MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190250

Identifying Diabetes in Clinical Notes in Hebrew: A Novel Text Classification Approach Based on Word Embedding

Maxim Topaz [a,b], Ludmila Murga [c], Chagai Grossman [d], Daniella Daliyot [d], Shlomit Jacobson [d], Noa Rozendorn [d], Eyal Zimlichman [d], Nadav Furie [d]

[a] *School of Nursing & Data Science Institute, Columbia University, New York City, NY, USA*
[b] *Visiting Nurse Service of New York, New York City, NY, USA*
[c] *Cheryl Spencer Department of Nursing, University of Haifa, Haifa, Israel*
[d] *Sheba Medical Center, Tel Hashomer, Israel.*

Abstract

NimbleMiner is a word embedding-based, language-agnostic natural language processing system for clinical text classification. Previously, NimbleMiner was applied in English and this study applied NimbleMiner on a large sample of inpatient clinical notes in Hebrew to identify instances of diabetes mellitus. The study data included 521,278 clinical notes (one admission and one discharge note per patient) for 268,664 hospital admissions to medical-surgical units of a large hospital in Israel. NimbleMiner achieved overall good performance (F-score =.94) when tested on a gold standard human annotated dataset of 800 clinical notes. We found 15% more patients with diabetes mentioned in the clinical notes compared with diagnoses data. Our findings about underreporting of diabetes in the coded diagnoses data highlight the urgent need for tools and algorithms that will help busy providers identify a range of useful information, like having a diabetes.

Keywords:
Natural language processing, Text classification, Diabetes.

Introduction

With the wide international adoption of electronic health records and other information technology in healthcare, there is an exponential growth in health data. It is estimated that as much as 50 to 80% of all healthcare data are captured as unstructured, mostly free text narrative data. Examples of health narrative data include daily inpatient progress notes that are completed by physicians and nurses, primary care clinics follow-up notes, radiology notes, surgical notes, etc. [1].

With large potential benefits that are projected from using these data, these exponentially growing narrative information sources pose several significant challenges. Busy clinicians are now required to go through an increasing number of narrative notes about a patient to understand their medical history and diagnose their conditions correctly [2]. Health researchers are dealing with an increasing number of clinical notes they need to review in order to identify important aspects of care and patient characteristics [2].

Some of these challenges can be solved through natural language processing (NLP) with a variety of techniques aimed at extracting meaning from narrative data [1]. NLP can be broadly divided into linguistic approaches and statistically based approaches. Linguistic NLP systems tend to use complex language characteristics, for example large vocabularies of pre-defined terms and expressions, to identify a concept of interest in clinical texts. For example, past studies extracted wound characteristics from outpatient notes [3] and identified patients with poor treatment adherence in inpatient notes [4] using pre-build vocabularies of clinical terms for these domains. Statistically-based NLP approaches - sometimes also called text classification- are often based on probabilistic estimations of a presence of a specific phenomenon in the clinical texts. For example, text classification was used to classify clinical notes in terms of information completeness and adequacy [5] or to identify depression and depression symptoms among patients [6].

Although significant advancements in clinical NLP were made over the last decades, most of the published NLP studies and existing NLP systems focus on processing information in English. In general, a vast majority of languages around the world are considered low-resource languages in terms of NLP [7]. More tools and approaches are urgently needed to overcome the resource barrier so advances in NLP can deliver more widespread benefits to health providers, researchers and patients [8].

This paper presents a novel open-source NLP system called NimbleMiner. NimbleMiner was previously tested in text classification tasks in English. For example, we successfully applied NimbleMiner to identify fall-related information in clinical notes, while achieving higher system performance in significantly less time compared to a previous rule-based systems [9]. By design, NimbleMiner implements a language-agnostic NLP approach and we estimated that our system can be applied to languages other than English. To evaluate this assumption, this study applied NimbleMiner to process clinical narratives in Hebrew.

Similar to many other languages, Hebrew has very scarce resources to implement NLP [8]. In healthcare, we found only one published effort that reported building Hebrew NLP healthcare pipeline in early 2010 [10] and since then, no other reports were identified. The aim of this study was to apply NimbleMiner on a large sample of inpatient clinical notes in Hebrew to identify instances of diabetes mellitus (both insulin dependent type I diabetes and insulin resistant type II diabetes). We worked with clinicians in the largest hospital in Israel (Sheba Medical Center) to develop the NLP pipeline and evaluate its performance.

