Article details: 2022-0235
Title: Development, validation and application of a machine learning algorithm to standardize antibiotic prescribing records in a pan-Canadian primary care EMR database: describing patterns of pediatric antibiotic prescribing
Authors: Stephanie Garies PhD, Matt Taylor, Boglarka Soos Mmath, Cliff Lindeman PhD, Neil Drummond PhD, Anh Pham PhD, Zhi Aponte-Hao MSc, Tyler Williamson PhD

Reviewer 1: Dr. Kevin Schwartz
Institution: Public Health Ontario, University of Toronto Dalla Lana School of Public Health
General comments (author response in bold)

1. I did not see that the authors made the developed code publicly available. The described concept is clearly important but would be of value to readers if the authors could include their programing/code within the supplement (or online) that other researchers could use and apply to their own EMRs. This would make this of much greater value to the community.
The code was developed specifically to be applied to the CPCSSN database, which will enable data requestors to access better quality data. The model described here may be less useful or not applicable for other general or raw EMR data, as CPCSSN contains specific common data elements that have been extracted and transformed prior to the application of the medication coding algorithm.

2. I think it is a strength of the manuscript that the authors demonstrated the potential application of the AI code to a research question. I am a bit unclear on the rationale for the denominator in Figure 2. Since the authors have all EMR data it would allow more comparison to the literature if you used a metric such as antibiotic prescriptions per patient population or patient visits. Since antibiotics make up 15-35% of pediatric prescriptions the denominator is quite susceptible to changes in the numerator. Perhaps the authors can provide some more justification and/or rationale why they chose this denominator and consider adding a population denominator as well.
We have decided to remove the pediatric antibiotic descriptive section and focus on the details of the machine learning classifier.

3. The pediatric antibiotic analysis does not really add novel data and just describes population trends with an atypical metric. It is perhaps beyond the scope of this project which was primarily to design and validate the algorithm, however a more in depth or novel research question related to pediatric antibiotic use would strengthen the manuscript
We agree and have chosen to omit the pediatric antibiotic prescribing section, which we will consider presenting in a separate publication.

4. I am curious if the authors' algorithm only identifies an antibiotic prescription (yes/no) or can it also identify the dose and duration information from the prescription? This could add value in terms of other prescribing metrics (eg; Days of therapy or duration of therapy).
The algorithm identifies the ATC code associated with the information within the medication record (e.g. name, dose, frequency, route, etc.). For each ATC code within the ATC system, it provides an estimate of the Defined Daily Dose (https://www.whocc.no/ddd/definition_and_general_considera/), which may be

useful when using the ATC-coded data for research or surveillance. The medication record itself within the CPCSSN database will often contain dose information, and duration information less so.

Reviewer 2: Dr. Balthasar Hug
Institution: Luzerner Kantonsspital
General comments (author response in bold)

3. Methodology

3.1 The data source "The Canadian Primary Care Sentinel Surveillance Network (CPCSSN)" is described on p. 2-3. It started in 2008, but older medical records since 1981 were also included (p.3, lines 54-55). The data source is impressive and works on a national level. According to the authors, 95% of their data stems from 2008 onwards (2008-2020= 12 years) and 5% from the 27-year period from 1981-2008. Did the authors use weights in their ML model to correct for this difference? How did that affect their model? And how do they explain this difference? And why did they include the former period knowing that there was so much less medication data available? Why not just use the latter period and going for a better data consistency?
The model took in all distinct values that were present in the entire CPCSSN database. It was not based on frequency, so weights were not relevant here. We also required the classifier to work on both old medication data and newer records, as the CPCSSN database is used for longitudinal research and may require use of historical medications.

3.2 Machine learning model (p.3-4, lines 63-77): We learn from the authors, that they used "The FastText open-source library (version 0.9.2)" to develop their ML model. Did they use Python for this procedure? The word "Python" does not appear in the paper. If used, please declare this including the version of Python.
Python 3.9 was used for text processing, and we have added this to the Methods.

