Improving the Utility of Tobacco-Related Problem List Entries Using Natural Language Processing

Daniel R. Harris, Ph.D[1,2], Darren W. Henderson[2], Alexandria Corbeau[2]
[1]Institute for Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Kentucky, Lexington, Kentucky 40506; [2]Center for Clinical and Translational Sciences, University of Kentucky, Lexington, KY 40506.

Abstract

We present findings on using natural language processing to classify tobacco-related entries from problem lists found within patient's electronic health records. Problem lists describe health-related issues recorded during a patient's medical visit; these problems are typically followed up upon during subsequent visits and are updated for relevance or accuracy. The mechanics of problem lists vary across different electronic health record systems. In general, they either manifest as pre-generated generic problems that may be selected from a master list or as text boxes where a healthcare professional may enter free text describing the problem. Using commonly-available natural language processing tools, we classified tobacco-related problems into three classes: active-user, former-user, and non-user; we further demonstrate that rule-based post-processing may significantly increase precision in identifying these classes (+32%, +22%, +35% respectively). We used these classes to generate tobacco time-spans that reconstruct a patient's tobacco-use history and better support secondary data analysis. We bundle this as an open-source toolkit with flow visualizations indicating how patient tobacco-related behavior changes longitudinally, which can also capture and visualize contradicting information such as smokers being flagged as having never smoked.

Introduction

Problem lists capture relevant health-related issues and are a natural component of a modern electronic health record (EHR) systems; conceptually the idea of a problem list has existed for around half of a century[1]. The importance of problem lists have been elevated in the past decade due to changes in meaningful use with the goal of capturing problem information electronically[2]. Before 2013, problem lists often used the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9-CM) if they were coded; ICD9-CM was largely inadequate for anything other than administrative purposes[3,4]. The Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) is an alternative to ICD9-CM and is better capable of capturing clinical issues[5]; a study determined that SNOMED-CT can describe 92.3% of clinical problems accurately[6]. Despite this, historical adoption for SNOMED-CT in EHRs has been slow[7–9].

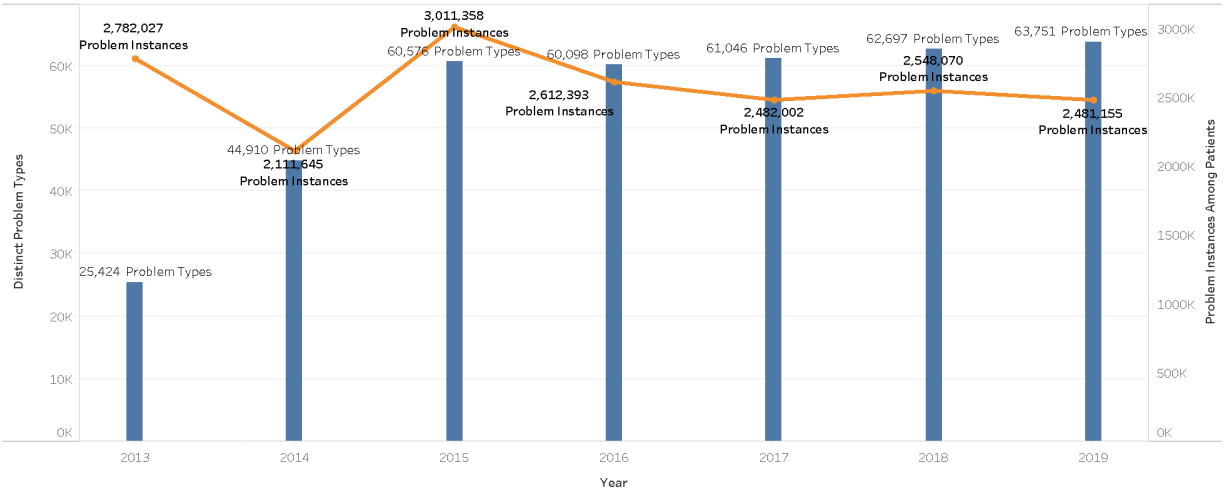

Figure 1: Distinct problem types and instances of problems over time (Kentucky)

Since 2013, the International Classification of Diseases, Tenth Revision, Clinical Modification/Procedure Coding System (ICD10-CM/PCS) has been used to describe entries in problem lists with the benefit of higher specificity than ICD9-CM. Despite this, information gaps still exist between ICD10-CM/PCS and SNOMED-CT where ICD10-CM/PCS is adequate for administrative purposes yet does not possess the depth and precision required for general clinical use[10]. In Figure 1, we show how problem lists have changed across time in the outpatient EHR of our university's hospital and clinic network. In 2014, a large increase in unique problem types stemmed from the inclusion of ICD10 terms to better support meaningful use initiatives. For 2016 and beyond, unique problem types steadily climb each year while instances of problems associated with patient records remain relatively stable; this phenomena might be explained by older generic problem types being replaced with newer problem types with higher specificity.

