ICD-9 tobacco use codes are effective identifiers of smoking status

Laura K Wiley,1,2 Anushi Shah,2 Hua Xu,2 William S Bush1,2

ABSTRACT

Objective To evaluate the validity of, characterize the usage of, and propose potential research applications for International Classification of Diseases, Ninth Revision (ICD-9) tobacco codes in clinical populations.

Materials and methods Using data on cancer cases and cancer-free controls from Vanderbilt’s biorepository, BioVU, we evaluated the utility of ICD-9 tobacco use codes to identify ever-smokers in general and high smoking prevalence (lung cancer) clinic populations. We assessed potential biases in documentation, and performed temporal analysis relating transitions between smoking codes to smoking cessation attempts. We also examined the suitability of these codes for use in genetic association analyses.

Results ICD-9 tobacco use codes can identify smokers in a general clinic population (specificity of 1, sensitivity of 0.32), and there is little evidence of documentation bias. Frequency of code transitions between ‘current’ and ‘former’ tobacco use was significantly correlated with initial success at smoking cessation (p<0.0001). Finally, code-based smoking status assignment is a comparable covariate to text-based smoking status for genetic association studies.

Discussion Our results support the use of ICD-9 tobacco use codes for identifying smokers in a clinical population. Furthermore, with some limitations, these codes are suitable for adjustment of smoking status in genetic studies utilizing electronic health records.

Conclusions Researchers should not be deterred by the unavailability of full-text records to determine smoking status if they have ICD-9 code histories.

BACKGROUND AND SIGNIFICANCE

Smoking is a well-recognized public health concern as it is a risk factor for multiple diseases such as cancer, asthma, chronic obstructive pulmonary disease, and heart disease.1 Because smoking alters the risk for disease, it is important to ascertain smoking status in epidemiological, clinical, and genetic studies.2 Traditional studies obtain this information through questionnaires or interviews, however many institutions are using de-identified versions of electronic medical records (EMRs) for clinical and genetic research.3–6 Smoking status is often stored in narrative text in EMRs, making identification more difficult.7 8 While documentation of smoking status as structured data is one of the objectives of ‘Meaningful Use’,9 it is important to be able to extract smoking status from records that may not be updated due to death or change of treatment facility.

Natural language processing (NLP) has shown promise in extracting smoking status from the text of medical records, and was the focus of the 2008 i2b2 Challenge.7 8 10 Research has demonstrated that NLP can determine specific smoking patterns (current, past, never, etc).11–13 Nevertheless, non-informaticians may be unable to use these algorithms due to technical limitations or privacy concerns since NLP requires medical record text.

Another potential method for identifying smoking status from EMRs is to use billing codes. The International Classification of Diseases, Ninth Revision (ICD-9) codes are frequently used in phenotype extraction algorithms.14 There are two primary ICD-9 tobacco use codes: 305.1—Current tobacco use and V15.82—History of tobacco use. Although some studies have reported limited sensitivity of these codes,15 no formal analysis of the validity of these codes has been performed.16

In this three-part study we evaluated the validity of, characterized the usage of, and propose potential research applications for ICD-9 tobacco use codes in clinical populations. To validate these codes we measured their accuracy in a general clinic population and a high smoking prevalence population (individuals with lung cancer). Characterization analyses evaluated documentation biases by providers/clinics and evaluated the consistency and appropriateness of assignment based on the temporal aspects of the codes. Finally, we explored the potential utility of these codes for genetic analyses by evaluating their suitability as controls for smoking status in a genetic association study of lung cancer.

MATERIALS AND METHODS

Study cohort
Given that smoking increases the risk of cancer, we analyzed a cohort originally ascertained to examine genetic associations with different forms of the following cancers: breast, colorectal, endometrial, lung, melanoma, non-Hodgkin’s lymphoma, ovarian, and prostate. This cohort of 17 514 individuals consists of 7050 cancer cases and 10 464 cancer-free controls collected from BioVU, the Vanderbilt DNA Biobank.4 To be included as a cancer case, a valid tumor registry entry (the Tennessee Cancer Registry) was required. For breast, prostate, and melanoma cancer, individuals were also considered cases if there were three instances of an appropriate ICD-9 code (174.*, 185.*, and 172.*, respectively). Age and sex matched controls were required to have two or more clinical narratives (clinical note, discharge summary, etc) in their electronic record. Colorectal cancer controls with at least one normal colonoscopy were preferentially selected. Control individuals were excluded if they had any ICD-9 codes...
between 140.8 and 239.8 (the general range for neoplasms) or an entry in the tumor registry. Additionally, individuals were excluded if their problem list contained any words that could indicate any form of cancer.

