The Importance of Data Cleaning when Reporting Statistics from Electronic Medical Record Data

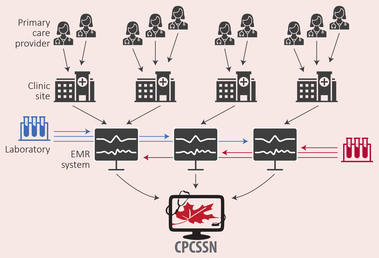
David Parker, Julia Langton, FeiFei Wang, Alexander Singer, Sabrina Wong
davidp@cpcssn.org | www.spor-bcphcrn.ca

CPCSSN RCSSSP
Canadian Primary Care Sentinel Surveillance Network
Réseau canadien de surveillance sentinelle en soins primaires

Background

The use of electronic medical record (EMR) systems continues to rise among Canadian primary care providers, but differences in data formats between EMR systems and coding practices between clinicians and laboratories can reduce the utility of these data for research and evaluation. To improve the usability and reliability of EMR data, the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), which extracts data from 11 different EMR systems used by 1,200 sentinel primary care providers, applies standardization protocols to data elements.

How CPCSSN works



Methods

We extracted EMR data from CPCSSN for patients with encounter records from June 1, 2005 to May 31, 2015 in British Columbia.

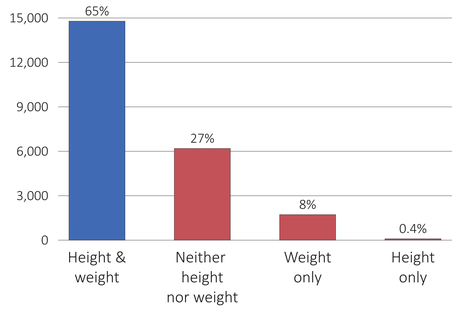
To demonstrate the value of data standardization on clinical, prescription, and behavioural data we:

1. Counted the number of patients for whom height and weight is recorded.
2. Compared (a) the number of test results for ten lab tests of interest and (b) the number of non-selective monoamine reuptake inhibitor (MRI) antidepressant prescriptions before and after the application of standardization protocols.
3. Assessed our ability to identify smokers.

Results

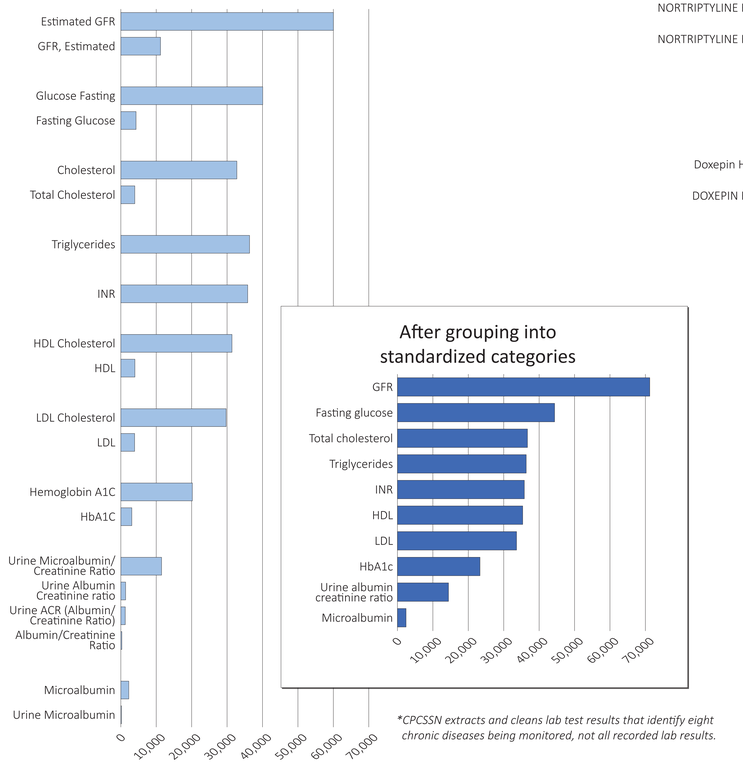
We extracted records for 22,790 patients, representing 28 sentinel primary care providers at eight sites in BC using three EMR systems (Wolf, Med Access, and OSCAR).

