How Confounder Strength Can Affect Allocation of Resources in Electronic Health Records

by Kristine E. Lynch, PhD; Brian W. Whitcomb, PhD; and Scott L. DuVall, PhD

Abstract

When electronic health record (EHR) data are used, multiple approaches may be available for measuring the same variable, introducing potentially confounding factors. While additional information may be gleaned and residual confounding reduced through resource-intensive assessment methods such as natural language processing (NLP), whether the added benefits offset the added cost of the additional resources is not straightforward. We evaluated the implications of misclassification of a confounder when using EHRs. Using a combination of simulations and real data surrounding hospital readmission, we considered smoking as a potential confounder. We compared ICD-9 diagnostic code assignment, which is an easily available measure but has the possibility of substantial misclassification of smoking status, with NLP, a method of determining smoking status that more expensive and time-consuming than ICD-9 code assignment but has less potential for misclassification. Classification of smoking status with NLP consistently produced less residual confounding than the use of ICD-9 codes; however, when minimal confounding was present, differences between the approaches were small. When considerable confounding is present, investing in a superior measurement tool becomes advantageous.

Keywords: electronic health records; confounding; natural language processing

Introduction

Electronic health records (EHRs) are widely used sources of data for health outcomes research.[1, 2] They offer a large amount of rich, codified, longitudinal data, allowing for elevated statistical power and provide information on temporality when evaluating exposure-disease relationships. Unlike prospective studies, which utilize and/or develop specific tools to ascertain primary data on exposures, outcomes, and covariates, EHR and administrative data include structured data (e.g., diagnosis and procedure codes, laboratory tests, and pharmacy data) and unstructured data (e.g., clinical notes) that are not primarily collected for research purposes. Consideration of the appropriateness and ability of a data source to support different research questions should take into account undocumented data and/or misclassification errors in the data that are documented.[3]

The consequences of misclassifying exposure and outcome in epidemiologic studies are well understood and widely described.[4, 5] Discussion of the likelihood, direction, and magnitude of such biases related to exposure and outcome misclassification are ubiquitous in published epidemiologic literature. The consequence that misclassifying a confounding factor has on effect estimates, though long understood and described in previous literature,[6–11] may be less well appreciated. When a confounding variable is subject to misclassification, the ability to control for its effect is reduced because some level of

confounding (i.e., residual confounding) still remains.[12] The magnitude and direction of the bias induced through misclassification of a confounder is a function of the difference in the prevalence of the confounder between the exposed and unexposed groups, the strength of the association between the confounder and the outcome, and the sensitivity and specificity of the tool used to evaluate confounder status.[13, 14] While sensitivity and specificity are frequently taken into account with regard to exposure and outcome, this may not be similarly true for the assessment of confounding factors.[15]

When an EHR data source is used, there may be more than one approach available for measuring the same confounding variable within the same data source. This scenario provides the investigator an opportunity to evaluate different approaches to potentially reduce the degree of misclassification. Although some methods have been shown to be more valid than others, some degree of misclassification in all methods is inevitable because they rely on the completeness and correctness of the underlying data source.[16] Knowing that an exposure-outcome effect estimate can be biased if confounding factors are not adequately addressed, a researcher can face the quandary of choosing which method to use for assessment of a confounder that balances the concern for control of confounding with the need for efficient utilization of resources.

This paper addresses residual confounding due to misclassification by presenting two studies. First, a simulation study is used to explore the magnitude of bias that occurs when a confounding variable is measured using two different methods. Second, an example using real EHR data is provided in order to demonstrate the likely impact of residual confounding when these imperfect measures are used in practice. Together these studies provide a framework for decision making.

In both analyses, we consider smoking status as the confounder of interest. Smoking has been shown to be a risk factor for a myriad of health outcomes, so evaluating its role as a potential confounder is common.[17, 18] The best method to classify smoking status using EHR data is unclear. Methods for classifying smoking status have relied on *International Classification of Diseases, Ninth Revision* (ICD-9) diagnostic codes,[19–21] extraction of relevant text from clinical notes using natural language processing (NLP),[22, 23] customizable flags associated with patient records, [24–26] and survey/screening data.[27] Each method has strengths and limitations, and the resources needed to carry out each method vary considerably. For example, ICD-9 codes are typically stored as structured data in EHRs.[28] This method makes the data relatively easy to obtain and use, but relies on clinicians and medical coders to properly document and assign codes for the diagnosis of tobacco-use history. Moreover, because ICD-9 codes specifically denote the presence of a condition, there are no codes for confirming the absence of tobacco use. Instead, nonsmokers are often defined as patients without documentation of smoking-specific ICD-9 codes. As a result, patients who may smoke but were not asked by a physician or were not assigned a code for smoking will be misclassified as nonsmokers. NLP has the ability to classify smoking status on a more granular level than ICD-9 codes do, including distinguishing among nonsmokers, former smokers, current smokers, and those without documentation of smoking status.[29] Unlike the finite set of ICD-9 codes used to classify smoking, NLP systems need to have the ability to identity and distinguish between the wide variety of ways smoking can be documented in the medical notes (e.g., *nonsmoker*; *does the patient use tobacco products?*; *25 pack/year history*; *daily cigarettes*). Thus, development of an NLP system that can accurately extract smoking status requires significant resources, including specialized programming skills and subject-matter experts to validate system output. Reuse of an NLP system that is already developed may reduce the resources needed, but there still may be a need for validation and possible modification when it is used with a different study population.[30] For these reasons, NLP may not be feasible for researchers facing time, technological, or funding constraints.

