Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record

**Yan Wang, PhD[1], Elizabeth S. Chen, PhD[4],
Serguei Pakhomov, PhD[1, 2], Elizabeth Lindemann, BS[1], Genevieve B. Melton, MD, PhD[1, 3]
[1]Institute for Health Informatics, [2]College of Pharmacy, and [3]Department of Surgery,
University of Minnesota, Minneapolis, MN;
[4]Center for Biomedical Informatics, Brown University, Providence, RI**

Abstract

The electronic health record (EHR) provides an opportunity for improved use of clinical documentation including leveraging tobacco use information by clinicians and researchers. In this study, we investigated the content, consistency, and completeness of tobacco use data from structured and unstructured sources in the EHR. A natural language process (NLP) pipeline was utilized to extract details about tobacco use from clinical notes and free-text tobacco use comments within the social history module of an EHR system. We analyzed the consistency of tobacco use information within clinical notes, comments, and available structured fields for tobacco use. Our results indicate that structured fields for tobacco use alone may not be able to provide complete tobacco use information. While there was better consistency for some elements (e.g., status and type), inconsistencies were found particularly for temporal information. Further work is needed to improve tobacco use information integration from different parts of the EHR.

Introduction

Social and behavioral factors such as tobacco, alcohol, and drug use are increasingly recognized as key factors for many causes of disease, disability, and mortality in the United States. A number of studies have been published describing the linkage between behavioral risk factors and their associated morbidity or mortality[1-4]. For example, worldwide, direct tobacco use is responsible for more than 5 million deaths each year[5]. The National Academy of Medicine (NAM; formerly Institute of Medicine) have emphasized the need for improving existing datasets, developing new data sources, and establishing strategies and models for incorporating social and behavioral factors and their interactions in its 2006 report on "Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate"[6]. In a recent NAM report on "Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2", tobacco use and exposure was featured among the domains recommended for inclusion in the electronic health record (EHR)[7].

The widespread adoption of EHR systems, in turn, provides an opportunity for clinicians and researchers to access a large amount of information about an individual's social history including substance use. Within EHR systems, documentation of social history, including tobacco use, can range from structured and coded data to free-text narrative. A number of studies have focused on the examination and representation of social history information[8-12]. Other studies have involved developing natural language processing (NLP) techniques for the automated identification and extraction of substance use information, with a particular emphasis on tobacco use[8, 13-18].

In this study, we sought to investigate tobacco use information collected in multiple structured and unstructured sources within an EHR system. NLP techniques, as previously described, were used to extract tobacco use statements from free-text for comparison with structured sources in order to characterize content, consistency, and completeness of this type of information for patients within a single system.

Background

In early work, a multi-institutional study was conducted to characterize social history information in clinical notes from different sources (MTSamples[19], University of Vermont Medical Center [UVMMC; formerly Fletcher Allen Health Care], and University of Minnesota-affiliated Fairview Health Services [FHS])[10]. We evaluated adequacy of several existing models including HL7 CDA-based models[20] and openEHR[21] archetypes for representing social history information. From this, initial models for tobacco, alcohol, and drug use were developed in this study. Table 1 illustrates part of the model used to represent information within a social history statement for tobacco use.

In a 2014 follow-up study, 525 tobacco use entries from the social history module of the Epic EHR at UVMMC, including structured fields (e.g., for smoking status, type, and frequency) and a free-text comment field, were manually reviewed to characterize the contents and quality issues of the free-text comments[12]. Results from the study showed a range of potential data quality issues between the structured fields and free-text comments.

To exploit the information from clinical text, we developed a NLP pipeline for detecting substance use (alcohol use, drug use and tobacco use) statements and extracting relevant elements of substance use[22]. Unlike many of the previous NLP studies included in the 2006 i2b2 challenge which focused on the extraction of smoking status[8], our goal was to extract additional semantics related to tobacco use including tobacco use beyond smoking (e.g., smokeless tobacco) and smoking status (e.g., pack-use and temporal information). In this work, we leveraged existing linguistic resources and domain knowledge from the earlier studies, as well as the Propbank[23] resource and the MiPACQ[24] corpus to boost extraction performance for tobacco use elements. The resulting NLP tool achieved good performance for extracting a wide breadth of substance use free text information.

Table 1. Elements and values for tobacco use statement type.

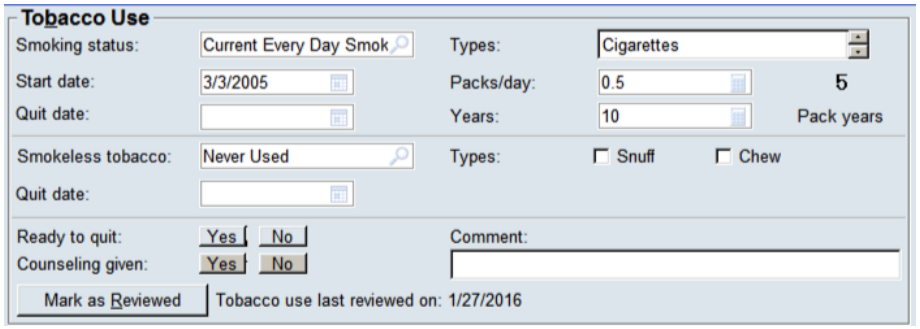
Tobacco use elements	Example value or pattern
Status	current, past, quit
Temporal	[in/since/until] <date>
Method	chew, use
Type	cigars, tobacco
Amount	1 pack per day, <#> ppd
Frequency	occasionally, daily, socially