**Smoking status definitions**

This study compares manually reviewed smoking status to two automated definitions of ever/never-smokers: ICD and NLP. ‘ICD ever-smokers’ had one or more ICD-9 tobacco use disorder (305.1) or history of tobacco use (V15.82) codes. ‘ICD never-smokers’ had no record of either of these codes. NLP smoking status was based on an implementation of the cTAKES smoking algorithm17 at Vanderbilt.18 To create the gold standard, a single reviewer read the full text of selected medical records to identify smoking status. Individuals were classified as ever-smokers if there was any mention of smoking in the record, regardless of duration, quantity, or time since cessation. Never-smokers were required to have explicit documentation of smoking status as none, never, or life-time non/never-smoker in their record.

**Assess validity of codes within clinical populations**

To determine the validity of ICD-9 tobacco use codes within a general clinic population, we examined 100 golden standard ever-smokers and 100 golden standard never-smokers from our control group of individuals without cancer. This sample size provided 85% power assuming an α of 0.05. We used this population because it is representative of the general clinic population at Vanderbilt. We categorized each individual as an ‘ICD ever/never-smoker’ and an ‘NLP ever/never-smoker’ and calculated the sensitivity, specificity, and accuracy of each definition. We also combined the two definitions so that ‘ever-smokers’ consisted of those labeled as smokers by either the ICD or NLP method (or both).

Next we investigated whether there was sufficient power to detect an association between smoking and lung cancer risk given the previously reported low sensitivity of the codes.15 We performed logistic regression on ICD ever/never-smokers for 731 lung cancer cases and 9003 controls adjusting for age and gender. Additionally, we expected that smoking status would be documented more frequently in lung cancer cases as smoking is a significant risk factor.1 As a result, we hypothesized that a lack of smoking codes for individuals with lung cancer could indicate never-smoking status. To evaluate this hypothesis, a single reviewer examined the records of all 288 ‘ICD never-smokers’ in our lung cancer population and classified individuals with the gold standard definition. Eighty of these records were reviewed a second time to confirm the accuracy of the classification. We then calculated the false negative rate (FNR) of the ICD definition (ie, ratio of gold standard never-smokers to ‘ICD never-smokers’). Individuals were excluded from this analysis if smoking status could not be identified by manual review. Given that the risk for lung cancer diminishes as a function of time since smoking cessation,19 we also hypothesized that healthcare providers may limit documentation of remote smoking. Thus we also considered a ‘relative risk’ definition where individuals who quit smoking prior to 1990 were considered non-smokers. This ‘cut point’ corresponds to at least 15 years since cessation since all individuals had one or more visits during or after 2007. We then compared the ICD smoking definition to the ‘relative risk’ gold standard with a FNR calculation. Former smokers whose quit date could not be determined from the record were excluded from the second analysis only.

**Characterize usage of codes**

To understand potential documentation biases, we examined which types of providers/clinics tend to document smoking status with ICD-9 codes. Unfortunately, in our research version of the EMR, it is not possible to identify providers or clinics that assign particular ICD-9 codes. However, we used contemporaneously assigned ICD-9 and Current Procedural Terminology (CPT) codes as surrogate variables for the type of provider. Thus, we extracted all ICD-9 and CPT codes for the 3752 ‘ICD ever-smokers’ in the entire study population. The codes were stratified based on co-documentation with tobacco use codes (ie, whether the code was assigned on the same date as the tobacco use code). We used the frequency of individuals with and without co-documentation to measure the difference in reporting of tobacco use.

