Published in final edited form as:
Ann Surg Oncol. 2023 April ; 30(4): 2069–2084. doi:10.1245/s10434-022-12890-6.

Breast Surgical Oncology Epidemiologic Research: A Guide and Comparison of Four National Databases

Robyn N. Rubenstein, MD,
Jonas A. Nelson, MD, MPH,
Saïd C. Azoury, MD,
Meghana G. Shamsunder, MPH,
Kathryn Haglich, BS, MS,
Shen Yin, PhD,
Carrie S. Stern, MD,
Evan Matros, MD, MMSc, MPH
Plastic and Reconstructive Surgery Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY

Abstract

Background.—National databases are a rich source of epidemiologic data for breast surgical oncology research. However, these databases differ in the demographic, surgical, and oncologic variables provided. This study aimed to compare the strengths and limitations of four national databases in the context of breast surgical oncology research.

Methods.—The study comprised a descriptive analysis of four national databases (the National Surgical Quality Improvement Program [NSQIP], the Nationwide Inpatient Sample [NIS], the Surveillance, Epidemiology and End Results [SEER] program, and the National Cancer Database [NCDB]) to assess their strengths and limitations in the context of breast surgical oncology. The study assessed the data available in each database for female patients with a breast cancer diagnosis between 2007 and 2017, and compared patient age, ethnicity, and race distributions.

Results.—Data from 3.9 million female patients were examined, with most patients being between 60 and 69 years of age, non-Hispanic, and white. Age, ethnicity, and race distributions were similar in the databases. The NSQIP includes data on operative details, comorbidities, and postoperative outcomes. The NIS provides health services and inpatient utilization information, but does not evaluate outpatient procedures. The SEER program provides population-based oncologic detail including stage, histology, and neoadjuvant/adjuvant treatment. The NCDB offers hospital-based oncologic information and the largest population in the study period, with approximately 2.5 million breast cancer patients.

J. A. Nelson, MD, MPH, nelsonj1@mskcc.org; E. Matros, MD, MMSc, MPH, matrose@mskcc.org.
Robyn N. Rubenstein and Jonas A. Nelson contributed equally to this work.

Supplementary Information The online version contains supplementary material.

DISCLOSURES There are no conflicts of interest.

Conclusion.—Epidemiologic datasets offer tremendous potential for the examination of oncologic breast surgery, with each database providing unique data useful for addressing different epidemiologic questions. Understanding the strengths and limitations of each database creates a more efficient and productive research environment.

Multiple national databases are available for epidemiologic examination of breast surgery patients, including the National Surgical Quality Improvement Program (NSQIP), the Nationwide Inpatient Sample (NIS), the Surveillance, Epidemiology and End Results (SEER) program, and the National Cancer Database (NCDB). With the benefit of large sample sizes, these resources allow researchers to assess trends in disease and treatment practices to improve patient outcomes.

These four databases vary in inclusion criteria, goals, and size, as well as in demographic, surgical, and oncologic data provided. The NSQIP, sponsored by the American College of Surgeons (ACS), is a national, risk-adjusted, outcomes-based database that measures the morbidity and mortality of surgical patients.[1] Originating in the 1980s out of concern for higher operative mortality rates in Veterans Affairs hospitals and without an existing national standard for baseline comparison, the goal of the NSQIP is to use these outcomes-based research measures to improve the quality of care for surgical patients (Table 1).[2,3] Using an ACS-validated sampling protocol to sample surgical cases systematically based on hospital surgical volume, the NSQIP reports 30-day outcomes for all major surgical procedures determined by current procedural terminology (CPT) codes and includes patients who are 18 years of age or older.[4] The NSQIP has grown from a 2005–2006 data file with a total of 152,490 cases from 121 hospitals into a 2019 data file with a total of 1,076,441 cases from 719 hospitals in 49 states.[2]

As part of the Healthcare Cost and Utilization Project (HCUP) sponsored by the Agency for Healthcare Research and Quality (AHRQ), the NIS is a publicly available, all-payer, inpatient-care database that reports estimates of inpatient utilization and outcomes.[5] The NIS includes data on patients of all ages admitted for any medical or surgical reason to one of the various HCUP-participating facilities, excluding rehabilitation centers and long-term acute care hospitals (Table 1).[5] The NIS data, available from 1988 to 2018, has grown from its initial inclusion of 8 states to its current inclusion of 47 states.[5]

Beginning in 2012, the NIS transitioned from including “all discharges from a sample of hospitals” to being a “sample of discharges” (not a comprehensive inclusion of all patients) from all HCUP-participating hospitals to reduce sampling error and provide more accurate national estimates.[6] The HCUP provides a collection of databases with varying focuses, including the State Inpatient Databases (SIDs), the Kids’ Inpatient Database (KID), the Nationwide Readmissions Database (NRD), the National (Nationwide) Inpatient Samples (NIS), and others. Containing data from more than 7 million hospitalizations annually, the NIS is sampled from the more granular SIDs, representing a 20% sample of discharges from U.S. hospitals.[7] The goal of NIS is to provide data for health care services and policy research, including “utilization of health services by special populations, hospital stays for rare conditions, variations in medical practice, health care cost inflation, regional and national analyses, quality of care and patient safety, impact of health policy changes, and access to care[5]".

The SEER program is sponsored by the Surveillance Research Program in the National Cancer Institute's Division of Cancer Control and Population Sciences. Beginning in 1975 with seven geographic locations, SEER is a cancer database that includes patients of all ages and collects comprehensive population-based data from U.S. cancer registries in order to be representative of the U.S. population (Table 1).[8,9] Inclusion is based on strategically chosen geographic locations and currently captures approximately 30% of new cancer diagnoses in the U.S. The goal of the SEER program includes collecting comprehensive data on cancers diagnosed within the geographic areas targeted by the SEER registries to provide epidemiologically appropriate information. An additional goal is to report on U.S. cancer burden, incidence, mortality, and overall survival,[9] aiming to identify changes in patterns of cancer occurrences, temporal changes in cancer incidence, occurrences of iatrogenic cancers, and cancer prevention patterns/outcomes.[9]

The NCDB, sponsored jointly by the American College of Surgeons and the American Cancer Society, is a hospital-based, clinical oncology database that includes patients of all ages with a cancer diagnosis and collects data from more than 1500 Commission on Cancer (CoC)-accredited facilities in the U.S. (Table 1).[10] With data available from 1989 to 2017, the NCDB tracks cancer patients whose diagnosis and treatment were received in a CoC-accredited program. The NCDB captures more than 70% of newly diagnosed cancers in the U.S., covering approximately 30% >1500/5000 of hospitals. Previously voluntary, the NCDB currently requires all CoC-accredited facilities to participate in data collection[11] and currently includes more than 34 million records.[10] The goals of NCDB are to collect accurate cancer data in order to track and analyze patients with malignant disease, explore trends in the treatment of cancer, facilitate improvement in the quality of cancer patient care, and improve outcomes.[10]

Although several studies have examined the strengths and limitations of various national databases,[11-23] none have compared the NSQIP, NIS, SEER, and NCDB with a specific focus on breast surgical oncology. This study aimed to compare the strengths and limitations of these four large national databases in the context of breast surgical oncology and create a guide to help researchers select the ideal database to address specific clinical research questions.