Method

Setting and Study Design

This study involved a retrospective analysis of tobacco use information collected from clinical notes and the social history module of an enterprise implementation of the Epic EHR (Epic Systems Corporation, Verona, WI)[25] at University of Minnesota-affiliated Fairview Health Services [FHS]). Fairview Health Services had been using EpicCare in one of its physician practice groups (Fairview Physicians) for over eight years. Other practices were supported by Allscripts, Eclipsys SCM, McKesson Paragon, and paper processes. Starting October 2010, two paper-based Fairview regional hospitals successfully went live with Epic clinical and revenue cycle applications. Then in March 2011, Fairview's two largest hospitals which are academically-based went live simultaneously.

In the social history module of Epic EHR, each entry includes structured fields for tobacco use information such as smoking status, start date, quit date, as well as a free-text entry for comments as shown in Figure 1. While some clinicians use templates that pull in information from the social history module, many clinicians continue to document tobacco use information within clinical notes in free text format outside of the tobacco use module most often as part of a social history section in the note (e.g., "The patient continued to smoke about half pack a day."). Most of the structured social history data is entered by nurses and medical assistants during ambulatory patient visits based on direct answers from patients or on questionnaires completed by patients. Tobacco use information within a note is usually entered during patient visits and is considered a required element for face-to-face ambulatory encounters in our healthcare system (as well as most others).



Tobacco Use

Smoking status: Current Every Day Smok

Types: Cigarettes

Start date: 3/3/2005

Packs/day: 0.5

5 Pack years

Quit date:

Years: 10

Smokeless tobacco: Never Used

Types: ☐ Snuff ☐ Chew

Quit date:

Ready to quit: Yes No

Counseling given: Yes No

Comment:

Tobacco use last reviewed on: 1/27/2016

Figure 1. Tobacco use data entry in the social history module including comments.
©2016 Epic Systems Corporation. Used with permission.

From 337,506 adult patients (age>=18) who had a most recent social history entry in 2015 and social history entries associated with at least two prior encounters, a set of 384 patients was randomly selected as the cohort for this study to provide for a confidence level of 95%, and margin of error of 5%. For each patient, all tobacco use entries (including structured fields and free-text comments) from the social history module and provider-authored clinical notes (e.g., progress notes, history and physical examinations, discharge summaries, and admission note) were obtained. Figure 2 shows the overall high-level process of the study.

All clinical notes used in this study were collected from the University of Minnesota research clinical data repository. The repository contains documents between 1993 and 2016 from the Epic EHR, as well as documents from affiliate clinics for variable time periods. Extracted notes were pre-processed to add newlines into appropriate places based on text features such as letter capitalization, punctuation, and special letters. Sections (e.g., "Assessment & plan" section in progress note) along with the section header (e.g., "Present illness", "Social history" and "Assessment & plan") within each note were extracted by an NLP section extraction component. Afterwards, text within each section was split into sentences by a sentence splitter component. The approach used in our earlier study[22] was used for processing sentence content as follows. First, a classifier was used to detect tobacco use statements from sentences. Tobacco use statements were then parsed by Stanford parser[26] to obtain constituent and dependency parses of each sentence. The constituent parse of a sentence provides syntactic cues for later tobacco use elements extraction while the dependency parse of a sentence provided the dependency structure of the sentence. A dependency structure represents a directed graph between the tokens of a sentence (e.g., subject, modifier and preposition). With constituent and dependency parses, tobacco use elements ("Amount", "Type", "Status", "Temporal", "Method" and "Frequency") were extracted by a substance use element extraction tool leveraging a large vocabulary of smoking-related items.

To extract the same tobacco use elements (e.g., "Amount" and "Status") from free-text tobacco comments in the tobacco use module (Figure 1), our approach was modified slightly. Values were first collected from structured fields in social history module of Epic EHR system and were mapped to appropriate elements (e.g., "Amount" and "Status). The resulting tobacco use information from structured tobacco use fields, free-text tobacco use comments, and clinical notes with tobacco use statements were the analyzed for content and quality issues within the cohort.