3.3 Have the authors thought about or tried other algorithms than FastText? If yes, which ones, if no, why not? Was there a discussion about this? If yes, that would be interesting for the reader and needs a description in the paper. Mapping of "unstructured information" (p.5, lines 118-9) with NLP programs in EMRs still is a great challenge. Therefore, this program selection process of these authors are of great interest. There are many publications on this topic, e.g., Afshar, M., et al., (Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study. The Lancet Digital Health, 2022. 4(6): p. e426-e435.).
CPCSSN's original medication coding algorithm used the TF-IDF algorithm and worked passably well, although was labour-intensive and not efficient for frequent medication updates. Some preliminary exploration with scikit-learn-based classifiers was carried out but were unsuitable for the CPCSSN database for reasons of speed, memory, and/or model size. We also included a brief summary of a review describing NLP approaches for medication classification in the Interpretation.

3.4 What is the validity of FastText? The authors mention solely ref #15 by Joulin et al. This paper mentions in the title, that it is about "Development…of a ML algorithm", but we don't read anything about this development, just that the authors chose this FastText

algorithm. There should be at least a paragraph on the explanation about why FastText is better than other natural language processing programs, its prediction capacity, pros, and cons.
Describing the validity of FastText itself is a bit outside the scope of the paper – it's a commonly used approach for machine learning and met our requirements for a fast, small, easy-to-use model that could run efficiently on standard computing infrastructure.

3.5 On p.4, line 96 the authors mention that they used PostGreSQL 11 for descriptive analysis of their data. They should explain this cloud service by Amazon, since the common reader of CMAJ will not be a clinical informatics specialist, but a clinician.
PostgreSQL is an open source object-relational database system built on the SQL language, https://www.postgresql.org. While there are examples of cloud-based services making use of PostgreSQL platform (e.g. AWS, Google GCP and Microsoft Azure), for privacy reasons, CPCSSN uses our own physical data centre at Queen's University (so it is not cloud based).

3.6 What type of software did the authors use for the confusion matrix in Figure 1? Python? R? Amazon cloud services? Please describe in the methods section and as footnote in Figure 1.
We added to the Methods that the confusion matrix was produced using PostgreSQL.

4. Results

4.1 The authors offer two paragraphs as results section (p.5, lines 102-114) referring to Figure 1 and Table 1. If we go back to p. 2 and have a look at the objectives of this study, the results section should be much more detailed and encompassing:

4.2 How was the ML model built and why was FastText chosen? See also above in the methods review.
We have addressed this in our response in 3.3 and in the Methods section of the manuscript.

4.3 How about the false positives and false negative words that the algorithm selected? It would be illustrating to see a few of these as examples.
We added a few examples of false positives in the Results section (page 5-6). These included: 1) 'hydrocortisone' from the raw medication data was sometimes coded as an antibiotic, because the algorithm had learned to code other hydrocortisone combinations with antibiotics as an antibiotic record) and the rare abbreviation (e.g. 'T1D', referring to type 1 diabetes, was found in the raw medication data and was erroneously coded as 'T-Stat', a topical acne treatment containing erythromycin).

4.4 The validation part is well described and easily understandable.

4.5 The "patterns of antibiotic prescribing" should be described in detail. What about the seasonality to be seen in Figure 2? This is discussed in the "Interpretation" section but not mentioned at all in the results section.

We have chosen to omit the pediatric antibiotic prescribing section, in order to spend more word count on the details of the classification algorithm and will consider submitting this as a separate paper.

4.6 As readers we would expect a table about the antibiotics used and how often. In Table 1 there are five lines about these results at the bottom. This is not enough, and the ATC-code allows to dig systematically deeper into the results of these authors. E.g., ATC J01 has ten sub-codes that could be described, see also below.
As above, we have chosen to omit the pediatric antibiotic descriptive analysis.

