BMJ Health & Care Informatics

Exploring the reliability of inpatient EMR algorithms for diabetes identification

Seungwon Lee ,[1,2] Elliot A Martin,[1,2] Jie Pan,[1,3] Cathy A Eastwood,[1,3] Danielle A Southern ,[3] David J T Campbell,[1,4] Abdel Aziz Shaheen,[1,4] Hude Quan,[1,3] Sonia Butalia[1,4]

To cite: Lee S, Martin EA, Pan J, *et al*. Exploring the reliability of inpatient EMR algorithms for diabetes identification. *BMJ Health Care Inform* 2023;**30**:e100894. doi:10.1136/bmjhci-2023-100894

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100894).

Received 06 September 2023
Accepted 04 December 2023

[1]Community Health Sciences, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada
[2]Provincial Research Data Services, Alberta Health Services, Edmonton, Alberta, Canada
[3]Centre for Health Informatics, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada
[4]Department of Medicine, University of Calgary Cumming School of Medicine, Calgary, Alberta, Canada

Correspondence to
Dr Seungwon Lee;
seungwon.lee@ucalgary.ca

ABSTRACT

Introduction Accurate identification of medical conditions within a real-time inpatient setting is crucial for health systems. Current inpatient comorbidity algorithms rely on integrating various sources of administrative data, but at times, there is a considerable lag in obtaining and linking these data. Our study objective was to develop electronic medical records (EMR) data-based inpatient diabetes phenotyping algorithms.

Materials and methods A chart review on 3040 individuals was completed, and 583 had diabetes. We linked EMR data on these individuals to the International Classification of Disease (ICD) administrative databases. The following EMR-data-based diabetes algorithms were developed: (1) laboratory data, (2) medication data, (3) laboratory and medications data, (4) diabetes concept keywords and (5) diabetes free-text algorithm. Combined algorithms used *or* statements between the above algorithms. Algorithm performances were measured using chart review as a gold standard. We determined the best-performing algorithm as the one that showed the high performance of sensitivity (SN), and positive predictive value (PPV).

Results The algorithms tested generally performed well: ICD-coded data, SN 0.84, specificity (SP) 0.98, PPV 0.93 and negative predictive value (NPV) 0.96; medication and laboratory algorithm, SN 0.90, SP 0.95, PPV 0.80 and NPV 0.97; all document types algorithm, SN 0.95, SP 0.98, PPV 0.94 and NPV 0.99.

Discussion Free-text data-based diabetes algorithm can yield comparable or superior performance to a commonly used ICD-coded algorithm and could supplement existing methods. These types of inpatient EMR-based algorithms for case identification may become a key method for timely resource planning and care delivery.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Identifying people with diabetes in databases is typically carried out by using International Classification of Disease codes, laboratory results and/or medications.

WHAT THIS STUDY ADDS

⇒ The diabetes identification algorithm based on free-text electronic medical records (EMR) notes shows excellent performance. This study further supports the idea that EMRs contain a wealth of details that can be leveraged to complement existing methods to identify people with diabetes within databases.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study provides evidence that free-text EMR data could enhance the flow of diabetes information in clinical care and improve associated downstream processes in case identification, surveillance and clinical outcome research.

INTRODUCTION

Accurate identification of chronic conditions, such as diabetes, within acute care facilities or hospitals is imperative for delivering optimal care.[1] Information regarding comorbidity status is typically gathered by healthcare professionals during their care encounters and stored within electronic medical records (EMRs). The collected information is subsequently conveyed to other care providers based on individual needs. Comorbidity information is not only useful in point-of-care clinical encounters but is also useful for research, quality improvement and resource planning. In the Canadian context, a key inpatient administrative health database, the discharge abstract database (DAD),[2] is used by>90% of hospitals, is coded by trained coding specialists reviewing physician documentations from the EMRs who assign International Classification of Diseases 10th Revision Canadian modification (ICD-10-CA) codes. This database is populated by these coding specialists who review discharge summaries from the EMRs and assign International Classification of Diseases (ICD) codes to each encounter.[3] The DAD serves various purposes such as care service planning activities, fiscal planning, operational planning, population surveillance and epidemiology. However, there exists a considerable delay in obtaining this data.

The increased adoption of EMRs within acute care facilities,[4] coupled with the integration of artificial intelligence techniques in healthcare,[5] has created the potential to extract chronic conditions and comorbidities directly from EMRs. This has the benefit of enhancing operational data practices in health systems and ensuring timelier information. For example, diabetes definitions have typically used ICD codes and laboratory and medication data to define diabetes in clinical datasets.[6] However, most detailed contextual information in healthcare is stored within free-text notes in paper charts or EMRs. The advancement of natural language processing (NLP) techniques now enables using EMR free-text data to refine condition definitions and facilitate the identification process, thereby enhancing various healthcare processes, including real-time point of care, research and care planning processes.

Our hypothesis is that a diabetes algorithm, using clinical free-text notes, can perform similarly or better than existing standard methods. We were also interested in assessing whether different components of free-text notes could contribute to phenotyping diabetes. The purpose of this study was to develop diabetes algorithms based on different EMR data modalities and compare their performances.

MATERIALS AND METHODS

Study population and design

This study is a retrospective cohort study covering the period of 1 January 2015 to 30 June 2015, from Alberta, Canada. This cohort was assembled from data sources listed below.

Data sources and linkage

The EMR and administrative data records were linked using Personal Health Number (PHN), and generated Patient Identification and Encounter Management details (eg, encounter number, health record number) sourced from the Clinibase system. These represent a unique set of identifiers for patient encounters[7] that are loaded into the EMR system. The combination ensured the linkage was pinpointed to the correct admission period contained within EMR data. We developed this linkage mechanism in a previous study[8] and subsequently created multiple EMR databases linking administrative health databases. The PHN and other personal identifiers were anonymised after the linkage was completed. The following sources of information were used: chart review database, Allscripts Sunrise Clinical Manager EMR and the DAD.

