NLP-derived information improves the estimates of risk of disease compared to estimates based on manually extracted data alone.

Fiona M Callaghan PhD
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
fiona.callaghan@nih.gov

Matthew T Jackson PhD
Food and Drug Administration (CDER/OTS/OB/DBVI), White Oak, MD, USA
matthew.jackson@fda.hhs.gov

Dina Demner-Fushman MD PhD, Swapna Abhyankar MD, and Clement J McDonald MD
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
ddemner@mail.nih.gov, swapna.abhyankar@nih.gov, ClemMcDonald@mail.nih.gov

Abstract

Natural language processing (NLP) enables researchers to extract large quantities of information from free-text that otherwise could only be extracted manually. This information can then be used to answer clinical research questions via statistical analysis. However, NLP extracts information with some degree of error – the sensitivity and specificity of state-of-the-art NLP methods are typically 80-90% – and most statistical methods assume that the information has been observed "without measurement error". As we show in this paper, if an NLP-derived smoking status predictor is used, for example, to estimate the risk of smoking-related cancer without any adjustment for measurement error, the estimate is biased. Conversely, if a smaller subset of manually extracted data is used alone, then the estimate is unbiased, but imprecise, and the corresponding inference methods tend to have low power to detect significant relationships. We propose using a statistical measurement error method – a maximum likelihood (ML) method – that combines information from NLP with manually validated data to produce unbiased estimates that also have good power to detect a significant signal. This method has the potential to open-up large free-text databases to statistical analysis for clinical research. With a case study using smoking status to predict smoking-related cancer and simulations, we demonstrate that the ML method performs better under a variety of scenarios than using either NLP or manually extracted data alone.

1 Introduction

Free-text fields are common in medical databases, for example clinical narratives (such as discharge summaries and progress notes) constitute approximately 10% of the fields in the database in our study. The notes often contain valuable information that may not be captured anywhere else in the structured part of the database and which may be essential to answering a research question. Traditionally, medical abstractors manually extract variables from unstructured text, which is a time- and labor-intensive process, and prone to subjectivity. Alternatively, natural language processing (NLP) methods can assist abstractors and greatly improve their efficiency, even replacing them in some cases. For example, the SHARPn project combines normalized NLP-derived observations with structured data for high-throughput phenotyping (Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG, 2012). NLP applications are successful in many tasks, for example, assisting medical coding (Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M, 2010), detecting complications (Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T, 2011; Wang X, Hripcsak G, Markatou M, Friedman C, 2009), and automatically classifying clinical records (Wilcox AB, Hripcsak G, 2003). The state-of-the-art performance of NLP applications ranges from high 80s to high 90s for both recall (sensitivity) and precision (positive predictive value). Specificity is rarely used for NLP tasks due

to the fact that it is practically impossible to enumerate true negatives. However, it can be used for classification problems because in such cases, it is reasonable to expect that we can enumerate true negatives. The levels of recall and precision of current NLP tools are quite satisfactory for many practical purposes. Due to the seemingly infinite number of ways clinicians describe clinical events, attempting to increase the accuracy measures above 96 or 97% is possible, but faces the problem of diminishing returns.

NLP extracts information with a degree of error, and this poses a problem to researchers who wish to use NLP-derived information as a predictor in a statistical model, as almost all statistical methods require that predictors are measured without error. If a predictor that has been measured with error is used, for example, in a regression model without adjusting for the measurement error, the estimates of the outcome suffer from the "triple whammy" of measurement error: the estimates of the outcome are prone to bias, the associated statistical hypothesis tests often suffer from lack of power, and important relationships between the predictors and the outcome are often obscured by the noise of the measurement error (Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM, 2006).

