Validity of Veterans Health Administration structured data to determine accurate smoking status

Sara E Golden
Elizabeth R Hooker
Sarah Shull
Matthew Howard
VA Portland Health Care System (VAPORHCS), USA

Kristina Crothers
VA Puget Sound Health Care System, USA; University of Washington, USA

Reid F Thompson
Oregon Health & Science University, USA; VA Portland Health Care System (VAPORHCS), USA

Christopher G Slatore
VA Portland Health Care System (VAPORHCS), USA; Oregon Health & Science University, USA

Abstract
We compared smoking status from Veterans Health Administration (VHA) structured data with text in electronic health record (EHR) to assess validity. We manually abstracted the smoking status of 5,610 VHA patients. Only those with a smoking status found in both EHR text data and VHA structured data were included (n=5,289). We calculated agreement and kappa statistics to compare structured data vs. manually abstracted EHR text smoking status. We found a kappa statistic of 0.70 and total agreement of 81.1% between EHR text data and structured data for Current, Former, and Never smoking categories. Comparing EHR text data and structured data between Never and Ever smokers revealed a kappa statistic of 0.62 and total agreement of 89.1%. For comparison between Current and Never/Former smokers, the kappa statistic...
was 0.80 and total agreement was 90.2%. We found substantial and significant agreement between smoking status in EHR text data and structured data that may aid in future research.

**Keywords**
informatics, public health, smoking/harm reduction

**Introduction**

Cigarette smoking is the cause of most preventable diseases in the United States and causes 80–90 percent of all lung cancers, the leading cancer killer in America. The best way to reduce deaths from lung cancer is to reduce smoking. Accordingly, many research and quality improvement studies, as well as tobacco programs and individual clinicians, rely on the electronic health record (EHR) for evaluation and tracking of patient smoking status. Similarly, smoking status is already used in myriad studies ranging from diabetes to lung cancer, therefore categorizing patients correctly in structured data remains imperative for research as well.

Determination of data quality is also variable; therefore, validation studies must be performed to establish a basis for the use of structured data. Currently, smoking information in Veterans Health Administration (VHA) standardized structured Health Factor data have been evaluated in one previous cohort study but never benchmarked against EHR text. Our aim was to establish a criterion standard for future use in multiple areas.

**Methods**

**Cohort selection**

As part of a larger study (EPID-007-15S/IIR 16-003, VA Portland Health Care System (VAPORHCS) IRB #3225), the cohort consisted of patients identified as receiving a chest computed tomography (CT) scan (not for screening) from 2012 to 2017 in the VAPORHCS or in the VA Midwest Health Care Network (VISN 23), which includes North Dakota, South Dakota, Nebraska, Minnesota, and Iowa. We first included patients from each location who had pulmonary nodules identified on their chest CT as recorded in structured data registries. VAPORHCS developed the unsuspected radiologic findings (URF) registry in the year 2000 to track patients identified by radiologists as having a pulmonary nodule using the diagnostic internal entry number code 10. Similar to VAPORHCS, VISN23 utilizes a Lung Nodule Registry that uses the diagnostic internal entry number codes 944 and 945 to include patients found to have a nodule.

We also included patients without a nodule on their CT scan, known hereafter as “non-nodule” patients, identified by including those without codes 10, 944, or 945 on the chest CT using the VA Corporate Data Warehouse (CDW). The CDW is a national data repository relational database that collects data from several VHA clinical and administrative systems, housed at the Austin Information Technology Center. It is a relational database organized into a collection of data domains, which represent logically or conceptually related sets of data tables. Patients eligible for this analysis were those with a \( \leq 3 \)-cm pulmonary nodule (as reported in radiology report text for patients with a nodule only) based on manual chart abstraction. Based on International Statistical Classification of Diseases and Related Health Problems (9th revision; ICD-9) diagnostic codes (Supplement 1), we excluded patients with a history of lung cancer prior to the baseline imaging study, a history of previous other cancers within 5 years of the baseline imaging study, and moderate immunosuppression (including HIV and those with organ transplants) at baseline. We utilized structured query language to apply the eligibility criteria.
**EHR text categorization—criterion standard**

