









For numbered affiliations see end of article.

#### Correspondence to

Dr Katherine Hicks-Courant;  
katherine.hicks-courant2@  
pennmedicine.upenn.edu

Received 5 May 2024  
Accepted 22 June 2024  
Published Online First  
23 July 2024

# Secondary databases in gynecologic cancer research

Katherine Hicks-Courant <sup>1,2</sup>, Emily Meichun Ko <sup>2,3,4</sup>, Koji Matsuo <sup>5,6</sup>, Alexander Melamed,<sup>7,8</sup> Dimitrios Nasioudis <sup>2</sup>, Jose Alejandro Rauh-Hain,<sup>9</sup> Shitanshu Uppal,<sup>10</sup> Jason D Wright <sup>11</sup>, Pedro T Ramirez <sup>12</sup>

## ABSTRACT

Observational and cohort studies using large databases have made important contributions to gynecologic oncology. Knowledge of the advantages and potential limitations of commonly used databases benefits both readers and reviewers. In this review, researchers familiar with National Cancer Database (NCDB), Surveillance, Epidemiology, and End Results Program (SEER), SEER-Medicare, MarketScan, Healthcare Cost and Utilization Project (HCUP), National Surgical Quality Improvement Program (NSQIP), and Premier, describe each database, its included data, access, management, storage, highlights, and limitations. A better understanding of these commonly used datasets can help readers, reviewers, and researchers to more effectively interpret and apply study results, evaluate new research studies, and develop compelling and practice-changing research.

## INTRODUCTION

While randomized controlled trials are considered the gold standard for cancer research, many observational and cohort studies using both primary<sup>1–6</sup> and secondary<sup>7–11</sup> data have contributed greatly to gynecologic oncology. Primary data are collected by a research team to answer a particular research question via a particular study design. Examples include institutional medical record studies and many prospective and retrospective cohort studies. Secondary data are observational data, collected for purposes other than the particular research question and design at hand. Examples include unplanned secondary analyses of data from completed trials, tumor registry data, and insurance claims data.

Secondary data are unique, as they can be used to generate hypotheses, provide real-world results in the context of randomized controlled trials, and investigate questions regarding rare diseases or rare outcomes that would be otherwise difficult to study in controlled settings. They can also be used to understand the effects of health reforms, policy changes, and changes and trends in practice over time. Secondary data provide information that is obtainable for a smaller cost than would be required to conduct a randomized controlled trial, and data collection is often simpler than that required for studies that use primary data. Secondary databases can often provide large amounts of data and patient

samples. Furthermore, many secondary databases are publicly available, making research more affordable and accessible to many researchers, such as trainees, early-career and international investigators. Studies of secondary data contribute significantly to the field, making up 10% of the most cited articles of the *International Journal of Gynecological Cancer*<sup>12</sup> and more than 15% of the most cited articles of the journal *Gynecologic Oncology* since 2020.<sup>13</sup>

Knowledge of the advantages and potential limitations of secondary data may be of benefit to both readers and reviewers, enabling them to have a clear perspective of how to interpret results published from secondary databases. Similarly, researchers must be aware of the limitations of each secondary database, especially variables that are missing, which can limit the study design, extent of analyses, and interpretation of results. In addition, it is important to highlight that the quality of data entered in secondary databases cannot always be verified and the qualifications of those entering the data is not routinely known. Furthermore, some may question the validity of the data entered in secondary databases as there are no clearly defined criteria for data auditing. Lastly, studies using secondary databases may identify association, but proving causation is very challenging, requiring complex study design and methodology.

The goal of this descriptive review article is to highlight secondary databases from the USA that are commonly used in research in the field of gynecologic oncology. Cancer registry databases include the National Cancer Database (NCDB), and Surveillance, Epidemiology, and End Results Program (SEER). Additionally, the National Institutes of Health, which oversee the SEER Registry, have created a Medicare-insurance claims database linked with SEER, known as the SEER–Medicare Database. Another insurance-claims database comprising commercial (private) insurance is MarketScan. Hospital administrative databases include the Healthcare Cost and Utilization Project (HCUP) and Premier. Lastly, a surgical outcomes database created by the American College of Surgeons, the National Surgical Quality Improvement Program (NSQIP), has been frequently used. A better understanding of these commonly used databases can help readers, reviewers, and researchers



© IGCS and ESGO 2024. No commercial re-use. See rights and permissions. Published by BMJ.

**To cite:** Hicks-Courant K, Ko EM, Matsuo K, *et al.* *Int J Gynecol Cancer* 2024;**34**:1619–1629.

## Review

**Table 1** Owner, data, access, and cost by database

	NCDB	SEER	SEER–Medicare	MarketScan	HCUP	Premier	NSQIP
Owner	ACS	NCI	NCI and CMS	Merative	AHRQ	Premier	ACS
Data source	Cancer Registry	Cancer Registry	Cancer Registry; insurance claims	Insurance claims and supplemental source for subsets	Administrative from hospitals	Administrative from hospitals	Medical records
Years available (as of 11/2023)	2004–2022	1975–2020	1999–2019	2006–2022	1988–2020 (NIS), 2016–2020 (NASS)	1997–2020	2005–2020
Delay to data availability	12–24 months	36–48 months	60 months	12–24 months	24–36 months	36–48 months	36–48 months
Cost*	Free	Free	US\$ (10–15K)	NA	US\$ (~1K per 1 year)	NA	Free
Application for use	Yes	Yes	Yes	Yes	Yes	Yes	Yes

\*Costs are estimates and may vary.

ACS, American College of Surgeons; AHRQ, Agency for Healthcare Research and Quality ; CMS, Centers for Medicare and Medicaid Services; HCUP, Healthcare Cost and Utilization Project; NASS, Nationwide Ambulatory Surgical Sample ; NCDB, National Cancer Database; NCI, National Cancer Institute; NIS, National Inpatient Sample; NSQIP, National Surgical Quality Improvement Program ; SEER, Surveillance, Epidemiology, and End Results Program.

to interpret and apply study results more effectively, evaluate new research studies, and develop compelling and practice-changing research.

### METHODS

Gynecologic oncologists who frequently use secondary data to conduct research were identified and invited to contribute as authors. As a group, the authors then selected databases that are commonly used in gynecologic oncology research to be included in this review. For each database, we have gathered and report information on specific features, such as available content, limitations, acquisition, data-use agreements, and examples of research questions applied. A gynecologic oncology researcher contributed a section about a database with which they have extensive experience conducting and publishing research. Owner, data, access, and cost by database are described in [Table 1](#). Data content available in each database is described in [Table 2](#).

### DATABASES

#### National Cancer Database

The National Cancer Database (NCDB) is one of the largest cancer registries in the world providing real-world data on the management and outcomes of patients with malignancies in the USA.<sup>14</sup> It has been established and curated jointly by the American Cancer Society and the Commission on Cancer of the American College of Surgeons. It is a hospital-based database that prospectively captures data from patients who received cancer care at a Commission on Cancer-accredited hospital. The database covers more than 1500 facilities (approximately 30% of all hospitals in the USA) and approximately 70% of all newly diagnosed cancer cases in the USA.

Dedicated registrars extract and collect patient data from medical charts.<sup>15</sup> Multiple electronic automated checks alert data registrars to missing data fields and internally inconsistent data to enhance data integrity. In addition, frequent auditing of the data ensures

quality control and improvement. For patients who receive care at multiple facilities, only the most complete record is maintained in the database. Data are de-identified and individual patients and providers cannot be identified. However, each hospital is assigned a unique identifier. To protect the privacy of patients, adult files include only those aged >18 years. Furthermore, any reporting of groups of fewer than 10 patients must be masked in order to protect confidentiality. For example, if a sample contains only eight patients with a certain diagnosis, it could not be published as ‘8’ but could be reported as ‘<10.’