Methods

Simulation Study to Assess the Effects of Misclassified Confounding

To illustrate the effects of misclassification of a confounder, we conducted a simulation study.[31] This simulation considered a dichotomous exposure, *x*; a dichotomous outcome, *y*; and smoking as a confounder of the exposure-outcome relationship. We used Monte Carlo methods to generate 100

synthetic data sets, each with 10,000 observations, with the data generation described as follows. In each data set, individuals were assigned true smoking status by sampling from a binomial distribution with prevalence of 45 percent. This prevalence was chosen to reflect smoking rates found in the veterans population.[32, 33] A binomial exposure was specified with a baseline probability of 10 percent and increased among smokers. A binomial outcome variable was similarly created, except with a baseline probability of 5 percent. Using 45 percent as the true smoking status proportion, we determined measured smoking status with misclassification according to values of sensitivity (32 percent, 78 percent) and specificity (100 percent, 88 percent) to represent the assessment of smoking data by ICD-9 codes and NLP, respectively. The values of these parameters were informed by measures previously reported in validation studies of ICD coding[34, 35] and NLP.[36] Each observation therefore contained data on exposure (yes, no), outcome (yes, no), ICD-9 smoking (yes, no), and NLP smoking (yes, no). The strength of confounding was varied according to the magnitude of the effect of smoking on exposure and outcome. Odds ratios describing the association of smoking with exposure and smoking with outcome were set at 1.5, 5.0, and 10.0 to denote weak, moderate, or strong confounding, respectively. Accordingly, nine confounding strengths were evaluated by this approach.

In order to evaluate the use of imperfectly classified smoking status to control for confounding by true smoking, we utilized logistic regression models in the simulated data. Using these logistic regression models, we derived estimates for the odds of the outcome among the exposed, compared with the unexposed, adjusting for imperfectly measured smoking using ICD-9 and NLP and, for comparison, adjusting for the actual smoking status. The latter scenario represented the true exposure-disease relationship and was thus used to calculate the amount of bias introduced by using imperfect assessments of smoking under each of the nine levels of confounding described earlier. For each scenario, bias was calculated using the average odds ratio estimates from the 100 data sets comparing models controlling for misclassified smoking status with those controlled for true smoking status.

For these analyses we assumed that exposure and disease were binary variables measured without error. All analyses, including the simulation, were conducted using SAS 9.4 (SAS Institute; Cary, NC).

Real-Data Motivating Example: Gender and Hospital Readmission Controlling for Smoking Status

To evaluate residual confounding when comparing multiple methods for confounder measurement in practice, we assessed the relationship of gender with hospital readmission risk among patients with chronic obstructive pulmonary disease (COPD) in the Department of Veterans Affairs (VA). The VA is the largest integrated healthcare delivery network in the United States, capturing healthcare information from patients across all medical specialties from both inpatient and outpatient settings with broad geographical coverage. The VA's Corporate Data Warehouse (CDW) is a nationwide repository with historical electronic medical records dating back to October 1, 1999, including patient encounters, pharmacy, lab/chemistry, microbiology, vital sign, and radiology-related data on more than 20 million veterans.[37] The data warehouse is dynamic, adding and updating data nightly and including more than 1 million text-based medical notes. The wealth of healthcare data in CDW is made available to VA-credentialed research teams, which affords VA researchers the opportunity to simultaneously control for many clinical, demographic, and lifestyle factors, making this data source an ideal setting for assessing the impact of residual confounding.

In order to evaluate predictors of hospital readmission, we identified patients with a COPD-related hospitalization between January 1, 2004, and July 1, 2014 (the index hospitalization). Of primary interest was the assessment of gender differences in the readmission risk. Patients who died during their index hospitalization stay were not at risk for readmission and were excluded from the study. Baseline eligibility was defined as having at least one clinical encounter, medication prescription or fill, or lab result at least 365 days before index hospitalization. The outcome of interest was 30-day hospital readmission, defined as having a subsequent inpatient stay within 30 days of the index hospitalization discharge date.

