



Published in final edited form as:

JAMA Surg. 2025 March 01; 160(3): 344–346. doi:10.1001/jamasurg.2024.6006.

Practical Guide to Clinical Big Data Sources

Oluwadamilola M. Fayanju, MD, MA, MPHS^{1,2,3,4}, Elliott R. Haut, MD, PhD^{5,6,7,8}, Kamal Itani, MD^{9,10,11}

¹Division of Breast Surgery, Department of Surgery, Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA

²Rena Rowan Breast Center, Abramson Cancer Center, Penn Medicine, Philadelphia, PA

³Penn Center for Cancer Care Innovation, Abramson Cancer Center, Philadelphia, PA

⁴Leonard Davis Institute of Health Economics (LDI), The University of Pennsylvania, Philadelphia, PA

⁵Department of Surgery, Division of Acute Care Surgery, Johns Hopkins University School of Medicine, Baltimore, MD

⁶Department of Emergency Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

⁷Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

⁸Department of Health Policy and Management, The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

⁹Department of Surgery, Veterans Affairs Boston Health Care System, Boston, MA

¹⁰Department of Surgery, Boston Medical Center, Boston University, Boston, MA

¹¹Brigham and Women's Hospital, Harvard Medical School, Boston, MA

Introduction

Clinical data sources are composed of patient-level data that are generated from interaction with and assessment of patients and are typically abstracted from institutional electronic health records (EHRs). They can be collated at the institution-level by health systems, by government agencies (e.g., Surveillance, Epidemiology, and End Results [SEER] by the National Cancer Institute [NCI]), by professional societies through participation that may or may not be tied to accreditation or data access (e.g., the American College of Surgeons [ACS] and the Society of Thoracic Surgeons [STS]), and by vendors that source data from multiple institutions sharing a common EHR platform (e.g., Epic Cosmos).¹ In addition, EHR-based networks such as All of Us by the National Institute of Health (NIH) and PCORNet by the Patient-Center Outcomes Research Institute (PCORI) work across different

Corresponding author: Dr. Oluwadamilola "Lola" Fayanju, 3400 Spruce St, Silverstein 4, Philadelphia, PA 19104, fayanju@upenn.edu, Fax: 215-662-7476.

EHR platforms using a Common Data Model for reporting.² All of these sources are distinct from claims-based data sets, which are generated from billing and insurance-related transactions rather than from interactions with people, and are addressed in a different guide in this series. Depending on the variables available, clinical datasets can be used to conduct a wide variety of health services research (HSR) including observational cohort studies, comparative effectiveness research (CER), and economic evaluations. In the case of clinical data sources for which prospective data collection can be centrally coordinated, they can also be used to conduct pragmatic clinical trials (e.g., ACS National Surgical Quality Improvement Program [NSQIP],³ Abdominal Core Health Quality Collaborative [ACHQC]).⁴ In this manuscript, we describe the strengths and weaknesses of clinical sources of “big data.” Specifically, we focus on datasets not previously included in the 2018 *JAMA Surgery* Practical Guide to Surgical Data Sets Series.³ A more expansive list is available on the website of the [Surgical Outcomes Club](#).

Limitations

Clinical datasets vary significantly based on included populations, how and from where content is collected, and data elements they contain. These factors may also limit generalizability and contribute to selection bias and should be considered carefully when deciding which dataset to use for a given research question. For example, data in the STS General Thoracic Surgery Database is provided by only 40–50% of eligible centers, thus clinical outcomes may be less generalizable than those conducted through the STS Congenital Heart Database, which is sourced from over 95% of the centers performing these operations.⁵ The Department of Veterans Affairs (VA) Corporate Data Warehouse (CDW) and the Million Veterans Program (MVP), although robust, may have limited generalizability given the preponderance of older men among the US veteran population.⁶ If datasets are procedure-based (e.g., Carotid Endarterectomy Registry from the Society of Vascular Surgery [SVS] Vascular Quality Initiative [VQI]) rather than disease-based, CER comparing surgical and non-surgical management is not possible. Finally, many datasets lack patient-reported outcomes (PROs) and long-term data on survival or recurrence, all of which should be considered alongside short-term outcomes (e.g., 30-day readmission, perioperative complications) when assessing a surgical procedure.