Chart review database

A previously conducted project assembled a chart review cohort of randomly selected patients in acute-care facilities in Calgary, Alberta.[9] The chart review data recorded patients' chronic disease status (binary) which included diabetes status, admission date and other system variables for linking it to the DAD and other data sources. The chart review included a total of 51 medical conditions and 3 healthcare-related adverse events. The chart review team consisted of six nurses who received training and followed a consistent protocol to review the charts. These reviewers were blinded to the ICD coding status.

Allscripts Sunrise Clinical Manager EMR

Sunrise Clinical Manager (SCM) has been used as the inpatient EMR for several acute-care sites operated by Alberta Health Services (AHS), the single health authority in the province of Alberta, since 2009. This EMR contains (but is not limited to) patient demographic information, laboratory information, medications, free-text history and physical notes, interdisciplinary progress notes, and discharge summaries for inpatient encounters. Detailed description of this EMR system is available in our previous work.[1]

Discharge abstract database

DAD is a national Canadian administrative health database which includes all inpatient separations (by discharge or death) through a collaborative system set up between provincial, and territorial governments, and the Canadian Institute for Health Information (CIHI). CIHI sets national training requirements for those responsible for coding the data. The utilisation of administrative health data, such as DAD, is widely acknowledged as the reference standard in Canada for both research activities[10] and public health initiatives,[11] from using ICD codes.

Data extraction

Once the coded patient records were deterministically linked to the EMR using PHN and Clinibase variables, linkage to subtables within EMR of interests was conducted through system variables (eg, table record identifier, health record number). We extracted and cleaned these EMR subtables that contained the following information: (1) inpatient laboratory subtable (contains all conducted laboratory tests within a patient encounter period), (2) inpatient medication subtable (contains all medications prescribed and fulfilled to the patient within a patient encounter period), and (3) subtable containing all clinical notes (free-text notes documented throughout the patient encounter) period. ICD codes were obtained from the linked DAD data. These EMR subtables were used to develop varying diabetes algorithms listed in the next section.

Diabetes algorithm development

Chart review labels served as the gold standard labels for algorithm development.

Operational standards—validated administrative data-based ICD codes algorithm

Current operational algorithm standards for surveillance and research are based on ICD-coded data. The National Diabetes Surveillance System (NDSS)[12] employs ICD-based code algorithm developed by Quan *et al*[13] and is inclusive of ICD-10-CA codes E10–E14 during

hospitalisation. We assessed the performance of the algorithm by Quan *et al* against the chart review labels.

EMR data-based algorithms

Various approaches were implemented for developing algorithms accounting for different data modalities. All algorithms were compared against the chart review labels for performance measurements.

Laboratory data-based clinical diagnosis algorithm

To identify diabetes, we used haemoglobin A1C (HbA1c) tests, oral glucose tolerance tests, random plasma glucose tests, or fasting plasma glucose tests, adhering to the thresholds outlined in Diabetes Canada's national guidelines for diagnosis. The criteria and thresholds for these tests have been published.[14] While Diabetes Canada requires at least two separate test types for a diabetes diagnosis, the varied prevalence of recommended tests for each patient led us to implement a single test meeting the diagnostic criteria[15] for performance reporting in this study.

Medication data-based clinical diagnosis algorithm

The medication clinical algorithm included any use of a single (or multiple) agent(s) that are commonly used to treat diabetes. The list of diabetes medications was derived from Diabetes Canada's national guidelines, reviewed by clinicians (endocrinologists), and validated on the Canada's Drug Product database[14] (online supplemental appendix table 1).

Inpatient laboratory and medication data-based clinical diagnosis algorithm

This clinical diagnosis algorithm included both laboratory and medications data. Specifically, the absence of diabetes was defined as the highest HbA1c laboratory result below 6.5%[16 17] with no evidence of prescribed or fulfilled medications. Pre-diabetes was defined by the highest HbA1c falling within the range of 6.0%–6.4% or through an oral glucose tolerance test, random plasma glucose test, or fasting plasma glucose test adhering to the thresholds listed in the Diabetes Canada guidelines, and no prescribed antidiabetic medications. Diabetes status was categorised as follows: as (1) HbA1c≥6.5%, if no evidence of medication, (2) meeting glycaemic targets: HbA1c values<7.0%, supported by evidence of both prescribed and dispensed medications, and (3) not meeting glycaemic targets: indicated by the highest HbA1c laboratory result closest to discharge>7.0 %.[18] Another subgroup of individuals with diabetes was identified as those with appropriately intensified therapy with agents known to confer cardiorenal benefit such as (1) GLP1RA if obese or with a history of cardiovascular disease or stroke, and (2) SGLT2 if chronic kidney disease (low GFR or albuminuria) or cardiovascular disease. These data were analysed using a time-series context, and all laboratory and medication records were used.

