Hypertension Identification Using Inpatient Electronic Medical Record Clinical Notes: An Explainable Data-Driven Algorithm Study

Elliot A. Martin PhD[1,2], email: eamartin@ucalgary.ca

Adam G. D'Souza PhD[1,2], email: adsouza@ucalgary.ca

Seungwon Lee MPH[1,2], email: seungwon.lee@ucalgary.ca

Chelsea Doktorchik MSc[1,3], email: ctadokto@ucalgary.ca

Cathy A. Eastwood PhD[1,3], email: caeastwo@ucalgary.ca

Hude Quan MD PhD[1,3], email: hquan@ucalgary.ca

1. Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
2. Alberta Health Services, Calgary, Alberta, Canada
3. Department of Community Health Science, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

All authors can be reached at:
Centre for Health Informatics
Cumming School of Medicine
University of Calgary
TRW Building, 3280 Hospital Drive NW
Calgary, Alberta, T2N 4Z6
Canada

Correspondence should be addressed to Dr. Elliot Martin at the above address, or

Telephone: (403) 220-2779
Fax: (403) 210-9744
Email: eamartin@ucalgary.ca

Key Words: Electronic medical records, case identification, machine learning, natural language processing, hypertension

ABSTRACT

Background: Case identification in inpatient environments is important for measuring health system performance and risk adjustment. The gold standard for case identification is chart review, which is costly and time consuming. Currently coded administrative data is used to capture conditions. Electronic Medical Records (EMR) can potentially be a superior rich data source for case identification compared to administrative data. Using machine learning, we developed an EMR-based hypertension case definition.

Methods: Chart review for inpatients was performed to identify documented hypertension status. Clinical notes in EMR were analyzed using natural language processing to extract their Unified Medical Language System concepts. The most important concepts and document-concept pairs were identified using machine learning. These were used to fit additional machine learning models, and to motivate a simpler concept search case identification algorithm. We compared EMR models against the commonly applied method of identifying cases using the International Classification of Diseases Tenth Revision codes abstracted from charts. The machine learning models were interpreted using Shapley Additive Explanations.

Results: Of our study sample (n=3040), 48.5% were hypertensive. Our final EMR-based models had higher sensitivity compared to ICD codes alone, >90% vs 47%, while maintaining high PPV, >90% vs 97%. Hypertension was best documented in nursing notes, which are not generally used in administrative data coding.

Interpretation: Our work demonstrates that hypertension tends to have clear documentation in EMR and is well classified by simple concept search on free text. Machine learning can provide insights into EMR documentation and can suggest simpler methods to implement.

INTRODUCTION

Condition identification is an essential part of a learning health system [1], monitoring health system performance, and risk adjustment. The gold standard for case identification is chart review, requiring trained clinicians to read each patient chart. This requires a substantial time commitment from well paid professionals -- often making it infeasible for population level research. To overcome this, coded administrative data is used to identify conditions.

Hospitalization rates for treatable conditions have been used as an indicator of appropriate primary care [3]. Hypertension is an example of an ambulatory care-sensitive condition. A lower hypertension hospitalization rate often indicates better access to primary care, or better quality of primary care services. However, it is often debated whether the hypertension hospitalization rate could be related to either true quality of care, data quality, or both. Accurate detection of hypertension in inpatient databases is therefore necessary when measuring health care performance.

Administrative health databases have been widely used to report hypertension hospitalization rates in many countries, because the data are routinely collected and cover wide geographic areas. After discharge, conditions are coded using the International Classification of Diseases (ICD). Canada uses ICD 10th Revision, Canadian Modification (ICD-10-CA). Unfortunately, hypertension is under-coded in ICD data, which can cast doubt on conclusions made when using administrative data to identify conditions and measure healthcare performance. Quan et al.[2] validated ICD hypertension data and reported a sensitivity of 68.3%, a positive predictive value (PPV) of 93.1%, a specificity of 97.8%, and a negative predictive value (NPV) of 87.7%. The observed under-coding of hypertension can potentially be attributed to coders having limited time (20-30 minutes in Canada) to abstract one chart and are therefore focused on identifying severe or main conditions, as mandated by reporting requirements [4,5].

Adopting EMRs to collect health information is a promising opportunity to improve the accuracy of identifying hypertension. However, methods for EMR-based hypertension identification are needed. Clinical notes are a rich source of information in EMRs, but are underutilized in automated processes like Machine Learning (ML) due to the difficulties in extracting information from them. The Unified Medical Language System (UMLS)[6] attempts to overcome some of these difficulties by mapping the varying lexical choices available in clinical documentation to a single Concept Unique Identifier (CUI). We believe CUIs can play an important role in creating interpretable models. Based on this concept, this study aimed to develop a hypertension case identification method using EMR inpatient clinical notes.

