Analysis of Primary Care Provider Electronic Health Record Notes for Discussions of Prediabetes Using Natural Language Processing Methods

INTRODUCTION

Studies examining primary care providers' (PCP) practices in managing patients with prediabetes using survey data[1] and structured data from electronic health records (EHRs)[2] suggest patients with prediabetes are not receiving evidence-based care. We developed and validated a natural language processing (NLP) tool to analyze unstructured data in EHR notes to identify prediabetes discussions and described these discussions.

METHODS

In phases 1 and 2, we included adults without diabetes with an in-person office visit at a primary care clinic ($n = 19$) at an academic medical center and at least one HbA1c 5.7–6.4% between 7/1/2016 and 12/31/2018. We based the initial keyword search strategy on the authors' clinical experience (Table 1). In phase 1, we identified and extracted PCP encounter notes matching ≥ 1 keyword from two clinics. Through random chart review of notes for patients meeting the inclusion/exclusion criteria but not containing any keyword, we identified additional keywords. The Supplement provides additional details.

In phase 2, using data from 17 other clinics, we extracted the first PCP visit note following lab results indicating prediabetes ($n = 1095$ encounters) and applied the updated keyword search strategy ($n = 391$ encounters). Two reviewers (E.T. and J.L.S.) manually annotated the notes to determine whether they contained clinical discussions of prediabetes. We applied NLP techniques using machine learning to replicate human annotation.[3] To reduce overfitting and classification bias and confirm internal validation, we applied 10-fold cross-validation to shuffle the training and test sets.[4] We selected logistic

regression and bi-directional recurrent neural networks based on performance. We evaluated classification results using sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

Two reviewers (E.T. and R.L.S.) reviewed each note from phase 2 to describe the prediabetes discussions: (1) labs ordered/reviewed (HbA1c or fasting glucose), (2) lifestyle counseling, (3) diabetes prevention program (DPP) discussion/referral, (4) nutrition discussion/referral, and (5) metformin discussion or ordering/continuation. We calculated proportions with the denominator being the number of patients with a documented discussion about prediabetes and numerator being the number of patients who had each type of discussion listed above (STATA, version 15). This study was approved by the Johns Hopkins IRB.

RESULTS

In phase 2, 322 of 391 encounter notes matching ≥ 1 keyword had documentation of prediabetes discussions. NLP and machine learning classification results were close to human performance. Logistic regression models revealed a sensitivity of 0.961, specificity of 0.923, PPV of 0.967, and NPV of 0.907. Convolutional neural networks revealed a sensitivity of 0.979, specificity of 0.956, PPV of 0.979, and NPV of 0.956. Table 1 describes the keywords in our sample.

PCPs commonly provided lifestyle counseling (78%), reviewed current labs (67%), and ordered follow-up labs (60%) (Table 2). PCPs discussed or referred patients to a nutritionist infrequently (3%). There were no discussions or referrals to a DPP. Metformin was discussed, ordered, or continued in < 2% of patients.

DISCUSSION

We developed and validated the first NLP tool that identifies clinical discussions about prediabetes from unstructured EHR data. PCPs underutilize the prediabetes diagnosis code (structured data); only 13% of patients were given a diagnosis code in a large EHR study.[2] Therefore, using structured EHR data is insufficient to identify visits where prediabetes is addressed.

Table 1 Keywords Included in Search Strategy and Frequency of Keywords Matching to Clinical Discussion About Prediabetes

Keyword	Frequency of occurrence in 322 notes* Number of patients (%)
Prediabetes	137 (43)
Hyperglycemia	55 (17)
Impaired fasting glucose	56 (17)
Pre-diabetes	41 (13)
Impaired glucose tolerance	15 (5)
Elevated hemoglobin A1c	14 (4)
Pre-DM	10 (3)
Glucose intolerance	4 (1)
Borderline diabetes	2 (1)
Elevated A1c	4 (1)
Elevated glucose	2 (1)
Elevated fasting glucose	2 (1)
PreDM	1 (0.3)
Increased risk of diabetes[†]	0
Elevated diabetes risk	0
Increased diabetes risk	0
Pre DM[†]	0
Dysglycemia	0
Pre diabetes	0