#### Data Included

All major gynecologic malignancy cases are collected in the NCDB in separate dataset files (uterus, ovary, vagina, vulva, cervix) that are renewed each year. The most up to date dataset includes cases diagnosed between 2004 and 2021. Data prior to 2004 have been collected but are not available in the public use files. The database includes basic demographic variables (patient age, race, type of insurance, median household income at postal code of residence), type of treatment facility, pathology variables (histology, tumor grade, number and status of regional lymph node examined), details of surgery performed (type of surgery, performance of lymphadenectomy, mode of surgery, inpatient hospital stay, peri-operative mortality) as well as receipt of chemotherapy and radiotherapy, follow-up, and patient status. A detailed variable dictionary is available online for prospective researchers.<sup>16</sup>

#### Access, Management, and Storage

Researchers can submit their proposals after obtaining a letter of support from the chairman of the cancer committee at the researcher’s institution through a dedicated website. A data analyst is required and must be located at the facility of the primary investigator. Data are free of cost; however, they are available only for researchers affiliated with one of the approximately 1500 participating Commission on Cancer institutions and thus are not available for researchers outside of the USA. The NCDB participant user files

**Table 2** Data availability by database

	NCDB	SEER	SEER– Medicare	MarketScan	HCUP	Premier	NSQIP
<b>Demographic data</b> eg, age, race, socioeconomic status	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<b>Primary tumor data</b> eg, site, histopathology, stage	Yes	Yes	Yes	Yes (site only)	Yes (site only)	Yes (site only)	Yes (site only)
<b>Treatment data</b> eg, systemic therapy, radiation	Limited	Limited	Yes	Yes	No	Yes	No
<b>Peri-operative claims</b>	Yes	Limited	Yes	Yes	Yes	Yes	Yes
<b>Recurrence data</b> eg, recurrence status, timing of recurrence	No	No	No	No	No	No	No
<b>Survival data</b> eg, progression-free survival, overall survival	Yes	Yes	Yes	No	Limited*	No	Limited†
<b>Cause of death</b>	No	Yes	Yes	No	No	No	Yes
<b>Inpatient claims</b>	No	No	Yes	Yes	Yes (NIS)	Yes	No
<b>Outpatient claims</b>	No	No	Yes	Yes	Yes (NASS)	Yes	No
<b>Physician-level data</b> eg, specialty, years of practice	No	No	Yes	Yes	No	Yes	Limited
<b>Hospital-level data</b> eg, teaching hospital status, urban status, size	Limited	No	Yes	Yes	Yes	Yes	No
<b>Cost data</b>	No	No	Yes	Yes	Yes	Yes	No
<b>Quality-of-life data</b>	No	No	No	No	No	No	No

\*In-hospital mortality for index admission (National Inpatient Sample).

†30-day mortality is available.

HCUP, Healthcare Cost and Utilization Project; NASS, Nationwide Ambulatory Surgical Sample; NCDB, National Cancer Database; NIS, National Inpatient Sample; NSQIP, National Surgical Quality Improvement Program; SEER, Surveillance, Epidemiology, and End Results Program.

are de-identified and Health Insurance Portability and Accountability Act compliant; however, local institutional review board review of the proposed research is required.

Following review of the research plan and approval of the proposed application, the researcher can sign the data sharing agreement and receive the data file requested in a format that can be downloaded and processed by all major statistical software programs. Data should be securely maintained until the proposed research plan is completed while only the primary and co-investigators listed in the proposal can have access to the data. Access or distribution of the data to outside parties not listed in the original agreement is prohibited. The American College of Surgeons and the Commission on Cancer do not verify, and are not responsible for, the analytical or statistical methodology employed in each study, or the conclusions drawn from these data.

#### Database Highlights

The NCDB is one of the most comprehensive cancer databases with a mechanism that permits continuous collection of high-quality data. It provides a unique platform to investigate patterns of care of patients with gynecologic malignancies in the USA and provide hypothesis generating data. The large number of patients included across multiple sites and histologic subtypes is a major advantage, while a relatively large number of variables collected permit meaningful analyses.

#### Database Limitations

The major limitation of the NCDB, similar to other observational databases, is patient selection bias since it includes only patients managed at Commission on Cancer accredited hospitals that are

required to demonstrate compliance with certain quality metrics of cancer-care delivery. The NCDB is hospital-based and thus not designed to represent the total US population overall. Moreover, other limitations include lack of genomic or imaging data, or information on tumor recurrence and cause of death, precluding calculation of recurrence, disease-free or cancer-specific survival. However, time from cancer diagnosis to last contact or patient death is provided and permits calculation of overall survival. Capturing patient's functional status and comorbidity profile is limited since only the Charlson-Deyo Comorbidity Index is provided, while other functional status indexes such as Eastern Cooperative Oncology Group are not available. In addition, details on chemotherapy administration (specific agents used, dose, and number of cycles) and specific details on the surgical procedures performed are lacking. Accuracy of staging information assigned to each case may also vary since over-coding has been described in other tumor sites, and care should be taken to avoid staging misclassifications when analyzing data for certain sites such as the cervix.<sup>17</sup> Similarly, accuracy of information related to radiation therapy administration may be limited.<sup>18</sup> Lastly, for certain variables depending on disease site there is a high prevalence of missing data that could potentially bias conclusions.<sup>19</sup>

#### Surveillance, Epidemiology, and End Results (SEER) Program

The Surveillance, Epidemiology, and End Results (SEER) Program was created following the National Cancer Act and is maintained by the National Cancer Institute. SEER is administered and funded by the Division of Cancer Control and Population Sciences at the National Cancer Institute. SEER currently collects and publishes

## Review

data on cancer incidence, prevalence, and survival from population-based cancer registries covering approximately one-third of the US population. It provides incidence, overall survival, and mortality data for histopathologic cancer subtypes.<sup>20 21</sup>

The SEER Database has been rigorously studied and validated. It is considered to be the gold standard of cancer registries and has become the standard for data quality among international cancer registries. SEER regions are included based on their capacity to maintain a high-quality cancer registry and having populations representative of the total US population.<sup>22</sup> Initially, seven registries were included, and incrementally were expanded to the current 22 registries. The SEER population tends to have an over-representation of people born in foreign countries and those residing in urban areas. Furthermore, there is a deliberate over-sampling of specific racial and ethnic minority groups, in order to improve the diversity of the SEER population and allow for evaluation of differences by race/ethnicity.<sup>23</sup> The population covered by SEER is representative of the general US population in regard to measures of poverty and education.

### Data Included

SEER data are now available for the years 1975–2019. Data collected include all primary invasive cancers and in situ carcinomas. SEER includes demographic information, such as age at diagnosis, date of diagnosis, gender, race/ethnicity, and county of residence. Additionally, it includes primary tumor information, such as stage, and pathologic findings. Cancer data are updated annually to capture vital status, survival time, and cause of death. Vital status is confirmed by linkage to the National Death Index and with supplemental data on date of last known contact obtained by medical record abstraction. Tumor recurrence data are currently not collected and therefore progression-free survival, correlates of local, regional, and distant control, and the effectiveness of salvage therapy cannot be assessed. For some non-gynecologic cancers, biomarkers are also included in most recent years. Although some information about upfront treatment is included, several important classes of data regarding treatment, including detailed chemotherapy (ie, specific drugs used and dosing) and radiotherapy information, are unavailable.