4.7 There should be a results section about the different application forms of the antibiotics. The ATC-code the authors use does make this distinction e.g., code "J" for "Anti-infectives for systemic use" (https://www.whocc.no/atc_ddd_index/?code=J&showdescription=no). As an example, gentamicin appears about five times in supplementary table 1, which probably has to do with mixtures, ointments etc. as described by the authors on p.5, lines 121-4. The mentioned table needs proper footnotes explaining this. These application forms analyses have also clinical consequences; whether a substance like gentamicin is applied in eye drops, as an ointment or systemically by iv route has not the same implications for antibiotic stewardship.
As above, the prescribing results table has been removed.

5. Interpretation, limitations, and conclusions: well written and should be expanded according to the above-mentioned adaptions to the manuscript.

Reviewer 3: Dr. Lisa Cook
Institution: Alberta Health Services, University of Lethbridge Faculty of Health Sciences
General comments (author response in bold)

My only two suggestions are to provide the number of sentinel sites from each of seven provinces and provide a list of 12 different EMR systems that were used in the analysis. I am curious as to whether there was a difference between the provinces, however that is separate question. Well done, this is a nicely designed study and a great written paper. I think the data the authors will generate from this work will be informative to the Canadian primary care system.
Thank you for this comment – we have added information about EMR system and number of sites in each of the regions/provinces as a supplementary table. We did not evaluate the algorithm by province, as it was trained to function on the national CPCSSN dataset. Any possible sources of variation would likely be attributed to differences in EMR system, in how each system captures medication information (but again, we needed the algorithm to be applicable and accurate for the entire dataset).

Reviewer 4: Dr. Wilson Pace
Institution: University of Colorado, DARTNet Institute
General comments (author response in bold)

1. This manuscript uses an open access text matching software to codify drug data in Canadian primary care EHRs. The FastText system can be trained rapidly from large nomenclatures and other standardized textual data files. The training process using the Health Canada Drug Product Database was based on an AI model and not hand

training. The advantage of this approach is that it can include a much larger domain of concepts with rapid learning, assuming the model proves reasonably accurate. It appears from the methods that the matching algorithm attempted to capture the medication name, strength, dose, frequency, and route. The classification of the non-standard text, such as dose and frequency, which is often free text in EHRs would be a major improvement for many groups involved in transforming native EHR data. The hand validation of all matches is an impressive amount of work. Unfortunately, this brief report currently leaves many unanswered questions related to the work reported upon.

2. It is unclear what components of the prescription are being codified. In the validation section of the Methods, it appears that the entire body of the prescription is being interpreted (as noted above) but in the results section only drug name matches are noted. The Validation was performed based on ATC codes, but the hierarchical level of ATC code used is not noted. Further, ATC nomenclature includes the Defined Daily Dose, which is the typical dose for the medication, but it does not include variable dosage information. Thus, the 159 codes would not appear to be able to classify all of the information concerning drug, dose, route, and frequency across the 16,116 different unique antibiotic prescriptions used to match on. Thus, it is not clear what the matching as shown in Figure 1 actually is reporting upon. This needs to be clarified. If the match is occurring across multiple variables, then it may be reasonable to report on the metrics related to correct antibiotic only and then separately for the full match. It appears that the text used for testing was all antibiotic names.
The medication string is being mapped to its corresponding ATC code, and the evaluation was determining whether the algorithm correctly identified the ATC code that matched with the information in the medication string/record. Many of the medication strings in the CPCSSN database have other information embedded within, dose being especially common. We have added clarification in the Methods and Results.

