cmajOPEN

Research

Using machine learning to standardize medication records in a pan-Canadian electronic medical record database: a data-driven algorithm study focused on antibiotics prescribed in primary care

Stephanie Garies PhD, Matt Taylor, Boglarka Soos MMath, Cliff Lindeman PhD, Neil Drummond PhD, Anh Pham PhD, Zhi Aponte-Hao MSc, Tyler Williamson PhD

Abstract

Background: Most antibiotics dispensed by community pharmacies in Canada are prescribed by family physicians, but using the prescribing information contained within primary care electronic medical records (EMRs) for secondary purposes can be challenging owing to variable data quality. We used antibiotic medications as an exemplar to validate a machine-learning approach for cleaning and coding medication data in a pan-Canadian primary care EMR database.

Methods: The Canadian Primary Care Sentinel Surveillance Network database contained an estimated 42 million medication records, which we mapped to an Anatomic Therapeutic Chemical (ATC) code by applying a semisupervised classification model developed using reference standard labels derived from the Health Canada Drug Product Database. We validated the resulting ATC codes in a subset of antibiotic records (16 119 unique strings) to determine whether the algorithm correctly classified the medication according to manual review of the original medication record.

Results: In the antibiotic subset, the algorithm showed high validity (sensitivity 99.5%, specificity 92.4%, positive predictive value 98.6%, negative predictive value 97.0%) in classifying whether the medication was an antibiotic.

Interpretation: Our machine-learning algorithm classified unstructured antibiotic medication data from primary care with a high degree of accuracy. Access to cleaned EMR data can support important secondary uses, including community-based antibiotic prescribing surveillance and practice improvement.

High-quality medical record data are essential for their primary purpose of supporting clinical care, as well as for important secondary uses such as clinical quality-improvement studies, disease surveillance and research. Pharmacoepidemiologic research traditionally has relied on administrative databases such as Canada's National Prescription Drug Utilization Information System[1] or provincial or regional sources such as the Pharmaceutical Information Network in Alberta.[2] Although these databases have the advantage of covering nearly entire populations, they contain information for dispensed medications only, and not prescriptions that were issued but not filled. Dispensing data can also be challenging to use, as they may be constrained by different types of drug coverage for various age groups across time and jurisdictions. With an estimated 85% of family physicians reporting use of an electronic medical record (EMR) system in their practice,[3] the data contained in primary care EMRs provide a novel opportunity to better understand prescribing patterns, health care delivery and patient care in the community, where the majority of encounters with the health care system occur.[4] However, the value of EMR data for any purpose, especially when derived from multiple providers, systems or jurisdictions, can be limited, as their quality is highly variable, and advanced processing and data standardization are usually required before analysis can take place.[5]

Machine learning is a novel way to accurately standardize prescribing records from large primary care EMR data sets. The objective of this study was to build and validate a machine-learning tool that would code medication text from

Competing interests: None declared.

This article has been peer reviewed.

Correspondence to: Tyler Williamson, tyler.williamson@ucalgary.ca

***CMAJ Open* 2023 October 31. DOI:10.9778/cmajo.20220235**

Canadian primary care EMR data to codes of the Anatomical Therapeutic Chemical (ATC) classification, a well-established medication classification system used internationally.[6] We selected a subset of coded medication records relating to antibiotics for manual validation. Our rationale for focusing on enhancing antibiotic medication data was threefold. First, antibiotic use and stewardship is an important and timely topic that requires high-quality data in order to be studied effectively. Second, given that nearly two-thirds of all antibiotics dispensed by community pharmacies in Canada are prescribed by family physicians,[7] this represents a key sector where we can focus practice-improvement reporting and initiatives. Third, antibiotic medications are one of the more complex categories within the ATC system; this is because antibiotics can span multiple target systems and include many combination products, making the coding process less straightforward and prone to error.

Primary care EMR data are the only source of information about prescribed medications (not just what was dispensed in a pharmacy), which is a critical gap in our understanding of antibiotic prescribing patterns and constitutes a necessary component of a robust antibiotic surveillance and research system for Canada.[8] We aimed to provide a fundamental first step toward advancing antibiotic prescribing surveillance and research in Canada by improving the quality of prescribed medication data available.