METHODS

Design

This is an EMR data-driven rule-based algorithm study design.

Setting and Participants

Our study cohort consisted of a random sample of 3,040 patients. We calculated that 3000 records are required to test the 10% difference in sensitivity of common comorbidities, such as hypertension (30.2%).

The patients were at least 18 years of age and were admitted to one of three acute care facilities in Calgary, Canada between January 1 and June 30, 2015. For patients with multiple admissions, one admission within the study period was randomly selected.

Data Sources

Sunrise Clinical Manager^{TM} (SCM)

AllScripts SCM is a city-wide, population-level EMR system currently in operation throughout all acute care facilities in Calgary, Canada. Alberta Health Services, the single health authority in Alberta, manages SCM and the associated electronic data warehouse [7].

Discharge Abstract Database (DAD)

The DAD is the administrative health database where diagnosis codes for all inpatient encounters are stored using ICD-10-CA [8]. The diagnosis codes are assigned by coders after discharge, based on the clinical documentation in patients charts. The database also contains basic demographic information about the patients (e.g., sex and age). The Canadian Institute for Health Information provides national coding standards and training programs for health information managers (i.e. coders) [9].

Medical Chart Review

We extracted the patient charts for each of these admissions from the hospital records departments [10]. Nurse reviewers looked for a listed diagnosis of hypertension in patients' History & Physical, Multidisciplinary Progress notes, Consult notes, and Discharge Summary. If a diagnosis was documented, the chart was labeled as hypertension present. The inter-rater reliability between reviewers was high (>0.8 kappa) [10]. We linked these three databases using Personal Health Number (a unique lifetime identifier), chart number (a unique number associated with a patient's admission), and admission date.

Defining Hypertension in DAD

In the DAD, hypertension was defined using the validated ICD-10 algorithm [8] through searching 25 diagnosis coding fields for each admission.

Defining Hypertension in EMR following Case Identification Pipeline

We outline the steps from extracting the EMR data to our final hypertension case-identification algorithms in Figure 1. The data was split into 80% training (n=2432) and 20% test (n=608). The training set was used for training and validation of all the machine learning models, and the test set was only used to compare the final EMR models with the ICD method.

Figure 1 Insert Here

Concept Extraction

We used the clinical Text Analysis and Knowledge Extraction System (cTAKES) [11], in particular its default clinical pipeline, to process all the clinical notes. We extracted clinical concepts in the form of CUIs from the UMLS. This method accounts for variation in terminology among EMRs, because UMLS maps synonymous terms to the same underlying concept. For example, in UMLS, the clinical concept "Hypertensive disease" is assigned the CUI "C0020538". The 2018AB UMLS release contains 67 synonyms for this clinical concept, including “BLOOD PRESSURE HIGH”, “HBP”, “HTN”, “Hyperpiesia”, “Hypertension”, and “systemic HTN”.

All these synonyms map to the same CUI, which allowed us to generate non-redundant (i.e., normalized) features. We used the negation and subject attribute annotators in cTAKES to label each CUI. These assessed whether the concept appeared in a negated context, e.g., ‘no evidence of hypertension’, and whether the subject to which the CUI was associated was the patient or someone else. The cTAKES outputs were then converted into a document-concept matrix containing the counts of each CUI for each document type (‘document’) and each chart. Only CUIs that had the patient as their subject, and that cTAKES determined were non-negated, were counted.

Feature Selection

Feature selection is the process of identifying the variables most relevant to the problem. Our features included both the CUIs and the clinical notes that could discriminate hypertensive cases. There were 58 types of clinical notes in our extracted EMR data, such as “Discharge Summary” and “History and Physical”. We used these to create two different types of feature sets. The first set of concept features contained only the number of times each concept occurred for each patient; the second set of document-concept features contained the number of times each concept appeared in a given document type. For example, the counts of history_and_physical-C0020538 and discharge_summary_medical-C0020538 would contribute to the same C0020538 feature in the first set, and would be separate features in the second set. The first set of features could illustrate the most reliable concepts used to identify hypertension, while the second could illustrate the most high yield and trustworthy documents to look at for future chart review.

The relative importance of each feature for determining hypertension was estimated using the gradient boosted algorithm XGBoost [12]. First, 20% of the patients were put aside to test the final algorithms, with the remaining 80% used for training and validation. For each feature set, five XGBoost models were fit, each using 5-fold cross-validation optimizing for AUC (See Table 1 for grid search parameters). This was done to ensure that only reliable features were selected, and to exclude those that only performed well on a subset of the data. The most important features that occurred in all 5 models were selected. Gain was used as the measure of feature importance (i.e., the improvement in accuracy of classification attributable to a feature).

