Ontario Cancer Data Linkage Project: "cd-link"

Craig C. Earle

Abstract

In the past, analysis of administrative data relevant to cancer in Ontario had to be conducted within a handful of secure physical locations, such as the Institute for Clinical Evaluative Sciences (ICES). While successful at preserving patient privacy and confidentiality, this approach was slow and expensive and was also unable to accommodate the necessary capacity for research in the province. In 2008, the Ontario Institute for Cancer Research, Cancer Care Ontario and ICES established a joint initiative to create a new cancer data release mechanism called "cd-link." Making the data more widely available brings the creativity of the broader research community to bear on what can be learned from these data, and will attract new researchers into the field of cancer health services research.

The Issue

Ontario's universal healthcare system offers several advantages for health services research. Large, truly population-based cohorts can be constructed including patients of all ages, with rich data on things such as radiation therapy or outpatient medications not usually available from other sources. Furthermore, studies of dissemination, quality of care and disparities from other jurisdictions are often confounded by insurance status, which is largely mitigated in Ontario. Ontario also has a superb infrastructure for research using administrative data. Most notably, the Institute for Clinical Evaluative Sciences (ICES) is an independent non-profit research organization, funded largely by the Ontario Ministry of Health and Long-Term Care, with a mandate to house and conduct research using administrative databases and registries in the province. To do this, ICES is one of only four organizations recognized as a "prescribed entity," a special privacy designation allowing it to access and link patient-level data. Capitalizing on all of these features can make Ontario researchers uniquely situated to answer important questions related to cancer service delivery.

In the past, analysis of administrative data relevant to cancer in Ontario, such as the Ontario Cancer Registry and Ontario Health Insurance Plan claims, had to be conducted within a handful of secure physical locations, such as ICES, and only aggregate data were released beyond the walls of these institutions. While successful at preserving patient privacy and confidentiality, this approach was slow and expensive and created barriers for researchers not affiliated with one of the specified organizations. It was also unable to accommodate the necessary capacity for research in the province. ICES recognized this problem, and over the past five years it has opened four satellite sites in Ontario, at Queen's University in Kingston, the University of Ottawa, the University of Toronto and Western University in London. There are also plans for a site at McMaster University in Hamilton and others. While this has improved data access, shifting data manipulation and analysis to investigators can further facilitate the timely and efficient conduct of this necessarily iterative form of research.

The Response

In 2008, the Ontario Institute for Cancer Research, Cancer Care Ontario and ICES established a joint initiative to create a new cancer data release mechanism called "cd-link." With this program, data sets are linked, de-identified and, with the protection of a comprehensive data use agreement, provided directly to investigators following a comprehensive review and approval process. The cd-link program is patterned after the highly successful Surveillance, Epidemiology, and End Results (SEER)–Medicare linked database program (Potosky et al. 1993) in the United States, which provides investigators with access to linked tumour registry and administrative claims data. In place for over 20 years, these data have been used by researchers to address questions of access, quality, equity, cost and outcomes of care.

The US Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (National Institutes of Health 2005) states that data can be de-identified by either (1) the "safe harbor" method of removing all of 18 specified variables or (2) applying "statistical and scientific principles and methods for rendering information not individually identifiable" and determining "that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by the anticipated recipient to identify

the subject of the information." Such data are no longer considered to be protected health information, and, as such, there are no limitations on their use. With cd-link, we have chosen to apply both safe harbor and statistical approaches to the de-identification of data.

Most of the *safe harbor approach* is fairly straightforward. Data are linked at a patient level using a secure anonymous encryption key. We remove or scramble the identifiers of patients and providers after linkage. We have established protocols that actually allow for scrambled identifiers to be created unique to each data release, removing the possibility of investigators linking up complementary data sets without permission. The main innovation in our process, however, relates to dates, which are commonly considered to be one of the most difficult data elements to remove. HIPAA considers any date that is more specific than a year to be protected health information. In the Ontario Cancer Registry, the date of diagnosis is defined as an actual date (usually the date of first pathological specimen such as a biopsy). This allows the anchoring of all dates in all data sets to that one date, which is known only within ICES. Any date, such as date of birth, could be used for the same purpose. As a result, with the data provided to an investigator through cd-link, a patient can be known to be born, for example, in 1946, diagnosed with breast cancer in 2007, hospitalized 21 days later for surgery on the day of admission, administered chemotherapy between days 42 and 166 and so on. For most research uses, this is sufficient information. Linked ecological data, such as census data, can be made available with figures rounded (e.g., median income to the nearest thousand dollars) to the point where they could not be directly mapped back to any group or area of less than 20,000 people, per HIPAA standards.

For *statistical de-identification*, the resultant data sets are evaluated with the Privacy Analytics Risk Assessment Tool. This software (www.privacyanalytics.ca) follows a risk-based approach to the de-identification of health data. It allows the data custodian to measure the risk of re-identification under different assumptions and conditions (Dankar et al. 2012; El Emam and Dankar 2008), and then uses optimization algorithms to transform the data to ensure that the risk is acceptably small while maintaining data quality (El Emam et al. 2009a). Examples of its application have been published for the de-identification of pharmacy data (El Emam et al. 2009b), discharge abstract data (El Emam et al. 2011) and claims data (El Emam et al. 2012). Variables in cd-link data sets are modified to ensure that there are at least three and preferably five patients in the data set with the same characteristics. Examples of the modifications are grouping year of birth into five-year bands, or suppressing or grouping geographical location. Such data are rendered anonymous to the point where they comply with both Canadian and American definitions of not being protected or personal health information. In this way, a new type of data are created that we call risk-reduced de-identified data, or "R2D2."