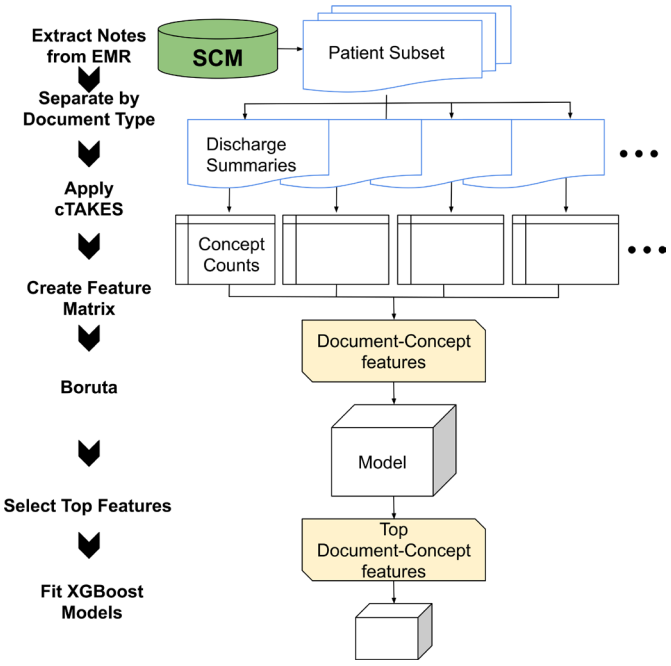


Figure 1 Clinical Text Analysis and Knowledge Extraction Systems (cTAKES)and XGBoost free-text algorithm. After free-text notes were extracted from the Sunrise Clinical Manager (SCM) electronic medical record (EMR), these notes were processed by document type using cTAKES. Boruta feature selection was employed and XGBoost classification model was fit. This diagram was adapted and modified from our previous work on hypertension.

NLP clinical notes-based machine learning (ML) algorithm

Free-text notes were cleaned and decoded into American Standard Code for Information Interchange (ASCII) to ensure extracted free-text notes were converted to an analyzable format. Then all free-text notes were stratified by document types. The default clinical pipeline of clinical Text Analysis and Knowledge Extraction Systems (cTAKES)[19] was used to process the raw text documents into unified medical language system's (UMLS) concept unique identifiers (CUIs) for each patient.[20] Two algorithms were developed: the first one was a CUI search of the diabetes concept which encompasses its synonyms (eg, diabetes, diabetes mellitus, hyperglycaemia), and the second algorithm was based on a data-driven model of all CUIs extracted from all document types. These CUIs covered anatomical sites, signs/symptoms, procedures, diseases/disorders and medications.

A data-driven supervised ML model on all document types and CUIs was developed (figure 1) and closely follows our previous work.[21] Boruta[22] feature selection algorithm was applied to reduce the dimension of CUIs. An XGBoost[23] algorithm was trained against the chart review cohort. The dataset was divided into 80:20 training ratio stratified by the diabetes outcome to ensure a similar ratio between the labels was maintained. Fivefold cross-validation was employed, and a grid search of hyperparameters was conducted. Feature importance assessed for the top predictive CUI document name pair (ie, a specific

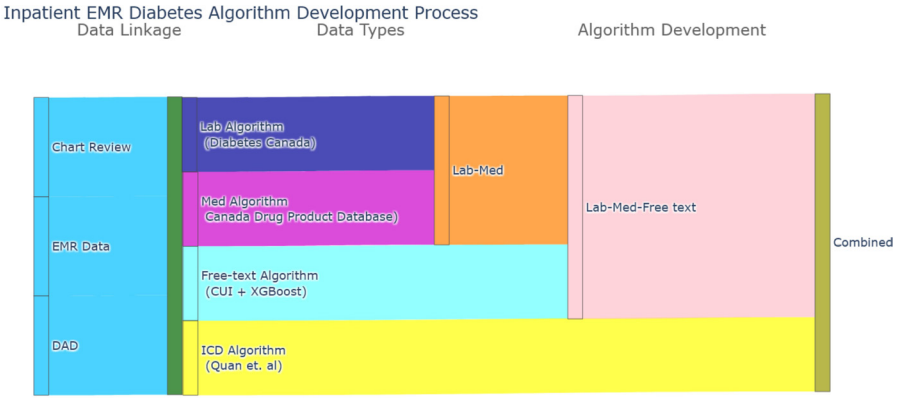


Figure 2 Flow process of algorithm development. Chart Review data were deterministically linked to discharge abstract database (DAD) and inpatient data. The International Classification of Diseases (ICD) algorithm was developed by Quan *et al*. Laboratory and medication algorithms used Diabetes Canada[3]'s established definitions. Medications were ascertained on Canada's drug product database. Free-text algorithm employed clinical Text Analysis and Knowledge Extraction Systems (cTAKES) for extracting concept unique identifiers (CUIs) from clinical notes and XGBoost was applied.

CUI in a specific document type) associated with diabetes. Top 20 document type—concept predictive features were identified after fitting the XGBoost algorithm.

Combined algorithms used *or* statements between the above algorithms.

Evaluation metrics and validation

Several evaluation metrics were calculated to assess the model performance. These metrics included sensitivity (SN), specificity (SP), positive predictive value (PPV), and negative predictive value (NPV). Statistical tests such as t-test, χ^2 and Kruskal-Wallis one-way analysis of variance test were applied for continuous, categorical and ordinal variables, respectively.

Figure 2 schematically presents the process flow from data linkage to algorithm development. Figure 1 depicts the detailed algorithm development process of applying cTAKES on the free-text data. We determined the best performing algorithm as the one that showed the high performance of SN and PPV.

RESULTS

Cohort overview

We analysed the charts of 3040 individuals, and their demographic details are summarised in table 1. The median age was 62.5 years and there was an equal distribution between males and females. The median body mass index of the cohort was 23.8 kg/m^2, and approximately 1617 individuals (53.2%) had no Charlson comorbidities. Among these 3040 individuals, 583 individuals (19.2%) had diabetes based on the chart review 'gold standard'. The cohort with diabetes was, on average, 10 years older than the overall chart review cohort (p<0.01). Within the diabetes cohort, there was a higher proportion of males than females (p<0.01). Additionally, the comorbidity profiles differed between the two groups, with the diabetes subcohort exhibiting a higher prevalence of comorbidities compared with the overall cohort (p<0.01).