Given the challenges of incorporating information from free-text into a statistical analysis, we are left with two possible sources of information: information extracted via NLP and manually extracted data. Our previous research has shown, and we demonstrate again in this study, that if NLP-derived predictors are used in a statistical model without adjusting for measurement error, the estimates of the outcome are subject to substantial bias, even when NLP delivers a sensitivity and specificity of 90%. For instance, when estimating the odds ratio (OR) risk of smoking-related cancer for people who smoke versus those who do not, the estimated increased of risk of cancer when using NLP-derived predictor of smoking status is between 20-50% less than the true risk, depending on the level of specificity, sensitivity, sample size and other factors (Callaghan FM, Jackson MT, Demner-Fushman D, Abhyankar S, McDonald CJ, 2012).

According to our simulations, it appears that, in almost all cases, neither manually extracted data alone nor NLP data alone produce estimates of risk that are both unbiased and yet powerful enough to detect significant relationships between the predictor and the outcome. Therefore, it would be useful to have a method that can leverage both the "accuracy" of the manually validated data and the power associated with the large sample size of the NLP-derived data. Fortunately, there are statistical methods for handling predictors that are measured with error, such as the NLP-derived smoking status predictor, and adjusting for that error. Our hypothesis was that if these methods were adapted to NLP-derived information and combined with manually extracted data, we could produce less biased estimates of risk of disease, and more powerful test procedures (i.e. tests that detect true differences more often). We propose a validation-adjusted NLP maximum likelihood (ML) method, new to the NLP literature, to control for misclassification rates in the NLP-derived predictor, and we illustrate the use of the ML method by using the NLP-derived predictor of smoking status (smoker/non-smoker) to predict smoking-related cancer risk (smoking-related cancer/no smoking-related cancer). Using this method, we estimate that the risk of smoking-related cancers for smokers compared to non-smokers. We also demonstrate in various simulation scenarios that the ML method performs better than methods based on manually extracted data alone or NLP-derived information alone, under a range of sensitivities and specificities for the NLP-derived predictor.

2 Methods

Our overall hypothesis is that using NLP-derived information via the ML method produces estimates of the outcome that are superior to estimates based on the manually extracted subset of the data alone. We formulated a validation-adjusted NLP maximum likelihood model to address the problem of misclassification in the predictor, i.e. subjects classified as being smokers by NLP when in fact they are non-smokers, and vice versa.

2.1 ML method

When a discrete predictor is measured with error, the problem is referred to in the statistical literature as misclassification. The effects of misclassification on

estimating the risk have been explored by a number of researchers (Gustafson P, 2004; Buonaccorsi JP, Laake P, Veirød M, 2005) and several methods have been proposed to address these problems (Cook JR, Stefanski LA, 1994; Kuchenhoff H, Mwalili SM, Lesaffre E, 2006; Carroll RJ, Stefanski LA, 1990; Gleser LJ, 1990; Stefanski LA, Buzas JS, 1995; Buzas JS, Stefanski LA, 1996; Stefanski LA, Carroll RJ, 1987). Maximum likelihood (ML) methods are a natural fit for misclassification of a binary predictor of a binary outcome, because the problem can be couched in terms of a series of relatively simple binomial probabilities relating the outcome Y to the "true" predictor, X, and the true predictor to the NLP-derived predictor. While some recent progress has been made in the field of political science to account for misclassification error in text-based document categorization (Hopkins D, King G, 2010; Benoit K, Laver M, Mikhaylov S, 2009; Grimmer J, Stewart BM, 2012), to our knowledge, modern statistical misclassification methods have not been applied to NLP-derived variables in order to predict estimates of risk for clinical research.

The ML method uses NLP-derived values of the predictor variable, W[1] (in our example this is 1 if NLP identifies that the subject is a smoker, and 0 otherwise), the outcome of interest Y (smoking-related cancer, yes/no), and a small subset is randomly selected to act as a validation sample. The "true" values of the predictor variable (X) for the validation sample are abstracted based on manual review of the free-text notes. The main purpose of the validation sample is to enable estimation of the relationship between the true predictor (smoking status) and the NLP-determined value of the predictor.