Table 1. Grid search parameters for XGBoost models

Parameter	Document-type XGBoost Models
Cross-validation folds	5
lambda (L2 regularization)	0, 0.5, 1
Alpha (L1 regularization)	0, 0.5, 1
Max depth	5, 8
Min child weight	4, 8
N_estimators (number of trees)	500, 1000
Subsample	0.8

Final Models

Two final XGBoost models were fit using the selected features, one for each reduced feature set, again using the parameters in Table 1. Interpretability of our algorithms was a key study objective. A new technique [13] was used to compute SHAP (SHapley Additive exPlanations) values [14] on trees. If a feature has a large positive SHAP value for a given patient, it would indicate that the feature makes a positive finding of hypertension much more likely, with a large negative SHAP value indicating the converse. Finally, we used these results to suggest a simpler concept search strategy for case identification, and provide insights for future chart review.

Ethics Approval

Ethical approval for this study was obtained from the Conjoint Health Research Ethics Board at the University of Calgary (REB15-0790).

RESULTS

Cohort characteristics are presented in Table 2. Almost half of the cohort was hypertensive (48.5%).

Table 2: Characteristics of Study Sample

Variable	All (3040)	Hypertensive (1474)	Non-Hypertensive (1566)
Median age (IQR)	62 (48-76)	71 (61-82)	52.5 (38-65)
Female	1529 (50.3%)	710 (48.2%)	819 (52.3%)
Surgical Patient	1102 (36.3%)	482 (32%)	620 (39.6%)

The performance of the initial Document-Concept and Concept models are shown in Table 3, where the models are compared using the same training and validation sets on different folds. It can be seen that concept models seemed too overfit on the training data, but have similar performance to the document-concept models on the validation data. All the models performed relatively well on the validation data, with sensitivities and PPVs close to 90% throughout.

To remove spurious features, we selected only those that were in the top 20 most important features across all folds, for both sets of models. We chose the top 20 as feature importance decayed rapidly for both sets of models, and thus captured the most relevant features. Ten document-concept and 8 concept features remained. These top features were then used to create new document-concept and concept models, again using the parameters from Table 1.

Table 3: Performance of initial XGBoost Document-Concept model (DC) models and Concept Models (C)

	Training Data*		Validation Data*	
Model: Document-Concept/Concept	Sensitivity (%) DC/C	PPV (%) DC/C	Sensitivity (%) DC/C	PPV (%) DC/C
Fold 0	90/100	93/100	89/89	88/89
Fold 1	85/100	94/100	86/92	94/93
Fold 2	90/100	94/100	87/91	90/91
Fold 3	90/100	94/100	88/91	92/92
Fold 4	89/100	94/100	91/94	91/89

Note: *Folds 0 and 1 have a Training n=1945 and Validation n=487, and Folds 2, 3, and 4 have Training n=1946 and Validation n=486

To examine how features impacted the classification of each patient in the training set, we show the relationship between feature values and SHAP values for the concept model in Figure 2, and the document-concept model in Figure 3, where a larger SHAP value means a higher likelihood of classifying the patient as hypertensive. Unsurprisingly, Figure 2 shows that the concept for hypertension, C0020538, is the most important feature in the concept model, and is the only feature whose absence results in a strong negative classification. In Figure 3, it can be seen that all but one of the features in the Document-Concept model involve C0020538, which amounts to a ranked set of documents to search for hypertension documentation, with the best document to search being “surgical_assessment_and_history_-_nursing”. The predominance of the hypertension concept in determining hypertension status indicated that a simple concept search for C0020538 could also perform well, and would have the benefit of being simpler to implement.

Figure 2 Insert Here

Figure 3 Insert Here

In Table 4, we show the results of the final ML models as well as the simpler concept search algorithm and the ICD-10 algorithm. We can see that the EMR algorithms have much higher sensitivities and NPVs compared to the ICD algorithm across all stratifications. This is offset by slightly worse PPVs, which are still above 90% for all groups except the youngest two age stratifications, where it drops as low as 82%. Interestingly the youngest age stratification is the only place where the ICD algorithm has a worse PPV than the EMR algorithms. The ICD algorithm also has a higher specificity than the EMR algorithms, which are still above 90% for all groups except the oldest age stratification, where they drop as low as 87%. In general, we see that the concept search algorithm has quite comparable performance to the ML algorithms, despite its simplicity.

