Comparison of Three Information Sources for Smoking Information in Electronic Health Records

Liwei Wang, Xiaoyang Ruan, Ping Yang and Hongfang Liu

Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.

ABSTRACT

OBJECTIVE: The primary aim was to compare independent and joint performance of retrieving smoking status through different sources, including narrative text processed by natural language processing (NLP), patient-provided information (PPI), and diagnosis codes (ie, International Classification of Diseases, Ninth Revision [ICD-9]). We also compared the performance of retrieving smoking strength information (ie, heavy/light smoker) from narrative text and PPI.

MATERIALS AND METHODS: Our study leveraged an existing lung cancer cohort for smoking status, amount, and strength information, which was manually chart-reviewed. On the NLP side, smoking-related electronic medical record (EMR) data were retrieved first. A pattern-based smoking information extraction module was then implemented to extract smoking-related information. After that, heuristic rules were used to obtain smoking status-related information. Smoking information was also obtained from structured data sources based on diagnosis codes and PPI. Sensitivity, specificity, and accuracy were measured using patients with coverage (ie, the proportion of patients whose smoking status/strength can be effectively determined).

RESULTS: NLP alone has the best overall performance for smoking status extraction (patient coverage: 0.88; sensitivity: 0.97; specificity: 0.70; accuracy: 0.88); combining PPI with NLP further improved patient coverage to 0.96. ICD-9 does not provide additional improvement to NLP and its combination with PPI. For smoking strength, combining NLP with PPI has slight improvement over NLP alone.

CONCLUSION: These findings suggest that narrative text could serve as a more reliable and comprehensive source for obtaining smoking-related information than structured data sources. PPI, the readily available structured data, could be used as a complementary source for more comprehensive patient coverage.

KEYWORDS: smoking status, smoking strength, natural language processing, ICD-9, patient-provided information

CITATION: Wang et al. Comparison of Three Information Sources for Smoking Information in Electronic Health Records. *Cancer Informatics* 2016:15 237–242 doi: 10.4137/CIN.S40604.

TYPE: Original Research

RECEIVED: August 01, 2016. **RESUBMITTED:** October 09, 2016. **ACCEPTED FOR PUBLICATION:** October 20, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1292 words, excluding any confidential comments to the academic editor.

FUNDING: The authors gratefully acknowledge the support from the National Institute of Health (NIH) grants R01GM102282-03 and R01 LM011934-02. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: liu.hongfang@mayo.edu; wang.liwei@mayo.edu

Introduction

Smoking is the leading cause of preventable death in the United States. More than 480,000 deaths per year are contributed to smoking, costing nearly $100 billion in direct medical care.[1] Only 30% of smokers have received evidence-based tobacco dependence treatment during a health care visit.[2] To motivate clinicians and health care systems to identify patients who use tobacco and to provide them with evidence-based treatment, recording smoking status for people aged 13 years or older is one of the core criteria for meaningful use of electronic medical records (EMRs).[2] Presently, smoking status is generally recorded as a structured data field with the following values: current every day smoker; current some day smoker; former smoker; never smoker; smoker, current status unknown; and unknown if ever smoked.

With the increasing use of longitudinal EMRs for clinical and translational research, extracting smoking information from EMRs is crucial since smoking is a risk factor for many conditions. Prior to meaningful use, smoking information available in EMRs is frequently recorded in narrative text. Natural language processing (NLP) can be used to extract smoking information,[3–5] especially after the organization of the 2006 i2b2 NLP Shared Task.[6] This task aims to evaluate the following two challenging questions: 1) "What is the state-of-the-art in automatic de-identification of clinical data?" and 2) "How accurately can automatic methods evaluate the smoking status of patients based on their medical records?". Both statistical- and pattern-based NLP techniques have been explored. Some systems adopt a hybrid approach. For example, the smoking module of cTAKES used smoking-related keywords to extract sentences and then used support vector machines to classify sentences as current smoker, former smoker, smoker, or never smoker. The patient's current smoking information is then summarized from all existing clinical documents using heuristics. The portability of the smoking module in cTAKES was examined on the Vanderbilt University Hospital's EMR data where a different summarization rule yields better performance.[7] Structurally, smoking information can be recorded as the International Classification of Diseases, Ninth Revision (ICD-9), tobacco use codes or patient-provided information (PPI). The evaluation of ICD-9 tobacco use codes indicates high precision but a low patient coverage rate in identifying ever smoker.[8] Some studies show that self-reported smoking status can be inconsistent.[9,10] Overall, the availability of structured smoking status is limited and the extraction of smoking information using NLP generally showed overall performance around 90% but is less accessible due to technical complexity.