The primary quantity of interest is the odds ratio (OR) of the outcome. This is a common quantity of interest in epidemiological studies that measures the extra risk of having the outcome for subjects with a risk factor compared to those without. In our example, the OR represents the extra risk of having smoking-related cancer for smokers compared to non-smokers.

For the ML method, the user has to supply the following variables in order to estimate the OR: 1) the overall proportion of subjects who are positive for the predictor (for example, the proportion of smokers); 2) the number of patients whose free-text reports were manually validated; 3) the overall sample size (N); and 4) the number of subjects broken down by outcome (yes/no), predictor (yes/no), and NLP predictor (yes/no) for the validation sample, and by outcome (yes/no) and NLP predictor (yes/no) for the non-validation sample.

We performed several simulations in order to compare the performance of the ML method to the estimates based on manually-validated data alone. We also included the estimates based on NLP data alone. The non-ML estimates were obtained using the Woolf large sample method (Woolf B, 1955). We designed the simulations to investigate how the estimates of the risk change with variations in NLP sensitivity and specificity, the size of the validation sample, and the magnitude of the odds ratio. The simulations were repeated for two values of the odds ratio representing small increased risk (OR=1.2 or 20% increased risk) and large increased risk (OR=2 or 100% increased risk), and three values of sensitivity and specificity (0.6, 0.8, and 0.9). The quantities that were fixed were the proportion of smokers (20%), the overall sample size ($N = 20,000$), the size of the validation sample (5% or $n_v = 1000$), and the baseline proportion of smoking-related cancers among the non-smokers (5%). The fixed values for sample size and prevalence of smoking-related cancers among non-smokers were based on the estimates from our study data. The proportion of smokers in the case study was approximately 30%, but we chose a value of 20% for the simulations to reflect the actual prevalence of smokers in the US, which was 19.3% in 2010 (Centers for Disease Control and Prevention, 2011).

A key assumption of our ML model (but not of ML models in general) is non-differential measurement error: once the true value of the predictor is known, then the NLP estimate of that predictor is assumed to not contain any extra information about the outcome. Non-differential measurement error is a common assumption among measurement error models and is a plausible assumption for our case study. In this setting, the non-differential measurement error assumption means that once a given patient's true smoking status is

[1] Following the notation of (Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM, 2006)

known, their NLP-derived smoking status is irrelevant for predicting their risk of developing smoking-related cancer. Further details about the ML method, as well as the R macro used to fit the model and estimate the parameters, are available from the first author.

2.2 NLP methods

We tested the method with a case study that used NLP-derived patient smoking status to predict the risk of developing smoking-related cancer. We applied rule-based NLP methods to the free-text hospital discharge summaries to extract each patient's smoking status. Our rule-based smoking extraction was based on the i2b2 observation that discharge summaries express smoking status using a limited number of textual features (e.g., "smok", "tobac", "cigar") in the Social History section (Uzuner O, Goldstein I, Luo Y, Kohane I, 2008). We first manually reviewed a small set of discharge summaries to put together a dictionary of smoking-related terms. We included terms that indicated positive smoking status, such as "smoker" and "pack-years" as well as those that indicated negative smoking status, such as "denies smoking" and "no history of tobacco". Our data dictionary had a total of 82 positive and 20 negative terms. We also found two pseudo-positive patterns: "smoker in the household" and "smoking crack". We then used regular expressions containing these terms to search the entire corpus of discharge summaries in order to assign a smoking status to each patient. We initially defined a smoker as someone who currently smoked or had a history of smoking in the past, a non-smoker as someone with specific information about having no current or past smoking history documented in the note, and an unknown as someone who had no documentation about smoking, either positive or negative. We made the assumption that the subjects with unknown smoking status as determined by NLP were non-smokers because the absence of their smoking status in the narrative likely implied that smoking was not an issue for that patient. For this reason and for practical purposes, we created a binary variable (smoker versus non-smoker) by including the patients with unknown smoking status in the non-smoker category. This conversion allowed us to use sensitivity and specificity as our measures of accuracy.