MATERIALS AND METHODS

We performed a descriptive analysis of the data available in the NSQIP, NIS, SEER, and NCDB, and assessed the strengths and limitations of each in the setting of breast surgical oncology. In our descriptive analysis, the variables available in each database were categorized into clinically relevant groups: Basic Demographics, Socioeconomics/Education/Insurance Information, Comorbidities, Facility Information, Hospitalization Information, Surgical Information, Diagnosis Information, and Oncologic Information. For patient data, we extracted age, ethnicity, and race information from female patients with a breast cancer diagnosis between 2007 and 2017 from each database. Annual NSQIP files from 2007 to 2017 and NIS data from annual “Inpatient Core” files from 2007 to 2017 were used. For SEER, the SEER 18 Incidence Registry from the November 2019 submission was used with a Case Listing Session in SEER*Stat. The NCDB data were delivered as a breast site-specific file. All patients with a breast cancer diagnosis (invasive or in situ) were included. International Classification of Diseases-9/10 (ICD9-/10) diagnosis codes were used to capture breast cancer patients for NSQIP and NIS data. For SEER, patients whose primary cancer site was breast (variable name: “site recode ICD-O-3/WHO 2008”) were selected. Because NCDB data are obtained in a site-specific manner, the breast cancer file was used with no further exclusion criteria. Institutional Review Board approval was not necessary for this study.

Age, ethnicity, and race were analyzed as categorical variables, and distributions were reported as counts and percentages, which were compared across the four databases using R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria). Basic filtering was performed using software compatible with each database, including RStudio (NSQIP), SEER*Stat Version 8.3.8 (SEER), and SPSS (NIS and NCDB).

Using the NSQIP, SEER, and NCDB, the volume of patients who underwent unilateral mastectomy (UM) or contralateral prophylactic mastectomy (CPM) in 2017 was determined using R version 4.1.2. To identify patients who underwent UM or CPM, CPT codes were used for the NSQIP, whereas site-specific surgery codes were used for SEER and the NCDB. We excluded NIS data from this analysis because the transition from the ICD-9-CM/PCS to the ICD-10-CM/PCS coding system in 2015 affected the quality of surgical trends data.

Obtaining the Data

The NSQIP participant use data file (PUF) incorporates all cases submitted to the NSQIP annually. Specific patients, hospitals, and providers are not identifiable.[24] The PUF is free and available to employees of NSQIP-participating hospitals. Available years include 2005–2019, with new PUF user guides distributed annually. All NSQIP data files are supplemented with a year-specific user guide and data dictionary.[24] The data files are available in SPSS, SAS, and delimited text file types (Table 1).

Available for purchase, NIS data are distributed as ASCII-formatted data files that can be loaded into SAS, SPSS, Stata, or other similar analysis software (Table 1).[7] The annual NIS data are distributed in four files: Inpatient Core, Diagnosis and Procedure Groups, Disease Severity Measures, and Hospital.[7] Given the transition from the ICD-9-clinical modification/procedure coding system (CM/PCS) to ICD-10-CM/PCS data in 2015, the first 9 months of 2015 uses ICD-9 codes, and the last 3 months uses ICD-10 codes. Data involving procedure and diagnosis codes are split into two files labeled as “Q1–Q3” and “Q4.”[5,7]

Access to SEER data is free, and viewing the data requires SEER*Stat software.[25] The most recent submission (November 2020) includes data from 1975 to 2018. However, the November 2019 submission was used for this study because it was the most current at the time of the analysis.[26] The SEER data are presented in registry groupings based on the number of registries/locations used, which include SEER 9, 13, 18, and 21 (Table 1).[27] This analysis used SEER 18, which covers approximately 28% of the U.S. population according to the 2010 census.[26,27]

The NCDB data are available to investigators associated with CoC-accredited cancer programs through a cancer site-specific application process, which includes a research project proposal and letter of support from the Cancer Committee Chair of the facility. The data are currently available from 1989 to 2017. The PUFs are provided in flat text, SPSS, SAS, and Stata files.[28]

RESULTS

Descriptive Analysis of the Databases

The data available in each database are detailed in Table 2 and described further in the following discussion.

Basic Demographics

All four databases include age, sex, race, and ethnicity (Table 2). However, the NSQIP includes only patients 18 years of age or older, whereas the others include all ages. The NSQIP and NIS classify race into more general categories (American Indian or Alaska Native, Asian, black or African American, Native Hawaiian or Pacific Islander, white, and unknown/not reported), whereas the NCDB and SEER offer more detailed groupings (30 categories). All four databases provide information regarding the ethnicity of the patient, with the NCDB offering slightly more information regarding ethnic origin (10 categories vs Hispanic: yes/no).

Socioeconomics, Education, Insurance Information

The NSQIP does not provide socioeconomics, education, or insurance information, whereas the NIS and NCDB provide the expected primary payer. Median household income based on patient zip code is provided by the NIS, SEER, and NCDB. Total hospital charges are provided by the NIS. As a proxy for educational attainment, the NCDB provides information about the percentage of people living in the patient’s zip code area who do not have a high school degree.

Comorbidities

For comorbidities, the NSQIP and NIS provide an extensive list of potential comorbidities per patient. The NSQIP uniquely offers information on the patient’s functional status, American Society of Anesthesiologists (ASA) classification, and preoperative laboratory values. The NSQIP and NIS include estimated probabilities of morbidity and mortality. Although the NCDB does not list patient comorbidities, it does provide a Charlson-Deyo comorbidity score, whereas SEER does not provide information on patient comorbidities.

Facility Information

The NSQIP does not provide facility information, whereas the NIS offers information on the ownership, bed size, and teaching status of the hospital. The NIS and NCDB provide the region/location of the treating facility. The SEER program offers information regarding the urban/rural status of the treatment center's location as well as the type of data reporting source. The NCDB provides facility type, urban/rural information, and more detailed location information than the NIS, and uniquely, the great-circle distance (distance in miles between the patient's residence and the reporting hospital).

Hospitalization Information

Regarding hospitalization data, "outpatient" is defined as surgery for which a patient did not stay in the hospital overnight, whereas "inpatient" is defined as observation either 23 h or longer. Whereas the NSQIP reports the inpatient/outpatient status of the procedure, the NIS includes only inpatient cases. The NSQIP provides the quarter of admission, days from admission to operation, days from operation to discharge, length of hospital stay, discharge destination, and readmission information.

The NIS clarifies whether the admission was on a weekday versus a weekend, month of admission, elective versus non-elective admission, number of days from admission to each listed procedure, length of hospital stay, discharge destination, discharge quarter, and whether the patient was transferred in from or out to another facility. Although the HCUP collects data on patients undergoing "observation services," these data are reported only in SIDs, not NIS. Because NIS data are sampled from the SIDs data, patients under 23-h observation are included in this sampling. However, researchers are unable to distinguish between 23-h observation and inpatient stay when using the NIS because it has no variable that distinguishes between 23-h observation and inpatient stay.

The NCDB specifies the length of surgical inpatient stay and whether there was a readmission to the same facility within 30 days after surgical discharge. The SEER data do not provide any information regarding specific hospitalizations.

Surgical Information

As a surgery database, the NSQIP provides the most detail regarding surgical procedures and outcomes, including the surgical specialty of the primary procedure, type of anesthesia, operative time, level of surgical resident present in the operating room, wound classification, and whether the surgery is elective or emergent. The NSQIP allows each patient to have up to 21 CPT codes, with associated relative value units (RVUs) for each procedure. It provides a primary CPT code as well as 10 "other" CPT codes (to classify any other procedures the primary surgical team is performing) and 10 "concurrent" CPT codes (to classify any additional procedure a separate/consulting surgical team is performing under the same anesthesia). With regard to postoperative complications and outcomes, the NSQIP provides an extensive list of potential postoperative complications as well as information regarding unplanned reoperations.

