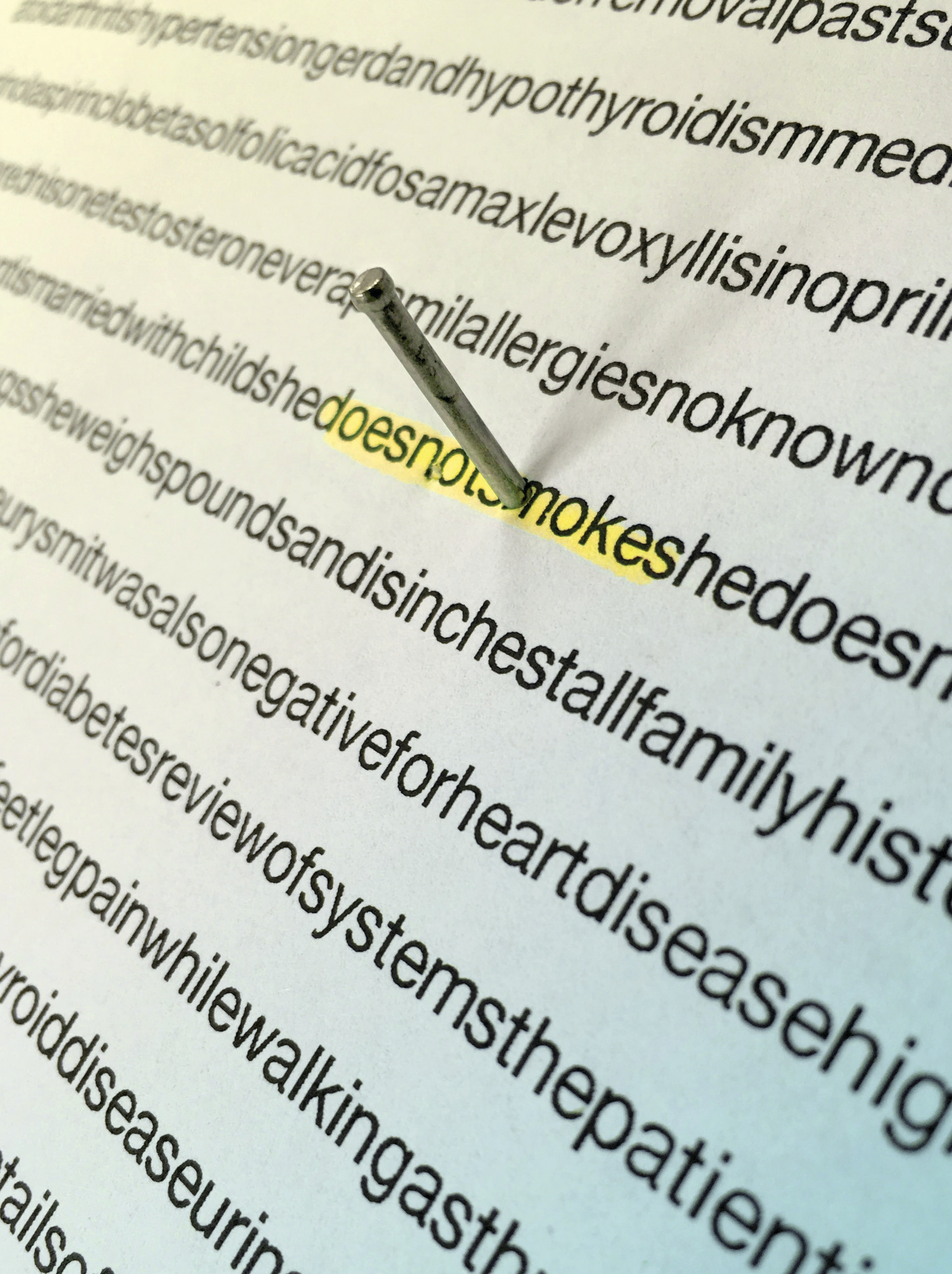


removalpastsu

zocarthritishypertensiongerdandhypothyroidismmed

olaspirinclobetasolfolicacidfosamaxlevoxyllisinopril

rednisonetestosteroneverapamilallergiesnoknown

ntismarriedwithchildshedoesnotsmokeshedoesn

gssheweighspoundsandisinchestallfamilyhist

urysmitwasalsonegativeforheartdiseasehig

fordiabetesreviewofsystemsthepatient

eetlegpainwhilewalkingasth

yroiddiseaseurin

ailsof

Uri Kartoun, IBM Research

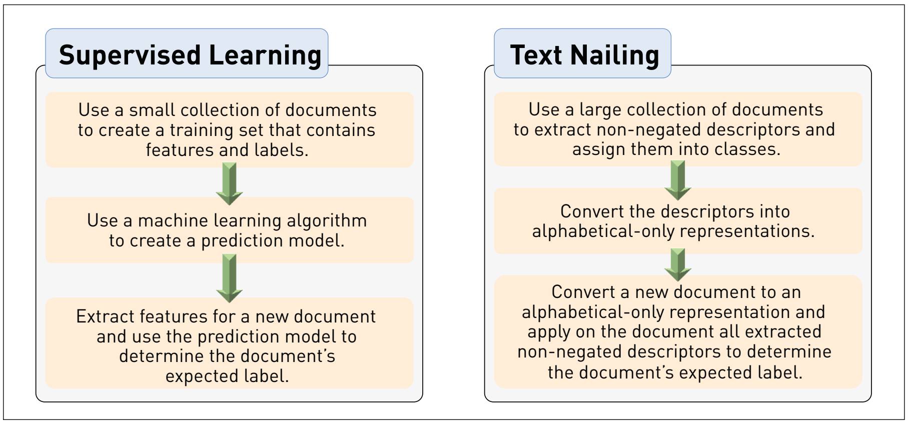
Text Nailing: An Efficient Human-in-the-Loop Text-Processing Method

Insights

- The available machine-learning text-classification methods show only fair levels of accuracy in extracting patients' medical conditions and behavioral descriptors.
- An easily adaptable human-in-the-loop big-data method with an interactive front end may improve classification accuracy of widely used text-classification techniques.

IMAGE BY ALICIA KUBISTA/ANDRIJ BORYS ASSOCIATES

A significant portion of my time as a research fellow at Massachusetts General Hospital (MGH) was dedicated to the exploration of a cohort of 314,292 patients at increased risk for metabolic syndrome [1]. Patients in this cohort had at least one type 2 diabetes mellitus (T2DM) diagnosis code, a T2DM medication, an HGB A1C level ≥ 6.5 percent, or plasma glucose ≥ 200 mg/dl. Of these patients, 65,099 were diagnosed with T2DM at a specificity of 97 percent and positive predictive value of 96 percent [2]. During my training years (2013–2016), my colleagues at MGH and Harvard and I implemented a variety of predictive-modeling methods and incorporated natural language processing techniques to better understand diseases and their complications. We focused on cardiovascular disease, liver disease, and physician-documented insomnia. This cohort contained the complete clinical details and demographics of patients who received care at MGH or Brigham and Women's Hospital between 1992 and 2010. The cohort was large, considering all clinical narrative notes (e.g., office, medication management, and operative notes) that accompanied the traditional electronic health record (EHR) elements (e.g., billing codes and medication prescriptions).


Supervised Learning

Use a small collection of documents to create a training set that contains features and labels.

Use a machine learning algorithm to create a prediction model.

Extract features for a new document and use the prediction model to determine the document's expected label.

Text Nailing

Use a large collection of documents to extract non-negated descriptors and assign them into classes.

Convert the descriptors into alphabetical-only representations.

Convert a new document to an alphabetical-only representation and apply on the document all extracted non-negated descriptors to determine the document's expected label.

Figure 1. Supervised learning versus Text Nailing.

