Published in final edited form as:

Cancer. 2019 September 01; 125(17): 2926–2934. doi:10.1002/cncr.32118.

The Opportunities and Shortcomings of Using Big Data and National Databases for Sarcoma Research

DR. Heather Lyu[1], **Adil Haider**[1], **Adam Landman**[2], **Chandrajit Raut**[1,3]

[1]Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115

[2]Department of Emergency Medicine, Brigham and Women's Hospital, Harvard Medical School Boston, MA, 02115

[3]Center for Sarcoma and Bone Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115

Abstract

The rarity and heterogeneity of sarcomas makes performing appropriately powered studies challenging and magnifies the significance of large databases in sarcoma research. Established large tumor registries and population-based databases have become increasingly more relevant to answer clinical questions regarding sarcoma incidence, treatment patterns, and outcomes. However, the validity of large databases has been questioned and scrutinized due to inaccuracy and wide variability of coding practices and absence of clinically relevant variables. Additionally, the utilization of large databases for the study of rare cancers like sarcoma may be particularly challenging secondary to known limitations of administrative data and poor overall data quality. Currently there are several large national cancer databases including the Surveillance, Epidemiology, and End Results (SEER) database, the American College of Surgeons' and American Cancer Society's National Cancer Database (NCDB), and the Center for Disease Control (CDC) National Program of Cancer Registries (NPCR). These are often used for sarcoma research but these databases are limited by a dependence on administrative or billing data, the lack of agreement between chart abstractors on diagnosis codes, and the use of preexisting documented hospital diagnosis codes for tumor registries leading to significant underestimation of sarcomas in large datasets. Current and future initiatives to improve databases and big data applications for sarcoma research include increasing the utilization of sarcoma-specific registries and encouraging national initiatives to expand on real-world evidence based datasets.

Precis:

The main aim of this article is to demonstrate the limitations of these databases specifically for sarcoma research. We also describe current initiatives formed to improve the application of big data for rare malignancies.

Corresponding Author: Heather Lyu, MD, Brigham and Women's Hospital, Department of Surgery, 75 Francis St, Boston, MA 02115, Phone: 703-965-9392, hlyu@bwh.harvard.edu.

Conflict of interest statement: All authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Keywords

Sarcoma; database; big data; NCDB; SEER

Introduction

In 2016, the United States presidential administration formed the Cancer MoonShot initiative to accelerate national efforts to prevent, diagnose, and treat cancer.[1] One of the primary recommendations of this initiative discussed the formation of a more collaborative approach to harness big data and the establishment of a high value national cancer data ecosystem integrating multiple if not all disciplines.[1] For instance, surgery-specific variables in addition to other therapeutic data should be combined with basic science and genomic information in order to provide a comprehensive picture of oncologic care.[1] Big data was defined in 2001 by Doug Laney by the three V's: volume: a large amount of data per transaction; velocity: fast pace of data delivered per transaction; and variety: heterogeneity of data types.[2] This emphasis on big data elevates the significance of high quality, population-based, clinical databases in cancer research. Currently, population-based cancer registries in the United States pool a vast amount of data from a heterogeneous group of institutions resulting in decreased selection bias and increased generalizability.[3] Data amassed in these registries can be utilized to fuel quality improvement initiatives, assess and evaluate the cost and effectiveness of oncologic treatment modalities, structure risk models for cancer patients, and reveal variations across multiple levels of care.[3] Moreover, the advantage of using large databases for studying rare outcomes and pathologies such as sarcomas has been cited in the literature.[4–7]

The rarity and heterogeneity of sarcomas makes performing appropriately powered studies challenging and magnifies the significance of large databases in sarcoma research. Single institution registries are unlikely to include enough patient data to make statistically significant inferences.[6] Yet, data acquisition from multiple institutions presents many challenges due to limited interoperability and variations in diagnostic methodologies.[7] Subsequently, established large tumor registries and population-based databases have become increasingly more relevant to answer clinical questions regarding sarcoma incidence, treatment patterns, and outcomes.[6–10] Nevertheless, the validity of large databases across multiple specialties has been questioned and scrutinized due to inaccuracy and wide variability of coding practices and absence of clinically relevant variables. Additionally, the utilization of large databases for the study of rare cancers like sarcoma may be particularly challenging secondary to known limitations of administrative data and poor overall data quality.[3,7]

The aim of this review is two-fold: 1) to present the current state of cancer databases often used for sarcoma research and to expand on the limitations of the databases and their potential consequences; and 2) to introduce current and future initiatives for improving sarcoma data collection and validity of large oncologic registries.