Feature selection on all document type ML model

The cTAKES system successfully processed a total of 59 document types and processed 692 918 free-text records within this cohort. The system also extracted negation status and experiencer details, distinguishing between

Table 1 Demographics of people with diabetes from the chart review cohort

	Chart review cohort (n=3040)	Diabetes cohort (n=583)	No diabetes cohort (n=2457)	P value
Demographics				
Age in years, median (IQR)	62.5 (28.0)	69.0 (19.0)	59.4 (19.6)	<0.01
Sex (F), proportion	1530 (50.3)	235 (40.3)	1161 (47.3)	<0.01
Body mass index, median (n, IQR)	23.8 (29.4)	24.7 (31.5)	23.8 (28.8)	0.56
Charlson comorbidities				<0.01
0	1617	57 (9.8)	1559 (63.5)	
1	896	240 (41.2)	654 (26.6)	
2	419	203 (34.8)	214 (8.7)	
3+	111	83 (14.2)	30 (1.2)	

Table 2 Performance of clinical and ML algorithms on the testing dataset (n=609)

Algorithm type	Sensitivity	Specificity	PPV	NPV	F1
ICD (Quan *et al*)[13]	0.84	0.98	0.93	0.96	0.88
SCM EMR data					
Laboratory	0.37	0.96	0.69	0.86	0.48
Medications	0.89	0.98	0.91	0.98	0.90
Free-text (CUI: keywords search)	0.73	0.93	0.70	0.93	0.71
Free-text (all documents; CUI and XGBoost)	0.95	0.98	0.94	0.99	0.95
Combinations					
Labs+Meds	0.90	0.95	0.80	0.97	0.85
Labs+Meds + Free text XGBoost	0.97	0.95	0.81	0.99	0.88
Labs+Meds + Free text XGBoost+ICD	0.97	0.94	0.79	0.99	0.87
Medications+free text XGBoost (SCM EMR)	0.97	0.98	0.90	0.99	0.94
Medications+free text XGBoost+ICD algorithm	0.97	0.96	0.87	0.99	0.92

CUIs, concept unique identifiers; EMR, electronic medical record; ICD, International Classification of Diseases; SCM, Sunrise Clinical Manager.

patients and family members. We retained only CUIs that were not negated, and had the patient as the experiencer, resulting in a total of 83 107 CUIs. Using the Boruta method, it recommended the inclusion of 42 ranked features, with an additional three features identified as tentative. Therefore, we considered the top 45 ranked features, which constituted the training dataset for the XGBoost model.

Algorithm performance

Table 2 presents the performance of the diabetes Clinical and ML algorithms on the testing dataset. The administrative database ICD-based algorithm yielded SN of 0.84, SP of 0.98, PPV of 0.93 and NPV of 0.96; medication data-based clinical algorithm, SN of 0.89, SP of 0.98, PPV of 0.91 and NPV of 0.98; selected keyword concepts from free-text notes, SN of 0.73, SP of 0.93, PPV of 0.70 and NPV of 0.93; ML algorithm based on free-text notes, SN of 0.95, SP of 0.98, PPV of 0.94 and NPV of 0.99. Various performance of the combined clinical and ML algorithms is also shown in table 2.

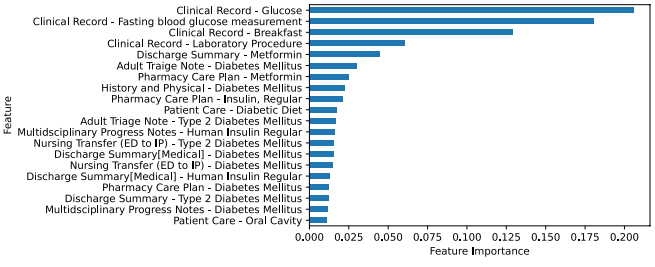


Figure 3 Top 20 document type—concept predictive features selected from the all-document types supervised free-text XGBoost algorithm. Consistent diabetes related terminologies were identified from multiple document types.

Top features from all document type ML model

Figure 3 presents the top 20 pairs of document type and feature from the free-text ML algorithm in the chart review cohort. Calibration plot of the free text (ie, all documents; CUI and XGBoost) is shown in online supplemental appendix figure 1. Grid search space and best hyperparameters are shown in online supplemental appendix table 2. The confusion matrix for the free-text XGBoost algorithm on testing dataset is shown in online supplemental appendix table 3.

Amongthe top 20 features of the XGBoost model, the most influential contributors to classifying individuals with a diagnosis were any glucose documentation or fasting blood glucose measurement recorded within SCM inpatient settings. Following closely was the mention of 'breakfast', in the free-text notes. Other captured top features included text of diabetes, medication administration (eg, metformin, insulin) and diabetic diet. Several document types consistently captured predictor variables or features.

DISCUSSION

This study explored various EMR data-based case definitions for diabetes, uncovering algorithms with excellent performance. We used chart review labels as our gold standard. While the validated administrative data-based ICD-code algorithm demonstrated strong performance, the findings support our hypothesis that harnessing free-text notes can yield comparable or superior results to existing standard methods. The ML algorithm that included all document types of free-text notes was the top performer in this study cohort, with 0.95 SN and 0.94 PPV. Meanwhile, the combination of free-text algorithm, medication, and ICD codes improved the SN to 0.97 but experienced a decline in PPV to 0.87.

The current operational standards for defining diabetes for surveillance (ie, NDSS)[12] and research purposes in Canada were shaped by the administrative data-based ICD code algorithm.[13] These methodologies rely on the utilisation of ICD-code databases, and rely on readily available standardised ICD-code databases, like the DAD, established at both national and international settings. In the Canadian context, these DAD records are reliant on the quality of ICD codes produced by the trained coders who review the charts. Diabetes is a chronic condition which is heavily emphasised for ICD coding in Alberta, and yet the algorithms that solely use these codes resulted in a lower SN compared with the free-text algorithm. This discrepancy stems from the fact that ICD coders primarily review physician documentations from free-text documents within the EMR system for ICD coding in Canada, as dictated by the system design. Challenges and limitations encountered in ICD coding have been described in previous studies[24] indicating the information overload experienced by the healthcare system and workers in various areas when dealing with EMR data.