Table 4: Stratified Validity Scores Across Population Characteristics for Classification Models Document-Concept model (DC)/Concept Model (C)/Concept Search (CS)/ICD

	Sensitivity (%) DC/C/CS/ICD	Specificity (%) DC/C/CS/ICD	PPV (%) DC/C/CS/ICD	NPV (%) DC/C/CS/ICD
All N = 608	95/91/95/47	92/93/92/98	91/93/92/97	95/91/95/66
By Age				
< 45 (n=123)	100/100/100/29	98/97/97/99	88/82/82/80	100/100/100/92
45-64 (n=206)	87/90/90/42	92/90/91/98	87/84/85/94	92/93/94/74
>64 (n=279)	91/96/97/50	88/87/87/97	95/95/95/98	79/89/90/42
By Service				
Surgical (n=213)	90/91/91/44	94/93/93/99	92/90/90/98	92/93/93/70
Non-Surgical (n=395)	91/96/97/48	93/91/92/98	93/92/92/96	91/96/96/64
By Sex				
Female (n=302)	90/94/94/45	94/92/92/98	92/91/91/95	92/95/95/68
Male (n=306)	91/95/96/48	93/91/92/99	93/92/93/97	91/94/95/64

INTERPRETATION

We examined how well hypertension could be identified in an inpatient population using UMLS concepts extracted from EMR clinical notes. We employed a data-driven approach to select relevant concepts and document-concept pairs in an automated way, thereby minimizing the need for clinical input. Our methods could be used with a common data model such as Observational Medical Outcomes Partnership (OMOP) [15,16]. OMOP makes use of a NOTE_NLP table where CUIs, their annotations, and their document types are referenced. Our algorithms can be executed from these fields.

While XGBoost is a powerful model, the potentially large number of trees makes it hard to determine how it arrives at a given classification. Therefore, we employed SHAP values to assess the impact of each feature on classification. They showed that the classification was dominated by the hypertension concept C0020538, which motivated us to try a simple concept search algorithm.

The simple concept search has comparable performance to the ML algorithms, and is the approach we recommend due to its simplicity. The concept ML algorithm did identify the hypertensive medication amlodipine (C0051696) as the second most relevant feature, but when the concept algorithm (C) was compared to the concept search algorithm (CS) in Table 4, it did not reliably improve classification. This indicates that medications do not robustly indicate hypertension status.

Our EMR-based method also provides insight into the underlying EMR data documentation. Our document-concept algorithm indicates that hypertension was documented most reliably in “Surgical Assessment and History - Nursing” followed by “Nursing Transfer Report - Emergency Department to Inpatient”. Canadian coders are not required to review these nursing documents and only review physician documentation [17]. In hospitals, nurses check patient blood pressures and document it in nursing notes, and they also collect patients' daily clinical information. Thus, our EMR-based method could be automated, which can avoid potential bias associated with coding guidelines and practice [18]. This has the potential to improve ICD databases with minimal cost.

Our study demonstrates EMR-based methods have higher sensitivity than the ICD-based method. This could result from coding practices in ICD data. Coders review charts to code hypertension. For many inpatients, hypertension as a comorbidity to the main condition does not contribute to hospital length of stay and clinical outcomes significantly. Coders are not mandatorily asked to code these secondary diagnoses, including hypertension. However, our EMR-based method searches clinical notes regardless of the contribution of hypertension to clinical care. It also captures documented hypertension efficiently, and may be a potential solution to improve DAD quality.

The presented EMR methods have various applications in clinical and research contexts, namely health system performance evaluation. While hypertension is most often diagnosed in a primary care setting, patients are admitted to hospital when the severity of the condition worsens. Therefore, identifying hypertension in an inpatient setting, without relying on primary care data, is essential for health system performance assessment. The ML approach presented in the current study can also be applied to identifying other conditions in inpatient EMR data, which may not have as straightforward documentation.

Limitations

We used inpatient documentation only, and are aware that hypertension is largely managed in outpatient settings. However, our study was aimed at developing EMR-based hypertension case identification in order to overcome under-coding issues in ICD databases. Of note, our methods performed comparably to state-of-the-art work using outpatient data [19].

Our reference standard identified cases based on clinician documentations and did not re-diagnose hypertension based on charts. It is challenging to follow hypertension diagnosis guidelines, because charts do not contain detailed information. Clinicians document diagnoses rather than supporting information. Although blood pressures are part of diagnoses, its criteria can vary across countries, including cut-off values for blood pressure to define hypertension [20,21]. Therefore, clinical rule-based algorithms may not be as robust when performing case identification in other contexts. Finally, we have not conducted external validation of our algorithm using data from other jurisdictions. This type of external validation study between multiple systems may be feasible using common data models, such as OMOP [15,16].

