REGISTRAR PIP

Visit SEER*Educate: A comprehensive training platform for registry professionals

## What is More Important than Collecting Cancer Data?

Introduction

Sure hope everyone's initial reaction to the article's title question was "using it"! While we are all very much aware of what we do, we may not realize the impact this data could potentially have if we keep four critical aspects of how we do our work in mind as we draft procedures to describe our approach to data collection. The most impactful data is not only relevant, complete, and accurate; it is also timely.

While the initial focus of registries was to describe cancer patterns and trends, it later evolved to calculate trends. With a little standard setter imagination and unilaterally declared expansion of the dataset, access to and use of registry data has proven to be an effective way to capture additional information needed for cancer control and epidemiologic research, public health planning, patient care improvement and survival analysis.

It's easy to conceptualize that if we capture more, we can offer more. For most of us it's overwhelming to think about being able to "capture more so we can offer more" when we currently feel tapped out. Our brains already seem to be operating in overdrive trying to learn the relevance of new data items and the rules associated with capturing these data items correctly while also trying to do this fast enough to keep those who want the data happy.

The juggling of growing expectations, nearly flat funding levels, and available human resources (aka trained registrars) is frustrating. And yet, it is in this environment we need to imagine the possibilities to ensure our ongoing existence. Let me be the first to say that the tough part is imagining "the how" as it relates to the approach of changing data collection, but it is the key to being able to meet the challenges
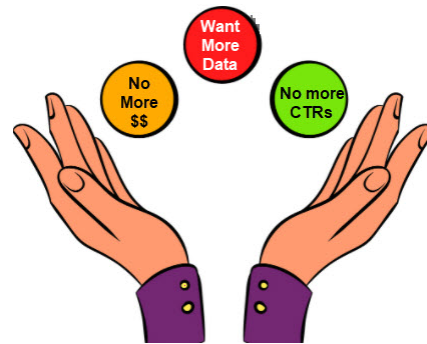
Technology Tools

As registrars, we tend to rely on the current abstract-centric model as the data collection model to use going forward because it has proven successful in the past. Maintaining the current abstracting model is simply not realistic given the financial constraints most registries are forced to operate under. However, clinging tightly to this data collection model in the face of growing expectations of registry data will not allow us to consider the advantages fast-changing technology can offer. Whenever this is uttered or written, many registrars' knee-jerk response is resistance to the idea of changing procedures because they believe embracing technology means they are headed toward the unemployment line.

Technology won't eliminate our jobs; it will help us keep them! Admittedly, our procedures will change, but for those of us who want employment in this field, there will still be jobs. We have to be open to the idea that doing

more with the same level of resources means we must embrace the advantages technology can bring to the profession. This is the only way we can meet the "capture more so we can offer more" challenge.

Baby Step #1 - Linkages

Many central cancer registries link data with other sources to address issues identified during the registry's data quality and completeness reviews. Such linkages identify missed cancer cases or data items and can help validate existing data. Linkages with other sources can also improve the usefulness of the registry data because it provides an opportunity to analyze data not typically abstracted for the cancer registry. The linkages help investigators with studies consider comorbidities, risk factors, treatment patterns, and the effectiveness of public health efforts to prevent cancer cases and improve cancer survival.

The following are a few examples of how linkages with outside sources have been used to enhance registry data while avoiding any abstracting cost associated with the collecting the additional data:

- Disease index files – to identify comorbidities and cancer recurrence
- Genealogy files – to better understand hereditary cancer burden
- Indian Health Service – to determine American Indian/Alaska Native race more accurately
- Insurance claims – to improve the capture of first course and to add subsequent courses of treatment
- Medicaid enrollment data - to allow researchers to evaluate stage and outcomes for some of the most vulnerable cancer patients
- Pathology and radiology files – to improve casefinding accuracy, data accuracy and completeness
- Screening programs (e.g., breast, cervical, colon) – to assess the effectiveness of screening
- State and National Death files - improves vital status and cause of death information
- Voter registration – to confirm residency and improve follow-up data items

Baby Step #2 - Natural Language Processing (NLP)

While linkages certainly helped central registries contain data collection costs, expand research options, and improve the completeness, accuracy, and timeliness of reporting incidence cases, it's what lies ahead that will transform registry operations. It is the work being done at the National Cancer Institute's (NCI) Surveillance Research Program (SRP) in the area of NLP tool and method development that will impact registrars most noticeably. Admittedly, this work is in its infancy, but noticeable progress is being made.

The challenge for the SRP is to assist registrars in handling the increasing demands placed on us for additional data items along with real-time submission to standard setters. To that end, it is clear we will need to turn to technology to meet these demands. The goal of the NLP project is to "improve the overall efficiency and quality of data abstraction and processing for cancer registries, and to enable acquisition of more detailed clinical data that may not be currently reported."

NLP tools can process free text documentation, including pathology reports, radiology reports, and oncology clinical notes, and can extract information. Experts train computer algorithms to complete tasks, including information extraction, de-identification, and classification. SRP is currently collaborating with four labs at the Department of Energy (DOE) to leverage the capabilities of high-performance computing to support implementation of more advanced NLP tools. The collaboration has already developed tools to:

- Semi-automatically abstract diagnosis date, site (including sub-site), histology, laterality, behavior, and AJCC stage from pathology reports
- Utilize datasets to train algorithms that will sort pathology reports based on reportability
- Perform quality assessments of SEER data
- De-identify data

NLP allows users to convert free text into machine readable, structured data, which are a growing part of the data in electronic health records (EHR) today. For example, NLP uses College of Pathology templates (CAP), an

example of structured data, from the EHR to derive the International Classification of Diseases for Oncology (ICD-O) codes for site and histology as well as diagnosis date and AJCC stage. Registrars using this technology can auto-populate fields from a CAP protocol into a commercialized registry system after verifying the information. The registrar's role will be one of quality assurance rather than manually entering data into a database. As mentioned previously, our jobs won't go away, they will simply change.

Data acquisition and information processing relies on expert cancer registrars. Current NLP efforts have targeted the following areas for development to assist us in our work:

- Case-finding – determining reportability
- Extracting currently collected clinical data elements from text documents (e.g., pathology reports)
- Extracting new clinical data elements from text documents (e.g., biomarkers from pathology reports, recurrence/progression from pathology or radiology reports)

Conclusion

Increased use of data linkages with other sources and the development of applicable NLP tools integrated into our daily processes has the potential to allow registries to create a dataset that is far more relevant, complete, accurate and timely than it is today. However, the "capture more so we can offer more" will simply remain a registry goal unless standard setters demonstrate an ongoing commitment to investing the necessary resources in negotiating the technical, legal, and ethical hurdles associated with linking potential datasets of interest to the registry and technology development streamlines hospital and central registry operations. When both things happen, that goal will become our reality.