Interestingly, the two ICD-9 tobacco use codes are divided into current and past use. We wanted to identify whether providers used the two codes interchangeably and whether temporal arrangements of these codes modeled actual behavior patterns (as recorded in the medical record). In order to capture individuals who had multiple smoking cessation attempts, we required individuals to have at least two current tobacco use ICD-9 codes (305.1) and at least two ‘history of tobacco use’ ICD-9 codes (V15.82) for this analysis. We ordered each individual’s smoking codes by date and then collapsed identical codes to create a summary of chronological code transitions (ie, switching from 305.1 to V15.82 and visa versa over time). The number of transitions for each subject was tallied and individuals were grouped based on whether they had one transition or more than one transition. Following manual review of the medical records, individuals were grouped into four categories: always smokers (ie, no notation of trying to quit), successful quitters (ie, smoked then quit and never resumed), unsuccessful quitters (eg, smoked, quit, and resumed in any pattern and frequency), and those who we were unable to classify. Individuals in the last category were removed from further analysis. We then performed a 2×3 χ2 analysis comparing the two types of transitions (single vs multiple) and actual smoking status (always, single successful quit attempt, unsuccessful quit attempt(s)).

**Investigate research applications of codes**

Finally, given the increased interest in using EMR-linked biobanks for genetic research,5 6 we investigated the suitability of ICD identified smoking status as a covariate in genetic analysis. We anticipated that the low ICD-9 documentation of smoking status would reduce our power to detect the effects of single nucleotide polymorphisms (SNPs). This reduction in power would likely reduce the accuracy of effect estimates for SNP analysis. To test these hypotheses, we performed logistic regression on 12 SNPs for 731 lung cancer cases and 9003 controls. There are well established relationships between SNPs that alter nucleotide polymorphisms (SNPs). This reduction in power is especially important for these analyses. The SNPs examined were previously reported in genome-wide association studies as associated with lung cancer. Our regression model controlled for gender, age, the top three principle components (reflecting genetic ancestry), and smoking status. We evaluated three models for each SNP: one using the NLP-defined smoking status, one using the ICD-defined smoking status (described earlier), and one unadjusted for smoking status. For comparisons, regression models using the NLP-defined smoking status are considered our ideal model. For the results of the regression
models we used a significance threshold of p<0.004 (the Bonferroni correction for 12 independent tests). Given that many of the associations may be robust against the effect of smoking, we examined all SNPs where there was non-concordance between the three models at a nominal significance level (p<0.05). As before, we used the NLP-adjusted model as the ideal and determined how the ICD-adjusted model compared. To assess how smoking definition alters statistical power to detect the effects of lung cancer SNPs, we performed post hoc power calculation using CaTS V0.2.21 We assumed a lung cancer prevalence of 10% based on available SEER data from 2009,22 and ORs from each regression model were used as point estimates of the genotype effect size. Across all SNPs, we calculated the average change in statistical power between the models adjusted using the ICD or NLP methods. The end result of this power calculation is the expected change in power to detect lung cancer-associated SNPs when using ICD-based smoking status versus NLP-based smoking status.

RESULTS
A breakdown of the cohort by cancer and control group type as well as ICD smoking status is presented in table 1. Additionally table 1 gives counts of individuals with multiple code transitions (two or more 305.1 codes and two or more V15.82 codes) and the average number of ICD-9 tobacco use codes for these individuals. The maximum number of codes any single individual in the transition group had was 43 codes (from the breast cancer control cohort).

Assess validity of codes within clinical populations
Performance metrics of the ICD and NLP definitions of ever-smokers on the sampled individuals are provided in table 2.

The logistic regression of ICD-defined smokers on lung cancer incidence showed a significant smoking effect with an OR (95% CI) of 9.69 (8.21 to 11.43). Results of the FNR analysis in individuals with lung cancer are presented in table 3. Of the 288 lung cancer cases without an ICD-9 code for tobacco use, we were unable to identify the smoking status of five individuals, who were excluded from both analyses. We calculated the FNR using 86 never-smokers and 197 ever-smokers. In the second analysis, eight individuals had quit smoking, but a quit date could not be determined, so they were excluded from the second calculation only. The low risk category consisted of 149 individuals (86 never-smokers and 63 ever-smokers with >15 years since smoking cessation). The high risk category consisted of 126 individuals (108 who quit after 1990 and 18 whose records indicate continuous smoking).