A recent scoping review highlighted that diabetes definitions typically incorporate laboratory and medications data, along with ICD codes.[8] Laboratory data typically employ values surpassing specific clinical thresholds to determine disease status. When a patient is being treated with antihyperglycaemic medication, these clinical values are presumed not reach that threshold due to the medication's effect. In our study, the combined clinical diagnosis algorithm of laboratory and medication had a 0.90 SN and 0.80 PPV, which is comparable to algorithms described in the above-mentioned review. In a systematic review[25] on the applications of NLP in diabetes care showed that out of 38 studies, 17 aimed to define diabetes, but most of these studies relied on single concept words or keyword-based definitions (ie, diabetes). In our cohort, the keyword algorithm had an 0.73 SN and 0.70 PPV, potentially reflecting the quality of documentation or the practice of data being entered into the EMR from the front end. Figure 3 showed that several consistent diabetes related medication terminologies (eg, metformin and insulin) were captured across multiple EMR document types. The ML-based algorithm which included all types of free-text documents performed the best in this study cohort, achieving a SN of 0.95 and PPV of 0.94 PPV, raising several important considerations. The ICD code algorithm had an 0.84 SN and 0.93 PPV. Combined algorithms often increased SN but reduced PPV, which was expected.

EMR systems, such as SCM[1] and Connect Care (Alberta's newly implemented province-wide clinical information system),[26] based on Epic software (Madison, WI), typically have a front-end graphical user interface for delivering clinical care. It is important to note that not all healthcare workers or providers have access to complete patient charts, and access is typically determined based on assigned roles in the system. Information overload from EMR data can occur if too much information is given,[27] and communication oversight could arise if insufficient information is provided.[28] Additionally, the quality of clinical notes documentation can be heavily influenced by interactions between the care providers and patients or their family members, potentially triggering varying sets of orders and interventions documented in the EMR system. This project extracted all free-text notes from the back end of the EMR system and processed these documents using a standardised medical terminology dictionary (ie, UMLS). Our findings demonstrated that various types of healthcare workers and providers are documenting similar medical concepts across multiple EMR document types for diabetes. Therefore, analysing the commonality in documentation across roles to consolidate and centralise information for shared awareness would enhance information flow in clinical care settings and improve downstream processes, such as improving the quality of the administrative health databases.

Current diabetes definitions based on ICD-code databases are not integrated into clinical practices within the Canadian context, as DAD coding systems and EMR systems operate separately from each other. Alberta's Connect Care clinical information system which includes EPIC-based EMR infrastructure, now in operations throughout AHS operated acute care and ambulatory facilities, has the capacity of integrating ML models,[29] with potential outputs incorporated into dashboards. The integration of inpatient data-specific case definitions could facilitate easier identification of comorbidities, designing automated risk prediction algorithms within EMR which could be implemented into point of care as needed. As EMR adoption in Canada continues to rise,[4] the implementation of EMR data-based diabetes case definitions from both inpatient and outpatient care[30] has the potential to enhance the quality of DAD data for diabetes. This, from a research operations standpoint, could assist with cohort selection for epidemiological and clinical studies. The subsequent improvement in DAD will, in turn, enhance the surveillance capabilities of the NDSS for Alberta in the long run.

This study is not without limitations. First, as we used a single geographic setting, external validation from a different geographical setting is needed. Second, our algorithms do not differentiate between type 1 and type 2 diabetes, the two most common forms of diabetes. With the prevalence of both types increasing, as well as differences in management and care, differentiating between these types is important, this will be an area of future work. Also, we appreciate the immaturity of the proposed application in real-life practice but importantly this study is foundational work for ML in healthcare systems. We appreciate the limited interpretability by the prediction model. Importantly, in our study, we demonstrated the explainbility by showing that top features (figure 3) are coinciding with what is documented within clinical practices. This strengthens the application of our model in real-world practice. We also appreciate the lack of system infrastructure to implement models with existing EMRs

not having the capacity to implement designed ML models. AHS has recently implemented EPIC-based clinical information system, which has the capacity to integrate ML models into EMR systems, in AHS-operated and partner acute and subacute care sites, ambulatory care locations, clinical lab services and diagnostic imaging areas. That being said, our study includes many strengths. Strengths include taking a multimodal EMR data approach to develop a case definition for diabetes and comparing to existing standards, integrating ML and NLP onto EMR data, and using the randomly selected chart review data as the gold standard.

Our future studies will expand to include Connect Care data and eventually validate this work in other jurisdictions. Furthermore, we will evaluate the implementation of our ML models into existing clinical information systems. Recent advancements in large language models have shifted the interest in developing such models for eventual deployment in healthcare systems from the NLP field perspective. While we acknowledge that we have not considered these deep learning NLP models for this study, a future study is in the planning stages, aiming to explore large language model methods on a study cohort with a much larger disease prevalence.[10]

CONCLUSION

As NLP techniques are advancing, there is the potential to leverage them in healthcare, particularly for using free text data within EMRs. As such, we assessed several algorithms and found the free-text algorithm performed the best in this cohort. Determining the ideal algorithm or combinations for implementation would be dependent on the needs, the clinical practice culture and data availability. These types of inpatient EMR-based algorithms for case identification are ideal for timely care delivery and resource planning.

Contributors SL, HQ and SB conceptualised this study. CAE and DAS provided the chart review data. EAM and SL conducted data extraction and linkage. SL conducted analysis. EAM and JP assisted with analysis. DJTC and AAS refined the laboratory and medications algorithm and reviewed the medications list. SL and SB drafted the manuscript. All authors reviewed the contents of the manuscript. SL, HQ, and SB are the guarnators of this study.