Conclusion

We have leveraged EMR clinical notes to create a data driven case identification pipeline for inpatients. We utilized ML models to identify the most relevant concepts and documents to examine in the EMR, and used those insights to create a simpler concept search case identification algorithm. This algorithm has great potential to improve hospital discharge abstract administrative data quality, and also to provide a tool to measure hypertension hospitalization rates for system performance evaluation. The ML models also provide insights into EMR documentation for future research, fulfilling the iterative feedback goal of a learning health system.

LIST OF ABBREVIATIONS

AUC: Area Under the receiver operating characteristic Curve

CART: Classification and Regression Trees

cTAKES: clinical Text Analysis and Knowledge Extraction System

CUI: Concept Unique Identifier

DAD: Discharge Abstract Database

EMR: Electronic Medical Record

ICD: International Classification of Diseases

ML: Machine Learning

NPV: Negative Predictive Value

PPV: Positive Predictive Value

SCM: Sunrise Clinical Manager

SHAP: SHapley Additive exPlanations

UMLS: Unified Medical Language System

Consent for publication

Not applicable.

Data Sharing

The datasets analysed during the current study are not publicly available due to the potential for identifying information to be exposed in the clinical notes. Working with the data is possible through collaboration with the Centre for Health Informatics and Alberta Health Services.

The python code used to perform this analysis can be found at the github repository https://github.com/centre-for-health-informatics/Hypertension-Case-Identification, DOI 10.5281/zenodo.4543942, under an open MIT license.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a Canadian Institutes for Health Research, Operating Project Grant 201809PJT-409935-HS1-CBBA-114817.

Authors' contributions

E.M. performed the majority of the analysis and writing. A.D. assisted with the analysis and writing. S.L. and C.D. provided subject matter expertise and assisted with writing. C.E. and H. Q. were responsible for study design and assisted in the writing.

Acknowledgements

The authors would like to thank Yuan Xu and Sang Min Lee for their helpful comments on the manuscript.

REFERENCES

1. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Science translational medicine.* 2010;2(57):57cm29.
2. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health services research.* 2008;43(4):1424-1441.
3. Nyweide DJ, Weeks WB, Gottlieb DJ, Casalino LP, Fisher ES. Relationship of primary care physicians' patient caseload with measurement of quality and cost performance. *Jama.* 2009;302(22):2444-2450.
4. Lucyk K, Tang K, Quan H. Barriers to data quality resulting from the process of coding health information to administrative data: a qualitative study. *BMC health services research.* 2017;17(1):766.
5. Tang KL, Lucyk K, Quan H. Coder perspectives on physician-related barriers to producing high-quality administrative data: a qualitative study. *CMAJ open.* 2017;5(3):E617-e622.
6. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research.* 2004;32(Database issue):D267-270.
7. Lee S, Xu Y, D'Souza A, et al. Unlocking the Potential of Electronic Health Records for Health Research. *International Journal of Population Data Science.* 2020;5(1).
8. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical care.* 2005;43(11):1130-1139.
9. Quan H, Smith M, Bartlett-Esquilant G, Johansen H, Tu K, Lix L. Mining administrative health databases to advance medical science: geographical considerations and untapped potential in Canada. *The Canadian journal of cardiology.* 2012;28(2):152-154.
10. Eastwood, CA, Southern, DA, Khair, S, et al. The ICD-11 Field Trial: Creating a Large Dually Coded Database, 13 May 2021, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-505934/v1]
11. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association: JAMIA.* 2010;17(5):507-513.
12. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, California, USA.
13. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell.* 2020;2(1):56-67.
14. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017; Long Beach, California, USA.
15. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Studies in health technology and informatics.* 2015;216:574-578.
16. Overhage J, Ryan P, Reich C, Hartzema A, Stang P. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association : JAMIA.* 2012;19(1):54-60.

17. Information CIfH. Canadian coding standards for version 2018 ICD-10-CA and CCI. In: CIHI Ottawa, ON; 2018.
18. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. *BMC medical informatics and decision making.* 2016;16(1):138.
19. Teixeira PL, Wei WQ, Cronin RM, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *Journal of the American Medical Informatics Association : JAMIA.* 2017;24(1):162-171.
20. Garies S, Hao S, McBrien K, et al. Prevalence of Hypertension, Treatment, and Blood Pressure Targets in Canada Associated With the 2017 American College of Cardiology and American Heart Association Blood Pressure Guidelines. *JAMA network open.* 2019;2(3):e190406.
21. Khera R, Lu Y, Lu J, et al. Impact of 2017 ACC/AHA guidelines on prevalence of hypertension and eligibility for antihypertensive treatment in United States and China: nationally representative cross sectional study. *BMJ (Clinical research ed).* 2018;362:k2357.