To date, there is little investigation on the comparison of smoking status information from three distinct sources: NLP, PPI, and ICD-9. Additionally, smoking amount or strength information is also crucial, especially for diseases that are directly caused by smoking such as lung diseases. Few of the existing NLP systems extract such information. In this study, we evaluated the performance of three sources of information (NLP, PPI, and ICD-9), independently and jointly, in identifying smoking status and two sources of information (NLP and PPI) in identifying smoking strength.

Materials and Methods

Our study leveraged an existing lung cancer cohort for smoking status, amount, and strength information, which was manually chart-reviewed as a background comparison. On the NLP side, smoking-related EMR data were retrieved at Mayo Clinic. A pattern-based smoking information extraction module was implemented to extract smoking-related information. We then utilized heuristic rules to obtain smoking status-related information. On the other hand, ICD-9- and PPI-based smoking information was obtained from structured data sources. Figure 1 shows the study design. In the following section, we describe the data set used, the algorithms developed, and performance assessment.

Manually reviewed smoking status. In total, 561 patients aged between 15 and 45 years who were diagnosed with one of the 21 categories of lung cancer subtypes from 1997 to 2011 were manually reviewed to obtain smoking status and strength information.[11] Mayo Clinic reached meaningful use stage 1 in 2011. Before stage 1, documentations of smoking status were meaningful in some cases. Information including smoker age, year start, year stop, pack per day, and pack-years was manually recorded. Date of diagnosis was incorporated to classify each patient in one of the five smoking status categories, including current (every/some) day smoker, former smoker, never smoker, smoker current status unknown, and unknown if ever smoked. Former smoker refers to an individual who has smoked at least 100 cigarettes during his/her lifetime but does not currently smoke. Never smoker is defined as an individual who has not smoked 100 or more cigarettes during his/her lifetime. Smoking strength contains two categories, heavy smoker and light smoker, where heavy smoker smokes 10 cigarettes or more per day.[12]

NLP-based identification of smoking status. Clinical notes in text format for 561 patients were obtained from the Mayo Clinic EMR. We extracted smoking-related information using MedTagger[13], integrated with heuristic rules and MedTime.[14] MedTime can parse temporal expression in text to the output of entity types such as "date", "time", and "duration". The following information was extracted:

- Smoking status – information relevant to smoking status (eg, current smoker, former smoker, never smoker)
- Amount information – the amount of smoking (eg, number of cigarettes per day)