A recent study comparing SEER with SEER–Medicare data, reported an overall sensitivity of 80% for SEER radiotherapy data and 68% for SEER chemotherapy data.<sup>24</sup> Details of surgical management during the first course of treatment are extracted from health records. However, specific details about the surgery are unavailable (as an example, whether a hysterectomy was laparoscopic or open). Additionally, the intent of the surgeon is not captured, and therefore it is uncertain whether a patient underwent surgery for palliative or curative intent.<sup>23</sup>

### Access, Management, and Storage

Access to the SEER website ([www.seer.cancer.gov](http://www.seer.cancer.gov)) is unrestricted. SEER data are free to use, but those who want to use it must get permission from SEER. This is accomplished through submitting a request on the SEER website and signing a data-use agreement.

SEER data can be accessed through SEER online software, or through SEER\*Stat and then downloaded in tabular form to allow for additional analyses. The online software includes SEER\*Stat, SEER PREP, JointPoint, and the Health Disparities Calculator. Use of all

databases and software is free and can be accessed by completing the online application form. SEER\*Stat is a versatile tool for making queries of SEER data. It is particularly useful for easily calculating age-adjusted rates. The Cancer Statistics Review option provides summaries of all cancers and site-specific cancers. The summaries include 5-year survival data, relative survival as compared with the general population, and survival by tumor stage and race/ethnicity.

In addition to cancer datasets, other datasets in the SEER program are the standard population data for SEER areas, US mortality data, and US population data linked to a Census Tract Socioeconomic Status (SES) Index or to county attributes. The specialized Census Tract-level SES and Ruralness Database (2006–2018) has five census tract-level attributes: two SES quintiles, two rurality variables, and persistent poverty. These data can be used for matched analyses with SEER cancer data.

### Database Highlights

The SEER Database is an excellent, increasingly used resource for clinical cancer investigation. Unlike the NCDB, SEER is a population-based database. Population-based databases like SEER are important for understanding the implications of pathology diagnoses across demographic groups, geographic regions, and time, and they provide unique insights into the practice of oncology in the USA that are not attainable from other sources.<sup>25 26</sup>

SEER is useful in the analysis of national demographic differences and regional trends in the diagnosis, treatment, and outcome of all cancers.<sup>23 27 28</sup> Furthermore, comparative effectiveness studies regarding the benefits of cancer treatment can be done, with the understanding that inherent limitations in the SEER Database should be acknowledged.<sup>24</sup> Details of the development of prognostic models based on initial demographic and clinicopathologic variables have also been published.<sup>29</sup> Because of its ability to provide large sample sizes, SEER also allows the opportunity to analyze rare diseases and risk of second malignancies.

### Database Limitations

While the SEER Database is a valuable tool for clinical cancer research, several limitations should be considered when analyzing and interpreting results from SEER. There are minimal data about non-cancer information, such as comorbidities, limiting the ability to adjust for some potential confounders. Many limitations revolve around under-reported and incomplete data regarding adjuvant therapy (specific chemotherapy agents and dosing, type of radiation therapy, fractionation, and dosing), unrecorded variables, variations in data reporting, migration of patients in and out of SEER registry areas, unmeasured confounding, and selection bias.<sup>23 30</sup> Finally, caution should be taken when using treatment information owing to limitations of the granularity of treatment agents, modalities, and dosing from SEER.<sup>24</sup> Detailed specific information about the type, dose, and duration of chemotherapy, radiation therapy, are not within the purview of SEER data collection. Knowledge gaps pertaining to treatment and follow-up may arise when individuals transition between SEER and non-SEER regions, potentially introducing bias into conclusions regarding cancer behavior. Notably, SEER does not capture data related to risk-reduction procedures (eg, risk-reducing salpingo-oophorectomy) or organ removal for non-cancer indications (eg, hysterectomy). Nevertheless, it is



possible to deal with this limitation by leveraging supplementary data sources discussed in this review.

### SEER–Medicare Database

The SEER–Medicare linked data resource is a national population-based data source of Medicare beneficiaries with cancer. This database combines Medicare claims information with the SEER cancer registry, such that each patient has full cancer-specific information linked with all their Medicare-billed medical care. The SEER–Medicare database is owned by the National Cancer Institute and Centers for Medicare and Medicaid Services. It is directly overseen by the Information Management System (National Cancer Institute's information contractor).

#### Data Included

The SEER–Medicare database specifically reports on Medicare patients who have been diagnosed with cancer. The SEER cancer registry collects cancer incidence data from population-based cancer registries covering approximately 50% of the US population. The paired Medicare data captures all insurance claims reimbursed by fee-for-service Medicare, but not any that are reimbursed by private payers and third-party payers. This database does not capture any treatment received before a patient was enrolled on Medicare, after a patient discontinues Medicare, or during lapses in Medicare coverage. Currently, the SEER–Medicare database contains information on cancers diagnosed between 1999 and 2019.

SEER–Medicare includes information on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, first course of treatment, vital status (death date), and cause of death. It captures treatments received in real-world practice, including the type of hysterectomy and surgical staging based on Current Procedural Terminology (CPT) codes (laparoscopic, open, cytoreductive surgery, radical hysterectomy, lymph node dissection, lymph node biopsy, type of colectomy), chemotherapy infusions through Healthcare Common Procedure Coding System infusion and drug codes (J codes, NDC codes), and radiation therapies through CPT codes (external beam, intensity-modulated radiation therapy, brachytherapy).

SEER–Medicare also captures oral cancer-directed therapies (PARP (poly-(ADP-ribose)-polymerase) inhibitors, mTOR (mammalian target of rapamycin) inhibitors) and other prescriptions. SEER–Medicare also provides reimbursement costs for all Medicare claims, allowing for calculation of global and specific costs of care. It does not, however, capture any out-of-pocket reimbursements, or payments made by individual institutions or individuals to any organizations outside of Medicare. It also captures care rendered by individual providers (who can be tracked through a designated de-identified provider variable), as well as hospital site and type where a service was rendered. It provides patient demographics, though items such as zip code require a special additional application. Health provider-based demographics can also be cross-linked to an American Medical Association-Physicians file, via an additional application and fee submitted to the American Medical Association. Dates are available for billed claims for all treatments. Researchers receive data for all Medicare claims for each patient during the years of interest, including Medicare claims unrelated to cancer care.

#### Access, Management, and Storage

SEER–Medicare data are not publicly available. Access requires a data use agreement application, for a specific research question. The purpose of the data use agreement is to ensure confidentiality of patients and providers in the SEER registry. The purchase fee for a comprehensive cohort of gynecologic cancers may range from US\$10 000 to US\$15 000, depending on which files are requested. Application review and approval may take several months. The SEER–Medicare application instructions and fee for obtaining SEER–Medicare data, as well as a comprehensive summary of database files, data dictionaries explaining the variables included, and available support services, are available online.<sup>31</sup>

The database itself is stored by the US government and provided to the researcher in a secure password-protected fashion after data use agreement approval. The principal investigator must provide a detailed storage and protection plan for the data (storage on a secure password-protected server). Data must be stored at the institution of the principal investigator listed on the agreement and may not be transferred or shared by other institutions. Data use agreements are 5 years from the date of approval and are not typically extendable. At the conclusion of the agreement, all data including analytic files must be destroyed.