Figure 1. Case identification pipeline flow chart.

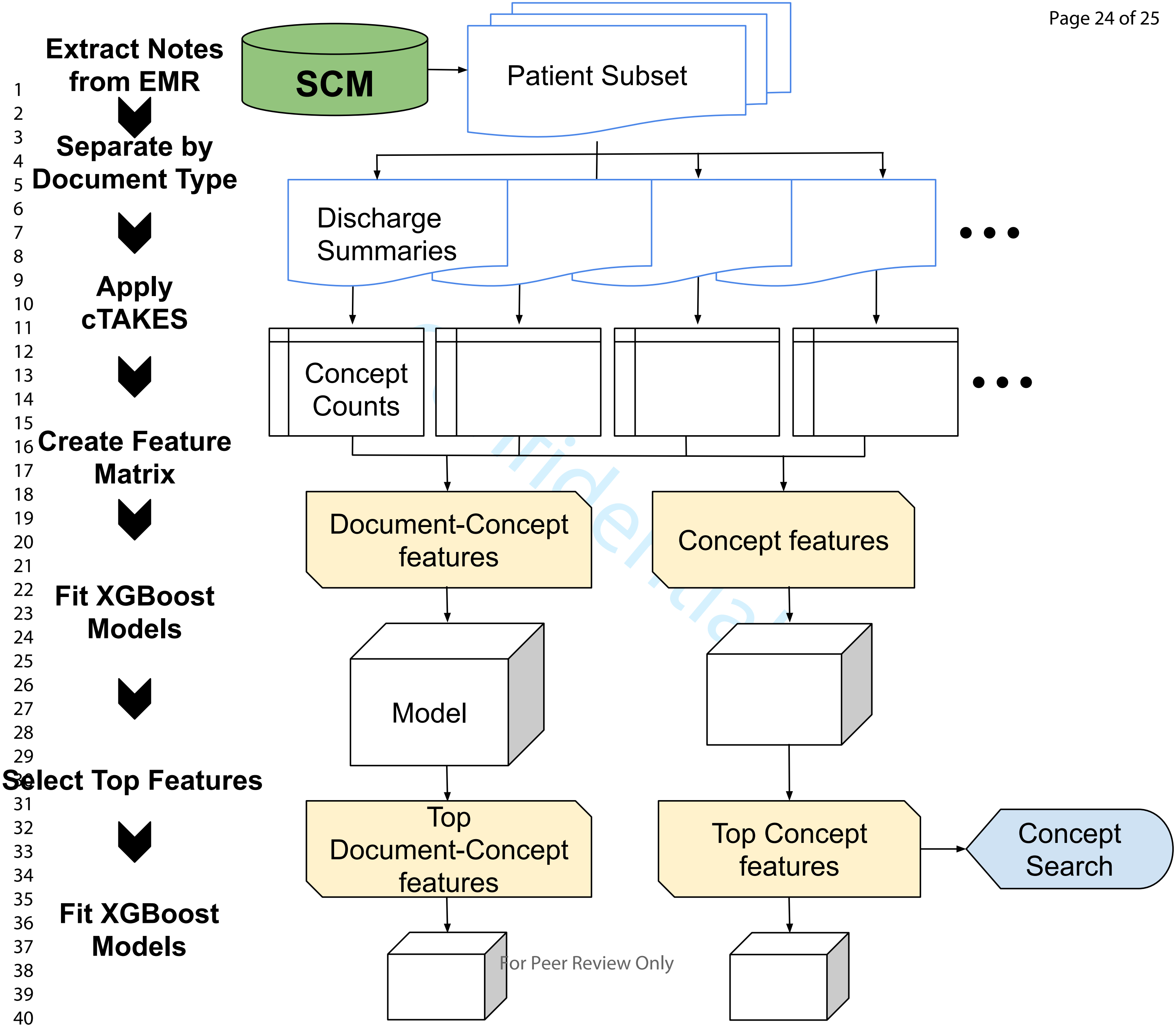
We randomly sampled 3040 inpatient charts from SCM and extracted their associated clinical notes, identifying UMLS concepts with cTAKES. XGBoost models were used to separately select the most important concept and document-concept pair features. The selected features were used to fit reduced concept and document-concept XGBoost models. The concept features were also used to implement a simple search algorithm for the hypertension concept C0020538.

