HHS Public Access

Author manuscript

Breast Cancer Res Treat. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

Breast Cancer Res Treat. 2023 January ; 197(1): 177–187. doi:10.1007/s10549-022-06764-4.

Implications of Missing Data on Reported Breast Cancer Mortality

Jennifer K. Plichta, MD, MS[1,2,3], **Christel N. Rushing, MS**[3,4], **Holly C. Lewis, MD, PhD**[1], **Marguerite M. Rooney, BS**[1], **Dan G. Blazer, MD**[1,3], **Samantha Thomas, MS**[3,4,5], **E. Shelley Hwang**[1,3], **Rachel A. Greenup, MD, MPH**[1,2,3]

[1)]Department of Surgery, Duke University Medical Center. Durham, NC.

[2)]Department of Population Health Sciences, Duke University Medical Center. Durham, NC.

[3)]Duke Cancer Institute. Durham, NC.

[4)]Biostatistics Shared Resource, Duke Cancer Institute. Durham, NC.

[5)]Duke University, Department of Biostatistics & Bioinformatics. Durham, NC.

Abstract

Background: National cancer registries are valuable tools to analyze patterns of care and clinical outcomes; yet, missing data may impact the accuracy and generalizability of these data. We sought to evaluate the association between missing data and overall survival (OS).

Methods: Using the NCDB (National Cancer Database) and SEER (Surveillance, Epidemiology, End Results Program), we assessed data missingness among patients diagnosed with invasive

Corresponding Author: Jennifer Plichta, MD, MS, FACS, DUMC 3513, Durham, NC 27710, jennifer.plichta@duke.edu, Phone: 919-681-9156, Fax: 919-660-8608.

AUTHOR CONTRIBUTIONS

- Jennifer K. Plichta: conceptualization, methodology, data analysis, writing (original draft, review, and editing), project administration
- Christel N. Rushing: methodology, resources, data curation, formal analysis, writing (review and editing)
- Holly C. Lewis: data review, writing (original draft, review, and editing)
- Marguerite M. Rooney: data review, writing (original draft, review, and editing)
- Dan G. Blazer: data analysis, writing (review, and editing)
- Samantha Thomas: data analysis, writing (review, and editing)
- E. Shelley Hwang: data analysis, writing (review, and editing)
- Rachel A. Greenup: conceptualization, methodology, resources, data analysis, writing (review, and editing), project administration

Presentation: Presented at the American Society of Clinical Oncology's Annual Meeting in June 2020.

DISCLOSURES

- The authors report no proprietary or commercial interest in any product mentioned or concept discussed in this article. The authors have no relevant financial or non-financial interests to disclose.
- Dr. J. Plichta is a recipient of research funding by the Color Foundation (PI: Plichta).
- The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

breast cancer from 2010–2014. Key variables included: demographic (age, race, ethnicity, insurance, education, income), tumor (grade, ER, PR, HER2, TNM stages), and treatment (surgery in both databases; chemotherapy and radiation in NCDB). OS was compared between those with and without missing data using Cox proportional hazards models.

Results: Overall, 775,996 patients in the NCDB and 263,016 in SEER were identified; missing at least 1 key variable occurred for 29% and 13%, respectively. Of those, the overwhelming majority (NCDB 80%; SEER 88%) were missing tumor variables. When compared to patients with complete data, missingness was associated with a greater risk of death: NCDB HR 1.23 (99% CI 1.21–1.25) and SEER HR 2.11 (99% CI 2.05–2.18). Patients with complete tumor data had higher unadjusted OS estimates than that of the entire sample: NCDB 82.7% vs 81.8% and SEER 83.5% vs 81.7% for 5-year OS.

Conclusions: Missingness of select variables is not uncommon within large national cancer registries and is associated with a worse OS. Exclusion of patients with missing variables may introduce unintended bias into analyses and result in findings that underestimate breast cancer mortality.

Keywords

breast cancer; data missingness; databases; cancer registry; survival; outcomes

INTRODUCTION

National cancer registries are increasingly utilized to analyze patterns of cancer care and clinical outcomes, resulting in numerous studies each year. However, within these registries, individuals with missing data may impact the accuracy and generalizability of study findings to a real-world breast cancer population. One of the largest registries of breast cancer patients in the United States (US) is the National Cancer Data Base (NCDB), which is managed by The American College of Surgeons (ACS) and the Commission on Cancer (CoC). The NCDB functions as a hospital-based registry, drawing from over 1,500 CoC-accredited clinical institutions across the US and includes select demographics, disease-related variables, treatment modalities, and oncologic outcomes; overall, the NCDB contains more than 34 million cases of cancer.[1] Furthermore, a recent analysis of the NCDB demonstrated inclusion of approximately 80% of incident breast cancer diagnoses in the US,[2] and prior research suggests >85% concordance with claims for chemotherapy and radiation receipt in breast cancer patients, although accuracy of endocrine therapy receipt was <65%.[3] In contrast to the nationwide hospital data in the NCDB, the Surveillance, Epidemiology, and End Results (SEER) program is a sample set, drawn from population-level registries for 19 geographic regions in the US. The National Cancer Institute (NCI) manages SEER, and it represents 35% of the US population, including over 9 million cases of cancer which are nationally-representative.[1]

Whether as hospital-based reporting or population-based sample sets, the accuracy and generalizability of research based on these cancer databases depends on the overall patient cohort, which variables of interest are included, and how often they are specified for each patient. Missing data in large-scale databases can bias conclusions toward null hypotheses,

particularly in smaller sub-group analyses, rare disease presentations, or non-participating centers.[4] In general, missing data can result in lower statistical power and biased estimates.[5, 6] Therefore, analyzing which patients are missing data and how the missingness may impact outcomes is of significant importance.

A recent study of breast cancer registries in Canada, the United Kingdom (UK), and Nordic countries showed high proportions of patients were missing staging data.[7] In the UK specifically, 25% of all patients were missing data on their clinical stage; furthermore, women with missing data had a 3-year survival that was 8.6% to 16.1% lower than in other countries studied.[7] Likewise, a study of colorectal cancer in these high-income countries showed that up to 50% of patients had no staging data; this was true for rectal cancers in the most populous Canadian province (Ontario) and Victoria (the second-most populous state of Australia).[8] As with breast cancer data missingness, the authors noted lower rates of colon and rectal cancer survival for patients who were missing staging data.[8] Considering the frequency with which patients are missing data, their exclusion may significantly bias the results from clinical outcome studies and the subsequent conclusions based on the study findings.

Given this prior research based on registries from other countries, we sought to investigate the data missingness in two widely used clinical oncology databases in the US, the NCDB and SEER database. Although some studies have previously sought to describe these databases and the completeness of data, we specifically aimed to analyze the potential association of data missingness on overall survival (OS) in the NCDB and SEER database.

METHODS

Patients diagnosed with invasive breast cancer in 2010–2014 were selected from the NCDB and SEER Database. Only those with histology codes defined by the World Health Organization (WHO) Classification of Tumors were included.[9] Those with “in situ” behavior codes were excluded.