3. Why there would be 2595 “negative” results or prescriptions that should not be matched to an antibiotic is unclear. How the system would perform against native EHR data which includes a full set of drug names, of which antibiotics would make up about 0.5% is not clear. But to try advance this work the system needs to be useful without prior processing of the source data. Correctly matching route is critical for antibiotic related research as many antibiotics have multiple routes of delivery and most would not be particularly relevant to the question being asked within this report. Alternative routes likely make up a small component of overall use but will differentially impact selected drugs. It is not clear what, if any, adjustments, or further inputs were provided between the rounds of training. It is not clear how the ability of the FastText system to handle morphemes came into play in this work since antibiotic names don’t typically appear with variations within a word for tense, gender, or part of speech, which is the primary use of morpheme-based text matching/natural language processing. Also, since text matching appears necessary for drug data in the CPCSNN EHR data, one presumes the drug names are hand entered. If this is the case, then misspellings are common. There is no mention of handling common misspellings of antibiotics.
We have added some clarification in the Methods. FastText is able to handle spelling errors fairly well. However, most of the medication entries in the EMR are supported by user interface assistance, like autocomplete or drop-down boxes (although this usually can be overridden), meaning that generally there are few spelling errors.

4. The Results are brief. It is unclear what the difference in between “some did not refer to antibiotics” (lines 107 and 108) and “some were the wrong medication entirely” (lines 108 and 108), particularly since misclassified antibiotics are not in either of these groups. The authors do not mention if all routes of antibiotic use were included in Figure 2. For many medication studies route of administration is critical and for many medications routes can be complex, for instance steroid medications have a very wide variety of routes of use which can confound work around a specific disease state. Furthermore, it is not entirely clear to what degree the FastText provided additional useful information for the simple text matching previously utilized. Given that medication data is typically highly codified within EHRs, particularly if prescribed out of the EHR, understanding the full complement of data elements addressed by the FastText system would help readers better judge the utility of the approach.
We have added more information in the Results, including some examples of false positives.

5. The Interpretation provides no reference as to whether the resultant metrics of sensitivity, specificity and positive/negative predictive value are reasonable and for what purposes the data could be used. For instance, presumably a specificity of 92.4% would suggest the codified data would not be appropriate to use for clinical purposes, such as clinical decision support. There is also no comment on whether sensitivity or specificity may be of greater importance when analyzing large data sets. This reviewer considers specificity more important than specificity as missing data within a large dataset typically has less impact that misclassified data. Others may feel differently, but the difference between the two metrics is considerable (15X different) and thus some comment on the impact of this difference warranted.
We have added more discussion around the potential uses of this coding algorithm and what validity metric might be more relevant or preferred in the Interpretation.

6. In line 118 of the Interpretation the word “accurately” is entirely open to an individual’s point of view. Given the lack of any alternative text matching accuracy information this approach may be better, the same or worse than the simple text matching to the ATC nomenclature. While this section mentions the issues of routes and formulations as being confounding issues, the lack of any comment on this in the Results section has been previously noted. Much of the information needed to clarify the ambiguous areas of the manuscript are likely available to be included in a future version. The construct appears potentially very interesting for many other areas of free text within EHRs, but the current report does not provide enough detail to warrant others exploring the approach.
We have clarified this in the Interpretation and revised the use of the term ‘accuracy’ to describe the findings in more objective terms.

Reviewer 5: Dr. Thomas Freeman
Institution: University of Western Ontario
General comments (author response in bold)

1. The data found within the Electronic Medical Records (EMRs) of family physicians holds great value and promise for a variety of purposes including quality improvement initiatives, surveillance and research. However, there is evidence that data quality in primary care databases is highly variable and establishment of and improvement of that quality is necessary. The first step in examining data quality must be to accurately

identify within the data those elements of interest. Beyond that, data quality has been defined as having 4 domains: comparability, completeness, correctness, timeliness (Terry, Stewart, Cejic et al. BMC Medical Informatics and Decision Making 2019;19:30).
While we agree that data quality is critically important, this study was designed to address the issue of unstandardized data in the CPCSSN database. It is very difficult to evaluate data quality or conduct analyses on unstructured, raw EMR data coming in from 12 EMR systems; this method will become part of our pre-processing pipeline to make better coded data available to researchers and lay the foundation for future data quality studies. Our CPCSSN team has also just completed a full data quality assessment of the national database, which will be available on the CPCSSN website by April 2023 (https://cpcssn.ca/).