Funding This work was supported by Canadian Institutes of Health Research, Foundation Grant FDN-167272, awarded to HQ.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Conjoint Health Research and Ethics Board of University of Calgary (REB15-0790, and approved 7 May 2023).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Restrictions apply to the availability of these data. Data were obtained from Alberta Health Services and are available with the permission of Alberta Health Services.

ORCID iDs

Seungwon Lee http://orcid.org/0000-0002-6532-5303
Danielle A Southern http://orcid.org/0000-0002-0006-0033

REFERENCES

1 Lee S, Xu Y, D Apos Souza AG, *et al*. Unlocking the potential of electronic health records for health research. *Int J Popul Data Sci* 2020;5:1123.
2 CIHI. Discharge abstract database metadata (DAD). 2023. Available: https://www.cihi.ca/en/discharge-abstract-database-metadata-dad
3 O'Malley KJ, Cook KF, Price MD, *et al*. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005;40:1620–39.
4 Chang F, Gupta N. Progress in electronic medical record adoption in Canada. *Can Fam Physician* 2015;61:1076–84.
5 Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019;6:94–8.
6 Lee S, Doktorchik C, Martin EA, *et al*. Electronic medical record–based case Phenotyping for the Charlson conditions: scoping review. *JMIR Med Inform* 2021;9:e23934.
7 Services AH. Patient care information system 2023. 2023. Available: https://albertahealthservices.ca/webapps/elearning/CB/Inpatient/a001_topic_1_introduction_to_clinibase_patient_care_information_system.html
8 Lee S, Li B, Martin EA, *et al*. CREATE: a new data resource to support cardiac precision health. *CJC Open* 2021;3:639–45.
9 Eastwood CA, Southern DA, Khair S, *et al*. Field testing a new ICD coding system: methods and early experiences with ICD-11 beta version 2018. *BMC Res Notes* 2022;15:343.
10 De Coster C, Quan H, Finlayson A, *et al*. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res* 2006;6:77.
11 Escorpizo R, Kostanjsek N, Kennedy C, *et al*. Harmonizing WHO's International Classification of Diseases (ICD) and International Classification of Functioning, Disability and Health (ICF): importance and methods to link disease and functioning. *BMC Public Health* 2013;13:742.
12 LeBlanc AG, Jun Gao Y, McRae L, *et al*. At-a-glance - twenty years of diabetes surveillance using the Canadian chronic disease surveillance system. *Health Promot Chronic Dis Prev Can* 2019;39:306–9.
13 Quan H, Li B, Couris CM, *et al*. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol* 2011;173:676–82.
14 Canada Go. Drug product database. 2023. Available: https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/drug-product-database.html
15 Punthakee Z, Goldenberg R, *et al*, Diabetes Canada Clinical Practice Guidelines Expert Committee. Classification and diagnosis of diabetes, prediabetes and metabolic syndrome. *Can J Diabetes* 2018;42 Suppl 1:S10–5.
16 Morris DH, Khunti K, Achana F, *et al*. Progression rates from Hba1C 6.0-6.4% and other prediabetes definitions to type 2 diabetes: a meta-analysis. *Diabetologia* 2013;56:1489–93.
17 Sherwani SI, Khan HA, Ekhzaimy A, *et al*. Significance of Hba1C test in diagnosis and prognosis of diabetic patients. *Biomark Insights* 2016;11:95–104.
18 Nunes JPL, DeMarco JP. A 7.0-7.7% value for Glycated Haemoglobin is better than a &Amp;Amp;Amp;Amp;Amp;Lt;7% value as an appropriate target for patient-centered drug treatment of type 2 diabetes mellitus. *Ann Transl Med* 2019;7:S122.
19 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture,

component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.
20 McInnes BT, Pedersen T, Carlis J, eds. Using UMLS concept unique Identifiers (Cuis) for word sense Disambiguation in the BIOMEDICAL domain. AMIA annual symposium proceedings; American Medical Informatics Association, 2007
21 Martin EA, D'Souza AG, Lee S, *et al*. Hypertension identification using inpatient clinical notes from electronic medical records: an explainable, data-driven algorithm study. *CMAJ Open* 2023;11:E131–9.
22 Kursa MB, Rudnicki WR. Feature selection with the Boruta package. *J Stat Softw* 2010;36:1–13.
23 Chen T, He T, Benesty M, *et al*. Xgboost: extreme gradient boosting [R package version 04-2]. 2015;1:1–4.
24 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:E617–22.
25 Turchin A, Florez Builes LF. Using natural language processing to measure and improve quality of diabetes care: a systematic review. *J Diabetes Sci Technol* 2021;15:553–60.
26 Services AH. Connect care. 2023. Available: https://www.albertahealthservices.ca/cis/cis.aspx
27 Nijor S, Rallis G, Lad N, *et al*. Patient safety issues from information overload in electronic medical records. *J Patient Saf* 2022;18:e999–1003.
28 Tiwary A, Rimal A, Paudyal B, *et al*. Poor communication by health care professionals may lead to life-threatening complications: examples from two case reports. *Wellcome Open Res* 2019;4:7.
29 Sendak M, Gao M, Nichols M, *et al*. Machine learning in health care: a critical appraisal of challenges and opportunities. *EGEMS (Wash DC)* 2019;7:1.
30 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.