As an inpatient hospitalization database, the NIS provides the number of procedures a patient undergoes during that inpatient stay. The NIS allows for each patient to have, most recently, up to 25 ICD-9/10-PCS procedure codes. The SEER program and the NCDB use a site-specific surgery coding system, assigning one surgery code per patient. The surgery code classification systems differ by anatomic cancer site (e.g., the breast coding system was used for this study). If applicable, SEER indicates the reason why a patient did not undergo surgery and also provides information on the scope of regional lymph-node examination and regional lymph-node surgery involvement.

The NCDB provides diagnostic and staging procedural information, days from diagnosis to staging procedure, days from diagnosis to first surgical treatment procedure and to definite surgical treatment procedure, surgical margins, scope of regional lymph node surgery, surgical procedure information to other sites than the primary site, and reason for no surgery of the primary site, if applicable. Although not nearly as detailed as the NSQIP, the NCDB also provides outcome information in the form of 30- and 90-day mortality after surgery.

Regarding the scope of regional lymph-node surgery information (e.g., sentinel lymph-node biopsy or axillary lymph-node dissection) provided by each database, the NSQIP provides lymph-node surgeries in the form of CPT codes and the NIS in the form of ICD9/10 procedure codes. As of 2011, SEER no longer provides the “scope of regional lymph-node surgery” for breast cancer cases.[29] Due to errors in data collection and reporting, the NCDB lymph-node surgery codes before 2012 are inaccurate and therefore available only from 2012 onward.[30]

Diagnosis Information

The NSQIP provides diagnosis information using the ICD-9/10-CM codes, restricting users to one code per patient. The NSQIP began transitioning from ICD-9-CM codes to ICD-10-CM codes in 2015, with full transition to ICD-10-CM codes alone by 2016. The NIS also uses the ICD-CM diagnosis codes, with a similar transition in 2015 from ICD-9-CM to ICD-10-CM. However, each patient can have multiple diagnosis codes, with as many as 40 per patient currently. Both the SEER program and the NCDB databases provide detailed diagnosis information from an oncologic standpoint (see later). Notably, bilateral breast cancers are not specified in the databases.

Oncologic Information

As cancer databases, SEER and NCDB provide extensive oncologic information (Table 2 and Table S1). Both SEER and NCDB use the ICD-O-3 (International Classification of Disease for Oncology, 3rd edition) classification,[31] which offers information on the primary site, behavior, histology, and grade of the disease. Both SEER and NCDB also have information on laterality of disease, tumor size, hormonal status of the tumor, method of diagnostic confirmation, number of primary tumors, number of regional lymph nodes examined, and presence/location of distant metastases. Whereas SEER provides combined pathologic and clinical tumor-node-metastasis (TNM) staging, the NCDB provides a distinct clinical TNM staging and a separate pathologic TNM staging. Both SEER and NCDB offer “site-specific factors” that, for breast cancer, include results of the estrogen receptor (ER) and progesterone receptor (PR) tests, number of positive ipsilateral levels 1 and 2 axillary lymph nodes, immuno-histochemistry (IHC) of regional lymph nodes, molecular studies of regional lymph nodes, and the Nottingham or Bloom-Richardson score/grade (Table S1).

With regard to treatment details, SEER provides surgical information as previously described, as well as the timing of systemic and radiation therapy relative to surgery. The SEER program does not differentiate between endocrine therapy, chemotherapy, and immunotherapy. In addition, the NCDB provides extensive information regarding treatment status (e.g., whether treatment was given or not, active surveillance), days from diagnosis to initiation of treatment, and detailed information on radiation, chemotherapy, and palliative care. Specific to radiation, the NCDB provides days from diagnosis to initiation, timing in relation to surgery, modality, anatomic location or locations of treatment, dose per treatment, number of treatments, and number of phases of treatment with subsequent details for each phase (Table S1). For systemic treatment information, the NCDB includes days from diagnosis to initiation of systemic therapy (chemotherapy, endocrine therapy, immunotherapy), as well as timing relative to surgery. Notably, human epidermal growth factor receptor 2 (HER2)-targeted therapy cannot be clearly delineated because it often is included in chemotherapy. Both SEER and NCDB provide the vital status of the patient and the time from diagnosis to death (survival months), if applicable. However, neither SEER nor NCDB provide disease-free survival or recurrence information.

For hormone receptor status, variation exists among datasets. Both the NSQIP and NIS provide ICD9/10 diagnosis codes, with the NSQIP providing only one diagnosis code per patient, which can lead to missingness and a lack of comprehensiveness. The NIS most recently provides up to 40 diagnosis codes per patient, making it more comprehensive with lower rates of missingness.

Detail provided by ICD9/10 codes regarding hormone receptor status is limited but includes ICD 9 code V86.1 (estrogen receptor [ER]-negative), ICD 9 code V86.0 (ER-positive), ICD 10 code Z17.1 (ER-negative), and Z17.0 (ER-positive), as well as ICD 10 code Z19 (hormone-sensitive malignancy status), which includes Z19.1 (hormone-sensitive malignancy status) and Z19.2 (hormone-resistant malignancy status). No ICD9/10 codes specifically address PR (progesterone receptor) status or HER2 status. Both SEER and NCDB do provide ER/PR status as well as HER2 receptor status (available for SEER since 2010 and for NCDB since 2004), and although the missingness has improved over time, it still exists.

Breast Cancer Patients by Dataset

The NCDB contained the largest number of breast cancer patients in the examined time period (2007–2017), followed by SEER, NIS, and finally NSQIP (Fig. 1). Overall, the NCDB recorded 2,439,315 patients with breast cancer compared with 262,103 in the NSQIP. The population of breast cancer patients in the NCDB, SEER, and NSQIP increased during the study period, but decreased in the NIS.

During the period examined, the NSQIP increased from 13,405 breast cancer patients per year to 34,220 patients (155.3% increase); the SEER program increased from 70,893 to 83,764 breast cancer patients (18.2% increase); and the NCDB increased from 188,479 to 244,733 breast cancer patients (29.8% increase) (Fig. 1). During the study period, the NIS population decreased from 41,481 to 35,409 patients (14.6% decrease). Histologic data (invasive vs in situ cases) by dataset are demonstrated in Fig. 2.

The demographics for the patients with breast cancer in each database are shown in Table 3. Age distribution was similar between the four cohorts, with the largest percentage of patients in all four databases ranging in age from 60 to 69 years. Notably, a larger percentage of patients in the NIS dataset were 80 years of age or older (13.16%) than in the other datasets (7.39% in the NSQIP, 10.32% in the SEER program, and 9.24% in the NCDB). The majority of breast cancer patients in all four databases were non-Hispanic and white.

An examination of the proportion of UM to CPM in the last year of the study (2017) for the NSQIP, SEER, and NCDB demonstrated differences in the use of CPM by each database population. In 2017, the NSQIP had the lowest proportion of patients undergoing CPM (28.3%), whereas the SEER program demonstrated a CPM rate of 34.3%, and the NCDB demonstrated the highest CPM rate (40.1%) (Table 4; Fig. 3).

DISCUSSION

Nationwide datasets can provide an epidemiologic perspective on current populations and treatments for breast cancer. To our knowledge, this is the first comprehensive review of nationwide databases specifically in the context of breast surgical oncology. These databases provide a vast range of information to guide oncologic breast surgery research, and the presented study provides researchers with a better understanding of the databases to address each of their clinical questions (Table 5). The authors provide examples showing the usefulness of each database in addressing various clinical questions in Table 5.