In Figure 2, we visualize frequency of new problems being assigned to patients within the EHR and again see a spike in 2014; in this picture, a problem type is counted if it has never been used in the EHR up until that point. Although this trend does begin decreasing after 2014, the numbers are still large enough in 2019 to suggest the need for automatic approaches for understanding and utilizing problem lists. The steady increase in unique problems being leveraged by the EHR further suggests the need for automatic approaches. We leverage natural language processing (NLP) to help fill this need.

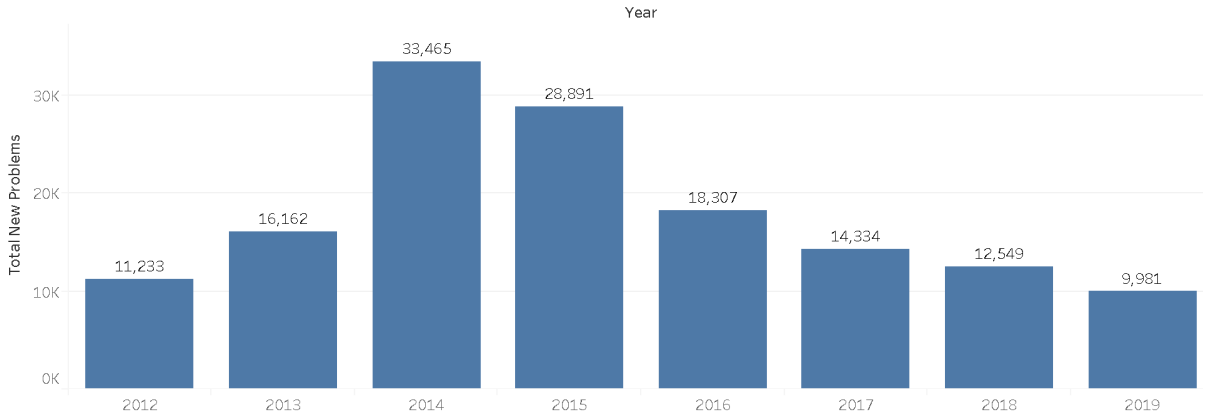

Figure 2: New problem types introduced over time (Kentucky)

The NLP research community has a long-held interest in problem lists. NLP has largely been used to automate or partially automate the generation of problem lists from clinical notes [11–15]. NLP has also been used to identify domain-specific problems[16], association discovery[17], data linkage from problems for decision support[18], and clustering of similar problems[19]. The importance of having an accurate tobacco-use status in an EHR is well-documented[20]. Tobacco-use status is a popular topic in the NLP community, where methods have focused on detecting the absence or presence of smoking within a patient's EHR[21–23]. Issues persist within EHRs which complicate the reliability of tobacco-related NLP processes[24] and may suggest targeted NLP as an easier method to contextualize and integrate tobacco-related findings. Our work focuses specifically on increasing the utility of tobacco-related entries of problem lists by classifying them into higher-level tobacco statuses (current, former, non) and in leveraging these statuses to generate tobacco-use time-spans for secondary data analysis and visualization. A previous study found that roughly a third of patients located at a well-established university medical center had conflicting information regarding their smoking status across time[20]; we also describe a rule-based method for creating time spans for smoking statuses and resolving contractions.

Methods

We leverage the Clinical Language Annotation, Modeling, and Processing (CLAMP) toolkit[25] to process problem lists from the outpatient EHR of our university's network of hospitals and clinics. CLAMP focuses on building reusable pipelines for specific NLP tasks. The smoking status pipeline was trained with sentences containing tobacco-related words; initial experiments yielded an accuracy of 0.95, 0.89, and 0.90 for non-user, current-user, former-user classes respectively[25]; these classes sufficiently met our needs as we did need the exact semantics or concepts extracted from

the problems. We hypothesized that CLAMP would perform well with the short text of problem lists given that it was trained with individual tobacco-related sentences. The general components of the smoking status pipeline is visualized in Figure 3. Problem lists are retrieved from our EHR and normalized on capitalization to avoid redundant processing; for example, small variants such as "Smokes two packs per day" and "smokes two packs per day" are merged. The problems are split into individual files for CLAMP to process. Each problem is sent through a smoking NLP pipeline. Despite each problem being a short-form sentence, the process begins with sentence detection in order to feed a downstream tokenizer and part-of-speech tagger. Tagged words are fed into a named entity recognizer and an assertion classifier for negation detection with NegEx[25,26]. Everything feeds the last step of the CLAMP pipeline: a rule-based text annotation engine based on UIMA Ruta[27].