Number of patients with and without height and weight measurements, 2015



Numbers of CPCSSN-collected* lab tests, before and after data cleaning, 2005-2015

Before grouping into standardized categories

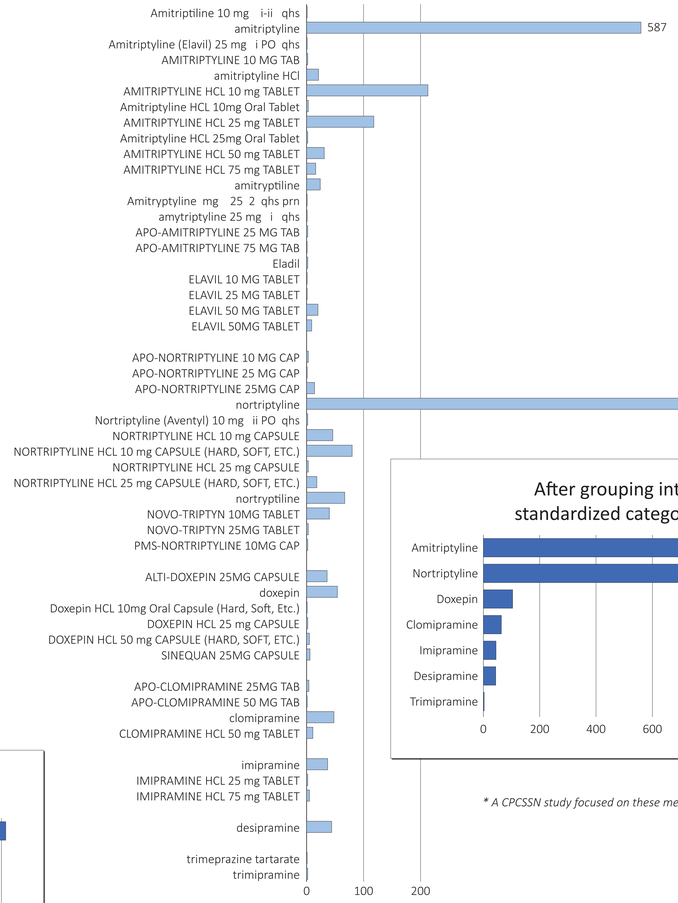


After grouping into standardized categories

*CPCSSN extracts and cleans lab test results that identify eight chronic diseases being monitored, not all recorded lab results.

Numbers of non-selective MRI antidepressant* prescriptions, before and after data cleaning, 2005-2015

Before grouping into standardized categories

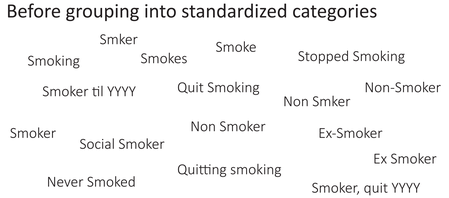


After grouping into standardized categories

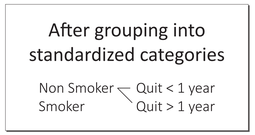
*A CPCSSN study focused on these medications (ATC code: N06AA).

Categorizing smoking status

Before grouping into standardized categories



After grouping into standardized categories



12,738 (65%) of patients had both height and weight recorded. 6,189 (27%) had neither measurement, and 1,818 (8%) were missing either height or weight.

We identified 333,598 individual test results for the ten tests before standardizing test names. 19 unique lab test names (range per test name: 178 to 60,029) matched the ten tests. After cleaning, the ten tests have 2,426 to 71,237 results per test.

We identified 2,379 prescriptions for non-selective MRIs with 51 unique drug names (range of prescriptions per drug: 1 to 783). After standardizing drug names, there are only only seven unique drugs (range of prescriptions per drug: 3 to 1,100).

Smoking status is often recorded as free text, with commonly used words overlapping in meaning. Cleanly delineated categories are difficult to create.

Conclusion

Variation and inconsistency in naming conventions used by clinicians, laboratories, and EMR systems mean that much of the 'raw' EMR data may not produce reliable or valid information. In many cases, the accuracy of the data can be improved by applying standardization protocols, such as those applied to the lab and medication data. Protocols can be applied across CPCSSN to enable cross-provincial comparisons of measures of primary health care. These data can then be used in more complex algorithms, such as those identifying medical complexity and frailty, which could be of great interest to primary health care providers.

Acknowledgements

Icons created by Ralf Schmitzer, Josy Dom Alexis, Souvik Bhattacharjee, and Hea Poh Lin from the Noun Project. This work is funded by:

MICHAEL SMITH FOUNDATION FOR HEALTH RESEARCH
Discover. Connect. Engage.

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherch en santé du Canada

SPOR SRAP
Strategy for Patient-Oriented Research
Stratégie de recherche axée sur le patient
"Putting Patients First"
Le patient d'abord