MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Representation of Smoking-related Concepts in an Electronic Health Record

Mollie R. Poynton[1 2], Lewis Frey[2], Hongying Tang[2]

[1]*University of Utah College of Nursing, Informatics Program, United States of America*
[2]*University of Utah School of Medicine, Department of Biomedical Informatics, United States of America*

Abstract

Data-driven models, created using the process of knowledge discovery in databases (KDD), can be used to predict characteristics or outcomes of patients, and thus aid decision-making. These models possess the capacity to augment evidence-based decision making related to smoking cessation on the basis of individual patient characteristics. This study (1) assessed the feasibility of a KDD project related to smoking cessation, and (2) mapped strengths and weaknesses in representation of smoking cessation related concepts in a commonly used electronic health record (EHR), EpicCare EMR. A sample of patient records was assayed and mapped to smoking cessation guidelines. Data quality issues related to coded representation of tobacco use were evident. Both strengths and weaknesses in representation of smoking-related concepts were identified. Analysis of free text patient notes may be helpful to adequately describe smoking and smoking cessation in patient records for the purpose of KDD.

Keywords:

Tobacco Use Cessation, decision support systems, clinical, medical records systems, computerized, data analysis, statistical, knowledge discovery in databases

Introduction

Previous work by the principal investigator demonstrated the capability to build predictive models of smoking cessation status from health survey data using the KDD process [1]. However, this capability has not been demonstrated using data from an ambulatory clinical information system. In order to create or augment clinical knowledge models related to smoking cessation using the KDD process, concepts relevant to smoking cessation must be represented by the data. Data quality must be assessed and considered. Moreover, the labor required to pre-process the large amount of data will be great. So, before engaging in a full-scale study, involving large amounts of data, it is prudent to determine the feasibility of using KDD/ data mining to create predictive and descriptive models of smoking cessation using the intended data. The purpose of this study was to determine the feasibility of using data mining methods to create predictive and descriptive models of smoking cessation using data from a primary care clinical information system.

Methods

This study examined data generated by EpicCare EMR (Epic) in a group of community outpatient clinics [2]. A random sample of 3000 electronic records, each corresponding to a single patient coded as a current or former smoker, and seen at a clinic within the past year, was extracted from the enterprise data warehouse. The Epic data dictionary was used to identify tables and attributes within those tables possibly relevant [3]. Each Epic attribute/ variable was assessed for representation of key concepts related to primary care management of smoking cessation, including the assessment of tobacco use/ dependence, tobacco dependence intervention, and relapse prevention. Attributes were manually mapped to concepts extracted from the VA/DoD clinical practice guideline for the management of tobacco use, and the United States Public Health Service (U.S.P.H.S.) Guidelines on Tobacco Use and Dependence. [4, 5]

Results

A series of forty-one individual reports was created, each containing descriptive statistics for each attribute identified as potentially relevant to a KDD study, and/or decode guides. The random sample of 3000 patients coded as current or former smokers, with at least one visit within the past year, was composed of 56% current smokers and 44% former smokers. Table 1 describes the presence and values of attributes used to describe intensity and duration of tobacco use in this sample of Epic records. The number of packs per day for current smokers ranged from 1-31. The most frequently used words in a free text "Tobacco Comment' attribute within the social history table were as follows: age, ago, day, quit, since, smoke, year, yr.

A concept map was generated – a single table mapping individual Epic attributes with smoking-relevant concepts extracted from the clinical practice guidelines and literature review. Table 2 summarizes strengths and weaknesses in representation of smoking cessation related concepts. Prescribed treatments and follow up visits within the system were clearly described in a coded format. However, other information was not

represented: over-the-counter treatments, social support, referrals for external treatment programs, and advice/ counseling offered in the course of a patient encounter.

Table 1 - Attributes describing intensity and duration of tobacco use

	Current User Records (n=1693)	*Former User Records (n=1307)*
"Packs Per Day" data entered	1220	521
"Packs Per Day": Mean (S.D.)	2.54 (3.48)	3.15 (3.69)
"Years": data entered	1012	542
"Years": Mean (S.D.)	15.71 (13.28)	15.18 (12.86)

Table 2 - Strengths and weaknesses in representation of smoking cessation related concepts

Strengths	*Weaknesses*
Prescribed medications	Advice/counseling to quit
ICD9 code "tobacco use disorder" [6]	Educational intervention
Willingness to quit	Self-help materials
Quit date	Referral to external counseling/ treatment programs
Presence and type of tobacco use	Type/ characteristics of program
Arrangement of follow up	OTC nicotine replacement
Demographic information	Social support
Family history	

Discussion

A fundamental description of a smoking habit is its intensity and duration. In the study sample, the records of most current smokers contained data describing intensity and duration of smoking. However, less than half the records of former smokers contained data describing intensity and duration of smoking. Clinicians entering the data appear to be mixing units of measurement, perhaps documenting number of cigarettes vs. number of packs. This interpretation is supported by the words most frequently entered into a "Tobacco Comment" attribute, words commonly used to describe duration. Many clinicians opted to use the free text field to describe intensity and duration of a smoking habit. Similar findings were encountered in other social history table attributes. Interviews with system users reveal that in many clinics, the social history attribute data are entered by medical assistants, not by nurse practitioners and physicians as they interview a patient.

Strengths and weaknesses in concept representation were noted, and may limit the ability of a KDD project to model smoking cessation. However, there are clear strengths in concept representation that could be leveraged to learn more about primary management of smoking cessation/ tobacco dependence (see table 2). Both data quality issues and weaknesses in coded concept representation may be informed by analysis of free text patient note data, not currently stored in the EDW. It is feasible to code smoking–related concepts in EpicCare free text patient notes with natural language processing (NLP). [7]

Conclusion

This feasibility study revealed significant data quality issues in coded social history attributes. These issues included missing data, particularly in the case of former smokers, and misrepresentation of patient characteristics with default attribute values. Unusual values may be related to clinicians using inappropriate units of measurement. The study also revealed key strengths and weaknesses in an ambulatory electronic health record's coded representation of concepts with known relevance to tobacco/ smoking cessation. The weaknesses in data quality and coded concept representation are anticipated to affect a KDD project using only coded data from this repository of ambulatory patient records, but may be ameliorated with analysis of additional data contained in free text patient notes.

Acknowledgements

This study was funded by the University of Utah College of Nursing Research Committee.

References

1. Poynton MR, McDaniel AM. Classification of smoking cessation status with a backpropagation neural network. J Biomed Inform. 2006 Mar 24.
2. Epic Systems I. EpicCare EMR. 2000-2006. p. ambulatory EMR.
3. Epic Systems I. Clarity Enterprise Reporting Spring 2005 Data Dictionary. 2005 [cited; Available from:
4. Fiore MCB, W.C.; Cohen, S.J., et al. . Treating Tobacco Use and Dependence. Clinical Practice Guideline. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service; 2000.
5. Veterans Administration DoD. VA/DoD clinical practice guideline for the management of tobacco use. Washington (DC) Department of Veteran Affairs; 2004 2004 June.
6. International Classification of Diseases, Ninth Edition, Clinical Modification (ICD-9-CM).
7. 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 Dec;29(5):434-9.

Address for correspondence

Mollie R. Poynton, PhD, APRN
10 S 2000 East
Salt Lake City, UT 84112-5880, U.S.A.
mollie.poynton@nurs.utah.edu
801.585.9740