Key variables in both databases were selected and sorted into three main categories: ***demographic*** (age, race, ethnicity, insurance, education, income), ***tumor*** (clinical TNM stages, ER, PR, HER2, grade), and ***treatment*** (NCDB: surgery, radiation, chemotherapy; SEER: surgery). In the SEER database, those who did not receive chemotherapy are included with those whose receipt of chemotherapy is unknown. The radiation summary variable is similarly coded in SEER. Therefore, these variables were excluded from our analysis of the SEER data. SEER high school education percentages and median household incomes were matched to the NCDB quartiles according to their cutoffs for the 2008–2012 version of the ACS. NCDB description of the TNM staging variables indicated that pathologic versions were not required to be recorded for diagnosis after 2008 and analysts should expect to see a decline in the availability of those variables as a result; as such, the clinical versions were used for this analysis. According to the SEER dictionary, the stage variables were derived from the CS coded fields.

Missing data indicators (complete vs missing data) were created at three levels: individual key variables, the main categories as described above, and overall. Patients with complete data for all 16 key variables in NCDB or all 14 key variables in SEER were categorized as having “complete” data. Total number of missing key variables are reported, as are the prevalence of missing data by category and over time.

Key variables and select characteristics were described using frequencies and percentages for all patients and by data missingness; p values were intentionally not reported for these comparisons, as they were not the focus of this study. Overall survival (OS) was defined as the time from diagnosis to death or last follow-up. The Kaplan-Meier method was used to estimate unadjusted OS curves. Single variable Proportional Hazards models were used to estimate the effect of data missingness on OS; hazard ratios (HRs) and 99% CIs are reported. Of note, a multivariate analysis was not possible, due to the requirement of excluding patients with missing values in these types of analyses; a multivariate analysis requires that all patients have a reported value for each adjusted variable. To evaluate the potential bias of complete data (listwise deletion) estimates, 5-year survival rates and 99% confidence intervals (CI) were calculated and are reported. Median follow-up was calculated using the reverse KM method.

Effective sample sizes are reported for each table/figure. No adjustments were made for multiple comparisons, and a p-value <0.01 was considered statistically significant. All statistical analyses were conducted using SAS, version 9.4 (SAS Institute, Cary NC) and R (R Core Team 2020). This study was deemed exempt by our institutional review board.

RESULTS

Overall, 775,996 patients in NCDB and 263,016 in SEER were identified (Table 1). Rate of missingness for at least one key variable was 29% and 13% in NCDB and SEER respectively, and the majority were missing only one variable (NCDB 17%; SEER 8%; Figure 1A). Tumor-related variables were missing most frequently (NCDB 23%; SEER 12%), while demographic and treatment variables were missing less often (Figure 1B). Among patients missing at least one key variable, this translated to 80% vs 88% missing a tumor variable in the NCDB and SEER, respectively. The proportion of patients with missing data decreased between 2010 and 2014 in both databases (NCDB 32% to 23%; SEER 15% to 12%; Figure 2). Because tumor information is often based on the findings from surgery, it is important to note that 13% of patients in NCDB with missing data were classified as having “no surgery of the primary site or autopsy only”, compared to 5.6% of those with no missing key variables. Similarly, 28.7% of patients in SEER with missing data were classified as having “no surgery of the primary site or autopsy only”, compared to 5.7% of those with complete data.

Median follow-up (IQR) was 51.2 months (35.4 – 68.2) in NCDB and 50 months (36 – 66) in SEER. For each database, 15% of the entire sample died by the time data was reported. Among those who were alive at that time, the maximum follow-up was 98.4 and 83 months, respectively. Five-year OS (99% CI) was estimated to be 81.8% (81.7–82.0) in the entire NCDB sample, but 82.7% (82.5–82.8) in patients with complete tumor data;

the corresponding estimates for the SEER sample were 81.7% (81.5–82.0) and 83.5% (83.3–83.8). When compared to patients with complete data, missingness was associated with a greater risk of death in both databases: NCDB 17% vs. 14% (HR 1.23, 99% CI 1.21–1.25) and SEER 27% vs 14% (HR 2.11, 99% CI 2.05–2.18; Figure 3A–B). Of note, death rate was not directly correlated with the extent of missingness and was similar whether the patient was missing 1 or ≥2 variables (Figure 3C–D). Regardless, the risk of death associated with missingness was greater for those in SEER than those in NCDB, even though more patients in NCDB were missing data; HRs (99% CI) for ≥2 vs 1 missing variable: NCDB 1.06 (1.03–1.09) and SEER 1.38 (1.31–1.46).

When stratified by the category of missing variable, differences in OS between those with and without missing data in the NCDB were small but consistently worse with HR estimates in the range of 1.04 to 1.27 (Figure 4A–D; Table 2). In SEER, reductions in OS were largest for those missing tumor variables (HR 2.26, 99% CI 2.19–2.33) or surgery data (HR 3.84, 99% CI 3.32–4.45; Figure 4E–H; Table 2). Among the tumor variables specifically (ER, PR, HER2, TNM), few clinically meaningful differences in OS were noted in the NCDB, and the largest differences were noted for those missing ER and PR status (ER missing: HR 1.60, 99% CI 1.52–1.68; PR missing: HR 1.55, 99% CI 1.48–1.63; Supplemental Figure 1; Table 2). In contrast, there were notable differences for most of the tumor variables in SEER, and the most significant differences were among those with missing T and N stage (Supplemental Figure 2; Table 2). Although information on the type of surgery was rarely missing (NCDB 0.2%, SEER 0.28%), missing this specific key variable was associated with a worse OS in both databases (NCDB: HR 2.11, 99% 1.83–2.44; SEER: HR 3.84, 99% CI 3.32–4.45). Of note, surgery was treated as a binary variable (missing or not missing), and as such, comparisons between the different types of breast surgery (lumpectomy vs mastectomy vs no surgery) were not evaluated. Regardless, our findings highlight the importance of complete data entry, such that exclusion of patients with missing data does not skew study results.

DISCUSSION

Our analysis of the NCDB and SEER breast cancer data corroborates findings from others, [10] demonstrating that missingness of key variables is a frequent occurrence within large national cancer registries. Our study found that the degree of missingness decreased over time, and that notably, missing data was associated with worse OS. We hypothesized that missingness may not be random and may introduce unintended bias into analyses of these data sets; furthermore, exclusion of patients with incomplete data may underestimate breast cancer mortality. Our results confirm this as 5-year OS was higher in patients with complete data.

Although the NCDB is nationwide in scope and contains a larger number of patients,[2] it is a hospital-based registry that draws only from clinical sites certified by the ACS CoC. Though the NCDB contains approximately 80% of all breast cancer diagnoses, CoC-accredited hospitals make up only ~30% of all hospitals in the US.[11] When compared to non-accredited hospitals, CoC-accredited hospitals were found to be larger, have more cancer-related services available, perform more operations, and more frequently be located

in urban settings, while also being less likely to be critical access hospitals or located in rural areas.[12] While accreditation is often thought to be associated with better outcomes, studies examining this have yielded mixed results [13–16]. In contrast, SEER is population based; though it draws from fewer geographic regions, yielding a smaller number of patients, its data comes from strategically selected cancer registries in an effort to be more representative of the US, and it includes all patients in that geographic area with a given diagnosis regardless of where they received care.[17] By providing data that is population-based, incidence can be ascertained, and therefore SEER may be more representative of the US population as a whole, though demographic differences do exist and should be taken into account when using this data.[18] Even with these differences, studies examining both SEER and NCDB data have found generally similar findings in certain disease types, such as breast and head and neck cancer.[1, 19]

Regardless of data provenance, we identified significant missingness in both databases with 29% of all NCDB patients and 13% of all SEER patients missing at least one key variable. Furthermore, we demonstrate that missingness of select variables is often associated with a worse survival. While the NCDB and SEER derive their data from different sources and vary in the manner in which they archive clinical outcomes, we observed similar trends in both databases. Although data presence does not directly contribute to causality for survival, the lack of data completeness may suggest a cohort of breast cancer patients worth additional attention. Missingness not only has implications for the accuracy and/or generalizability of study findings, but, more importantly, may also reflect a vulnerable population of breast cancer patients that lacks resources or access to complex multidisciplinary care at participating centers. Exclusion of these individuals from observational studies risks further exacerbation of existing cancer disparities, by neglecting the treatment experiences and cancer outcomes of our most vulnerable patients. While our study sheds light on an important limitation of these data sets, it also raises more questions about *who* is missing data. Unfortunately, demographic data is limited in NCDB and SEER, and some of the data provided is not patient-specific, such as the income and education data. Based on the available data, demographic data was only missing in 6.1% of NCDB cases and 2.6% of SEER cases, with markedly low rates of missingness for age and race specifically (<1% for both, in both databases).