Data Management Considerations

The quality of data collection and collation varies across datasets. Data in the VA Surgical Quality Improvement Program (VASQIP) and NSQIP (modeled after VASQIP) are collected and iteratively updated by trained clinical reviewers.² The National Cancer Database (NCDB) is sourced from tumor registries of institutions that are accredited by the Commission on Cancer (CoC) and must submit data to maintain CoC accreditation and enable NCDB access for investigators.³ Although tumor registrars are expected to reference the Facility Oncology Registry Data Standards (FORDS) manual while coding and entering registry data, no central review of NCDB data is performed, so coding errors occurring at the institutional level may be perpetuated in the NCDB.

Most of the data in clinical sources is primary, i.e., comes from patients themselves via self-report (e.g., race, gender identity, PROs, survey responses), objective measurements (e.g., labs, radiology exams), and/or direct observation by clinicians (e.g., palpable lymphadenopathy). However, a small proportion of data in clinical sources may be secondary, i.e., collected by a third party, and sometimes linked to primary data (e.g., administrative Medicare claims data [secondary] linked to SEER data [primary], vital status information from the National Death Index and State Vital Statistic Registries [secondary] linked to data in the VA CDW [primary]), with varying levels of transparency as to what is primary and what is secondary data. Because all research using clinical datasets involves information collected from human participants, investigators should always seek approval from their local institutional review board. If only deidentifiable data is used, exempt status is typically granted. Some clinical datasets, including ones that include deidentified but potentially identifiable data, require data use agreements (DUAs) either with participating institutions and/or the organization overseeing the dataset or with individual investigators. Precluding the publication of table cells below a certain number of participants (e.g., 10 for ACHQC and NCDB) is another measure intended to limit the likelihood of participant re-identification.

Analytic considerations

Some clinical data sources require that all analyses be conducted by credentialed analysts and biostatisticians within the administering organization or may require that all analyses be done within a specified trusted research environment. Others, such as SEER, require specific software to view the data. As data in the VA system is spread across over 60 databases within the CDW, researchers typically need to work with VA-affiliated analysts with programming expertise to cross-match the various databases at the level of patients and timeframes.⁶ Finally, large sample sizes may potentially lead to overestimation of statistical significance and data precision even when there is little or no clinically meaningful significance.

Health equity considerations

Care must be paid to both the population in a clinical dataset as well as the origin of collected data to anticipate and avoid selection bias. Regression and other techniques can address many but not all sources of confounding. Careful dataset selection based on the question and population of interest is the best way to minimize selection bias and to ensure the validity, generalizability, and real-world applicability of the research findings. For example, hospital-based databases such as the NCDB are not appropriate for describing population-wide, epidemiological trends. In addition for some research questions, particularly for rare conditions or in marginalized groups, mixed-methods analyses incorporating qualitative approaches may be more appropriate than running regression analyses on large but non-representative samples. With greater focus on redressing disparities and promoting equity in surgical care, clinical data are increasingly used to identify ways in which traditionally marginalized groups experience disparities in disease prevalence, access to diagnosis/treatment, and outcomes. It is important, however, that investigations regarding disparities be pursued with intentionality and not as an

analytical afterthought. For such work, as for all high-quality HSR, a statistical analysis plan should be drafted prior to study initiation that makes clear the hypotheses to be tested, the variables to be included, and the a priori subset and interaction analyses that are planned to avoid “fishing expeditions.” Furthermore, investigators should be aware of how nonrandom missingness may disproportionately affect minoritized groups, potentially amplifying bias and distorting results.

Funding

Dr. Fayanju is supported by the National Institutes of Health (NIH) under Award Numbers 7K08CA241390-03 (PI: Fayanju), P50CA244690 (PI: Bekelman), and 2P30CA016520-45 (PI: Vonderheide); the Centers for Disease Control and Prevention (CDC) under Award Number U48DP006801 (PI: Fayanju); the Breast Cancer Research Foundation, and philanthropic funds from the Haas family. She also reports research support unrelated to this work from Gilead Sciences, Inc. Dr. Haut reports grant funding from NIH/NHLBI, AHRQ, and PCORI. The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and other funding agencies.