Current State of National Oncologic Databases

Currently there are several large national cancer databases that are most often used for sarcoma research including the Surveillance, Epidemiology, and End Results (SEER) database, the American College of Surgeons' and American Cancer Society's National Cancer Database (NCDB), and the Center for Disease Control (CDC) National Program of Cancer Registries (NPCR) (Table 1). These databases are very relevant in sarcoma research. For instance, there are 302 sarcoma publications accessible in Pubmed using the search terms, "Sarcoma" and "SEER". There are several advantages to large database studies for sarcoma. These large registries are relatively accessible and inexpensive to use. Moreover, multiple registries can be pooled together to study various aspects of rare pathologies and outcomes.[4,13] The advantages of utilizing large registries for rare cancers such as sarcoma or mesothelioma have been cited in several studies.[4,14] In general, large population-based databases provide high statistical power and precise effect estimates over smaller single institution or multi-institutional registries.[15] Moreover, since data are manually collected by trained abstractors, most population-based datasets undergo extensive quality control and audits to ensure appropriate case selection.[16] For instance, the SEER database undergoes reliability studies to test the skills of registry personnel and assess the consistency of coding data across registries.[16] Reliability testing of SEER abstractors consists of web-based studies that require participants to code information from a uniform set of medical records under standardized testing conditions. Strict monitoring and review of databases can ensure accuracy of the data. For example, the hip and knee replacement codes in the Veterans Administration database were shown to be accurate with an excellent positive predictive value of 98%.[17] Furthermore, tumor registry staff often undergo the same training and education on American Joint Commission on Cancer (AJCC) guidelines during the annual meeting of the national cancer registrar's association.[16] While the advantages of large databases for the study of rare pathologies such as dermatofibrosarcoma protuberans (DFSP), scrotal malignancies, and sarcomatoid carcinoma of the breast has been cited, there have been limited studies on the veracity of these databases for sarcoma research and the potential consequences of inappropriate data collection and utilization are unclear.[6]

Limitations of Current Large Databases

Previous studies have questioned the use of large databases due to their dependence on administrative or billing data. There are many warnings cited in the literature against using administrative databases such as the National Inpatient Sample (NIS) or Medicare registries for clinical research as many variables are not intended for research purposes and key clinical elements are often missing.[13,15,18, 19] For example, a study in 2014 on ICD-9 codes for patients undergoing lumbar fusions showed that 48% of cases did not have a primary diagnosis code to reflect the primary indication for surgery.[20] The reliance on coding of medical diagnoses and procedures in all large databases can create significant variability and undermine the accuracy of information.[3, 21, 22] A review of non-Hodgkin's lymphoma cases in one study using the SEER database showed that the agreement on the classification of specific histologies amongst experts vary from 5 to 100%.[5] The agreement between chart abstractors on diagnosis codes is difficult to achieve.[15] In Great Britain and Ireland, a regional colorectal cancer dataset was shown to have a large proportion of missing data as well as a lack of agreement between medical charts and data points.[23] Another independent review of large cell lung cancer cases in the Iowa Cancer registry, which feeds in to the large SEER database, noted low sensitivity and positive predictive value. [5] Other data elements are not exempt from similar challenges. Treatments and procedures are also vulnerable to coding issues as another study showed that 21% of patients who reported receiving radiotherapy were coded as not receiving radiotherapy in the SEER database. [5] Analysis of sarcoma data from these large databases may lead to inaccurate predictions and outcomes. In a study using data from both the NCDB and a multi-institutional database of a consortium of seven institutions (Emory, Stanford, Wake Forest, Medical College of Wisconsin, University of Chicago, Ohio State, and Washington University in St. Louis), propensity score matched cohorts were used to analyze the impact of radiation therapy on oncologic survival. Consortium data showed that patients who underwent radiation therapy did not have improved outcomes while a confirmatory NCDB analysis showed that patients who received radiation had improved survival.[7] It is uncertain which database is more accurate but this discrepancy brings into question the validity of large registry studies. The benefit of a multi-institutional database is that the data may include more granular, clinically relevant data that are often replaced by billing or coded data in larger, population-based databases. Despite these findings, one may still argue for large national databases given the auditing and validation processes put in place by the institutions that can significantly reduce coding errors and variability.

Reliance on Coding

Yet, strict auditing and validation of databases only ensures the accuracy of the data already collected; there is no way of validating the initial coding process itself as many of the larger databases use preexisting documented diagnosis codes from institutions. A recent study by the authors of this review showed that only 62% of all sarcoma procedures performed at a single institution were coded accurately by ICD-9/10 codes representing a significant under-reporting of the true sarcoma operative volume.[24] Furthermore, only 60% of patients were logged accurately by ICD-O-3 codes in the institutional tumor registry demonstrating that the primary diagnoses entered by surgical coders and tumor registrars do not accurately describe the specific disease process that serves as the indication for resection. The under-reporting of sarcoma volume was largely due to coding of malignancies by their organ site of origin rather than the actual malignancy histology (e.g., gastric gastrointestinal stromal tumors were coded as “gastric cancer,” and breast angiosarcomas were coded as “breast cancer”). Radiation-associated breast angiosarcomas are commonly miscoded as breast cancer despite it being a completely unique pathological category. The identification of errors and variability of coding practices may have larger implications on the validity and accuracy of larger oncology databases and registries. While this study reports an under-representation of sarcomas due to coding issues at one institution (a high-volume sarcoma center), coding inaccuracies may be a more widespread problem as all tumor registrars use the same American Joint Committee on Cancer coding guidelines. The consequences of these findings, if validated by other institutions, suggest that sarcoma case volumes may be significantly higher than those reported by studies using these datasets. For example, a 2017 NCDB study by Corey et al. reports 63,714 soft tissue sarcomas over 12 years, or approximately 5,310 sarcomas per year.[12] These only included the top 34 of 239 soft tissue sarcomas reported to the NCDB as commonly encountered by orthopedic oncologists. If up to 40% of cases were not reported accurately, then there could be up to 3,540 more cases per year and 106,200 reported cases over the same time period. A 2011 SEER study by Ferrari et al. covering 1973 to 2006 reported 72,972 cases, for an incidence of 2,211 cases annually. With the same extrapolation, 1,474 more cases would be reported annually, and 121,620 cases would be reported over the same 33 years.[8] The data represented in these datasets are not necessarily inaccurate, but potentially not comprehensive about the total volume of disease prevalence. These numbers are significantly lower than the American Cancer Society estimate of 12,298 new diagnoses of soft tissue sarcomas annually, however, this includes rhabdomyosarcoma, which is typically studied separately from non-rhabdomyosarcoma soft tissue sarcomas (NSRTS). [25] The American Cancer Society data are derived from the North American Association of Central Cancer Registries (NAACCR) which consolidates data from every central and state cancer registry reporting.[25] Since this is an aggregate of all registry data, it may be more accurate, but underestimation of the true national sarcoma burden is still highly likely and warrants further study. Oncologic guidelines and policy recommendations based on national databases may need to be further examined. If coding inaccuracies for other sarcoma centers are similar, there could be significant impact on the utility of NCDB and SEER data. At present, the net effect of coding errors on analysis from these large databases is unknown.[20, 26]

