Original Paper

Identification of Prediabetes Discussions in Unstructured Clinical Documentation: Validation of a Natural Language Processing Algorithm

Jessica L Schwartz[1,2*], MD, MHS; Eva Tseng[1,3*], MD, MPH; Nisa M Maruthur[1,3,4*], MD, MHS; Masoud Rouhizadeh[5,6*], MS, PhD

[1]Division of General Internal Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States
[2]Division of Hospital Medicine, Johns Hopkins Hospital, Baltimore, MD, United States
[3]Welch Center for Prevention, Epidemiology, & Clinical Research, Johns Hopkins University, Baltimore, MD, United States
[4]Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, United States
[5]Department of Pharmaceutical Outcomes and Policy, University of Florida College of Pharmacy, Gainesville, FL, United States
[6]Division of Biomedical Informatics and Data Science, Johns Hopkins University School of Medicine, Baltimore, MD, United States
[*]all authors contributed equally

Corresponding Author:
Jessica L Schwartz, MD, MHS
Division of General Internal Medicine
Johns Hopkins School of Medicine
2024 E Monument St.
Ste 2-604D
Baltimore, MD, 21205
United States
Phone: 1 973 722 8552
Fax: 1 410 955 0476
Email: jschwa64@jhmi.edu

Abstract

Background: Prediabetes affects 1 in 3 US adults. Most are not receiving evidence-based interventions, so understanding how providers discuss prediabetes with patients will inform how to improve their care.

Objective: This study aimed to develop a natural language processing (NLP) algorithm using machine learning techniques to identify discussions of prediabetes in narrative documentation.

Methods: We developed and applied a keyword search strategy to identify discussions of prediabetes in clinical documentation for patients with prediabetes. We manually reviewed matching notes to determine which represented actual prediabetes discussions. We applied 7 machine learning models against our manual annotation.

Results: Machine learning classifiers were able to achieve classification results that were close to human performance with up to 98% precision and recall to identify prediabetes discussions in clinical documentation.

Conclusions: We demonstrated that prediabetes discussions can be accurately identified using an NLP algorithm. This approach can be used to understand and identify prediabetes management practices in primary care, thereby informing interventions to improve guideline-concordant care.

(JMIR Med Inform 2022;10(2):e29803) doi: 10.2196/29803

KEYWORDS

prediabetes; prediabetes discussions; prediabetes management; chronic disease management; physician-patient communication; natural language processing; machine learning

Introduction

Prediabetes affects 88 million US adults [1,2], and evidence-based interventions focusing on lifestyle modification can prevent type 2 diabetes [3-12]. In particular, the Diabetes Prevention Program is an effective lifestyle intervention that decreases diabetes incidence, with the most recent data showing a 27% risk reduction compared with the placebo arm over 15 years of follow up [5]. Unfortunately, up to 89% of patients do not know they have prediabetes [13,14], and many patients are unaware of interventions to decrease their risk of diabetes—relying on their primary care providers (PCPs) to initiate discussions about diabetes prevention, including the importance of lifestyle changes [8,9]. However, survey data demonstrate that many providers feel that they lack the resources to effectively implement evidence-based prediabetes treatment [8,9]. Focused primary care interventions to support decision-making and education may be able to improve diagnosis of prediabetes and delivery of guideline-concordant care.

Rigorous quality improvement interventions require evaluation using measurement before and after implementation of a project to determine whether there is a demonstrable change in target outcomes. Unfortunately, it is difficult to identify changes and improvement in prediabetes management through structured data alone. Relying on diagnosis codes is insufficient; one study showed that only 13% of patients with prediabetes had an International Classification of Diseases (ICD)-9 diagnosis of prediabetes or hyperglycemia [14]. Although labs, orders, and referrals provide some insight, this information lacks detail about management, particularly lifestyle counseling, which is better captured in narrative documentation. This content is not easily queried and requires innovative research methods to accurately reflect delivery of prediabetes care.

Prior studies have shown that natural language processing (NLP) can be used to diagnose chronic conditions, like diabetes, but few focus on disease management [15]. Similarly, NLP studies in prediabetes have primarily focused on disease detection, screening, and predictive modeling, with no studies applying machine learning (ML) techniques to determine how prediabetes is managed [16-27]. Our goal was to develop a method to identify when providers discuss prediabetes management and treatment, which could later be used to determine if care delivered meets evidence-based guidelines and compare outcomes before and after an intervention. Therefore, we developed and validated NLP pipelines to identify primary care discussions about prediabetes in clinical documentation.