Methods

Our methods are summarized in Figure 1. We first manually reviewed a random sample of admission and discharge narrative notes for 400 patients, identified diabetes cases, and created a gold standard of human annotated notes with the help of clinical experts. We then trained and tested our NLP system—called NimbleMiner—and classification algorithms. Finally, we assessed our automated approach by comparing system-generated classification against the gold standard. We also compared diabetes cases extracted from narrative data versus coded diagnoses for pateints included in the gold standard.

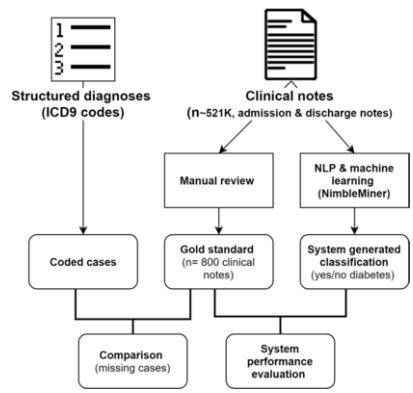

Figure 1– *Study Methods Overview*

Dataset

We extracted a large sample of clinical notes of all the patients admitted to the Sheba Medical Center over two years 2016-17. Sheba Medical Center located in the Tel Aviv District, is the largest hospital in Israel with about 1,700 beds, over 1,400 physicians, and 2,600 nurses. Only patients admitted to medical-surgical units were included in this study. We excluded any pediatric patients or women who were admitted for pregnancy/delivery reasons. Over the study time period, there were 268,664 admissions for 127,851 unique patients. Patients' average age was 65.1.

In this project, we used a subset of admission and discharge narrative clinical notes for all the patients in the sample. There were 521,278 clinical notes (one admission and one discharge note per patient's admission) after excluding missing data. Admission notes consisted of narrative description of patient's state at the hospital admission, including previous medical history, patient's diagnoses, medications, reason for admission, patient's signs and symptoms, relevant social history, etc. Discharge notes included information similar to the admission note, with addition of hospital treatment course and discharge recommendations. On average, clinical notes had roughly 2,000 characters, with admission notes being slightly longer than discharge summaries (about 200 charters longer on average). It is important to note that the written Hebrew alphabet does not include vowels, thus writing is more concise compared to other languages. This study received an Institutional Review Board approval from the University of Haifa, Israel.

Domain Definitions

We used clinical literature [11] and our team's expertise to define the domain of interest. In this project, we focused on identifying clinical notes where diabetes is very likely to be described. To accomplish that, we defined several categories of terms that will suggest high likelihood of diabetes presence, as follows:

- **Diagnosis of diabetes**: any mention of patient's having diabetes in the text (e.g., terms like "diabetes mellitus", abbreviations like "DM", etc.).
- **Mention of one or more diabetic medications** in the text: including – Insulin derivatives (e.g., Lantus, Apidra, Novorapid, Actrapid), Insulin pump, Glucomin, Glucophage, Methformin, Januet, Jardiance, Sulfanyl urea, Gluben, Forxiga, DPP4, Januvia, GLP1, Victoza, Trajenta, etc.).
- **Indication of hemoglobin A1c** >6.5% or **fasting glucose levels** > 125 mg/dL in the text.
- **Diabetes complications** in the text (e.g., diabetic foot, diabetic neuropathy, diabetic retinopathy, etc.).

NLP System Description

NimbleMiner is an open source NLP system developed by our team [9]. User manual and download options can be accessed at: http://github.com/mtopaz/NimbleMiner. Other research or clinical teams can use the system under the GNU General Public License v3.0. NimbleMiner includes several methodological stages of clinical note processing that are briefly described below and presented in Figure 2.

