Improving usability of smoking data in EMR systems

Stephanie Garies MPH Dave Jackson Babak Aliarzadeh MD MPH
Karim Keshavjee MD CCFP Ken Martin MSc Tyler Williamson PhD

The development of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) has created a rich longitudinal database of anonymized patient information extracted from electronic medical records (EMRs). These data are intended for surveillance and research activities, as well as clinical improvement. However, the usability of EMR data for these purposes is highly dependent on the proper coding, or standard entry, of patient information. Many elements of EMRs, such as prescribing and diagnostic information, contain high-quality data that are specific and largely coded. Risk factor data, however, tend to be of lower quality, in part because they are documented less frequently and with less specificity, but also because they are often recorded in noncoded or free-text fields.[1] For instance, addressing tobacco use is important in clinical care, as it is a preventable cause of morbidity and mortality. Clinical guidelines recommend the documentation of smoking status for all patients, and whether cessation advice has been offered to tobacco users. This does not happen regularly, and when this information is recorded, it is usually captured in a way that severely limits the ability to estimate smoking prevalence in practices or to easily identify smokers for targeted cessation advice.[2,3]

The EMR data become more usable through standardized data entry methods that record coded patient information in a disciplined manner. When data are not entered using a structured approach, coding algorithms are needed to map the data into usable formats. A data coding algorithm was developed by CPCSSN for cleaning and classifying smoking risk factor data extracted from multiple EMR systems. However, these methods should not be viewed as supplanting the need for high-quality data collection and input procedures and processes. The CPCSSN coding algorithm is only functional if data are complete and accurate.

The CPCSSN coding algorithm takes long blocks of text that describe risk factor history and extracts important information detailing patient smoking behaviour, such as status, duration, exposure, and cessation. Data were extracted from 10 different EMR systems up to September 30, 2011, and included 157569 patients aged 12 and older who had visited primary care clinics at least once in the previous year. There were 11909 distinct ways that their smoking information was recorded, with an average length of 8 words (maximum 52) in each entry. There were 52942 patients (34%) who had smoking information recorded.

After the coding algorithm was applied, 21% of those with smoking information entered were considered "current" smokers, 17% were "past" smokers, and 14% were "never" smokers. A total of 47% were ambiguously labeled "not current," meaning they were not considered current smokers, but it was unclear whether they had quit or never smoked. The algorithm was unable to classify slightly more than 1% of patients with recorded smoking information. The algorithm also calculates the lifetime cumulative cigarette exposure for each patient in pack-years. Of the 11202 patients classified as current smokers, approximately 47% had associated frequencies recorded and only 7% had durations recorded.

Other studies also demonstrated the problematic nature of smoking documentation.[2,3] Coding has a substantial effect on the usability of these data for research and surveillance, and for health care providers as part of high-quality patient care. This method demonstrates CPCSSN's ability to code data to a format that makes it usable for research and surveillance, and for practice reflection and management. However, low overall documentation rates preclude estimations of smoking prevalence or registries that would be expected to include all smokers in a practice.

Suggestions for augmenting EMR data quality and emphasizing standardized data entry procedures include arranging continuing medical education courses for physicians on various practical data-quality topics; imploring EMR vendors to create structured text fields for collecting comprehensive and specific risk factor data; and creating standardization across the different EMR systems to ensure consistent data elements that support the capture of usable smoking data.

Ms Garies is CPCSSN Research Associate and **Mr Jackson** is CPCSSN Data Manager for the Southern Alberta Primary Care Research Network in Calgary. **Dr Aliarzadeh** is Data Manager for the North Toronto Primary Care Research Network. **Dr Keshavjee** is CPCSSN Data Architect and EMR Consultant. **Mr Martin** is CPCSSN IT Manager. **Dr Williamson** is Senior Epidemiologist for CPCSSN, and Assistant Professor at Queen's University in Kingston, Ont.

Acknowledgment
Funding for this publication was provided by the Public Health Agency of Canada. The views expressed do not necessarily represent the views of the Public Health Agency of Canada.

Competing interests
None declared

References

1. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ* 2003;326(7398):1070.
2. Murray RL, Coleman T, Antoniak M, Fergus A, Britton J, Lewis SA. The potential to improve ascertainment and intervention to reduce smoking in primary care: a cross sectional survey. *BMC Health Serv Res* 2008;8:6.
3. Dhoul N, van Vlymen J, de Lusignan S. Quality of smoking data in GP computer systems in the UK. *Inform Prim Care* 2006;14(4):242-5.

La traduction en français de cet article se trouve à **www.cfp.ca** dans la table des matières du numéro de janvier 2013 à la page **e60**.

Sentinel Eye is coordinated by CPCSSN, in partnership with the CFPC, to highlight surveillance and research initiatives related to chronic illness prevalence and management in Canada. Please send questions or comments to **Anita Lambert Lanning**, CPCSSN Project Manager, at **all@cfpc.ca**.