Evidence from published literature warranted consideration of smoking as a confounder of the association of interest in our cohort of COPD patients. Prior studies have suggested that smoking is

associated with lung disease–related hospital readmission[38, 39] and longer length of hospital stay.[40, 41] Moreover, men are more likely to be current or former smokers, compared to women.[42, 43] Under this scenario, we consider the decision to assess smoking via ICD-9 codes or through the development of an NLP system that abstracts smoking status from clinical notes.

We assessed the potential for smoking to confound the gender-readmission relationship using the conventional criteria for confounding:

1. smoking must be associated with gender (that is, odds ratio of exposure-confounder does not equal 1.0),
2. smoking is an independent risk factor for readmission (that is, odds ratio of outcome-confounder does not equal 1.0), and
3. the confounder cannot be an intermediate step in the causal pathway from gender to readmission.

Using logistic regression, we derived estimates for the odds of 30-day hospital readmission among men, compared to women, adjusting for smoking as classified by ICD-9 codes or as classified by NLP, and we assessed the differences between the two methods and with regard to the unadjusted estimate.

Results

Results of the Simulation Models

Results of the simulation models are presented in Table 1 and illustrated in Figure 1. The crude association between exposure and outcome ranged from 1.04 to 2.31, depending on the degree of confounding present. With adjustment for the true smoking status, confounding of the association between exposure and outcome was completely addressed, and the adjusted odds ratio (AOR) was correctly estimated as 1.0. As expected, in models adjusting for measured smoking with misclassification, the effect of exposure on outcome was consistently biased away from the null because of residual confounding, with bias varying by the degree of confounding present and the degree of misclassification (i.e., which assessment tool [ICD-9 or NLP] was used). Adjustment for smoking as measured by ICD-9 or NLP appeared minimally effective for controlling for smoking status, with greater bias for ICD-9 related to the greater degree of smoking status misclassification compared to NLP. The disparity between the two methods existed across all nine confounding scenarios, but with weak to moderate confounding, the difference between the two tools was negligible. Bias ranged from 0.02 to 0.17 for ICD-9 and from 0.02 to 0.12 for NLP. With moderate to strong confounding, bias ranged from 0.52 to 0.99 for ICD-9 and from 0.36 to 0.68 for NLP. See Table 1 for a complete list of confounding scenarios and the corresponding bias present after imperfect adjustment.

The leftmost side of Figure 1 depicts the weakest confounding scenario. After controlling for smoking via NLP, the effect of exposure on outcome is 1.02, whereas the effect of exposure on outcome controlling for ICD-9 smoking is 1.03. As shown on the far right-hand side of Figure 1, under the strongest confounding scenario, after controlling for smoking using NLP, the odds of the outcome were 1.68 times higher for the exposed than for the unexposed, whereas after controlling for smoking using ICD-9 codes, the odds of the outcome were 1.99 times higher for the exposed than for the unexposed.

Results of the Real-Data Example

In our data, 89,502 patients had an index COPD-related hospitalization, of which 3,046 patients (3.4 percent) were female. Overall, 11,977 patients (13.4 percent) were readmitted to the hospital within 30 days of their index hospitalization. The majority of patients were classified as smokers by both methods, with 59.1 percent (n = 52,881) using ICD-9 codes and 89.0 percent (n = 79,654) using NLP. No documentation of smoking status was found for 5.7 percent (n = 5,070) of patients using NLP. The two methods provided concordant classification for 63 percent of the remaining 84,432 patients when both methods were assessed. Specifically, 49,809 patients (59.0 percent) were defined as smokers by both methods, and 3,628 patients (4.3 percent) were defined as nonsmokers by both methods. The most common discordant scenario occurred for patients defined as nonsmokers according to ICD-9 codes but

as smokers according to NLP (*n* = 29,845, 35.4 percent), whereas only 1,150 (1.3 percent) of the patients were classified as smokers according to ICD-9 codes and as nonsmokers according to NLP.

In statistical analyses evaluating the potential of smoking to confound the relationship of interest (see Table 2), there were slight associations between smoking (NLP and ICD-9) and both exposure and outcome, suggesting that smoking is. at most, a weak confounder. Using ICD-9 codes, smokers had a slightly decreased risk of readmission compared to nonsmokers (odds ratio [OR] = 0.94, 95 percent confidence interval [CI] = 0.90–0.98), and males were 30 percent less likely to be smokers than females were (OR = 0.68, 95 percent CI = 0.63–0.74). In line with previous literature, when using NLP, smokers were more likely to be readmitted, and males were more likely to be smokers, although neither finding reached statistical significance (OR = 1.04, 95 percent CI = 0.96–1.14; OR = 1.08, 95 percent CI = 0.92–1.26, respectively).