Methods

Data source

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) was established in 2008 with a mandate to build a repository of de-identified primary care EMR data available for research, surveillance and quality improvement.[4,9] The network currently extracts EMR data from more than 1400 sentinel family practices across Canada, including longitudinal information from nearly 2 million Canadians. The CPCSSN data included in this study were collected in 7 provinces: British Columbia, Alberta, Manitoba, Ontario, Quebec, Nova Scotia, and Newfoundland and Labrador (further information about the number of sites and types of EMR systems are provided in Appendix 1, Supplementary Table S1, available at available at www.cmajopen.ca/content/11/5/E1020/suppl/DC1). The data for this study were extracted from 12 different EMR systems on June 30, 2020, as part of CPCSSN's routine biannual extractions. Included were patient demographic characteristics, physical measurements (e.g., height, weight, blood pressure), prescribed medications, symptoms and diagnoses recorded during patient visits, billing claims and laboratory results.[10]

Since EMRs are designed for direct patient care and not for secondary purposes, CPCSSN has developed an extensive suite of cleaning and processing algorithms to convert raw EMR data into coded information and to apply disease case definitions to the database. These processes have been described elsewhere.[5,10,11] Briefly, they include algorithms that map text strings to standard classification and terminology codes, including those of the *International Classification of Diseases, Ninth Revision* for diagnoses and Logical Observation Identifier Names and Codes for laboratory results, in addition to general data cleaning (e.g., standardizing dates) and trimming values (e.g., blood pressure, height, weight) to fit within plausible ranges. Medication data were previously standardized to the ATC system[6] by means of a simple pattern-matching approach that assigneed ATC codes to medication names found in the original EMR text, creating a new "CPCSSN-coded" field of ATC codes and names. Although this approach was seemingly effective, it is a cumbersome process that is expensive to both implement and update regularly; for instance, new medications would need to be added manually to the pattern-matching algorithm code.

In the present study, we used all medication data in the CPCSSN database (about 42 million records containing 2.4 million unique medication names), which included records dating from 1981 to 2020, although most medication records (95%) were from 2008 onward. In general, most EMR systems offer a drop-down box or user-assisted auto-complete functions for entering medication information, resulting in relatively structured data. Many systems also provide a user override mechanism to support nonstructured data entry.

Stage 1: Map all medication data to Anatomical Therapeutic Chemical codes

For the machine learning–based classifier developed for this study, we used the Health Canada Drug Product Database (DPD)[12] as the reference standard to which all medications in the CPCSSN database were mapped. The DPD, which is updated nightly, contains all drug products approved for human use in Canada and is available in comma-separated-values format from the Health Canada website.[12]

We used the fastText open-source library version 0.9.2 (Facebook) to develop the machine-learning model for cleaning and coding the unstructured medication text in the CPCSSN database.[13] We considered other, similar models, such as scikit-learn,[14] but fastText was most suited for the CPCSSN database in that it is fast and efficient when standard computing infrastructure is used, it is able to handle new (or "out of vocabulary") words, and it can leverage morphemes such as prefixes and suffixes when training word-embedding models. Word embeddings are a useful technique in language modelling that can represent words from a document in a machine-readable (i.e., numeric) way, while also capturing the context of those words in relation to other words.

We used a large corpus (n = 2 419 786) of uncoded medication name text selected from the CPCSSN database to include any and all references to antibiotic medications to train a skip-gram word representation model. Skip-gram models are a type of word embeddings that use an unsupervised or semisupervised learning approach to predict the context (i.e., surrounding) words given a target (i.e., input) word. We then used the skip-gram model to build a semisupervised classification model using multinomial logistic regression, with labels derived from the DPD. The labelled data set used

for training and validation (n = 151 296 records) included values sourced from the DPD, such as brand and generic names, in addition to medication names in the CPCSSN database that we had previously coded using simple pattern and prefix matching from the DPD. We iteratively refined the model through 5 rounds of review. At each round, the team was engaged to evaluate whether the model was appropriate and to review sources of potential disagreement between the classification from the model and the reference standard (DPD).