Although missing tumor variables can be intrinsic to large datasets, the reasons behind missingness may be multifactorial. In general, missingness may be related to data collection/input problems, incomplete documentation by providers, non-standard care, or patient factors. For example, missingness may represent non-operative patients whose tumors were never biopsied or removed. For patients with advanced or metastatic malignancies diagnosed by imaging alone, biopsies may not alter clinical decision-making[20] and tissue-diagnosis may only be available at autopsy, if at all. In addition to advanced and metastatic cancers, other populations may also be less likely to undergo tissue-based diagnostic staging due to non-standard care or patient refusal. In a recent study of SEER patients with stage I and II lung, prostate, breast, and colon cancer (diagnosed 2007–2014), not undergoing surgery was associated with increasing age, non-Hispanic Black race/ethnicity, uninsured status, marital status, and stage.[21] Furthermore, not undergoing surgery in this study was associated with an increased risk of death, suggesting that more vulnerable populations may be at

an increased risk of not receiving the recommended surgical care, potentially resulting in worse outcomes.[21] Similarly, a study of Australian patients with prostate cancer noted that nearly 33% of patients had no clinical staging data.[22] Exclusion of these patients estimated higher mortality rates for patients from lower socioeconomic status (SES) or rural areas when compared to nationwide averages.[22]

In our study, 13% of NCDB patients and 29% of SEER patients with missing variables were due to non-operative status. For patients with missing surgery data specifically, survival outcomes were notably worse for both NCDB and SEER patients. This may suggest that despite their cancer diagnoses, such patients may have been lost to follow-up, declined standard treatment (such as surgery), did not receive appropriate medical advice, or were appropriately managed in the context of overall poor health or patient choice. Older studies suggest a low rate of overall treatment refusal, <1% in some populations,[23] while other studies in the US suggest potentially higher rates of refusal, up to 7.2% in a study of women ≥50y with late-stage breast cancer[24]. Although the NCDB and SEER databases used for this study do not consistently provide details on why certain treatments were or were not performed/received, it is possible that additional patient and provider education aimed at improving communication may help minimize the number of the patients receiving non-standard care.

In oncology, missing variables are of particular importance, as they may represent prognostic factors that serve as the basis for inclusion/exclusion in research studies and clinical trials. For many studies, complete case analyses are performed, which exclude patients with missing data and analyze only those with complete datasets. Although this can be an adequate statistical technique for some situations, it is known to result in lower statistical power and biased estimates in others. For example, when analyzing the health impact via OS in our analysis, patients with any missing data were observed to have a greater risk of death. Therefore, complete case analysis resulted in an overestimation of 5-year OS in both databases confirming that excluding patients with missing data from analyses, which is common practice, creates biased estimates of survival. Moreover, the impact of that bias is likely unpredictable (whether it equally affects all study groups, potentially some more than others, or some in a different direction than others). Lastly, cancer treatment centers participating in national registry collection may inherently be biased in their commitment to guideline-concordant oncology care. Differences in cancer outcomes between registry patients comprehensively captured in the NCDB and SEER, without missing data, and individuals receiving cancer treatment in many non-participating centers are unknown.

While the problem of missingness is pervasive in studies based on NCDB and SEER populations, statistical strategies have been developed to address this significant problem. As bioinformatics has evolved, democratized programming software has enabled researchers to analyze large datasets despite missingness, using tools such as imputation or complete case analysis. Imputation is the replacement of a missing variable with a new value that is mathematically carried forward as though it were a true observation. Multiple imputation by chained equations (MICE) is a method whereby missing variables are imputed by regression with other observed values.[25, 26] MICE has been used for imputation of missing data in

breast, kidney, and hepatopancreaticobiliary cancer registries, to varying effects.[4, 27, 28] One study of breast cancer patients in the NCDB analyzed a three year period in which 56% of patients were missing data on clinical stage, due to a change in reporting procedures (that has since been revised).[27] Using MICE, researchers analyzed the data despite missingness to demonstrate that neoadjuvant chemotherapy was being increasingly used for stage II and III breast cancers. Such a study demonstrates the value of MICE, to fill in missing holes in an otherwise robust database. Imputation of missingness for OS analyses can be more challenging to parse, as seen in a study of cytoreductive nephrectomy for patients missing histological tumor grade.[28],[29] In a SEER-based analysis with imputation, Egleston *et al* showed surgery was protective, but the degree to which it was therapeutic depended on the nature of *a priori* statistical assumptions about missingness. Using a sensitivity analysis, they demonstrated that whether grade was missing at random or non-randomly was a significant modifier; perceived surgical benefit may be due to confounding by grade rather than a true surgical effect. Such confounding and effect modification can significantly impact the quality of OS estimates, particularly in clinical databases with high missingness. As such, it is critically important for researchers to carefully consider the most appropriate inclusion/exclusion criteria (i.e. based on missing data) and variables to analyze for studies that are conducted with data from large national cancer registries, and to consider statistical methods that may limit the impact of data missingness on outcomes.

While some missingness is unavoidable, minimizing avoidable missingness would improve the quality of the datasets and decrease the need for such statistical methods (though it is not clear how much missingness is from such sources). Both SEER and NCDB dedicate substantial resources to data collection, optimization, and quality. In regards to SEER, NCI staff work with the North American Association of Central Cancer Registries (NAACCR) and other collaborating organizations to set standards to guide state registries in the data collection process.[30] In contrast, NCDB is jointly sponsored by ACS and the American Cancer Society; data is submitted by CoC-accredited hospitals yearly with set standards and conditions as part of maintaining accreditation status. [31] Though literature is limited, described strategies to minimize missing data center on multimodal preventative strategies, including clear documentation, adequate study personnel, adequate training of study personnel, and timely data entry, among others.[32, 33] Continuing to emphasize preventative strategies and adhere to rigorous reporting standards is critical to minimize avoidable missingness.

In conclusion, our analysis shows that a significant number of patients with breast cancer in the NCDB and SEER are missing data, and, importantly, that missingness of select variables was associated with worse OS. When interpreting OS studies with frequently missing data, researchers and clinicians must remain judicious in selecting inclusion/exclusion criteria as well as potential explanatory variables for outcomes studies to avoid bias and improve accuracy and/or generalizability. Future research is needed to elucidate which patients are most often missing data and why survival differences may be observed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The National Cancer Data Base (NCDB) is a joint project of the Commission on Cancer (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 SOURCES

- This work was in part supported by Duke Cancer Institute through NIH grant P30CA014236 (PI: Kastan) for the Biostatistics Core.