Four trained abstractors manually abstracted smoking status. Each began with 100 practice patients with results vetted by the study principal investigator (PI) (C.G.S., a board-certified pulmonologist) to ensure reliability. We held weekly meetings with the group, including the PI, to discuss problems and discrepancies throughout the abstraction process. The process of abstraction was the same for all patients (i.e. nodule and non-nodule) for both sites. All clinical encounter notes were searched in the EHR system using the text search term “smok.” Using the date of the patient’s chest CT scan that first identified a nodule, or their first chest CT scan for non-nodule patients, as the starting index date, we reviewed encounter notes 1 year prior, with the purpose of finding the documented smoking status closest to the index scan date. If the smoking status could not be found in the previous year, we (1) reviewed encounter notes 1 year after the index scan date, then (2) reviewed all encounter notes prior to the index scan date, and finally, (3) reviewed all notes after the index scan date, always starting with the note closest to the index scan date. We categorized smoking status as Current, Former, or Never. We defined Former smokers as those who stopped smoking more than 3 months prior to the note date. For the Ever compared with Never smoking analysis we defined Ever smokers as those who were categorized as Current or Former smokers. Hereafter, we refer to the clinician-documented smoking status as recorded in a textual encounter note in the EHR as “EHR text data.”

**Structured data categorization**

We collected structured data on smoking status derived from clinical reminders (known as Health Factors in VHA), which are automated requirements that providers must complete regularly within the EHR. These structured data from the EHR are housed in the CDW and refreshed daily. The exact content of these prompts, the frequency, and possible response entries vary by site and over time. Similar methods have been previously used in collecting smoking data. Hereafter, we refer to data recorded from Health Factors as “structured data.”

The Health Factor lookup table consists of a very large number of potential Health Factors since they are text values representing an answer to a clinical question. The types of Health Factors available in this table are associated with a unique patient and visit date in the table. These data were used to classify each patient as a Current, Former, or Never smoker at each date when a Health Factor related to smoking status was available. We used the smoking_health_factor_lookup_table_2017 created and available through the Veterans Aging Cohort Study (https://medicine.yale.edu/intmed/vacs/) to classify the large number of possible health factor types (Table 1). This resource classifies each Health Factor regarding smoking into Current, Former, Never, or Unknown smoking categories.

**Statistical analysis**

Our statistician (E.R.H.) worked with the team to develop, refine, and test the analytic protocols. We compared each patient’s structured data smoking status to our criterion standard, the clinician-documented smoking status in EHR text data. We used the smoking status recorded closest in time to the index scan from the EHR, while structured data smoking status results were based on that closest to the EHR smoking status date. Concordance between data sources was determined by percent agreement and kappa statistics comparing (1) Current, Former, and Never smokers (non-weighted and weighted), (2) Current/Former (Ever) and Never smokers (non-weighted), and (3) Current and Former/Never smokers (non-weighted). Non-weighted statistics value agreement as
perfect = 1 (e.g. Current vs Current) or not perfect = 0 (e.g. Current vs Former). Weighted statistics can be used when there are three or more levels, in our case valuing agreement where perfect = 1 (e.g. Current vs Current), one category different = 0.5 (e.g. Current vs Former), or two categories different = 0 (e.g. Current vs Never).5

In kappa analyses, strength of agreement was based on well-accepted cut-points.11 For percent agreement, p < 0.05 was used to test for statistical significance. For Ever smokers compared with Never smokers, and Current smokers compared with Never/Former smokers, we calculated four performance indicators: sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV). For Ever versus Never smokers, sensitivity is the proportion of Veterans who were Ever smokers in their structured data among those who were Ever smokers in their EHR text data. Specificity is the proportion of Veterans who were Never smokers in their structured data given that they were Never smokers in their EHR text data. Positive predictive value is the proportion of Veterans who were Ever smokers according to their EHR text data given that they were marked as Ever smokers in their structured data. Negative predictive value is the proportion of Veterans who were Never smokers according to their EHR text data given that they were marked as Never smokers in their structured data.12

We also conducted sensitivity analyses by restricting our sample to (1) individuals with ≤5 years (n = 5079) and (2) ≤2 years (n = 4499), between their EHR text data and structured data date, and then comparing individuals between (3) VAPORHCS and VISN 23, and (4) nodule and non-nodule patients. If a patient’s smoking history could not be found in either source, including if the structured data status found was not listed in the smoking_health_factor_lookup_table_2017

Table 1. Structured data top smoking status health factors by location.

