**Project Title:** American Cancer Databases

**Names & Email Addresses of Investigators/Team Members:**

Bethel Kumsa; Ben Goodwin | [bkumsa@smu.edu](mailto:bkumsa@smu.edu); [bgoodwin@mail.smu.edu](mailto:bgoodwin@mail.smu.edu)

**Introduction:**

We started this project with an interest in learning how cancer diagnoses in the United States have changed by demographic. In our search for tools to support this effort, we discovered that cancer patient data is heavily regulated, and in many cases proprietary.

The institutional goals of the governing bodies expending considerable resources to create and maintain clinical cancer and cancer research databases heavily impact the availability of the data and tools instituted. These governing bodies stringently regulate the access approval and submissions process for their data and data systems as well as their specialized software in an attempt to provide security and confidentiality as well as protect the integrity of their systems, which not only require extensive training to maintain but also to interact with properly.

**Background/Research:**

The National Cancer Database (NCDB) was created as a tool for documenting, researching, and improving cancer diagnoses, treatments, and outcomes in the United States. In the decades since its 1989 establishment through a partnership between the Commission on Cancer and the American College of Surgeons, the NCDB has undergone various changes in support of its foundational goals.

The database’s first structural changes took place in 1996 and 2001. Prior to 1996, hospitals were able to voluntarily submit clinical cancer data to the NCDB. However, beginning in 1996 submission of cancer data to the NCBD became a requirement of members of the Commission on Cancer (CoC). Following 2001, database submissions and privileges were limited exclusively to member institution of the CoC. NCDB’s next largest systemic changes took place in 2002 and 2005, during which the quality constraints on data submissions were greatly increased.

NCDB has honed other various notable strengths. Amongst these strengths is geographically diverse data representing 49 states and Puerto Rico. Further, collected data is highly standardized. To meet its high-quality constraints, the NCDB employs rigorous internal annual quality testing as well as trained, Certified Tumor Registrars (CTRs) who maintain and correct data in real-time.

Despite these strengths, there are concerns that NCDB’s sole reliance on CoC member institutions for data collection may bias the data’s representativeness of the larger American population. Fortunately, amongst NCDB’s growing strengths are large and growing sample sizes, which may effectively offset such bias. Recent research found that the NCDB records 72% of new cancer diagnoses in the United States. This is an increase from previous estimates. Notably, the percent of cancer diagnoses represented by the database vary by cancer type, geographical location, and demographics.

**Case Study:**

A possible major area of improvement is the NCDB’s accessibility. NCDB limits both the submission and accessibility of its data to members of its governing institutions. To more clearly understand how NCDB’s accessibility compares to related national databases with similar goals, we attempt querying and analyzing breast cancer data through the National Cancer Database (NCDB), Surveillance, Epidemiology, and End Results Program (SEER), and the National Cancer Institute’s Genomic Data Commons (GDC).

National Cancer Database

The National Cancer Database is an offshoot of an organization started in 1989 by the American College of Surgeons (ACoS) and the Commission on Cancer (CoC) that acts as a multidisciplinary consortium of professional organizations that strive to improve cancer care through setting standards, prevention, research, education, and the monitoring of comprehensive cancer care. The CoC also accredits hospitals as cancer centers based on their ability to provide a broad range of cancer-related services and specialists. The goal of the ACDB is to improve cancer care through setting standards, prevention, research, education, and the monitoring of comprehensive cancer care.

To gain access to the NCDB, the requesting entity must be a hospital from a CoC accredited program. Once that initial criteria is established, the entity must agree to the data use agreement and provide a letter of support from CoC accredited program, all in addition to a completed application. When the ACDB was first established, any hospital could voluntarily report data to the NCDB, however, in 1996, all CoC-approved hospitals were required to report cancer cases to the NCDB, and in 2001, participation and the associated advantages of reporting to the NCDB were limited to hospitals who earned CoC approval.

The data within the NCDB is contained in PUFs (Participant User Files), of which are HIPPA compliant. There is a limited subset of the data available for public use, however, the institutional data is robust and access comes with a variety of software packages to evaluate the data depending on the needs of the user. Due to the stringent access approval process, only CoC approved hospitals are permitted to access the data. The data is walled and only exists in detail with the correct credentials. The full dataset contains data on 21 million cancer patients diagnosed between 1985 and 2005, the data comes from 1,430 hospitals.

To explore the data (publicly accessible) The online data exploration tool allows the user to select cancer site (breast) and case type (All diagnosis types) and allows the user to select up to three analysis variables along with the year. As a baseline the cancer selected across databases was breast cancer.