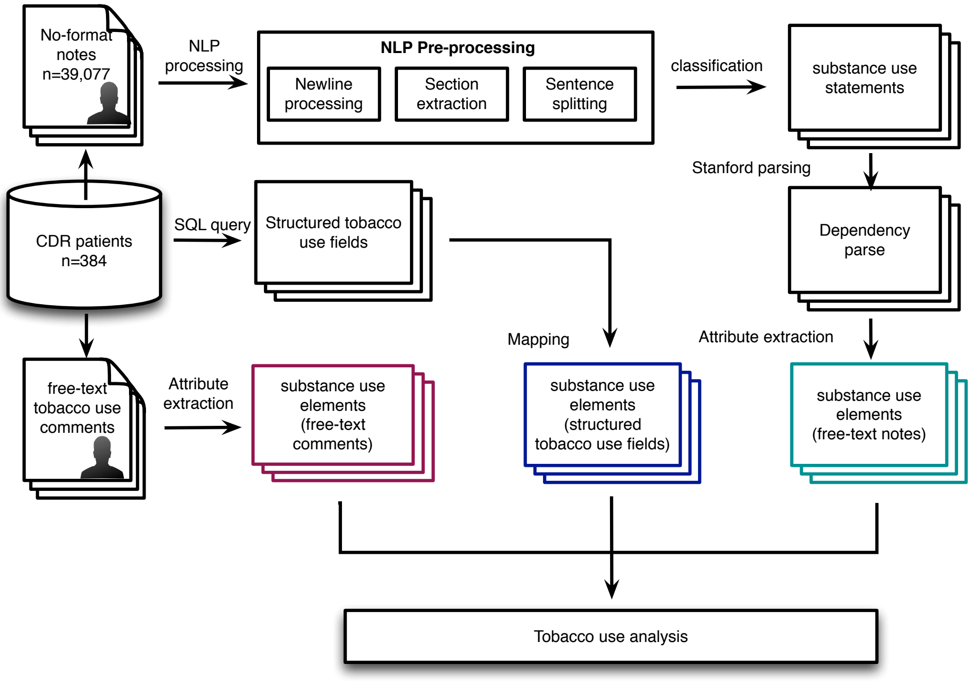


Figure 2. Overall high-level study process resulting in information from free-text comments, structured tobacco use fields, and clinical notes.

Analysis of Tobacco Use Information

The content of tobacco use elements collected from the structured tobacco use fields, free-text tobacco use comments and extracted from clinical notes with tobacco use were characterized as distributions of patients, tobacco use elements and new tobacco use elements patterns of all three tobacco use information sources. A similar coding schema developed and applied in a previous study was used to examine the consistency between structured data and information extracted from unstructured data[12] as shown in Table 2.

As observed in the previous study, a wide range of data consistency issues exists with the tobacco use fields from the social history module and free-text comments. For example, patients may have conflicting information even within a short time span. One patient with respect to smoking status was recorded as "Passive Smoker" in the structured smoking status field while the clinical note from the same day states "trying to quit smoking" and a comment states "quit date 10/30/2009".

Table 2. Examples of observed inconsistencies and discrepancies in tobacco use data at the individual patient level.

Field	Description	Example
Smoking status	Tobacco use statement and tobacco use comment inconsistent with smoking status field	***Smoking status***: Passive Smoker ***Comment***: quit date M1/D1/Y1 ***Statement***: The patient is trying to quit smoking
Packs/day	Tobacco use statement and tobacco use comment inconsistent with Packs/day field	***Packs/day***: 0.5 ***Comment***: half a pack a day ***Statement***: He currently smokes around 1/4 pack of cigarettes per day
Years	Tobacco use statement and tobacco use comment inconsistent with Years of smoking field	***Years***: 20 ***Comment***: chewed x 12 years, 2 tins per day ***Statement***: She smokes about a pack a day for about 12 years
Type	Tobacco use statement and tobacco use comment inconsistent with Type field	***Type***: Cigarettes ***Comment***: Occasional cigar ***Statement***: Significant for tobacco use
Start or quit date	Tobacco use statement and tobacco use comment inconsistent with Start or Quit date field	***Quit date***: M2-D2-Y2 00:00:00 ***Comment***: quit 6 1/2 yrs ago ***Statement***: The patient is a former smoker who quit on M3/D3/Y3

To compare tobacco use data from different sources, tobacco use elements extracted from clinical notes and comments were mapped to structured fields. The "Amount" element extracted from tobacco use statements includes broader information than "pack/year" field. For example, the value for amount in the tobacco use statement "Significant for tobacco use" is "Significant". Also, "Temporal" element can hold more information than "start/quit date" and "years", such as the statement "The patient is a life long nonsmoker" does not provide start or quit date.

In addition, tobacco use elements extracted in free-text comments or statements can be expressed in numerous ways (e.g., "10/02/79", "18 Aug 2008", "x 10 yrs", "1 ppd" and "half a pack"). In this study, we developed mapping strategies to normalize numbers and map different expressions of "Amount", "Packs/year", "Start/Quit date", "Years" and "Type" in free-text tobacco use comments and statements into the structured tobacco use fields.