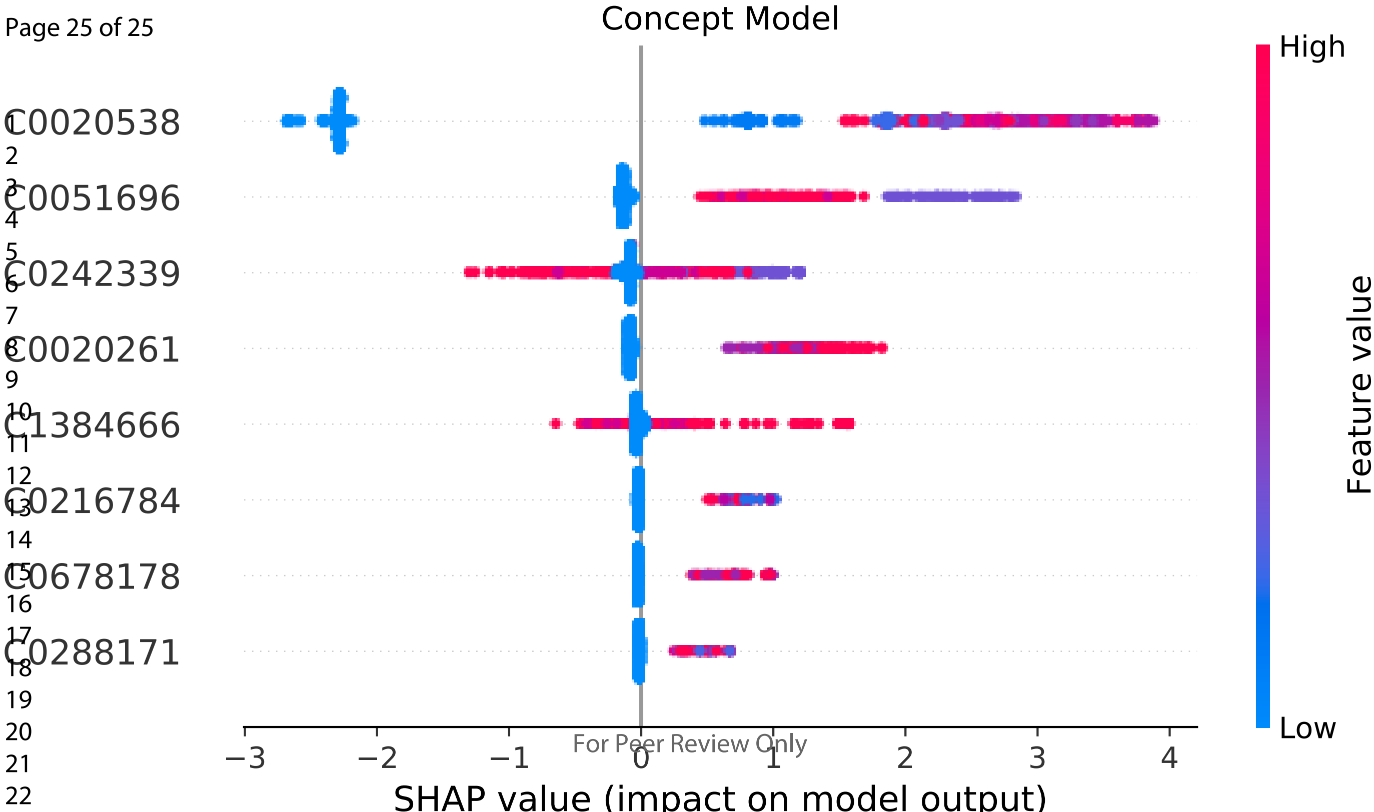
Figure 2. SHAP values for final Concept Model