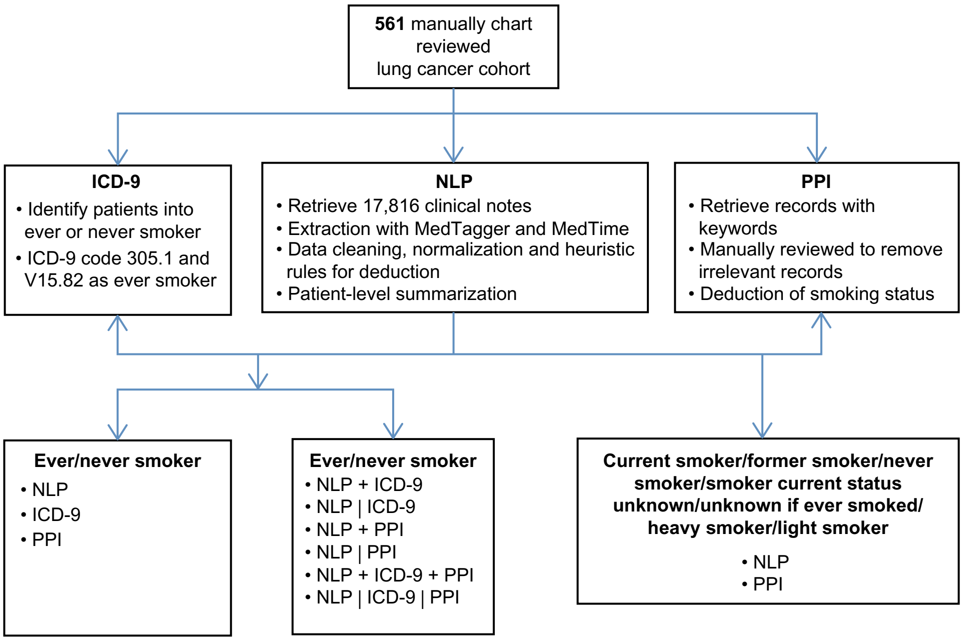
561 manually chart reviewed lung cancer cohort
ICD-9
• Identify patients into ever or never smoker
• ICD-9 code 305.1 and V15.82 as ever smoker
NLP
• Retrieve 17,816 clinical notes
• Extraction with MedTagger and MedTime
• Data cleaning, normalization and heuristic rules for deduction
• Patient-level summarization
PPI
• Retrieve records with keywords
• Manually reviewed to remove irrelevant records
• Deduction of smoking status
Ever/never smoker
• NLP
• ICD-9
• PPI
Ever/never smoker
• NLP + ICD-9
• NLP | ICD-9
• NLP + PPI
• NLP | PPI
• NLP + ICD-9 + PPI
• NLP | ICD-9 | PPI
Current smoker/former smoker/never smoker/smoker current status unknown/unknown if ever smoked/ heavy smoker/light smoker
• NLP
• PPI

Figure 1. Workflow of NLP-based identification of smoking-related information.

- Quit tense – the keyword quit and the corresponding tense (eg, will quit, has quit, should quit)
- Temporal expression or age – temporal information relevant to date, frequency (eg, repeat two times a day), or duration (eg, smoked for 10 years)

We then used the above smoking-related information to detect smoking status within 30 days after cancer diagnosis.

The parsed information was then standardized by formatting date and smoking status into consistent formats and converting frequency, duration, and smoking amount to consistent units of measure. For example, for cases where date contains only year and month, the 15th day of the month was assumed. For smoking status, the co-occurrence of quit with current or past tense indication words "have", "had", "has", and "is" was considered quitting smoking (ie, former smoker). Frequency and duration were standardized to hours. Smoking frequency was standardized to pack per day for identification of smoking strength.

The standardized information was stored separately for each patient. All information available for each patient was processed according to the order of clinical note record date. We only considered the paragraphs with section codes that belonged to "social and behavior history" category or those paragraphs having the presence of pack per day, pack-years, or smoking status (eg, former smoker or current smoker). Information including pack per day, pack-years, and smoking duration was utilized to deduce the corresponding missing values not mentioned/captured from the original clinical notes. For example, given a patient smoked 1.5 pack per day for 6 years, it can be calculated that the patient smoked 9 pack-years.

We traced each patient through the record date timeline to assign a smoking status for that record date. The general rule was that every patient starts with unknown smoking status on the first record date and switches status to nonsmoker, former smoker, or current smoker, based on the smoking-related information throughout the timeline. Smoking status at the time of diagnosis was determined by using the smoking status that was recorded no later than 30 days after the diagnosis date.

For each patient, smoking amount-related information was used to determine the smoking strength (ie, heavy smoker [≥10 cigarettes per day]; light smoker [<10 cigarettes per day]) for each record date. The strength results from all record dates were summarized to determine the final smoking strength of a patient (the smoking strength with the highest occurrence was used, or unknown if the occurrences are equal).