At the end of 2014, cardiologist Dr. Stanley Shaw introduced me to Dr. Kathleen Corey, a hepatologist with whom I started interrogating the cohort to identify new biomarkers associated with outcomes in individuals suffering from liver diseases and associated comorbidities. Dr. Corey was interested in extracting smoking-status information for use as an important covariate in our prediction models. The effects of tobacco use are significant in the study of patient outcomes and have been studied extensively over the past several decades. Tobacco use is linked to an increased risk for and severity of a variety of diseases, including cardiovascular disease, respiratory illness, psychiatric conditions, and cancers.

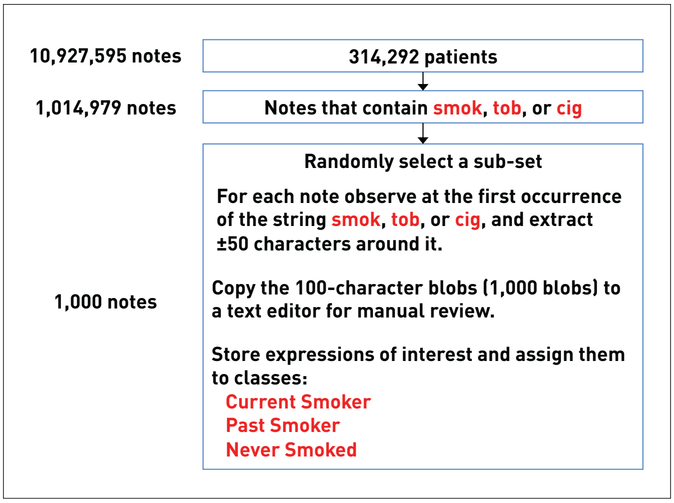
I pointed out that smoking status is a data element that is not captured sufficiently in a structured form in this cohort. Smoking status is typically documented in clinical narrative notes as free text. The available smoking-status extraction methods (which are commonly based on supervised learning) are only moderately accurate as reported in many publications, which could result in misclassifications. Using support-vector machines (SVM), for instance, and several hundred documents, yielded an accuracy of 85.57 percent, meaning that 14.43 percent of the documents were misclassified [3].

The most significant scientific moment during my training years at MGH was when, inspired by Dr. Corey's request to extract smoking statuses, I thought to implement a new, highly accurate text-classification method to extract the statuses from notes. Motivated by my Ph.D. dissertation in human-robot collaboration to accomplish learning tasks [4], I hypothesized that following a simple human-in-the-loop approach could achieve better results than many widely used computational approaches. I dedicated a few days to implementing my method; in the subsequent months, my colleagues and I evaluated its accuracy and performance.

We have extensively tested my method, which I call Text Nailing (TN). I came up with the notion of Text Nailing to allude to a metaphorical hammer that uses metaphorical nails to fasten characters in a fixed position. Any alphabetical letter must precede or follow another alphabetical letter. Figure 1 illustrates the difference between the widely adopted supervised-learning approach for text classification and TN. I had the opportunity to briefly present TN at the American Medical Informatics Association's 2016 Annual Symposium [5]. In all use cases, nurses and physicians manually validated our performance results using clinical chart reviews to guarantee high levels of accuracy. Typically, micro and macro F-measures (weighted averages of precision and recall frequently used in information retrieval) were above 0.95 for the extracted descriptors. In contrast, using other approaches in the task of classifying smoking status yielded lower performance, for example, micro F-measures of up to 0.90 and macro F-measures of up to 0.76 [6].

HUMAN-IN-THE-LOOP FOR IDENTIFYING SMOKING-RELATED EXPRESSIONS

Classification of whether a clinical narrative note contains an indication for smoking status (i.e., current, past, or never) requires the identification of smoking-related expressions, which need to be manually assigned into classes. To identify unique expressions that distinctively define smoking status, we implemented


10,927,595 notes — 314,292 patients

1,014,979 notes — Notes that contain smok, tob, or cig

1,000 notes — Randomly select a sub-set

For each note observe at the first occurrence of the string smok, tob, or cig, and extract ±50 characters around it.