Characterize usage of codes
Figures 1 and 2 describe the relative frequency of ICD-9 and CPT codes documented concurrently (ie, on the same date) with the smoking ICD-9 codes. ICD-9 codes are broken down by general code type in figure 1. The ICD-9 codes most frequently co-documented with tobacco use ICD-9 codes are: 17.42—laparoscopic robotic assisted procedure, 40.3—regional lymph node excision, 40.53—radical excision of iliac lymph nodes, 60.5—radical prostatectomy, and V38.66—long-term (current) use of aspirin. The three codes least frequently documented with smoking are: 780.79—other malaise and fatigue, V67.09—following other surgery, and V72.83—other specified preoperative examination. CPT codes are plotted in figure 2. The most frequently co-documented codes are: 83571—laparoscopic surgery with bilateral total pelvic lymphadectomy, 53866—laparoscopic prostatectomy, and 88309—level VI surgical pathology gross and microscopic examination. The least frequently co-documented codes are: 36415—radical excision of iliac lymph nodes, 36586—laparoscopic robotic assisted procedure, 53866—laparoscopic prostatectomy, and 88309—level VI surgical pathology gross and microscopic examination. The least frequently co-documented codes are: 36415—collection of venous blood by venipuncture, 99024—postoperative follow-up visit, and 99213—office outpatient visit.

Of the 116 eligible individuals, six were excluded from the analysis because their smoking pattern could not be determined after manual review of their medical records. A contingency

### Table 1

<table>
<thead>
<tr>
<th>Cohort summary (n=17514)</th>
<th>ICD ever-smokers,* n (%)</th>
<th>Individuals with multiple codes,† n (average number of codes)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cases</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast cancer (n=1377)</td>
<td>217 (15.8%)</td>
<td>3 (7)</td>
</tr>
<tr>
<td>Colorectal cancer (n=912)</td>
<td>270 (29.6%)</td>
<td>4 (7)</td>
</tr>
<tr>
<td>Endometrial cancer (n=238)</td>
<td>33 (13.9%)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Lung cancer (n=924)</td>
<td>588 (63.6%)</td>
<td>50 (11)</td>
</tr>
<tr>
<td>Melanoma (n=1348)</td>
<td>275 (20.4%)</td>
<td>4 (5)</td>
</tr>
<tr>
<td>Non-Hodgkin’s lymphoma (n=348)</td>
<td>68 (19.5%)</td>
<td>5 (7)</td>
</tr>
<tr>
<td>Ovarian cancer (n=186)</td>
<td>46 (24.7%)</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Prostate cancer (n=2145)</td>
<td>569 (26.5%)</td>
<td>10 (10)</td>
</tr>
<tr>
<td><strong>Controls</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast cancer (n=2753)</td>
<td>358 (13.0%)</td>
<td>9 (13)</td>
</tr>
<tr>
<td>Colorectal cancer (n=1822)</td>
<td>333 (18.3%)</td>
<td>7 (12)</td>
</tr>
<tr>
<td>Melanoma (n=2630)</td>
<td>472 (17.9%)</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Ovarian cancer (n=342)</td>
<td>51 (14.9%)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>Prostate cancer (n=4288)</td>
<td>931 (21.7%)</td>
<td>17 (11)</td>
</tr>
</tbody>
</table>

*One or more 305.1 or V15.82 ICD-9 codes.
†Two or more 305.1 ICD-9 codes and two or more V15.82 ICD-9 codes.

### Table 2

<table>
<thead>
<tr>
<th>Performance of ICD, NLP, and combined definitions of ever-smokers from group of ever-smokers (n=100) and never-smokers (n=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitivity (95% CI)</strong></td>
</tr>
<tr>
<td>ICD only</td>
</tr>
<tr>
<td>NLP only</td>
</tr>
<tr>
<td>ICD+NLP*</td>
</tr>
</tbody>
</table>

*Ever-smokers if either ICD or NLP (or both) classify as ever-smoker.
ICD, International Classification of Diseases; NLP, natural language processing.