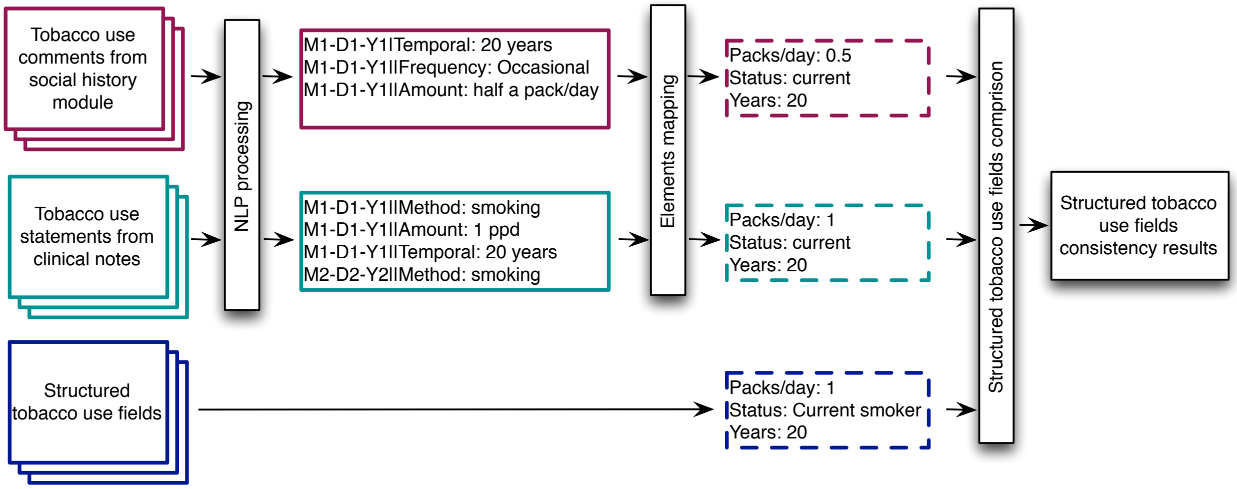

Tobacco use comments from social history module
Tobacco use statements from clinical notes
Structured tobacco use fields
NLP processing
M1-D1-Y1|Temporal: 20 years
M1-D1-Y1||Frequency: Occasional
M1-D1-Y1||Amount: half a pack/day
M1-D1-Y1||Method: smoking
M1-D1-Y1||Amount: 1 ppd
M1-D1-Y1||Temporal: 20 years
M2-D2-Y2||Method: smoking
Elements mapping
Packs/day: 0.5
Status: current
Years: 20
Packs/day: 1
Status: current
Years: 20
Packs/day: 1
Status: Current smoker
Years: 20
Structured tobacco use fields comprison
Structured tobacco use fields consistency results

Figure 3. Tobacco use information extraction and mapping for consistency analysis.

Results

For the 384 patients, 377 (98.2%) had tobacco use entries with the smoking status specified as a value other than "Not Assessed" (e.g., "Former smoker" and "Never smoker"); 201 (52.3%) had clinical notes with tobacco use statements, and 68 (17.7%) of patients have tobacco use comments. Figure 4 shows the overlap of the presence of notes, comments and tobacco use entries at the patient level. As shown in Table 3, only 17.7% of the patients had tobacco use comments data; fewer patients had both clinical notes with tobacco use information and tobacco use comments; and a large portion of patients had clinical notes with tobacco use information. Figure 5 shows the distribution of elements within each of the three sources.

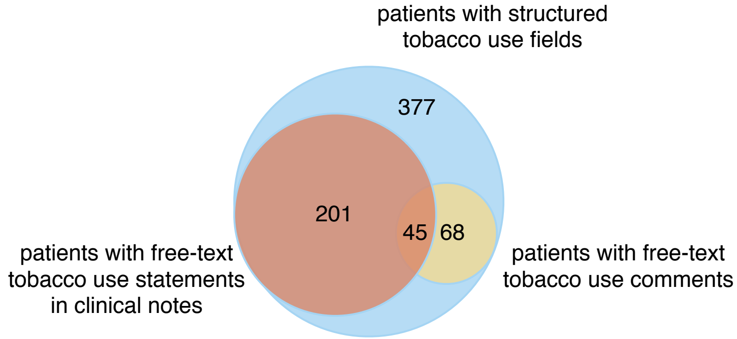


Figure 4. Overlap of tobacco use documentation at the patient level.

The numbers of elements extracted from each source are listed in Table 3. We observed that a large amount of duplicate tobacco use information exists within the structured tobacco use entries. Of 5,754 entries, only 944 (16.4%) of them were unique data with different dates since information in the social history module for patients appears often to remain unchanged across encounters and therefore the same information is propagated between encounters. For clinical notes, 590 (65.8%) out of 896 tobacco use statements were unique, and only 153 (7.9%) of tobacco use comments were unique (i.e., again due to propagation of the entry between encounters).

Table 3. Tobacco use information in difference sources.

Total patients	384			Total elements
Patients with notes	201 (52.3%)	Total statements	896	1,989
Patients with comments	68 (17.7%)	Total comments	1,937	3,233
Patients with structured tobacco use fields	377 (98.2%)	Total entries	5,754	11,098

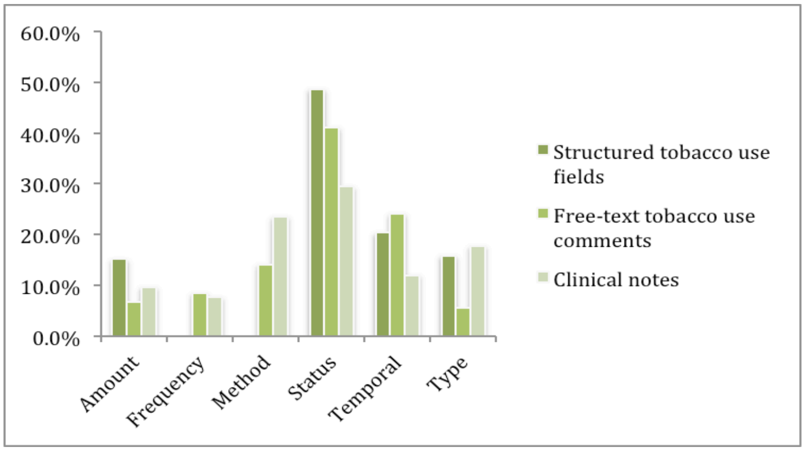


Figure 5. Distribution of tobacco use elements from each source.