| Location     | Health factor                                      | Count   |
|--------------|----------------------------------------------------|---------|
| VAPORHCS     | Current tobacco user                                | 73,144  |
| VAPORHCS     | Former tobacco user                                 | 12,876  |
| VAPORHCS     | Quit tobacco (>7 years ago)                         | 12,352  |
| VAPORHCS     | Cigarette user                                      | 9315    |
| VAPORHCS     | Tobacco offered stop smoking clinic                 | 8843    |
| VAPORHCS     | Tobacco offered PT meds (provider)                 | 8838    |
| VAPORHCS     | Lifetime non-user of tobacco                        | 8405    |
| VAPORHCS     | Tobacco offered PT meds (provider)                 | 8398    |
| VAPORHCS     | Tobacco counseling offered                          | 7897    |
| VAPORHCS     | Lifetime non-tobacco user                           | 7419    |
| VISN 23      | Current tobacco user                                | 57,411  |
| VISN 23      | Former tobacco user (7 years or more)               | 23,169  |
| VISN 23      | Current smoker                                      | 15,881  |
| VISN 23      | Lifetime non-tobacco user                           | 13,055  |
| VISN 23      | Former tobacco use (>1 year and <7 years)           | 10,709  |
| VISN 23      | Current cigarette smoker                            | 10,029  |
| VISN 23      | Former tobacco use (<1 year)                        | 5864    |
| VISN 23      | Smoking cessation referral declined                 | 5682    |
| VISN 23      | PBM Pharmd tob cess nonpharm interv                 | 4535    |
| VISN 23      | Lifetime non-smoker                                 | 3434    |

VISN 23: VA Midwest Health Care Network; VAPORHCS: VA Portland Health Care System.

*Unknown factors excluded.
(5.8%), they were dropped from these analyses (Figure 1). All analyses were conducted in Stata 15 (StataCorp LP, College Station, TX, USA).

**Results**

We included 5289 patients in our sample. Most were non-Hispanic White (67.6%) and male (95.3%). The majority of our cohort had some smoking history, with only 15–20 percent being Never smokers (Table 2). Time between participants’ EHR date and their matched structured data date was a median of 81 days (IQR = 4, 364), with 75.5 percent of participants having less than 1 year difference, and 85.1 percent of participants having less than 2 years difference.

We found that for EHR text data and structured data comparison for Never, Former, and Current smoking categories, there was substantial agreement (non-weighted kappa = 0.70 and weighted kappa = 0.73). The total agreement was 81.1 percent non-weighted and higher when weighted at 89.7 percent (both p < 0.01). Of those classified as Former or Current smokers in their EHR text data, 15 percent and 2 percent (respectively), were classified as Never smokers in structured data.

For comparison between EHR text data and structured data for Never compared with Ever smokers, the kappa statistic was substantial at 0.62 and total agreement was 89.1 percent (p < 0.01). Only 9 percent of Ever smokers identified in the EHR were classified as Never smokers in structured data; 21 percent of Never smokers were classified as having Ever smoked (Table 3). Using EHR data as the criterion standard, we found 0.91 sensitivity, 0.79 specificity, 0.96 PPV, and 0.61 NPV.

For comparison between Current and Never/Former smokers, the kappa statistic was substantial at 0.80 and total agreement was 90.2 percent (p < 0.01). Only 11 percent of Current smokers identified in EHR text data were classified as Never/Former smokers in structured data; 9 percent of Never/Former smokers were classified as Current smokers (Table 3). We found 0.89 sensitivity, 0.91 specificity, 0.87 PPV, and 0.93 NPV. In our first sensitivity analysis, (1) patients with ≤5 years and (2) patients with ≤2 years between EHR text data date and structured data date, we found negligible increases (<1%) in both kappa and percent agreement values. In addition, we stratified
analyses by location and found negligible differences between sites (site 1 (n = 3008): agreement = 87.5%, kappa = 0.59, sensitivity = 0.89, specificity = 0.79, PPV = 0.96, NPV = 0.58; site 2 (n = 2281): agreement = 91.2%, kappa = 0.67, sensitivity = 0.93, specificity = 0.80, PPV = 0.97, NPV = 0.66). Finally, we assessed for differences between nodule and non-nodule patients and found negligible differences in kappa and percent agreement between these groups.

Table 2. Demographic characteristics of Veterans study sample, restricted to participants with structured and EHR text data (n = 5289).

| Characteristic                                      | n (%) or mean(± SD) |
|----------------------------------------------------|----------------------|
| Age EHR (years)                                    | 63.3 (±11.6)         |
| Race/ethnicity                                     |                      |
| Non-Hispanic White                                 | 3576 (67.6%)         |
| Non-Hispanic Black                                 | 120 (2.3%)           |
| Hispanic                                           | 45 (0.9%)            |
| Other/unknown                                      | 1548 (29.2%)         |
| Male                                               | 5039 (95.3%)         |
| Smoking status from EHR text data                  |                      |
| Never                                              | 797 (15.1%)          |
| Former                                             | 2371 (44.8%)         |
| Current                                            | 2121 (40.1%)         |
| Smoking status from structured data                |                      |
| Never                                              | 1043 (19.7%)         |
| Former                                             | 2076 (39.3%)         |
| Current                                            | 2170 (41.0%)         |
| Time between structured data and index date (years) | 0.9 (±1.8)           |

EHR: electronic health record.