Methods

Population and Ethics Approval

We identified patients with prediabetes who had an internal medicine primary care visit within an academic center with multiple ambulatory locations in Maryland and Washington, DC. Eligible patients were adults (≥18 years old) covered by 1 of 3 major insurers who completed an in-person visit and had a hemoglobin A_{1c} (HbA_{1c}) level between 5.7% and 6.4% between July 1, 2016 and December 31, 2018. Patients with diabetes (any type) based on billing codes or documentation in the problem list or past medical history were excluded. Data cleaning and analyses were performed using Stata 15. This study was approved by the Johns Hopkins Institutional Review Board (IRB00196984).

Keyword Search Refinement (Phase 1)

Based on clinical experience, we developed a list of keywords used to describe "prediabetes" (Table S1 in Multimedia Appendix 1). We identified visit notes containing these keywords using Python string matching and dictionary look-up, accounting for variations like spelling errors and morphological differences. We extracted a ±25-word concordance window ("note snippet") for each match to provide textual context. Multiple snippets could come from the same note if multiple matching keywords were present.

We selected 2 ambulatory clinics from our overall population. Of 315 patients meeting inclusion criteria, 40.6% (128/315) had at least one matching keyword during the study period. These patients had a total of 637 keyword matches across 324 encounters with 25 providers. We conducted manual annotation to determine which of the 637 note snippets represented true clinical discussions of prediabetes (yes or no). Outpatient provider documentation typically includes chief complaint, history of present illness, medical and family history, objective data including physical exam, and an assessment and plan. We considered use of a section identification pipeline to exclude specific sections of the notes (eg, past medical history) in which keywords would not represent prediabetes discussions. However, section identification pipelines are less generalizable, and the providers in our sample did not use standardized templates, making section boundaries difficult to define [28]. Instead, note snippets were designated "no" during manual review if the keyword was only present in past medical history, a list of diagnoses outside of the assessment and plan, family history, or the description of a lab result.

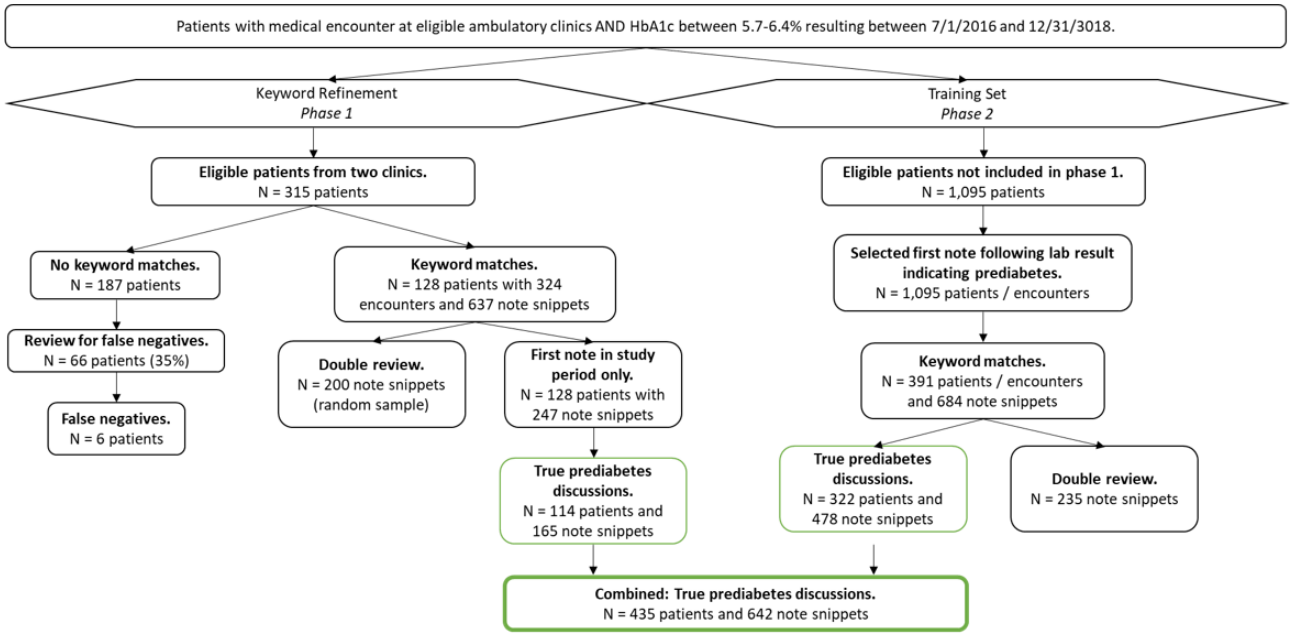
We double-reviewed a random sample of 200 note snippets. Interrater reliability (IRR) was 95%. Discrepancies between annotators were resolved via consensus to refine the definition of "prediabetes discussion." We then manually reviewed patient records for 35.3% (66/187) of charts without a keyword match to identify false negatives. We reviewed all notes written by the patient's PCP within the inclusion timeframe, and 9% (6/66) of patients had prediabetes discussions that were not captured. We added 3 keywords ("dysglycemia," "hyperglycemia," and "pre diabetes") to the lexicon (Table S1 in Multimedia Appendix 1).