The process behind sarcoma coding deserves further scrutiny as coding errors and inaccuracies in large population databases may be amplified for sarcomas and other rare pathologies. First, the heterogeneity and large number of histologic subtypes of sarcomas (with over 70 described subtypes) render accurate characterization of these cases challenging.[27,28] Second, the nomenclature of sarcomas can further confound accurate classification. For instance, GISTs were most commonly mischaracterized as GI malignancy rather than sarcoma by both the ICD9/10 and ICD-O-3 codes.[24] This may be due in part to the name (“gastrointestinal”). Furthermore, for tumor registrars, GISTs cannot be classified as a sarcoma unless the pathology report specifically labels the tumor as “malignant GIST” or “gastrointestinal stromal sarcoma.” Third, most malignancies are organ-specific, and thus they are classified based on organ of origin. Sarcomas on the other hand can arise in almost any tissue. While they should be classified based on the type of malignancy, they can often get classified based on organ of origin instead. For instance, radiation-associated breast angiosarcomas were typically characterized as breast cancers as resections were often performed on previous breast cancer patients and the procedure codes were similar for both types of malignancies.[24] Not surprisingly, sarcomas arising in the extremities (representing a larger percentage of cases performed by orthopedic oncology) were classified more accurately.[24]

Future of cancer databases

Rare diagnoses like sarcoma have the potential to benefit greatly from big data. Nevertheless, the challenges presented in this review emphasize the variability and vague nature of definitions for procedures and diseases that may lead to coding inaccuracies that can be propagated through various local and national datasets. Popular national datasets, which many researchers, including the authors, have used may not be as comprehensive as expected for studying population-based outcomes for rare malignancies such as sarcomas which are not organ-based (Table 2). Future endeavors to improve data quality may require validation of codes to appropriately identify certain diagnoses. As the World Health Organization develops further definitions of sarcoma for future ICD10 and ICD-O-3 codes, there should be multidisciplinary discussions and agreements on guidelines and definitions to help both physicians and coders better identify these malignancies. We can also improve on current database architecture models to make population-based registries more clinically relevant. Innovative methods to construct large, time-sensitive, population-based cancer registries that collect clinically relevant variables need to be explored and implemented in order to ensure good data quality.

Moving toward Sarcoma-Specific Databases

Problems with validating data for sarcomas in large population-based oncologic databases call attention to the benefits of sarcoma-specific databases. In one study, a dataset compiled from six dedicated sarcoma centers in six countries was used to validate a nomogram that predicts disease free survival and survival after primary resection for retroperitoneal sarcoma (RPS).[29] The validation of this RPS nomogram provides oncologists with a more accurate tool for calculating prognosis than the current AJCC classification.[29] Other successful sarcoma-specific databases have been cited in the literature. A multidisciplinary Danish Sarcoma group developed a national, centralized Danish Sarcoma Database (DSD) in 2009 and since its inception, there have been 2,000 patients registered in the database.[30] All sarcoma patients in Denmark are referred to two high volume centers where all clinicians are obligated to report to the database and consent is not required for data collection and registration, making this a very comprehensive dataset. [30] Data for this database are entered by the clinicians responsible for the patient and audited by dedicated database managers who validate all data with medical records.[30] The database allows linkage to the Danish Civil Registration System for mortality and demographic variables as well as other databases to capture gynecological sarcomas that are otherwise not included in the DSD.[30] However, sarcoma patients in the DSD are registered using ICD codes. If sarcomas are under-coded in the Danish centers as they have been found in our own study, there may still be a significant underestimation of sarcomas in the DSD requiring further validation of coding practices.[24] Also, since a vast majority of sarcomas are treated at two nationally approved sarcoma centers in Denmark, data collection and aggregation is much more streamlined than in other countries.[30] This efficient and well-organized approach is not easily generalizable. For instance, building a national sarcoma database in the United States may be much more time-consuming and near impossible due to the decentralized nature of sarcoma care.

However, well-curated single institutional or multi-institutional sarcoma databases may be as informative and significant as national databases. Sarcoma-specific databases in the United States have been largely limited to a single institution or consortiums. The Sarcoma Alliance for Research through Collaboration (SARC), a non-profit organization, collects and pools prospective data from SARC institutions to create multiple datasets such as the SARC Clinical Data Repository which includes 12 clinical trials.[31] Large single institution databases from high-volume centers have also demonstrated substantial merit in sarcoma research. The Memorial Sloan-Kettering Cancer Center manages a prospectively collected sarcoma database that was started in 1982 and as of 2013, 10,000 patients were entered in the database.[32] While it only includes one institution, the large number of patients in the registry allow for the development of disease-specific nomograms and predictive analytics that are typically limited to large, population-based databases.

