

Using Clinical Notes to Assess the Exposure of Tobacco and Alcohol Use in the Electronic Health Record

Jessica Chen¹, Abel Ermias², Cosmin A. Bejan PhD³

¹Georgia Institute of Technology, Atlanta, GA, USA; ²Lipscomb University, Nashville, TN, USA; ³Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

Tobacco and alcohol use have adverse effects on a person's body systems and are determinants for many diseases including respiratory and cardiovascular diseases. However, in general, details on usage and consumption are not standardized in the Electronic Health Record (EHR). This results in difficulty following the progression of any patient's tobacco and alcohol use history.

While clinical notes encode various forms of information on tobacco and alcohol use, identifying a detailed signature of these determinants of health as well as their longitudinal exposure has been less well studied. Our goal was to evaluate the feasibility of creating a comprehensive timeline of each patient's tobacco and alcohol use history from the clinical notes of patients admitted at Vanderbilt University Medical Center (VUMC).

Materials and Methods

The dataset used in our study included patient notes from the Synthetic Derivative (SD), a de-identified version of the Vanderbilt EHR. From this dataset, we randomly selected 105 patients with records spanning across at least 5 years and performed manual annotation of all their corresponding clinical notes with "social history" mentions. All annotations were performed using the brat rapid annotation tool. Our annotation guidelines consisted of 6 status categories for smoking use (*Current Smoker*, *Past Smoker*, *Never Smoker*, *Unknown Smoker*, *Smoker*, and *Secondary Smoker*) and 5 status categories for alcohol use (*Current Drinker*, *Past Drinker*, *Never Drinker*, *Unknown Drinker*, and *Drinker*). Of note, similar to the annotation guidelines proposed for the i2b2 challenge on smoking status identification,[1] we annotated as *Smoker/Drinker* those expressions that indicate current or past tobacco/alcohol use, but do not provide specific information for either category (eg, "history of alcohol use"). Furthermore, for each status category, we selected for annotation the following 9 attributes: *Assertion* ("denies"), *Type* ("cigarette", "wine"), *Method* ("smoke"), *Amount* ("2-3 drinks"), *Frequency* ("occasionally"), *Exposure History* ("2 yrs"), *Quit History* ("a year ago"), *Start Date* ("1991"), and *Quit Date* ("2013"). Our analysis of the annotated data included: 1) extracting descriptive statistics of annotation categories and corresponding text mentions, and 2) extracting trends of tobacco and alcohol use history from the patient timelines encoding the two determinants of health.

Results

We manually annotated a total of 2,220 clinical notes in which most of the status categories were *Never Smoker* (N=1,300) and *Never Drinker* (N=742). Less frequently annotated status categories were *Current Smoker* (127), *Past Smoker* (236), *Current Drinker* (439), and *Past Drinker* (118). Among the annotated attributes, *Assertion* (2,007), *Frequency* (811), *Amount* (532), and *Type* (499) were found as the most prevalent in the annotated dataset. Aggregation of status categories and construction of patient timelines enabled us to extract information regarding the longitudinal exposure of tobacco and alcohol use history. From the 105 analyzed patients: 1) 9 (8.6%) of them did not have tobacco use information in their notes; 2) 23 (21.9%) did not have alcohol use information; 3) 55 (52.4%) were *Never Smoker*; 4) 6 (5.7%) transitioned from *Never Smoker* to *Current Smoker*; 5) 32 (30.5%) were *Never Drinker*, and 6) 16 (15.2%) transitioned from *Never Drinker* to *Current Drinker*.

Conclusions

We presented preliminary results from a dataset of patient notes manually annotated with tobacco and alcohol use information. We found that, for most of the patients, their clinical notes contain information regarding the two determinants of health, despite the fact the majority of the annotated categories were *Never Smoker* and *Never Drinker*. Preliminary analysis indicated that timelines of tobacco and alcohol use history can be reliably constructed.

References

1. 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.