DATA AVAILABILITY

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

REFERENCES

1. Janz TA, Graboyes EM, Nguyen SA, Ellis MA, Neskey DM, Harruff EE, Lentsch EJ: A Comparison of the NCDB and SEER Database for Research Involving Head and Neck Cancer. Otolaryngol Head Neck Surg 2019, 160(2):284–294. [PubMed: 30129822]
2. Mallin K, Browner A, Palis B, Gay G, McCabe R, Nogueira L, Yabroff R, Shulman L, Facktor M, Winchester DP et al. : Incident Cases Captured in the National Cancer Database Compared with Those in U.S. Population Based Central Cancer Registries in 2012–2014. Ann Surg Oncol 2019.
3. Mallin K, Palis BE, Watroba N, Stewart AK, Walczak D, Singer J, Barron J, Blumenthal W, Haydu G, Edge SB: Completeness of American Cancer Registry Treatment Data: implications for quality of care research. J Am Coll Surg 2013, 216(3):428–437. [PubMed: 23357724]
4. An MW, Tang J, Grothey A, Sargent DJ, Ou FS, Mandrekar SJ: Missing tumor measurement (TM) data in the search for alternative TM-based endpoints in cancer clinical trials. Contemp Clin Trials Commun 2020, 17:100492. [PubMed: 31872158]
5. Newman DA: Missing Data:Five Practical Guidelines. Organizational Research Methods 2014, 17(4):372–411.
6. Graham JW: Missing data analysis: making it work in the real world. Annu Rev Psychol 2009, 60:549–576. [PubMed: 18652544]
7. Walters S, Maringe C, Butler J, Rachet B, Barrett-Lee P, Bergh J, Boyages J, Christiansen P, Lee M, Wärnberg F et al. : Breast cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000–2007: a population-based study. Br J Cancer 2013, 108(5):1195–1208. [PubMed: 23449362]
8. Maringe C, Walters S, Rachet B, Butler J, Fields T, Finan P, Maxwell R, Nedrebø B, Påhlman L, Sjövall A et al. : Stage at diagnosis and colorectal cancer survival in six high-income countries: a population-based study of patients diagnosed during 2000–2007. Acta Oncol 2013, 52(5):919–932. [PubMed: 23581611]
9. WHO/IARC Classification of Tumours, vol. 4, 4 edn: World Health Organization; 2012.
10. Yang DX, Khera R, Miccio JA, Jairam V, Chang E, Yu JB, Park HS, Krumholz HM, Aneja S: Prevalence of Missing Data in the National Cancer Database and Association With Overall Survival. JAMA Netw Open 2021, 4(3):e211793. [PubMed: 33755165]
11. Boffa DJ, Rosen JE, Mallin K, Loomis A, Gay G, Palis B, Thoburn K, Gress D, McKellar DP, Shulman LN et al. : Using the National Cancer Database for Outcomes Research. JAMA Oncology 2017, 3(12):1722. [PubMed: 28241198]
12. Bilimoria KY, Bentrem DJ, Stewart AK, Winchester DP, Ko CY: Comparison of Commission on Cancer–Approved and –Nonapproved Hospitals in the United States: Implications for Studies That Use the National Cancer Data Base. Journal of Clinical Oncology 2009, 27(25):4177–4181. [PubMed: 19636004]

13. Schlick CJ, Yang AD: Is there value in cancer center accreditation? Am J Surg 2020, 220(1):27–28. [PubMed: 32416939]
14. Brubakk K, Vist GE, Bukholm G, Barach P, Tjomsland O: A systematic review of hospital accreditation: the challenges of measuring complex intervention effects. BMC Health Serv Res 2015, 15:280. [PubMed: 26202068]
15. Fong ZV, Chang DC, Hur C, Jin G, Tramontano A, Sell NM, Warshaw AL, Fernandez-Del Castillo C, Ferrone CR, Lillemoe KD et al. : Variation in long-term oncologic outcomes by type of cancer center accreditation: An analysis of a SEER-Medicare population with pancreatic cancer. Am J Surg 2020, 220(1):29–34. [PubMed: 32265013]
16. David EA, Cooke DT, Chen Y, Perry A, Canter RJ, Cress R: Surgery in high-volume hospitals not commission on cancer accreditation leads to increased cancer-specific survival for early-stage lung cancer. The American Journal of Surgery 2015, 210(4):643–647. [PubMed: 26193801]
17. SEER Cancer Statistics Review, 1975–2014, National Cancer Institute [https://seer.cancer.gov/csr/1975_2014/]
18. Kuo T-M, Mobley LR: How generalizable are the SEER registries to the cancer populations of the USA? Cancer Causes & Control 2016, 27(9):1117–1126. [PubMed: 27443170]
19. Bleicher RJ, Ruth K, Sigurdson ER, Beck JR, Ross E, Wong Y-N, Patel SA, Boraas M, Chang EI, Topham NS et al. : Time to Surgery and Breast Cancer Survival in the United States. JAMA Oncology 2016, 2(3):330. [PubMed: 26659430]
20. Gradishar WJ, Anderson BO, Abraham J, Aft R, Agnese DM, Allison KH, Blair SL, Burstein HJ, Dang C, Elias AD et al.: NCCN Clinical Practice Guidelines in Oncology: Breast Cancer. In., Version 1.2019 edn. Online; 2019.
21. Rapp J, Tuminello S, Alpert N, Flores RM, Taioli E: Disparities in surgery for early-stage cancer: the impact of refusal. Cancer Causes Control 2019, 30(12):1389–1397. [PubMed: 31630307]
22. Luo Q, Egger S, Yu XQ, Smith DP, O'Connell DL: Validity of using multiple imputation for "unknown" stage at diagnosis in population-based cancer registry data. PLoS One 2017, 12(6):e0180033. [PubMed: 28654653]
23. Huchcroft SA, Snodgrass T: Cancer patients who refuse treatment. Cancer Causes Control 1993, 4(3):179–185. [PubMed: 8318634]
24. Weinmann S, Taplin SH, Gilbert J, Beverly RK, Geiger AM, Yood MU, Mouchawar J, Manos MM, Zapka JG, Westbrook E et al. : Characteristics of women refusing follow-up for tests or symptoms suggestive of breast cancer. Journal of the National Cancer Institute Monographs 2005(35):33–38. [PubMed: 16287883]
25. van Buuren S, Groothuis-Oudshoorn K: mice: Multivariate Imputation by Chained Equations in R. 2011 2011, 45(3):67.
26. Azur MJ, Stuart EA, Frangakis C, Leaf PJ: Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011, 20(1):40–49. [PubMed: 21499542]
27. Hoskin TL, Boughey JC, Day CN, Habermann EB: Lessons Learned Regarding Missing Clinical Stage in the National Cancer Database. Ann Surg Oncol 2019, 26(3):739–745.
28. Egleston BL, Wong YN: Sensitivity analysis to investigate the impact of a missing covariate on survival analyses using cancer registry data. Stat Med 2009, 28(10):1498–1511. [PubMed: 19235263]
29. Motzer RJ, Jonasch E, Agarwal N, Alva A, Bhayani S, Choueiri TK, Costello BA, Derweesh IH, Gallagher TH, George S et al.: NCCN Clinical Practice Guidelines in Oncology: Kidney Cancer. Version 2.2020 In., Version 2.2020 edn. Online; 2020.
30. Overview of the SEER Program [https://seer.cancer.gov/about/overview.html]
31. National Cancer Database [http://www.facs.org/quality-programs/cancer/ncdb]
32. Mercieca-Bebber R, Palmer MJ, Brundage M, Calvert M, Stockler MR, King MT: Design, implementation and reporting strategies to reduce the instance and impact of missing patient-reported outcome (PRO) data: a systematic review. BMJ Open 2016, 6(6):e010938.
33. Wisniewski SR, Leon AC, Otto MW, Trivedi MH: Prevention of missing data in clinical research studies. Biol Psychiatry 2006, 59(11):997–1000. [PubMed: 16566901]

Fig 1A.