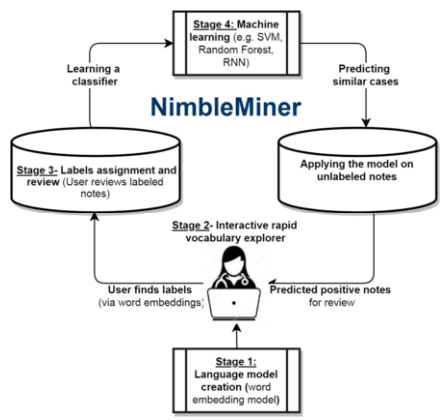

Figure 2– *NimbleMiner System Process Stages*

Stage 1- Language model creation: The user selects a large corpus of clinical notes and defines language model characteristics. We use a word embedding model (specifically skip-gram model) for language model generation and users can change model settings based on their preferences. **Stage 2- Interactive rapid vocabulary explorer**: The user enters a query term of interest, and the system returns a list of similar terms it identified as relevant. The list of suggested similar terms is based on the cosine term similarity metric extracted from the word embedding model. The user selects and saves the relevant terms by clicking on them. Negated or other irrelevant terms and expressions that are not selected by the user are saved by the system for further tasks, such as negation detection. Figure 3 describes the steps of the vocabulary explorer stage. **Stage 3- Labels assignment and review**: The system uses the stage 2 discovered similar terms to assign labels to clinical notes (while excluding notes with negations and other irrelevant terms). Assigning a positive label means that a concept of interest is present in the clinical note. When needed, the user reviews and updates lists of similar terms and negated similar terms. The user reviews the clinical notes with assigned labels for accuracy. This weakly supervised rapid labeling approach

is based on a positive labels learning framework validated in previous research [12,13]. **Stage 4- Machine learning:** The user choses a machine learning algorithm to be applied to create a predictive model (e.g., Support vector machines, recurrent neural network, random forest). The model is then applied to predict which clinical notes might have the concept of interest. The user reviews the predicted notes and can go through stages 2-4 again to add new labels.

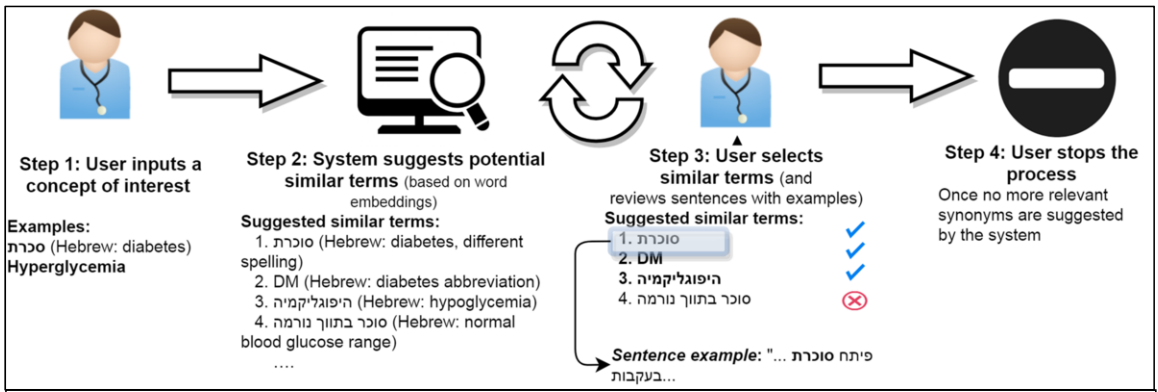

Figure 3– *NimbleMiner Stage 2 -Interactive Rapid Vocabulary Explorer Process Overview*

Legend*: Interactive rapid vocabulary explorer: The user enters a query term of interest (step 1), and the system returns a list of similar terms it identified as relevant (step 2). The list of suggested similar terms is based on the cosine term similarity metric extracted from the word embedding model. The user selects and saves the relevant terms while negated or other irrelevant terms that were not selected by the user are also saved in the system for further tasks (step 3). The system iteratively identifies new potential similar terms and presents them to the user for review (steps 2-3) and the process continues until there no new similar terms to suggest (step 4)*

NLP System Settings

Narrative data were pre-processed by removing punctuation and lowercasing all English words (Hebrew letters do not have upper case). All numeric symbols were kept in the text. Additionally, we converted frequently co-occurring words in the clinical notes into phrases with lengths of up to four words (4-grams) [14]. This is a common process in NLP where sets of co-occurring words are combined into phrases. For example, "pt has diabetes" might be a common 3-gram. We used a phrase2vec algorithm with default settings to implement this in NimbleMiner.

NimbleMiner has a user interface that is implemented in R statistical package. We used a skip-gram model implementation called word2vec and phrase2vec to create a word embedding model in R [14]. Parameters of the word embedding model were held constant based suggestions in other studies of word embedding [15]. Specifically, we used a model with window width size = 10, vector dimension = 100, minimum word count = 5, negative sample size = 5, and sub-sampling = 1e-3. We used all the available clinical documents (~ 521K clinical notes) to train the model.

