Better—Not Just Bigger—Data Analytics

Brahmajee K. Nallamothu, MD, MPH

Although the phrase Big Data Analytics is relatively new, the practice of leveraging large data sources is not. Outcomes researchers are familiar with secondary data analyses of diverse data sources that include administrative claims, electronic health records, and clinical registries.[1] Among the most prominent examples is the National Inpatient Sample (NIS).

The NIS is constructed by sampling nationwide discharge records from acute care hospitals. It is part of the Healthcare Cost and Utilization Project and is the largest publicly available all-payer inpatient healthcare database in the United States, including annual use and cost data on >7 million hospital stays (see https://www.hcup-us.ahrq.gov/nisoverview.jsp). The NIS is available from the Agency for Healthcare Research and Quality at a minimal cost and supported with web-based tutorials, software, and user groups. Hundreds of articles using the NIS are published each year. I have used it in the past and suspect many of you have as well.

Yet the NIS, like most other secondary data sources, must be used with caution. In this issue, we are publishing an important Perspective focused on examples of potential misuse of the NIS. In this provocative piece, Khera and Krumholz[2] highlight 4 types of errors in published NIS studies: (1) not accounting for its complex sampling methodology; (2) ignoring the fact that it is limited to hospitalization data only; (3) incorrectly applying statewide and provider-level inferences to its results; and (4) using its diagnosis and procedure codes without sufficient validation.

This article emphasizes the critical role of journals for assessing and publishing secondary data analyses in general. It is appropriate, therefore, that this month we are also publishing a notable example of robust work from the NIS. Zaiean et al[3] used it to examine rates of acute heart failure hospitalizations across the United States—a question that is important for policy makers and providers. They found an overall decrease in heart failure hospitalizations but persistent disparities among black men and women. These are key observations, and it is hard to imagine a study like this being done without the NIS.

In that context, the article serves as a successful model of secondary data analyses: heart failure experts framed an important research question, obtained the NIS, performed high-quality analyses, and then submitted for publication a balanced and clear report of their findings.

The analyses by Zaiean et al[3] were particularly nuanced and detailed. For example, they appropriately accounted for shifts in the sampling strategy of the NIS over time and used recommended imputation procedures for handling missing race/ethnicity data to obtain stable population estimates. And we were pleased to see that the study avoided each of the 4 errors raised in the Perspective.

Based on this process, we worked with Khera and Krumholz[2] to create a simple checklist of good practices with the NIS. We applied it to the manuscript by Zaiean et al[3], and we will implement it moving forward with any future studies submitted to *Circulation: Cardiovascular Quality and Outcomes* that use that data source as part of our internal review. This is a starting point for us, and we hope other journals will consider doing the same. But what practical implications does this Perspective have beyond the NIS? In my mind: plenty.

In addition to their review of errors, Khera and Krumholz[2] also report that the number of published studies using the NIS grew by >500% during the past decade and approached ≈600 in 2016. This is not surprising to many of us who have seen a surge in Big Data Analytics and secondary data analyses in recent years across many disciplines. This growth is being fueled by a greater availability of public and private data sources that are now being combined with increasingly powerful and accessible analytic resources.[4] Anyone with a computer and a statistical program can now analyze data that may stimulate more science but also risks producing poor science. Many are already pushing back against Big Data Analytics and feel it is overhyped.[5] Just this past month, *The New Yorker* published a must-read piece irreverently entitled, "How to Call B.S. on Big Data: A Practical Guide."[6]

We certainly do not want to discourage use of the NIS, which has pioneered many important principles of open data sharing by creating this invaluable resource. Rather, the potential concerns raised by Khera and Krumholz[2] are fundamentally broader. They are applicable to other data sources, both public (eg, Medicare and Medicaid data) and proprietary (eg, OptumInsight, Marketscan, electronic health record data). It is not always easy to understand where and when those data sources may fall short. The concerns also will be relevant for how we debate policies around open data sharing from clinical trials and registries.

One might argue that it is up to the reviewers and editors to catch errors. This view ignores the reality of how the peer review process works or the scarce resources most journals often have at their disposal. Ideally, investigators should consider the strengths and limitations of data sources before data

From the Department of Internal Medicine, Michigan Integrated Center for Health Analytics and Medical Prediction, University of Michigan, Ann Arbor; and the Center for Clinical Management and Research, Ann Arbor VA Medical Center, MI.