Appendix Table 1. List of diabetes medication, types, and DIN

Diabetes Drug Type	Drug Identification Number (DIN)
Short-acting insulin	Regular insulin (Humulin and Novolin) 01962639, 00889113, 01962655, 00889105, 00795879, 01959212, 01962647, 00889091, 01962663, 00889121, 00587737, 01959239, 02403447, 02241310, 00586714, 02415089, 01959220, 00733075
Rapid-acting insulins	Insulin aspart 02460408, 02460416, 02460424, 02520974, 02520982, 02265435, 02244353, 02245397, 02377209
	Insulin glulisine 02279460, 02279479, 02279487, 02294346
	Insulin lispro (Humalog) 02229704, 02470152, 02229705, 02403412, 02439611, 02240294, 02403420, 02240295, 02240297, 02403439, 02241283, 02469898, 02469901, 02469871, 02506564, 02506572, 02529254
Intermediate-acting insulin	Insulin isophane (Humulin N, Novolin N) 01962639, 00889113, 01962655, 00889105, 00795879, 01959212, 01962647, 00889091, 01962663,

	00889121, 01959239, 02403447, 02024306, 02024217, 02025248, 02024314, 02024322, 02024225, 02024268, 02024446, 02024403, 02024322, 02024225, 02024268, 02024233, 02024284, 02024314
Long-acting insulins	Insulin degludec 02467860, 02467879, 02467887, 02474875, 02474875
	Insulin detemir 02412829, 02271842
	Insulin glargine 02444844, 02461528, 02245689, 02251930, 02294338, 02526441, 02478293, 02478293, 02493373, 02441829
Combination insulins	Humalog Mix 75/25 (insulin lispro protamine-insulin lispro) 02240294, 02403420, 02240295
	Humalog Mix 50/50 (insulin lispro protamine-insulin lispro) 02240297, 02403439

	Humulin 70/30 (human insulin NPH-human insulin regular) 00795879, 01959212
	Novolin 70/30 02024217, 02025248
Biguanides	Metformin (Glucophage, Metformin Hydrochloride ER, Glumetza, Riomet, Fortamet) 02099233, 02162849, 02446065, 02162822, 02229517, 02231389, 02238827, 02242793, 02242794, 02246965, 02284782, 02284790, 02343606, 02343614, 02353377, 02353385, 02378841, 02378868, 02385341, 02385368, 02268493, 02300451, 02268507 Metformin-alogliptin (Kazano) 02417219, 02417227, 02417235
	Metformin-canagliflozin (Invokamet) 02455404, 02455412, 02455420, 02455439, 02455447, 02455455, 02477394, 02477408, 02477416, 02477424
	Metformin-dapagliflozin (Xigduo XR) 02449935, 02449943

	Metformin-empagliflozin (Synjardy) 02456575, 02456583, 02456591, 02456605, 02456613, 02456621
	Metformin-linagliptin (Jentadueto) 02403250, 02403269, 02403277
	Metformin-rosiglitazone (Avandamet) 02247085, 02247086, 02247087, 02248440, 02248441
	Metformin-saxagliptin (Kombiglyze XR) 02389169, 02389177, 02389185
	Metformin-sitagliptin (Janumet) 02333864, 02333872, 02416786, 02416794, 02416808
Dipeptidyl peptidase-4 (DPP-4) inhibitors	Alogliptin (Nesina) 02417189, 02417197, 02417200
	Alogliptin-metformin (Kazano) 02417219, 02417227, 02417235
	Alogliptin-pioglitazone (Oseni)

	02419300, 02419319, 02419327, 02419335, 02419343, 02419351
	Linagliptin (Tradjenta) 02370921
	Linagliptin-empagliflozin (Glyxambi) 02459752, 02459760
	Linagliptin-metformin (Jentadueto) 02403250, 02403269, 02403277
	Saxagliptin (Onglyza) 02333554, 02375842
	Sexagliptin (Apo-Sexagliptin) 02507471, 02507498,
	Sexagliptin (Sandoz-Sexagliptin) 02468603, 02468611
	Saxagliptin-metformin (Kombiglyze XR) 02389169, 02389177, 02389185
	Sitagliptin (Januvia) 02303922, 02388839, 02388847

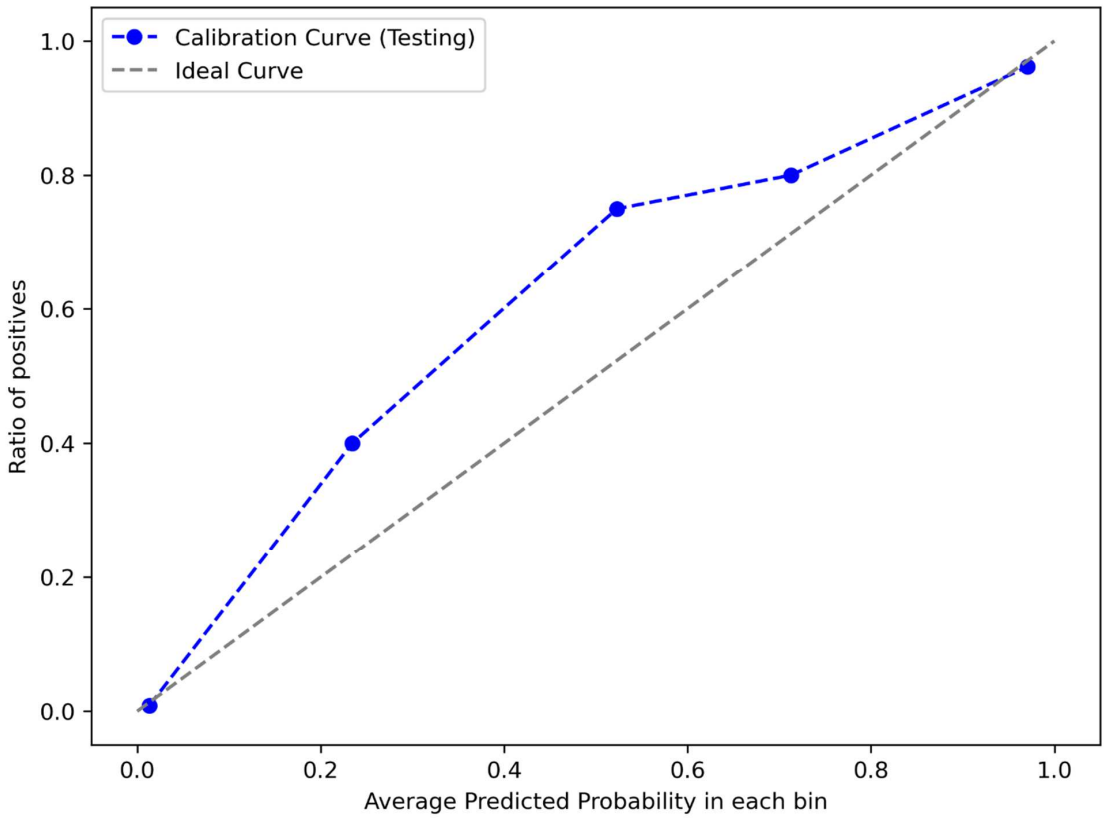
	Sitagliptin-metformin (Janumet and Janumet XR) 02333856, 02333864, 02333872, 02416786, 02416794, 02416808
Glucagon-like peptide-1 receptor agonists (GLP-1 receptor agonists)	Dulaglutide (Trulicity) 02448572, 02448580, 02448599, 02448602, 02530163, 02530171
	Exenatide (Byetta) 02361809, 02361817
	Exenatide extended release (Bydureon) 02448610, 02483203
	Liraglutide 02437899, 02351064, 02474875
	Semaglutide 02471469, 02471477, , 02523930, 02497581, 02497603, 02497611, 02522551, 02522578, 02522586, 02522594, 02522608, 02528509, 02528517, 02528525, 02528533, 02528541
GLP-1 insulin combinations	(Xultophy, Soliqua) 02474875, 02478293
Meglitinides	Nateglinide (Starlix)