Sarcoma-specific databases may also beneficial since it can include genomic data as well as biospecimen banking, critical to understanding the basic science behind sarcoma occurrence and cure. In the Cancer Moonshot Initiative, one of the priorities listed was the development of a Premalignant Cancer Atlas (PMCA) that would allow for linkage of genomic data with clinical information.[1] Currently, the National Institute of Health (NIH) houses The Cancer Genome Atlas Project (TCGA) which, to date, has identified a multitude of molecular alterations in 206 cases of adult soft tissue sarcomas.[33] Similarly, the benefits of genomic data and precision medicine has been seen on individual patients or a small select group of cases but expanding these advancements on a large-scale has not yet been done and has been proven to be challenging. High-volume sarcoma centers may benefit from a collaborate effort to create a multidisciplinary and comprehensive sarcoma-specific database that allows for linkage of specimens with genomic and clinical data.

Benefits of Real World Evidence

Challenges of using existing large population-based cancer databases for sarcoma research can be addressed with a unified effort to improve sarcoma database design and data collection. Whether it is a single institution, consortium, or nationally led effort, the formation of large sarcoma databases with meaningful clinical variables is imperative to future research initiatives. Registries are more valuable when there are uniform standards for data quality, efficient and automated data collection methods, and allowance for data linkage across databases.[34,35] The rise of the electronic health record (EHR) has provided researchers with the opportunity to leverage tehcnology for more efficient and automated data extraction. The American Society of Clinical Oncology (ASCO) recently released a position statement stating the significance of observational data and its role in representing a more realistic and generalizable view of patient care and treatment patterns.[36,37] Clinically relevant data from EHRs may fill gaps left behind by administrative databases. Furthermore, the collection of real world evidence (RWE), which can represent an accurate picture of actual clinical practice, can supplement data acquired in strictly controlled clinical trial settings.[38] Some practical examples of successful utilization of EHR data for quality improvement and research include the Veteran’s Affairs (VA) inpatient evaluation center and the MIMIC (Multiparameter Intelligent Monitoring in Intensive Care) II databases; the VA inpatient evaluation center collected inpatient data from the VA patient records from over 100 hospitals and the MIMIC II database collected physiologic data for over 30,000 intensive care unit patients at Beth Israel Deaconness Medical Center over an 8 year period.[19,39–41] These initial successes highlight the great potential of using real world evidence to complete databases and validate existing administrative data.

Currently, there are widespread efforts led by both academic and industry leaders to improve oncologic data quality using RWE. The most notable is the recent American Society of Clinical Oncology (ASCO) led effort called the Cancer Learning Intelligence Network for

Quality (CancerLinQ).[41] CancerLinQ, a nonprofit subsidiary of ASCO, captures and aggregates data from EHRs via direct feeds and automatically enters the information in to a series of cloud-based databases.[42,43] The database stores a combination of structured data that are automatically captured and unstructured data retrieved through manual data abstraction. Currently, the data are mainly utilized to track quality improvement metrics.[44] CancerLinQ is a massive undertaking by ASCO with the potential to provide providers and researchers with access to real world data from a variety of practice types representing thousands of oncologists across the country.[45]

A similar industry approach has been taken on by Flatiron Health, a for-profit company that was recently acquired by Roche, that is curating EHR data from more than 200 US cancer centers including oncology clinics and academic medical centers. Similarly to CancerLinQ, Flatiron Health created an automated collection and aggregation system to extract structured data that are then de-identified and entered in to a standard data model. The database is supplemented by the collection of unstructured data by human staff trained in chart abstraction; inter and intra-abstractor agreement is continuously monitored in order to maintain data quality.[38] The major advantage of this particular database is its ability to be linked to other important oncologic databases such as the Foundation Medicine laboratory's genomic database and mortality data from the National Death Index (NDI).[38,46] Linkage to the NDI permitted the validation of mortality data available in EHRs that was collected by Flatiron Health, creating a more comprehensive resource.[46] For sarcoma researchers and oncologists who choose to participate in their initiative, the Flatiron Health database can provide high-quality, real world, longitudinal, and genomic data to advance sarcoma care and drug development. Sarcoma research, which is often limited by small numbers accompanied by the challenges of using large population-based databases, can benefit greatly from the development of the CancerLinQ and Flatiron Health databases that present researchers with access to a larger patient population linked to high quality clinical and genomic data.