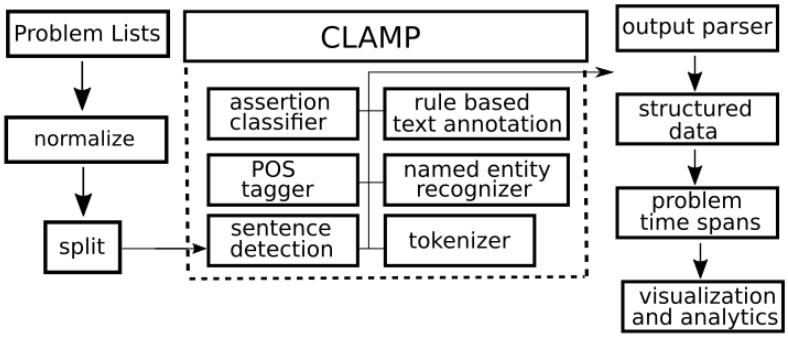

Figure 3: A high-level overview of the CLAMP smoking pipeline

The output of CLAMP is parsed and collated into a structured data set. This data set is used to develop time spans of smoking statuses per patient, where each status is associated with a start and stop date. Figure 4 shows a simple plot of smoking-related statuses. We use these spans to determine smoking status at a given point in time and to do basic quality checks in order to detect where contradictions may occur. For example, Figure 4 shows a patient being tagged as never having used tobacco after previous problems indicated that they have some type of tobacco use history. This contradicting data point can be safely removed from consideration. We later show an aggregated flow diagram in our discussion section and review its utility.

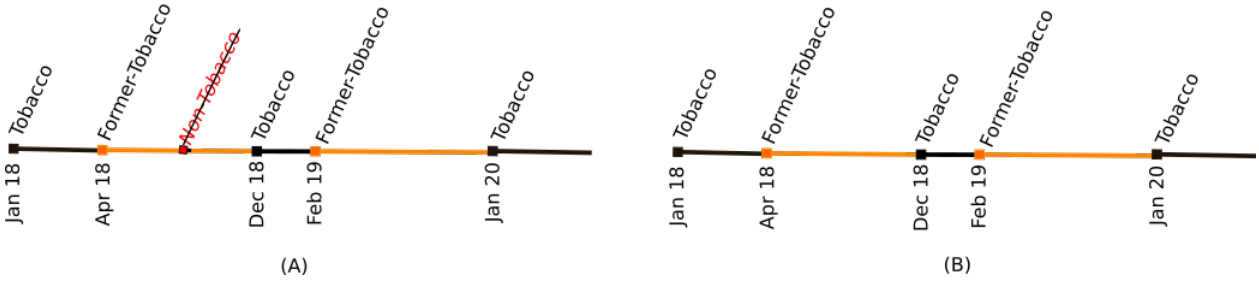

Figure 4: Visualizing a single patient's tobacco-related statuses over time (A) and fixing contradictions (B)

It is known that careful selection of pre-processing and post-processing steps greatly impacts the accuracy of text classification tasks[28]. We processed all problems listed in our EHR's dictionary of problems; 1667 problems came back associated with tobacco use. Our initial pass of this resulted in many results unrelated to smoking and informed us on how we may filter as a pre-processing step. We excluded a subset of these problems by using simple pattern matching against the problem's text.

Because our interest in tobacco use status is specific to a patient's behavior, we performed a manual review of initial results and identified eight categories of terms that should be excluded. These categories for exclusion are summarized in Table 1 with examples of each. We specifically exclude passive tobacco use through either exposure, second-hand, or maternal-use smoking references; these health-related environmental factors can be handled separately. Our goal

Table 1: Exclusionary terms for processing problem list

Exclusion Terms Category	Matching Problems	Example
exposure	145	daily exposure to tobacco smoke
smoke (from fire)	127	toxic effect of smoke, unintentional
second-hand	62	2nd hand tobacco smoke
maternal	31	fetus and newborn affected by maternal use of tobacco
unknown	23	current smoking status unknown
drug-related	10	smokes drugs through pipe
family history	10	family history of tobacco abuse
smoke detectors	7	no smoke detectors in home

was to identify direct tobacco use to supplement administrative tobacco-use billing codes to accurately reconstruct someone's tobacco use history for data warehousing purposes.