Training Set (Phase 2)

We developed a training set to test our prediabetes lexicon against patients from clinics not included in phase 1 (Figure 1). We included a single note per patient (n=1095), choosing the first encounter after the HbA_{1c} result that met inclusion criteria. We applied the finalized keyword search, which resulted in 684 matches for 381 patients seen by 73 providers. We abstracted the 684 note snippets and annotated the notes using a similar process as above. We double-reviewed 34% of the note snippets

with an IRR of 97% for manual annotation, resolving to 100% agreement upon review. We combined these results with note snippets from phase 1. To avoid overselection of a single patient or provider, we included note snippets from 1 encounter per patient for a total of 930 note snippets written by 96 unique providers.

Figure 1. Diagram depicting selection and review during keyword search refinement (Phase 1) and training set development (Phase 2). Eligible patients were adults (≥18 years old) covered by 1 of 3 major insurers who completed an in-person visit at a Johns Hopkins clinic and had an HbA_{1c} level between 5.7% and 6.4% (39-46 mmol/mol) between July 1, 2016 and December 31, 2018. Note, double review indicates that 2 providers reviewed the keyword matches to identify whether the surrounding text represented a true prediabetes discussion.



Rule-Based System

Rule-based systems are frequently used for clinical concept extraction and text classification systems because of their ease of implementation and minimal computational requirements. To establish a strong baseline, we tested the feasibility of identifying prediabetes discussions with a rule-based classification scheme. Using the spaCy EntityRuler module [29], we created 42 expert-developed patterns that, if present, would represent prediabetes discussions. The spaCy EntityRuler module facilitates various pattern, keyword, and regular expression searching and matching and allows us to account for morphological variations (eg, singular vs plural forms, conjunctions), as well as substitutions of different prepositions (eg, about vs for) and synonyms (eg, prediabetes, impaired fasting glucose). Table S2 in Multimedia Appendix 1 provides our expert-developed patterns for this rule-based system. We randomly sampled 90% of the note snippets to develop and revise the rule-based system and evaluated the system on the remaining 10%.

Machine Learning

Feature Selection

Note snippets from the training set were stemmed using the Porter stemmer, and common stop words were removed using the Natural Language Toolkit (NLTK) stop word list [30]. We used the Python scikit-learn library [31] to extract word ngram sequences (1-5 grams), weighted by term frequency-inverse document frequency (TF-IDF) [32]. We applied logistic regression with L1 regularization [33] to reduce the dimensionality of the feature vectors.

Computational Environment

Deep learning and ML experiments were conducted on the Johns Hopkins University (JHU) Precision Medicine Analytics Platform (PMAP), a high-performance, cloud-based, big-data platform to accelerate biomedical discovery and translate discovered knowledge to improve patient-centered care. PMAP pulls data from the Johns Hopkins Medicine electronic health record (EHR) to support processing by ML and NLP technologies. Statistical analysis and manual annotation were done in the JHU Secure Analytic Framework Environment, a virtual desktop that provides JHU investigators with a secure platform for analyzing and sharing sensitive data (including protected health information) with colleagues.

Classification

We used the labeled note snippets to train multiple ML classifiers to replicate human annotation for prediabetes discussions. We applied 6 binary classification models: logistic regression [34], linear support vector machines (SVM) [35], stochastic gradient descent (SGD) [36], decision tree [37], random forest [38,39], and Gaussian naïve Bayes (NB) [40]. To reduce overfitting, each model was evaluated using 10-fold cross-validation by training, randomly, on 90% of the data and holding out 10% for testing. All modeling was performed in scikit-learn [31].

We also applied convolutional neural networks (CNNs) for sentence categorization [41], a well-established deep learning method in NLP for text classification [42] using Python spaCy 2.1 implementation [29]. We started with the tokenization of each note snippet and creating an embedding vector of each token using scispaCy large models (~785,000 vocabulary and

600,000 word vectors), pretrained on biomedical and clinical text [43]. Next, to represent the tokens in context, these vectors were encoded into a sentence matrix by computing the vector for each token using a forward pass and a backward pass. After that, a self-attention mechanism was applied to reduce the dimensionality of the sentence matrix representation into a single context vector. Finally, these vectors were average-pooled and used as features in a simple feed-forward network for predicting true discussions of prediabetes. For the CNN model, we used the spaCy 2.2 default network architecture and parameters [44].