Nevertheless, there are many limitations with the current derivation of RWE-based databases that warrants further scrutiny. While clinical trial data are produced from adherence to strict conditions such as eligibility criteria, certain diagnostic and treatment regiments, EHR data is not confined to such limitations. While some view this as a benefit as it allows observation and monitoring of the real-world patterns of oncologic treatment patterns and management, it can also lead to poor data quality due to the heterogeneous nature of EHR utilization, differences in workflows, or incomplete recording of data in discrete fields.[37] Inappropriate data collection is exacerbated by the expansion of large EHR systems that were fundamentally designed to assist with billing rather than to provide clinical decision support or facilitate research.[47] Unstructured data captured in EHRs remains a major limitation of RWE; data existing outside of structured or discrete fields are difficult to capture and require manual chart abstraction by experienced curators. This process is often slow, expensive, and lack appropriate quality control. [47] Yet, structured data as it currently exists has its own set of challenges. Requiring physicians to spend more time clicking through structured data fields, which is often slower than free text documentation, may lead to increased administrative burden and burnout. Natural language processing solutions are showing promise to help automate the conversion of free text data to structured data for research.[48]

Moreover, data standards for oncology have not been widely adopted and despite attempts to compensate for widespread differences in data models and ontologies across specialties, it is still difficult to collect complete and uniform data from many institutions. ASCO recently issued a position statement calling for legislation to mandate interoperability of EHRs amongst hospital systems and oncology practices.[49] The Office of the National Coordination for Health Information Technology (ONC-HIT) has addressed ASCO's requests with a ten year roadmap that predicts the expansion of interoperable health information technology and the development of a learning health system enabled by nationwide interoperability by 2024.[50] Finally, the ability to scale technology driven data collection methods globally must be considered as a significant limitation. As of 2015, 57 out of 125 (46%) countries reported having any kind of national EHR system.[51] For low and middle income class countries with limited or no EHR systems, RWE-based data collection may not feasible or very difficult to implement. Although electronic information systems are being increasingly adopted by developing countries,[51] particularly with the support of outside funding and donors, the scalability and feasibility of RWE-based databases will remain a challenge for many countries. Given all of these limitations, RWE-based databases may require further development and validation prior to widespread use for sarcoma research.

Conclusion:

Sarcoma accounts for less than 1% of new adult cancer cases annually. While recent data[52–54] show that patients who present in a specialized sarcoma center have better outcomes (fairly common in Europe), many sarcoma patients in the US continue to present to both low and high volume centers alike making the consolidation of sarcoma data from a few centers nearly impossible. Consequently, large population-based databases have been increasingly utilized to perform highly powered sarcoma research. Nevertheless, the various limitations of population-based databases for sarcoma research have stunted more widespread adoption. For example, sarcoma remains one of the least studied types of cancer using the NCDB.[14]

The question remains then, what is the best approach to collect, aggregate, and share sarcoma data? Some high volume sarcoma centers mentioned in our review have attempted to answer this by creating single institution or consortium based sarcoma-specific registries. ASCO and Flatiron Health have been more innovative and have leveraged EHRs to create large, prospective databases to capture real world evidence along with important clinical variables often not captured with structured data fields. Unfortunately, all of these solutions are accompanied by unique limitations that make it impossible to safely draw definitive conclusions about sarcoma incidence, diagnostic and treatment methods, and prognosis from just one source. It is essential that when utilizing any large database, queries must be appropriate for the types of data that are available and data elements should be studied carefully to prevent introducing significant biases to the analyses. A checklist providing the strengths and limitations of widely used datasets was recently published with an aim to provide a guide for all researchers wishing to use large databases.[55] Specifically, for those using the NCDB, the published guide on this particular dataset should be reviewed carefully to understand both the benefits and challenges for sarcoma research.[56]

Consequently, what this review emphasizes is the importance of breaking down data silos and fostering a collaborative environment to answer important questions for sarcoma patients. This is particularly important as technology enabled data collection exponentially grows without the capability to harness the ever-increasing volume of the data to our advantage. To put this in perspective, more than 90% of the existing digital data across all fields, including medicine, were collected in the past 2 years and only 1% has been analyzed.[1,57] Therefore, efforts to pool resources to collect and analyze oncologic data should be encouraged, if not mandated. The advantages of the varying cancer initiatives and programs including the large population-based databases, CancerLinQ, Flatiron Health, and single institution efforts can be merged together to establish a more effective national cancer data ecosystem[1,58] and alleviate the limitations of current databases and registries. Collaboration and multidisciplinary efforts to create these databases will require extensive resources, not just from stakeholders but from policy-makers and future oncologists and data scientists. Building well-defined and appropriately structured sarcoma databases will not be possible without Center for Medicare and Medicaid Services (CMS) involvement to develop policies that mandate interoperability and data-sharing across providers and health systems. High-volume sarcoma centers that already have well-established comprehensive databases must collaborate to expand outside of the single institution model. Moreover, formal training in clinical informatics should be encouraged for sarcoma oncologists interested in database work should be encouraged to help data scientists create a clinically relevant sarcoma database. Since sarcoma is not only a difficult disease to treat, but also to study comprehensively, researchers and physicians must push for increased collaboration and multidisciplinary efforts to create high-value oncology data sources.

Acknowledgments

Funding sources: Dr. Heather Lyu was supported by the National Library of Medicine Institutional training grant for research training in biomedical informatics and data science (T15) under award number T15LM007092

