disparities ■ observational research ■ race

The opinions expressed in this article are not necessarily those of the editors or of the American Heart Association.

Correspondence to: Akshay Pendyal, MD, MHS, University of North Carolina School of Medicine, Novant Health Charlotte Campus, Presbyterian Medical Tower, Suite 601, 1718 E 4th St, Charlotte, NC 28204. Email: akshay_pendyal@med.unc.edu

For Sources of Funding and Disclosures, see page 2.

JAHA is available at: www.ahajournals.org/journal/jaha

sorts of methodological limitations, which are considerable, should elicit certain questions; chief among them, what is the overall utility of documenting racial and ethnic health disparities in privately held data sets that, "representative" though they may be, remain inaccessible to most researchers and ascertain race via algorithms that rely on crude proxies, such as name and geography (proxies that, given rapidly changing demographic patterns in the United States,[10,11] may be outliving their usefulness)?

Indeed, many disparities researchers are already attuned to these difficulties, often providing certain caveats pertaining to observational findings, citing, for example, an inability "to rule out misclassification of demographics such as race, ethnicity, or…social risk factors," that "race and ethnicity are highly heterogeneous and may not be adequately captured by the categories used in [the] analysis," and that "administrative data are an indirect estimation, and as such…racial and ethnic groups may not be consistent with those identified as self-report."[3] But qualifications such as these are significant, and should do little to assuage the concerns of the reader; I would contend that they deserve more than passing mention in articles whose core conclusions often rely on models in which race and ethnicity are, in fact, incorporated as the primary exposure of interest.

The question of how race is "determined" in administrative claims data sets is a vexing one. Contrary to being a naturally existing category in any meaningful sense, race and ethnicity are, as is frequently acknowledged, "nuanced social constructs."[3] And although self-report is widely regarded as the most useful form of data collection on race in the health care setting,[12] a large body of social science literature suggests that, in an everyday sense, race is even more complex, depending, as I have written elsewhere, "to varying degrees in different contexts, on an individual's self-conception and on society's labeling of that individual."[13]

And if this more nuanced and contingent view of race is true, if race is not simply "out there," waiting to be discovered and categorized (or, for that matter, is not simply a function of how individuals themselves perceive, and therefore report, their own identity[14]), then what are computerized algorithms, which glean something called "race" from a collection of "ethnically unique" names and geographic areas, actually detecting? Clearly, there is a need to improve race and ethnicity collection and reporting practices in administrative claims data sets, especially commercially available ones; I would argue that methodological transparency is a necessary first step.

As health services researchers increasingly rely on nonpublic data sources in which the construction of key variables is undertaken by proprietary algorithms whose inner workings go largely unquestioned, clarification of concepts such as race (concepts that, historically, we have taken for granted as a priori facts) will become even more vital.

ARTICLE INFORMATION

Affiliation

University of North Carolina School of Medicine, Charlotte, NC

Source of Funding

None.

Disclosures

None.

REFERENCES

1. Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, de Ferranti S, Despres JP, Fullerton HJ, Howard VJ, et al. Heart disease and stroke statistics–2015 update: a report from the American Heart Association. *Circulation*. 2015;131:e29–e322. doi: 10.1161/CIR.0000000000000152
2. Smedley BD, Stith AY, Nelson AR. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. The National Academies Press; 2003.
3. Tertulien T, Broughton ST, Swabe G, Essien UR, Magnani JW. Association of race and ethnicity on the management of acute non–ST-segment elevation myocardial infarction. *J Am Heart Assoc.* 2022;11:e025758. doi: 10.1161/JAHA.121.025758
4. Amsterdam EA, Wenger NK, Brindis RG, Casey DE, Ganiats TG, Holmes DR, Jaffe AS, Jneid H, Kelly RF, Kontos MC, et al. 2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation.* 2014;130:2354–2394. doi: 10.1161/CIR.0000000000000133
5. Pendyal A, Rothenberg C, Scofi JE, Krumholz HM, Safdar B, Dreyer RP, Venkatesh AK. National trends in emergency department care processes for acute myocardial infarction in the United States, 2005 to 2015. *J Am Heart Assoc.* 2020;9:e017208. doi: 10.1161/JAHA.120.017208
6. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, Hripcsak G, Smith S, Tatonetti NP, Vawdrey DK. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc.* 2019;26:730–736. doi: 10.1093/jamia/ocz113
7. Nead KT, Hinkston CL, Wehner MR. Cautions when using race and ethnicity in administrative claims data sets. *JAMA Health Forum.* 2022;3:e221812. doi: 10.1001/jamahealthforum.2022.1812
8. Garfein J, Guhl EN, Swabe G, Sekikawa A, Barinas-Mitchell E, Forman DE, Magnani JW. Racial and ethnic differences in cardiac rehabilitation participation: effect modification by household income. *J Am Heart Assoc.* 2022;11:e025591. doi: 10.1161/JAHA.122.025591
9. DeFrank JT, Bowling JM, Rimer BK, Gierisch JM, Skinner CS. Triangulating differential nonresponse by race in a telephone survey. *Prev Chronic Dis.* 2007;4:A60.
10. Alba R, Beck B, Basaran SD. The rise of mixed parentage: a sociological and demographic phenomenon to be reckoned with. *Ann Am Acad Political Soc Sci.* 2018;677:26–38. doi: 10.1177/0002716218757656
11. Logan JR, Stults B. *"The Persistence of Segregation in the Metropolis: New Findings from the 2020 Census" Diversity and Disparities Project*, Brown University, 2021. https://s4.ad.brown.edu/Projects/Diversity. Accessed October 7, 2022.
12. Cooper RS, Nadkarni GN, Ogedegbe G. Race, ancestry, and reporting in medical journals. *JAMA.* 2018;320:1531–1532. doi: 10.1001/jama.2018.10960
13. Pendyal A. The price of metaphor: seeking conceptual clarity in racial disparities discourse. *Bioethics.* 2022;36:816–817. doi: 10.1111/bioe.13066
14. Maghbouleh N, Schachter A, Flores RD. Middle Eastern and North African Americans may not be perceived, nor perceive themselves, to be White. *Proc Natl Acad Sci USA.* 2022;119:e2117940119. doi: 10.1073/pnas.2117940119