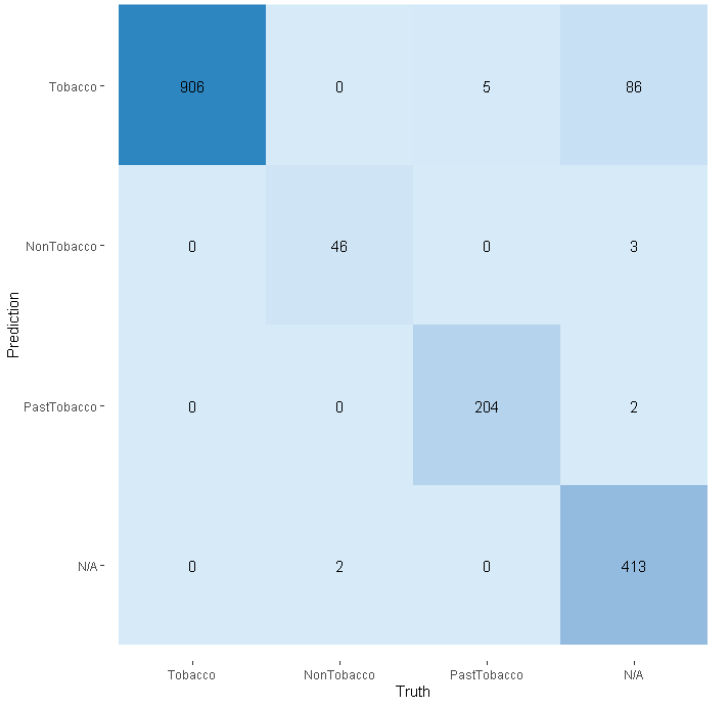

Figure 5: Confusion matrix for tobacco status classification

Results

Three subject-matter experts from our clinical research team independently reviewed the classification assigned per problem by CLAMP. Fleiss' kappa was 0.675, indicating significant agreement among reviewers[29]. Reviewers judged correctness of CLAMP's tobacco-status assignment and annotated the problem as being context sensitive. A context-sensitive problem is too ambiguous to truly judge smoking status; additional information from the EHR may contextualize the problem and resolve ambiguity. For example, "age started smoking" may either refer to current tobacco-use or former-tobacco use; however, additional information from the EHR may be able to resolve the ambiguity, such as an additional tobacco-related problem or a tobacco-related diagnosis code.

The majority vote determined the collective reviewer's decision on ground truth on a problem's true class and subsequently determined the correctness per CLAMP classification. Figure 5 contains a confusion matrix detailing how the predicted class and true classes align; the correctness of the assigned classes after each step of our workflow is

detailed in Table 2. The list of exclusions described in Table 1 were used to filter the raw data; the appropriateness of our filters was further verified by having no loss of true-positive classifications for problems post-filtering. The bulk of the problems classified incorrectly were truthfully indeterminate of tobacco status (N/A) and were safely removed from consideration during later analysis and visualization steps.

Table 2: Performance of classification of problem lists

		Precision			Recall	
Class	**Orig. Count**	**Raw Majority**	**Post-Exclusions**	**Post-Rules**	**Post-Rules**	**Final Count**
Tobacco	1499	0.59	0.80	0.91	1.0	997
Former-Tobacco	102	0.77	0.88	0.99	0.98	206
Non-Tobacco	66	0.59	0.75	0.94	0.96	49

In addition to removing obviously wrong candidates with exclusion rules, we successfully created very simple rules for adjusting the class assigned to each problem. For example, the occurrence of "non-user" and "abstinence" would force the class assigned to non-tobacco user, regardless of the class assigned by CLAMP. In total, there were 3 simple rules for adjusting non-tobacco problems, 12 rules for adjusting former-tobacco use, and 10 rules for adjusting current tobacco-use. The rules are documented in our online package for NLP with tobacco-related problems.