2.3 Data

The case study was based on information extracted from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-II) database (Saeed M, Lieu C, Raber G, Mark RG, 2002), which is maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT). MIMIC-II contains de-identified data from patients hospitalized in the intensive care unit (ICU) at Beth Israel Deaconess Medical Center from 2001 to 2008. The database includes clinical information in both structured and unstructured formats. Structured data include patients' discharge ICD-9 codes. Unstructured data include physician narrative discharge summaries containing a wealth of information, including the patient's smoking status. There were a total of 18207 subjects in the database, and 739 of those had their smoking status manually validated.

We used data from the National Cancer Institute to define smoking-related cancers (National Cancer Institute, National Institutes of Health, 2012): lung, esophagus, larynx, floor of mouth, mouth (other), oropharynx, hypopharynx, kidney, bladder, pancreas, stomach, cervix, and acute myeloid leukemia. We classified patients having any of these codes in their list of discharge diagnoses as having smoking-related cancer.

3 Results

3.1 Case Study

What is the added value of incorporating the NLP information (W), once we have information on the cancer-status Y and the subset of the predictor that has been validated X, smoking-status? It may be reasonable to think that all the useful information about the risk is contained in the variables that we know "without error" (Y and X), and adding the NLP-derived information only adds "noise" to our estimates. However, this is not the case: the NLP information greatly improves the accuracy of the estimates (see Figure 1). When we calculated the estimates of risks using only the 739 validation sample values (that we know without error), the estimate of the odds ratio was 2.67 (very similar to the ML estimate of 2.65). This means that under either method smokers are estimated to have approximately 2.7 times the risk of having smoking-related

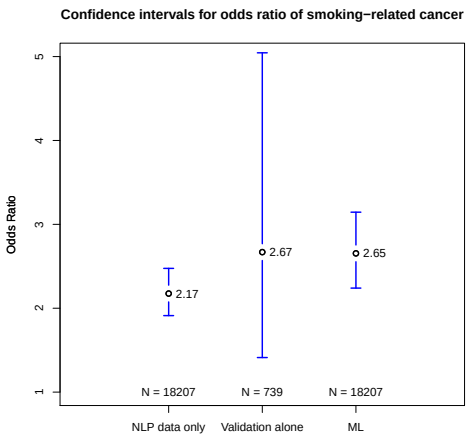

Figure 1: Results from the analysis where smoking status was used to predict smoking-related cancer. We compare 3 different methods to predict smoking-related cancer: 1) NLP-derived smoking status only, 2) smoking status taken from the validation sample only, and 3) the ML method that use the validation and NLP information. The non-ML estimates are estimated using Woolf's method (Woolf B, 1955).

cancer compared to non-smokers.

However, the confidence interval based on the validation sample ranges from 1.4 to 5.0, which is substantially wider than the confidence interval generated using the ML method (95% CI 2.2 to 3.1). In other words, based only on the manually extracted data, the range of plausible values for risk of smoking-related cancer for smokers is 1.4 to 5 times the risk of non-smokers, whereas the ML method estimates the additional risk of smoking to be between 2.2 to 3.1 times the risk of a non-smoker. Consequently, we can conclude that the NLP-derived variable contains information that is essential to incorporate into the estimate. Without it, we lose power and the resulting confidence interval is very wide. Although we may get a "good" (unbiased) estimate for the odds ratio itself, if we were to only use the subset of the data that has been manually-validated, we may lose significance and accuracy. Note that, based on the validation sample, the sensitivity and specificity for the NLP smoking predictor were good (0.84 and 0.95, respectively), and similar to values reported in the literature.