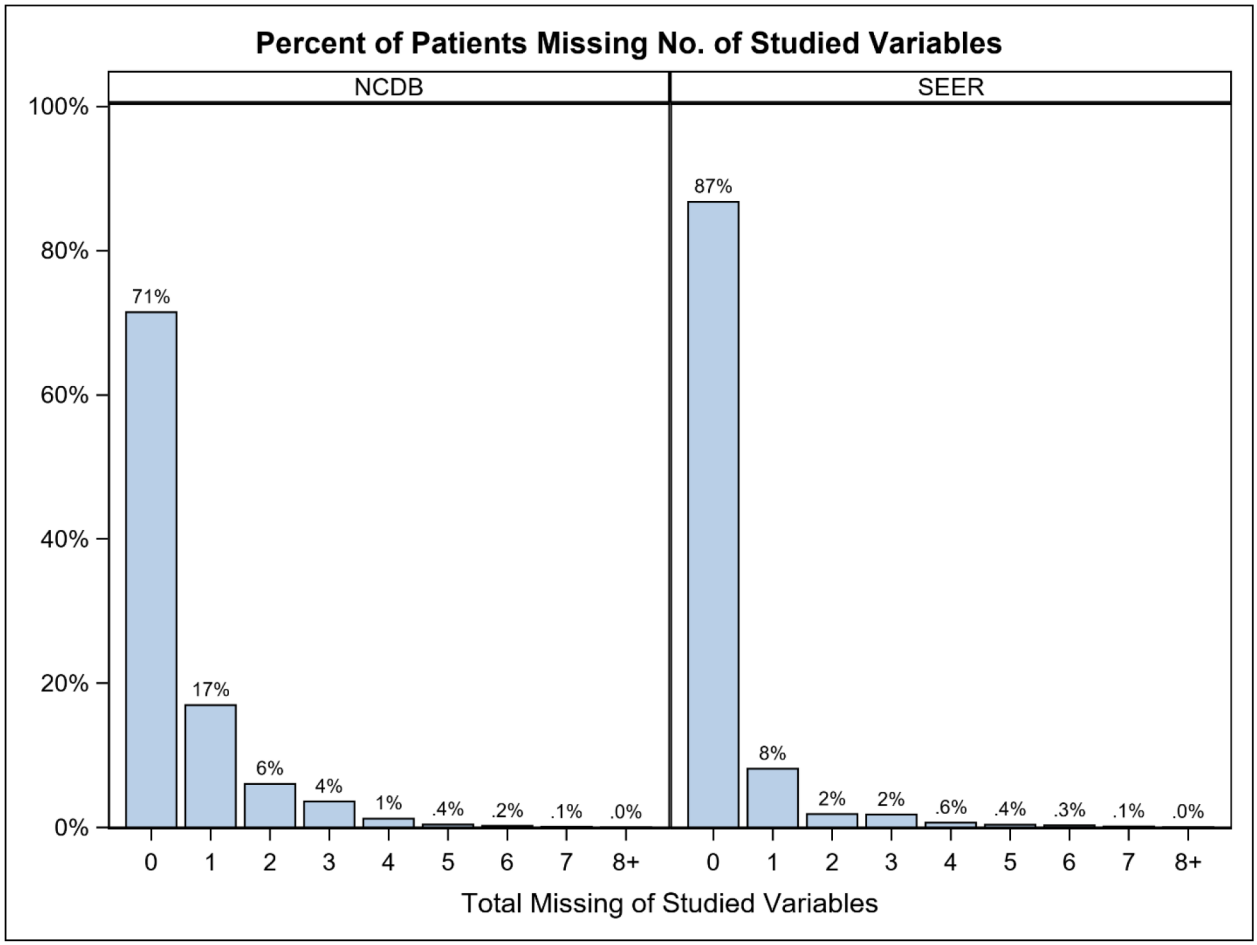

Fig 1B.

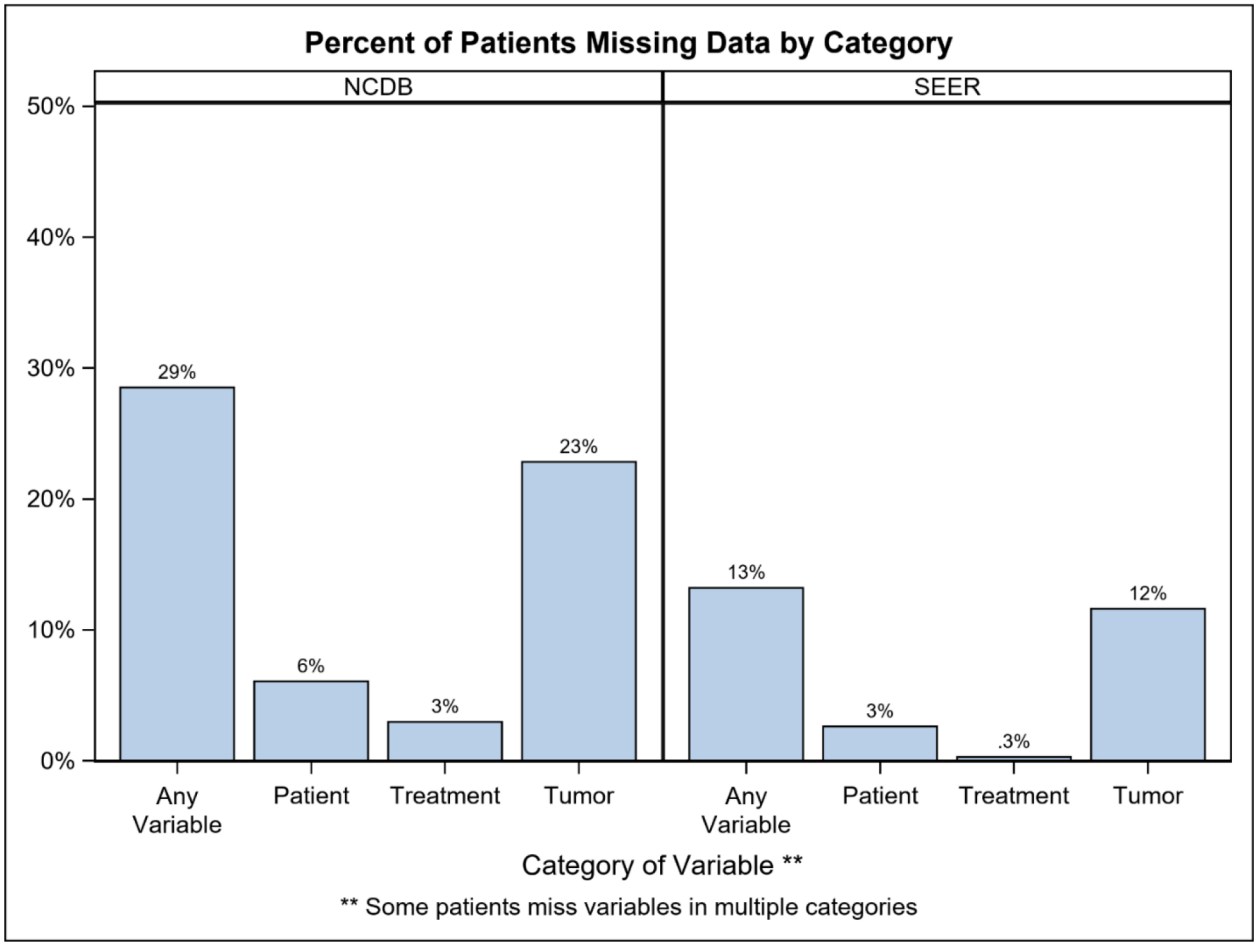

Figure 1.

