Ann Surg Oncol (2019) 26:739–745
https://doi.org/10.1245/s10434-018-07128-3

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

ORIGINAL ARTICLE – HEALTH SERVICES RESEARCH AND GLOBAL ONCOLOGY

Lessons Learned Regarding Missing Clinical Stage in the National Cancer Database

Tanya L. Hoskin, MS[1], Judy C. Boughey, MD[2], Courtney N. Day, BS[1], and Elizabeth B. Habermann, PhD[1,2,3]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN; [2]Department of Surgery, Mayo Clinic, Rochester, MN; [3]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN

ABSTRACT

Background. The National Cancer Database (NCDB) is a valuable resource for studying national cancer treatment patterns. However, data abstraction rules from 2004 to 2007 resulted in missing clinical stage for a high percentage of cases. We investigated how this missingness can bias results in breast cancer studies including patients treated with neoadjuvant chemotherapy (NAC).

Methods. The impact of missing clinical stage on the estimated percentage of breast cancers treated with NAC versus adjuvant chemotherapy (AC) was examined from 2004 to 2013. Trends in NAC use were presented, excluding those cases with missing clinical stage, and compared with trends after multiple imputation, performed using the chained equations approach with predictive mean matching.

Results. Clinical stage was missing for 56% of cases in 2004–2007, versus 12% in 2008–2013, and was missing more than twice as often for AC patients versus NAC patients (31% vs. 12% overall), with the largest difference occurring in 2004–2007 (60% vs. 27% missing). Because stage was more frequently missing in AC patients, excluding those missing clinical stage introduced bias when considering NAC versus AC trends. With multiple imputation, significant increases in NAC use between 2004 and 2013 were identified for each stage: I (2–5%), II (11–24%), and III (34–46%), in contrast to an analysis excluding those missing stage, which suggested little or no increase within any stage.

Conclusion. NCDB data abstraction rules from 2004 to 2007 resulted in missing clinical stage for > 50% of breast cancers, which may introduce substantial bias. Multiple imputation or exclusion of the years 2004–2007 should be considered to mitigate the problem of missing clinical stage in NCDB.

The National Cancer Database (NCDB) is a collaborative database from efforts of the American College of Surgeons and the American Cancer Society that includes hospital registry data from over 1500 Commission on Cancer (CoC)-accredited facilities. NCDB data represent over 70% of newly diagnosed cancer cases nationwide[1] and provide vast opportunities to investigate cancer treatment practices.[2] The NCDB was the source of more than 200 publications in 2016 alone.[3]

The NCDB began in 1988, but 2004 was the first year many relevant site-specific factors and collaborative staging elements were collected;[3] thus, most contemporary NCDB studies use the data from 2004 onward. As with any large registry, there are many caveats to using the data appropriately. The NCDB provides both an extensive data dictionary[4] and some analysis recommendations.[5] Boffa et al.[3] also published a review article focused on using the NCDB for research, and others have recently published recommendations for working with the NCDB and other large databases.[6–8] Despite the availability of these resources, users may overlook an important issue regarding clinical stage.

American Joint Committee on Cancer (AJCC) staging variables are among the most important information captured in the NCDB, and most analyses of NCDB data necessarily incorporate stage in some way, whether as part of inclusion criteria, for stratification, adjustment, or as a covariate of interest. NCDB captures both clinical (pretreatment) stage and pathologic stage based on final

First Received: 24 July 2018;
Published Online: 4 January 2019

T. L. Hoskin, MS
e-mail: Hoskin.tanya@mayo.edu

pathology from definitive surgery. For patients who undergo surgical resection as their first treatment, pathologic and clinical stage are similar, but for patients treated with neoadjuvant therapy, pathologic stage is post-treatment and thus clinical stage, as well as pathologic stage, is critically important.

The CoC made important changes to the rules for registrars regarding staging between 2004 and 2007.[4] For cases diagnosed from 1 January 2004 through 31 December 2007, the CoC required registrars to copy the staging elements from a standardized document, as recorded by the managing physician; deriving staging from other clinical notes was not allowed during these years. As a result, the proportion of patients with missing stage, particularly clinical stage fields,[3] was high until this rule was changed in 2008. For cases diagnosed since 1 January 2008, registrars are required to complete staging fields, using all information available in the patient record, even if the attending physician does not document a stage.[4] Therefore, missingness for staging variables in NCDB varies greatly by year of diagnosis, which may introduce substantial bias into analyses. Multiple imputation is well-established as a valid and appropriate statistical approach to deal with missing data bias[9–11] and may be useful to address the issue of missing clinical stage in NCDB.