Figure 6 shows the number of tobacco use elements and the date entered or documented for all patients over a 14-year period (2002 to 2016). Green data points signify free-text tobacco use comments and show a similar pattern as data points from structured tobacco use fields. This is likely related to free-text tobacco use comments being entered within the same user interface as the structured data fields. Compared with the other two sources, tobacco use data points from clinical notes did not show as much of an increase with time.

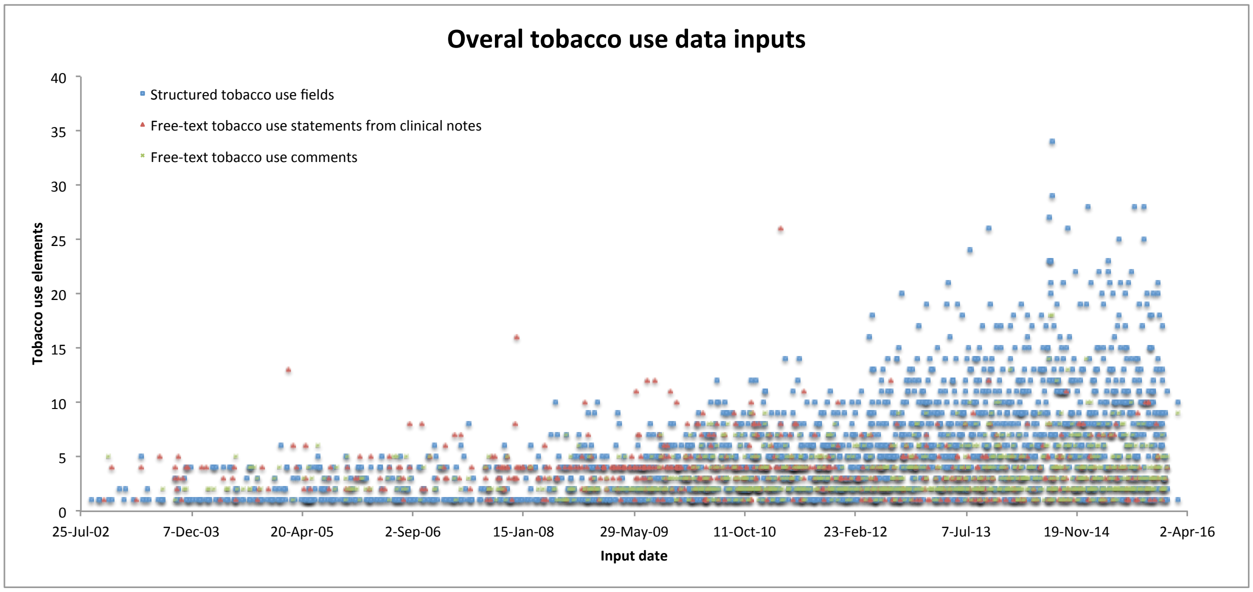


Figure 6. Tobacco use elements input patterns of all patients.

Figure 7a shows longitudinal distinct tobacco use information each year in an example patient start 2004 -14 with and without tobacco use elements extracted from free-text tobacco use comments and clinical notes. Each data point represents the number of distinct tobacco use elements documented within the year. The number of unique tobacco use elements for this patient increases with inclusion of information from the unstructured data sources. New tobacco use information can be plotted for the same patient with time (Figure 7b) resulting in a similar pattern. Figure 7c summarizes new tobacco use information aggregated for the patient cohort. As shown in the three figures, tobacco use information extracted from the structured tobacco use fields combined with information extracted from tobacco use comments and clinical notes provide more complete information than data from a single data source.

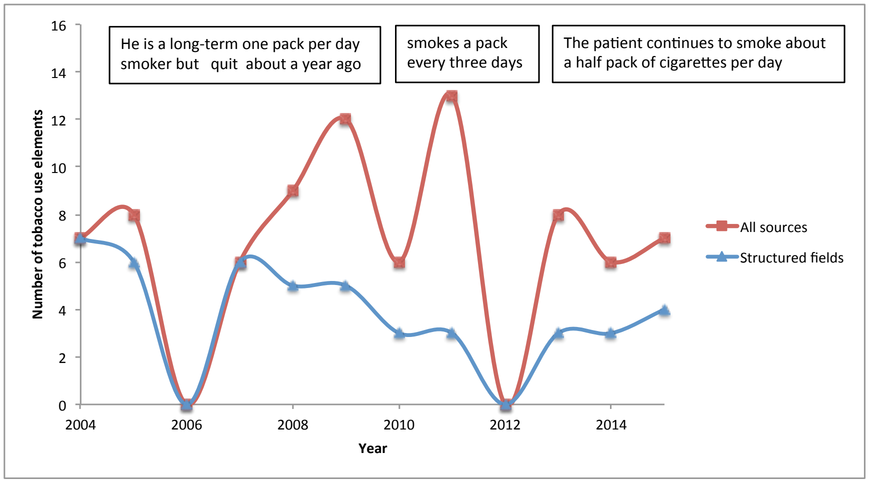


Figure 7a. Distinct tobacco use elements of an example patient.