### Table 3

<table>
<thead>
<tr>
<th>False negative rate (FNR) of International Classification of Diseases definition of never-smokers in a lung cancer case cohort (n=288)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
</tr>
<tr>
<td>Gold standard</td>
</tr>
<tr>
<td>Ever-smokers</td>
</tr>
<tr>
<td>Never-smokers</td>
</tr>
<tr>
<td>Unable to classify</td>
</tr>
<tr>
<td>Relative risk</td>
</tr>
<tr>
<td>Ongoing smoking</td>
</tr>
<tr>
<td>Quit smoking during/after 1990</td>
</tr>
<tr>
<td>Quit smoking prior to 1990</td>
</tr>
<tr>
<td>Unable to assign quit date</td>
</tr>
</tbody>
</table>
table comparing the number of code transitions (single or multiple) to behavior patterns (always smoker, single successful quit attempt, and one or more unsuccessful quit attempts) for the remaining 110 individuals is presented in table 4. The mean numbers of smoking codes for single and multiple code transitions are 11 and 10 codes per individual, respectively.

**Investigate research applications of codes**

Results from the logistic regression of 12 SNPs on lung cancer are presented in figure 3. Each SNP has three regression results: one unadjusted for smoking status, one adjusted for NLP-derived smoking status, and the other adjusted using ICD-defined smokers. All SNP models—even those that are not significant—have the same direction of effect and overlapping 95% CIs. Two SNP associations had non-concordant significance levels between the three different models (at a nominal significance cut-off of \( p < 0.05 \)). rs402710 was listed as nominally significant in both the unadjusted and ICD-adjusted models (\( p = 0.027 \) and \( p = 0.044 \), respectively). The NLP-adjusted model for this SNP was not significant (\( p = 0.065 \)). rs7626795 was nominally significant in the unadjusted model (\( p = 0.047 \)) and not significant in either the NLP or ICD-adjusted models (\( p = 0.15 \) and \( p = 0.17 \), respectively). Post hoc power calculations using these ORs averaged a 0.8% decrease in power (SD 4.6%) for models using ICD-defined smoking status.

**DISCUSSION**

Previous studies have suggested that tobacco use ICD-9 codes are an unreliable method of determining smoking status. In this study we sought to validate the accuracy of these codes and investigate potential uses of these data for future research studies. Our study supports the notion that tobacco use codes are not particularly sensitive within a general (without cancer) clinic population, although our study showed a higher sensitivity than previously reported. This could be due to sampling differences in clinic populations as Vanderbilt Medical Center is a tertiary care facility treating patients from mid Tennessee and Kentucky—areas with a high prevalence of smoking. An important feature of detection methods is preventing false positives. The specificity of the tobacco use ICD-9 codes was perfect, indicating the exceptional utility of these codes for identifying true smokers. Indeed, none of the ~150 records with a tobacco use ICD-9 code examined was an identifiable false positive (five records had no explicit documentation of tobacco use). This high specificity supports the use of these codes for the identification of smokers for clinical studies, making participant selection easier and increasing patient privacy as this method does not require the full text of the medical record. While the significance and interpretation of the results of the NLP methods alone are discussed elsewhere, our analysis demonstrates that combining ICD-9 codes with the NLP method increased the sensitivity over the NLP method alone.

It is well known that smoking is a significant risk factor for lung cancer. As a measure of the efficacy of tobacco codes to identify smokers, we replicated the association of smoking with lung cancer using ICD-identified smokers. Published ORs for the effect of smoking on lung cancer vary widely based on gender, age, duration, and amount of smoking as well as lung cancer histologic type, with values ranging from 10 to 30-fold increased risk. The OR from our logistic regression model was on the low end of these estimates, an expected result given the granularity of our tobacco exposure measure as ever/never-smokers rather than the more specific pack-years measure. Additionally, we found that women had a slightly higher OR than men when we stratified our analysis by gender (data not shown). One potential confounding element in this analysis is