Stage 2: Validation with a subset of antibiotic records

We drew all 16 119 unique medication names present in any antibiotic category from the CPCSSN data; this accounted for 159 unique ATC codes for antibiotics. A reviewer with a background in public health and nursing was trained on the organization and function of the ATC coding system and, after several practice sessions with the study team, was asked to manually indicate in an Excel (Microsoft) spreadsheet whether the algorithm (once applied to the data set) had produced the correct ATC code, given the medication name, strength, dosage, frequency and route of administration information available in each of the 16 119 medication records from the raw (unprocessed) EMR data. In cases in which the reviewer was unsure whether the match was correct, a consensus was sought with 2 other study team members (M.T., S.G.). In cases in which the reviewer was unsure whether the match was correct, a secondary review was conducted with 2 other study team members (M.T., S.G.) to achieve consensus on whether the algorithm correctly coded the medication text as an antibiotic or not an antibiotic. We calculated sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) from the binary comparison of the algorithm-derived ATC code to the reviewer's verification (i.e., whether the ATC code correctly identified an antibiotic medication).

We used Python 3.9 (Python Software Foundation) for text processing. We generated the confusion matrix with PostgreSQL 11 (PostgreSQL Global Development Group), an-open source relational database system hosted on physical servers at Queen's University, Kingston, Ontario.

Ethics approval

This study received approval from the University of Calgary Conjoint Health Research Ethics Board (REB20-1316).

Results

The validity of the machine-learning algorithm compared to manual review of each of the antibiotic medication records (i.e., whether the correct ATC code was assigned to the record) is shown in Table 1. Overall, the algorithm performed very well, with a sensitivity of 99.5%, specificity of 92.4%, PPV of 98.6% and NPV of 97.0%. In total, 270 records were inaccurately classified (false-positive and false-negative results). In a simple post hoc analysis of these 270 records, we found that some did not refer to antibiotics, some were identified as an antibiotic but were misclassified as the incorrect type of antibiotic, and some were the wrong medication entirely. For example, hydrocortisone in the original (raw) EMR data was sometimes coded as an antibiotic because the algorithm learned to code combination medications that included both an antibiotic and hydrocortisone. Another false-positive issue (although much rarer) was when the algorithm coded certain abbreviations in the raw EMR data, such as T1D (type 1 diabetes) and T3 (Tylenol 3 [acetaminophen–codeine, Janssen Pharmaceuticals]), as the topical antibiotic acne treatment T-Stat (erythromycin, Bristol-Myers Squibb Canada).

Interpretation

This study presents a new method for standardizing medication records in a Canadian primary care EMR database, with validation conducted on the subset of antibiotic records. Other medication-coding tools have been developed with the use of similar artificial intelligence methods applied to EMR databases, primarily from the United States and often trained on hospital or specialist data.[15] The high sensitivity (99.5%) and specificity (92.4%) of our algorithm is comparable to those of many of these medication classifiers, including the MedXN (Medication Extraction and Normalization) system

Table 1: Confusion matrix comparing the machine-learning algorithm to the reviewer confirmation of Anatomic Therapeutic Chemical–coded antibiotic medications in the Canadian Primary Care Sentinel Surveillance Network database

	Human review; no. of records			
CPCSSN algorithm	Yes (antibiotic)	No (not an antibiotic)		
Yes (antibiotic)	13 451	197	PPV 98.6%	True positive 13 648
No (not an antibiotic)	73	2398	NPV 97.0%	True negative 2471
	Sensitivity 99.5%	Specificity 92.4%		

Note: CPCSSN = Canadian Primary Care Sentinel Surveillance Network, NPV = negative predictive value, PPV = positive predictive value.

(F-measure 0.9 for dosage and 0.84 for frequency), Merki (0.94 precision, 0.825 recall), MTERMS (0.9 precision and recall) and MedEx (95% precision, 92% recall), although these vary by setting, data source, types of medications and natural language–processing method.[15,16]

The output of this algorithm (i.e., mapping the correct ATC code to a medication record) was intended primarily to support more robust secondary analysis of medication data within the CPCSSN database. Although the high PPV also suggests potential for use in clinical practices, the ideal scenario would be integration of this algorithm within clinical information systems to facilitate better coding functionalities in the backend, which could then be used by clinical staff for easier medication searching by ATC code and, subsequently, would help to enhance practice-improvement initiatives.