Table 3: Common types of classification mistakes and examples

Error Category	**Example**	**Computed Class**	**Actual Class**
Missed negation	not ready to quit smoking	Non-Tobacco	Tobacco
Verb tense confusion	has never tried to quit using tobacco	Non-Tobacco	Tobacco
Negation confusion	treatment not carried out due to patient smoking	Non-Tobacco	Tobacco
Missed negative verb	declined smoking cessation	Non-Tobacco	Tobacco
Missed negative adjective	ex-heavy cigarette smoker (20-39 per day)	Tobacco	Past-Tobacco
Missed "remission" context	severe tobacco dependence in remission	Tobacco	Past-Tobacco
Missed "history" context	history of prior cigarette smoking	Tobacco	Past-Tobacco
Missed "non-user" context	tobacco non-user	Tobacco	Non-Tobacco
Missed qualification	ready to quit smoking	Past-Tobacco	Tobacco
Missed negative adjective	tried unsuccessfully to quit smoking in the past	Past-Tobacco	Tobacco
Verb tense confusion	would like to quit tobacco use	Past-Tobacco	Tobacco
Unexplained	has reduced amount of tobacco smoked	Past-Tobacco	Tobacco

Recall that our manual reviewers also annotated the problems as being context-sensitive or not. This potentially helps us understand why a human judge might disagree with CLAMP's classifier. In Table 4, we described each class computed by CLAMP in terms of what percentage were incorrect (post-exclusions and post-rules); of those labeled as incorrect, we report what percentages were viewed as context-sensitive (CS) by the majority vote of reviewers.

Table 4: Impact of context-sensitive problems

Computed Class	**Incorrect**	**C.S.**	**Example**
Tobacco	0.09	0.62	age of onset of smoking
Former-Tobacco	0.01	1.0	lung cancer screening for patient with less than a 30 pack year history
Non-Tobacco	0.06	1.0	has not smoked cigarettes within the last year

All erroneous mappings for non-tobacco and former-tobacco problems were context-sensitive; 62% of tobacco problems were context-sensitive. It is our belief that neither an automated method nor a human could definitively determine that this problem should be classified as tobacco, former-tobacco, or non-tobacco. For example, "lung cancer screening for patient with less than a 30 pack year history" is likely either a current smoker or former smoker; it is not

absolutely clear which one may be the actual case and in fact, non-smoker is technically possible as well due to the "less than" phrasing. "has not smoked cigarettes within the last year" may be interpreted as a former tobacco user or a non-tobacco user; it is not clear from this limited context if the patient has ever smoked.

Discussion

We discuss additional considerations when using natural language processing. Namely, favorable gaps may exist between theoretical implementation and practice, where the commonly-used problems are the easiest for machines and humans to understand. Administrative data may not align well with problem list data; this makes problem lists an attractive and necessary addition to understanding a patient's tobacco profile. Additionally, contradictions may occur within the EHR. We will discuss each of these in turn.

Observed vs Theoretical Results

The results presented in Table 2 describe how accurately our methods map problem lists to higher-level classes. Not every problem listed in our EHR's dictionary of problems has been associated with an actual patient. Table 5 describes the perceived precision given the subset of the problems from the problem dictionary which are actually assigned to patients in the EHR. All problems assigned to patients and also deemed to be related to former tobacco use were correct; the context-sensitive problems related to former-tobacco status were not used in practice. The tobacco class saw the largest increase (8%) in theoretical dictionary precision (91%) versus the observed in-practice precision (99%). These results indicate the practical utility of NLP for problems currently attached to patient data. We primarily focus our efforts on processing the entire dictionary of problems because healthcare providers are free to select any existing problems or submit a new problem description for their patients.

Table 5: Precision of problems being used with patients

Computed Class	Precision	Precision (Used Only)	N=Correct & Used
Tobacco	0.91	0.99	261
Former-Tobacco	0.99	1.0	110
Non-Tobacco	0.94	0.96	26

Administrative Data

Our motivation for analyzing smoking statuses from problem lists was due in part to the lack of reliable structured fields in our local EHR. Limitations exist in using ICD10 billing diagnoses for research. In particular, tobacco-related codes are associated with limited sensitivity[30,31] which can be improved when combined with NLP[31]. We cross-referenced a patient's problem list with their billing diagnosis codes and checked for inconsistencies; we looked within a two week window of the problem's date for any tobacco-related diagnosis codes. We summarize our findings in Table 6.

Table 6: Comparison of problem lists and administrative/billing data

Class	Matched	Mismatch	Missing	Correct (Among Non-Missing)
Tobacco	0.43	0.57	0.53	0.93
Former-Tobacco	0.19	0.81	0.73	0.70
Non-Tobacco	0.01	0.99	0.96	0.38

The largest issue encountered was patients with tobacco-related problems missing any tobacco-related diagnosis codes in their billing data. For all three tobacco-related statuses, missing data accounted for the majority of the errors. If a tobacco-related code was found for a current tobacco-related problem, it was likely to be correct. The minority of errors are those regarding whether smoking status is present or past. There are examples of a patient with a 'tobacco use disorder' problem being billed as someone with "nicotine dependence in remission"; the lack of specificity in the problem creates a semantic mismatch. Former-tobacco problems were less reliable than current tobacco-use problems; the disagreements again center around timing of smoking. As an example of a mismatch between billing and the

problem list, former-smoker problems occasionally map to unspecific nicotine dependence.