---

**Figure 1** Frequency of International Classification of Diseases, Ninth Revision (ICD-9) co-documentation with tobacco use codes. Each point represents a single ICD-9 code. Points to the right of the equal documentation line (indicated in white) have an increased frequency of co-documentation with an ICD-9 tobacco use code (e.g., 300 more individuals had their code documented on the same day than those whose code was documented on different days). Points to the left of the center line have a reduced frequency of co-documentation with tobacco use codes.
the increased documentation rates of smokers with lung cancer as compared to populations without cancer. We did not specifically measure the sensitivity of ICD tobacco use codes in our lung cancer population, and we made the assumption that codes would continue to have perfect specificity as in populations without cancer and only examined individuals without tobacco use ICD-9 codes. Our hypothesis was that if documentation increased in this population, it was possible that a lack of documentation was informative of never-smoking status. Unfortunately, although tobacco use has a great effect on recurrence risk and treatment protocols for lung cancer, our data do not support that the lack of a smoking code is reflective of actual never-smoking status. Interestingly, the rate of false negatives declined when we considered relative risk from tobacco use. It is well known that the risk of cancer decreases with increased time since cessation of smoking. It is possible that as the risk decreased so did the impetus to document smoking status with ICD-9 codes. To support this hypothesis, the time since cessation for former smokers who did receive an ICD-9 code would need to be examined. Still, our data do not support using the lack of a smoking code as an indicator of never-smoking or remote smoking history.

A potential concern in using these codes for study ascertainment is the induction of an unknown bias towards certain clinical populations. For example, if tobacco use ICD-9 codes were preferentially documented in a cancer clinic, studies ascertaining smokers using the codes would be enriched for cancer patients, potentially biasing study results. Our study is limited because we cannot link the ICD-9 codes to any particular clinic or physician. This effect may be compounded by billing practices at Vanderbilt. Typically in outpatient clinics the physician assigns ICD-9 codes for visits, whereas patients admitted to the hospital are assigned codes by a team of professional billers in consultation with the physician and the medical record. With this caveat, our results demonstrated a slight increase in co-documentation of tobacco codes with surgical procedures. This could reflect the increased risk of surgery in smoking populations or justification for billing for additional services. Performing this analysis across the clinical EMR (where clinic and provider level data are available) may provide better resolution of potential biases. Additionally, future analyses should be performed at a different medical center to assess the generalizability of these trends.

The two tobacco ICD-9 codes have a temporal element reflecting current use (305.1) and past use (V15.82). However, it was unknown whether these codes are used interchangeably. Transitions between these codes showed remarkable similarity to behavioral changes recorded in the EMR, indicating that the temporal component of the code appears to be used appropriately. Discrepancies between assigned ICD-9 codes and actual usage could reflect differences in clinical practice or billing policies. Future research should consider implementing electronic portals that automatically record smoking status to improve documentation rates and reduce bias in study ascertainment.
smoking behavior were often due to either persistent duplication of a single smoking misclassification, or a failure to document updated smoking status in patient problem lists. This error may be inflated by in-patient professional medical coders using the problem list for billing rather than physician notes. Interestingly, the history of the tobacco use code was often documented when heavy smokers reduced cigarette consumption, perhaps reflecting the physician anticipating complete cessation. Additionally, given the significant interest in identifying factors affecting smoking cessation, we evaluated the potential of code transitions to model smoking behavior as single successful quit attempts or one or more failed quit attempts. Extracting even this granular classification of behavior from a clinical record would be difficult given the complexity of identifying sequences of events with NLP. Our results support the hypothesis that multiple code transitions are associated with one or more failed quit attempts. Thus researchers could use the frequency of transitions between ICD-9 tobacco use codes to identify individuals who have had difficulty with smoking cessation.

An important caveat of this analysis is that we are matching to smoking status as recorded in the medical record. Given that the accuracy and completeness of smoking status in medical records is variable, it is unclear how code transitions relate to actual behavioral patterns. Excitingly, however, this work serves as a proof of concept that temporal transitions in codes can be used to accurately model transitions in the medical record. Further research is needed to determine how this temporal modeling works on other types of ICD-9 codes, but our data suggest that using transitions among ICD-9 codes may be useful in phenotyping algorithms.