Percent of breast cancer patients (diagnosed 2010–2014) missing (A) 0 to 8+ of the 14 (SEER) or 16 (NCDB) key variables; or (B) data by variable category. Key variables included: demographic (age, race, ethnicity, insurance, education, income), tumor (grade, ER, PR, HER2, TNM stages), and treatment (surgery in both databases; chemotherapy and radiation in NCDB). NCDB: National Cancer Database. SEER: Surveillance, Epidemiology, End Results Program.

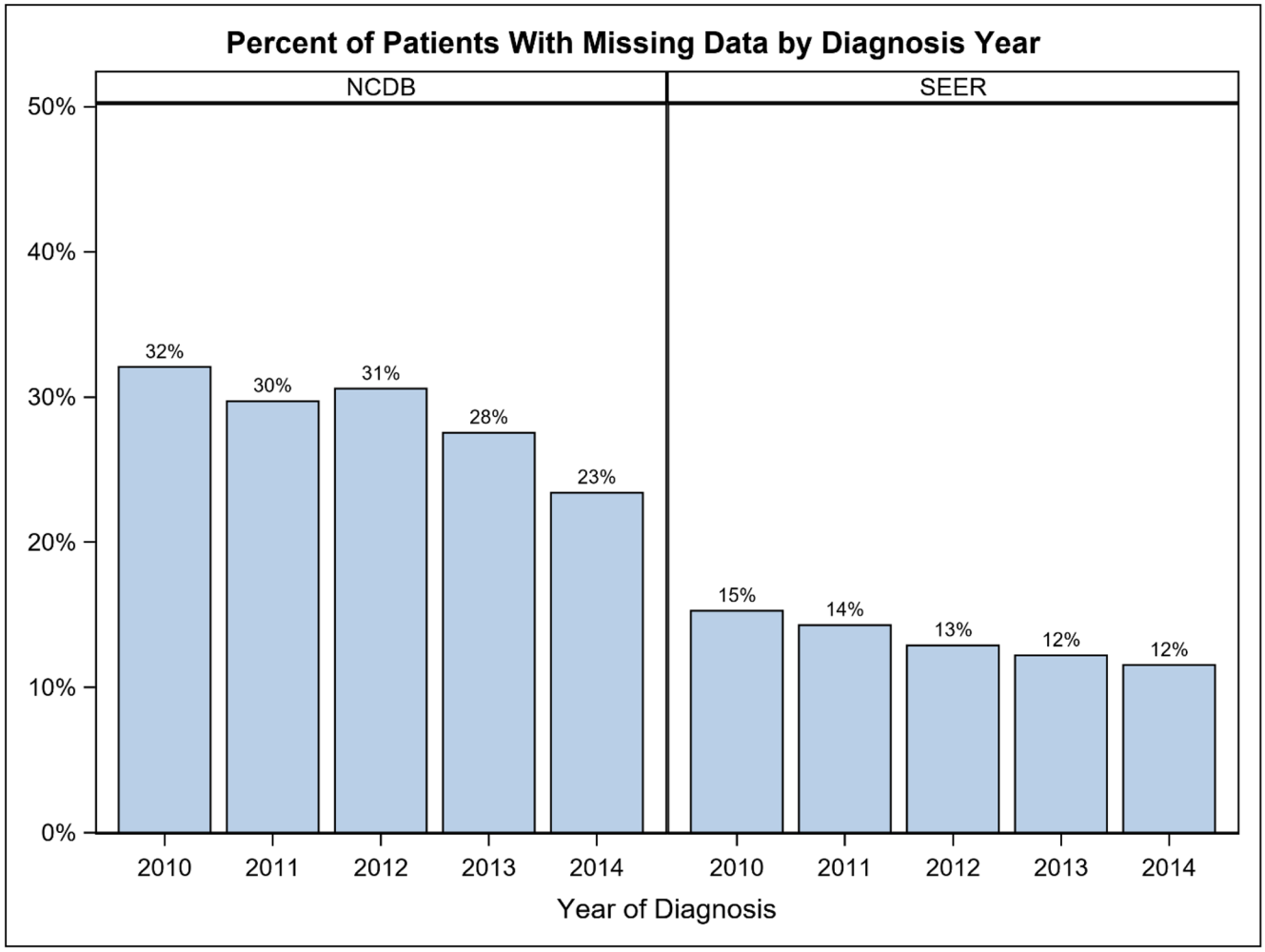

Figure 2.

Percent of breast cancer patients with missing data by diagnosis year (2010–2014). NCDB: National Cancer Database. SEER: Surveillance, Epidemiology, End Results Program.