	02245439, 02245440, 02245438
	Repaglinide 02239926, 02239925, 02239924, 02355663, 02355671, 02355698, 02424258, 02424266, 02424274, 02321475, 02321483, 02321491, 02354926, 02354934, 02354942, 02357453, 02357461, 02357488
Alpha-glucosidase inhibitor	Acarbose 02493780, 02493799, 02190885, 02190893, 02494078, 02494086
Sodium-glucose transporter (SGLT) 2 inhibitors	Dapagliflozin (Farxiga) 02435462, 02435470
	Dapagliflozin-metformin (Xigduo XR) 02449935, 02449943
	Canagliflozin (Invokana) 02425483, 02425491
	Canagliflozin-metformin (Invokamet) 02455404, 02455412, 02455420, 02455439, 02455447, 02455455, 02477394, 02477408, 02477416, 02477424
	Empagliflozin (Jardiance)

	02443937, 02443945
	Empagliflozin-linagliptin (Glyxambi) 02459752, 02459760
	Empagliflozin-metformin (Synjardy) 02456575, 02456583, 02456591, 02456605, 02456613, 02456621
	Ertugliflozin (Steglatro) empagliflozin-metformin (Synjardy) 02456575, 02456583, 02456591
Sulfonylureas	Glimepiride 02245272, 02245273, 02245274, 02269589, 02269597, 02269619
	Glimepiride-rosiglitazone (Avandaryl) 02258781, 02258803, 02258811
	Gliclazide 02483300, 02483319, 02245247, 02297795, 02407124, 02363518, 00765996, 02242987, 02356422, 02248210, 02287072, 02155850, 02248453, 02429764, 02429772, 02423286, 02423294, 02229519, 02438658, 02449765, 02336316, 02294400, 02254719, 02461323, 02461331, 02439328, 02463571, 02238103

	Glyburide 01913654, 01913662, 02234514, 00720941, 01959352 02350459, 02350467, 02485664, 02236734, 01913670, 01913689 02224550, 02224569, 00012599, 00454753, 01987836, 01987534
	Chlorpropamide 00399302, 00312711, 00024708, 00024716
	Tolbutamide 00013889, 00021849, 00012602, 00012610, 00312762, 00156663, 00431168
Thiazolidinediones	Rosiglitazone 02403366, 02403374, 02403382, 02241112, 02241113, 02241114
	Rosiglitazone-glimepiride (Avandaryl) 02258781, 02258803, 02258811
	Pioglitazone 02339587, 02339595, 02391600 , 02302861, 02302888, 02302896 , 02374587, 02374595, 02302942, 02302950 , 02302977, 02365529, 02365537, 02397307, 02326477, 02326485, 02326493, 02303124, 02303132, 02303140, 02389290,

	02389304, 02389312, 02242572, 02242573, 02242574
	Pioglitazone-alogliptin (Oseni) 02419300, 02419319, 02419327, 02419335, 02419343, 02419351


Calibration Curve (Testing)
Ideal Curve
Ratio of positives
Average Predicted Probability in each bin
0.0
0.2
0.4
0.6
0.8
1.0

Appendix Table 2. Grid Search Space and Best Hyperparameters of XGBoost Model.

Factor	Value
Lambda	0, 0.3, 0.5, 0.7, 1
Alpha	0, 0.5, 1
Learning Rate	0.01, 0.03, 0.05, 0.10
Objective	Binary: logistic
Max Depth	1,3,5,7
Min Child Weight	1,3,5
Subsample	0.5, 0.7, 1
Colsample	0.7
Best Hyperparameter	
Factor	Value
Lambda	0.7
Alpha	0.5
Learning Rate	0.03
Objective	Binary: logistic
Max Depth	3
Min Child Weight	1
Subsample	1

Appendix Table 3. Confusion Matrix for the Free-text XGBoost Algorithm on the Testing Dataset (n=609).

	Predicted Positive	Precited Negative
True Positive	111	6
True Negative	7	485