Patient-provided smoking information. Patient-provided smoking information was obtained from structured PPI in EMRs, where 49% of patients complete the structured data. To find smoking-related information, we limited the search criteria to retrieve only smoking-related PPI entries. The results were manually reviewed to keep entries that can be reasonably used to deduce smoking status or strength and remove irrelevant records such as "Are you interested in more information about safety (seat belts, smoke detectors, firearms)?". To deduce smoking status, the records were filtered to keep those recorded no later than 30 days after the recorded diagnosis date. Records from all available dates were used to determine the smoking strength.

Diagnosis code-based smoking status. The diagnosis code extracted from hospital billing information was used to group patients into two categories of ever smoker or never smoker. Specifically, patients with one or more occurrences of the ICD-9 codes, tobacco use disorder (305.1) and history of tobacco use (V15.82), were considered ever smoker.

Standard smoking status type and comparison between automatic and manual results. The comparison between automatically generated and manually reviewed results was conducted at two levels. The first level categorizes patients into ever smoker or never smoker. At this level, we evaluated the performance of automatically generated results using NLP, ICD, PPI alone, and combinations including NLP + ICD, NLP | ICD, NLP + PPI, NLP | PPI, NLP + PPI + ICD, and NLP | PPI | ICD. For combinations, the "+" rule considers a patient as ever smoker when any single strategy identifies the patient as ever smoker. The "|" rule uses NLP as the primary source of information and other source(s) as a supplement when NLP failed to identify the status. At the second level, the NLP-, PPI-, and NLP | PPI-based results and manually reviewed smoking status were compared by converting both parts to standard smoking status type, which assigns each patient to one of the five smoking status categories and one of the two smoking strength categories. Comparison results were presented as sensitivity, specificity, and accuracy of measurements.

For NLP and PPI, patient coverage is defined as the proportion of patients whose smoking status/strength can be effectively determined. For diagnosis codes, the patient coverage is 100% since all patients have ICD-9 diagnosis codes available. Patients with smoking-related ICD-9 codes are categorized as ever smoker, otherwise never smoker.

For each method, sensitivity, specificity, and accuracy are measured using patients with coverage (ie, excluding patients without original documentation or whose smoking status and/or strength cannot be effectively determined from existing documentation). The sensitivity is defined as $\frac{|\text{True positives}|}{|\text{True positives}|+|\text{False negatives}|}$, specificity is defined as $\frac{|\text{True negatives}|}{|\text{True negatives}|+|\text{False positives}|}$, and accuracy is defined as $\frac{|\text{True positives}+\text{True negatives}|}{|\text{True positives}|+|\text{False positives}|+|\text{True negatives}|+|\text{False negatives}|}$.

Results

The number of patients grouped by cancer subtypes and manually reviewed smoking status is shown in Table 1. We manually identified 206 (37%) as never smokers and 355 (63%) as ever smokers, of which 26% were current every/some day smokers, 15% were former smokers, and 22% were smokers

Table 1. Smoking status and strength by manual review.

	DETAILED SMOKING STATUS	COUNT	PERCENTAGE
Detailed Smoking Status	Current smoker[a]	143	26
	Former smoker	86	15
	Smoker current status unknown	126	22
	Never smoker	206	37
	Unknown if ever smoked	0	0
Smoking Strength	Heavy tobacco smoker[b]	293	52
	Light tobacco smoker	36	6

Notes: [a]Includes current every day or some day smoker. [b]Smoke more than or equal to 10 cigarettes per day.

with current status unknown. For smoking strength, we manually identified 52% as heavy smokers and 6% as light smokers, which accounts for 91% of patients with smoking history (Table 1).