In unadjusted analyses, males were 1.40 times as likely as females to be readmitted to the hospital within 30 days of discharge (95 percent CI = 1.23–1.57); see Table 3. After adjusting for misclassified smoking status with NLP and then again with ICD-9 codes, none of the observed association between gender and readmission was explained by smoking status. Smoking did not appear to be a confounder or not a strong enough confounder to explain the observed association. We can assume that both ICD-9 and NLP to some extent misclassify smoking status, and thus the odds ratio of exposure-confounder and odds ratio of outcome-confounder calculated to evaluate criterion 1 and 2 from above, are biased.

Discussion

The simulation results illustrate circumstances in which substantial reductions in bias can be achieved by choosing NLP over ICD-9 codes when measuring smoking status as a confounder. Only when the variable of interest considerably confounds the exposure-outcome relationship do differences between measurement methods in regard to control for confounding become apparent. Notably, in our example of VA data, although prior literature suggested that controlling for smoking is important, it did not appear to be the case in our large data set. Smoking very weakly confounded the relationship between gender and 30-day hospital readmission. As a result, ICD-9 codes, although a less valid measure of smoking status than NLP, had a similar impact on the effect estimate, suggesting that directing additional resources to NLP in this specific circumstance would provide no added benefit to the study. In a circumstance where there is strong confounding, however, NLP may be beneficial.

This study used values of sensitivity and specificity that were reported in previous literature on the validity of ICD-9 codes[44, 45] and NLP[46] for smoking status. It should be noted that these values could differ from cohort to cohort. The validity of ICD-9 codes is influenced by the accuracy of the medical notes. Similarly, the sensitivity and specificity of an NLP algorithm are functions of both the recall and precision of the system itself and the completeness and accuracy of the medical notes. The accuracy of the medical notes is affected by patients' recall of their smoking status and clinicians recording of patients' response. For example, one study found that clinical notes documented smoking status more frequently for patients with diseases in which symptoms are exacerbated by smoking exposure or for patients with conditions that contraindicate smoking.[47] Another study used urinary cotinine concentrations to validate self-reported smoking status among patients with pulmonary disease and found that patients with pulmonary disease are likely to misreport smoking status.[48] Therefore, while COPD patients may have smoking history recorded in medical notes, taking into account the potential inaccuracies documented in the notes may mean that the overall sensitivity and specificity may be lower than the system in our study validated against the notes themselves. These considerations are true also for variables that can be captured in administrative databases, where varying levels of sensitivity and specificity can be established and investigators may be interested in using more expensive and methodologically involved assessment methods.

We used ICD-9 codes and NLP to assess smoking status as one example when an investigator has at least two options for a measurement approach in administrative databases, with one measured variable thought to have less misclassification than the other variable(s). Other examples are plausible within and outside of the COPD domain. For example, in a study seeking to adjust for respiratory morbidities, COPD can be defined by an ICD-9 code or by a combination of ICD-9 codes and pharmacy data among

spirometry-tested patients.[49] Similarly, rheumatoid arthritis can be defined as the presence of an ICD-9 code or a positive rheumatoid factor and prescription for a disease-modifying antirheumatic drug.[50] While these examples do not involve NLP, they do involve methods beyond standard ICD-9 code extraction, which may unnecessarily deplete study resources if they are used to evaluate confounder status when confounding is weak or may greatly benefit study results if confounding is substantial.

Consideration of a potential confounder for inclusion in an analysis typically begins with an assessment of prior knowledge of the variable as a potential confounder and its relationship with the exposure and outcome, followed by formal statistical assessment.[51, 52] Published research can help investigators gain knowledge of potential confounders and anticipate the magnitude and direction of the impact that such factors might have on their results, but direct extrapolation is discouraged.[53] Our simulation study showed that when relationships between smoking and outcome and smoking and exposure are of substantial magnitude, differences between assessment tools are more likely to become apparent. Utilizing available data may help inform decisions regarding when use of additional resources for better measured proxies of confounders is merited. Consideration of the potential strength of confounding is an important step prior to determining the most appropriate measurement method.

Conclusion

To draw appropriate conclusions regarding a study hypotheses, it is important to consider the potential impact of misclassification of confounding variables, in addition to that of the exposure(s) and outcome of interest. Because investigators must rely on measured variables subject to error, this paper serves as a guide for research using similar administrative databases. For researchers with the resources and means to search clinical notes, or employ a method beyond surveying simple structured data, we recommend formally assessing the magnitude of confounding most likely to be present before investing in a resource-intensive method.