The worst performing of the three classes was non-tobacco users which is mostly due to the lack of an appropriate ICD10 billing code. We counted *Z78.9-Other specified health status* as correct for non-tobacco user due to its approximate synonyms list containing "current non-smoker", "never smoked", and "not currently a smoker". This is not ideal since the phrase "current non-smoker" does not supply sufficient information about the patient's past. Furthermore, *Z78.9* contains many synonyms completely unrelated to smoking status, such as "impaired mobility"; this code does not appear to be commonly leveraged within our EHR.

ICD10 Annotations for Problems

The absence of a tobacco-related diagnosis code does not necessarily imply that the patient does not use tobacco and a computed tobacco status of non-user would be stronger evidence based on EHR documentation for that particular person. Problems within our EHR's dictionary are annotated with ICD10 codes, yet these annotations appear incomplete and occasionally inaccurate. Table 7 describes the coverage of these ICD10 annotations in comparison to our NLP-designated classes.

Table 7: Coverage of ICD10-annotations within problem lists

Class	Matched	Mismatch	Missing
Tobacco	0.777	0.015	0.208
Former-Tobacco	0.699	0.003	0.297
Non-Tobacco	0.706	0.000	0.293

Much like our comparison to ICD10 billing codes associated with patient visits, the biggest issue with the ICD10 codes from the problem list dictionary is that codes can be missing. There may exist a lag in the assignment of ICD10 codes associated with problems; our process expedites this assignment and yields broader tobacco statuses which will assist in detecting contradicting information when looking at problem list entries across time.

Contradictions and Limitations

Problem lists assist healthcare providers in documenting any health-related issues per patient during a visit. Although problem lists act as a great organization tool, their utility can be weakened by inaccuracies[32]. One drawback from research leveraging longitudinal analysis of problem lists is that patients must have a visit in order for the list to be updated. Additionally, the healthcare professional must update the list upon seeing a patient. For a topic such as smoking, it is plausible that a patient may change statuses any number of times between visits. The accuracy of problem lists in general is questionable and potentially problematic in any downstream use of the data; most of these issues are not impossible to solve and research continues to provide motivation in finding solutions[32].

Another issue is that problem lists may contradict themselves across time. For example, someone who was once recorded as being a smoker may accidentally be flagged in the future as someone who has never smoked. As a post-processing step, we generate time spans for a patient's tobacco-use profile using the tobacco-status of problems recorded during visits over time. These time spans demonstrated a significant amount of erroneous transitions in the form of tobacco or past-tobacco users being switched to non-tobacco users. We visualize the flow of changes between tobacco statuses as an alluvial diagram in Figure 6. This figure was constructed by taking a subset of our patients having tobacco-related problems with healthcare visits in 2017, 2018, and 2019; a patient's status for the year was their most charted status based on their historical problem list. Contradicting status transitions from current-tobacco status to non-tobacco status were seen in patients (N) during both 2017 to 2018 (N=534) and 2018 to 2019 (N=497); additionally past-tobacco status transitioned to non-tobacco status in patients (N) during both 2017 to 2018 (N=516) and 2018 to 2019 (N=494).

To combat contradicting information, we deployed rules during post-processing in order to assert a patient's correct status based on evidence from prior problems. Once tobacco use is observed, a patient's minimum allowable class