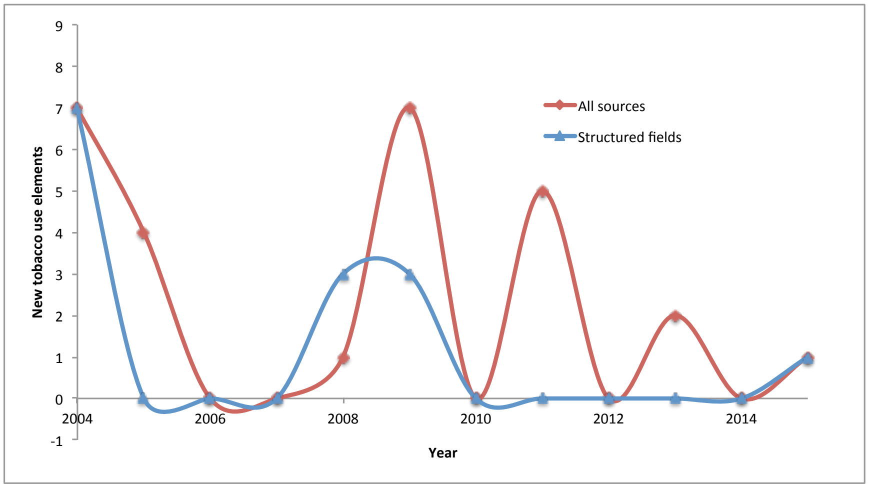


Figure 7b. Tobacco use new information patterns of the same example patient.

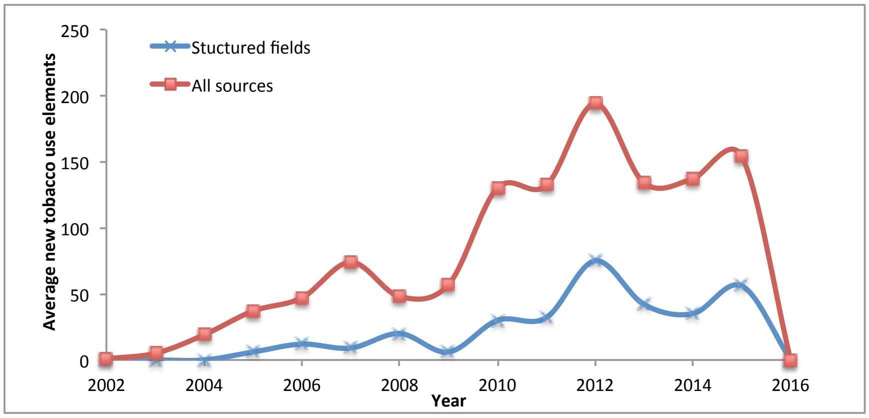


Figure 7c. New tobacco use elements patterns of all patients.

With respect to the quality and consistency of tobacco use information from the three sources, Figure 8 shows the consistency rate of five fields between tobacco use elements collected from the structured fields, comments, and clinical notes. Compared with other data fields, there was a large amount of inconsistency with "start/quit date" between EHR module entries, comments, and notes. Other data fields demonstrated much better consistency, which implicate the overall reliability of tobacco use information.

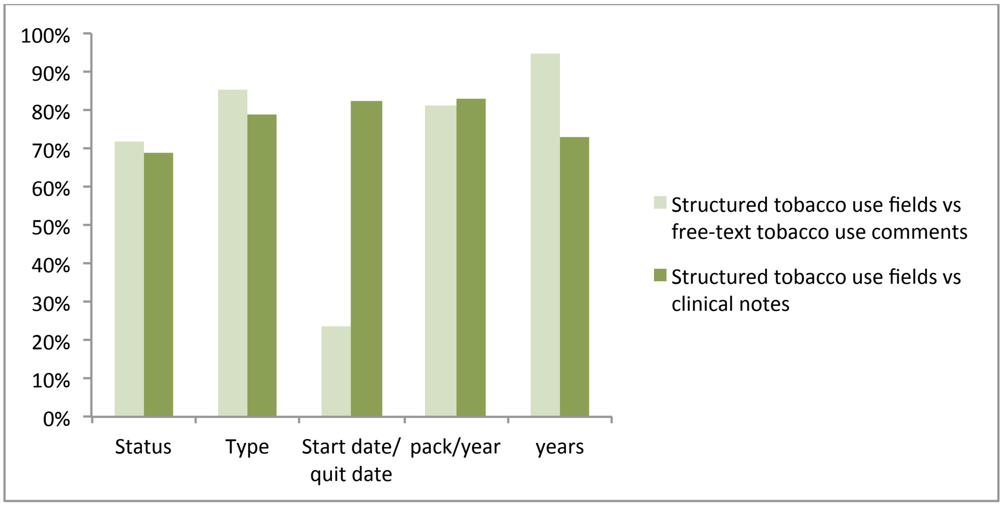


Figure 8. Data consistency between structured data and unstructured data.