At the first level, patients were categorized into ever/never smoker (Table 2). Patient coverage rates for NLP, ICD-9, and PPI were 88%, 100%, and 49%, respectively. NLP alone showed the best sensitivity (0.97) and accuracy (0.88) compared to ICD-9 and PPI. ICD-9 alone had the best specificity (0.98) but the worst sensitivity (0.25). The performance of PPI alone was between ICD-9 and NLP. For combinations, the "+" rule generally had poorer performance than the "|" rule that used NLP as the primary source of information. NLP | PPI and NLP | PPI | ICD-9 had the best performance by improving the patient coverage over NLP alone (88% for NLP to 96% and 100% for NLP | PPI and NLP | PPI | ICD-9, respectively). ICD-9 generally did not provide additional improvement to NLP and NLP | PPI. Using PPI or ICD as the primary source supplemented with NLP generally showed poor performance (data not shown).

For 329 patients with smoking strength identified by manual review, NLP alone was able to identify 184 (56%) patients with smoking strength-related information and achieved sensitivity of 0.74, specificity of 0.88, and accuracy of

Table 2. Performance of NLP, ICD-9, and PPI in identifying smoking status and strength.

		MANUAL		SENSITIVITY (95% CI)	SPECIFICITY (95% CI)	ACCURACY (95% CI)
Smoking status (561 total)		**Ever**	**Never**			
NLP	Ever	311	53	0.97(0.95–0.99)	0.70(0.63–0.77)	0.88(0.84–0.90)
	Never	8	124			
ICD-9	Ever	89	4	0.25(0.2–0.3)	0.98(0.95–1)	0.52(0.48–0.56)
	Never	266	202			
PPI	Ever	223	51	0.73(0.68–0.78)	0.72(0.64–0.78)	0.73(0.68–0.77)
	Never	82	129			
NLP + ICD-9[a]	Ever	315	53	0.89(0.85–0.92)	0.74(0.68–0.80)	0.83(0.80–0.86)
	Never	40	153			
NLP | ICD-9[b]	Ever	313	53	0.88(0.84–0.91)	0.74(0.68–0.80)	0.83(0.80–0.86)
	Never	42	153			
NLP + PPI	Ever	333	82	0.97(0.94–0.98)	0.58(0.51–0.65)	0.83(0.79–0.86)
	Never	12	114			
NLP | PPI	Ever	329	57	0.95(0.93–0.97)	0.71(0.64–0.77)	0.87(0.83–0.89)
	Never	16	139			
NLP + PPI + ICD	Ever	334	82	0.94(0.91–0.96)	0.6(0.53–0.67)	0.82(0.78–0.85)
	Never	21	124			
NLP | PPI | ICD-9	Ever	329	57	0.93(0.89–0.95)	0.72(0.66–0.78)	0.85(0.82–0.88)
	Never	26	149			
Smoking strength (329 total)		**Heavy**	**Light**			
NLP	Heavy	123	2	0.74(0.66–0.80)	0.88(0.64–0.99)	0.75(0.68–0.81)
	Light	44	15			
PPI	Heavy	45	0	0.66(0.54–0.77)	1.0(0.59–1)	0.69(0.58–0.79)
	Light	23	7			
NLP | PPI	Heavy	136	2	0.73(0.66–0.80)	0.9(0.68–0.99)	0.74(0.68–0.80)
	Light	52	18			

Notes: [a]Categorized as ever smoker when either NLP or ICD identifies as ever smoker. Same rule applies to other combinations. [b]Used NLP as primary source, and if not available, used ICD. Same rule applies to other combinations.

0.75. While PPI alone had poor patient coverage rate (23%), it was improved when combined with NLP. NLP | PPI together had 63% patient coverage rate, with sensitivity of 0.73, specificity of 0.9, and accuracy of 0.74 (Table 2).

At the second level of comparing standard smoking status types, NLP identified 263 (51%) patients who have exactly the same detailed smoking status as in manual review (Table 3). Among them, 135 patients are current or former smokers, 4 are smokers with unknown current status, and 124 are never smokers. For the 248 (49%) patients without exactly the same smoking status, the majority (n = 172) was caused by the misidentification among current smoker, former smoker, and smoker with unknown current status. There were 8 cases where smokers were misidentified as never smokers and 53 cases where never smokers were misidentified as smokers. The performance of NLP (51% exact match) was generally better than PPI. Supplementing PPI to NLP slightly improved the percentage of exact match (from 51% to 53%).