For each classification method, we reported on agreement, sensitivity and recall, specificity, positive predictive value and precision, and F measure using manual annotation as the gold standard. To test statistical significance between classification methods, we used MLxtend Python library to perform a 5x2 cross-validation paired *t* test [45]. A *P* value <.05 indicated that we could reject the null hypothesis that both models performed equally to classify prediabetes discussions.

Results

We identified 1410 patients with prediabetes; 518 (36.74%) had at least one keyword match. Among these patients, 435 (84.0%) had a true discussion about prediabetes in the manually reviewed documents (Figure 1).

The rule-based system was inadequate for replicating human performance, with 72.5% recall and 42.6% specificity (Table 1). ML and CNN classification, however, were close to human performance across all models (Table 1). When comparing conventional classifiers with logistic regression (which had the highest agreement), only linear SVM and NB had similar performance (*P*=.11 and *P*=.15, respectively). CNN outperformed all conventional ML classifiers (logistic regression: *P*=.04; SVM: *P*=.02; SGD: *P*=.002; random forest: *P*=.002; decision tree: *P*=.001; NB: *P*=.03).

Table 1. Performance of machine learning methods to approximate manual annotation in identifying prediabetes discussions from primary care note snippets (n=930).

Method	Instances classifier agreed with manual annotation, n (%)	Sensitivity/recall	Specificity	PPV[a]/precision	F measure
Rule-based system					
Expert-developed patterns	588 (63.2)	0.725	0.426	0.737	0.731
Machine learning					
Logistic regression	885 (95.2)	0.966	0.921	0.965	0.965
Linear support vector machines	878 (94.4)	0.962	0.903	0.957	0.960
Stochastic gradient descent	858 (92.3)	0.926	0.915	0.96	0.943
Random forest	863 (92.8)	0.961	0.854	0.937	0.948
Decision tree	832 (89.5)	0.923	0.83	0.925	0.924
Gaussian naïve Bayes	883 (95.0)	0.966	0.912	0.96	0.963
Convolutional neural networks	910 (97.9)	0.984	0.966	0.984	0.984

[a]PPV: positive predictive value.

Manual annotation revealed a variety of linguistic patterns that did and did not represent clinical discussions of prediabetes (Table 2). Most commonly, true discussions were found in the assessment and plan, and those that did not were auto populated from structured fields. ML did result in 5% misclassification based on logistic regression, the best performing conventional classifier; a pattern was not apparent on review of these misclassified note snippets.

Table 2. Example text from clinical documentation containing keywords matching the "prediabetes" extraction lexicon, stratified by whether the text represents documentation of a prediabetes discussion.

Location in note	Representative text from note snippets[a]
Text containing keyword matches representing prediabetes discussions.	
Chief complaint	• Chief complaint: Patient is a 42 y.o. female here with questions about prediabetes. • Patient presents to the visit for an annual physical and reevaluation of HTN[b] and impaired fasting glucose.
History of Present Illness	• Has a treadmill but not using regularly. Recent a1c was 6.2 consistent with pre-diabetes.
Visit Problem List	• Problem List Items Addressed This Visit Asthma Borderline diabetes Essential hypertension • Assessment Order Plan 1. Hyperlipidemia ... 7. Impaired fasting glucose 8. Health care maintenance
Assessment & Plan	• Hyperglycemia Lifestyle modification including diet and exercise discussed. 6. Elevated blood pressure. • Pre-diabetes Assessment: recent A1C in good range. Plan: exercise and healthy food changes.
Text containing keyword matches not representing prediabetes discussions.	
One-liner	• Patient with history of HTN, HLD[c], prediabetes, scleroderma here for routine health assessment.
Past Medical History	• Past Medical History: Diagnosis Date Asthma 5/14/2008 ... Prediabetes 2/6/2012 Osteoporosis 5/14/2008
Problem List	• ... Hyperlipidemia E78.5 Impaired fasting glucose R73.01 Overweight E66.3 ...
Diagnosis list	• Diagnoses of Essential hypertension, Osteoporosis, ..., Prediabetes, Asthma, ...
Family history	• Family History Problem Relation Age of Onset Diabetes Father Prediabetes Paternal Grandfather...
Pertinent positive	• Diagnosis remains unclear. He has prediabetes. Reports 2-3 months of intermittent palpitations.
Pertinent negative	• Likely has peripheral neuropathy. Negative RPR[d], HIV, pre-diabetes.
Follow up reasons	• Follow up in 1 month for flu shot and prediabetes discussion.
Results[e]	• For someone without known diabetes, a hemoglobin A_{1c} value between 5.7 % and 6.4 % is consistent with prediabetes and should be confirmed.
General guidelines[e]	• Type 2 diabetes or prediabetes All men beginning at age 45 and men without symptoms at any age who are overweight or obese and have 1 or more other risk factors.