References

1. Jaffee EM, Dang CV, Agus DB, Alexander BM, Anderson KC, Ashworth A, Barker AD, Bastani R, Bhatia S, Bluestone JA, Brawley O, Butte AJ, Coit DG, Davidson NE, Davis M, DePinho RA, Diasio RB, Draetta G, Frazier AL, Futreal A, Gambhir SS,Ganz PA, Garraway L, Gerson S, Gupta S, Heath J, Hoffman RI, Hudis C,Hughes-Halbert C, Ibrahim R, Jadvar H, Kavanagh B, Kittles R, Le QT, Lippman SM, Mankoff D, Mardis ER, Mayer DK, McMasters K, Meropol NJ, Mitchell B, Naredi P, Ornish D, Pawlik TM, Peppercorn J, Pomper MG, Raghavan D, Ritchie C, Schwarz SW, Sullivan R, Wahl R, Wolchok JD, Wong SL, Yung A. Future cancer research priorities in the USA: a Lancet Oncology Commission. Lancet Oncol. 2017 11;18(11):e653–e706. doi: 10.1016/S1470-2045(17)30698-8. Epub 2017 Oct 31. Review. [PubMed: 29208398]
2. Laney D 3D Data Management: Controlling Data Volume, Velocity, and Variety. META group Inc., 2001 http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
3. Murphy M, Alavi K, Maykel J. Working with existing databases. Clin Colon Rectal Surg. 2013 3;26(1):5–11. doi: 10.1055/s-0033-1333627. Review. [PubMed: 24436641]
4. Wild S, Fischbacher C, McKnight J; (on Behalf of the Scottish Diabetes Research Network Epidemiology Group). Using Large Diabetes Databases for Research. J Diabetes Sci Technol. 2016 8 22;10(5):1073–8. doi: 10.1177/1932296816645120. Print 2016 Sep. Review. [PubMed: 27127207]

5. Park HS, Lloyd S, Decker RH, Wilson LD, Yu JB. Limitations and biases of the Surveillance, Epidemiology, and End Results database. Curr Probl Cancer. 2012 Jul-Aug;36(4):216–24. Doi: 10.1016/j.currproblcancer.2012.03.011. [PubMed: 22481009]

6. Lloyd S, Park HS, Decker RH, Wilson LD, Yu JB. Using the Surveillance, Epidemiology, and End Results database to investigate rare cancers, second malignancies, and trends in epidemiology, treatment, and outcomes. Curr Probl Cancer. 2012 Jul-Aug;36(4):191–9. doi: 10.1016/j.currproblcancer.2012.03.008. Epub 2012 Apr 10. Review. [PubMed: 22495057]

7. Johnson AC, Ethun CG, Liu Y, Lopez-Aguiar AG, Tran TB, Poultsides G, Grignol V, Howard JH, Bedi M, Gamblin TC, Tseng J, Roggin KK, Chouliaras K, Votanopoulos K, Cullinan D, Fields RC, Delman KA, Wood WC, Cardona K, Maithel SK. Studying a Rare Disease Using Multi-Institutional Research Collaborations vs Big Data: Where Lies the Truth? J Am Coll Surg. 2018 6 12 pii: S1072-7515(18)30393-4. doi:10.1016/j.jamcollsurg.2018.05.009. [Epub ahead of print].

8. Ferrari A, Sultan I, Huang TT et al. Soft tissue sarcoma across the age spectrum: A Population-Based Study from the Surveillance Epidemiology and End Results Database. Pediatr Blood Cancer. 2011; 57(6): 943–949. [PubMed: 21793180]

9. Peng KA, Grogan T, Wang MB. Head and Neck Sarcomas: Analysis of the SEER Database. Otolarngol Head Neck Surg. 2014; 151(4): 627–633.

10. Ca Thiels, Bergquist KR, Krajewski AC, et al. Outcomes of Primary Colorectal Sarcoma: A National Cancer Data Base (NCDB) Review. 2017; 21(3): 560–568.

11. Datar M and Khanna R. Inpatient burden of gastrointestinal stromal tumors in the United States. J Gastrointest. Oncol 2012; 3(4): 335–341. [PubMed: 23205310]

12. Corey RM, Swett K, Ward WG. Epidemiology and survivorship of soft tissue sarcomas in adults: a national cancer database report. Cancer medicine. 2014;3(5):1404–1 [PubMed: 25044961]

13. Jagsi R, Bekelman JE, Chen A, Chen RC, Hoffman K, Shih YC, Smith BD, Yu JB. Considerations for observational research using large data sets in radiation oncology. Int J Radiat Oncol Biol Phys. 2014 9 1;90(1):11–24. Doi: 10.1016/j.ijrobp.2014.05.013. Review. [PubMed: 25195986]

14. Su C, Peng C, Agbodza E, Bai HX, Huang Y, Karakousis G, Zhang PJ, Zhang Z. Publication trend, resource utilization, and impact of the US National Cancer Database: A systematic review. Medicine (Baltimore). 2018 3;97(9):e9823. doi:10.1097/MD.0000000000009823. Review. [PubMed: 29489679]

15. Garland A, Gershengorn HB, Marrie RA, Reider N, Wilcox ME. A Practical, Global Perspective on Using Administrative Data to Conduct Intensive Care Unit Research. Ann Am Thorac Soc. 2015 9;12(9):1373–86. doi: 10.1513/AnnalsATS.201503-136FR. Review. [PubMed: 26148250]

16. Scosyrev E, Messing J, Noyes K, Veazie P, Messing E. Surveillance Epidemiology and End Results (SEER) program and population-based research in urologic oncology: an overview. Urol Oncol. 2012 Mar-Apr;30(2):126–32. doi: 10.1016/j.urolonc.2009.11.005. Epub 2010 Apr 3. Review. [PubMed: 20363162]

17. Singh JA, Ayub S. Accuracy of VA databases for diagnoses of knee replacement and hip replacement. Osteoarthritis and cartilage. 2010;18(12):1639–42. [PubMed: 20950694]

18. Sarrazin MS, Rosenthal GE. Finding pure and simple truths with administrative data. JAMA 2012; 307:1433–1435. [PubMed: 22474208]