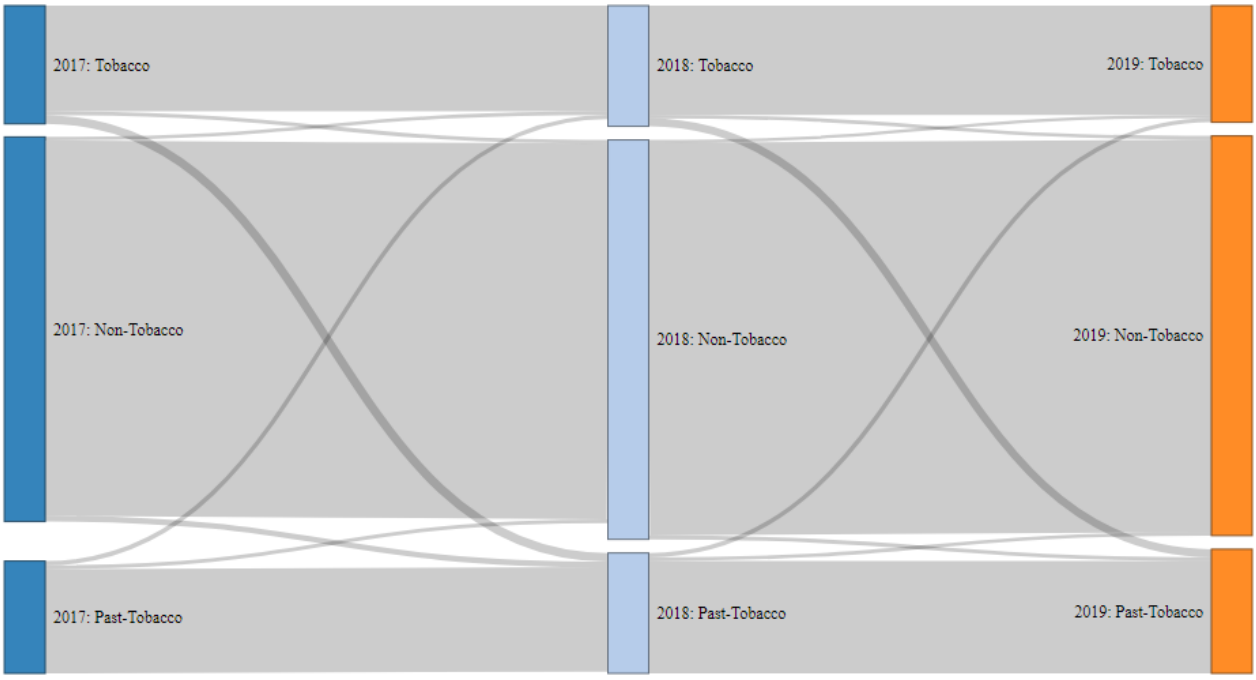

Figure 6: Capturing patients and their recorded tobacco-use changes over time

becomes former-tobacco user and problems that suggest non-tobacco status are flagged for review and hidden. The weakness of this approach is that if a prior problem erroneously flagged someone as a tobacco user, then this erroneous judgement would be carried downstream and would impact the patient's smoking status time spans. We suspect this would be a rare occurrence, yet we have not validated our assumption. The time span logic assists us accomplish what is visualized in Figure 4; the removal of contradicting information helps create clean spans of smoking statuses. Validation of the time spans is difficult because smoking documentation is often limited to the problem lists themselves; a preliminary scan indicated that only a small fraction of smokers had ancillary diagnoses or medications on record as further evidence for smoking,

Figure 6 can also be used to visualize positive and negative changes in collective tobacco-use behavior across all patients. In both periods 2017 to 2018 and 2018 to 2019, approximately 8% of patients transitioned from current tobacco users to former tobacco users. Similarly, both 2017 to 2018 and 2018 to 2019 transitions saw 4% of former-tobacco users become current tobacco users again. The net result of these transitions yield a shrinking tobacco-using population and a growing population of former tobacco users.

Our code for pre-processing and post-processing CLAMP results is included in our open-source toolkit, a CLAMP companion named the **T**obacco-**R**elated **A**nalyses for **P**roblem-Lists (TRAP) toolkit[33]. We also bundle code for creating time spans, adjusting time spans to correct contradicting information, and code for creating the alluvial chart from Figure 6.

Conclusion

We demonstrate that the utility of problem lists commonly found within EHR systems may readily be improved for research purposes by leveraging easily-accessible natural language processing tools. In particular, the results of processing problem lists can greatly benefit from simple rule-based post-processing. We plan to compare our results against retraining the tobacco status classifier using problem list data; the benefit of post-processing is that the NLP tool can be used out of the box with very few barriers to success. In the future, we wish to marry classification and concept extraction in order to obtain more details per problem. Additionally, we aim to validate our findings by replicating our work using data from alternate healthcare systems and EHRs. We plan to explore other healthcare

statuses that are temporal in nature; drug and/or alcohol use statuses may benefit from a similar time span analyses and contradiction detection.