Kristine E. Lynch, PhD, is a research associate at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Brian W. Whitcomb, PhD, is an associate professor at the University of Massachusetts Amherst in Amherst, MA.

Scott L. DuVall, PhD, is an assistant professor at the VA Salt Lake City Health Care System in Salt Lake City, UT.

Notes

1. Dean, B. B., J. Lam, J. L. Natoli, Q. Butler, D. Aguilar, and R. J. Nordyke. "Review: Use of Electronic Medical Records for Health Outcomes Research: A Literature Review." *Medical Care Research and Review* 66, no. 6 (2009): 611–38.
2. Lau, E. C., F. S. Mowat, M. A. Kelsh, J. C. Legg, N. M. Engel-Nitz, H. N. Watson, et al. "Use of Electronic Medical Records (EMR) for Oncology Outcomes Research: Assessing the Comparability of EMR Information to Patient Registry and Health Claims Data." *Clinical Epidemiology* 3 (2011): 259–72.
3. Hersh, W. R., M. G. Weiner, P. J. Embi, J. R. Logan, P. R. Payne, E. V. Bernstam, et al. "Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research." *Medical Care* 51, no. 8, supp. 3 (2013): S30–S37.
4. Flegal, K. M., P. M. Keyl, and F. J. Nieto. "Differential Misclassification Arising from Nondifferential Errors in Exposure Measurement." *American Journal of Epidemiology* 134, no. 10 (1991): 1233–44.
5. Greenland, S., and P. Gustafson. "Accounting for Independent Nondifferential Misclassification Does Not Increase Certainty That an Observed Association Is in the Correct Direction." *American Journal of Epidemiology* 2006;164(1): 63–68.
6. Fewell, Z., G. Davey Smith, and J. A. Sterne. "The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study." *American Journal of Epidemiology* 166, no. 6 (2007): 646–55.
7. Greenland, S. "The Effect of Misclassification in the Presence of Covariates." *American Journal of Epidemiology* 112, no. 4 (1980): 564–69.
8. Marshall, J. R., and J. L. Hastrup. "Mismeasurement and the Resonance of Strong Confounders: Uncorrelated Errors." *American Journal of Epidemiology* 143, no. 10 (1996): 1069–78.
9. Marshall, J. R., J. L. Hastrup, and J. S. Ross. "Mismeasurement and the Resonance of Strong Confounders: Correlated Errors." *American Journal of Epidemiology* 150, no. 1 (1999): 88–96.
10. Savitz, D. A. *Interpreting Epidemiologic Evidence*. New York, NY: Oxford University Press, 2003.
11. Savitz, D. A., and A. E. Baron. "Estimating and Correcting for Confounder Misclassification." *American Journal of Epidemiology* 129, no. 5 (1989): 1062–71.
12. Szklo, M., and F. J. Nieto. *Epidemiology beyond the Basics*. Burlington, MA: Jones & Bartlett Learning, 2014.
13. Greenland, S. "The Effect of Misclassification in the Presence of Covariates."
14. Budtz-Jorgensen, E., N. Keiding, P. Grandjean, P. Weihe, and R. F. White. "Consequences of Exposure Measurement Error for Confounder Identification in Environmental Epidemiology." *Statistics in Medicine* 22, no. 19 (2003): 3089–3100.
15. Fewell, Z., G. Davey Smith, and J. A. Sterne. "The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study."
16. Hersh, W. R., M. G. Weiner, P. J. Embi, J. R. Logan, P. R. Payne, E. V. Bernstam, et al. "Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research."
17. An, R. "Health Care Expenses in Relation to Obesity and Smoking among U.S. Adults by Gender, Race/Ethnicity, and Age Group: 1998–2011." *Public Health* 129, no. 1 (2015): 29–36.