Correspondence to Brahmajee K. Nallamothu, MD, MPH, University of Michigan Cardiovascular Center, CVC Cardiovascular Medicine - SPC 5869, 1500 E, Medical Center Dr, Ann Arbor, MI 48109-5869. E-mail bnallamo@med.umich.edu

(*Circ Cardiovasc Qual Outcomes.* 2017;10:e004019. DOI: 10.1161/CIRCOUTCOMES.117.004019.)

***Circ Cardiovasc Qual Outcomes* is available at http://circoutcomes.ahajournals.org**
DOI: 10.1161/CIRCOUTCOMES.117.004019

analysis, not during the manuscript review phase. They are better informed than reviewers about the nuances of their data sources. Investigators could be asked for their code, but that is not always feasible, and it is impossible to expect journals or reviewers to rerun every analysis even when data sources are open.

In the case of databases that are not publicly available, the challenge of ensuring the veracity of findings becomes even more arduous, especially because analytic tools become more complex. Other checklists created for observational studies (e.g. Strengthening the Reporting of Observational Studies in Epidemiology)[7] and secondary data analyses (eg, Reporting of Studies Conducted Using Observational Routinely-Collected Data)[8] attempt to capture these concerns, but these have been variably applied. Currently, *Circulation: Cardiovascular Quality and Outcomes* does not require those other checklists, but we are considering it. A checklist does not prevent all errors, but it could potentially help researchers avoid common mistakes. We welcome your input as we recognize the addition of more checklists may lead to barriers toward submission.

Finally, readers will recognize that although Khera and Krumholz[2] review examples, no specific articles are cited. They shared a list of flawed studies with us, and we reviewed these but chose along with the authors not to call out individuals. This may strike some as unusual. However, we felt that the purpose of their Perspective was to encourage better research practices and not to arbitrarily impugn the reputations of a handful of researchers when many others might have made similar mistakes. (In full disclosure and with complete humility, the first thing I did when I read their piece was look for my name on the author lists of the studies. We are all vulnerable to these errors.)

The NIS has been a valuable resource for outcomes research. It is a particularly striking example of open data sharing—a necessary idea whose time has finally come. We need to make sure that the NIS and other widely available data sources are used in a way that does not compromise our faith in the results of the work. Moving forward, we need to think more about Better Data Analytics— not just big ones.

Acknowledgments

The author acknowledges Patrick Lohier for his help with reviewing an earlier draft of this manuscript.

Sources of Funding

Dr Nallamothu receives funding from the Michigan Institute for Data Science and the American Heart Association Institute for Precision Cardiovascular Medicine.

Disclosures

None.

References

1. Nallamothu BK. Moving from big data to vital insights. *Circ Cardiovasc Qual Outcomes*. 2016;9:615. doi: 10.1161/CIRCOUTCOMES.116.003375.
2. Khera R, Krumholz HM. With great power comes great responsibility: "Big Data" research from the National Inpatient Sample. *Circ Cardiovasc Qual Outcomes*. 2017;10:e003486.
3. Zaiean B, Kominski GF, Ong MK, Mays VM, Brook RH, Fonarow GC. National differences in trends for heart failure hospitalizations by sex and race/ethnicity. *Circ Cardiovasc Qual Outcomes*. 2017;10:e003552.
4. Krumholz HM. The promise of big data: opportunities and challenges. *Circ Cardiovasc Qual Outcomes*. 2016;9:616–617. doi: 10.1161/CIRCOUTCOMES.116.003366.
5. Groeneveld PW, Rumsfeld JS. Can big data fulfill its promise? *Circ Cardiovasc Qual Outcomes*. 2016;9:679–682. doi: 10.1161/CIRCOUTCOMES.116.003097.
6. How to Call B.S. on Big Data: A Practical Guide. The New Yorker [Internet]. http://www.newyorker.com/tech/elements/how-to-call-bullshit-on-big-data-a-practical-guide. Accessed June 20, 2017.
7. Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ*. 2007;335:806–808.
8. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*. 2015;12:e1001885. doi: 10.1371/journal.pmed.1001885.

KEY WORDS: data accuracy ■ data mining ■ outcomes research ■ statistical data analysis

Better—Not Just Bigger—Data Analytics

Brahmajee K. Nallamothu

Circ Cardiovasc Qual Outcomes. 2017;10:
doi: 10.1161/CIRCOUTCOMES.117.004019
Circulation: Cardiovascular Quality and Outcomes is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1941-7705. Online ISSN: 1941-7713