19. Cooke CR, Iwashyna TJ. Using existing data to address important clinical questions in critical care. Crit Care Med. 2013 3;41(3):886–96. doi:10.1097/CCM.0b013e31827bfc3c. Review. [PubMed: 23328262]

20. Gologorsky Y, Knightly JJ, Chi JH, Groff MW. The Nationwide Inpatient Sample database does not accurately reflect surgical indications for fusion. Journal of neurosurgery Spine. 2014;21(6): 984–93 [PubMed: 25325170]

21. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. Ann Intern Med 1993;119(8):844–850 [PubMed: 8018127]

22. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. Agreement of Medicare claims and tumor registry data for assessment of cancer-related treatment. Med Care 2000;38(4):411–421 [PubMed: 10752973]

23. Kelly M, Lamah M. Evaluating the accuracy of data entry in a regional colorectal cancer database: implications for national audit. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland. 2007;9(4):337–9. [PubMed: 17432986]

24. Lyu HG, Stein LA, Saadat LV et al. Assessment of the Accuracy of Disease Coding Among Patients Diagnosed with Sarcoma. JAMA Oncology 2018;

25. American Cancer Society. Cancer Facts and Figures 2017. Atlanta, Ga: American Cancer Society; 2017.

26. Faciszewski T, Broste SK, Fardon D. Quality of data regarding diagnoses of spinal disorders in administrative databases. A multicenter study. The Journal of bone and joint surgery American volume. 1997;79(10):1481–8. [PubMed: 9378733]

27. Dancsok AR, Asleh-Aburaya K, Nielsen TO. Advances in sarcoma diagnostics and treatment. Oncotarget. 2017;8(4):7068–93. [PubMed: 27732970]

28. Jo VY, Fletcher CD. WHO classification of soft tissue tumours: an update based on the 2013 (4th) edition. Pathology. 2014;46(2):95–104. [PubMed: 24378391]

29. Raut CP, Miceli R, Strauss DC. External Validation of a Multi-Institutional Retroperitoneal Sarcoma Nomogram. Cancer. 2016; 122(9):1417–1424. [PubMed: 26916507]

30. Jørgensen PH, Lausten GS, Pedersen AB. The Danish Sarcoma Database. Clin Epidemiol. 2016 10 25;8:685–690. eCollection 2016. Review. [PubMed: 27822116]

31. Sarc: collaborating to cure sarcoma. https://sarctrials.org/who-we-are Accessed July 24, 2018.

32. Brennan MF, Antonescu CR, Moraco N, Singer S. Lessons learned from the study of 10,000 patients with soft tissue sarcoma. Ann Surg. 2014 9;260(3):416–21; 10.1097/SLA.0000000000000869. [PubMed: 25115417]

33. The Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcoma. Cell. 2017 11 2; 171(4):950–965. [PubMed: 29100075]

34. MacCallum C, Skandarajah A, Gibbs P, Hayes I. The Value of Clinical Colorectal Cancer Registries in Colorectal Cancer Research: A Systematic Review. JAMA Surg. 2018 6 20. doi: 10.1001/jamasurg.2018.1635. [Epub ahead of print].

35. Loke YK. Use of databases for clinical research. Arch Dis Child. 2014 6;99(6):587–9. doi: 10.1136/archdischild-2013-304466. Epub 2014 Jan 31. Review. [PubMed: 24489362]

36. Visvanathan K, Levit LA, Raghavan D, et al. Untapped potential of observational research to inform clinical decision making: American Society of Clinical Oncology research statement. J Clin Oncol. 2017;35(16):1845–1854. [PubMed: 28358653]

37. Miksad RA, Meropol NJ. Carcinoembryonic Antigen-Still More to Learn From the Real World. JAMA Oncol. 2018 3 1;4(3):315–316. doi: 10.1001/jamaoncol.2017.4408. [PubMed: 29270625]

38. Agarwala V, Khozin S, Singal G, O'Connell C, Kuk D, Li G, Gossai A, Miller V, Abernethy AP. Real-World Evidence In Support Of Precision Medicine: Clinico-Genomic Cancer Data As A Case Study. Health Aff (Millwood). 2018 5;37(5):765–772. doi: 10.1377/hlthaff.2017.1579. [PubMed: 29733723]

39. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med. 2011; 39(5):952–960. [PubMed: 21283005] [PubMed: 21283005]

40. Render ML, Hasselbeck R, Freyberg RW, et al. Reduction of central line infections in Veterans Administration intensive care units: an observational cohort using a central infrastructure to support learning and improvement. BMJ Qual Saf. 2011; 20(8):725–732.

41. Jain R, Kralovic SM, Evans ME, et al. Veterans Affairs initiative to prevent methicillin-resistant Staphylococcus aureus infections. N Engl J Med. 2011; 364(15):1419–1430. [PubMed: 21488764] [PubMed: 21488764]

42. Miller RS, Wong JL. Using oncology real-world evidence for quality improvement and discovery: the case for ASCO's CancerLinQ. Future Oncol. 2018 1;14(1):5–8.doi: 10.2217/fon-2017-0521. Epub 2017 Oct 20.

43. Miller RS. CancerLinQ Update. J Oncol Pract. 2016 10;12(10):835–837. [PubMed: 27531380]

44. Blayney DW, McNiff K, Eisenberg PD et al. Development and future of the American Society of Clinical Oncology's Quality Oncology Practice Initiative. J. Clin. Oncol 32, 3907–3913 (2014). [PubMed: 25225418]

45. West Cancer Center joins CancerLinQ® as participating practice. www.asco.org/about-asco/press-center/news-releases/west-cancer-center-joins-cancerlinq-participating-practice.