The findings demonstrate that the NCDB offers the largest population of breast cancer patients overall, followed by the SEER, NIS, and NSQIP databases. It is important to reiterate that the four databases examined in this study have varying inclusion criteria, focuses, goals, and scopes (Tables 1, 2). Therefore, comparison of demographics among these databases must be performed with this concept in mind. Differences in the demographics reflect differences in data collection, epidemiologic aims, and goals of each database and represent transitions in data reporting and facility inclusion over time for each database. Over time, annual breast cancer cases in the NCDB, SEER, and NSQIP (invasive, in situ, and overall) have increased, with the NSQIP having the largest relative increase throughout the study period. This trend is likely due to an increase in facilities included in data collection and more consistent, improved reporting.[11] In contrast, the annual population of breast cancer patients reported by the NIS database decreased during the study period, which is likely a representation of the decrease in inpatient stays associated with surgical breast patients. This is reinforced by the observed decrease in in-situ cases (45.9%) during the study time frame, which was three times greater than for invasive cases (13.3%) in the NIS cohort. Because invasive breast cancer generally is treated more aggressively than in-situ cases, patients with in-situ pathology are less likely to require an inpatient stay than invasive cancer patients who undergo more invasive procedures.

Age, ethnicity, and race were the only variables that could be directly compared among the four databases. Despite the varying focus of each database, the age distributions overall were relatively similar between the datasets in the time frame examined. However, the NIS cohort had a higher percentage of patients 80 years of age or older. This can likely be explained by the notion that the NIS records inpatient stays, so it is reasonable that patients 80 years of age or older would more likely be hospitalized for breast cancer treatment.

Slight differences were noted with regard to race and ethnicity. Of the four databases, the SEER program contained the highest percentage of Hispanic patients as well as patients whose race was categorized as Asian, Pacific Islander, Hawaiian Native, or "other," suggesting that the SEER program reports a more diverse population than the other databases, consistent with previous literature comparing the SEER program and the NCDB.[12,20] This finding also is consistent with the focus of SEER to provide epidemiologically appropriate information by targeting strategically chosen geographic locations to represent the population.[8]

Ultimately, the patient demographics did not differ meaningfully among the four databases. The proportions of UM and CPM differed among the NSQIP, SEER, and NCDB, with the highest proportion of CPM use at the end of the study period (2017) in the NCDB population and lowest in the NSQIP population. The factors influencing the increased CPM use in SEER and NCDB should be further investigated in future studies.

Our descriptive analysis supplemented with the comparison of demographics across the four databases allowed us to assess the strengths and limitations of each database in the context of breast surgical oncology. The NSQIP, a risk-adjusted outcomes-based quality improvement database, aims to monitor the quality of postoperative care for surgical patients.[18,19] It provides a surgery-centered focus and is most useful in evaluating patient comorbidity profiles, operative details, and postoperative outcomes and complications. The NSQIP is well-suited for evaluating trends data for surgical procedures and early postoperative complications, but its oncologic data and health services/facility information are limited. Although the smallest of the four databases, the NSQIP still provides substantially large sample sizes, which can be of benefit compared with single-institution studies, and can provide clinically relevant, nationally based results.[18]

As a hospital system-centered, inpatient database, the NIS provides patient and facility information to guide health services and policy inquiries.[18] Health services research is a multidisciplinary field that includes examination of health care costs, patient outcomes based on cost, access to health care, and institutional processes with the goal of improving the management of health-related finances while providing the highest quality medical care for patients. The NIS is useful for health services inquiries because it is the only one of the four databases that provides information on total hospital charges, control/ ownership of inpatient facilities, facility bed size, and discharge quarter. This information enables researchers to examine health services topics, such as breast cancer patient burden on inpatient facilities, hospital costs associated with breast cancer treatment, costs of complications related to breast cancer treatment, and breast cancer inpatient burden by quarter and how it relates to insurance type. However, the NIS collects data during only a single hospital admission, so the costs associated with complications after discharge are not included, which is a limitation of the database.

The NIS also offers detailed comorbidity information, but it is limited with regard to the operative details provided compared with the NSQIP.[19] Researchers can use multiple diagnosis codes per patient for a better understanding of the diagnosis burden on patients with breast cancer. However, the 2015 transition from the ICD-9-CM/ PCS to the ICD-10-

CM/PCS coding system affected the quality of trends data,[32,33] which can be assessed more accurately with the other databases that did not experience a large conversion in the procedure coding system.

Because research in breast surgical oncology often involves analysis of diagnoses and procedures related to breast cancer, this transition must be kept in mind when 2015 NIS data are used. As an inpatient database, the NIS is well-suited to evaluate the hospitalization burden of both surgical and non-surgical breast cancer patients. However, researchers must be aware that the NIS is less useful when evaluating procedures largely performed on an outpatient basis (e.g., lumpectomy). Furthermore, because the NIS does include some 23-h observation cases as sampled from SIDs, researchers are unable to decipher between 23-h observation and inpatient stay in the NIS database. For this reason, the NIS will be increasingly less relevant for breast cancer patients as practices move toward more outpatient and 23-h observation procedures. Although most autologous breast reconstruction patients and many implant-based breast reconstruction patients will remain included in the database, some mastectomy patients, some implant-based reconstruction, and most lumpectomy patients are not included in this database.

The SEER program is an epidemiologic oncologic database used to monitor cancer incidence and mortality, assist in early detection and cancer prevention, and monitor outcome patterns.[11,20,34] In the context of breast cancer, the SEER program’s epidemiologic focus and extensive oncologic detail can guide research on population-based breast surgery patterns according to TNM stage, grade/behavior, histology, hormonal status, and non-surgical treatment. Although the SEER program does not provide comorbidity data and is limited to one surgical code per patient, it does incorporate some surgical variables, albeit fewer than the NSQIP or the NIS. Population-based cancer registries, such as the SEER program, can assist in epidemiologic assessment of diagnostic and treatment trends in breast cancer, with the aim of reporting and reducing cancer burden.[9,11,34]

The NCDB is the largest of the four databases and assists effectively in cancer surveillance and quality improvement research.[11,22,34] In contrast to the SEER program, the NCDB provides comorbidity, insurance, and facility/hospitalization data and offers the most detailed information regarding systemic and radiation treatment. The NCDB would prove useful to clinicians when evaluating radiation burden and systemic (chemo and endocrine) therapy use for surgical breast patients. Mallin et al.[21] determined that among the top 10 major anatomic cancer sites in the U.S., breast was the primary cancer site covered most extensively by the NCDB, reinforcing the NCDB as a useful database for breast surgical oncology. Notably, NCDB data are collected from CoC-accredited hospitals only, which often are larger than non-CoC facilities and offer more cancer-related services in more urban locations, potentially limiting its applicability for certain populations.[20,35]

Notably, a limitation of all four databases for those performing oncologic breast surgery research is that none of the four databases examined in this study provide the menopausal status of the patient. Rather, age (e.g., \50 or C50 years) can be used as a proxy. Additionally, in the NSQIP and NIS, the ICD 9/10 diagnosis codes may indicate genetic susceptibility to breast malignancy or a family history of breast cancer, but specific BRCA1/2 information is not provided by any of the four databases.

Another limitation of this study was that related to comparing the demographics of breast cancer cohorts from four different national databases with varying inclusion criteria and focuses. Although this must be kept in mind when the data are interpreted, the differences between the patient cohorts and the changes over time within each cohort are representative of the varying results researchers may produce based on which database the researcher chooses to examine. Further, missingness is a limitation of all national databases.