The aim of this study was to examine the impact of missing clinical stage on trend over time in the use of neoadjuvant chemotherapy (NAC) versus adjuvant chemotherapy (AC) for breast cancer in order to raise awareness of the issue of missing clinical stage in NCDB and to evaluate some strategies to mitigate the problem.

METHODS

Patients from NCDB with invasive breast cancer diagnosed from 2004 to 2013 and treated with chemotherapy were included. However, patients with inflammatory breast cancer or stage IV disease were excluded, as were patients who did not undergo surgical treatment. Patients were classified as undergoing NAC if their chemotherapy start time was 30–365 days prior to the definitive surgery timing, and as AC if their chemotherapy start time was 1–365 days after surgery.

The frequency of missingness of clinical stage was assessed over time. Trends in the use of NAC were presented, limited to clinical TNM staging-complete cases (excluding cases with missing clinical stage), and compared with trends when all cases were included. Multiple imputation was then performed on the entire data set. The multiple imputation algorithm included clinical TNM stage and the outcome[12] of NAC versus AC, as well as age, sex, calendar year, histology, pathologic stage, any other available components of staging, estrogen receptor (ER) status, progesterone receptor (PR) status, and treatment variables, including mastectomy versus lumpectomy, radiation therapy, and hormone therapy. An interaction between pathologic stage and NAC versus AC was also included in the imputation model since pathologic stage would be a better representation of clinical stage in patients undergoing surgery before chemotherapy than in patients with NAC, a significant percentage of whom will downstage between clinical and pathologic staging. We followed the suggestions of Eisemann et al.,[13] who demonstrated that multiple imputation with a chained equations approach,[14] using predictive mean matching (PMM) to impute stage, showed good performance in a national cancer registry. Multiple imputation was performed using the mice package[15] for R (version 3.2.3) to create five multiply imputed datasets. Estimates of the proportion with NAC were calculated for each year and clinical stage stratum using PROC GEMOND, and then pooled across imputations using PROC MIANALYZE in SAS (version 9.4). Imputation was performed separately for cases staged under the AJCC 6th edition (2004–2009) versus the AJCC 7th edition (2010–2013).

RESULTS

A total of 492,242 cases were included, of which 140,796 (29%) were missing clinical stage overall. However, the percentage of cases missing clinical stage varied substantially by year of diagnosis, with 61% missing clinical stage in 2004, 59% in 2005, 56% in 2006, and 48% in 2007. This percentage dropped dramatically to 24% in 2008 and then to only 7–15% in 2009–2013, demonstrating the effect of the change in abstraction rules on missingness in clinical stage (Fig. 1).

AC patients were more than twice as likely to be missing clinical stage as NAC patients (31% vs. 12% overall, $p < 0.001$) with the largest absolute differences in missingness occurring in the years 2004–2007. In 2004, 65% of AC patients versus 31% of NAC patients were missing clinical stage, with years 2005 through 2007 showing similar differences, but by 2013 the percentage missing clinical stage was only 8% for AC versus 3% for NAC patients (Fig. 1).

To assess trends in the use of NAC, we first calculated the percentage of chemotherapy patients treated in the neoadjuvant setting for each calendar year among all cases (i.e. including cases without regard for whether clinical stage was missing). Next, we calculated the same percentage excluding those with missing clinical stage for comparison. When patients with missing clinical stage were excluded, the percentage treated with NAC was