Copy the 100-character blobs (1,000 blobs) to a text editor for manual review.

Store expressions of interest and assign them to classes:
Current Smoker
Past Smoker
Never Smoked

Figure 2. An interactive human-in-the-loop method to identify smoking status.

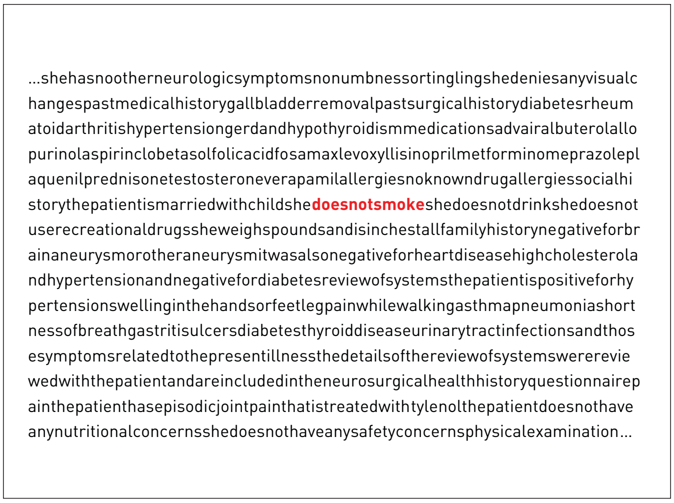

...shehasnootherneurologicsymptomsnonumbnessortinglingshedeniesanyvisualc
hangespastmedicalhistorygallbladderremovalpastsurgicalhistorydiabetesrheum
atoidarthritishypertensiongerdandhypothyroidismmedicationsadvairalbuterolallo
purinolaspirinclobetasolfolicacidfosamaxlevoxyllisinoprilmetforminomeprazolepl
aquenilprednisonetestosteroneverapamilallergiesnoknowndrugallergiessocialhi
storythepatientismarriedwithchildshe**doesnotsmoke**shedoesnotdrinkshedoesnot
userecreationaldrugssheweighspoundsandisinchestallfamilyhistorynegativeforbr
ainaneurysmorotheraneurysmitwasalsonegativeforheartdiseasehighcholesterola
ndhypertensionandnegativefordiabetesreviewofsystemsthepatientispositiveforhy
pertensionswellinginthehandsorfeetlegpainwhilewalkingasthmapneumoniashort
nessofbreathgastritisulcersdiabetesthyroiddiseaseurinarytractinfectionsandthos
esymptomsrelatedtothepresentillnessthedetailsofthereviewofsystemswererevie
wedwiththepatientandareincludedintheneurosurgicalhealthhistoryquestionnairep
ainthepatienthasepisodicjointpainthatistreatedwithtylenolthepatientdoesnothave
anynutritionalconcernsshedoesnothaveanysafetyconcernsphysicalexamination...

Figure 3. An example of an alphabetical-only converted note.

a human-in-the-loop procedure (Figure 2). The procedure initially considered all available 10,927,595 clinical narrative notes associated with the 314,292 patients and ignored notes that did not contain any smoking-related keywords (e.g., “smok,” “tobac,” and “cig”). Once we identified a smaller set of notes, we used a randomly selected sample to observe small text blobs located next to smoking-related keywords. This allowed for the quick identification of expressions associated with smoking status and the classification of the expressions into the smoking-status classes. The manual evaluation continued until we identified a subjectively defined significant portion of the smoking-related distinctive expressions.