CONCLUSION

Each of the four national databases evaluated in this study provides data that lend themselves to specific clinical focuses and research questions in breast surgical oncology. Based on the analysis, the authors find the NSQIP ideal for investigations of surgical trends, comorbidity profiles, operative information, and postoperative outcomes; whereas they find the NIS well-suited for health services and policy research, the SEER for population-based oncologic studies, and the NCDB for hospital-based, large-scale oncology research. Thus, the analysis provides a guide that enables researchers to understand the strengths and limitations of the NSQIP, NIS, SEER, and NCDB databases in the context of breast surgical oncology to facilitate more efficient epidemiologic research in the field.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748, which supports Memorial Sloan Kettering Cancer Center's research infrastructure.

REFERENCES

1. Surgeons Aco. ACS National Surgical Quality Improvement Program. https://www.facs.org/quality-programs/acs-nsqip. Accessed 1 Oct 2021.
2. Surgeons Aco. ACS NSQIP History. https://www.facs.org/quality-programs/acs-nsqip/about/history. Accessed 1 Oct 2021.
3. Surgeons Aco. ACS National Surgical Quality Improvement Program. https://www.facs.org/quality-programs/acs-nsqip. Accessed 21 July 2021.
4. Surgeons Aco. About ACS NSQIP. https://www.facs.org/quality-programs/acs-nsqip/about. Accessed 1 Dec 2021.
5. (HCUP) HDHCaUP. Overview of the National (Nationwide) Inpatient Sample (NIS). 2021. https://www.hcup-us.ahrq.gov/nisoverview.jsp. Accessed 1 Oct 2021.
6. (HCUP) HDHCaUP. 2012 NIS Redesign Report Available on HCUP Website (May 2014). 2014 https://www.hcup-us.ahrq.gov/news/announcements/nisredesign_2012.jsp. Accessed 1 Oct 2021.
7. (HCUP) HDHCaUP. Introduction to the HCUP National Inpatient Sample (NIS), 2017. 2020. https://www.hcup-us.ahrq.gov/db/nation/nis/NIS_Introduction_2017.jsp#table2app1. Accessed 1 Oct 2021.
8. National Cancer Institute S, Epidemiology, and End Results Program. Overview of the SEER Program. https://seer.cancer.gov/about/overview.html. Accessed 1 Oct 2021.

9. National Cancer Institute S, Epidemiology, and End Results Program. Goals of the SEER Program. https://seer.cancer.gov/about/goals.html. Accessed 1 Oct 2021.

10. Surgeons Aco. National Cancer Database. https://www.facs.org/quality-programs/cancer/ncdb. Accessed 1 Oct 2021.

11. Mohanty S, Bilimoria KY. Comparing national cancer registries: the National Cancer Data Base (NCDB) and the Surveillance, Epidemiology, and End Results (SEER) program. J Surg Oncol. 2014;109:629–30. [PubMed: 24464362]

12. Mettlin CJ, Menck HR, Winchester DP, Murphy GP. A comparison of breast, colorectal, lung, and prostate cancers reported to the National Cancer Data Base and the Surveillance, Epidemiology, and End Results program. Cancer. 1997;79:2052–61. [PubMed: 9149035]

13. Winchester DP, Stewart AK, Bura C, Jones RS. The National Cancer Data Base: a clinical surveillance and quality improvement tool. J Surg Oncol. 2004;85:1–3. [PubMed: 14696080]

14. 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;15:683–90. [PubMed: 18183467]

15. Feig B. Comprehensive databases: a cautionary note. Ann Surg Oncol. 2013;20:1756–8. [PubMed: 23508583]

16. Lerro CC, Robbins AS, Phillips JL, Stewart AK. Comparison of cases captured in the National Cancer Database with those in population-based central cancer registries. Ann Surg Oncol. 2013;20:1759–65. [PubMed: 23475400]

17. Bohl DD, Basques BA, Golinvaux NS, Baumgaertner MR, Grauer JN. Nationwide Inpatient Sample and National Surgical Quality Improvement Program give different results in hip fracture studies. Clin Orthop Relat Res. 2014;472:1672–80. [PubMed: 24615426]

18. Kwaan MR. Using NSQIP as a research tool: how to answer questions that are not amenable to local data. Semin Colon Rectal Surg. 2016;27:83–6.

19. Alluri RK, Leland H, Heckmann N. Surgical research using national databases. Ann Transl Med. 2016;4:393. [PubMed: 27867945]

20. Boffa DJ. Using the National Cancer Database for outcomes research: a review. JAMA Oncol. 2017;3:1722–8. [PubMed: 28241198]

21. Mallin K, Browner A, Palis B, 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;26:1604–12. [PubMed: 30737668]

22. McCabe RM. National Cancer Database: the past, present, and future of the Cancer Registry and its efforts to improve the quality of cancer care. Semin Radiat Oncol. 2019;29:323–5. [PubMed: 31472733]

23. Janz TA, Graboyes EM, Nguyen SA, et al. A comparison of the NCDB and SEER database for research involving head and neck cancer. Otolaryngol Head Neck Surg. 2019;160:284–94. [PubMed: 30129822]

24. Surgeons Aco. ACS NSQIP Participant Use Data File. Retrieved 1 October at https://www.facs.org/quality-programs/acs-nsqip/participant-use.

25. National Cancer Institute S, Epidemiology, and End Results Program. How to Request Access to SEER Data. 1 October 2021 at https://seer.cancer.gov/data/access.html.

26. National Cancer Institute S, Epidemiology, and End Results Program. SEER*Stat Databases: November 2020 submission. Retrieved 1 October 2021 at https://seer.cancer.gov/data-software/documentation/seerstat/nov2020/#ss-variables.

27. National Cancer Institute S, Epidemiology, and End Results Program. Registry Groupings in SEER Data and Statistics. Retrieved 1 October 2021 at https://seer.cancer.gov/registries/terms.html.

28. Surgeons Aco. Participant User Files. Retrieved 1 October 2021 at https://www.facs.org/quality-programs/cancer/ncdb/puf.

29. National Cancer Institute S, Epidemiology, and End Results Program. Scope of Regional Lymph Node Surgery. Retrieved 25 May 2022 at https://seer.cancer.gov/seerstat/variables/seer/regional_ln/.

30. American college of Surgeons NCD. National Cancer Database Participant User File, 2017 Data Dictionary, Includes patients diagnosed in 2004–2017. 2020; pp 139–44. Retrieved 25 May 2022 at: https://www.facs.org/media/khro23pr/puf_data_dictionary_2017.pdf.

31. Fritz AG. International Classification of Diseases for Oncology: ICD-O. 3rd edn. Geneva: First revision. World Health Organization; 2013.

32. (HCUP) HHCaUP. Healthcare Cost and Utilization Project (HCUP) Recommendations for Reporting Trends using ICD-9-CM and ICD-10-CM/PCS Data. 2017. Retrieved 1 October 2021 at https://www.hcup-us.ahrq.gov/datainnovations/HCUP_RecomForReportingTrends_070517.pdf.

33. Mandelbaum AD, Thompson CK, Attai DJ, et al. National trends in immediate breast reconstruction: an analysis of implant-based versus autologous reconstruction after mastectomy. Ann Surg Oncol. 2020;27:4777–85. [PubMed: 32712889]

34. National Cancer Institute STM. Types of Registries. Retrieved 1 October 2021 at https://training.seer.cancer.gov/registration/types/.

