Article details: 2021-0170
Title: Hypertension identification using inpatient electronic medical record clinical notes: an explainable data-driven algorithm study
Authors: Elliot A. Martin PhD, Adam G. D'Souza PhD, Seungwon Lee MPH, Chelsea Doktorchik MSc, Cathy A. Eastwood PhD, Hude Quan MD PhD

Reviewer 1: Dr. Wilson Pace
Institution: University of Colorado, DARTNet Institute
General comments (author response in bold)

This manuscript reports on the use of machine learning (ML) approaches based on data abstracted from EHR inpatient notes via natural language processing (NLP) using USMLS constructs to identify pre-existing hypertension (HTN) in hospitalized individuals. The resultant information was compared to discharge codes (using ICD-10 CA) applied by hospital coders as well as to manual chart review outcomes by research nurse coders. The authors assert that identifying pre-existing HTN in the in-patient environment has value, that it is often not recognized by hospital discharge coders, and that by using NLP and ML, individuals with unrecognized pre-existing HTN can be identified and properly coded.

The sample for the analysis consisted of 3040 randomly selected hospital records. The sample size was powered on a 10% difference in detection sensitivity for common conditions, which is a clinically reasonable difference. Diagnostic codes for these admissions were applied by hospital coders as part of routine administrative activities and abstracted from an administrative database. These codes were linked with the hospital records. All records underwent manual review by nurse reviewers. The use of the UMLS concept unique identifiers to group textual constructs uncovered by NLP is being used more widely in the NLP world and is a strong approach for grouping and coding NLP textual strings. The ML modeling with an 80% training and 20% test split and training using a 5 fold methodology is a standard, reasonable approach. The exploration of "concept features" at the overall chart level and by document type was logical. The use of SHAP values to help provide explanatory ML outcomes from the two final XGBoost models is a strong approach to interpreting the models and helped to uncover the strength of simple text matching as opposed to full NLP.

There was a high rate of pre-existing hypertension in the overall sample size, clearly higher than the general population, which would help the PPV of any approach to diagnosis. The ML algorithm performed well compared to the manual chart review, but a simple text matching algorithm looking for a diagnosis of HTN in any note (nursing notes provided the highest return) performed as well (perhaps even slightly better) than either the document-concept model or the concept model using ML. All of the automated approaches produced more false positives than the hospital coders but found approximately 50% more people with pre-existing HTN.

1. While the work appears to be well carried out and demonstrates that expanding the search for pre-existing HTN to nurses' notes, in particular, would pick up many more people with this condition, the question of overall value is not well laid out. The background [Editor's note: the reviewer is referring to the Introduction in the main document] and Interpretation (though to a lesser extent than the background) wander between condition identification being important to monitoring health system performance, to risk adjustment (presumably of inpatients) to monitoring of ambulatory

sensitive conditions (which HTN clearly is.) The problem with all of these constructs, except perhaps the risk adjustment one, is that there is no evidence that HTN (malignant HTN) as the primary cause of a hospital admission is currently missed (thus the exercise does not inform health system performance overall (typically measured in this area by the percent of people whose blood pressure is controlled in the ambulatory setting) nor does it inform about performance related to ambulatory sensitive conditions (presumably none of the study sample were admitted for malignant HTN.)

The paragraph in the Background about HTN hospitalization rates as well as the paragraph concerning HTN hospitalization rates across the world have virtually nothing to do with the current study and simply confuse the reader. The finding that pre-existing HTN is often not included in discharge coding tracks with what the authors note in the Interpretation section, that the diagnosis of HTN was irrelevant to the hospitalization. Thus, perhaps the diagnosis should not be considered as part of risk stratification and is likely not related to any payment adjustments. It appears that adding the information at the time of discharge is an academic exercise (note that second definition of academic is "irrelevant".) Thus, it is not clear why a health system would spend the time and effort to adopt even the text matching approach to finding pre-existing HTN diagnoses.

In any case, given the higher false positive rate of all the automated approaches the findings would still require manual confirmation, but that could be focused to a specific note based on the automated output. Perhaps if this approach was expanded to include an array of diagnoses that would impact risk adjustments and payments then it would make sense to consider for a Canadian health system. At this point this appears to be a solution looking for a problem to solve. The authors need to better elucidate why a health system would care about the current performance of hospital discharge coders related to existing, non-relevant conditions, related to the inpatient stay. [Editor's note: please revise the Introduction and Interpretation to address the reviewer's comments.]
Note: I have no conflicts, financial or otherwise with this manuscript.
The ideal surveillance method for identifying hypertension cases would be a prospective cohort study with assessment of blood pressure measurements and/or other physiological parameters at repeated intervals, such a method is expensive and may be impractical in many settings. Over the past 20 years administrative data have been promising data sources for surveillance of chronic conditions. Many validation studies have assessed the validity of ICD coded algorithms to identify specific conditions (e.g., hypertension).
The value of computable phenotyping has recently been recognized in medicine. In the context of electronic medical records (EMR), a "computable phenotype," is a clinical condition or characteristic that can be ascertained by means of a computerized query to an EMR system using a defined set of data elements and logical expressions. It can be determined solely from data in EMRs and not require chart review or interpretation by a clinician. Computable phenotypes are also sometimes referred to as "EMR condition definitions," "EMR-based phenotype definitions," or simply "phenotypes".
Computable phenotype definitions can support reproducible queries of EMR data. These queries can then be replicated at multiple sites in a consistent fashion, enabling efficiencies and ensuring that populations identified from different healthcare organizations have similar features, or at least identified in the same way. Standard phenotype definitions can enable direct identification of cohorts based on population characteristics, risk factors, and complications, allowing decision makers to identify and target patients for screening tests and