HOMOGENEOUS SMOKING-RELATED EXPRESSIONS

We converted all identified expressions and notes into alphabetical-only representations (i.e., removed all numbers, spaces, and characters outside the 26 English letters) to create homogeneous representations of the expressions and to allow one-to-one matching when searching for an expression in a note. In clinical documentation, notes typically are heterogeneous and may contain dozens of similar expressions that describe the same smoking status. For instance, “smoked 2 packs per week” (an indication for past smoker) may be similar to variations in other notes, such as “smoked 5 packs   per   week” (a different number of packs and a tab) or “smoked: 4-6 packs   per   week” (a range for the number of packs, a hyphen, and multiple spaces). As such, we converted all identified smoking-related expressions to homogeneous representations of the expressions. The three exemplary sentences are represented by one homogeneous expression: “smokedpacksperweek.” An example for a note converted to an alphabetical-only representation is shown in Figure 3.

Of the hundreds of smoking-related expressions that we identified, we selected a cutoff threshold to highlight the expressions that are useful to the identification of smoking status in a variety of cohorts. We sampled 10,000 notes to present the prevalence of the

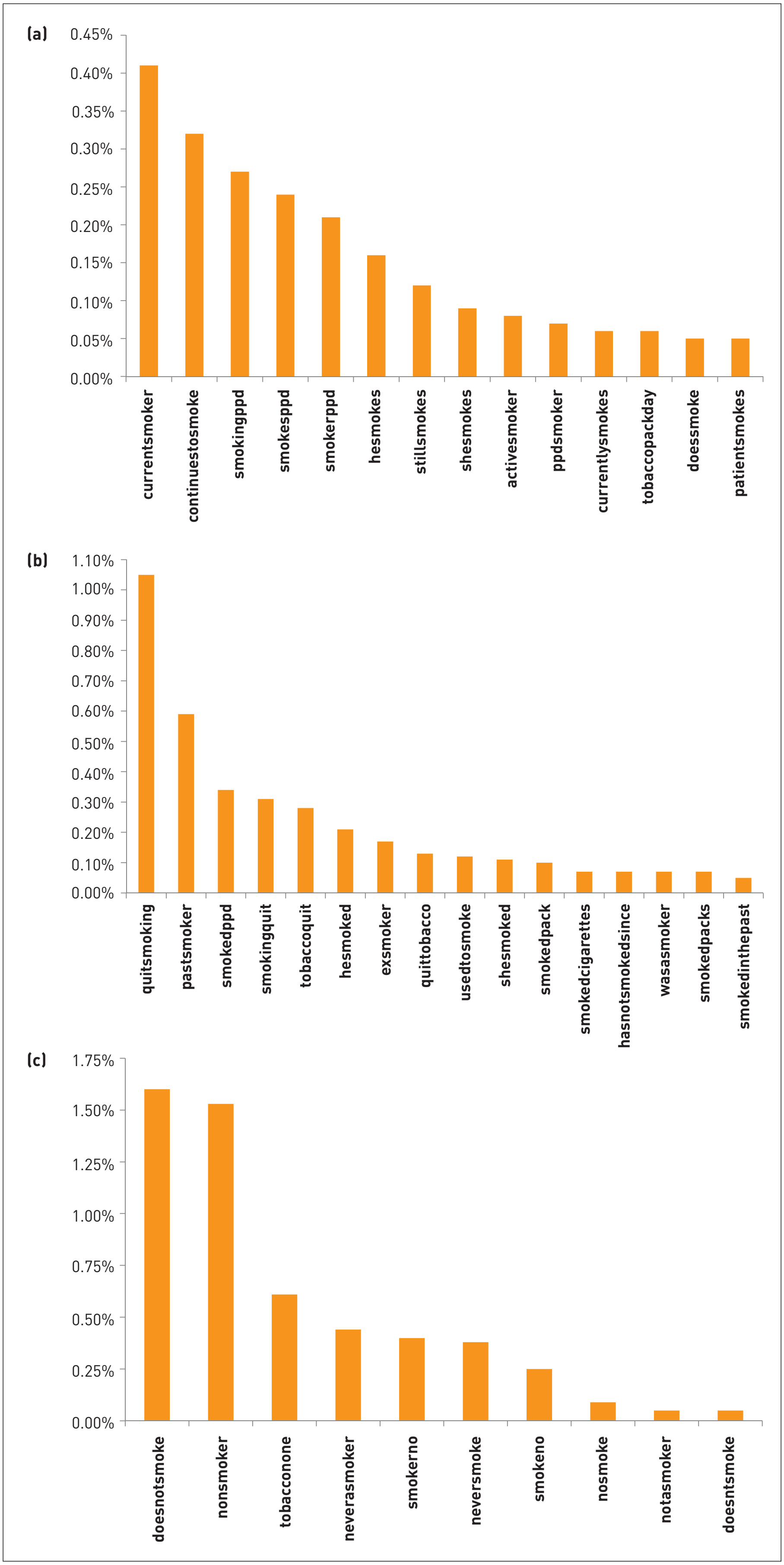


Figure 4. Most prevalent smoking-related expressions. The y-axis represents the percentage of notes that contain the expression within a sample of 10,000 randomly selected notes. PPD=packs per day. (A) Current smoker. (B) Past smoker. (C) Never smoked.

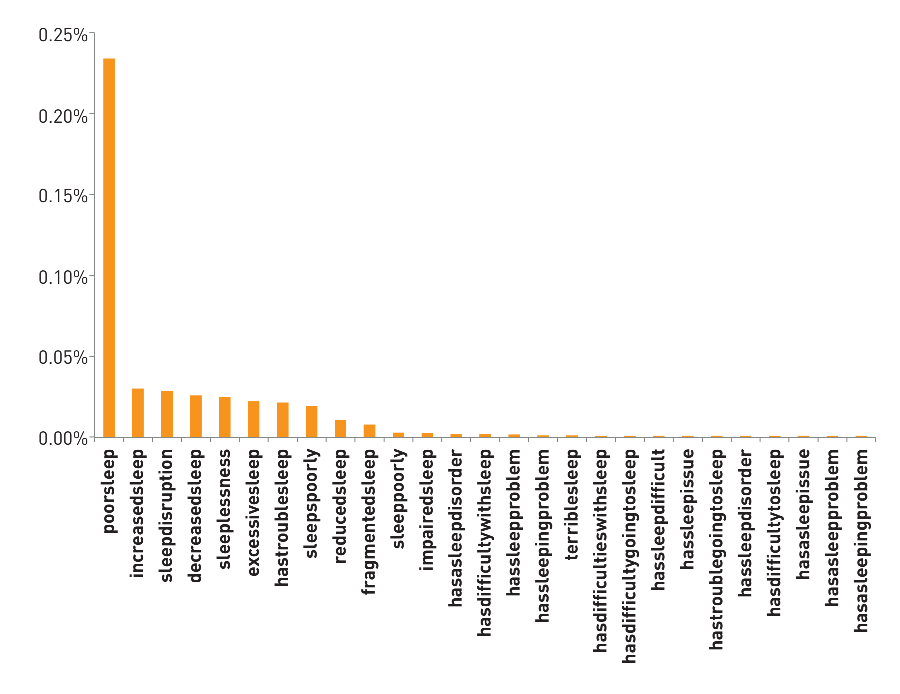


Figure 5. Most prevalent sleep disorder expressions. The y-axis represents the percentage of notes that contain the expression within a sample of 1 million randomly selected notes.

expressions that were found in at least 0.05 percent of the notes, as shown in Figure 4.