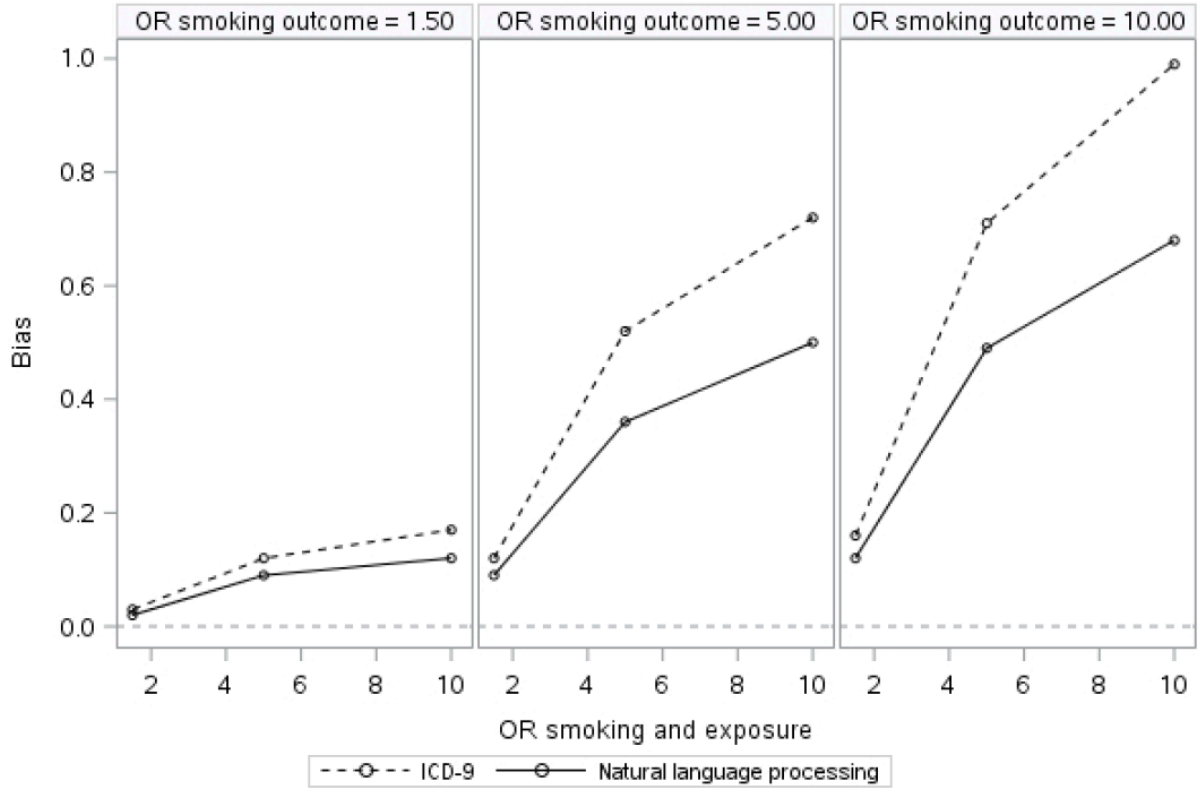
18. Centers for Disease Control and Prevention (CDC). "Smoking-Attributable Mortality, Years of Potential Life Lost, and Productivity Losses—United States, 2000–2004." *Morbidity and Mortality Weekly Report* 57, no. 45 (2008): 1226–28.
19. Eapen, Z. J., L. Liang, J. H. Shubrook, M. A. Bauman, V. J. Bufalino, D. L. Bhatt, et al. "Current Quality of Cardiovascular Prevention for Million Hearts: An Analysis of 147,038 Outpatients from The Guideline Advantage." *American Heart Journal* 168, no. 3 (2014): 398–404.
20. Kim, H. M., E. G. Smith, C. M. Stano, D. Ganoczy, K. Zivin, H. Walters, and M. Valenstein. "Validation of Key Behaviourally Based Mental Health Diagnoses in Administrative Data: Suicide Attempt, Alcohol Abuse, Illicit Drug Abuse and Tobacco Use." *BMC Health Services Research* 12 (2012): 18.
21. Thompson, W. H., and S. St-Hilaire. "Prevalence of Chronic Obstructive Pulmonary Disease and Tobacco Use in Veterans at Boise Veterans Affairs Medical Center." *Respiratory Care* 55, no. 5 (2010): 555–60.
22. De Silva, L., T. Ginter, T. Forbush, N. Nokes, B. Fay, T. Mikuls, et al. "Extraction and Quantification of Pack-years and Classification of Smoker Information in Semi-structured Medical Records." Paper presented at the 28th International Conference on Machine Learning, Bellevue, WA, 2011.
23. Xu, H., M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, et al. "Validating Drug Repurposing Signals Using Electronic Health Records: A Case Study of Metformin Associated with Reduced Cancer Mortality." *Journal of the American Medical Informatics Association* 22, no. 1 (2015): 179–91.
24. Thompson, W. H., and S. St-Hilaire. "Prevalence of Chronic Obstructive Pulmonary Disease and Tobacco Use in Veterans at Boise Veterans Affairs Medical Center."
25. Huetsch, J. C., J. E. Uman, E. M. Udris, and D. H. Au. "Predictors of Adherence to Inhaled Medications among Veterans with COPD." *Journal of General Internal Medicine* 27, no. 11 (2012): 1506–12.
26. McGinnis, K. A., C. A. Brandt, M. Skanderson, A. C. Justice, S. Shahrir, A. A. Butt, et al. "Validating Smoking Data from the Veteran's Affairs Health Factors Dataset, an Electronic Data Source." *Nicotine and Tobacco Research* 13, no. 12 (2011): 1233–39.
27. Kruse, G. R., and N. A. Rigotti. "Routine Screening of Hospital Patients for Secondhand Tobacco Smoke Exposure: A Feasibility Study." *Preventive Medicine* 69 (2014): 141–45.
28. O'Malley, K. J., K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton. "Measuring Diagnoses: ICD Code Accuracy." *Health Services Research* 40, no. 5, pt. 2 (2005): 1620–39.
29. Uzuner, O., I. Goldstein, Y. Luo, and I. Kohane. "Identifying Patient Smoking Status from Medical Discharge Records." *Journal of the American Medical Informatics Association* 15, no. 1 (2008): 14–25.
30. Chapman, W. W., P. M. Nadkarni, L. Hirschman, L. W. D'Avolio, G. K. Savova, and O. Uzuner. "Overcoming Barriers to NLP for Clinical Text: The Role of Shared Tasks and the Need for Additional Creative Solutions." *Journal of the American Medical Informatics Association* 18, no. 5 (2011): 540–43.
31. Burton, A., D. G. Altman, P. Royston, and R. L. Holder. "The Design of Simulation Studies in Medical Statistics." *Statistics in Medicine* 25 (2006): 4279–92.
32. Xu, H., M. C. Aldrich, Q. Chen, H. Liu, N. B. Peterson, Q. Dai, et al. "Validating Drug Repurposing Signals Using Electronic Health Records: A Case Study of Metformin Associated with Reduced Cancer Mortality."
33. Brown, D. W. "Smoking Prevalence among US Veterans." *Journal of General Internal Medicine* 25, no. 2 (2010): 147–49.