The cd-link data use agreement addresses the following issues: purpose limitation; confidentiality and prohibition of re-identification, external linkage or re-contact; data security; research ethics approval; onward transfer or sharing with third parties; cell-size suppression; prepublication review; acknowledgement; data ownership; data destruction; breach notification; responsibility to educate the research team; and the cd-link program's right to on-site inspection and audit. The cd-link program is currently available to investigators at Ontario academic institutions. Data cuts are provided free of charge, although a modest cost-recovery fee is planned for the future. The data sets are provided to investigators within a target of six weeks after receiving a complete application package. They are delivered on encrypted CDs by a fully tracked courier.

The Results

The first cd-link data release occurred in March 2010. To date, the program has received and approved more than 40 requests from academic researchers, clinician scientists and post-doctoral fellows on topics such as these: variation in the surgical management of renal tumours; healthcare settings, transitions and services used by cancer patients in the last year of life; the effect of adjuvant hormonal treatment on bone health in older breast cancer survivors; the impact of adherence to HER2 testing, treatment and monitoring guidelines in early-stage breast cancer; phase-specific and lifetime costs of cancer in Ontario; and the epidemiology and burden of illness associated with hepatocellular carcinoma. The cd-link data sets have been used for postgraduate theses and post-doctoral projects and at least two successful national grants. Applications have come from non-traditional users such as business school academics and healthcare administrators in northern Ontario.

A major goal of cd-link has been to create data sets efficiently so that cost does not pose a barrier to access. The average cost of cd-link releases has been steadily declining. Our average cost per release is under $2,500, less than 10% of the cost of comparable in-house ICES projects. To further support users, we host a website (www.cd-link.ices.on.ca), hold workshops for data users and support the time of expert ICES analysts to address questions about the data sets as they arise. For these reasons, cd-link projects now account for about half of the cancer research approved at ICES.

Lessons Learned

The public has an interest in seeing the maximal benefit from data collected with public funds. The Ontario Cancer Data Linkage Project is contributing to achieving this goal by responsibly enhancing access to Ontario's rich data resources for research, thereby facilitating timely, efficient and creative studies

addressing access, quality, cost and outcomes of care across the cancer control continuum. Making the data more widely available brings the creativity of the broader research community to bear on what can be learned from these data, and will attract new researchers into the field of cancer health services research. The cd-link concept is in the process of being expanded beyond cancer; and in the near future, collaboration will be even easier as data sets may become accessible to researchers outside Ontario through VPN (virtual private network) access. The more we are able to study our health system, the more transparent it will become. This "democratization" of research increases our ability to shine a light on problems, inform policy and, ultimately, optimize the delivery of health services and resultant outcomes for the citizens of Ontario. HQ

References

Dankar, F.K., K. El Emam, A. Neisa and T. Roffey. 2012. "Estimating the Re-identification Risk of Clinical Data Sets." *BMC Medical Informatics and Decision Making* 12: 66.

El Emam, K., D. Paton, F. Dankar and G. Koru. 2011. "De-identifying a Public Use Microdata File from the Canadian National Discharge Abstract Database." *BMC Medical Informatics and Decision Making* 11: 53.

El Emam, K. and F.K. Dankar. 2008. "Protecting Privacy Using k-Anonymity." *Journal of the American Medical Informatics Association* 15(5): 627–37.

El Emam, K., F.K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo et al. 2009a. "A Globally Optimal k-Anonymity Method for the De-identification of Health Data." *Journal of the American Medical Informatics Association* 16(5): 670–82.

El Emam, K., F.K. Dankar, R. Vaillancourt, T. Roffey and M. Lysyk. 2009b. "Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records." *Canadian Journal of Hospital Pharmacy* 62(4): 307–19.

El Emam, K., L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri et al. 2012. "De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset." *Journal of Medical Internet Research* 14(1): e33.

National Institutes of Health. 2005. *Health Services Research and the HIPAA Privacy Rule* (NIH Pub. No. 05-5308). Retrieved December 5, 2013. <http://privacyruleandresearch.nih.gov/pdf/HealthServicesResearchHIPAAPrivacyRule.pdf>.

Potosky, A.L., G.F. Riley, J.D. Lubitz, R.M. Mentnech and L.G. Kessler. 1993. "Potential for Cancer Related Health Services Research Using a Linked Medicare–Tumor Registry Database." *Medical Care* 31(8): 732–48.

About the Author

Craig Earle, MD, MSc, FRCPC, is a health services researcher and medical oncologist specializing in gastrointestinal malignancies. He is director of the health services research program for Cancer Care Ontario and the Ontario Institute for Cancer Research, a senior scientist at the Institute for Clinical Evaluative Sciences, and a professor in the Department of Medicine at the University of Toronto, in Toronto, Ontario. Dr. Earle may be reached at craig.earle@ices.on.ca.