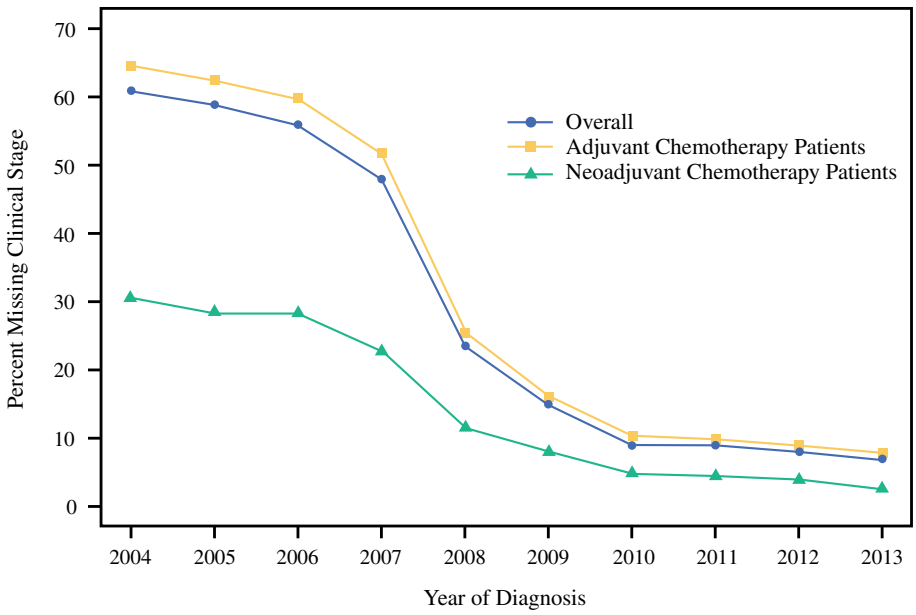

FIG. 1 Percentage of patients in the NCDB with chemotherapy-treated invasive breast cancer who were missing clinical TNM stage, both overall and separately for those treated with adjuvant versus neoadjuvant chemotherapy, in relation to year of diagnosis. *NCDB* National Cancer Database

overestimated because AC patients were more often excluded due to missing clinical stage. The magnitude of bias was directly related to the degree of missingness and was thus largest from 2004 to 2007 and much smaller from 2008 to 2013 (Fig. 2). When all cases were included, there was a clear increasing trend in NAC use, from 11% in 2004 to 19% in 2013.

Trends by stage also differed between an analysis excluding cases missing clinical stage versus using imputed clinical stage to include all cases (Fig. 3). Excluding those missing clinical stage, we concluded that the use of NAC increased modestly between 2004 and 2013 in stage I (2.5–4.8%) and stage II (19.9–25.1%), but did not increase in stage III (51.4–48.2%). These findings of little or no increase for any stage were inconsistent with the overall trend including all cases that the use of NAC increased from 11 to 19% (Fig. 2). Using the multiply imputed data instead, a clear increasing trend was observed over time, particularly within clinical stage II (10.8% in 2004 to 23.7% in 2013) and stage III (33.8% in 2004 to 46.5% in 2013) [each $p < 0.001$].

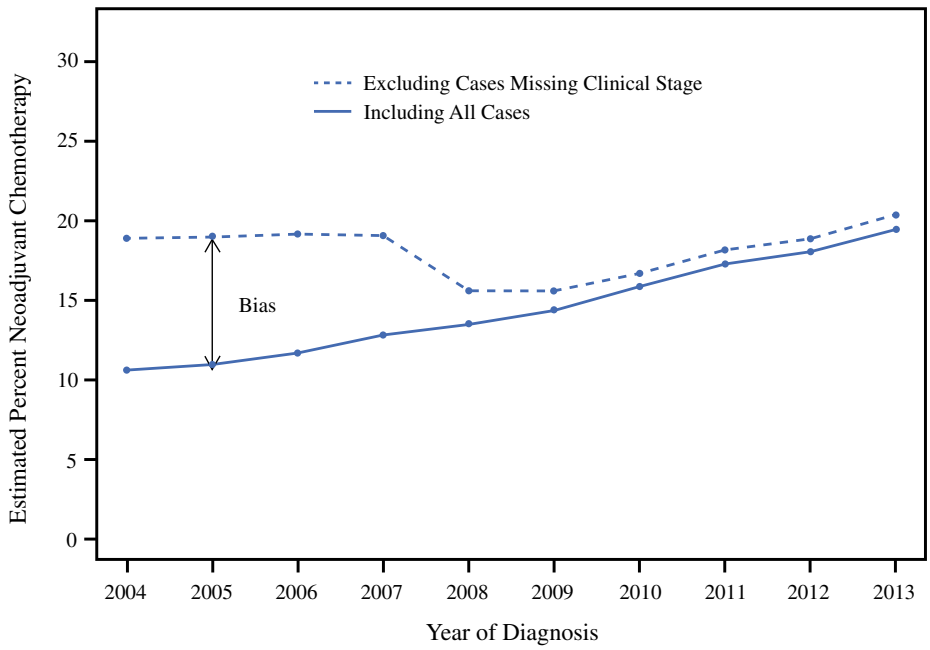

FIG. 2 Estimated percentage of stage I–III breast cancer patients undergoing neoadjuvant versus adjuvant chemotherapy over time, including all patients versus excluding those missing clinical stage