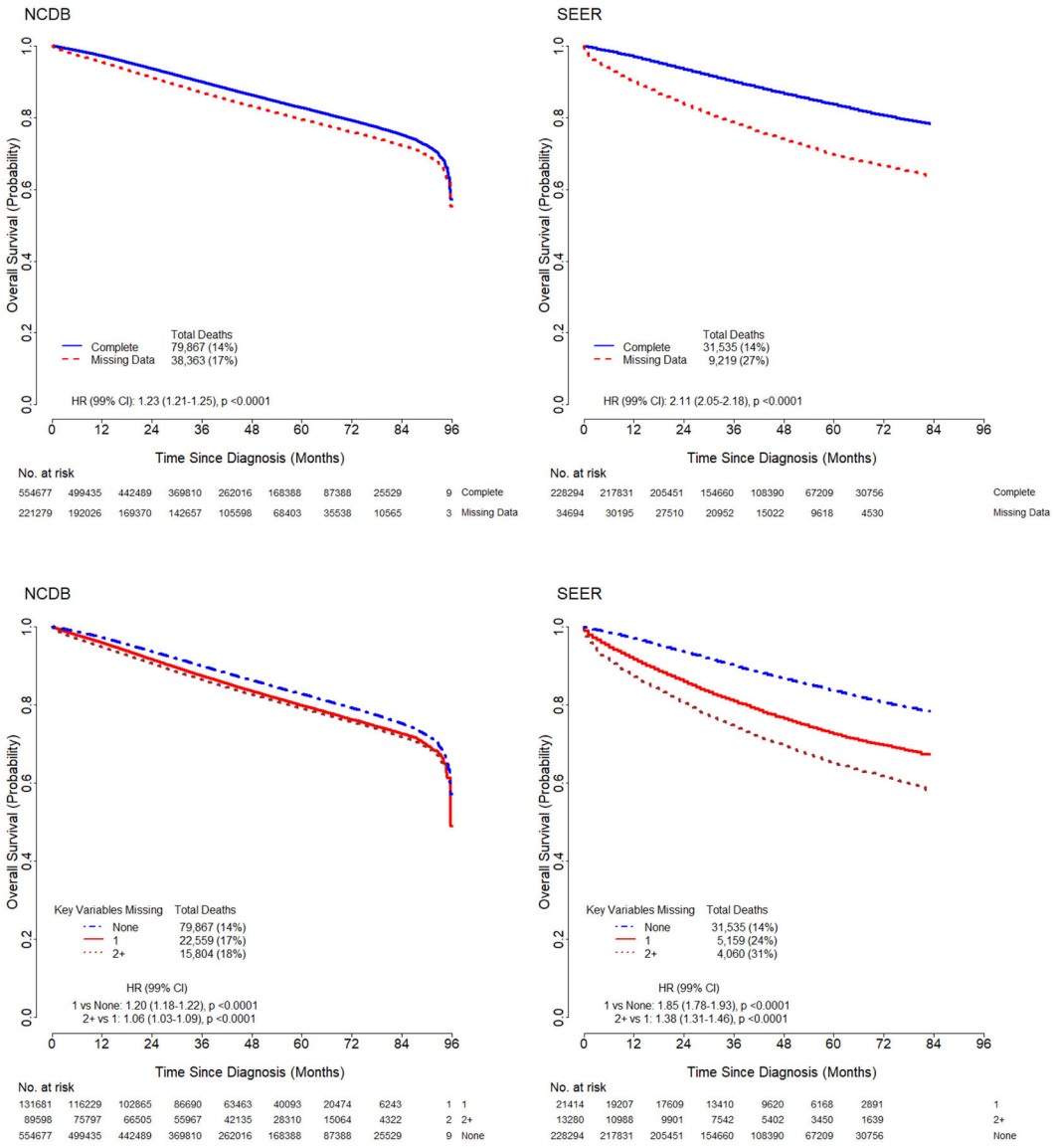

Figure 3.

Unadjusted overall survival for breast cancer patients (diagnosed 2010–2014) compared by missingness of any key variable (3A: NCDB, 3B: SEER), or compared by number of missing key variables (3C: NCDB; 3D: SEER). NCDB: National Cancer Database. SEER: Surveillance, Epidemiology, End Results Program.

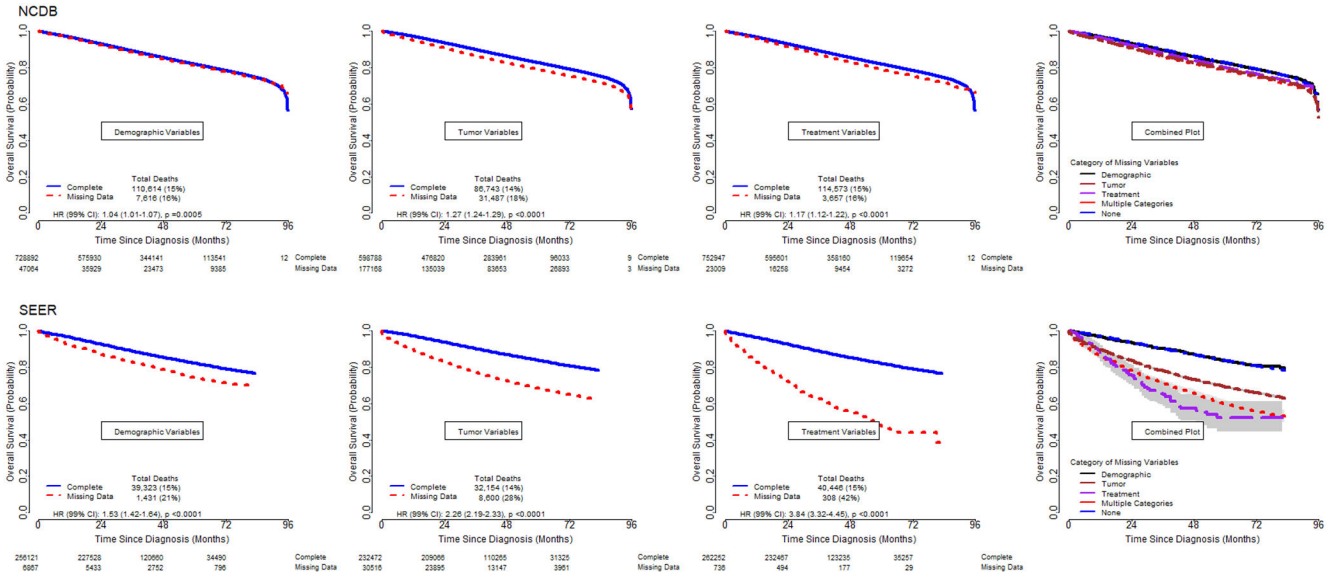

Figure 4.

Unadjusted overall survival for breast cancer patients (diagnosed 2010–2014) compared by missingness within and across categories. (4A) NCDB, Demographic variables; (4B) NCDB, Tumor variables; (4C) NCDB, Treatment variables; (4D) NCDB, combined; (4E) SEER, Demographic variables; (4F) SEER, Tumor Variables; (4G) SEER, Treatment Variables, (4H) SEER, combined. NCDB: National Cancer Database. SEER: Surveillance, Epidemiology, End Results Program.

Table 1.

Demographic, tumor, and treatment data for breast cancer patients (diagnosed 2010–2014) in the NCDB and SEER. NCDB: National Cancer Database. SEER: Surveillance, Epidemiology, End Results Program. NOS: not otherwise specified. ER: estrogen receptor. PR: progesterone receptor. HER2: human-epidermal-growth-factor-receptor-2.

	NCDB n=775,996				SEER n=263,016			
	Complete		Missing Data		Complete		Missing Data	
	N	%	N	%	N	%	N	%
Sex								
Female	549860	99.1	218877	98.9	226598	99.3	34437	99.2
Male	4836	0.9	2423	1.1	1696	0.7	285	0.8
Age (years)								
<40	27376	4.9	9876	4.5	11091	4.9	1569	4.5
40–70	367911	66.3	143679	64.9	150847	66.1	21370	61.5
>70	140941	25.4	60471	27.3	58783	25.7	10709	30.8
missing	18468	3.3	7274	3.3	7573	3.3	1074	3.1
Race								
White	466936	84.2	176560	79.8	181692	79.6	26169	75.4
Black	64022	11.5	26806	12.1	25256	11.1	4327	12.5
Other	23738	4.3	10813	4.9	21346	9.4	3091	8.9
Unknown	0	0	7121	3.2	0	0	1135	3.3
Spanish/Hispanic Origin								
Hispanic	27857	5.0	14176	6.4	24285	10.6	4141	11.9
Not Hispanic	526839	95.0	178351	80.6	204009	89.4	30581	88.1
Unknown	0	0	28773	13.0	0	0	0	0
Insurance								
Private Insurance/Managed Care	285577	51.5	102232	46.2				
Insured					167666	73.4	18898	54.4
Insured/No specifics					30229	13.2	4754	13.7
Any Medicaid					26551	11.6	4000	11.5
Medicare	213912	38.6	82617	37.3				
Medicaid	37553	6.8	14788	6.7				
Not Insured	11867	2.1	5376	2.4	3848	1.7	814	2.3
Other Government	5787	1.0	2150	1.0				
Insurance Status Unknown	0	0	14137	6.4	0	0	6256	18.0
Percent of area of residence with less than high school diploma								
>=21.0%	78887	14.2	37317	16.9	46284	20.3	7218	20.8
13.0–20.9%	131697	23.7	53373	24.1	70507	30.9	11792	34.0
7.0–12.9%	185307	33.4	70137	31.7	94912	41.6	13337	38.4
<7.0%	158805	28.6	58647	26.5	16591	7.3	2310	6.7