Table 3. Cross-tabulation of smoking status as listed in structured and EHR text data (n = 5289).

| EHR text data = criterion standard                  | Never | Former | Current |
|----------------------------------------------------|-------|--------|---------|
| Structured data                                    |       |        |         |
| Never                                              | 632   | 366    | 45      |
| Former                                             | 117   | 1770   | 189     |
| Current                                            | 48    | 235    | 1887    |
| Never                                              |       |        |         |
| Ever                                               | 632   | 411    |         |
| Never/Former                                       | 165   | 4081   |         |
| Current                                            | 2885  | 234    |         |

EHR: electronic health record.
Results presented as n.
Discussion

In a cohort of Veterans from multiple VHA facilities with high rates of Current and Former cigarette use, we found substantial agreement between VHA structured data and the criterion standard of EHR clinician-documented smoking status for Current, Former, and Never smoking categories, as well as for Ever compared with Never, and Current compared with Never/Former smoking categories. The VHA, the nation’s largest integrated health system, has a long-standing history of using a high-quality EHR system that focuses on “clinically relevant record keeping that improves patient care.” While the ease of obtaining and using standardized structured data taken routinely from the EHR compared to manual EHR text data abstractions is paramount for time- and cost-effectiveness for research and clinical studies in all settings, it still must be validated. For example, a recent provocative study used structured data to evaluate DNA methylation sites that are often associated with smoking. The VHA is poised to be a leader in accurate and useful data sets, especially for assessments of smoking status given our findings, but other medical systems have similar datasets that can be equally as useful.

One previous Veterans Administration (VA) Health Services Research report documented 70 percent of VHA users had timely information on smoking status in the administrative database. This same report concluded that structured data are useful for determining tobacco use status for epidemiologic studies but also has limitations that make the data questionable for determining smoking prevalence. For instance, the study did not validate administrative data with the EHR or other data sources, although subsequent studies used similar methods to report on cessation trends as well as use of pharmacotherapy for tobacco cessation. Our study is the first to provide validation of structured smoking data with EHR text data. We also have less missing data across sites compared to a previous validation study of VHA data, and include sensitivity analyses not included in prior research. Our results overcome existing limitations and resolve questions regarding the usability of these data.

Limitations

The widely used and generally agreed upon cut-off for kappa statistics does not determine which data source is “correct,” but rather is a measure of concordance. In addition, all data are subject to reporting errors and may lead to underreporting of smoking status, especially for Never smokers, since they may be documented less frequently than Former or Current smokers. However, our number of missing observations are relatively low (5.8%) and similar to McGinnis et al. Since documentation of smoking status is based on self-report, it is vulnerable to recall, moderator, and documentation biases. A recent study among patients referred for lung cancer screening found very high discordance between pack-years recorded in structured data and during an intake interview with a dedicated licensed nursing professional but it is not clear if smoking status would be subject to same discordance. All participants had a CT scan in a VHA facility so our results should be validated in other VHA and non-VHA populations. VHA data may also misclassify some patients as cigarette smokers based on use of pipes or cigars, although we did exclude those who were specifically classified as such. During EHR text data review, we limited our search term to “smok,” so we may have missed documentation of “tobacco” or “cigarettes.” We also did not calculate inter-rater reliability statistics for the trained abstractors, however we did conduct 100 practice patient charts each, reviewed by the PI to ensure consistency, and we incorporated weekly team meetings to discuss problems and discrepancies.

Conclusion

Our study showed substantial agreement between abstracted EHR text data smoking status and readily accessible structured data smoking status in a large cohort of Veterans. Assessing the
validity of smoking status in structured data aids in future research and for identification of patients eligible for clinical interventions. Structured data is cost-effective, quicker to obtain, and is generally standardized for a large population of patients, making it an attractive form of information for researchers and policy-makers. As we move toward increased information sharing and the use of “big data” evolves, our study provides important validation on the accessibility of smoking status data in administrative structured data sources.