34. Kim, H. M., E. G. Smith, C. M. Stano, D. Ganoczy, K. Zivin, H. Walters, and M. Valenstein. “Validation of Key Behaviourally Based Mental Health Diagnoses in Administrative Data: Suicide Attempt, Alcohol Abuse, Illicit Drug Abuse and Tobacco Use.”
35. Wiley, L. K., A. Shah, H. Xu, and W. S. Bush. “ICD-9 Tobacco Use Codes Are Effective Identifiers of Smoking Status.” *Journal of the American Medical Informatics Association* 20, no. 4 (2013): 652–58.
36. Ibid.
37. Fihn, S. D., J. Francis, C. Clancy, C. Nielson, K. Nelson, J. Rumsfeld, et al. “Insights from Advanced Analytics at the Veterans Health Administration.” *Health Affairs* 33, no. 7 (2014): 1203–11.
38. Hunter, L. C., R. J. Lee, I. Butcher, C. J. Weir, C. M. Fischbacher, D. McAllister, et al. “Patient Characteristics Associated with Risk of First Hospital Admission and Readmission for Acute Exacerbation of Chronic Obstructive Pulmonary Disease (COPD) Following Primary Care COPD Diagnosis: A Cohort Study Using Linked Electronic Patient Records.” *BMJ Open* 6, no. 1 (2016): e009121.
39. Ogawa, F., Y. Satoh, A. Iyoda, H. Amano, Y. Kumagai, and M. Majima. “Clinical Impact of Lung Age on Postoperative Readmission in Non–Small Cell Lung Cancer.” *Journal of Surgical Research* 193, no. 1 (2015): 442–48.
40. Haapanen-Niemi, N., S. Miilunpalo, I. Vuori, M. Pasanen, and P. Oja. “The Impact of Smoking, Alcohol Consumption, and Physical Activity on Use of Hospital Services.” *American Journal of Public Health* 89, no. 5 (1999): 691–98.
41. Iversen, L., S. Fielding, and P. C. Hannaford. “Smoking in Young Women in Scotland and Future Burden of Hospital Admission and Death: A Nested Cohort Study.” *British Journal of General Practice* 63, no. 613 (2013): e523–e533.
42. Brown, D. W. “Smoking Prevalence among US Veterans.”
43. Curry, J. F., N. Aubuchon-Endsley, M. Brancu, J. J. Runnals, and J. A. Fairbank. “Lifetime Major Depression and Comorbid Disorders among Current-Era Women Veterans.” *Journal of Affective Disorders* 152–154 (2014): 434–40.
44. Kim, H. M., E. G. Smith, C. M. Stano, D. Ganoczy, K. Zivin, H. Walters, and M. Valenstein. “Validation of Key Behaviourally Based Mental Health Diagnoses in Administrative Data: Suicide Attempt, Alcohol Abuse, Illicit Drug Abuse and Tobacco Use.”
45. Wiley, L. K., A. Shah, H. Xu, and W. S. Bush. “ICD-9 Tobacco Use Codes Are Effective Identifiers of Smoking Status.”
46. Ibid.
47. Silfen, S. L., J. Cha, J. J. Wang, T. G. Land, and S. C. Shih. “Patient Characteristics Associated with Smoking Cessation Interventions and Quit Attempt Rates across 10 Community Health Centers with Electronic Health Records.” *American Journal of Public Health* 105, no. 10 (2015): 2143–49.
48. Stelmach, R., F. L. Fernandes, R. M. Carvalho-Pinto, R. A. Athanazio, S. Z. Rached, G. F. Prado, and A. Cukier. “Comparison between Objective Measures of Smoking and Self-reported Smoking Status in Patients with Asthma or COPD: Are Our Patients Telling Us the Truth?” *Brazilian Journal of Pulmonology* 41, no. 2 (2015): 124–32.
49. Cooke, C. R., M. J. Joo, S. M. Anderson, T. A. Lee, E. M. Udris, E. Johnson, and D. H. Au. “The Validity of Using ICD-9 Codes and Pharmacy Records to Identify Patients with Chronic Obstructive Pulmonary Disease.” *BMC Health Services Research* 11 (2011): 37.
50. Singh, J. A., A. R. Holmgren, and S. Noorbaloochi. “Accuracy of Veterans Administration Databases for a Diagnosis of Rheumatoid Arthritis.” *Arthritis and Rheumatism* 51, no. 6 (2004): 952–57.