3.2 Simulations

The results for the simulations are given in Table 1. In general, the estimate of the odds ratio ($\widehat{\text{OR}}$) based on the validation sample was similar to the ML estimate in most cases. For instance, when the sensitivity and specificity are high (0.9) and the true odds ratio is 2, the mean estimate of the OR across 1000 replications using the ML method, is 2.00. The corresponding mean estimate using only the data from the validation sample is 2.05. The mean estimate of the odds ratio based on NLP data alone is the most biased (1.64) and this is the case in all the simulation scenarios. Based on the goal of minimizing bias alone, there is little difference between the ML and validation-data estimates. However, the standard deviation of the estimate of the odds ratio or "standard error" (SE) is much higher for the validation estimates than for ML (0.30 versus 0.09). Consequently, the ML method detects significant differences more often than validation alone: again, for high sensitivity and specificity and OR=2, the ML method is able to detect the difference in risk between the smokers and non-smokers 100% of the time, whereas the validation data estimates only detect a difference 62% of the time. The coverage percentage (the percent of times that the 95% confidence interval includes the true value) should be close to 95, and we see that this is true for both methods. In short, for high sensitivity and specificity, the ML method achieves about the same or better bias than the validation-only method, but, because it has lower variability, the ML method detects significant differences far more often than estimates based on manually validated data alone. Estimates based on NLP alone, are clearly the most biased.

When the specificity of the NLP process is low (0.6), the ML method underestimates the true OR even when the sensitivity is high (0.9): for OR $= 2$, the mean values of $\widehat{\text{OR}}$ are 2.05 and 1.81 for the validation sample and ML methods, respectively. This is unsurprising, since with low specificity, the ML method must allow for large numbers of non-smokers being misclassified as smokers; In fact, with low specificity, the majority of subjects classified as smokers by the NLP method will be non-smokers. This effect generates a "bias towards the null", which translates into a tendency to systemati-

cally underestimate the odds ratio. However, the effect seems modest in these cases. Estimates based only on the manually extracted values do not experience such an effect; again, this is unsurprising, since this method is independent of the NLP process. However, the ML method generally produces estimates that are known with greater accuracy, in a familiar phenomenon of the bias-variance trade-off: sometimes a small amount of bias is a good price to pay for knowing an estimate with much greater accuracy. Therefore, again for specificity=0.6 and sensitivity=0.9, the method based on the manually extracted data detected a significant difference 64% of the time whereas the ML method detected a significant difference 100% of the time. Generally, the advantage in power of the ML method has over the method based only on the manually extracted data should be worth the price of the small bias in the estimate of the risk. By contrast, NLP-method alone produces an estimate that is highly biased (for example, for sensitivity=0.9, specificity=0.6, $\widehat{\text{OR}} = 1.30$ where $\text{OR} = 2$) and seldom includes the correct odds ratio in the confidence interval (in our simulations the coverage percentage is often zero).

We see similar results for $\text{OR} = 1.2$, except that, as expected, the ML method detects a significant difference less often when the true effect size is smaller. However, the ML method still performs better than the other approaches.

In summary, regardless of sensitivity and specificity, the ML method achieves about the same level of bias as the method based on the validation sample. However, because the ML method is based on a large sample size and has lower variability, the ML method detects significant differences far more often than estimates based on manually validated data alone.

4 Discussion and Conclusion

Extraction of data from free-text notes in a large dataset can be limited by the time- and labor-intensive nature of data abstraction; it is frequently only practical to manually abstract data from a small fraction of the dataset. In such a case, statistical techniques suffer from the small sample size, leading to underpowered tests and imprecise estimates. However, when such data are used in conjunction with an NLP algorithm applied to the entire dataset, the statistical tests become considerably more powerful, although at the cost of possibly introducing a small amount of bias.