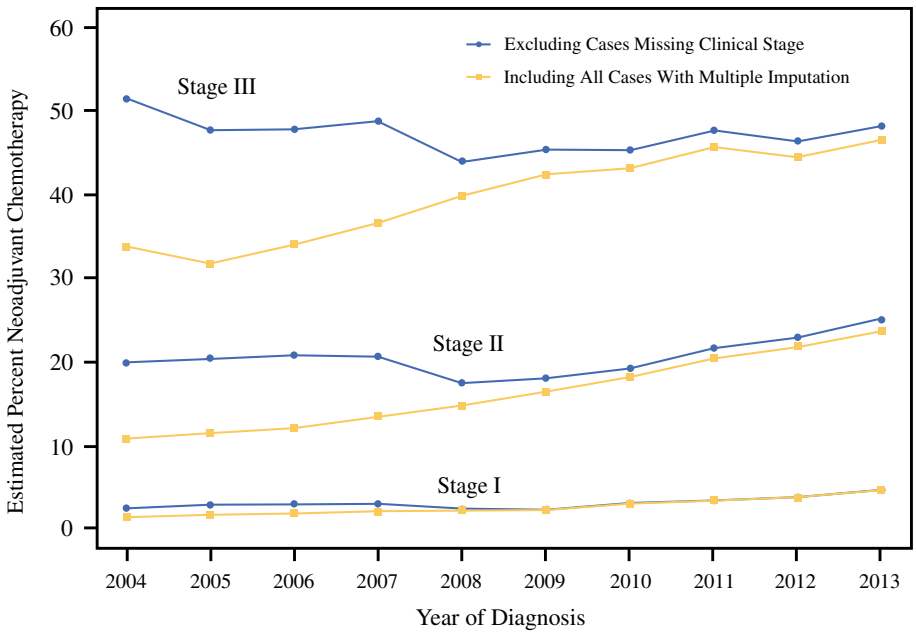

FIG. 3 Estimated percentage of patients undergoing neoadjuvant versus adjuvant chemotherapy by clinical stage, excluding patients missing clinical stage versus using multiply imputed clinical stage to include all cases

DISCUSSION

Clinical stage was missing in over 50% of patients with breast cancer from 2004 to 2007 in the NCDB. Excluding cases missing clinical stage disproportionately excluded patients treated with AC compared with NAC, which can result in bias in studies performed using this approach.

When using NCDB to study trends in the use of NAC, ignoring missingness in NCDB clinical stage substantively affected the study conclusion. This highlights the potentially large limitation of using only cases with non-missing clinical stage in NCDB from 2004 to 2007, years that have a large proportion of missing clinical stage due to the coding rules in that era. The problem was compounded when trying to compare across time, to years 2008 and later, when different coding rules resulted in much more complete clinical stage data, and by our study question since missingness was also strongly related to our outcome of whether the patient had NAC or AC. The finding that clinical stage was less likely to be missing in NAC patients was likely because clinical stage is more important for evaluating treatment and prognosis in neoadjuvant patients, and thus was more likely to be documented by the physician.

The problem of missing clinical stage affects every anatomic site; thus, all investigators analyzing NCDB data should take appropriate steps to identify and address missing data concerns in their analyses (we offer guidance in Table 1). First, carefully review documentation to identify potential areas of concern. As described in this commentary, all investigators should become familiar with the difference in NCDB AJCC stage coding rules during 2004–2007, versus 2008 and later,[4] and incorporate exploration of the degree of missingness over time and its impact on study conclusions in their particular analysis. Most variables in NCDB have some degree of missingness, but clinical stage deserves special attention because of its importance to many analyses, the high degree of missingness for one particular era (2004–2007), and the change in abstraction rules over time.

Many NCDB studies use particular clinical stages as inclusion criteria and thus exclude patients missing stage before analysis, which may result in an analysis biased in ways that cannot be explored without the larger dataset. One potential approach is to define the sample broadly initially and then impute clinical stage, at least for a sensitivity analysis. For example, rather than selecting breast cancers coded clinical stage I–III, one could instead include breast cancers with invasive behavior and then exclude those with distant metastasis. The second approach would result in a larger dataset, with the opportunity to explore the impact of missing clinical stage, than would the first approach.