The tool is extremely intuitive, albeit a bit limited in scope (no ability to model relationships).

Surveillance, Epidemiology, and End Results

The Surveillance, Epidemiology, and End Results program of the National Cancer Institute is a source of epidemiologic information on the incidence and survival rates of cancer in the United States. The SEER program collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 34.6% of the population of the United States. The SEER program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data. SEER began collecting data in 1973 in a limited number of states and cities within the United States. The National Cancer Institute funds for the program come from the Centers for Disease Control and Prevention through the National Program of Cancer Registries and gets additional funding from participating states. SEER is supported by the Surveillance Research Program in NCI’s division of Cancer Control and Population Services. Access to SEER is similar to the NCDB and requires an application and an institutional association.

The data is somewhat publicly accessible and can be accessed from the SEER website using a web browser. The public can access and visualize the data through the SEER\*Explorer application, which is web based and connects to the SEER database, although this web interface does not allow public users access to the raw data. The publicly accessible and available as a public service in print and electronic formats.

The SEER\*Explorer tool allows the user to select cancer site (breast) and a statistic to explore (SEER Incidence) and allows the user to select multiple explanatory variables along with the year, gender, state at diagnosis, rate type, and a precision option. As a baseline the cancer selected across databases was breast cancer.

The tool is extremely intuitive, user friendly and has robust data export options (export data, export plots, or share the web generated chart). The tool has many more customization options compared to the NCDB and features many statistical tools.

SEER is an authoritative source for cancer statistics in the United States and provides information on cancer statistics in an effort to reduce the cancer burden among the U.S population and its tool reflects that mission.

National Cancer Institute’s Genomic Data Commons

The National Cancer Institute’s Genomic Data Commons (GDC) was launched in June of 2016 through the National Cancer Institute. It coalesces data from The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) standardizing these datasets through shared bioinformatics pipelines and allowing their direct comparison for the first time. It houses over 2 petabytes of secured data encompassing a multitude of cancer patients and tumors. The GDC has established itself as an ever-growing knowledge network with the many tools as part of its data system, some of which include its data transfer tool, data portal, API, its organizations and collaborators, reports, data submission tool and more. Further, the GDC maintains the raw genomic data garnered through its collaborators for direct reanalysis through future computational techniques and innovations. The GDC provides a robust online data portal allowing public access, exploration, and analysis of approved data. In line with its central goal to accelerate, support, and drive precision oncology while maintaining privacy safeguards, controlled data can be accessed by biostatistician and cancer researchers through a rigorous approval process.

As our project was initially prompted by an interest in American breast cancer trends, we utilized the Genomic Data Commons to explore and analyze breast cancer data. In exploring the variables for the 1,384 breast cancer patients documented in the GDC’s publicly accessible data, we were intrigued by the fact that 98% of these cases were representative of women. As a result, we also extended our analysis and exploration to cover data on patients afflicted by prostate cancer and cancer of the kidney. This data set is represented by the 965 kidney cancer patients and 493 prostate cancer patients for whom data is publicly accessible through the GDC. Of the 965 kidney cancer patients, 66.94% represented males and 33.06% represented females, and all prostate cancer patients represented are males as females do not have prostates.

In addition to exploring gender-based insights for each of these diagnoses, we also analyzed other variables regarding treatment measures, patients’ medical history, and diagnoses. We learned that a majority of women, approximately 60% are diagnosed with cancer between 40 and 70 years of age, with the largest percentage of diagnoses taking place between 50 to 60 years of age. Notably, 70.95% of the breast cancer patients represented in this dataset, received a primary diagnosis of infiltrating duct carcinoma, which is a form of cancer rupturing milk ducts in the breast and invading surrounding tissue. Further, of the patients documented through the GDC, 63.44% of those diagnosed with breast cancer underwent treatment or therapy with 33.74% foregoing medical intervention. Disparately, 76.67% of prostate cancer patients documented in this dataset forewent treatment with 24.34% choosing to undergo treatment or therapy. For comparison, 50.98% individuals diagnosed with kidney cancer, are documented as undergoing treatment or therapy with 30.78% having this variable undocumented, and 13.47% documented as opting out of treatment or therapy. We were curious to research if age at diagnosis played a mediating factor in the decision to undergo or forego treatment, but the analysis tool does not allow this flexibility.

However, as compared to the NCDB and SEER analysis tools, the Genomic Data Commons allows more flexible and powerful analysis and exploration through its public, online data portal. As such, we also performed analysis in which we compared treatment and survival outcome data between cohorts. While cohorts are editable and may be uniquely formed to answer questions of interest, we maintained our varied diagnoses groups as distinct cohorts for comparison. Through the cohort analysis tool, we found that survival rates amongst breast cancer patients were higher than those of kidney cancer patients. This gap in survival is especially pronounced during the first 10 years following diagnoses. An opposing pattern is found when comparing survival rates of breast cancer and prostate cancer patients with prostate cancer patients displaying stronger survival outcomes from the first to the fourteenth year following diagnoses. A point of interest in our comparisons of these cancer diagnoses cohorts was the diagnoses age of these diseases. The data indicates a trend in which breast and kidney cancer are more likely to be diagnosed in individuals over the age of 70 than is prostate cancer with breast, kidney, and prostate cancer diagnoses over the age of 70 representing 18%, 25%, and 11% of their respective cohorts. The GDC provides over 100 variables for analyzing and exploring the many cancer diagnoses and genomic profiles available through their data system. While the public data portal tool is powerful for driving some insight, it is only the periphery of this powerful knowledge network.

**Conclusion:**

In exploring our topic, we chose to investigate three databases, the National Cancer Database (NCDB), Surveillance, Epidemiology, and End Results Program (SEER), and the National Cancer Institute’s Genomic Data Commons (GDC). Through this investigation, we learned that institutions housing and providing data requiring specialized oncology, biology, or genomic training were more likely to provide data in more granularity to the public than institutions providing clinical data that is more easily linked to a patient. As such, we found that the NCDB provided the most limited, summarized data and insights with the GDC offering a more robust, flexible public data analysis portal. At the completion of our investigation, we have learned the importance of bioinformaticians, biostatisticians, and oncology research groups composed of data scientists and cancer as well as genomic specialists.

**References:**

1. Bilimoria, K.Y., Stewart, A.K., Winchester, D.P. *et al.* The National Cancer Data Base: A Powerful Initiative to Improve Cancer Care in the United States. *Ann Surg Oncol* **15,**683–690 (2008). <https://doi-org.proxy.libraries.smu.edu/10.1245/s10434-007-9747-3>
2. Burge, Sarah et al. “Biocurators and Biocuration: Surveying the 21st Century Challenges.” *Database : the journal of biological databases and curation* 2012 (2012): bar059–bar059. Web.
3. Lerro, C.C., Robbins, A.S., Phillips, J.L. *et al.* Comparison of Cases Captured in the National Cancer Data Base with Those in Population-based Central Cancer Registries. *Ann Surg Oncol* **20,**1759–1765 (2013). <https://doi.org/10.1245/s10434-013-2901-1>
4. Huang, H. (2015). Domain knowledge and data quality perceptions in genome curation work.*Journal of Documentation, 71*(1), 116-142. doi:http://dx.doi.org.proxy.libraries.smu.edu/10.1108/JD-08-2013-0104
5. Boca SM, Panagiotou OA, Rao S, McGarvey PB, Madhavan S. Future of Evidence Synthesis in Precision Oncology: Between Systematic Reviews and Biocuration. JCO Precis Oncol. 2018;2:PO.17.00175. doi: 10.1200/PO.17.00175. Epub 2018 Apr 10. PMID: 31930186; PMCID: PMC6953752.
6. Schlick, C.J.R., Bentrem, D.J. Thirty Years of Counting Cancer Cases: The Collective Effort and Efficacy of the National Cancer Database. *Ann Surg Oncol* **26,**1583–1584 (2019). <https://doi-org.proxy.libraries.smu.edu/10.1245/s10434-019-07215-z>
7. Mallin, K., Browner, A., Palis, B. *et al.* Incident Cases Captured in the National Cancer Database Compared with Those in U.S. Population Based Central Cancer Registries in 2012–2014. *Ann Surg Oncol* **26,**1604–1612 (2019). <https://doi.org/10.1245/s10434-019-07213-1>
8. National Institutes of Health (NIH). 2021. *Newly launched Genomic Data Commons to facilitate data and clinical information sharing*. [online] Available at: <https://www.nih.gov/news-events/news-releases/newly-launched-genomic-data-commons-facilitate-data-clinical-information-sharing> [Accessed 20 February 2021].
9. Gdc.cancer.gov. 2021. *GDC Overview | NCI Genomic Data Commons*. [online] Available at: <https://gdc.cancer.gov/about-gdc/gdc-overview> [Accessed 20 February 2021].