Patients' smoking status was a good candidate for this method as it is an essential variable in analyses of many diseases, and is usually recorded only in the patients' admission and discharge summaries. In most epidemiological studies, adjusting for smoking status is considered a prerequisite for the analysis to be considered plausible. This variable is so important that one of the first tasks in the i2b2 NLP challenges was extraction of the patients' smoking status from discharge summaries (Uzuner O, Goldstein I, Luo Y, Kohane I, 2008). The best systems in the i2b2 evaluation achieved microaveraged F-measures (a harmonic mean of recall and precision) above 0.84. Subsequent studies report improvements in smoking status detection up to almost 90% F-score (Sohn S, Savova GK, 2009).

We made the assumption that the NLP-derived smoking predictor was binary, i.e. patients were classified as being either smokers or non-smokers, and any patients categorized as "unknown" were included in the non-smoker category. This assumption is reasonable since absence of information in the clinical narrative about smoking status is likely to signal that smoking is not an issue for that patient. However, we are looking to extend this method to handle multinomial predictors: for example, smoker, non-smoker, and unknown smoking status.

In summary, using the ML misclassification method will enable researchers to incorporate NLP-derived variables into their analysis and thereby largely avoiding the problems of bias and loss of power. Our method provides a new source of predictors for research by accounting for the error in NLP variable extraction.

Acknowledgments and Disclosures

The authors would like to thank Samantha Tate for manual validation. The opinions presented here do not necessarily represent those of the US Food and Drug Administration. This research was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Specificity	Method	Sensitivity											
		0.6				0.8				0.9			
		$\widehat{OR}$	SE	% Cov.	% Sig.	$\widehat{OR}$	SE	% Cov.	% Sig.	$\widehat{OR}$	SE	% Cov.	% Sig.
		True odds ratio = 1.2											
0.6	NLP	1.03	0.06	32	5	1.05	0.06	42	10	1.07	0.06	55	18
	Validation	1.22	0.35	96	9	1.24	0.35	95	7	1.25	0.35	95	8
	ML	1.21	0.26	97	10	1.17	0.17	95	14	1.17	0.14	95	24
0.8	NLP	1.07	0.07	59	14	1.09	0.07	69	24	1.10	0.07	75	30
	Validation	1.25	0.35	96	9	1.25	0.35	95	9	1.24	0.35	95	9
	ML	1.20	0.17	96	18	1.19	0.12	95	26	1.18	0.11	94	33
0.9	NLP	1.10	0.08	82	21	1.13	0.07	86	37	1.13	0.07	86	43
	Validation	1.24	0.35	96	9	1.23	0.35	96	7	1.22	0.35	95	9
	ML	1.20	0.14	95	25	1.20	0.11	95	39	1.19	0.10	95	43
		True odds ratio = 2											
0.6	NLP	1.11	0.06	0	43	1.23	0.06	0	93	1.30	0.06	0	99.5
	Validation	2.05	0.30	95	60	2.09	0.30	95	66	2.08	0.30	95	64
	ML	1.94	0.23	94	80	1.86	0.15	90	99	1.81	0.12	86	100
0.8	NLP	1.28	0.06	0	97	1.41	0.06	0	100	1.48	0.06	0	100
	Validation	2.09	0.30	96	67	2.08	0.30	94	65	2.05	0.30	96	65
	ML	1.94	0.15	94	99	1.91	0.11	93	100	1.88	0.10	88	100
0.9	NLP	1.45	0.07	0	100	1.58	0.06	4	100	1.64	0.06	12	100
	Validation	2.06	0.29	95	64	2.07	0.30	95	65	2.05	0.30	95	62
	ML	2.01	0.13	97	100	2.01	0.10	96	100	2.00	0.09	95	100

Table 1: Results from the simulations. The "NLP" method uses only the information from the NLP smoking predictor (W) to predict the outcome (Y), smoking-related cancer; the "Validation" method uses only the information from the validation sample ("true" smoking predictor X) to predict Y; and the "ML" method uses both the validation information X and the NLP information about smoking status W to predict Y. Each cell contains: 1) Mean estimate of the odds ratio $\widehat{OR}$; 2) Mean estimate of the standard deviation of $\widehat{OR}$ (Standard Error); 3) the "Coverage Percentage": the percent of times the true odds ratio falls in the estimated 95% confidence interval (should be close to 95); and 4) The percent of times that a significant result is detected (the percent of times the confidence interval does *not* include 1). Each simulation is based on 1000 replications. Size of the validation sample is $n_v = 1000$ or 5% of the total sample, $N = 20,000$

References

Benoit K, Laver M, Mikhaylov S. 2009. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, 53(2):495–513.