[a]Text was modified for length and content to serve as general examples while protecting patient anonymity.

[b]HTN: hypertension.

[c]HLD: hyperlipidemia.

[d]RPR: rapid plasma reagin.

[e]Populated in notes from clinical decision support tools.

Discussion

Principal Findings

We utilized NLP and ML techniques to identify prediabetes discussions from unstructured narrative documentation with up to 98% precision and recall. To date, NLP techniques have been used in prediabetes for screening, diagnosis, risk stratification, predictive modeling, and intervention design [16-27,46-50]. To our knowledge, this is the first NLP tool to identify prediabetes discussions. NLP methods have been applied in health care in many ways including in EHR free-text clinical notes to classify disease phenotype, with most studies using simple methods like shallow classifiers or combined with rule-based methods [15,51]. Compared with these studies, our NLP methods are not novel, but our application to disease management distinguishes our study from those that primarily focus on condition identification for chronic diseases [15].

In our study, a simple rule-based system was inadequate to identify prediabetes discussions due to poor specificity. In contrast, all ML methods performed well, with 89% to 98% accuracy. This result demonstrates that prediabetes discussions, despite a variety of documentation styles, can be identified using NLP pipelines. Logistic regression, an efficient conventional classifier with minimal technical dependencies, was statistically outperformed by CNN, a deep learning technique. However, both identified >95% of prediabetes discussions, suggesting that either method could be applied depending on system needs.

Our NLP tool has multiple applications. The simplicity of logistic regression allows for deployment in operational settings,

particularly clinical decision support. The tool can also simplify the analytic process before and after a clinical intervention intended to change provider practices. For example, it can isolate discussions about prediabetes, a task that otherwise requires time-consuming manual review. The context of these discussions could then be reviewed to understand the impact of an intervention. This process would strengthen the evaluation of quality improvement programs for prediabetes to promote guideline-concordant care, which includes lifestyle counseling [3-7]. These methods should be replicable to identify conversations about behavioral interventions for other conditions, such as obesity, polysubstance abuse, or tobacco use, that rely heavily on counseling in addition to medication management and referrals.

Strengths

Our study has several strengths. The keyword refinement stage was rigorous. We validated the initial keyword list against a random sample from 2 ambulatory clinics, ensuring we reviewed a variety of documentation styles. Manual annotation was performed by 2 experts to standardize our definition of "prediabetes discussion," leading to improvement in IRR scores during training set development. We also identified false negatives and revised our initial keyword list accordingly to ensure capture of prediabetes discussions. Finally, we applied the search criteria developed during keyword refinement to a new set of notes from unique clinics to reduce overfitting. There was a total of 96 different providers included in the 930 unique note snippets, which allowed the model to learn the vocabulary and writing styles of many different clinicians.

Limitations

Limitations of our study include collection of data from a single health system. However, the clinics included represent urban and suburban sites serving patients of different socioeconomic levels and disease burden, improving generalizability. Providers at other institutions may use different medical terminology, not considered in this study, to describe "prediabetes." This could limit generalizability outside of the home-trained institution. However, we took several steps to reduce institutional bias, including rigorous keyword refinement and application of the final lexical search to multiple clinics that do not share standardized templates to include many linguistic styles and patterns. We limited our note selection to the first encounter following the abnormal HbA_{1c} result; although this could miss some dialogue about prediabetes, logically these discussions are most likely to occur close to the time of the abnormal result, and this decreased bias in our models. Finally, the note selection process, requiring at least one prediabetes keyword to enter our data set, limited our ability to calculate true recall. We minimized this issue by performing manual review on a subset of the charts that did not enter our data set, to ensure we did not have selection bias in our keyword search. Future studies may consider applying our NLP pipeline against a random sample of notes without requiring keyword selection to perform additional validations. Additionally, our study provides a baseline framework for identifying discussions of prediabetes. Next steps could apply NLP pipelines to identify when discussions about prediabetes meet the threshold for delivery of guideline-concordant care.

Conclusion

Our NLP pipeline successfully identified prediabetes discussions in unstructured notes with precision approximating human annotation. This approach can be used to evaluate prediabetes counseling during patient visits and describe prediabetes management in primary care. Gathering these data is a critical step to inform interventions to improve the delivery of evidence-based prediabetes care to reduce the incidence of type 2 diabetes.

Acknowledgments

This work was supported by the Johns Hopkins Institute for Clinical and Translational Research Core Coins Award 2018. ET was supported by the National Institute of Diabetes and Digestive and Kidney Diseases [K23DK118205]. JLS was supported by the National Heart, Lung, and Blood Institute [5T32HL007180, PI: Hill-Briggs].