Finally, as more genetic studies use phenotypes derived from EMRs, we investigated the suitability of these codes for use as a covariate in genetic models. We performed association analyses with lung cancer, as controlling for smoking status is crucial in these models. In data to be presented elsewhere, we have shown that the regression models using NLP-derived smoking status as a covariate provide similar results to traditionally collected genetic epidemiology studies. When this ideal model is compared to models using the ICD-defined smokers for adjusted and unadjusted models, many of the SNP associations appear robust to the effect of smoking (as evidenced by no difference between unadjusted and adjusted models). While the similarity of ORs and significance levels between NLP and ICD-adjusted models is encouraging, our data are insufficient to definitively declare the equivalence of these methods. To better understand some of the potential biases of the two correction models, we reduced the significance threshold to nominal significance (p<0.05) and examined associations where the models were not uniformly concordant. One association (rs402710) was identified as not significant in the NLP-adjusted model and significant in the unadjusted and ICD-adjusted models. It is unclear whether this result is an effect of reduced power to adjust smoking by the ICD-based method or a feature of the underlying SNP association. While the original report of this associate showed only slight attenuation of the signal based on tobacco use status, another SNP in moderate linkage disequilibrium (rs31489, r^2 = 0.687) was highly affected by smoking status as the association was only significant in current smokers. The second non-concordant association was for rs7626795, where both smoking adjusted models eliminated the

Figure 3 Effects of ICD-based tobacco use covariates on genetic associations. Single nucleotide polymorphism (SNP) ORs estimated by logistic regression for unadjusted (black) and adjusted for natural language processing (blue) or ICD (green) defined smoking status are plotted with their 95% CIs. Significant SNP models (Bonferroni corrected, p<0.004) are indicated in bold. Published ORs for the significant SNPs are represented by the red points. ICD, International Classification of Diseases; NLP, natural language processing.


Research and applications
nominal significance of the association from the unadjusted model. Interestingly, this association failed to replicate in the second stage of the reporting study, supporting our lack of replication under both adjusted models. Unfortunately, these data are inconclusive for evaluation of correction of SNP associations for smoking status. Ideally we would test these adjustment methods on SNPs that are highly sensitive to tobacco use status. Nevertheless, the totality of evidence presented in this paper leads us to cautiously support the usage of ICD-9-defined smoking status for use in genetic association studies. Given the decreased sensitivity of this method compared to NLP, we expected there would be a decrease in power to detect SNP effects. However, we observed a less than 1% average reduction in power using the ICD-9 classification method. While NLP-based methods or manual retrieval of smoking status will provide the most accurate results, unavailability of free-text records should not deter investigators from using structured data to determine smoking status.

CONCLUSION
This study demonstrates that ICD-9 tobacco use codes are very specific, if not particularly sensitive, for identifying smoking status in a general clinic population. Importantly, absence of these codes does not indicate a lack of smoking, even in a high smoking prevalence population (e.g., lung cancer). Documentation procedures among different types of physicians and clinics appear to be similar. Temporal analysis reveals that transitions between current tobacco use and history of tobacco use codes correlate with behavior as recorded in patients’ medical records. Finally, use of these methods to determine smoking status for genetic studies does not appear to significantly alter regression outcomes or reduce power.

Acknowledgements. We would like to thank Robert Goodloe for scripting support for the association analyses and Drs Joshua Denny and Trent Rosenbloom for their insightful feedback.

Funding. This work was supported in part by the National Institute of General Medical Sciences Training grant T32GM080178, the National Cancer Institute grant R01CA141307, and by the National Human Genome Research Institute grant Medical Sciences Training grant T32GM080178, the National Cancer Institute grant R01CA141307, and by the National Human Genome Research Institute grant R01CA141307, and by the National Human Genome Research Institute grant R01CA141307, and by the National Human Genome Research Institute grant R01CA141307, and by the National Human Genome Research Institute grant R01CA141307.

Competing interests. None.

Provenance and peer review. Commissioned; externally peer reviewed.

REFERENCES