46. Curtis MD, Griffith SD, Tucker M, Taylor MD, Capra WB, Carrigan G, Holzman B, Torres AZ, You P, Arnieri B, Abernethy AP. Development and Validation of a High-Quality Composite Real-World Mortality Endpoint. Health Serv Res. 2018 5 14. doi: 10.1111/1475-6773.12872. [Epub ahead of print].

47. Berger ML, Curtis MD, Smith G, Harnett J, Abernethy AP. Opportunities and challenges in leveraging electronic health record data in oncology. Future Oncol. 2016 5;12(10):1261–74. doi: 10.2217/fon-2015-0043. Epub 2016 Mar 8. [PubMed: 27096309]

48. Kreimeyer K, Foster M, Pandey A et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. Jounrl of Biomedical Informatics. 2017: 73:14–29. [PubMed: 28729030]

49. American Society of Clinical Oncology. Cancer-specific data sharing standards for communication, collaboration, and coordination of care. www.asco.org

50. The Office of the National Coordinator for Health Information Technology. Connecting Health and Care for the Nationa: A Shared Nationwide Interoperability Roadmap. https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf

51. Silverstre E How Electronic Health Records Strength the Health Systems of Low- and Middle-Income Countries: Learning from Eswatini and Mexico. Chapel Hill, NC: MEASURE Evaluation 2018 9.

52. Gronchi A, Crago A, Raut CP. Minimally Invasive Surgery for Retroperitoneal Sarcoma: Just Because We Can Does Not Mean We Should. Ann Surg Oncol. 2018 8;25(8):2129–2131. doi: 10.1245/s10434-018-6572-9. Epub 2018 Jun 14. [PubMed: 29948424]

53. Toulmonde M, Bonvalot S, Méeus P, et al. Retroperitoneal sarcomas: patterns of care at diagnosis, prognostic factors and focus on main histological subtypes: a multicenter analysis of the French Sarcoma Group. Ann Oncol. 2014;25(3):735–742. [PubMed: 24567518]

54. Gronchi A, Strauss DC, Miceli R, et al. Variability in patterns of recurrence after resection of primary retroperitoneal sarcoma (RPS): a report on 1007 patients from the multi-institutional collaborative RPS Working Group. Ann Surg. 2016;263(5):1002–1009. [PubMed: 26727100]

55. Haider AH, Bilimoria KY, Kibbe MR. A Checklist to Elevate the Science of Surgical Database Research. JAMA Surg. 2018;153(6):505–507

56. Merkow RP, Rademaker AW, Bilimoria KY. Practical Guide to Surgical Data Sets: National Cancer Database (NCDB). JAMA Surg. Published online April 4, 2018.

57. Toga AW, Dinov D. Sharing big biomedical data. J Big Data 2015; 2: 7. [PubMed: 26929900]

58. Miller K NIH launches a united ecosystem for big data. 12 big data to knowledge centers of excellence funded. Jan 8, 2015 http://biomedicalcomputationreview.org/content/nih-launches-united-ecosystem-big-data (accessed May 9, 2017).

Table 1:

Summary of the Large National Databases Currently Available for Sarcoma Research

Database Name	**Description/purpose**	**Method of Data Collection**	**Accessibility**	**Linkage to other databases**	**No. of publications**	**Years of data currently available**
Surveillance, Epidemiology, and End Results Program Database ***(SEER)***	Develop and report national estimates of cancer incidence and mortality Monitoring annual cancer incidence trends Providing continuous information on trends in therapy and changes in patient survival	Manual chart abstraction by trained registrars	Deidentified data are publicly available	Medicare database Medicare Health Outcomes Sruvey Consumer Assessment of Healthcare Providers and Systems	302	1975–2015
National Program of Cancer Registries ***(NPCR)***	Same purpose as SEER but in different states with separate submission dates	Manual chart abstraction by trained registrars	Deidentified data are publicly available	Same as SEER	5	2001–2015
National Cancer Database ***(NCDB)***	Analyze and track treatments and outcomes for patients with cancer diagnoses Provide performance metrics for participating centers	Manual chart abstraction by trained registrars	Investigators in participating organizations must apply for deidentified participant user files	Claims data	25	2003–2014

Table 2:

Challenges and Opportunities for Big Data in Sarcoma Research

Challenge	Opportunity
Dependence on administrative/billing data that may not have high clinical accuracy	Improve data quality by validating codes to appropriately identify key diagnoses
Difficulty validating initial coding process based on preexisting documented diagnosis codes	Multidisciplinary discussions and agreements on data elements, such as ICD10 and ICD-O-3 codes, prior to submission to large databases Reduce variability in sarcoma nomenclature and cancer classification
Errors in large databases may be amplified for rare diseases, such as Sarcoma	Improve on current database architecture models to make population-based registries more clinically relevant Utilization of sarcoma-specific databases with more granular, clinically relevant data
Data silos created due to lack of information sharing amongst multiple institutions and databases	Automated aggregation of real world data (both structure) supplemented by NLP-assisted manual curation of unstructured clinical documentation such as the ASCO CancerLinQ initiative and Flatiron Health's database. Linkage to administrative databases to validate information in real world evidence based datasets.