The data are stored in separate files including the Cancer File, and multiple Medicare claims files organized by inpatient, outpatient, drug, home health agency, durable medical equipment, hospice, hospital, census tract, and zip-code characteristics. Patients are identified by a unique identifier across these files to permit linking and merging of the data for analysis. Thus, extensive computer programming and coding is required to collate and merge information to answer a particular research question. The data files can be analyzed using a statistical software program such as SAS, Stata, R, or other similar computing capabilities.

#### Database Highlights

SEER–Medicare is particularly useful for identifying and validating actual treatments received by patients immediately preceding or following a cancer diagnosis across the USA. For example, one can determine the chemotherapy drug infused (paclitaxel, carboplatin, cisplatin, or other) during adjuvant therapy, whether it was given concurrently or serially with or without radiation, for endometrial cancer.<sup>32</sup> It allows researchers to conduct research questions in rarer but more aggressive subtypes of gynecologic cancers, such as uterine carcinosarcoma.<sup>33</sup> It also permits capture of rare outcome events, such as deep venous thromboembolism, or stroke, that might be difficult to study using smaller institutional or local databases.<sup>34</sup> It allows for evaluation of outcomes related to cancer and treatment-based factors, such as overall survival, and cancer-specific survival. It also permits analyses regarding healthcare use, national and regional patterns of care, and cost of care rendered.<sup>35</sup>

#### Database Limitations

The SEER–Medicare database has a population of patients with cancer that is predominantly over 65 years of age because SEER–Medicare specifically reports on fee-for-service Medicare patients who have been diagnosed with cancer. Nearly all Medicare enrollees are 65 years or older because adults in the USA may start enrollment at age 65 according to federal regulations. A small subset of adults who are younger than 65 may also enroll in Medicare, if they qualify

## Review

based on meeting criteria for a disability, end-stage renal disease, or have amyotrophic lateral sclerosis. Thus, SEER–Medicare is less useful for cancers that primarily occur in younger populations. The SEER–Medicare database does not include patients enrolled in Medicare Advantage plans. As enrollment in Medicare Advantage plans increased to over 50% in 2023,<sup>36</sup> a smaller proportion of patients are included in SEER–Medicare.

SEER–Medicare does not specifically capture measures regarding functional status or quality of life. It also does not directly capture metastases occurring after diagnosis. Thus recurrence, progression, and derived outcomes, such as disease-free survival, are not available. It does not acquire intent of treatment, physician treatment recommendations, or patient preferences related to treatment. Pathology-specific limitations include lack of surgical margin status and lymphovascular space invasion. It also does not include laboratory or imaging test results, body mass index, or American Society of Anesthesiology score. SEER–Medicare does not directly obtain treatment-related toxicities, although acute events can be identified if a clinician has billed for it or if patient received treatment for it. It does not necessarily specify the start date of chronic conditions (since that could have started in years prior to a patient enrolling on Medicare). SEER–Medicare provides a thorough explanation of the limitations of its dataset.<sup>37</sup>

To protect confidentiality, any reporting of groups of fewer than 11 patients must be masked. For example, if a sample contains only eight patients with a certain diagnosis, it could not be published as '8' but could be reported as '<11.' When working with specific gynecologic cancer populations, some of these resultant variable fields may frequently be quite small, and care must be taken to report these in a confidentially appropriate fashion. SEER–Medicare provides guidance on appropriate reporting of data for small cell counts.<sup>38</sup>

### MarketScan Database

Merative MarketScan is a widely used commercial source of healthcare data composed of insurance claims for individuals enrolled in employer-sponsored commercial health insurance plans and supplemental Medicare plans, as well as Medicaid beneficiaries from 11 states.<sup>39</sup> MarketScan data include claims for more than 270 million unique patients. Following a 2022 reorganization, MarketScan data are owned and licensed by Merative, a standalone healthcare data and analytics company that was previously a division of IBM.<sup>39</sup> Merative provides services to employers, health plans, and state Medicaid agencies who agree to contribute data to MarketScan. These data are standardized and de-identified and can be licensed for commercial or academic research.

### Data Included

MarketScan covers all cancer types, and includes records dating back to 1995.<sup>39</sup> Data are updated regularly with an approximate 12–24 month delay to data availability. As of 2023, the most recent year available was 2022. MarketScan is one of the most contemporaneous large data sources in the US. MarketScan data include claims submitted by hospitals, outpatient clinics, pharmacies, and other healthcare facilities. Claims data include *International Classification of Disease (ICD)* and *Current Procedural Terminology (CPT)* codes as well as charges, reimbursement and out-of-pocket costs.<sup>40</sup> It includes health plan enrollment and demographic data,

claims, and benefit plan information. For some populations, data on disability and workplace absence, inpatient drug administration, dental care, and laboratory results are also available through linkages to specialized databases. Custom linkages are also possible.

### Access, Management, and Storage

To conduct research using MarketScan, investigators and their intuitions must work with Merative (previously IBM Watson Health and Truven) to obtain a license. Cost can be prohibitive to researchers without extramural funding or significant institutional support and range up to US\$50 000. Some academic intuitions may provide MarketScan access to faculty through an existing license. Researchers desiring to access these data should contact Merative through their website.<sup>41</sup> Licenses are finite and must be renewed after expiration. Furthermore, querying and analyzing the data may require an experienced analyst or programmer.

### Database Highlights

MarketScan data are useful for describing trends and patterns of use of cancer-related care, measuring the financial burden of such care, and investigating clinical outcomes that are amenable to measurement in claims data.

### Database Limitations

The availability of some demographic characteristics, such as race and ethnicity and socioeconomic variables, is limited.<sup>42</sup> Furthermore, while cancer care can be identified through diagnosis and procedure codes, histologic information, stage, recurrence, and cause of death are difficult to ascertain reliably from claims data, limiting the usefulness of MarketScan to studies focused on long-term cancer outcomes.<sup>43</sup> Changes in a patient's insurance status limit subsequent follow-up. Like other sources of coded data, claims in MarketScan were impacted by the transition of the 10th revision of *ICD* in 2015. Finally, MarketScan data are drawn from a convenience sample that over-represents employees of large employers and their dependents, leading to the systematic under-representation of individuals employed in small- and medium-sized companies.<sup>39</sup>

### Healthcare Cost and Use Project Database (HCUP)

The Healthcare Cost and Utilization Project (HCUP) is the United States health service data platform that is supported by the Agency for Healthcare Research and Quality (AHRQ), one of the twelve federal agencies within the United States Department of Health and Human Service.<sup>44</sup> The AHRQ's central mission is to promote patient safety and quality improvement.

Launched more than three decades ago in 1988 and developed through federal, state, and industry partnership, HCUP is the largest all-payer encounters-based healthcare database in the USA. HCUP includes a group of several databases capturing national and state-level data. The national-level databases include the National Inpatient Sample (NIS), the Nationwide Ambulatory Surgical Sample (NASS), the Nationwide Readmission Database, the Nationwide Emergency Department Sample, and the Kid's Inpatient Database. The state-level databases include the State Inpatient Databases, the State Ambulatory Surgery and Services Databases, and State Emergency Department Databases. This article specifically focuses on NIS and NASS owing to their relevance to surgical care in gynecologic oncology.

NIP is HCUP's national-level inpatient database.<sup>45</sup> It approximates a stratified-sample of 20% of discharges in each center from all the HCUP-participating hospitals across 48 states and the District of Columbia. Every year the dataset captures more than 7 million inpatient admissions. In 2020, a total of 4580 hospitals participated in the HCUP program. When weighted for national survey estimates, NIS covered more than 97% of the US population.