Some inconsistencies were also found within structured tobacco use fields. Below examples show part of the pipe-delimited entries of the structured tobacco use fields limited to status information, in a format of "*date |tobacco_use*", of a patient. When examining three structured tobacco use fields of same patient for different encounters, the "Smoke status" of the last entry is inconsistent with the first two entries.

D1-M1-Y1 00:00:00 |Former Smoker
D2-M1-Y1 00:00:00 |Former Smoker
D3-M2-Y1 00:00:00 |Never Smoker

In another example, 52 of 377 patients are recorded with the same "Years" of tobacco used (30.0) along different encounters as shown in below example tobacco use fields, in a format of "*date|tobacco_use|quit_date|chew|cigarettes|cigars|pipe|snuff|packs_year|years|smokeless_tobacco*". There are also a few cases of inconsistent values for "Amount" in structured tobacco use fields. The example as follows also shows the large amount of duplicated tobacco use within the structured tobacco use fields.

D1-M1-Y1 00:00:00|Former Smoker|D1-M1-Y1 00:00:00|N|N|N|N|N|0|30.0|Never Used
D2-M2-Y2 00:00:00|Former Smoker D1-M1-Y1 00:00:00|N|Y|N|N|N|0|30.0|Never Used
D3-M3-Y3 00:00:00|Former Smoker| D1-M1-Y1 00:00:00|N|Y|N|N|N|0|30.0|Never Used
D4-M4-Y3 00:00:00|Current Every Day Smoker| D1-M1-Y1 00:00:00|N|Y|N|N|N|.5|30.0|Never Used
D5-M5-Y3 00:00:00|Current Every Day Smoker| D1-M1-Y1 00:00:00|N|Y|N|N|N|.5|30.0|Never Used
D6-M6-Y3 00:00:00|Current Every Day Smoker| D1-M1-Y1 00:00:00|N|Y|N|N|N|.5|30.0|Never Used
D7-M7-Y4 00:00:00|Current Every Day Smoker| D1-M1-Y1 00:00:00|N|Y|N|N|N|.5|30.0|Never Used

Discussion

The results in this study provide good insight into the characteristics of tobacco use content and quality issues from structured tobacco use fields in social history module, tobacco use comments in the same module and tobacco use statement in clinical notes. This work is an initial step to extract and integrate large amount of substance use information stored within EHR systems using automatic tools based NLP techniques. The approach used in this study can be easily generalized for other substance use types such as alcohol use and drug use and potentially more broadly to data represented in both text and structure fields over time. The results illustrate some of the fundamental issues of integrating substance use information from different sources. The data used in this study is limited to a single health care system of one institution with a medium sized cohort. The results will, therefore need to be validated in other settings. Also, the automated methods will likely need modifications before being applied to other EHR systems.

We observed data consistency issues between tobacco use information from different sources, which particularly presents challenges in integrating tobacco use information from different sources and using the combined date for decision support, research, public health, and other primary and secondary uses. While we expect for data to potentially change with time as smoking status changes (e.g., a smoker who subsequently quit), we encountered discrepancies on the same date and changes in status over time that were nonsensical. For example, the smoke status of a patient as follows changed between structured tobacco use records within structured tobacco use fields and free-text tobacco use statements. The structured tobacco use fields also failed to capture this status change. The change was documented into a clinical notes created at some point in between. More efforts are needed to determine adequate strategies for addressing and integrating inconsistent data from different sources, which may also need to be based off of use cases for how the data will subsequently be used.

D1-M1-Y1 00:00:00|Yes|Current Every Day Smoker||N|Y|N|N|N|.5|25.0|Unknown|

D2-M2-Y2 00:00:00|Yes|Current Every Day Smoker||N|Y|N|N|N|0.3|25.0|Unknown|

D3-M3-Y2|Sleeps poorly since May, after quitting smoking temporarily

Besides inconsistency between tobacco use information from different sources, we also noticed inconsistencies within the structure tobacco used fields such as unchanged "Years" of tobacco use or slightly different amount for "Packs/day" (e.g., 0.5 vs. 0.3). These issues indicate that the tobacco use information extracted from structured need to be further processed and normalized before integrating with tobacco use information from clinical notes and comments. Also, these issues may indicate the need for better EHR user training or improved user interface to help user entering valid data values.

Issues were also observed with comparing certain data such as "0.5 pack" vs. "10 cigarettes" or slightly different amounts "0.75 pack" vs. "0.5 pack", "6 years ago" vs. "6.5 years ago". Moreover, in tobacco use statement from notes and comments, tobacco use elements can be documented using various expressions with different levels of detail (e.g. "10/02/90" vs. "Feb 90" and "half pack per day" vs. "0.5 pack/day"). These issues highlight the requirement for good mapping (e.g., "former smoker" to "quit") and normalization (e.g., "half" to "0.5") strategies (e.g., "0.5 pack" is same as "0.75 pack") for NLP components.