interventions that have been demonstrated to be effective in similar populations. This identification process can be integrated with the EMR for real-time clinical decision support. Further, EMR phenotyping can be used to support a variety of purposes, including population management, quality measurement, and observational and interventional research.
This paper is indeed part of a larger research program on EMR phenotyping, which we now explain at the start of the introduction and cite some additional work of ours in this area, Pg. 3. Instead of packaging all the phenotypes together, we have decided to publish them individually as methods vary somewhat between them, and different cohorts were utilized for some conditions. We have added the following statement to the first paragraph of the introduction,
"On the other hand, Electronic Medical Record (EMR) phenotypes can be automated, making them relatively inexpensive to implement, and have the potential to have both high sensitivity and precision. We believe that EMR phenotyping, like we propose here, represents the future of case identification. It could supplement administrative data for health research, and its timely nature means it could also be used in clinical decision making. This work is part of our larger research program on EMR phenotyping".
We have also added further discussion of this in the first paragraph of the Interpretation, Pg. 18, and removed the portion of the introduction about hypertension being an ambulatory sensitive condition.

Reviewer 2: Dr. Jonathan Howlett
Institution: Calgary Foothills Hospital, Calgary, Alta.
General comments (author response in bold)

In the present study, the investigators evaluate a machine learning algorithm for diagnosis of hypertension in a cohort of inpatients from 3 hospitals in a large Canadian city. To do this, they analyzed a random sample of patient charts in 3 ways. For the gold standard diagnosis of hypertension, they (I think) utilized a formal chart review by trained abstractors. For a comparator, they used a previously validated method of ICD code abstractions of administrative data derived at hospital discharge and for the intervention, they utilized an algorithm based on analysis of all available electronic medical records, with application of a Unified Medical Language System to identify a Concept Unique Identifier.

They employed a standard array of data extraction and analytic techniques, and a newer method to determine Shapley Additive exPlanations (SHAP) to associate the presence or absence of hypertension in any given unique patient record.

From 3040 unique records, the authors determined that 48% of the cohort had hypertension and the the ML analysis of EMR-based models outperformed ICD-based diagnosis in terms of sensitivity with a comparable, albeit slightly lower positive predictive value. They conclude that hypertension tends to have clear documentation in the EMR and is best seen in nursing notes, suggesting the ML based methods may be an effective and low cost manner to detect hypertension for hospital inpatients.

Overall, the study is well designed and executed. The data is clearly shown and appears to support the conclusions. The strengths and limitations of the study and its design were adequately presented.

The accurate and practical identification of important clinical conditions is an important objective and this study shows that ML methods may, at least for this diagnosis, be used as a low cost alternative to typical abstracted ICD codification, at least in electronic medical records.

A few comments are noted below:

1) The format of the paper is non-standard and is confusing to the reader. It is unusual to begin a paper with Figures for example, and inclusion of a table of contents is not necessary [Editor's note: reviewer is referring to the TRIPOD checklist]. In addition, the figures are poorly labelled and annotated.
We have now reformatted the sections as suggested by the editor, and edited the figures for clarity.

2) The authors describe in detail the manner in which the ML algorithms were applied (except the manner in which the SHAP values were obtained). However, they did not clearly outline how the 'gold standard' hypertension diagnosis was made. I surmised that a physical chart review was employed to determine the presence or absence of hypertension, but this may be correct. This must be remedied.
We now clarify in the "Medical Chart Review" subsection of "Data Sources" that we use medical chart review as our gold standard, and expand our explanation on how these were collected, Pg. 8.

3) The key Table 4 is very dense, making it difficult to determine if any differences occurred between diagnostic methods. This table should be redone to clearly show where significant differences were noted between methods.
We have now altered this table to display the results for each model in its own column, and have bolded the best performing model in each case and italicized them when there was a tie, Pg. 16, 17.

4) The authors suggest the most reliable source for ML analysis was the Surgical Nursing note. It is likely that this form was available in a minority of cases. It would be useful to present the availability of each note in the charts (i.e. what percent of charts contained a Surgical Nursing note?). This will undoubtedly impact the choice of information source in future studies, since a broadly applicable method will require a data source that is universally present in all charts. This would bear some discussion by the authors.
We have now included a paragraph after Figure 3 in the Results, where we give the frequency with which the top identified notes occur, Pg. 15, 16. The "Surgical Assessment and History - Nursing" note was available in 37% of admissions for instance. We have also added a short discussion of this in the "Interpretation", where we again note the frequency of the top 2 documents chosen, and that the availability of documentation will also depend on the type of hospital visit, Pg. 19. For instance, our rationale for stratifying on the type of admission (surgical vs. non-surgical admission) was motivated by only 54% of surgical admissions having a discharge summary vs. 86% of non-surgical admissions.