	NCDB n=775,996				SEER n=263,016			
	Complete		Missing Data		Complete		Missing Data	
	N	%	N	%	N	%	N	%
missing	0	0	1826	0.8	0	0	65	0.2
Median household income								
>=$63,000	205624	37.1	80075	36.2	89059	39.0	12375	35.6
$48,000-$62,999	149812	27.0	57065	25.8	95490	41.8	15628	45.0
$38,000-$47,999	118043	21.3	46573	21.0	31434	13.8	4640	13.4
< $38,000	81217	14.6	35470	16.0	12311	5.4	2014	5.8
missing	0	0	2117	1.0	0	0	65	0.2
Grade								
Well differentiated, differentiated, NOS	124217	22.4	37609	17.0	51061	22.4	4631	13.3
Moderately differentiated, moderately well differentiated, intermediate differentiation	242518	43.7	72278	32.7	99126	43.4	8858	25.5
Poorly differentiated	186653	33.6	52008	23.5	77417	33.9	6729	19.4
Undifferentiated, anaplastic	1308	0.2	460	0.2	690	0.3	217	0.6
Cell type not determined, not stated or not applicable, unknown primaries, high grade dysplasia	0	0	58945	26.6	0	0	14287	41.1
ER status								
Positive/elevated	451488	81.4	171621	77.6	188273	82.5	22334	64.3
Negative/normal	102961	18.6	38462	17.4	39914	17.5	5252	15.1
Borderline, undetermined whether positive or negative	247	0.0	113	0.1	107	0.0	30	0.1
Missing/Unknown	0	0	11104	5.0	0	0	7106	20.5
PR status								
Positive/elevated	396889	71.6	148844	67.3	163842	71.8	18246	52.5
Negative/normal	157167	28.3	59785	27.0	64160	28.1	8448	24.3
Borderline, undetermined whether positive or negative	640	0.1	326	0.1	292	0.1	64	0.2
Missing/Unknown	0	0	12345	5.6	0	0	7964	22.9
HER2 status								
Positive/elevated	80224	14.5	28205	12.7	33729	14.8	4056	11.7
Negative/normal	459108	82.8	153574	69.4	189584	83.0	17341	49.9
Borderline, undetermined whether positive or negative	11913	2.1	5132	2.3	4981	2.2	757	2.2
Not applicable and not collected	3451	0.6	1038	0.5	-	-	-	-
Missing/Unknown	0	0	33351	15.1	0	0	12568	36.2
T stage								
0	12461	2.2	6896	3.1	83	< 0.1	254	0.7
1	342490	61.7	86833	39.2	136474	59.8	15275	44.0
2	149295	26.9	38060	17.2	69447	30.4	6658	19.2
3	28631	5.2	8612	3.9	13468	5.9	1778	5.1
4	21819	3.9	8743	4.0	8822	3.9	2252	6.5
missing	0	0	72156	32.6	0	0	8505	24.5

	NCDB n=775,996				SEER n=263,016			
	Complete		Missing Data		Complete		Missing Data	
	N	%	N	%	N	%	N	%
N stage								
0	452476	81.6	127369	57.6	156198	68.4	20365	58.7
1	78820	14.2	23031	10.4	52389	22.9	6806	19.6
2	14316	2.6	4519	2.0	12193	5.3	1272	3.7
3	9084	1.6	3280	1.5	7514	3.3	980	2.8
missing	0	0	63101	28.5	0	0	5299	15.3
M stage								
0	533397	96.2	161905	73.2	219386	96.1	30399	87.5
1	21299	3.8	13529	6.1	8908	3.9	4306	12.4
missing	0	0	45866	20.7	0	0	17	0.0
Surgical Treatment								
Lumpectomy	295114	53.2	101053	45.7	119913	52.5	12483	36.0
Mastectomy	228460	41.2	89480	40.4	95258	41.7	11359	32.7
No surgery	30843	5.6	28757	13.0	12945	5.7	9972	28.7
Unknown type of surgery	279	0.1	429	0.2	131	0.1	153	0.4
Missing	0	0	1581	0.7	47	0.0	755	2.2
Chemotherapy								
Yes	247427	44.6	81006	36.6	-	-	-	-
No	307269	55.4	120512	54.5	-	-	-	-
Unknown/Missing	0	0	19782	8.9	-	-	-	-
Radiation								
Yes	326567	58.9	107341	48.5	-	-	-	-
No	228129	41.1	108392	49.0	-	-	-	-
Unknown/Missing	0	0	5567	2.5	-	-	-	-
Charlson-Deyo Score								
0	456545	82.3	186820	84.4	-	-	-	-
1	78034	14.1	27231	12.3	-	-	-	-
2	15387	2.8	5466	2.5	-	-	-	-
>=3	4730	0.9	1783	0.8	-	-	-	-

Table 2.

Single variable Proportional Hazard models estimates of missingness effect on overall survival for breast cancer patients (diagnosed 2010–2014) in the NCDB and SEER. NCDB: National Cancer Database. SEER: Surveillance, Epidemiology, End Results Program. HR: hazards ratio. CI: confidence interval. ER: estrogen receptor. PR: progesterone receptor. HER2: human-epidermal-growth-factor-receptor-2.

	NCDB		SEER	
Variable/Category	**% missing**	**HR (99% CI) for missing vs not**	**% missing**	**HR (99% CI)**
Demographic	6.1	1.04 (1.01 – 1.07)	2.6	1.53 (1.42 – 1.64)
Age	0	--	<0.01	--
Race	0.92	0.76 (0.70 – 0.84)	0.43	0.12 (0.07 – 0.23)
Hispanic Ethnicity	3.7	1.04 (1.01 – 1.08)	0	--
Insurance	1.8	1.10 (1.04–1.16)	2.4	1.69 (1.58 – 1.82)
Education	0.24	1.06 (0.90 – 1.25)	0.02	1.34 (0.60 – 3.04)[a]
Income	0.27	1.11 (0.95 – 1.28)	0.02	1.34 (0.60 – 3.04)[a]
Tumor	23	1.27 (1.24 – 1.29)	12	2.26 (2.19 – 2.33)
Grade	7.6	1.45 (1.42 – 1.49)	5.4	2.00 (1.91 – 2.09)
ER status	1.4	1.60 (1.52 – 1.68)	2.7	1.97 (1.85 – 2.09)
PR status	1.6	1.55 (1.48 – 1.63)	3.0	1.84 (1.74 – 1.95)
HER2 status	4.3	1.24 (1.20 – 1.29)	4.8	1.56 (1.48 – 1.64)
T stage	9.3	1.12 (1.09 – 1.14)	3.2	3.42 (3.27 – 3.59)
N stage	8.1	1.29 (1.26 – 1.32)	2.0	4.94 (4.69 – 5.20)
M stage	5.9	0.99 (0.96 – 1.02)	0.01	NE (1 event among 17 pts missing M stage)
Treatment	3.0	1.17 (1.12 – 1.23)	0.28	3.84 (3.32 – 4.45)
Surgery Type	0.2	2.11 (1.83 – 2.44)	--	--
Chemotherapy	2.6	1.09 (1.04 – 1.14)	--	--
Radiation	0.2	1.71 (1.59 – 1.84)	--	--

[a]Note: the same patients are missing education and income in SEER