Acknowledgment

The project described was supported by the NIH National Center for Advancing Translational Sciences through grant number UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Weed LL. Medical records that guide and teach (concluded). Yearbook of Medical Informatics. 1968;212:1.
2. Department of Health and Human Services. Health information technology: initial set of standards, implementation specifications, and certification criteria for electronic health record technology. Final rule. Federal Register. 2010;75(144):44589–654.
3. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR, on Codes & Structures CBPRIWG. The content coverage of clinical classifications. Journal of the American Medical Informatics Association. 1996;3(3):224–233.
4. Payne TH, Murphy GR, Salazar A. How well does ICD9 represent phrases used in the medical record problem list? In: Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association; 1992. p. 654.
5. Campbell JR, Payne T. A comparison of four schemes for codification of problem lists. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association; 1994. p. 201.
6. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. In: Mayo Clinic Proceedings. vol. 81. Elsevier; 2006. p. 741–748.
7. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. Journal of the American Medical Informatics Association. 2014;21(e1):e11–e19.
8. Elhanan G, Perl Y, Geller J. A survey of direct users and uses of SNOMED CT: 2010 status. In: AMIA Annual Symposium Proceedings. vol. 2010. American Medical Informatics Association; 2010. p. 207.
9. Lee D, Cornet R, Lau F, De Keizer N. A survey of SNOMED CT implementations. Journal of biomedical informatics. 2013;46(1):87–96.
10. Steindel SJ. A comparison between a SNOMED CT problem list and the ICD-10-CM/PCS HIPAA code sets. Perspectives in Health Information Management/AHIMA, American Health Information Management Association. 2012;9(Winter).
11. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC medical informatics and decision making. 2005;5(1):30.
12. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. Journal of biomedical informatics. 2006;39(6):589–599.
13. Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. International journal of medical informatics. 2008;77(9):602–612.
14. Solti I, Aaronson B, Fletcher G, Solti M, Gennari JH, Cooper M, et al. Building an automated problem list based on natural language processing: lessons learned in the early phase of development. In: AMIA Annual Symposium Proceedings. vol. 2008. American Medical Informatics Association; 2008. p. 687.

15. Devarakonda MV, Mehta N, Tsou CH, Liang JJ, Nowacki AS, Jelovsek JE. Automated problem list generation and physicians perspective from a pilot study. International journal of medical informatics. 2017;105:121–129.

16. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. International journal of medical informatics. 2015;84(12):1057–1064.

17. Wang L, Wang Y, Shen F, Rastegar-Mojarad M, Liu H. Discovering associations between problem list and practice setting. BMC medical informatics and decision making. 2019;19(3):69.

18. Jao CS, Hier DB, Galanter WL. Using clinical decision support to maintain medication and problem lists A pilot study to yield higher patient safety. In: 2008 IEEE International Conference on Systems, Man and Cybernetics. IEEE; 2008. p. 739–743.

19. Kreuzthaler M, Pfeifer B, Ramos JAV, Kramer D, Grogger V, Bredenfeldt S, et al. EHR problem list clustering for improved topic-space navigation. BMC medical informatics and decision making. 2019;19(3):72.

20. Polubriaginof F, Salmasian H, Albert DA, Vawdrey DK. Challenges with collecting smoking status in electronic health records. In: AMIA Annual Symposium Proceedings. vol. 2017. American Medical Informatics Association; 2017. p. 1392.

21. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. Journal of the American Medical Informatics Association. 2008;15(1):25–28.

22. Heinze DT, Morsch ML, Potter BC, Sheffer Jr RE. Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology. Journal of the American Medical Informatics Association. 2008;15(1):40–43.

23. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association. 2008;15(1):14–24.

24. Chen ES, Carter EW, Sarkar IN, Winden TJ, Melton GB. Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record. In: AMIA Annual Symposium Proceedings. vol. 2014. American Medical Informatics Association; 2014. p. 366.

25. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2018;25(3):331–336.

26. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301–310.

27. Kluegl P, Toepfer M, Beck PD, Fette G, Puppe F. UIMA Ruta: Rapid development of rule-based information extraction applications. Natural Language Engineering. 2016;22(1):1–40.

28. Uysal AK, Gunal S. The impact of preprocessing on text classification. Information Processing & Management. 2014;50(1):104–112.

29. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin. 1971;76(5):378.

30. Desai RJ, Solomon DH, Shadick N, Iannaccone C, Kim SC. Identification of smoking using Medicare data—a validation study of claims-based algorithms. Pharmacoepidemiology and drug safety. 2016;25(4):472–475.

31. Wang L, Ruan X, Yang P, Liu H. Comparison of three information sources for smoking information in electronic health records. Cancer informatics. 2016;15:CIN–S40604.

32. Holmes C. The problem list beyond meaningful use: part 2: fixing the problem list. Journal of AHIMA. 2011;82(3):32–35.

33. CLAMP-TRAP. Bitbucket.org; 2020. Available from: `https://bitbucket.org/_harris/clamp-trap`.