Manual review of 10 patients who were ever smokers in the gold standard but misidentified as never smokers by NLP showed that five patients had no smoking-related labels in the original clinical notes before diagnosis or within 30 days after diagnosis. Four patients had at least one never or nonsmoker label possibly due to incorrect documentation. One patient had a former smoker label appeared 6 years after the diagnosis date.

On the other hand, manual review of 10 never smokers misidentified as ever smokers by NLP indicated that two patients had never smoker tags coexisting with smoking amount tags. Four had coexistence of never smoker and current/former smoker tags. Three of them had coexistence of never/nonsmoker and quit smoking tags. One patient had both a former smoker tag and more than one smoking frequency- or duration-related tags.

A considerable number (n = 83) of those classified as smoker with unknown current status were identified as current smoker by NLP. This is primarily due to our algorithm that considers a patient as current smoker if ever appeared in history and has no quit tag. While in manual review, ever smokers without a smoking quit year listed were considered as without known current status. However, the manual review is very subjective and it is not clear how the abstractor distinguished the smoking information that appeared 5 years ago and the smoking information that appeared a month ago. In our NLP algorithm, we simplified the algorithm so that ever smoker without a quit tag is considered as current smoker.

Discussion

In this study, we compared independent and joint performance of extracting smoking information through different sources including narrative text (parsed by NLP), PPI, and ICD-9. Results showed that NLP and PPI could provide adequate information for smoking detection, while ICD-9 had almost no additional contribution. Our analysis indicates that using NLP as the primary source of information, supplemented with other sources, has better performance, suggesting narrative text as a more reliable and comprehensive source for obtaining smoking-related information. While adding PPI information did not improve the sensitivity and specificity, it remarkably

Table 3. Performance of NLP and PPI in identifying accurate smoking status.

		MANUAL				UNKNOWN	EXACT MATCH
		CURRENT SMOKER	FORMER SMOKER	SMOKER CURRENT STATUS UNKNOWN	NEVER		
NLP	Current smoker	**85**	15	83	11	0	0.51
	Former smoker	36	**50**	22	39	0	
	Smoker current status unknown	7	9	**4**	3	0	
	Never	2	5	1	**124**	0	
	Unknown	3	2	3	7	**0**	
PPI	Current smoker	**42**	17	49	40	0	0.42
	Former smoker	48	**25**	28	7	0	
	Smoker current status unknown	2	4	**8**	4	0	
	Never	31	33	18	**129**	0	
	Unknown	0	0	0	0	**0**	
NLP | PPI	Current smoker	**88**	17	86	13	0	0.53
	Former smoker	41	**52**	23	39	0	
	Smoker current status unknown	7	9	**6**	5	0	
	Never	4	7	5	**139**	0	
	Unknown	0	0	0	0	**0**	

Note: Numbers in bold indicate accurate results.

improved the patient coverage rate over NLP alone. This suggested that PPI, the readily available structured information, could be used as an extra source for more comprehensive patient coverage.

Smoking status was determined by using clinical information in EMRs no later than 30 days after the diagnosis (lung cancer) date. The performance of NLP with respect to patient coverage is dependent on the availability of smoking information and the associated details. For example, the majority of the patients with no smoking information extracted by NLP had no smoking-related information in their original clinical notes.

Unlike previous studies that primarily focused on smoking status,[4,15,16] our study features smoking strength detection. Despite the incomplete original clinical notes, some missing values regarding smoking strength could be inferred. The final smoking strength was determined through longitudinal EMRs for each patient, which maximized the possibility of detecting the true smoking strength. NLP correctly identified current smokers more than PPI. In PPI, about half of the patients did not provide smoking-related information and some of the remaining patients provided incomplete or contradictive information. We investigated the current smokers by gold standard, for example, one patient answered "No" to the question of "Cigarettes current use", while in clinical notes, a status of current smoker was recorded. Therefore, PPI requires further verification.