More than one keyword may have matched in each note so frequency of occurrence is greater than total number of notes
[†]*These keywords were identified in phase 1 but did not match any notes in the phase 2 sample*

Few studies have used NLP methods to identify discussions about chronic conditions[5] or lifestyle counseling.[6]

Consistent with prior findings, PCPs most commonly addressed prediabetes through lifestyle change counseling.[1] However, PCPs infrequently referred to nutrition or DPPs. Although this institution has a DPP, it is community-based and not integrated into the clinical setting, which may explain the lack of DPP referrals. PCPs commonly reviewed and ordered follow-up labs, although prior work suggests low completion rates.[2]

Our results are from one academic medical center and from a small sample of a larger population of patients with prediabetes. Our descriptive outcomes are based on provider visit documentation and providers may not have documented all the details of their verbal discussions. We did not capture discussions that may have occurred through EHR patient-provider messaging or telephone documentation.

In conclusion, we developed and validated an NLP tool identifying clinical discussions about prediabetes in the EHR. As diabetes prevention grows, this novel tool may help in tracking PCP practices outside of identifiable tasks in structured data.

Eva Tseng, MD[1,2]
Jessica L. Schwartz, MD[1]
Masoud Rouhizadeh, PhD[3]
Nisa M. Maruthur, MD[1,2,4]

[1]Division of General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2]Welch Center for Prevention, Epidemiology, & Clinical Research, Johns Hopkins University, Baltimore, MD, USA
[3]Institute for Clinical and Translational Research, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[4]Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA

Corresponding Author: *Eva Tseng, MD; Division of General Internal Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA (e-mail: etseng3@jhmi.edu).*

Compliance with Ethical Standards:

Conflict of Interest: *Dr. Maruthur is the co-inventor of a virtual diabetes prevention program. Under a license agreement between Johns Hopkins HealthCare Solutions and the Johns Hopkins University, Dr. Maruthur 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.*

Supplementary Information: *The online version contains supplementary material available at https://doi.org/10.1007/s11606-020-06400-1.*

Table 2 PCP Management of Prediabetes Documented in Clinical Encounters (N = 322 Encounters)

Outcome	N (%) Percentage (N = 322)
Labs discussion*	
Labs reviewed	215 (66.8)
Labs ordered or plan to order	196 (60.1)
Lifestyle change discussion	
Behavioral/lifestyle management recommended	250 (77.6)
Nutrition referral placed	10 (3.1)
Nutrition visit discussed[†]	9 (2.8)
Diabetes prevention program discussed or referral placed	0
Metformin discussion	
Metformin discussed	6 (1.9)
Metformin ordered	5 (1.6)
Metformin continued	4 (1.2)

Labs included hemoglobin A1c and glucose if fasting was noted
[†]*Discussion about prior completed nutrition visit*

REFERENCES

1. **Tseng E, Greer RC, O'Rourke P, Yeh HC, McGuire MM, Albright AL, et al.** National Survey of Primary Care Physicians' Knowledge, Practices, and Perceptions of Prediabetes. J Gen Intern Med. 2019;34(11):2475-81.
2. **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;37(2):565-8.

3. **Deng L, Liu Y**, editors. Deep Learning in Natural Language Processing. Germany: Springer Singapore; 2018.
4. **Martin JH, Jurafsky D.** Speech and Language Processing, 2nd Edition. United Kingdom: Pearson Prentice Hall; 2008.
5. **Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H.** Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. BMC Geriatr. 2017;17(1):248.
6. **Hazlehurst BL, Lawrence JM, Donahoo WT, Sherwood NE, Kurtz SE, Xu S, et al.** Automating assessment of lifestyle counseling in electronic health records. Am J Prev Med. 2014;46(5):457-64.