Conflicts of Interest

NMM is the co-inventor of a virtual diabetes prevention program. Under a license agreement between Johns Hopkins HealthCare Solutions and the Johns Hopkins University, NMM and the University are entitled to royalty distributions related to this technology. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict of interest policies. This technology is not described in this study. JLS is a co-investigator on a research project funded by NovoNordisk Inc. The primary aim of the project is to create and pilot a clinical decision support tool to assist clinicians when talking to their patients about weight and obesity treatment. This project is not addressed or referenced in this publication.

Multimedia Appendix 1

Supplementary methods and tables.
[DOCX File , 30 KB-Multimedia Appendix 1]

References

1. National Diabetes Statistics Report, 2020: Estimates of Diabetes and Its Burden in the United States. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf [accessed 2022-01-31]

2. National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States. Centers for Disease Control and Prevention. 2014. URL: https://www.cdc.gov/diabetes/data/statistics-report/index.html [accessed 2022-01-31]
3. Diabetes Prevention Program Research Group, Knowler WC, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, et al. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. Lancet 2009 Dec 14;374(9702):1677-1686 [FREE Full text] [doi: 10.1016/S0140-6736(09)61457-4] [Medline: 19878986]
4. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002 Mar 07;346(6):393-403 [FREE Full text] [doi: 10.1056/NEJMoa012512] [Medline: 11832527]
5. Diabetes Prevention Program Research Group. Long-term effects of lifestyle intervention or metformin on diabetes development and microvascular complications over 15-year follow-up: the Diabetes Prevention Program Outcomes Study. Lancet Diabetes Endocrinol 2015 Dec;3(11):866-875 [FREE Full text] [doi: 10.1016/S2213-8587(15)00291-0] [Medline: 26377054]
6. American Diabetes Association. 3. Prevention or Delay of Type 2 Diabetes. Diabetes Care 2020 Jan;43(Suppl 1):S32-S36. [doi: 10.2337/dc20-S003] [Medline: 31862746]
7. National Diabetes Prevention Program. Centers for Disease Control and Prevention. 2017. URL: http://www.cdc.gov/diabetes/prevention/index.htm [accessed 2022-01-31]
8. Tseng E, Greer RC, O'Rourke P, Yeh H, McGuire MM, Clark JM, et al. Survey of primary care providers' knowledge of screening for, diagnosing and managing prediabetes. J Gen Intern Med 2017 Dec;32(11):1172-1178 [FREE Full text] [doi: 10.1007/s11606-017-4103-1] [Medline: 28730532]
9. Tseng E, Greer RC, O'Rourke P, Yeh H, McGuire MM, Albright AL, et al. National survey of primary care physicians' knowledge, practices, and perceptions of prediabetes. J Gen Intern Med 2019 Nov;34(11):2475-2481 [FREE Full text] [doi: 10.1007/s11606-019-05245-7] [Medline: 31502095]
10. Rhee MK, Herrick K, Ziemer DC, Vaccarino V, Weintraub WS, Narayan KMV, et al. Many Americans have pre-diabetes and should be considered for metformin therapy. Diabetes Care 2010 Jan;33(1):49-54 [FREE Full text] [doi: 10.2337/dc09-0341] [Medline: 19808929]
11. Karve A, Hayward RA. Prevalence, diagnosis, and treatment of impaired fasting glucose and impaired glucose tolerance in nondiabetic U.S. adults. Diabetes Care 2010 Dec;33(11):2355-2359 [FREE Full text] [doi: 10.2337/dc09-1957] [Medline: 20724649]
12. Moin T, Li J, Duru OK, Ettner S, Turk N, Keckhafer A, et al. Metformin prescription for insured adults with prediabetes from 2010 to 2012: a retrospective cohort study. Ann Intern Med 2015 May 21;162(8):542-548 [FREE Full text] [doi: 10.7326/M14-1773] [Medline: 25894024]
13. Centers for Disease Control and Prevention (CDC). Awareness of prediabetes--United States, 2005-2010. MMWR Morb Mortal Wkly Rep 2013 Mar 22;62(11):209-212 [FREE Full text] [Medline: 23515058]
14. Schmittdiel JA, Adams SR, Segal J, Griffin MR, Roumie CL, Ohnsorg K, et al. Novel use and utility of integrated electronic health records to assess rates of prediabetes recognition and treatment: brief report from an integrated electronic health records pilot study. Diabetes Care 2014 Mar;37(2):565-568 [FREE Full text] [doi: 10.2337/dc13-1223] [Medline: 24271190]
15. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 May 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
16. De Silva K, Jönsson D, Demmer RT. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. J Am Med Inform Assoc 2020 Mar 01;27(3):396-406 [FREE Full text] [doi: 10.1093/jamia/ocz204] [Medline: 31889178]
17. Chung JW, Kim WJ, Choi SB, Park JS, Kim DW. Screening for pre-diabetes using support vector machine model. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:2472-2475. [doi: 10.1109/EMBC.2014.6944123] [Medline: 25570491]
18. Maeta K, Nishiyama Y, Fujibayashi K, Gunji T, Sasabe N, Iijima K, et al. Prediction of glucose metabolism disorder risk using a machine learning algorithm: pilot study. JMIR Diabetes 2018 Dec 26;3(4):e10212 [FREE Full text] [doi: 10.2196/10212] [Medline: 30478026]
19. Anderson JP, Parikh JR, Shenfeld DK, Ivanov V, Marks C, Church BW, et al. Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. J Diabetes Sci Technol 2015 Dec 20;10(1):6-18 [FREE Full text] [doi: 10.1177/1932296815620200] [Medline: 26685993]
20. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y, et al. Screening for prediabetes using machine learning models. Comput Math Methods Med 2014;2014:618976 [FREE Full text] [doi: 10.1155/2014/618976] [Medline: 25165484]
21. Acciaroli G, Sparacino G, Hakaste L, Facchinetti A, Di Nunzio GM, Palombit A, et al. Diabetes and prediabetes classification using glycemic variability indices from continuous glucose monitoring data. J Diabetes Sci Technol 2018 Jan;12(1):105-113 [FREE Full text] [doi: 10.1177/1932296817710478] [Medline: 28569077]
22. Shankaracharya, Odedra D, Samanta S, Vidyarthi AS. Computational intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India. Rev Diabet Stud 2012;9(1):55-62 [FREE Full text] [doi: 10.1900/RDS.2012.9.55] [Medline: 22972445]