References

1. Tarabichi Y, Frees A, Honeywell S, et al. The Cosmos collaborative: a vendor-facilitated electronic health record data aggregation platform. *ACI open*. 2021;5(01):e36–e46. [PubMed: 35071993]
2. Murtha JA, Jawara D, Funk LM. Optimization of the Electronic Health Record for Research. *Ann Surg Open*. Jun 2023;4(2):e297. doi:10.1097/as9.0000000000000297 [PubMed: 37601479]
3. Practical Guide to Surgical Data Sets. *JAMA Surgery*. 2018;153(6)doi:10.1001/jamasurg.2018.0628
4. Poulouse BK, Roll S, Murphy JW, et al. Design and implementation of the Americas Hernia Society Quality Collaborative (AHSQC): improving value in hernia care. *Hernia*. Apr 2016;20(2):177–89. doi:10.1007/s10029-016-1477-7 [PubMed: 26936373]
5. Towe CW, Servais EL, Brown LM, et al. The Society of Thoracic Surgeons General Thoracic Surgery Database: 2023 Update on Outcomes and Research. *Ann Thorac Surg*. Mar 2024;117(3):489–496. doi:10.1016/j.athoracsurg.2023.11.021 [PubMed: 38043852]
6. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. Feb 2016;70:214–23. doi:10.1016/j.jclinepi.2015.09.016 [PubMed: 26441289]
7. Malamon JS, Kaplan B. Validation of the Integrity of the OPTN/UNOS Transplantation Registry Data. *Transplantation*. Dec 1 2023;107(12):e324–e325. doi:10.1097/tp.0000000000004793 [PubMed: 37726887]

Box Summary**Clinical “Big Data” Sources**

- Clinical data sources are large datasets composed mainly of primary data collated by health systems, government agencies, professional societies, and third-party vendors.
- These datasets can be used to conduct a wide variety of health services research including observational cohort studies, comparative effectiveness research, and economic comparisons as well as pragmatic clinical trials.
- Investigators should carefully consider data content, origin, and structure as well participant characteristics to ensure they are using the most appropriate dataset for their research question, to optimize the generalizability of their findings, and to minimize selection bias.
- Some clinical data sources require dedicated software, significant and highly specific programming expertise, and/or that all analyses be conducted within a trusted research environment.
- Health equity research using clinical big data should be pursued with intentionality and a priori attention to nonrandom missingness and other factors that disproportionately affect minoritized individuals, leading to distorted findings and perpetuated bias.

Table 1.

Examples of Clinical “Big Data” Sources*

Data Source / Database	Clinical Disciplines	Organization(s)	Patient Population	Access	Analytic Variables**
All of US	All (EHR-based Research Network)	National Institutes of Health (NIH)	33,000 (goal 1 million) US volunteers	3 tiers of data access to the Curated Data Repository (CDR): (a) <i>Public</i> (no login required, aggregate participant data only), (b) <i>Registered</i> (login required, deidentified participant-level data), and (c) <i>Controlled</i> (additional approval required, most granular and potentially re-identifiable participant-level data).	All except cost. Also includes gender identity, disability
Epic Cosmos	All (EHR-based Research Network)	Epic	246 million patients from contributing health systems with Epic-based EHRs in US + American University of Beirut Medical Center	Free to individuals from paying, contributing organizations	All except cost; SES – (Insurance, SDOH), Longitudinal (>15 million patients with ≥3 years of follow-up, >1 million with ≥10 years)
PCORnet	All (EHR-based Research Network)	Patient-Centered Outcomes Research Institute (PCORI)	140 million US patients from contributing US health systems	Free to individuals from paying, contributing organizations	All except Cost; SES (Insurance)
VA Million Veterans Program (MVP)	All (EHR-based Research Network)	Veterans' Administration	US veteran volunteers who receive care through the Veterans Health Administration	Free to VA-employed or Without Compensation (WOC) investigators	All; SES (Insurance, specifically, Eligibility for Medicaid)
Abdominal Core Health Quality Collaborative (ACHQC)	Hernias, Abdominal Wall Surgery	Abdominal Core Health Quality Collaborative (ACHQC)	Patients of ACHQC members (> 450 hernia surgeons in US)	Free to ACHQC member surgeon in good standing (25 cases with 30d follow-up within preceding 12 months)	All except Labs, Pharmacy, Cost Targeted hernia related data
Standard Transplant Analysis and Research (STAR) database ⁷	Transplant Surgery	United Network for Organ Sharing (UNOS) Organ Procurement and Transplantation Network (OPTN)	Transplant recipients, deceased and living donors, and waiting list candidates dating back to October 1987	Public but fee required	All except Cost; SES (area-level MH income), Longitudinal (duration of graft survival)

* Representing a selection of datasets in the United States and with a focus on sources not previously included in the 2018 JAMA Surgery Practical Guide to Surgical Data Sets Series. For a more expansive compendium of clinical big data sources, please visit the Surgical Outcomes Club website.

** Analytic Variables: Operative, Labs, Pharmacy, Cost, Age, Race, Ethnicity, Sex, SES, Longitudinal, Comorbidities, Vital Status