As shown in the Figure 6, more tobacco use data were entered into the social history module in recent years. We speculate that the increased entry of data in the social history module is due to a combination of factors, likely including increased attention to entry of social history information and the requirement of asking this by certain regulations (i.e., Meaningful Use requirements).

This study is limited in one substance use type, and a next step will therefore include more substance use types such as alcohol use, drug use, or caffeine. We also noticed large amount of free-text comments and statements with smoke exposure information and use of e-cigarettes, for which we did not address in this study and which represents an area of further development. Also, except for the structured tobacco use fields in the social history module, other potential EHR fields that may also include tobacco use information were not included in our analysis.

Conclusion

Overall, we applied an NLP pipeline with components to detect tobacco use information from tobacco use statements in clinical notes and tobacco use comments from social history module of Epic EHR, as well as retrieved structured information from the tobacco use EHR module. The extracted tobacco use information was analyzed over a cohort of patients, characterizing tobacco use EHR content and quality issues. Our results provide insights into the challenges with reconciling and integrating this data for secondary uses. The results indicate the structured tobacco use data alone like do not provide complete tobacco use information. Further work is needed to improve approaches for integration of tobacco use information from different parts of the EHR.

Acknowledgements

The National Institutes of Health through the National Library of Medicine (R01LM011364 and R01GM102282), Clinical and Translational Science Award (8UL1TR000114-02) supported this work.

References

1. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. *Jama*. 2004;291(10):1238-45.

2. Jane-Llopis E, Matytsina I. Mental health and alcohol, drugs and tobacco: a review of the comorbidity between mental disorders and the use of alcohol, tobacco and illicit drugs. *Drug Alcohol Rev*. 2006;25(6):515-36.
3. Huang FY, Ziedonis DM, Hu HM, Kline A. Using information technology to evaluate the detection of co-occurring substance use disorders amongst patients in a state mental health system: implications for co-occurring disorder state initiatives. *Community Ment Health J*. 2008;44(1):11-27.
4. Babor TF, Sciamanna CN, Pronk NP. Assessing multiple risk behaviors in primary care. Screening issues and related concepts. *Am J Prev Med*. 2004;27(2 Suppl):42-53.
5. Centers for Disease Control and Prevention. Smoking & Tobacco Use Fast Facts. [updated April 24, 2014]; Available from: http://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/.
6. In: Hernandez LM, Blazer DG, editors. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. Washington (DC)2006.
7. Institute of Medicine . Capturing social and behavioral domains and measures in electronic health records: Phase 2. *Washington, DC: The National Academies Press*. 2014.
8. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*. 2008;15(1):14-24.
9. Melton GB, Manaktala S, Sarkar IN, Chen ES. Social and behavioral history information in public health datasets. *AMIA Annu Symp Proc*. 2012;2012:625-34.
10. Chen ES, Manaktala S, Sarkar IN, Melton GB. A multi-site content analysis of social history information in clinical notes. *AMIA Annu Symp Proc*. 2011;2011:227-36.
11. Chen E, Garcia-Webb M. An analysis of free-text alcohol use documentation in the electronic health record: early findings and implications. *Appl Clin Inform*. 2014;5(2):402-15.
12. 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. *AMIA Annu Symp Proc*. 2014;2014:366-74.
13. Wu C-Y, Chang C-K, Robson D, Jackson R, Chen S-J, Hayes RD, et al. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PLoS One*. 2013;8(9):e74262.
14. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. *J Am Med Inform Assoc*. 2005;12(5):517-29.
15. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, et al. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. *Am J Prev Med*. 2005;29(5):434-9.
16. Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, et al. A study of transportability of an existing smoking status detection module across institutions. *AMIA Annu Symp Proc*. 2012;2012:577-86.
17. McCormick PJ, Elhadad N, Stetson PD. Use of semantic features to classify patient smoking status. *AMIA Annu Symp Proc*. 2008:450-4.
18. Wicentowski R, Sydes MR. Using implicit information to identify smoking status in smoke-blind medical discharge summaries. *J Am Med Inform Assoc*. 2008;15(1):29-31.
19. MTSamples. [March 1, 2015]; Available from: http://www.mtsamples.com/.
20. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. *J Am Med Inform Assoc*. 2006 Jan-Feb;13(1):30-9.
21.OpenEHR Specifications. [March 1, 2015]; Available from: http://www.openehr.org/svn/specification/TAGS/Release-1.0.2/publishing/architecture/rm/ehr_im.pdf.
22. Wang Y, Chen ES, Pakhomov S, Arsoniadis E, Carter EW, Lindemann E, et al. Automated Extraction of Substance Use Information from Clinical Texts. *AMIA Annu Symp Proc 2015*. 2015 (Accepted).
23. Palmer M, Gildea D, Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput Linguist*. 2005;31(1):71-106.
24. Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ Clinical Question Answering System. *AMIA Annu Symp Proc*. 2011;2011:171-80.
25. Epic Systems Corporation [March 1, 2016]; Available from: http://www.epic.com/.
26. Stanford NLP tools. [March 1, 2015]; Available from: http://nlp.stanford.edu/software/index.shtml.