For each similar term entered by the user, the system presented 50 potentially similar terms based on the cosine term similarity. Our previous experiments [9] showed that the random forest algorithm outperforms other approaches (e.g., J48 Decision trees, Support Vector Machines), hence we used this algorithm in the machine learning stage of this study. Random Forest algorithm was used with default settings (number of iterations = 100, minimum number of instances = 1, minimum variance for split = 1e-3, depth = unlimited).

To identify negation terms for medical domain in Hebrew, we started from translating a commonly used vocabulary of negation terms in English, called NegEx [16]. Next, we used our clinical and Hebrew linguistic knowledge to identify additional negation terms. We then used NimbleMiner's vocabulary explorer to identify other potential negations and variations in negations spelling. Our final list of medical negations in Hebrew includes 118 terms, which might appear before or/and after a diabetes term (e.g., "שולל"- denies or "ללא"- has no). n-grams (irrelevant terms not selected by the user) were used to detect family and past history contexts.

NLP Implementation

Similar terms were explored by a team of 2 nurses and 2 physicians who used NimbleMiner. First, two nurses and one physician used NimbleMiner to identify and review potentially relevant terms independently (NimbleMiner stage 2) and then the lists of terms were combined into one list. Another physician then reviewed the combined list and finalized any disagreements or additional terms that appeared in one list but not in the other.

System Evaluation- Gold Standard Creation

We extracted a random sample of admission and discharge notes for 400 patients in our database (800 clinical notes total). Each note was reviewed by 3 clinicians on our team (2 nurses and 1 doctors) for presence of diabetes. Clinicians were asked to use their clinical judgment and indicate if each specific note has a mention of diabetes. There was moderate inter-rater agreement (Cohen's Kappa inter-rater agreement = .69). A senior internal medicine physician then reviewed each of the cases and adjudicated whether diabetes was present or absent in cases of disagreement. This corpus of documents was not used for algorithm training but only as a gold standard for algorithms evaluation.

We applied our NLP algorithms on the gold standard dataset to predict presence of diabetes in each of the documents. We calculated precision (defined as the number of true positives out of the total number of predicted positives), recall (defined as the number of true positives out of actual number of positives) and F-score (F1, weighted harmonic mean of the precision and recall) to evaluate and compare the performance of our algorithms.

Comparison with Structured Diagnosis Data

In this project, we only had access to coded diagnoses of patients included in the gold standard set. Based on expert judgment, we labeled each note as either positive or negative for presence of diabetes based on information identified in the clinical notes. We then compared our results (per hospital admission and discharge) with structured data on patient diagnoses. Diagnoses data included primary diagnosis at admission, a list of secondary diagnoses, and discharge diagnoses (that might be different from admission diagnoses because they were updated during hospitalization). All diagnoses data were stored as International Calcification of Diseases version 9 (ICD 9) codes [17] and we used all codes from category 250.XX (diabetes mellitus) and V58.67 (Long term, current insulin use).

Results

Overall, 1,427 terms and expressions were discovered for diabetes using NimbleMiner. There was a moderate inter-rater agreement (Cohen's Kappa inter-rater agreement = .69) on the relevant terms between the three clinicians. Another physician reviewed all the terms, forming the final terms corpus. Table 1 presents the number of terms identified for each diabetes category. Since the system is using cosine similarity as a measure of term relatedness, our approach identified terms in Hebrew, English, misspellings in both languages, abbreviations and other lexical variants specific to Hebrew (for example specific Hebrew suffixes indicative of male or female).

Table 1– Categories and Examples of Diabetes Expressions

Domain	Number of terms	Examples
Diagnosis of diabetes	928	"dm 2", "סכרת type 2" (diabetes type 2), "ברקע גבולית סכרת" (diabetic background)
Diabetic medications	358	"באינסולין מטופלת" (uses insulin), "lantus במינון", "ב glucophage"
Hemoglobin A1c >6.5% or fasting glucose levels > 125 mg/dL	49	"hba1c>6.5", "מעל סוכר 180" (blood sugar above 180)
Diabetes complications	92	"diabetic foot", "נפרופתיה סוכרתית" (diabetic nephropathy)
Total	**1,427**	

NLP System Performance Evaluation

We first evaluated NimbleMiner's performance using just the labeling process (see NimbleMiner stage 3, weakly supervised labeling) and then compared these results with random forest algorithm performance. Table 2 shows that NimbleMiner's labeling alone showed better results than further machine learning or codes alone.