35. 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 Database. J Clin Oncol. 2009;27:4177–81. [PubMed: 19636004]

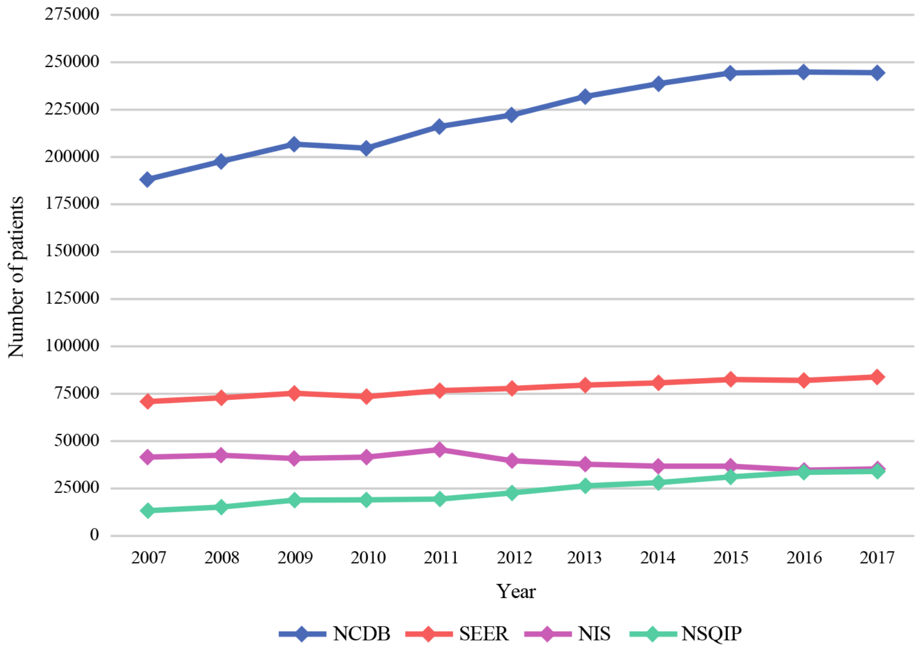

FIG. 1.

Breast cancer patients per year by database

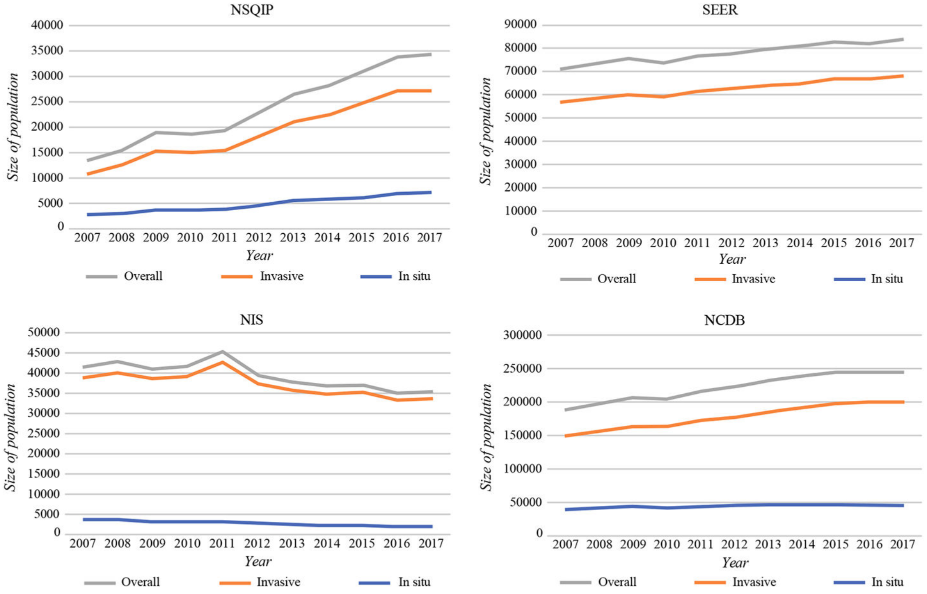

FIG. 2.

Invasive versus in situ breast cancer cases per year by database

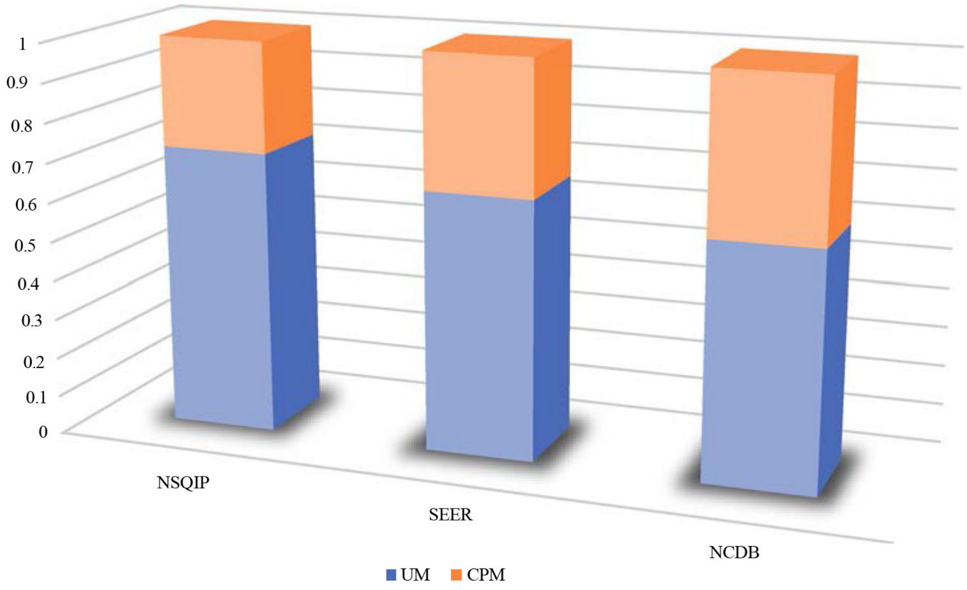

FIG. 3.

Proportion of unilateral mastectomy (UM) and contralateral prophylactic mastectomy (CPM) rates by dataset (2017)