23. Wang L, Mu Y, Zhao J, Wang X, Che H. IGRNet: a deep learning model for non-invasive, real-time diagnosis of prediabetes through electrocardiograms. Sensors (Basel) 2020 May 30;20(9):1 [FREE Full text] [doi: 10.3390/s20092556] [Medline: 32365875]
24. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 2019 Nov 06;19(1):211 [FREE Full text] [doi: 10.1186/s12911-019-0918-5] [Medline: 31694707]
25. Cahn A, Shoshan A, Sagiv T, Yesharim R, Goshen R, Shalev V, et al. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. Diabetes Metab Res Rev 2020 Feb;36(2):e3252. [doi: 10.1002/dmrr.3252] [Medline: 31943669]
26. Garcia-Carretero R, Vigil-Medina L, Mora-Jimenez I, Soguero-Ruiz C, Barquero-Perez O, Ramos-Lopez J. Use of a K-nearest neighbors model to predict the development of type 2 diabetes within 2 years in an obese, hypertensive population. Med Biol Eng Comput 2020 May;58(5):991-1002. [doi: 10.1007/s11517-020-02132-w] [Medline: 32100174]
27. Jin B, Liu R, Hao S, Li Z, Zhu C, Zhou X, et al. Defining and characterizing the critical transition state prior to the type 2 diabetes disease. PLoS One 2017;12(7):e0180937 [FREE Full text] [doi: 10.1371/journal.pone.0180937] [Medline: 28686739]
28. Pomares-Quimbaya A, Kreuzthaler M, Schulz S. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. BMC Med Res Methodol 2019 Jul 18;19(1):155 [FREE Full text] [doi: 10.1186/s12874-019-0792-y] [Medline: 31319802]
29. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolution neural networks and incremental parsing. spacy.io 2017:1 [FREE Full text]
30. Loper E, Bird S. NLTK: the natural language toolkit. 2002 Presented at: ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics; July 7, 2002; Philadelphia, PA URL: https://doi.org/10.3115/1118108.1118117 [doi: 10.3115/1118108.1118117]
31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 2011;12:2825-2830 [FREE Full text] [doi: 10.1145/2786984.2786995]
32. Rouhizadeh M, Jaidka K, Smith L, Schwartz HA, Buffone A, Ungar LH. Identifying locus of control in social media language. 2018 Presented at: Conference on Empirical Methods in Natural Language Processing; October 31 - November 4, 2018; Brussels, Belgium p. 1146-1152 URL: https://www.aclweb.org/anthology/D18-1145.pdf [doi: 10.18653/v1/d18-1145]
33. Park MY, Hastie T. L1 regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 2007;69(4):659-677 [FREE Full text] [doi: 10.1111/j.1467-9868.2007.00607.x]
34. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression, 3rd Edition. Hoboken, NJ: John Wiley & Sons, Ltd; 2013.
35. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;20:273-297 [FREE Full text] [doi: 10.1007/bf00994018]
36. Bottou L. Large-Scale Machine Learning with Stochastic Gradient Descent. In: Lechevallier Y, Saporta G, editors. Proceedings of COMPSTAT'2010. Heidelberg, Germany: Physica-Verlag HD; 2010:177-186.
37. Safavian S, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans. Syst., Man, Cybern 1991;21(3):660-674 [FREE Full text] [doi: 10.1109/21.97458]
38. Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2/3:18-22 [FREE Full text]
39. Breiman L. Random forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
40. Rish I. An Empirical Study of the Naive Bayes Classifier. 2001 Presented at: IJCAI 2001 workshop on empirical methods in artificial intelligence; August 4-6, 2001; Seattle, WA p. 41-46 URL: https://www.cc.gatech.edu/fac/Charles.Isbell/classes/reading/papers/Rish.pdf
41. Johnson R, Zhang T. Semi-supervised convolutional neural networks for text categorization via region embedding. Adv Neural Inf Process Syst 2015 Dec;28:919-927 [FREE Full text] [Medline: 27087766]
42. Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Cornell University. 2015. URL: https://arxiv.org/abs/1510.03820 [accessed 2022-01-31]
43. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Cornell University. 2019. URL: https://arxiv.org/abs/1902.07669 [accessed 2022-01-31]
44. Honnibal M. Embed, encode, attend, predict: The new deep learning formula for state-of-the-art NLP models. Explosion AI. 2016 Nov 9. URL: https://explosion.ai/blog/deep-learning-formula-nlp [accessed 2022-01-31]
45. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. JOSS 2018 Apr;3(24):638. [doi: 10.21105/joss.00638]
46. Hu X, Reaven PD, Saremi A, Liu N, Abbasi MA, Liu H, ACT NOW Study Investigators. Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance. EURASIP J Bioinform Syst Biol 2016 Dec;2016(1):14 [FREE Full text] [doi: 10.1186/s13637-016-0049-6] [Medline: 27642290]
47. Garcia-Carretero R, Vigil-Medina L, Barquero-Perez O, Ramos-Lopez J. Pulse wave velocity and machine learning to predict cardiovascular outcomes in prediabetic and diabetic populations. J Med Syst 2019 Dec 09;44(1):16. [doi: 10.1007/s10916-019-1479-y] [Medline: 31820120]

48. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized nutrition by prediction of glycemic responses. Cell 2015 Dec 19;163(5):1079-1094 [FREE Full text] [doi: 10.1016/j.cell.2015.11.001] [Medline: 26590418]
49. Popp CJ, St-Jules DE, Hu L, Ganguzza L, Illiano P, Curran M, et al. The rationale and design of the personal diet study, a randomized clinical trial evaluating a personalized approach to weight loss in individuals with pre-diabetes and early-stage type 2 diabetes. Contemp Clin Trials 2019 Apr;79:80-88. [doi: 10.1016/j.cct.2019.03.001] [Medline: 30844471]
50. Liu Y, Wang Y, Ni Y, Cheung CKY, Lam KSL, Wang Y, et al. Gut Microbiome Fermentation Determines the Efficacy of Exercise for Diabetes Prevention. Cell Metab 2020 Jan 07;31(1):77-91.e5 [FREE Full text] [doi: 10.1016/j.cmet.2019.11.001] [Medline: 31786155]
51. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: A methodology review. J Biomed Inform 2020 Sep;109:103526 [FREE Full text] [doi: 10.1016/j.jbi.2020.103526] [Medline: 32768446]

Abbreviations

CNN: convolutional neural network
EHR: electronic health record
HbA_{1c}: hemoglobin A_{1c}
ICD: International Classification of Diseases
IRR: interrater reliability
JHU: Johns Hopkins University
ML: machine learning
NB: Gaussian naïve bayes
NLP: natural language processing
NLTK: Natural Language Toolkit
PCP: primary care provider
PMAP: Precision Medicine Analytics Platform
SGD: stochastic gradient descent
SVM: support vector machines
TF-IDF: term frequency-inverse document frequency

Edited by C Lovis; submitted 21.04.21; peer-reviewed by M Peeples, M Burns, M Elbattah, O Serban; comments to author 23.09.21; revised version received 15.11.21; accepted 04.12.21; published 24.02.22

Please cite as:
Schwartz JL, Tseng E, Maruthur NM, Rouhizadeh M
Identification of Prediabetes Discussions in Unstructured Clinical Documentation: Validation of a Natural Language Processing Algorithm
JMIR Med Inform 2022;10(2):e29803
URL: https://medinform.jmir.org/2022/2/e29803
doi: 10.2196/29803
PMID: