USING CLINICAL INFORMATION FROM ELECTRONIC HEALTH RECORDS TO SUPPLEMENT NATIONAL CANCER REGISTRY DATA

Caroline A. Thompson, PhD, MPH
SDSU Graduate School of Public Health
UCSD Clinical and Translational Research Institute
Palo Alto Medical Foundation Research Institute

The Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute (NCI) was established in 1973 as a result of the National Cancer Act of 1971. The National Program of Cancer Registries (NPCR) was established by Congress through the Cancer Registries Amendment Act in 1992, and administered by the Centers for Disease Control and Prevention (CDC). NPCR and SEER together collect cancer data for the entire U.S. population. CDC and NCI, in collaboration with the North American Association of Central Cancer Registries, have been publishing annual federal cancer statistics in the United States Cancer Statistics: Incidence and Mortality report. Today, the SEER program is one of the premier cancer surveillance programs in the world, composed of population-based cancer registries covering 30% of the total US population. Information maintained in the cancer registry includes: demographic information, medical history, diagnostic findings, cancer therapy and follow up details. These data are used to evaluate cancer patient outcomes, quality of life, provide follow-up information, calculate survival rates, analyze referral patterns, allocate resources at the regional or state level, report cancer incidence as required under state law, and evaluate efficacy of treatment modalities.

In recent years, the SEER population-based cancer registries have been facing challenges in providing optimal support of cancer research. Most of these challenges relate to the increasing complexity of health care information. For example, a growing number of gene and molecular biomarkers have been discovered that are used by physicians to direct patient care, and it is crucial for cancer registries to collect this information in order to sustain the relevance of registry data to current cancer research. Hospital medical records are the traditional information source for cancer registries. But with more care being delivered at outpatient facilities, it is increasingly difficult for cancer registries to collect relevant information on all cancer cases diagnosed in their catchment areas. With the recent evolution of widespread electronic health tracking systems, such as electronic health records (EHR), there is great interest in understanding how the cancer registration systems described above may improve their efficiency and accuracy by supplementing their data collection efforts with external data from health care delivery systems. In support of the need to combine and streamline big data sources for the purpose of cancer reporting, there are a number of approaches that warrant exploration. First, the feasibility of supplementing cancer registries with new or more detailed data items, from existing data sources or from linkages with novel data sources, e.g. EHR. Second, the development, validation, evaluation of scalable tools to facilitate automatic or unsupervised extraction of specific data from medical records. In this study we conduct a large scale linkage of patients treated at Sutter Health with the California Cancer Registry, and evaluate the utility of supplementing cancer registry data from elements of the Sutter Health EHR.

Sutter Health is a non-profit health system that delivers healthcare coverage in 18 northern California counties across 150 ambulatory medical clinics, with more than 7 million patients and over 10 million outpatient visits per year. The demographics of the patient population are generally representative of the underlying population with respect to sex, age and race/ethnicity. The EpicCare EHR system, designed by the Epic Systems Corporation (Verona, WI), has been in use for nearly 15 years at Sutter Health. It is designed to collect details of all patient encounters, including laboratory results, procedures, medication orders, diagnoses, immunizations, radiologic reports, and routine testing, as well as demographics, medical and surgical history, and transactional detail about care utilization (providers seen, physician notes, dates

and times, communications between providers and patients, etc). We have created a de-identified dataset based on linking California Cancer Registry (CCR) data with information on 6 million adult Sutter Health patients. We will characterize socio-demographic, clinical, and outcome characteristics among the Sutter Health cancer patients identified by this linkage to establish a descriptive profile of this population. We will also link Sutter Health cancer patients to their neighborhood attribute measures to characterize their social and built environments and compare the neighborhood, socio-demographic, clinical, and outcome characteristics among Sutter Health cancer patients with non-Sutter Health cancer patients to determine the extent to which Sutter Health cancer patients are representative of the underlying cancer patient population.

After the linkage, we will conduct a feasibility study to evaluate the utility of supplementing specific selected cancer registry variables (race/ethnicity, birthplace, treatment, tumor biomarkers) with data from electronic health records (EHR) data, and; evaluate the potential for improving standard cancer reporting statistics (e.g., incidence, prevalence, treatment-specific survival) by augmenting race/ethnicity, birthplace, treatment, and tumor biomarkers registry data with supplementary data by:

a. performing a series of validation studies based on augmentation of registry data with EHR;

b. comparing the populations used in the validation studies to the overall Greater Bay Area Cancer Registry (GBACR) and California Cancer Registry (CCR) populations in order to assess the generalizability of the validation statistics to the underlying population-based sample; and

c. performing sensitivity analyses for misclassification of the standard cancer reporting statistics using the validation study results.

This study is a collaborative effort between the CPIC Greater Bay Area Cancer Registry and the Palo Alto Medical Foundation Research Institute. This collaborative effort will leverage the unique resources and expertise available at the two research institutes to form a rich, multiethnic resource for investigations of factors related to cancer etiology, treatment, survival, and other outcomes (e.g., treatment symptoms). Our objective is to produce a large-scale resource that will facilitate multi-disciplinary, cells-to-society research addressing key research questions in cancer etiology, survival, and outcomes in diverse populations in future collaborative grant applications.

In an era of need to maximize existing resources, it is important to consider all opportunities to achieve better accuracy and completeness of cancer registry data by leveraging additional, external (i.e., outside of SEER) sources. However, the implications of supplementing data in this less systematic way must be fully examined and understood first. We expect that this study will inform the cancer registry community of when it is appropriate to supplement cancer registry data with external sources, and what data items are amenable to being supplemented without bias in resultant statistics. We also anticipate that this approach (using sensitivity analysis for misclassification bias to evaluate the robustness of SEER data-derived statistics) will contribute substantially to the growing body of research already using quantitative bias analysis, and will provide examples for other SEER sites utilizing validation data to quantitatively describe uncertainty in their reports.

Acknowledgements

This study is funded by the National Cancer Institute (HHSN261201300005I/HHSN26100009).
Dr. Thompson is supported by the National Center for Advancing Translational Sciences (1KL2TR001444)