TABLE 1

Obtaining the data, inclusion criteria, and goals by database

	NSQIP	NIS	SEER program	NCDB
Access	Free and available to those affiliated with an NSQIP-participating hospital	Available for purchase	Free and accessed via eRA Commons; available to all regardless of association with an institution	Available to those associated with CoC-accredited cancer programs; application required, which includes a research project proposal and letter of support from the Cancer Committee Chair of the CoC-accredited facility
Data files	Annual participant use data file with associated use guide and data dictionary	Annual data distributed in 4 files: Inpatient core (1988–2018) Diagnosis and procedure groups (2005–2015) Disease severity measures (2002–2018) Hospital (1988–2018)	Four SEER registry grouping files: SEER 9 (1975–2018) SEER 13 (1992–2018) SEER 18 (2000–2018) SEER 21 (2000–2018)	Cancer site-specific files provided
File format	Data files available in SPSS, SAS, and delimited text files	Distributed as ASCII-formatted data files that can be loaded into SAS, SPSS, Stata, or similar analysis software	SEER*Stat Software is required	Data files provided in flat text, SPSS, SAS, and Stata files
Inclusion criteria	Includes patients C18 years of age undergoing all major surgical procedures determined by CPT code and treated at >700 ACS-NSQIP-participating hospitals in 49 states in the U.S.	Includes patients of all ages admitted for any medical or surgical reason to one of the various HCUP-participating facilities (currently in 47 states), excluding rehabilitation centers and long-term acute care hospitals. Beginning in 2012, NIS transitioned from including all discharges from a sample of hospitals to a sample of all discharges (not a comprehensive inclusion of all patients). The NIS includes samples from the state inpatient databases (SIDs), and the data represented a 20% sample of discharges from U.S. hospitals	Includes patients of all ages from population-based cancer registries in the U.S. Beginning in 1975 with 7 geographic locations, SEER collects comprehensive data on patients with a cancer diagnosis. Data inclusion is based on geographic location and captures approximately 30% of newly diagnosed cancers in the U.S.	Includes patients of all ages with a cancer diagnosis, from >1500 CoC-accredited facilities in the U.S. NCDB captures >70% of newly diagnosed cancers in the U.S. Whereas data reporting was previously voluntary, inclusion criteria currently also involves a mandatory reporting of all cancers diagnosed at any COC-accredited facility
Goals	Originating in the 1980s out of concern regarding higher operative mortality rates in VA hospitals without an existing national standard for baseline comparison, the goal of NSQIP is to use risk-adjusted outcomes-based research to measure morbidity and mortality of surgical patients in order to improve the quality of care of surgical patients	The aim of the NIS database is for the data to be used for health care services and policy research, including “utilization of health services by special populations, hospital stays for rare conditions, variations in medical practice, healthcare cost inflation, regional and national analyses, quality of care and patient safety, impact of health policy changes, and access to care”	The goals of the SEER database and program are to collect comprehensive data on cancer diagnosed within the geographic areas targeted by the SEER registries. Because SEER is population-based and intended to be representative of the U.S. population, data are collected from strategically chosen regions to provide epidemiologically appropriate information. An additional goal is to report on U.S. cancer burden, incidence, mortality, and overall survival. SEER also aims to identify changes in patterns of cancer occurrences, temporal changes in cancer incidence, occurrences of iatrogenic cancers, and cancer prevention patterns/outcomes	The goals of NCDB include collecting accurate cancer data from >30% of U.S. hospitals in order to track and analyze patients with malignant disease, exploring trends in the treatment of cancer, facilitating improvement in the quality of cancer patient care, and improving outcomes

NSQIP National Surgical Quality Improvement Program, *NIS* Nationwide Inpatient Sample, *SEER* Surveillance, Epidemiology and End Results program, *NCDB* National Cancer Database, *eRA* electronic Research Administration, *CoC* Commission on Cancer, *CPT* current procedural terminology, *ACS* American College of Surgeons, *HCUP* Healthcare Cost and Utilization Project, *SIDs* State Inpatient Databases, *VA* Veterans Affairs

TABLE 2

Comparison of the databases

	NSQIP	NIS	SEER	NCDB
Sponsor/ association	Sponsored by the American College of Surgeons	Part of the HCUP sponsored by the AHRQ	Sponsored by the Surveillance Research Program in the National Cancer Institute's Division of Cancer Control and Population Sciences	Sponsored jointly by the American College of Surgeons and the American Cancer Society
Brief description	National, risk-adjusted, outcomes-based database that aims to measure morbidity and mortality to improve the quality of care of surgical patients	All-payer inpatient-care database, which reports estimates of inpatient utilization and outcomes; a 20% stratified sample of discharges from all HCUP-participating hospitals	Population-based cancer database; data inclusion is based on the geographic location and captures ~30% of newly diagnoses cancers	Hospital-based cancer database; data are collected from Commission on Cancer-accredited facilities in the U.S.; captures >70% of newly diagnosed cancer cases and ~30% of U.S. hospitals (>1500 of 5000)
Time period	2005–2019	1988–2018	1975–2018	1989–2017
Basic demographics	Sex, race (5 options), ethnicity (Hispanic/Latino), weight/height	Sex, race/ethnicity (5 options)	Sex, race (30 options), ethnicity (Hispanic/ Latino)	Sex, race (30 options), ethnicity (Hispanic/ Latino)
Age (years)	≥18	All ages	All ages	All ages
Socioeconomic, education, and insurance information	• None	• Expected primary payer • Median household income by zip code • Total hospital charges	• Median household income by zip code	• Primary payer • Median household income by zip code • Educational attainment in area of residence (estimated % in zip code without high school degree)
Comorbidities	• Comorbidities • Functional status • ASA classification • Preoperative laboratory values • Estimated probability of mortality and morbidity • Year of death	• Comorbidities (2002–2015) • All patient refined diagnosis-related groups (APR DRGs) • APR DRG mortality risk classification • APR DRG severity of illness classification	• None	• Charlson-Deyo comorbidity score
Facility Information	• None	• Control/ownership of hospital (e.g., government, private) • Bed size of hospital • Teaching status of hospital	• Urban-rural continuum of treatment facility • Type of reporting source (e.g., hospital, radiation center, laboratory, outpatient office, autopsy)	• Urban-rural continuum of treatment facility • Facility type (Community Cancer Program, Comprehensive Community Cancer Program, Academic-Research Program, Integrated Network Cancer Program)

	NSQIP	NIS	SEER	NCDB
		• Region of the hospital (4 categories)		• Facility location (9 categories) • Great circle distance (distance in miles between patient’s residence and reporting hospital)
Hospitalization Information	• Inpatient vs outpatient • Quarter of admission • Days from admission to operation • Days from operation to discharge • Length of hospital stay • Discharge destination • Readmission information	• Inpatient only • Admission weekday/ weekend and month • Elective vs non-elective admission • Length of hospital stay • Death during hospitalization • Discharge quarter • Discharge destination • No. of days from admission to each procedure • Transfer from or to another facility	• No information regarding specific hospitalizations	• Length of surgical inpatient stay • Readmission to the same hospital within 30 days after surgical discharge
Surgery Information	• Surgical specialty performing primary operation • Anesthesia type • Operative time • Level of resident in operating room • Elective vs emergent operation • CPT codes (21 codes available per patient) • RVU associated with each CPT code • Surgical wound closure • Wound classification • Postoperative outcomes and complications	• No. of procedures per discharge • ICD9 and ICD10 procedure codes (25 procedure codes available per patient) • NIS provides information to determine whether a procedure is uni- or bilateral. As of 2015, with the transition to ICD 10 procedure codes, researchers are also able to determine if a procedure was performed on the right or left. • The 25 procedure codes per patient provided in NIS allows researchers to indicate immediate oncoplastic breast procedures (performed at the time of lumpectomy) and immediate breast	• Site-specific surgery coding system (1 code per patient) • Reason for no surgery of primary site, if applicable • Scope of regional lymph node surgery involvement is available for certain cancers but is not available for breast cancer cases • SEER provides laterality of the disease (right vs left). Site-specific surgery codes indicate whether the patient had no surgical intervention, lumpectomy, or mastectomy (with or without contralateral prophylactic mastectomy, and with or without breast reconstruction during the first course of treatment). Immediate and delayed post-lumpectomy oncoplastic procedure	• Site-specific surgery coding system (1 code per patient) • Surgical diagnostic and staging procedure information • Days from diagnosis to first surgical procedure and days from diagnosis to definitive surgical procedure • Surgical margins of primary site • Scope of regional lymph node surgery • Surgical procedure to sites other than primary site • Reason for no surgery of primary site, if applicable • 30- and 90-day mortality after surgery • NCDB provides laterality of the disease (right vs left). Site-specific surgery codes