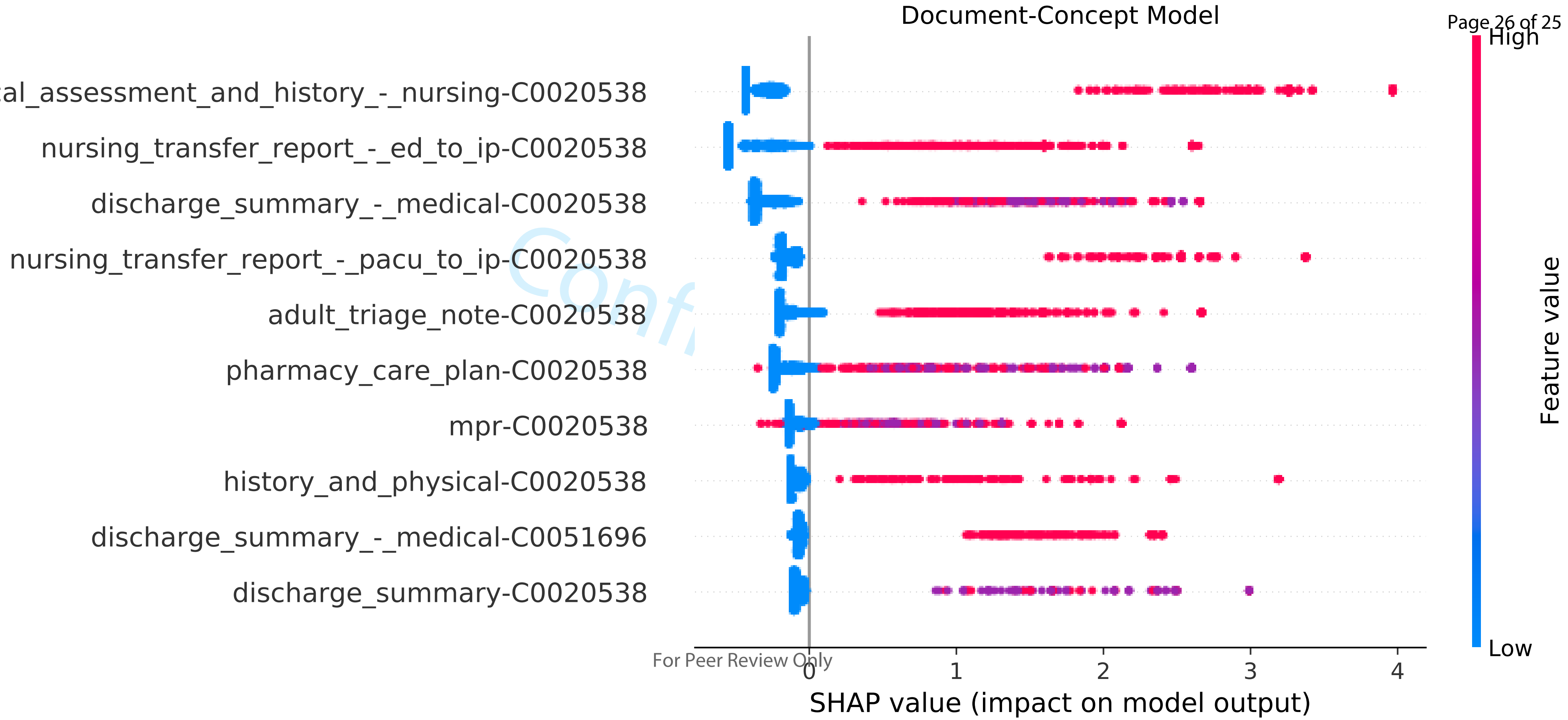
The SHAP values for each patient in the training set (n=2432) by model feature. Each dot represents a patient, with the SHAP value on the x-axis, and feature value given by its color. The higher the SHAP value the more likely the patient will be classified as hypertensive. The hypertension concept C0020538 is the most predictive feature, and the only one where a low count results in a significant likelihood of the patient not being classified as hypertensive.

Figure 3. SHAP values for final Document-Concept Model

The SHAP values for each patient in the training set (n=2432) by model feature. Each dot represents a patient, with the SHAP value on the x-axis, and feature value given by its color. The higher the SHAP value the more likely the patient will be classified as hypertnesive. All the features represent different places to search for the hypertension concept C0020538, except the second to last important which looks for the amlodipine concept (C0051696) in the "Discharge Summary - Medical" document.


Extract Notes from EMR
Separate by Document Type
Apply cTAKES
Create Feature Matrix
Fit XGBoost Models
Select Top Features
Fit XGBoost Models
SCM
Patient Subset
Discharge Summaries
Concept Counts
Document-Concept features
Concept features
Model
Top Document-Concept features
Top Concept features
Concept Search


Concept Model
C0020538
C0051696
C0242339
C0020261
C1384666
C0216784
C0678178
C0288171
−3
−2
−1
0
1
2
3
4
SHAP value (impact on model output)
High
Low
Feature value


Document-Concept Model
surgical_assessment_and_history_-_nursing-C0020538
nursing_transfer_report_-_ed_to_ip-C0020538
discharge_summary_-_medical-C0020538
nursing_transfer_report_-_pacu_to_ip-C0020538
adult_triage_note-C0020538
pharmacy_care_plan-C0020538
mpr-C0020538
history_and_physical-C0020538
discharge_summary_-_medical-C0051696
discharge_summary-C0020538
0
1
2
3
4
SHAP value (impact on model output)
High
Feature value
Low