Ann Surg Oncol (2014) 21:1506–1507
DOI 10.1245/s10434-014-3519-7

Annals of
SURGICAL ONCOLOGY
OFFICIAL JOURNAL OF THE SOCIETY OF SURGICAL ONCOLOGY

EDITORIAL – HEALTHCARE POLICY AND OUTCOMES

Big Data and Clinical Research in Oncology: The Good, the Bad, the Challenges, and the Opportunities

Christopher M. Pezzi, MD

Department of Surgery, Abington Memorial Hospital, Abington, PA

National cancer registries provide an ever-growing volume of data and increasing access to that data for the purpose of clinical research in oncology. In the United States, three national cancer-specific registries have been developed to collect data on cancer patients, their cancers, how they are treated, and their outcomes. Beginning with the passage of the National Cancer Act of 1971 and funded since 1973, the National Cancer Institute (NCI)'s Surveillance Epidemiology and End Results (SEER) program is a population-based registry from 20 U.S. geographic areas, covering ~28 % of the U.S. population.[1] In 1989, the American College of Surgeons (ACoS) Commission on Cancer (CoC) started a joint program with the American Cancer Society: the National Cancer Data Base (NCDB). Approximately 70 % of all new cancer cases diagnosed in the United States each year are currently captured by the NCDB, which contains the records of ~29 million patients from ~1,500 institutions (making the NCDB a hospital-based, not population-based, registry).[2] Finally, the National Program of Cancer Registries (NPCR) was established in 1992 and is administered by the U.S. Centers for Disease Control and Prevention (CDC). NPCR supports cancer registries in 45 states, representing 96 % of the U.S. population.[3] The data entered into each of these three national cancer registries are not collected completely independently of the others. The processes, system, and rules that govern the data collection for all three registries significantly overlap, as do the professionals (cancer registrars/CTRs) who actually collect and enter the data. The North American Association of Central Cancer Registries Inc. (NAACR), established in 1987, is a collaborative umbrella organization that develops and promotes uniform data standards for the cancer registries. All central cancer registries in the United States (and Canada) are members of NAACR,[4] although each registry may require a different subset of data elements to be reported.

In this issue of *Annals of Surgical Oncology,* In et al.[5] from the ACoS CoC and nearby Chicago hospitals, bring to light one of several important weaknesses of the data currently collected and reported in these national cancer registries: accurate information on local, regional, and distant recurrence rates and the timing of those recurrences after a first course of treatment is completed. The authors point out that the data reported by each of these registries have traditionally concentrated on the initial presentation and first course of treatment, with little follow-up information reliably collected, except death.

The NCDB does attempt to collect data on the time of first recurrence as well as the type of recurrence (local, regional, distant), and the authors examined the completeness of these data points in the NCDB for more than 700,000 patients with five common tumor types diagnosed between 2002 and 2005. Disappointingly, they report that complete information to allow an accurate determination if and when a recurrence had occurred after completion of the first course of treatment, and the type of recurrence, was lacking at a majority of the more than 1,400 hospitals for more than half of their patients. On average, hospitals had incomplete recurrence information on 56.7–66.7 % of patients studied. Only 9.0 % of hospitals collected recurrence information well on all five of the cancer sites examined.

The absence of reliable information on recurrence after treatment of primary cancer makes the determination of

This is an editorial to the article available at doi:10.1245/s10434-014-3516-x.

First Received: 20 January 2014;
Published Online: 20 February 2014

C. M. Pezzi, MD
e-mail: chmario@comcast.net

disease-free survival, disease-specific survival, distant-disease free survival, local recurrence rates, or regional nodal failure rates essentially impossible to calculate at the current time using the NCDB and the other registries. The data are often not collected, and they are so unreliable that information on recurrence is not made available. Clinical researchers are left with crude or overall survival as the only reliable long-term outcome in the NCDB. While crude survival is valuable, especially in the absence of more specific data on recurrence, crude survival is greatly influenced by the age of the patient at diagnosis, socio-economic factors, and comorbidities. In many cases, these non-treatment-related variables are more likely to explain differences in overall survival between institutions or regions than the treatment rendered. Sophisticated risk-adjustment methodologies that attempt to adjust observed survival rates in cancer registries for these non-treatment-related variables are complex, but such efforts are ongoing at the CoC and other groups.

The authors correctly suggest that one major reason for the high rate of incomplete information about recurrence is the lack of a mandatory requirement to report this information to the NCDB to maintain CoC accreditation. Patient factors (such as increasing comorbidities, higher cancer stage, nonprivate insurance, and longer distance lived from the hospital) and hospital factors (larger tertiary hospitals) were found to have significance, but their impact explained only a small amount of the variation. One solution would be a new CoC standard mandating accurate recurrence information. However, this would place a large additional burden on cancer registrars (at a time when there is a national shortage of CTRs), and it therefore comes with a significant cost. Nevertheless, this is a discussion that needs to be undertaken in collaboration with cancer registrars. Another equally important weakness of the NCDB and other national registries is a failure to accurately capture specific systemic chemotherapy agents and specific biologic agents (such as trastuzumab).

Despite limitations, including the lack of reliable recurrence information documented by In et al., the three national cancer registries remain incredibly valuable resources for clinical research in oncology in their current form and still contain vast amounts of unexplored but important information for future study. Detailed and accurate information on patient demographics, disease stage at presentation, patterns of first treatment, and overall survival can be found in over 100 data elements for each case entered in the NCDB. The NCDB makes available a participant user data file (PUF) to investigators at CoC-accredited cancer programs through an annual application period.[6] Interestingly, one of the most common errors in PUF applications in 2013 related to lack of knowledge about the absence of recurrence data in the NCDB and PUF.

This article raises good points and informs a discussion about the need to strengthen the data collected by national cancer registries in the future to further improve these data sets. A complete revision of the Facility Oncology Registry Data Standards (FORDS) manual is overdue to keep up with changes that rapidly occur in surgical procedures, radiation oncology techniques, and systemic therapies, including biologic therapies. In the future, linkage of cancer registries to hospital electronic medical records and other data sources may allow the more seamless transfer and entry of accurate data that cancer registrars currently must enter by hand, thus providing registrars with electronic tools to make them more efficient and free them to expand the amount of data captured as well as further increase the quality. These enhancements will come with a price, however, for our national cancer registries to get even better and realize their fullest potential in the future will require significant investment. The payoff is well worth the price.

REFERENCES

1. Surveillance Epidemiology and End Results (SEER). http://seer.cancer.gov/about/overview.html. Accessed 18 Jan 2014.
2. National Cancer Data Base (NCDB). http://www.facs.org/cancer/ncdb/. Accessed 18 Jan 2014.
3. National Program of Cancer Registries (NPCR). http://www.cdc.gov/cancer/npcr/about.htm. Accessed 18 Jan 2014.
4. North American Association of Central Cancer Registries (NAACCR). http://www.naaccr.org/AboutNAACCR/NAACCRMission.aspx. Accessed 18 Jan 2014.
5. In H, Bilimoria K, Stewart AK, et al. (2014) Cancer recurrence: an important but missing variable in national cancer registries. *Ann Surg Oncol.* doi:10.1245/s10434-014-3516-x.
6. National Cancer Data Base Participant User Data File (PUF). http://www.facs.org/cancer/ncdb/participantuserfiles.html. Accessed 18 Jan 2014.