	NSQIP	NIS	SEER	NCDB
	• Unplanned reoperation information • NSQIP provides information to determine whether a procedure is uni- or bilateral, but does not indicate if a unilateral procedure is on the right or left • The 21 CPT codes per patient provided in NSQIP allows researchers to indicate immediate oncoplastic breast procedures (performed at the time of lumpectomy) and immediate breast reconstruction procedures (performed at the time of mastectomy). Researchers are also able to examine delayed breast reconstruction procedures, but the previous mastectomy procedure details will not be associated with that patient encounter	reconstruction procedures (performed at the time of mastectomy). Researchers are also able to examine delayed breast reconstruction procedures, but the previous mastectomy procedure details will not be associated with that patient encounter unless it occurred during the same inpatient admission (which is unlikely)	information is not provided. Delayed breast reconstruction information is not provided	indicate whether the patient had no surgical intervention, lumpectomy, or mastectomy (with or without contralateral prophylactic mastectomy, and with or without breast reconstruction during the first course of treatment). Immediate and delayed post-lumpectomy oncoplastic procedure information is not provided. Delayed breast reconstruction information is not provided
Diagnosis Information	• Diagnosis code (ICD-9 and ICD-10 diagnosis codes); 1 code per patient • Laterality of disease is not provided • ICD 9/10 diagnosis codes may indicate genetic susceptibility to breast malignancy of family history or breast cancer, but specific BRCA1/2 information is not provided	• Diagnosis code (ICD9 and ICD10 diagnosis codes) ; 40 codes available per patient • No. of diagnoses per discharge • MDC on day of discharge • Laterality of *disease* is not provided (laterality of *procedure* is provided beginning in 2015). • ICD 9/10 diagnosis codes may indicate genetic susceptibility to breast malignancy or family history of breast cancer, but specific BRCA1/2	• See oncologic data	• See oncologic data

	NSQIP	NIS	SEER	NCDB
		information is not provided.		
Oncologic Information	• No specific oncologic information other than the diagnosis code	• No specific oncologic information other than diagnosis code	• Uses ICD-O-3 classification • Behavior, histology, grade, tumor size • Laterality of disease if applicable • Method of diagnostic confirmation (e.g., histology, cytology) • No. of primary tumors (in situ and/or invasive) • Regional lymph nodes examined (number) and positivity of regional lymph nodes • Presence of distance metastases at time of diagnosis (bone, brain, liver, lung, distant lymph nodes, or other) • Combined (clinical and pathologic) TNM staging • Site-specific factors for breast cancer (see Supplemental Table for additional information) • ER and PR assays • HER2: summary result of testing • BRCA1/2 status is not provided • Treatment details–surgery, systemic therapy* (and timing relative to surgery), radiation therapy (and timing relative to surgery) • *Regarding systemic therapy, SEER does not differentiate between chemotherapy, endocrine therapy, and immunotherapy.	• Uses ICD-O-3 classification • Behavior, histology, grade, tumor size • Laterality of disease if applicable • Method of diagnostic confirmation (e.g., histology, cytology) • No. of primary tumors (in situ and/or invasive) • Regional lymph nodes examined (number) and positivity of regional lymph nodes • Presence of distance metastases at time of diagnosis (bone, brain, liver, lung, distant lymph nodes, or other) • Separate clinical TNM staging and pathologic TNM staging • Site-specific factors for breast cancer (see Supplemental Table for additional information) • ER and PR) assays • HER2: summary result of testing • BRCA1/2 status is not provided. • Location of initial diagnosis vs location of treatment • Treatment details-surgery, systemic therapy (and timing relative to surgery), radiation therapy (and timing relative to surgery) • Detailed information on timing, type, and dose of radiation (see Supplemental Table) • Detailed information on timing and type of systemic therapy (chemotherapy, endocrine therapy, and immunotherapy) (see Supplemental Table)

NSQIP	NIS	SEER	NCDB
		- Vital status of patient (overall survival) - Survival (months)	- Palliative care treatment information - Vital status of patient (overall survival) - Survival (months)

[*] The asterisks refer to extra information/added information

NSQIP National Surgical Quality Improvement Program, *NIS* Nationwide Inpatient Sample, *SEER* Surveillance, Epidemiology and End Results program, *NCDB* National Cancer Database, *HCUP* Healthcare Cost and Utilization Project, *AHRQ* Agency for Healthcare Research and Quality, *ASA* American Society of Anesthesiologists, *APR DRG* all patient refined diagnosis-related group, *CPT* current procedural terminology, *RVU* relative value unit, *ICD* International Classification of Diseases, *MDC* major diagnostic category, *ER* estrogen receptor, *PR* progesterone receptor, *HER2* human epidermal growth factor receptor

TABLE 3

Demographics of breast cancer patients by dataset

	NSQIP (%)	NIS (%)	SEER (%)	NCDB (%)
Breast cancer patient population size	262,103	433,380	855,288	2,439,315
Age (years)				
<40	5.30	5.60	4.33	4.40
40–49	18.15	15.49	16.29	16.32
50–59	25.76	22.98	24.45	24.72
60–69	26.60	24.54	26.65	27.17
70–79	16.60	18.21	17.97	18.14
≥80	7.39	13.16	10.32	9.24
Unknown/not reported	0.19	0.02	N/A	N/A
Ethnicity				
Hispanic	5.73	7.51	10.84	5.47
Non-Hispanic	84.63	92.49	89.16	90.31
Unknown	9.64	N/A	N/A	4.22
Race				
White	72.15	62.56	68.43	82.49
Black/African American	10.62	14.40	10.95	11.69
Asian, Pacific Islander, or Hawaiian Native	5.04	2.74	8.64	3.62
American Indian or Alaska Native	0.47	0.40	0.53	0.28
Other	0.00	10.08	10.84	0.89
Unknown	11.72	9.81	0.61	1.03

NSQIP National Surgical Quality Improvement Program, *NIS* Nationwide Inpatient Sample, *SEER* Surveillance, Epidemiology and End Results program, *NCDB* National Cancer Database, *N/A* not available

TABLE 4

Unilateral mastectomy (UM) and contralateral prophylactic mastectomy (CPM) rates by dataset (2017)

	Total no. of mastectomies	No. of UM	UM rate (%)	No. of CPMs	CPM rate (%)
NSQIP	16,494	11,831	71.7	4663	28.3
SEER	23,540	15,474	65.7	8066	34.3
NCDB	68,291	40,879	59.9	27,412	40.1

NSQIP National Surgical Quality Improvement Program, *SEER* Surveillance, Epidemiology and End Results program, *NCDB* National Cancer Database

TABLE 5

Usefulness of each of the four databases for various clinical research questions

	NSQIP	NIS	SEER	NCDB
Rates of lumpectomy vs mastectomy over time	X	X (before 2015)	X	X
Rates of contralateral prophylactic mastectomy over time	X	X (before 2015)	X	X
Trends in immediate breast reconstruction	X	X (before 2015)	X	X
Radiation burden in breast reconstruction patients				X
Relationship between insurance status and breast cancer treatment		X		X
Hospital costs associated with breast cancer patients		X		
Comorbidity profiles of breast cancer patients	X	X*		X*
Postoperative complication profile of surgical breast cancer patients	X			
Effects of anesthesia type on postoperative outcomes in breast cancer surgeries	X			
Patterns in use of chemotherapy, endocrine therapy, and immunotherapy for breast cancer patients				X
Pathologic vs clinical lymph node status of breast cancer patients			X*	X
Effects of geographic location on breast cancer treatment patterns		X*		X
Granular race data on breast cancer patients			X	X
Length of hospital stay for surgical breast cancer patients	X	X		X
Patterns in breast cancer treatment based on granular oncologic data (tumor behavior, grade, histology, ER/PR/HER2 status)			X	X

NSQIP National Surgical Quality Improvement Program, *NIS* Nationwide Inpatient Sample, *SEER* Surveillance, Epidemiology and End Results program, *NCDB* National Cancer Database, *ER* estrogen receptor, *PR* progesterone receptor, *HER2* human epidermal growth factor receptor 2

X = appropriate database for this clinical inquiry

X* = less ideal