51. Evans, D., B. Chaix, T. Lobbedez, C. Verger, and A. Flahault. “Combining Directed Acyclic Graphs and the Change-in-Estimate Procedure as a Novel Approach to Adjustment-Variable Selection in Epidemiology.” *BMC Medical Research Methodology* 12 (2012): 156.
52. Weng, H. Y., Y. H. Hsueh, L. L. Messam, and I. Hertz-Picciotto. “Methods of Covariate Selection: Directed Acyclic Graphs and the Change-in-Estimate Procedure.” *American Journal of Epidemiology* 169, no. 10 (2009): 1182–90.
53. Savitz, D. A. *Interpreting Epidemiologic Evidence*.

Figure 1

Residual Confounding Due to Imperfect Measurement of Smoking across Different Degrees of Confounding



Abbreviations: ICD-9, *International Classification of Diseases, Ninth Revision*; OR, odds ratio.

Table 1

Simulation Results of the Quantity of Residual Confounding across Different Degrees of Confounding Assuming the True Association between Exposure and Outcome Is Null (Odds Ratio = 1.0)

True Smoke-Exposure OR	True Smoke-Outcome OR	Crude OR	NLP AOR	ICD-9 AOR	ICD-9 Smoke-Outcome OR	ICD-9 Smoke-Exposure OR
1.50	1.50	1.04	1.02	1.03	1.29	1.31
	5.00	1.15	1.09	1.12	1.31	3.09
	10.00	1.20	1.12	1.16	1.32	5.43
5.00	1.50	1.16	1.09	1.12	3.12	1.31
	5.00	1.67	1.36	1.52	3.10	3.10
	10.00	1.93	1.49	1.71	3.09	5.57
10.00	1.50	1.20	1.12	1.17	5.42	1.30
	5.00	1.93	1.50	1.72	5.53	3.09
	10.00	2.31	1.68	1.99	5.46	5.44

Abbreviations: AOR, adjusted odds ratio; ICD-9, *International Classification of Diseases, Ninth Revision*; NLP, natural language processing; OR, odds ratio.

Table 2

Misclassified Estimates of the Effect of Smoking on Exposure and Outcome Using ICD-9 Codes and Natural Language Processing

Estimates	OR	95% CI
ICD-9 and readmission (outcome), Yes vs. No	0.94	0.90–0.98
ICD-9 and gender (exposure), Male vs. Female	0.68	0.63–0.74
NLP and readmission (outcome)	1.04	0.96–1.14
NLP and gender (exposure)	1.08	0.92–1.26

Abbreviations: CI, confidence interval; ICD-9, *International Classification of Diseases, Ninth Revision*; NLP, natural language processing; OR, odds ratio.

Table 3

Logistic Regression Results of the Effect of Gender on 30-Day Hospital Readmission

Model	OR	95% CI
Unadjusted	1.40	1.23–1.57
ICD-9 adjusted	1.40	1.23–1.57
NLP adjusted	1.39	1.23–1.58

Abbreviations: CI, confidence interval; ICD-9, *International Classification of Diseases, Ninth Revision*; NLP, natural language processing; OR, odds ratio.