**Author’s note**

The Department of Veterans Affairs did not have a role in the conduct of the study, in the collection, management, analysis, interpretation of data, or in the preparation of the manuscript. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the US Government.

**Acknowledgements**

The authors would like to acknowledge the help and support of Tara Thomas, BS, Sujata Thakurta, MPA: HA, and Philip Tostado, MA for their countless hours of manual chart review.

**Author contributions**

All authors have made substantial contributions to the (1) conception and design, acquisition of data, or analysis and interpretation of data; (2) have contributed to drafting the article for important intellectual content; and (3) have provided final approval of the version to be published. S.G. takes responsibility for the content of the manuscript, including data and analysis.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study and Dr Slatore are supported by an award from the Department of Veterans Affairs (CSRD EPID-007-15S and HSRD IIR 16-003). It was also supported by resources from the Center to Improve Veteran Involvement in Care, VA Portland Health Care System, Portland, OR, USA.

**ORCID iD**

Sara E Golden [https://orcid.org/0000-0003-0536-5853](https://orcid.org/0000-0003-0536-5853)

**Supplemental material**

Supplemental material for this article is available online.

**References**

1. Centers for Disease Control and Prevention. Smoking & tobacco use, https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm (2018, accessed 8 March 2018).
2. Centers for Disease Control and Prevention. What are the risk factors for lung cancer?, https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm (2018, accessed 14 December 2018).
3. Siegel RL, Miller KD and Jemal A. Cancer statistics, 2016. *CA Cancer J Clin* 2016; 66(1): 7–30.
4. Anthonisen NR, Skeans MA, Wise RA, et al. The effects of a smoking cessation intervention on 14.5-year mortality: a randomized clinical trial. *Ann Intern Med* 2005; 142: 233–239.
5. McGinnis KA, Brandt CA, Skanderson M, et al. Validating smoking data from the Veteran’s Affairs health factors dataset, an electronic data source. *Nicotine Tob Res* 2011; 13(12): 1233–1239.

6. Uzuner O, Goldstein I, Luo Y, et al. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc* 2008; 15(1): 14–24.

7. Bernardes C. Australian indigenous cancer patients’ self-report of diagnosis, treatment and co-morbidity data: how does it compare to medical chart review? *Eur J Pers Cent Healthc* 2016; 4: 330–339.

8. Holden WE, Lewinsohn D, Osborne ML, et al. Use of a clinical pathway to manage unsuspected radiographic findings. *Chest* 2004; 125(5): 1753–1760.

9. Shelver J, Wendt CH, McClure M, et al. Effect of an automated tracking registry on the rate of tracking failure in incidental pulmonary nodules. *J Am Coll Radiol* 2017; 14(6): 773–777.

10. U.S. Department of Veterans Affairs, Health Services Research and Development Service, VA Information Resource Center. VIREC resource guide: VA corporate data warehouse, https://www.virec.research.va.gov/Resources/RUGs.asp (2016, accessed 8 March 2018).

11. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; 33(1): 159–174.

12. Apostolović M, Toplaović A and Golubović M. Understanding sensitivity, specificity and predictive values. *Vojnosanit Pregl* 2014; 71(12): 1167.

13. The Plone Foundation. About VistA, http://worldvista.org/AboutVistA (accessed 8 March 2018).

14. McGinnis J, Justice AC, Tate JP, et al. Using DNA methylation to validate an electronic medical record phenotype for smoking. *Addict Biol* 2019; 24: 1056–1065.

15. Barnett PG, Chow A and Flores NE. *Using tobacco health factors data for VA health services research*. Technical report 28, February 2014, Menlo Park, CA: VA Palo Alto Healthcare System.

16. Barnett PG, Chow A, Flores NE, et al. Changes in Veteran tobacco use identified in electronic medical records. *Am J Prev Med* 2017; 53(1): e9–e18.

17. Ignacio RV, Barnett PG, Kim HM, et al. Trends and patient characteristics associated with tobacco pharmacotherapy dispensed in the veterans health administration. *Nicotine Tob Res* 2017; 20(10): 1173–1181.

18. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012; 22(3): 276–282.

19. Pannucci CJ and Wilkins EG. Identifying and avoiding bias in research. *Plast Reconstr Surg* 2010; 126(2): 619–625.

20. Modin HE, Fathi JT, Gilbert CR, et al. Pack-year cigarette smoking history for determination of lung cancer screening eligibility. Comparison of the electronic medical record versus a shared decision-making conversation. *Ann Am Thorac Soc* 2017; 14(8): 1320–1325.