The limitation of the study is that heuristic rules to obtain smoking status-related information in NLP were developed on the basis of existing research data from a lung cancer cohort, and the performance was only evaluated based on these data; no cross evaluation was conducted.

Conclusion

Our results indicate NLP alone has the best sensitivity, and using joint sources of information primarily improved patient coverage over NLP alone. NLP is a relatively comprehensive and reliable resource for extracting both smoking status and strength information. The NLP rules we used are very generic, not specifically trained and not cancer specific. Therefore, they can be applied on other data sets. Nevertheless, the overall performance can be improved further by appending information from PPI.

Acknowledgment

The authors would like to thank Mrs. Katelyn N Cordie for her editing assistance with the manuscript.

Author Contributions

All authors are justifiably credited with authorship, according to the authorship criteria. In detail, LW and XR – analysis of data, interpretation of results, and drafting of the manuscript; PY – interpretation of results, critical revision of manuscript; HL – conception, design, development, interpretation of results, and critical revision of manuscript. All authors reviewed and approved the final version.

REFERENCES

1. Smoking & Tobacco Use. *Centers for Disease Control and Prevention*. [cited 2016 April 25]. 2008;15(1):29–31. Available from: http://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/
2. Meaningful Use and Tobacco Cessation. University of Wisconsin. Center for Tobacco Research & Intervention. Available at: http://quitworksnh.org/wp-content/uploads/2016/05/UW_CTR_Meaningful-Use_2012.pdf. Accessed May 20, 2016.
3. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. *J Am Med Inform Assoc*. 2008;15(1):25–8.
4. Wicentowski R, Sydes MR. Using implicit information to identify smoking status in smoke-blind medical discharge summaries. *Journal of the American Medical Informatics Assoc*. 2008;15(1):29–31.
5. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak*. 2006;6:30.
6. Uzuner O, Szolovits P, Kohane I. i2b2 workshop on natural language processing challenges for clinical records. Paper presented at: Proceedings of the Fall Symposium of the American Medical Informatics Association; 2006. Washington, DC.
7. Liu M, Shah A, Jiang M, et al. A study of transportability of an existing smoking status detection module across institutions. Paper presented at: AMIA Annual Symposium Proceedings; 2012. Chicago.
8. Wiley LK, Shah A, Xu H, Bush WS. ICD-9 tobacco use codes are effective identifiers of smoking status. *J Am Med Inform Assoc*. 2013;20(4):652–8.
9. Stelmach R, Fernandes FL, Carvalho-Pinto RM, et al. Comparison between objective measures of smoking and self-reported smoking status in patients with asthma or COPD: are our patients telling us the truth? *J Bras Pneumol*. 2015;41(2):124–32.
10. Wong SL, Shields M, Leatherdale S, Malaison E, Hammond D. Assessment of validity of self-reported smoking status. *Health Rep*. 2012;23(1):47–53.
11. Deng B, Wang Y, Xie D, Stoddard SM, Yang P. Metformin use and young age lung cancer: a case series report. *Oncol Lett*. 2016;11(4):2899–902.
12. Test Procedure for §170.314(a)(11) Smoking status. Available at: https://www.healthit.gov/sites/default/files/standards-certification/2014-edition-draft-test-procedures/170-314-a-11-smoking-status-2014-test-procedure-draft-v1.0.pdf. Accessed May 20, 2016.
13. Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. *AMIA Jt Summits Transl Sci Proc*. 2013;2013:249.
14. Sohn S, Wagholikar KB, Li D, et al. Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. *J Am Med Inform Assoc*. 2013;20(5):836–42.
15. McCormick PJ, Elhadad N, Stetson PD. Use of semantic features to classify patient smoking status. AMIA Annual Symposium Proceedings; 2008: *American Medical Informatics Assoc.*; 2008. p. 450. Washington, DC.
16. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Assoc*. 2008;15(1):14–24.