We focused the evaluation component on antibiotic prescribing, as it is a common activity in primary care and is also one of the more complex types of medication records in that there are several drug classes and multiple routes of administration (e.g., creams, pills), and these drugs are often compounded with other medications. Although the algorithm showed overall high PPV and sensitivity, one of the more challenging areas was in mapping correct ATC codes to combination medications. There is much less consistency in how these types of medications are documented in the EMR, owing to the unique customization that can occur with drug compounding.

In the future, we aim to conduct a manual validation of additional medication classes. Since machine learning is highly dependent on the training data used, our access to a large pan-Canadian data source reinforces confidence that our machine-learning algorithm will be able to accurately classify additional medication data from various regions and different EMR systems. With the use of the DPD as the reference standard, this ensures that newly approved medications will be immediately included in our classifier in the future. The machine-learning method described here will also be useful in expanding our standardization processes to include other fields in the CPCSSN database, such as laboratory values and diagnoses.

The use of coded EMR data from primary care practices provides a novel and efficient way to conduct antibiotic prescribing surveillance in Canada. In the future, we may link regional CPCSSN data with health administrative data, such as the Pharmaceutical Information Network, to create more robust data sets that capture a more complete trajectory of diagnoses, prescriptions and dispensed medications. This work will also inform the development of machine-learning methods to code and classify the rest of the prescribed medication data in CPCSSN, as well as other types of data in EMRs such as diagnoses, medical procedures and referrals. The potential of this approach to improve EMR data quality for almost every secondary purpose is clearly substantial.

Limitations

The CPCSSN is not without limitations: it does not include all practices, providers or patients; rather, it is a sample of providers willing to contribute de-identified EMR data for surveillance and research. Generally, the CPCSSN database is reasonably representative of patients (with slight overrepresentation of females and older adults) and providers (who are more often younger, female and in an academic practice).[17] Although the validation was conducted on 1 medication class only (antibiotics), this is a more difficult set of medication records to classify with an ATC code and, thus, more prone to errors. However, our model produced statistics with high validity, which lends some confidence that this method will be accurate for other types of medication classes. Our machine-learning algorithm may not be directly portable to other data sources or settings, or in practices with low antibiotic prescribing, which would likely reduce the PPV. However, given that 12 different primary care EMR systems from multiple provincial contexts were included in the CPCSSN database, we are relatively confident in the robustness of our model. Last, the machine-learning model classifies only prescribed medications listed in the DPD. There may be other, nonprescription medications in the medication table, as well as notes unrelated to medications (e.g., use of compression stockings, massage therapy recommendation), that would not be classified with this approach.

Conclusion

Our machine-learning algorithm classified unstructured antibiotic medication data from primary care with a high degree of accuracy. When applied to the national CPCSSN database, the use of this algorithm will help to provide more robust data for pharmacoepidemiologic research and clinical quality improvement, and will be transferrable to other conditions and other data in the record.

References

1. National Prescription Drug Utilization Information System metadata. Ottawa: Canadian Institute for Health Information. Available: https://www.cihi.ca/en/national-prescription-drug-utilization-information-system-metadata (accessed 2022 Apr. 4).
2. Overview of administrative health datasets. Edmonton: Analytics and Performance Reporting Branch, Alberta Health; 2017. Available: https://open.alberta.ca/dataset/657ed26d-eb2c-4432-b9cb-0ca2158f165d/resource/38f47433-b33d-4d1e-b959-df312e9d9855/download/research-health-datasets.pdf (accessed 2022 Apr. 4).
3. CMA Workforce Study, 2017: national results by FP/GP or other specialist, gender, age, and province/territory. Ottawa: Canadian Medical Association; 2017. Available: https://surveys.cma.ca/viewer?file=%2Fmedia%2FSurveyPDF%2FCMA_Survey_Workforce2017_Q22_ElectronicRecords-e.pdf#page=1 (accessed 2018 Oct. 31).
4. Birtwhistle R, Williamson T. Primary care electronic medical records: a new data source for research in Canada. *CMAJ* 2015;187:239-40.
5. Garies S, Cummings M, Forst B, et al. Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction, and processing in Alberta. *Int J Popul Data Sci* 2019; 4:1132.
6. ATC/DDD Index 2023. Geneva: WHO Collaborating Centre for Drug Statistics Methodology, World Health Organization. Available: https://www.whocc.no/atc_ddd_index/ (accessed 2017 Nov. 30).
7. Canadian Antimicrobial Resistance Surveillance System: 2017 report. Ottawa: Public Health Agency of Canada; 2017, modified 2018 May 24. Available: https://www.canada.ca/en/public-health/services/publications/drugs-health-products/canadian-antimicrobial-resistance-surveillance-system-2017-report-executive-summary.html (accessed 2022 Jan. 28).
8. Handle with care: preserving antibiotics now and into the future — Chief Public Health Officer of Canada's 2019 spotlight report. Ottawa: Public Health Agency of Canada; 2019. Available: https://www.canada.ca/content/dam/phac-aspc/documents/corporate/publications/chief-public-health-officer-reports-state-public-health-canada/preserving-antibiotics/Final_CPHO_Report_EN_June6_2019.pdf (accessed 2023 Mar. 22).