Table 2– Codes and NLP performance on the gold standard

	Diagnoses codes	NimbleMiner labeling	Random forest
Precision	.78	**.91**	.87
Recall	.85	**.96**	.86
F-score	.82	**.94**	.87

Comparison with Structured Diagnosis Data

Overall, we found 176 patients with diabetes in the coded diagnoses data while our experts found 159 patients with diabetes based on clinical notes review (800 clinical notes included in the gold standard). This difference is explained by the fact that diabetes for these patients (n=17) was coded, but was not mentioned in the text. On the other hand, we found that our experts identified an additional 15% (n= 26) more patients with diabetes described in the text compared to the diagnoses data.

Discussion

This study is one of the first to implement NLP in medical domain in Hebrew. In general, NimbleMiner's approach proved useful for clinical narratives in Hebrew. First, our team of clinicians have successfully interacted with NimbleMiner to identify a large vocabulary of terms and expressions describing diabetes in clinical notes. This list included misspellings, expressions in both Hebrew and English, specific Hebrew lexical variants, etc. Compiling such a vocabulary is often a challenge in NLP projects and it is usually conducted thorough literature review, reading large amounts of clinical notes, etc. [8]. Our approach enables clinicians to identify a comprehensive vocabulary of terms specific to their domain within a short period of time, while being assisted by machine learning components.

This comprehensive vocabulary of terms is then applied to label clinical notes. During labeling, negated terms are removed from positively labeled cases. We also remove irrelevant terms that were reviewed by the user during the vocabulary exploration process (stage 2). For example, similar terms related to presence of diabetes, like "dm 2" or are sometimes included in a larger irrelevant term, like "family history of dm 2", "son has dm 2", or "screened for dm2". These irrelevant terms are excluded, which helps our system to conduct high quality labeling. Our results support the current literature suggesting that machine learning with human-in-the-loop is an effective approach in healthcare.[18,19] Human experts can interact with machine learning to reduce NLP task complexity and improve machine learning speed.

Machine learning is then performed using the set of positively labeled clinical notes and similar in size sample of notes labeled as "unknown" to train the algorithm. In this project, our labeling approach outperformed random forest algorithm. One possible explanation is that for domains with relatively straightforward words and expressions, like presence of a diagnosed disease, rule-based approaches like ours work better than machine learning. On the other hand, machine learning can work better for domains where more ambiguity is present. Further research is needed to validate these claims. Other studies of NLP in the diabetes domain in English achieved similar NLP system performance.[20,21]

Importantly, our results indicate that about 15% of patients with diabetes mentioned in the clinical notes, did not have a diabetes-related diagnoses in the coded data. This is a serious concern for various reasons. First, not having a diabetes diagnosis might affect patient safety. For example, diabetic patients might be given a contra-indicated medication or provided with an inappropriate diet. Additionally, there are several significant implications for the organization in terms of reimbursement for patient care. When diagnoses are underreported to the health services payer—whether it is an insurance or ministry of health—medical institutions can potentially receive less reimbursement for the care they

provided. Third, underreported major diagnoses—like diabetes—pose significant limitations for any research that is conducted with clinical data.

Limitations: This study has several important limitations. First, we have only experimented with one data mining algorithm and more work is needed to examine more algorithms, such as Support Vector Machines. In addition, the use of NimbleMiner might be restricted to classifying documents based on phrases with high positive predictive value, such as diabetes terms and expressions, and further work is needed to explore the generalizability of our approach.

Conclusions

Our pioneering results indicate the feasibility of an NLP approach that can span across different languages, like English and Hebrew. Our finding about underreporting of diabetes in the coded diagnoses data highlight the urgent need in tools and algorithms that will help busy providers identify a range of useful information, like having diabetes. Identifying this information can also help health organizations receive optimal reimbursement for their services while health researchers will glean a more accurate picture of the patient population for research.

Acknowledgements

Funding for this work was provided through Sheba Innovation Fund.

References

[1] D. Demner-Fushman, W.W. Chapman, and C.J. McDonald, What can natural language processing do for clinical decision support?, *J. Biomed. Inform.* **42** (2009) 760–72. doi:10.1016/j.jbi.2009.08.007.

[2] N. Griffon, J. Charlet, and S.J. Darmoni, Managing Free Text for Secondary Use of Health Data, *Yearb. Med. Inform.* **23** (2014) 167–169. doi:10.15265/IY-2014-0037.