NASS is HCUP's national-level outpatient database that was started in 2016.<sup>46</sup> It approximates a stratified-sample of 67% of ambulatory surgery encounters in each hospital-owned center every year. In 2020, nearly 2900 hospital-owned centers across 35 states and the District of Columbia participated in the HCUP data-capturing mechanism, collecting 7.8 million outpatient surgical encounters in that year. The research implications for gynecologic malignancies include the strength of these databases in the area of surgical care and outcomes' evaluation. These databases have been used widely in gynecologic oncology, including a number of studies examining trends, patterns of care, and outcomes.

#### Data Included

Key data elements of the NIS program are patient demographic data, hospital parameters, length of stay, total charges, and mortality data during the admission. In addition, the program captures a maximum of 40 diagnoses and 25 procedures for the index admission in each encounter. These diagnoses and procedures are recorded via WHO's ICD codes. As of the fourth quarter of 2015, the program transitioned to the 10th revision (ICD-10): clinical modification (ICD-10-CM) for diagnostic codes and procedure coding system (ICD-10-PCS) for procedure codes. The ICD-10-PCS data includes time to the index procedure. HCUP also provide their own codes for clinical categories (diagnosis-related group).

Unlike the NIS program, NASS captures the surgical procedure data via the American Medical Association's CPT codes. The diagnosis codes in NASS follow the ICD-10-CM. NASS also provides HCUP's own clinical categorical schema (clinical classifications software-services and procedures).<sup>47</sup>

#### Access, Management, and Storage

Both NIS and NASS are publicly available, de-identified databases and are available through the HCUP central distributor.<sup>48</sup> Investigators take a review course to go over the outline and content of the data platform, data handling and storage, and reporting guidelines. A dataset is available per one full year basis from January to December, and the cost of the dataset varies across the program.

#### Database Highlights

First, these databases can provide temporal national trends of patient demographics, surgical practice, and peri-operative outcomes—for example, temporal trends of hysterectomy route after the LACC trial report).<sup>49</sup> These databases are useful for assessing the characteristics in certain areas, such as (i) uncommon clinical entities (pregnancy and malignancy,<sup>50</sup> pelvic organ dysfunction surgical treatment in the setting of gynecologic malignancy),<sup>51</sup> (ii) diseases that are not captured in other tumor registries (cervical and endometrial pre-malignancies),<sup>52 53</sup> (iii) outpatient surgical practices (same-day hysterectomy, operative hysteroscopy), and (iv) rare surgical procedures (pelvic exenteration,<sup>54</sup> cesarean radical hysterectomy).<sup>55</sup> Third, these databases may be useful to examine peri-operative

morbidity and mortality. Unique information in these databases to assess peri-operative outcomes includes patient factors (obesity, medical comorbidity, and frailty) and hospital parameters (bed capacity, location, and teaching status). Failure-to-rescue (mortality risk after peri-operative morbidity)<sup>56</sup> is an important quality metric of peri-operative care, and NIS program can be used to assess this outcome. The impact of a national-level healthcare crisis can be also assessed (COVID-19 inpatient case fatality in patients with cancer).<sup>57</sup>

#### Database Limitations

Important limitations of these databases include lack of information for cancer stage, histology, and post-discharge data, including long-term oncologic outcome (recurrence and death) and adjuvant therapy. Despite these limitations, the HCUP databases are valuable research tools for investigators for whom surgical care and peri-operative management are of interest. The data available through these platforms allow for a broad understanding of national outcomes that has proved valuable for clinical practice in a number of scenarios.

#### Premier Healthcare Database (PHD)

The Premier Healthcare Database (PHD) is a large, all-payer database that captures data from hospitals across the USA. The origins of the PHD began in 1997 and the data source has evolved over time. Currently, PHD collects data annually from over 700 hospitals, with the exact number of participating hospitals varying each year. The hospitals sampled are from across the USA and include non-profit and non-governmental teaching and community hospitals in both urban and rural settings.<sup>58</sup>

#### Data Included

PHD captures both inpatient and outpatient encounters at participating hospitals and, as of 2020, contained over 231 million unique patients. PHD includes comprehensive inpatient data for all participating hospitals. Each year, more than 10 million inpatient visits, approximately 25% of inpatient admissions in the USA, are captured in the PHD. It also captures hospital-based outpatient visits, including from ambulatory surgery centers, emergency departments, and alternative sites of care. Each year, more than 93 million outpatient visits are logged in the PHD.<sup>58</sup>

Within the PHD, patients are assigned a unique identifier and can be tracked across visits within a given hospital. Each PHD encounter records clinical and demographic data, physician and hospital data, and service data from the encounter. Clinicodemographic data that are available include patient sex, race and ethnicity, payer, and age. PHD obtains ICD codes, including both ICD-9 and ICD-10 codes, for both diagnoses and procedures, and CPT codes and Healthcare Common Procedure Coding System codes. Admission and discharge diagnoses can be discerned and the disposition of the patients at discharge is available.<sup>58</sup>

Each encounter includes data elements characterizing the treating facility, including region of the country in which the hospital is located, hospital bed size, hospital teaching status, and whether the hospital is in an urban or rural location. Comparative analyses have demonstrated that the make-up of hospitals in PHD is roughly similar to survey data provided by the American Hospital Association, although there are fewer smaller hospitals in PHD. In



## Review

addition to hospital characteristics, the specialty of the admitting and attending physician are available in PHD.

The PHD captures charge coding, which allows identification of medications, tests, services, and devices that were used during a given patient encounter. These data in turn are used to determine costs associated with each encounter.<sup>58</sup> Details of chemotherapy treatment and radiation therapy are obtained if they are administered at a hospital-owned facility.

### Access, Management, and Storage

The PHD is a proprietary database that can be licensed for use.<sup>59</sup> All data are de-identified. The database is purchased from the company for use. Analysis can be performed using standard statistical software.

### Database Highlights

The PHD has been used widely in numerous studies in healthcare. The comprehensive capture of inpatient data and hospital-based outpatient care is well suited to studies examining safety and quality.<sup>60 61</sup> Similarly, the PHD has been widely used to explore real-world comparative effectiveness of treatments.<sup>62 63</sup>

The PHD also has a number of unique features that have been leveraged. Only a limited number of data sources capture prescription drug use in hospitalized patients. The pharmacologic data within PHD can be used to examine inpatient drug prescribing. Likewise, inclusion of devices and services used by patients is unique and has been leveraged to study device use where an ICD or CPT billing code may not yet exist.<sup>64 65</sup> Finally, unlike many data sources that capture charges, the PHD obtains the actual cost to a hospital to deliver a service. Some hospitals report actual cost data, whereas others report cost-to-charge ratios.<sup>66</sup>

### Database Limitations

The PHD has several limitations. Importantly, while an individual patient's encounters within a given hospital can be linked, encounters across different hospitals from the same patient cannot be linked. Second, although the PHD records hospital-based outpatient treatments, such as emergency department visits, other outpatient care, such as physician office visits and outpatient prescription drug use, are generally not captured. As such, the PHD is often best suited to studies examining episodic, hospital-based care as opposed to longitudinal care that spans inpatient and outpatient settings. Finally, like many administrative databases, the PHD lacks granular disease characteristics (eg, histopathologic type and staging) that may be available in tumor registries that are linked to disease-specific data.

## National Surgical Quality Improvement Program (NSQIP)

After the historic success of the Veterans Affairs surgical database in reducing surgical morbidity and mortality, the American College of Surgeons (ACS) developed the NSQIP Database for the private sector to replicate the success of the Veterans Affairs.<sup>67</sup> The NSQIP now operates in more than 700 hospitals across the USA in 49 out of 50 states. The exception is Michigan, where a similar program—the Blue Cross Blue Shield of Michigan—funds the Michigan Surgical Quality Collaborative.

These programs are unique compared with other datasets discussed in this article for several reasons. First, these programs deploy abstractors within the participating hospitals to abstract

patient-level data of a subset of patients. Both programs deploy sampling methodology to capture a representative subset of patients. Details of their sampling methods have been described.<sup>68</sup> Second, the abstractors (or nurses in some cases) call the patients at 30 days to determine surgical morbidity and mortality. Finally, both programs report the hospital-level de-identified information to compare participating hospitals and encourage best practices to reduce surgical morbidity and mortality. Given that Michigan Surgical Quality Collaborative data are not widely available to anyone outside of Michigan, the remainder of this section will focus on NSQIP data only.

### Access, Management, and Storage

The NSQIP dataset resides with the ACS. Participating organizations can obtain the dataset from the ACS website.<sup>69</sup> These datasets are available from 2005 to 2020 for the entire dataset. However, in 2014 a special procedure-targeted dataset was created by the ACS for hysterectomies. The application process for the NSQIP dataset is relatively simple if the requesting person uses their institutional email address, as long as their institution participates in the NSQIP. Once the application is filled and the applicant's institutional leadership approves the request, the dataset is available for download free of cost. Only participating institutions can use the data.

Management and storage of data requirements for NSQIP are not as clearly specified on the ACS website as for the Medicare data or the NCDB datasets. However, given that these datasets are compliant with the Health Insurance Portability and Accountability Act, we recommend that users should store them on a secure drive with encrypted access to prevent unauthorized access to these datasets. The ACS website does not specify any limits to the time the data can be kept with institutions. However, if the user moves out of the institution enrolled in NSQIP to one not enrolled in the program, that individual loses access to the data.

### Data Included

The primary dataset includes several variables about pre-operative workup, diagnosis codes, type of surgery performed, operative time, length of stay, and 30-day outcomes measures, such as surgical complications, readmissions, and death. Basic pre-operative comorbidity information is also available in the NSQIP dataset.

The NSQIP hysterectomy datasets contain additional information specific to hysterectomy procedures, such as the subspecialty of the surgeon, the size of the uterus, and whether or not cystoscopy was performed. More details of these site-specific variables are available on the ACS website by accessing the participant user file guides.<sup>70</sup> Unfortunately, the official linkage of NSQIP to any other dataset is unavailable. Moreover, attempts to join NSQIP and NCDB to study the impact of short-term (30-day) outcomes on long-term survival outcomes have failed in the past.<sup>71</sup> No other linkage has been widely used for this dataset.

### Database Highlights

The NSQIP Database is uniquely not administrative. In other words, the data are not collected from the insurance claims (unlike Medicare or the NIS datasets). The fact that every sampled patient is followed up for 30 days to inquire about the outcomes makes the validity of this dataset far superior to that of many other administrative datasets. However, in recent years, individual departments



have tried to validate the NSQIP data with their internal registries and found that the dataset has several inconsistencies, especially in the disease-specific participant user files.<sup>72</sup> Nevertheless, given the paucity of large registries linking surgical outcomes in the USA, NSQIP is the best available dataset for monitoring the quality of participating institutions. A large sample size (close to 900 000 cases in the 2020 file) gives the dataset adequate statistical power to study rare events and complications.

### Database Limitations

Like any sizeable national dataset, NSQIP has several limitations. The NSQIP dataset focuses solely on the 30-day outcomes of a procedure. The appropriateness of surgery is not shown in this dataset, as no variables capture the indication of surgery and the alternatives offered. Moreover, the dataset does not obtain any patient-reported outcomes to elucidate the success of surgery. For example, a patient undergoing a hysterectomy for pelvic pain might not have experienced any complications, but there might not have been any improvement of her pain as result of surgery. NSQIP, or any other large dataset available in the USA, cannot answer this question.

NSQIP variables frequently change, evolve, or get dropped completely from the dataset—researchers should carefully refer to the yearly participant user file guides before linking multiple years for trend analysis. NSQIP does not provide hospital-level data. Therefore, researchers seeking answers about the relationships between the procedure volumes and outcomes should use something other than the NSQIP dataset. Finally, one of the most significant limitations of the ACS NSQIP dataset is that no prospective study has shown that participation in NSQIP translates to better patient outcomes at those hospitals. A 2015 quasi-experimental study showed that surgical quality improvement and Medicare payments were unrelated to participation in NSQIP. In other words, hospitals not participating in NSQIP performed just as well as those participating in this program.<sup>73</sup>

## DISCUSSION

In this article, we describe common registry, insurance claims and hospital administrative databases from the USA that are used in gynecologic oncology research. The data contained in these databases are extensive and can be used to answer a broad range of questions. However, it is important that researchers and research consumers understand the limitations of these datasets and the study designs they require when interpreting and applying study results.

Many of the datasets and databases described would benefit from additional information or variables. Many do not have data on recurrence, molecular or genetic testing, intent of treatment, and important patient factors, such as body mass index. Furthermore, as patient-reported outcomes become more encouraged and prevalent in research and clinical contexts, patient-reported outcomes and quality of life metrics would be valuable to expand the usefulness of these datasets.

### Author affiliations

<sup>1</sup>Ann B. Barshinger Cancer Institute, Lancaster General Health, Lancaster, Pennsylvania, USA

<sup>2</sup>Division of Gynecologic Oncology, University of Pennsylvania Health System, Philadelphia, Pennsylvania, USA

<sup>3</sup>Leonard Davis Institute of Health Economics, Philadelphia, Pennsylvania, USA

<sup>4</sup>Penn Center for Cancer Care Innovation, Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, USA

<sup>5</sup>Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Southern California, Los Angeles, California, USA

<sup>6</sup>USC Norris Comprehensive Cancer Center, Los Angeles, California, USA

<sup>7</sup>Vincent Department of Obstetrics and Gynecology, Massachusetts General Hospital, Boston, Massachusetts, USA

<sup>8</sup>Department of Obstetrics Gynecology and Reproductive Biology, Harvard Medical School, Boston, Massachusetts, USA

<sup>9</sup>Department of Gynecologic Oncology and Reproductive Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

<sup>10</sup>Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, Michigan, USA

<sup>11</sup>Department of Obstetrics and Gynecology, Columbia University, New York, New York, USA

<sup>12</sup>Department of Obstetrics and Gynecology, Houston Methodist Hospital, Houston, Texas, USA

