Ann Surg Oncol (2019) 26:709–710
https://doi.org/10.1245/s10434-018-07130-9

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

EDITORIAL – HEALTH SERVICES RESEARCH AND GLOBAL ONCOLOGY

What's Lost in What's Missing: A Thoughtful Approach to Missing Data in the National Cancer Database

Daniel J. Boffa, MD

Yale University School of Medicine, New Haven, CT

It is the art and responsibility of observational researchers to distill the 'truest truth'. Interactions between variables and outcomes of interest are recognized to distort observed relationships, leading to a litany of corrective strategies (e.g. adjustment, matching, etc.). However, each step in an approach to mitigate bias may itself introduce variability that could impact conclusions. As a result, slight variations in analytic approach may allow the same dataset to support different, even opposing, conclusions. Missing data within a study set exemplifies a challenge whose solution could itself introduce bias. The study by Hoskin et al.[1] represents a classic example of missing data leading to a potentially misleading conclusion. The authors discovered that the 'clinical stage' variable was missing in half the breast cancer cases captured and submitted between 2004 and 2007, but was twice as likely to be missing in patients treated with surgery as their first treatment, as opposed to patients receiving chemotherapy then surgery. One could question the clinical relevance of their *specific* discovery as failure to catch this phenomenon in their data may not have impacted the field (the cases in question were more than 10 years old, it was a descriptive study, etc.). However, this would be missing the point. The importance of this study lies in the way in which missing data were discovered and the realization that missing data may not be an homogeneous problem.

Two years ago, the timeframe in question (2004–2007) was described to be problematic for the study of clinically staged patients in the NCDB,[2] yet little has changed. A cursory search of PubMed using the search terms 'clinical stage' and 'NCDB' identified 53 publications in the first 11 months of 2018, of which 45 (85%) included this problematic timeframe. Most of these studies excluded patients with missing data, and none appeared to have imputed missing staging data. Other than supporting the assertion that far more information is published than is possible to read (an admittedly ironic point to make in writing), the ongoing inclusion of clinically staged patients from the 2004–2007 timeframe highlights the challenges relating to sample size in observational research. Many tumors and oncologic scenarios are uncommon. Given its large size, the NCDB offers a tremendous perspective of less common cancers. That being said, researchers may be reluctant to restrict their population by years, tempting them to simply study the best of what is available (the non-missing data). However, as Hoskin et al.[1] discovered, the differential rates of missing data across patient and tumor strata may bias results. Ideally, investigators would employ a strategy for missing data that was maximally inclusive and minimally biased.

Hoskin et al.[1] present a nice summary of the accepted alternatives to handle missing data. While there is no absolute rule as to the proportion of data that could be imputed, the overall proportion imputed in the study by Hoskin et al. is within the reported range. However, as the authors note, the distribution of missing data is not uniform. For the earliest years of their study they are imputing more than half of the data, which may be less obviously acceptable to investigators and their audiences. Furthermore, one should consider the importance of the imputed variable to the specific outcomes of interest. For example, had the study instead examined survival, imputing clinical stage (which is so tightly linked to prognosis) for more than half the study population, may diminish the confidence that many readers would have in the results.

First Received: 2 December 2018;
Published Online: 2 January 2019

D. J. Boffa, MD
e-mail: daniel.boffa@yale.edu

Perhaps the most important message of the study by Hoskin et al. is the recognition that the prevalence of missing data is just the beginning, and that authors should look for associations between missing data rates and other variables. As an example of a potentially more problematic case, Rosen et al.[3] encountered discordant rates of missing data in a propensity-matched study comparing surgery versus stereotactic radiation in healthy patients with early-stage lung cancer. The variable ‘tumor grade’ was missing in 10% of patients. However, grade was missing for half of the radiation patients (likely because the biopsies of radiation patients were performed by fine-needle aspirations, which may not provide sufficient tissue to determine grade). Excluding these patients would cut the radiation cohort in half (which was not desirable). On the other hand, including grade as a variable in the propensity match, but allowing ‘missing’ as an option, is also problematic. Radiation patients who were missing grade (a common scenario) would have been matched with the few surgical patients missing grade (a highly unusual scenario), which could introduce bias (e.g. if less-experienced centers were leaving out grade). Ultimately grade was left out of the matching.

Unfortunately, there is not a universal statistical approach that would remedy all issues with missing data. On the other hand, there should be a common approach to *progressively* screen datasets for missing data, particularly as populations are stratified and subgroup analyses are performed. The outlined strategy in the paper by Hoskin et al. does this beautifully. Ultimately, attempts should be made to explain discordant rates of missing data in the context of patient care, as the explanation may help identify the most appropriate strategy to manage missing data.

REFERENCES

1. Hoskin TL, Boughey JC, Day CN, Habermann EB. Lessons learned regarding missing clinical stage in the national cancer database. *Ann Surg Oncol*. 2018. https://doi.org/10.1245/s10434-018-07128-3.
2. Boffa DJ, Rosen JE, Mallin K, et al. Using the national cancer database for outcomes research: a review. *JAMA Oncol.* 2017;3(12):1722–8.
3. Rosen JE, Salazar MC, Wang Z, et al. Lobectomy versus stereotactic body radiotherapy in healthy patients with stage I lung cancer. *J Thorac Cardiovasc Surg*. 2016;152(1):44–54.e49.