DISCUSSION

This article presents a novel method of determining smoking status using clinical narrative notes. TN substantially relies on the fact that physicians tend to use similar expressions to describe medical conditions and, further, tend to use these expressions consistently. Converting all expressions and notes to alphabetical-only representations eliminates the heterogeneity in the descriptions of the medical descriptors and allows a perfect

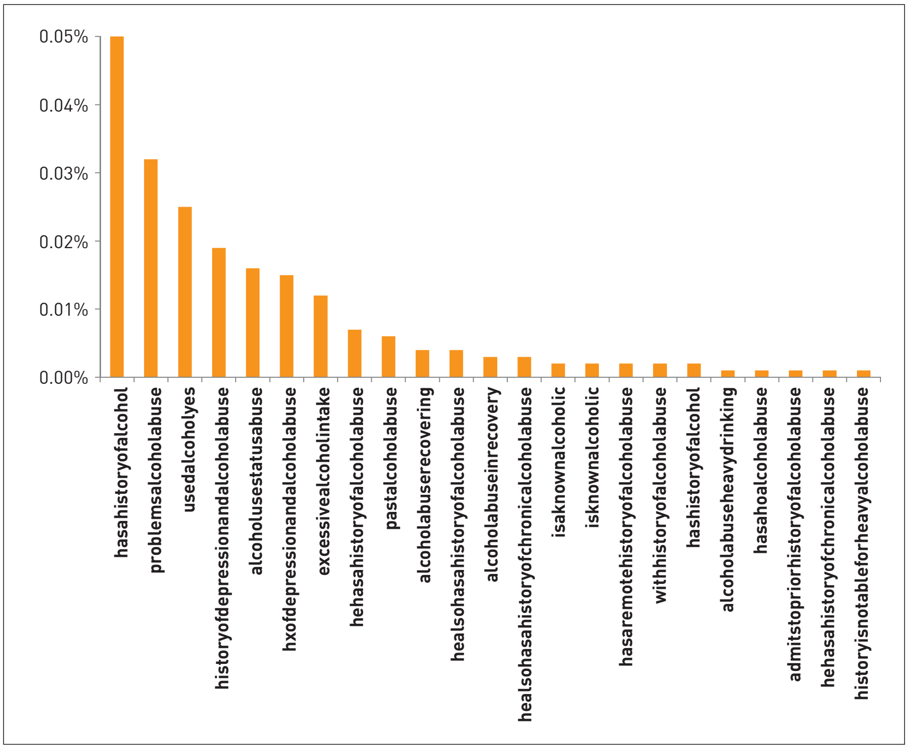


Figure 6. Most prevalent alcohol-use expressions. The y-axis represents the percentage of notes that contain the expression within a sample of 10,000 randomly selected notes.

match between an expression and a note that may contain the expression (e.g., "hastroublegoingtosleep" and "hasahistoryofalcohol," as presented as examples in Figures 5 and 6, respectively). In traditional machine-learning approaches for text classification, a human expert is required to label phrases or entire notes, and then a supervised-learning algorithm attempts to generalize the associations and apply them to new data. In contrast, using non-negated distinct expressions eliminates the need for an additional computational method to achieve generalizability, as the expressions have consistently been found highly prevalent across multiple clinical conditions by considering more than 10 million clinical narrative notes. TN thus provides distinct classifications and is thereby expected to provide robust results.

TN yields a high performance for the determination of a variety of clinical conditions using rapid processing (i.e., approximately 1 millisecond per note on average); however, no benchmark for processing performance comparison has yet been reported for similar tasks. We also extended TN for uses beyond extracting smoking status. For instance, we used TN to extract family history of coronary artery disease, classify patients with sleep disorders, improve the accuracy of the Framingham risk score for patients with nonalcoholic fatty liver disease, and classify nonadherence to T2DM (see past projects: http://researcher.ibm.com/researcher/view_person_pubs.php?person=ibm-Uri.Kartoun).

Further, TN could be used to enhance the standard regular expression pattern language (in which a sequence of characters defines a search pattern). Applying standard regular expressions relies on knowing a priori the patterns to search for, and this is exactly what TN's human-in-the-loop step addresses (Figure 2). The regular expression pattern language can benefit from TN's initial identification of a collection of phrases to match.