**X** Katherine Hicks-Courant @khickscourant and Pedro T Ramirez @pedroramirezMD

**Contributors** All authors contributed to concept development, writing, and revising the manuscript.

**Funding** The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

### ORCID iDs

Katherine Hicks-Courant <http://orcid.org/0000-0003-1395-405X>

Emily Meichun Ko <http://orcid.org/0000-0002-9438-1566>

Koji Matsuo <http://orcid.org/0000-0002-6232-8701>

Dimitrios Nasioudis <http://orcid.org/0000-0001-6260-5353>

Jason D Wright <http://orcid.org/0000-0001-6390-825X>

Pedro T Ramirez <http://orcid.org/0000-0002-6370-8052>

## REFERENCES

- Oonk MH, van Hemel BM, Hollema H, *et al.* Size of sentinel-node metastasis and chances of non-sentinel-node involvement and survival in early stage vulvar cancer: results from GROINSS-V, a multicentre observational study. *Lancet Oncol* 2010;11:646–52.
- Giuntoli RL 2nd, Metzinger DS, DiMarco CS, *et al.* Retrospective review of 208 patients with leiomyosarcoma of the uterus: prognostic indicators, surgical management, and adjuvant therapy. *Gynecol Oncol* 2003;89:460–9.
- Cantrell LA, Havrilesky L, Moore DT, *et al.* A multi-institutional cohort study of adjuvant therapy in stage I-II uterine carcinosarcoma. *Gynecol Oncol* 2012;127:22–6.
- Todo Y, Kato H, Kaneuchi M, *et al.* Survival effect of para-aortic lymphadenectomy in endometrial cancer (SEPAL study): a retrospective cohort analysis. *Lancet* 2010;375:1165–72.
- Rossi EC, Kowalski LD, Scalici J, *et al.* A comparison of sentinel lymph node biopsy to lymphadenectomy for endometrial cancer staging (FIRES trial): a multicentre, prospective, cohort study. *Lancet Oncol* 2017;18:384–92.
- Gouy S, Morice P, Narducci F, *et al.* Prospective multicenter study evaluating the survival of patients with locally advanced cervical cancer undergoing Laparoscopic para-aortic lymphadenectomy before chemoradiotherapy in the era of positron emission tomography imaging. *J Clin Oncol* 2013;31:3026–33.
- Crugan JM, Havrilesky LJ, Calingaert B, *et al.* Retrospective analysis of selective lymphadenectomy in apparent early-stage endometrial cancer. *J Clin Oncol* 2005;23:3668–75.
- Melamed A, Margul DJ, Chen L, *et al.* Survival after minimally invasive radical hysterectomy for early-stage cervical cancer. *N Engl J Med* 2018;379:1905–14.

- 9 Kapp DS, Shin JY, Chan JK. Prognostic factors and survival in 1396 patients with uterine leiomyosarcomas: emphasis on impact of lymphadenectomy and oophorectomy. *Cancer* 2008;112:820–30.
- 10 Baldwin LA, Huang B, Miller RW, et al. Ten-year relative survival for epithelial ovarian cancer. *Obstet Gynecol* 2012;120:612–8.
- 11 Chan JK, Kapp DS, Cheung MK, et al. The impact of the absolute number and ratio of positive lymph nodes on survival of endometrioid uterine cancer patients. *Br J Cancer* 2007;97:605–11.
- 12 Top cited articles | *International Journal of Gynecological Cancer*. Available: <https://ijgc.bmj.com/pages/top-cited-articles> [Accessed 14 Nov 2023].
- 13 *Gynecologic Oncology*. Most cited articles. Available: <https://www.journals.elsevier.com/gynecologic-oncology/most-cited-articles> [Accessed 7 Apr 2023].
- 14 American College of Surgeons. National cancer database. Available: <https://www.facs.org/quality-programs/cancer-programs/national-cancer-database> [Accessed 2 Jan 2023].
- 15 Nelson H, Winchester DP, McCabe RM. The journey from end results to the national cancer database. Chicago, 2022. Available: <https://bulletin.facs.org/2022/01/the-journey-from-end-results-to-the-national-cancer-database> [Accessed 1 Feb 2023].
- 16 American College of Surgeons. National cancer database participant user file 2020 data dictionary. Chicago, 2020.
- 17 Boffa DJ, Rosen JE, Mallin K, et al. Using the National cancer database for outcomes research: a review. *JAMA Oncol* 2017;3:1722–8.
- 18 Jacobs CD, Carpenter DJ, Hong JC, et al. Radiation records in the national cancer database: variations in coding and/or practice can significantly alter survival results. *JCO Clin Cancer Inform* 2019;3:1–9.
- 19 Yang DX, Khara R, Miccio JA, et al. Prevalence of missing data in the national cancer database and association with overall survival. *JAMA Netw Open* 2021;4:e211793.
- 20 Melamed A, Rauh-Hain JAA, Schorge J. Clinical outcomes research in gynecologic oncology. *Gynecol Oncol* 2017;146:653–60.
- 21 Melamed A, Wright JD, Rauh-Hain JA. Minimally invasive or abdominal radical hysterectomy for cervical cancer. *N Engl J Med* 2019;380:793–5.
- 22 Daly MC, Paquette IM. Epidemiology, and end results (SEER) and SEER-medicare databases: use in clinical research for improving colorectal cancer outcomes. *Clin Colon Rectal Surg* 2019;32:61–8.
- 23 Duggan MA, Anderson WF, Altekruse S, et al. The surveillance, epidemiology, and end results (SEER) program and pathology: toward strengthening the critical relationship. *Am J Surg Pathol* 2016;40:e94–102.
- 24 Noone A-M, Lund JL, Mariotto A, et al. Comparison of SEER treatment data with medicare claims. *Med Care* 2016;54:e55–64.
- 25 Howlader N, Noone A, Krapcho M, et al. SEER cancer statistics review, 1975–2016. Bethesda, MD, 2019. Available: [https://seer.cancer.gov/csr/1975\\_2016/](https://seer.cancer.gov/csr/1975_2016/)
- 26 Sharma A, Jung MK, Polce SA, et al. Surveillance epidemiology and end results (SEER) reported survival outcomes have a high accuracy and positive predictive value when tested using randomized controlled trials (RCTs) as gold standard. *Int J Radiat Oncol Biol Phys* 2019;105:E467–8.
- 27 Ward E, Jemal A, Cokkinides V, et al. Cancer disparities by race/ethnicity and socioeconomic status. *CA Cancer J Clin* 2004;54:78–93.
- 28 Rauh-Hain JA, Clemmer JT, Bradford LS, et al. Racial disparities in cervical cancer survival over time. *Cancer* 2013;119:3644–52.
- 29 Wang H, Chen X, Zhao J, et al. Predictive nomogram for midterm to long-term prognosis in patients with papillary renal cell carcinoma based on data from the surveillance, epidemiology, and end results (SEER) program. *Med Sci Monit* 2020;26:e921859.
- 30 Yu JB, Gross CP, Wilson LD, et al. NCI SEER public-use data: applications and limitations in oncology research. *Oncology (Williston Park)* 2009;23:288–95.
- 31 National Cancer Institute. Required documents & instructions for submitting applications. Available: <https://healthcaredelivery.cancer.gov/seermedicare/obtain/requests.html> [Accessed 11 Feb 2023].
- 32 Ko EM, Brensinger CM, Cory L, et al. Utilization and survival outcomes of sequential, concurrent and sandwich therapies for advanced stage endometrial cancers by histology. *Gynecol Oncol* 2020;159:394–401.
- 33 Cory L, Brensinger C, Burger RA, et al. Patterns of adjuvant treatment and survival outcomes in stage I uterine carcinosarcoma. *Gynecol Oncol Rep* 2022;39:100930.
- 34 Latif N, Oh J, Brensinger C, et al. Lymphadenectomy is associated with an increased risk of postoperative venous thromboembolism in early stage endometrial cancer. *Gynecol Oncol* 2021;161:130–4.
- 35 Ko EM, Bekelman JE, Hicks-Courant K, et al. Association of gynecologic oncology versus medical oncology specialty with survival, utilization, and spending for treatment of gynecologic cancers. *Gynecol Oncol* 2022;164:295–303.
- 36 Ochieng N, Fuglesten Biniek J, Freed M, et al. Medicare advantage in 2023: enrollment update and key trends | KFF. Available: <https://www.kff.org/medicare/issue-brief/medicare-advantage-in-2023-enrollment-update-and-key-trends> [Accessed 20 Feb 2024].
- 37 Measures that are limited or not available in the data. Available: <https://healthcaredelivery.cancer.gov/seermedicare/considerations/measures.html> [Accessed 11 Feb 2023].
- 38 CMS cell size suppression policy | ResDAC. Available: <https://resdac.org/articles/cms-cell-size-suppression-policy> [Accessed 11 Feb 2023].
- 39 Merative. Healthcare data, technology and analytics. Available: <https://www.merative.com> [Accessed 15 Aug 2023].
- 40 Kulaylat AS, Schaefer EW, Messaris E, et al. Truven health analytics MarketScan databases for clinical research in colon and rectal surgery. *Clin Colon Rectal Surg* 2019;32:54–60.
- 41 Merative. Contact us. Available: <https://www.merative.com/contact> [Accessed 14 Feb 2023].
- 42 Center for Disease Control. MarketScan. Available: <https://www.cdc.gov/visionhealth/vehss/data/claims/marketscan.html> [Accessed 15 Jan 2023].
- 43 Cooper GS, Yuan Z, Stange KC, et al. The utility of medicare claims data for measuring cancer stage. *Med Care* 1999;37:706–11.
- 44 Agency for Healthcare Research and Quality. HCUP overview. Healthcare cost and utilization project (HCUP). 2022. Available: [www.hcup-us.ahrq.gov/overview.jsp](http://www.hcup-us.ahrq.gov/overview.jsp) [Accessed 1 Feb 2023].
- 45 Agency for Healthcare Research and Quality. HCUP-US databases. Healthcare cost and utilization project (HCUP). 2022. Available: <https://www.hcup-us.ahrq.gov/nisoverview.jsp> [Accessed 1 Feb 2023].
- 46 Agency for Healthcare Research and Quality. NASS overview. Healthcare cost and utilization project (HCUP). 2021. Available: <https://www.hcup-us.ahrq.gov/nassoverview.jsp> [Accessed 1 Feb 2023].
- 47 Agency for Healthcare Research and Quality. HCUP CCS-services and procedures. Healthcare cost and utilization project (HCUP). 2021. Available: [https://www.hcup-us.ahrq.gov/toolssoftware/ccs\\_svcsproc/ccsvproc.jsp](https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccsvproc.jsp) [Accessed 1 Feb 2023].
- 48 Agency for Healthcare Research and Quality. Purchase HCUP data. Healthcare cost and utilization project (HCUP). 2022. Available: [www.hcup-us.ahrq.gov/tech\\_assist/centdist.jsp](http://www.hcup-us.ahrq.gov/tech_assist/centdist.jsp) [Accessed 1 Feb 2023].
- 49 Matsuo K, Mandelbaum RS, Klar M, et al. Decreasing utilization of minimally invasive hysterectomy for cervical cancer in the United States. *Gynecol Oncol* 2021;162:43–9.
- 50 Matsuo K, Klar M, Youssefzadeh AC, et al. Assessment of severe maternal morbidity and mortality in pregnancies complicated by cancer in the US. *JAMA Oncol* 2022;8:1213–6.
- 51 Matsuo K, Rau AR, Ciesielski KM, et al. Concurrent minimally invasive gynecologic procedures at the time of laparoscopic cholecystectomy. *Obstet Gynecol* 2023;142:1491–5.
- 52 Matsuo K, Violette CJ, Mandelbaum RS, et al. Substantial variability in ovarian conservation at hysterectomy for endometrial hyperplasia. *Am J Obstet Gynecol* 2022;227:255.
- 53 Matsuo K, Tavakoli A, Donovan KM, et al. Utilization of lymph node evaluation at hysterectomy for cervical carcinoma in situ. *J Surg Oncol* 2022;126:1543–50.
- 54 Matsuo K, Mandelbaum RS, Adams CL, et al. Performance and outcome of pelvic exenteration for gynecologic malignancies: a population-based study. *Gynecol Oncol* 2019;153:368–75.
- 55 Matsuo K, Mandelbaum RS, Matsuzaki S, et al. Cesarean radical hysterectomy for cervical cancer in the United States: a national study of surgical outcomes. *Am J Obstet Gynecol* 2020;222:507–11.
- 56 Lee MW, Vallejo A, Mandelbaum RS, et al. Temporal trends of failure-to-rescue following perioperative complications in vulvar cancer surgery in the United States. *Gynecol Oncol* 2023;177:1–8.
- 57 Matsuo K, Mandelbaum RS, Vallejo A, et al. Assessment of gender-specific COVID-19 case fatality risk per malignant neoplasm type. *JAMA Oncol* 2023;9:1113–8.
- 58 Premier Healthcare Database. 2020. Available: <https://www.aha.org/system/files/2018-05/2018-chartbook> [Accessed 1 Feb 2023].
- 59 Applied Research & Analytics | Applied Sciences | PINC AI. Available: <https://www.pinc-ai.com/applied-sciences/solutions/applied-research-and-analytics> [Accessed 14 Feb 2023].
- 60 Rosenthal N, Cao Z, Gundrum J, et al. Risk factors associated with in-hospital mortality in a US national sample of patients with COVID-19. *JAMA Netw Open* 2020;3:e2029058.