[3] M. Topaz, K. Lai, D. Dowding, V.J. Lei, A. Zisberg, K.H. Bowles, and L. Zhou, Automated identification of wound information in clinical notes of patients with heart diseases: Developing and validating a natural language processing application, *Int. J. Nurs. Stud.* **64** (2016). doi:10.1016/j.ijnurstu.2016.09.013.

[4] M. Topaz, K. Radhakrishnan, S. Blackley, V. Lei, K. Lai, and L. Zhou, Studying Associations Between Heart Failure Self-Management and Rehospitalizations Using Natural Language Processing., *West. J. Nurs. Res.* (2016). doi:10.1177/0193945916668493.

[5] W. Al Assad, M. Topaz, J. Tu, and L. Zhou, The application of machine learning to evaluate the adequacy of information in radiology orders, in: 2017 IEEE Int. Conf. Bioinforma. Biomed., IEEE, 2017: pp. 305–310. doi:10.1109/BIBM.2017.8217668.

[6] L. Zhou, A.W. Baughman, V.J. Lei, K.H. Lai, A.S. Navathe, F. Chang, M. Sordo, M. Topaz, F. Zhong, M. Murrali, S. Navathe, and R.A. Rocha, Identifying Patients with Depression Using Free-text Clinical Documents, 2015. doi:10.3233/978-1-61499-564-7-629.

[7] S. Abney, and S. Bird, The Human Language Project: Building a Universal Corpus of the World's Languages, *Proc. 48th Annu. Meet. Assoc. Comput. Linguist.* (2010) 88–97. https://aclanthology.coli.uni-saarland.de/papers/P10-1010/p10-1010 (accessed November 21, 2018).

[8] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, Clinical Natural Language Processing in languages other than English: opportunities and challenges, *J. Biomed. Semantics*. **9** (2018) 12. doi:10.1186/s13326-018-0179-8.

[9] M. Topaz, L. Murga, K.M. Gaddis, M. V. McDonald, O. Bar-Bachar, Y. Goldberg, and K. Bowles, Mining fall-related information in clinical notes: comparison of rule-based and novel word embedding-based machine learning approaches, *J. Biomed. Inform.* (2019) 103103. doi:10.1016/J.JBI.2019.103103.

[10] R. Cohen, Y. Goldberg, and M. Elhadad, Transliterated pairs acquisition in medical Hebrew, in: Mach. Transl. Morphol. Lang. Work., 2011.

[11] American Diabetes Association, Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2018., *Diabetes Care*. **41** (2018) S13–S27. doi:10.2337/dc18-S002.

[12] Y. Halpern, S. Horng, Y. Choi, and D. Sontag, Electronic medical record phenotyping using the anchor and learn framework., *J. Am. Med. Inform. Assoc.* **23** (2016) 731–40. doi:10.1093/jamia/ocw011.

[13] C. Elkan, and K. Noto, Learning Classifiers from Only Positive and Unlabeled Data, *KDD* . (2008).

[14] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, Distributed Representations of Words and Phrases and their Compositionality, (2013) 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality (accessed October 2, 2017).

[15] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo, How to Train good Word Embeddings for Biomedical NLP, in: Proc. 15th Work. Biomed. Nat. Lang. Process., 2016: pp. 166–174. doi:10.18653/v1/W16-2922.

[16] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, *J. Biomed. Inform.* **34** (2001) 301–310. doi:10.1006/jbin.2001.1029.

[17] WHO, WHO | International Classification of Diseases (ICD), (2014). http://www.who.int/classifications/icd/en/ (accessed February 23, 2015).

[18] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, *Brain Informatics*. **3** (2016) 119–131. doi:10.1007/s40708-016-0042-6.

[19] A. Holzinger, From Machine Learning to Explainable AI, in: 2018 World Symp. Digit. Intell. Syst. Mach., IEEE, 2018: pp. 55–66. doi:10.1109/DISA.2018.8490530.

[20] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, *Comput. Struct. Biotechnol. J.* **15** (2017) 104–116. doi:10.1016/J.CSBJ.2016.12.005.

[21] S.M. Meystre, and P.J. Haug, Comparing natural language processing tools to extract medical problems from narrative text., *AMIA ... Annu. Symp. Proceedings. AMIA Symp.* **2005** (2005) 525–9. http://www.ncbi.nlm.nih.gov/pubmed/16779095 (accessed March 19, 2019).

Address for correspondence

Max Topaz, mt3315@cumc.columbia.edu.