An additional advantage of TN is that it is not sensitive to negations and is capable of ignoring them efficiently (e.g., "no longer smokes," "not a cigarette smoker," and "has not smoked since"). Further, TN does not require setting a priori, subjectively selected, or data-dependent configuration parameters such as those required when using SVMs, which was the most popular approach used at the i2b2 smoking status natural language processing challenge [6]. Another advantage unique to TN is its potential applicability to languages other than English because the human-in-the-loop procedure is not tied to a specific language. For instance, the phrase "smokes two packages per day" would be "smokestwopackagesperday" in English, "מעשןשתיחפיסותביום " in Hebrew, "每天抽兩包煙" in Chinese, and "fumadoispacotespordia" in Portuguese.

TN has several limitations. First, TN requires that a human dedicate time to identifying candidate expressions. While TN requires only a few human hours for the task of classifying smoking status, performing more complex tasks (e.g., identifying complications after a surgery) would require additional time. However, when this effort is complete, the identified expressions can be generalizable and could be deployed on any database and used by the research community. In addition, the language describing an individual's smoking status might be quite diverse in various places. Per Zipf's law, English has an infinite number of possible expressions, so one cannot enumerate all the ways to describe smoking status. However, across varied conditions, the results demonstrate that the tail of the distributions can be ignored (as seen in Figures 4–6).

Humans are the ones who created letters and languages, and therefore we are capable of accurately identifying highly descriptive non-negated expressions. Similar to my Ph.D. dissertation in which I described how a collaboration between a human and a robot can expedite a learning task [4], my research on TN demonstrated that an interactive human-in-the-loop extension, applied here on large collections of clinical narrative notes, can produce high classification accuracy across distinct medical conditions. In conclusion, TN is a rapid, accurate, and easily adaptable method of identifying patients' clinical descriptors by interacting with clinical narrative notes. The use of TN allows for accurate and comprehensive identification of many medical conditions, which will improve precision and recall values in studies that rely on textual data.

ACKNOWLEDGMENTS

I would like to thank Professor Peter Szolovits of MIT's Department of Electrical Engineering and Computer Science for his constructive feedback regarding Text Nailing. I further wish to express my deepest gratitude to Drs. Stanley Shaw and Kathleen Corey for their priceless guidance during my fellowship at MGH.

Endnotes

1. Kartoun, U. The man who had them all. *ACM Interactions 24*, 4 (July–Aug. 2017), 22–23.
2. Kartoun, U., Kumar, V., Cheng, S.C., Yu, S., Liao, K., Karlson, E., Ananthakrishnan, A., Xia, Z., Gainer, V., Cagan, A., Savova, G., Chen, P., Murphy, S., Churchill, S., Kohane, I., Szolovits, P., Cai, T., and Shaw, S. Demonstrating the advantages of applying data mining techniques on time-dependent electronic medical records. *Proc of AMIA 2015 Annual Symposium.*
3. Savova, G.K., Ogren, P.V., Duffy, P.H., Buntrock, J.D., and Chute, C.G. Mayo clinic NLP system for patient smoking status identification. *Journal of the American Medical Informatics Association 15*, 1 (2008), 25–28.
4. Kartoun, U. Human–robot collaborative learning methods. Ph.D. Dissertation. Ben-Gurion University of the Negev, Israel. 2008.
5. Kartoun, U.*, Beam, A.*, Pai, J., Chatterjee, A., Fitzgerald, T., Kohane, I.*, and Shaw, S*. The spectrum of insomnia-associated comorbidities in an electronic medical records cohort. *Proc. of AMIA 2016 Annual Symposium.* *Contributed equally.
6. Uzuner, O., Goldstein, I., Luo, Y., and Kohane, I. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association 15*, 1 (2008), 14–24.

Uri Kartoun is a research staff member at IBM Research in Cambridge, MA. Previously he was a research fellow at Harvard Medical School/Massachusetts General Hospital. His Ph.D. from Ben-Gurion University of the Negev, Israel, focused on human-robot collaboration.
→ uri.kartoun@ibm.com

DOI: 10.1145/3139488 © 2017 ACM 1072-5520/17/11 $15.00