When exploring missingness, one of the most important steps is to plot the percentage missing for each important variable as a function of calendar year. If a particular variable is missing for a high proportion of cases early in the time period, consider shortening the range of years included in the analysis. In addition to calendar year, consider other variables that may be associated with missingness (e.g. NAC vs. AC) and the potential impact on your study. If missingness is related to other observed variables, the data are not missing completely at random and complete case analysis is generally not appropriate;

TABLE 1 Recommended steps to approach missing data in the NCDB

Recommendation	Specifics
Carefully review the NCDB PUF Data Dictionary[4] and Supplemental Documents[5]	Be aware of the fact that there were different rules for registrars in collecting staging data from 2004 to 2007 versus 2008 onward, and consider the impact on your particular study. For example, the rules in place from 2004 to 2007 resulted in > 50% of cases missing clinical stage in the breast PUF
	Make a specific note of other variables that have not been collected for the entire study period or have been collected differently over time
	Determine when site-specific factors relevant to your project were required to be collected and only use cases from that year forward if the variable is important to your analysis
	Follow analysis recommendations contained within the Supplemental Documents[5], such as excluding cases with Class of Case = 00 (cases diagnosed at the reporting facility but not receiving any treatment there)
Define your sample broadly initially (including subjects who are missing data) to allow opportunity to explore missingness	For example, rather than restricting a sample to clinical stage I–III initially (which will exclude all those missing clinical stage), start with invasive cancers and exclude those clearly not meeting the criteria (e.g. stage IV, distant metastasis) from there
	Identify variables missing in a substantial (e.g. > 5–10%) percentage of cases without a reasonable explanation (e.g. pathologic variables missing in non-surgical cases)
Explore patterns of missingness	Plot the percentage missing versus calendar year for key variables. If important variables are missing early in the time period, consider whether the series timeframe could be shortened to use only more recent cases with more complete data
	In addition to calendar time, consider other variables that may be associated with missingness and the impact on your study. For example, we found missing clinical stage to be associated with whether the patient was treated with adjuvant versus neoadjuvant chemotherapy in the breast PUF
	If missingness is related to other variables, then the data are not missing *completely* at random but may still be missing at random (missingness that can be explained by observed variables); missing at random data may be addressed with multiple imputation
Consider missing data remedies	Limit the cohort timeframe to more recent cases with more complete data if feasible
	Consider not using variables with substantial missing data if they are not integral to the analysis
	Use sensitivity analyses to consider the impact of missing data on study conclusions (e.g. compare complete case analysis, deleting those with missing data, with approaches that do not exclude patients due to missing data)
	Use multiple imputation approaches, following recommendations such as those of Sterne et al.[10] in performing and reporting your multiple imputation analysis

NCDB National Cancer Database, *PUF* Participant User File

however, the data might still be considered to be missing at random, in which case multiple imputation using other observed variables to complete missing data may be a solution.[10]

Multiple imputation is well-established as an appropriate statistical approach to deal with missing data.[9–11] However, practically speaking, multiple imputation may not always be feasible. Most publications utilizing NCDB data do not use multiple imputation to deal with missing data, generally preferring simpler approaches, for example excluding cases missing key variables such as stage or using an unknown level in the analysis to retain cases with missing values. Such approaches would not have been effective in the situation presented here, and other authors have noted the inadequacies of these approaches in general.[16] There are also examples in the literature of a simpler

imputation approach using pathologic stage to fill in missing clinical stage;[17] this approach is problematic, particularly for studies trying to address neoadjuvant treatment questions. In general, it has been reported that multiple imputation is underutilized in medical literature and that imputation methods are often inadequately performed or reported.[18]

Sterne et al.[10] published a set of recommendations for conducting and reporting multiple imputation in clinical and epidemiologic data that provides a list of considerations and potential pitfalls for those considering multiple imputation, while Eisemann et al.[13] have made recommendations specifically for imputing stage in a national cancer registry. However, multiple imputation is not necessarily a simple or straightforward solution for a number of reasons: (1) it requires more analyst time and effort; (2) requires specification of an imputation method or model; (3) can have problems with convergence that must be monitored; (4) requires assumptions about the mechanism of missingness (i.e. whether the variable is missing at random); (5) must be carefully checked to make certain that it does not introduce new biases and problems into the analysis; and (6) may be of questionable validity when a very large proportion (e.g. $> 50\%$) of data are missing.