- 61 Lindenauer PK, Pekow P, Wang K, *et al.* Lipid-lowering therapy and in-hospital mortality following major noncardiac surgery. *JAMA* 2004;291:2092–9.
- 62 Wright JD, Burke WM, Tergas AI, *et al.* Comparative effectiveness of minimally invasive hysterectomy for endometrial cancer. *J Clin Oncol* 2016;34:1087–96.
- 63 Jeong IG, Khandwala YS, Kim JH, *et al.* Association of robotic-assisted vs laparoscopic radical nephrectomy with perioperative outcomes and health care costs. *JAMA* 2017;318:1561.
- 64 Wright JD, Tergas AI, Burke WM, *et al.* Uterine pathology in women undergoing minimally invasive hysterectomy using morcellation. *JAMA* 2014;312:1253–5.
- 65 Wright JD, Ananth CV, Lewin SN, *et al.* Robotically assisted vs laparoscopic hysterectomy among women with benign gynecologic disease. *JAMA* 2013;309:689–98.
- 66 Wright JD, Ananth CV, Tergas AI, *et al.* An economic analysis of robotically assisted hysterectomy. *Obstet Gynecol* 2014;123:1038–48.
- 67 About ACS NSQIP | ACS. Available: <https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/about-acs-nsqip> [Accessed 1 Feb 2023].
- 68 Eisenstein S, Stringfield S, Holubar SD. Using the national surgical quality improvement project (NSQIP) to perform clinical research in colon and rectal surgery. *Clin Colon Rectal Surg* 2019;32:41–53.
- 69 Participant Use Request Form | ACS, Available: <https://www.facs.org/quality-programs/data-and-registries/acs-nsqip/participant-use-data-file/participant-use-request-form> [Accessed 1 Feb 2023].
- 70 American College of Surgeons, National Surgical Quality Improvement Program. User guide for the 2020 ACS NSQIP procedure torted participant use data file (PUF). 2021.
- 71 Matsen CB, Luther SL, Stewart AK, *et al.* A match made in heaven? Trying to combine ACS-NSQIP and NCDB databases. *J Surg Res* 2012;175:6–11.
- 72 Anderson KT, Bartz-Kurycki MA, Austin MT, *et al.* Room for “quality” improvement? Validating national surgical quality improvement program-pediatric (NSQIP-P) appendectomy data. *J Pediatr Surg* 2019;54:97–102.
- 73 Osborne NH, Nicholas LH, Ryan AM, *et al.* Association of hospital participation in a quality reporting program with surgical outcomes and expenditures for medicare beneficiaries. *JAMA* 2015;313:496.