9. Birtwhistle R, Keshavjee K, Lambert-Lanning A, et al. Building a pan-Canadian primary care sentinel surveillance network: initial development and moving forward. *J Am Board Fam Med* 2009;22:412-22.
10. Garies S, Birtwhistle R, Drummond N, et al. Data resource profile: national electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). *Int J Epidemiol* 2017;46:1091-2f.
11. Williamson T, Green ME, Birtwhistle R, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med* 2014;12:367-72.
12. Drug Product Database: access the database. Ottawa: Health Canada; updated 2015 June 18. Available: https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/drug-product-database.html (accessed 2022 Jan. 28).
13. Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*; 2017 Apr. 3–7, Valencia, Spain. Vol. 2, Short Papers. Stroudsburg (PA): The Association for Computational Linguistics; 2017:427-31.
14. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825-30.
15. Kreimeyer K, Foster M, Pandey A, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017;73:14-29.
16. Xu H, Stenner SP, Doan S, et al. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;17:19-24.
17. Queenan JA, Williamson T, Khan S, et al. Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study. *CMAJ Open* 2016;4:E28-32.

Affiliations: Department of Family Medicine (Garies, Soos, Drummond), University of Calgary, Calgary, Alta.; Departments of Family Medicine (Taylor, Lindeman, Drummond) and Physical Therapy (Lindeman, Pham), University of Alberta, Edmonton, Alta.; Department of Community Health Sciences (Drummond, Aponte-Hao, Williamson) and Centre for Health Informatics (Williamson), University of Calgary, Calgary, Alta.

Contributors: Tyler Williamson, Stephanie Garies and Matt Taylor conceived the study. Tyler Williamson, Stephanie Garies and Matt Taylor designed the study, with contributions from Boglarka Soos, Cliff Lindeman, Neil Drummond, Anh Pham and Zhi Aponte-Hao. Matt Taylor analyzed the data. Tyler Williamson, Stephanie Garies and Matt Taylor interpreted the data. Stephanie Garies drafted the manuscript. Tyler Williamson, Matt Taylor, Boglarka Soos, Cliff Lindeman, Neil Drummond, Anh Pham and Zhi Aponte-Hao revised the manuscript critically for important intellectual content. All of the authors approved the final version to be published and agreed to be accountable for all aspects of the work.

Funding: This study received funding from the Alberta Children's Hospital Research Institute Health Outcomes theme (Collaborative Research Initiative Grant).

Data sharing: The Canadian Primary Care Sentinel Surveillance Network database is accessible to researchers for approved uses. To submit a data access request or find out more about the data access fees, data dictionary and other information, visit http://cpcssn.ca/dar/.

Acknowledgements: The authors thank Ms. Sara Alvarado for her assistance with the machine-learning validation process. They are grateful to all primary care providers and patients who contribute to the CPCSSN database across Canada.

Disclaimer: The funder had no role in the design or conduct of this study.

Supplemental information: For reviewer comments and the original submission of this manuscript, please see www.cmajopen.ca/content/11/5/E1020/suppl/DC1.