Since multiple imputation is complex and has its own limitations in situations with large volumes of missing data (such as clinical missing stage in NCDB in 2004–2007), another alternative is to consider the simpler approach of excluding the years 2004–2007 from studies using NCDB data if those years are not crucial to the question being investigated and if the question can be adequately addressed with the sample size available using cases from 2008 onward. This will lessen the problem to a great degree and, since the NCDB provides a large sample size, may be the easiest and most reliable solution. Table 1 summarizes the above recommendations to approach missing data in the NCDB.

This study has several limitations, including its narrow scope that dealt only with cancer at one anatomic site and the impact of missing clinical stage on a specific research question regarding trends in NAC use. However, the purpose of this report is to raise awareness for those who may be analyzing NCDB data, including cases from 2004 to 2007, and to suggest that they examine the missingness in clinical stage over time, as well as its effect on their particular analysis, and consider methods to reduce bias. Another limitation of this study was that HER2-neu status, which is known to be associated with NAC, was only added to NCDB as a site-specific factor in 2010. With this factor not available over the entire study period, we did not include it in the imputation. Finally, this study is limited by our use of a single method of multiple imputation rather than systematically evaluating the variety of methods available, yet such an investigation was beyond our scope and others have performed this work and found good performance for the methods used here.[13]

CONCLUSION

The NCDB data abstraction rules for AJCC staging variables were different for cases accessioned in 2004–2007, versus 2008 and later. As a result, clinical stage was missing for 56% of cases from 2004 to 2007, but only 12% of cases from 2008 to 2013, in the breast cancer dataset utilized here. Missingness was also strongly related to whether the patient was treated with NAC or AC, and thus introduced substantial bias when considering NAC versus AC trends. Multiple imputation or exclusion of the years 2004–2007 should be considered to mitigate the problem of missing clinical stage in NCDB.

ACKNOWLEDGMENT The NCDB is a joint project of the CoC of the American College of Surgeons and the American Cancer Society. The CoC's NCDB, and the hospitals participating in the CoC NCDB, are the source of the de-identified data used herein; they have not verified and are not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

FUNDING The Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery provides salary support for Dr. Habermann, Ms. Hoskin, and Ms. Day. No external funding was used.

REFERENCES

1. Raval MV, Bilimoria KY, Stewart AK, Bentrem DJ, Ko CY. Using the NCDB for cancer care improvement: an introduction to available quality assessment tools. *J Surg Oncol.* 2009;99(8):488–90.
2. Bilimoria KY, Stewart AK, Winchester DP, Ko CY. The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Ann Surg Oncol.* 2008;15(3):683–90.
3. 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.
4. National Cancer Database. Participant Use Data File (PUF) data dictionary. https://www.facs.org/~/media/files/quality%20programs/cancer/ncdb/puf_data_dictionary_puf_2015.ashx. Accessed 13 Aug 2018.
5. National Cancer Database. Participant Use Data File (PUF) supplemental documents. https://www.facs.org/~/media/files/quality%20programs/cancer/ncdb/puf_supplemental_documentation.ashx. Accessed 13 Aug 2018.
6. Merkow RP, Rademaker AW, Bilimoria KY. Practical guide to surgical data sets: National Cancer Database (NCDB). *JAMA Surg.* 2018;153(9):850–1.
7. Haider AH, Bilimoria KY, Kibbe MR. A checklist to elevate the science of surgical database research. *JAMA Surg.* 2018;153(6):505–7.

8. Kaji AH, Rademaker AW, Hyslop T. Tips for analyzing large data sets from the JAMA surgery statistical editors. *JAMA Surg.* 2018;153(6):508–9.
9. Little RJ, Rubin DB. Statistical analysis with missing data. Wiley, London; 2014.
10. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ.* 2009;338:b2393.
11. Horton NJ, Kleinman KP. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat.* 2007;61(1):79-90.
12. Moons KG, Donders RA, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol.* 2006;59(10):1092–101.
13. Eisemann N, Waldmann A, Katalinic A. Imputation of missing values of tumour stage in population-based cancer registration. *BMC Med Res Methodol.* 2011;11(1):129.
14. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med.* 2011;30(4):377–99.
15. Van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw.* 2011;45(3):1–67.
16. Knol MJ, Janssen KJ, Donders ART, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol.* 2010;63(7):728–36.
17. Mougalian SS, Soulos PR, Killelea BK, et al. Use of neoadjuvant chemotherapy for patients with stage I to III breast cancer in the United States. *Cancer.* 2015;121(15):2544–52.
18. Mackinnon A. The use and reporting of multiple imputation in medical research: a review. *J Intern Med.* 2010;268(6):586–93.