Buonaccorsi JP, Laake P, Veirød M. 2005. On the effect of misclassification on bias of perfectly measured covariates in regression. *Biometrics*, 61:831–36.

Buzas JS, Stefanski LA. 1996. Instrumental variable estimation in probit measurement error model. *Journal of Statistical Planning and Inference*, 55:47–62.

Callaghan FM, Jackson MT, Demner-Fushman D, Abhyankar S, McDonald CJ. 2012. Misclassification model for NLP derived variables: a case study and a simulation. [Manuscript].

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. 2006. *Measurement error in nonlinear models: A modern perspective*. Chapman & Hall/CRC, 2nd edition.

Carroll RJ, Stefanski LA. 1990. Approximate quasilikelihood estimation in models with surrogate predictors. *JASA*, 85:652–63.

Centers for Disease Control and Prevention. 2011. Vital signs: Current cigarette smoking among adults aged greater than or equal to 18 years – United States, 2005–2010. *Morbidity and Mortality Weekly Report*, 60(33):1207–12. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6035a5.htm?s_cid=%20mm6035a5.htm_w.

Cook JR, Stefanski LA. 1994. Simulation-extrapolation estimation in parametric measurement error models. *JASA*, 89:1314–28.

Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. 2010. caTIES: A grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. *J Am Med Inform Assoc*, 17(3):253–64.

Gleser LJ. 1990. Improvements of the naïve approach to estimation in non-linear errors-in-variables regression models. In Fuller WA Brown PJ, editor, *Statistical analysis of measurement error models and application*. American Mathematics Society, Providence.

Grimmer J, Stewart BM. 2012. Text as data: The promise and pitfalls of automatic content. www.stanford.edu/j̃grimmer/tad2.pdf.

Gustafson P. 2004. *Measurement error and misclassification in statistics and epidemiology*. Chapman & Hall/CRC.

Hopkins D, King G. 2010. Extracting systematic social science meaning from text. *American Journal of Political Science*, 54(1):229–47.

Kuchenhoff H, Mwalili SM, Lesaffre E. 2006. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*, 62(1):85–96.

Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T. 2011. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, 306(8):848–55.

National Cancer Institute, National Institutes of Health. 2012. Factsheet: Harms of smoking and health benefits of quitting. http://www.cancer.gov/cancertopics/factsheet/Tobacco/cessation, [reviewed 2011 Jan 12; cited 2012 Mar 14].

Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, Tao C, Parker CG, Haug PJ, Huff SM, Chute CG. 2012. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. *J Biomed Inform*. [Epub ahead of print].

Saeed M, Lieu C, Raber G, Mark RG. 2002. MIMIC II: A massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, 29:641–44. http://mimic.physionet.org/.

Sohn S, Savova GK. 2009. Mayo clinic smoking status classification system: extensions and improvements. *AMIA Annu Symp Proc*, 15(1):619–23.

Stefanski LA, Buzas JS. 1995. Instrumental variable estimation in binary regression measurement error models. *JASA*, 90:541–50.

Stefanski LA, Carroll RJ. 1987. Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, 74:703–16.

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

Wang X, Hripcsak G, Markatou M, Friedman C. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc*, 16(3):328–37.

Wilcox AB, Hripcsak G. 2003. The role of domain knowledge in automating medical text report classification. *J Am Med Inform Assoc*, 10(4):330–38.

Woolf B. 1955. On estimating the relation between blood group and disease. *Ann Hum Genet*, 19:251–3.