2. As I understand the trajectory of EMR data to the national database (CPCSSN) it begins with initial entry at the point of practice into one of the 12 EMR platforms across the 8 provinces; some of this data is extracted biannually and uploaded to CPCSSN where coding and cleaning occur. Presumably the machine learning model examined in this study would be applied at this stage. Once cleaned and coded the data becomes part of the CPCSSN database.
Yes, this new coding approach will become part of the CPCSSN processing pipeline, which will be applied to the national dataset once all regional data submissions have come in. The intent is that it will create more accurate coded medication data for use in research and surveillance.

3. It is sometimes difficult to follow the stages of this study and a flow chart would be very helpful. As I understand it there were 42 million individual records containing 2.4 million medication names out of which 151,296 records were selected for the training set using the DPD and previously coded values. This was then used to build a classification model using DPD labels. This was refined through 5 rounds of iteration and review resulting in 16,119 prescriptions for antibiotics. These antibiotic prescription names were then manually validated against the DPD names and found a high degree of sensitivity and specificity.
We agree this was not described clearly and have included a data flow diagram in the supplementary materials.

4. On line 84 reference is made to the ‘original record’ and it needs to be clarified what this refers to. I assume it means the record found in the CPCSSN database, but in fact the ‘original’ record is the clinical record in the EMR. Later the term prescription record is used. It would help to settle on one defined term for 'record' and use it consistently throughout.
We have revised the terms so that we are consistently using ‘medication’ record rather than both ‘medication’ and ‘prescription’ interchangeably. The ‘original’ record refers to the raw, unprocessed data straight out of the EMR.

5. All of the records identified by the algorithm as antibiotics were accurate to a high degree when compared to the reference standard DPD. But, how would we know if the algorithm missed any a/b records? A comparison to another method of identifying a/b records would be needed. It would be necessary to compare the new algorithm with another method of extracting the same information. The previous method of standardization mentioned using a pattern matching approach would seem to be most relevant for this purpose. The new algorithm is less cumbersome and effective, but is it as accurate?

It is possible that some antibiotic records were missed in the evaluation, although this subset of records were selected in an overly-inclusive way, as to reduce the chance of missing any mention of an antibiotic. We have informally checked the results of the pre-existing coder over the years as it's been used in the CPCSSN database, but unfortunately are not able to provide validity metrics without conducting another study (which our current resources do not support at the moment).

6. Under Limitations it is mentioned that because the CPCSSN data is derived from 12 different EMR platforms the algorithm model is robust, but there is no information about how the algorithm performed on individual EMR platforms. There is evidence that these different platforms show great variability in the recording and extraction from the clinical record prior to uploading to CPCSSN. There may be significant differences between what happens in the clinical world and recorded in individual records and its representation in the CPCSSN database. This is important to know for surveillance, quality improvement, research, and improvement in data quality.
We agree this would be an interesting sub-analysis, however it is outside the scope of this current paper. One of the benefits of the machine learning model described here is that it was designed for general use in the national CPCSSN database and includes all EMR systems, rather than creating up to 12 custom algorithms trained on individual EMR systems. Further, because we didn't conduct the training on data for each distinct EMR system, we did not feel it was appropriate to report the validity metrics by EMR system.

7. In summary this study makes an important contribution to one of the steps necessary to establish data quality in primary care EMRs by developing and validating a method for deriving antibiotic prescription records with a high degree of accuracy. However, this is only one element of data quality, and I would encourage the authors to also look at the remaining data quality aspects: completeness (in addition to name, strength, dose, frequency, route), correctness, comparability and timeliness.
We would be interested in conducting a separate study to evaluate different data quality metrics for medication data within CPCSSN. However, this is outside the scope of the current paper, as to adequately describe results and nuances for each of the proposed data quality indicators would require much more word space. A separate data quality report will be published on our website in the next month, which assesses the entirety of the national database: www.cpcssn.ca.