Controlling for Unobserved Confounding with Large Language Model Classification of Patient Smoking Status

Samuel Lee
Department of Computer Science
Northwestern University
Evanston, IL 60201
samlee2023@u.northwestern.edu

Zach Wood-Doughty
Department of Computer Science
Northwestern University
Evanston, IL 60201
zach@northwestern.edu

Abstract

Causal understanding is a fundamental goal of evidence-based medicine. When randomization is impossible, causal inference methods allow the estimation of treatment effects from retrospective analysis of observational data. However, such analyses rely on a number of assumptions, often including that of no unobserved confounding. In many practical settings, this assumption is violated when important variables are not explicitly measured in the clinical record. Prior work has proposed to address unobserved confounding with machine learning by imputing unobserved variables and then correcting for the classifier's mismeasurement. When such a classifier can be trained and the necessary assumptions are met, this method can recover an unbiased estimate of a causal effect. However, such work has been limited to synthetic data, simple classifiers, and binary variables. This paper extends this methodology by using a large language model trained on clinical notes to predict patients' smoking status, which would otherwise be unobserved. We then apply a measurement error correction on the categorical predicted smoking status to estimate the causal effect of transthoracic echocardiography on mortality in the MIMIC dataset.

1 Introduction

Evidence-based medicine seeks to use data from clinical research to inform best practices [Sackett et al., 1996, Masic et al., 2008]. While randomized control trials (RCTs) are the gold standard of clinical research, randomization is often impossible or unethical [Sanson-Fisher et al., 2007]. In the absence of randomization, retrospective causal analyses must control for confounding variables that could influence both a patient's treatment assignment and their outcome [VanderWeele and Shpitser, 2013]. For example, if a certain medication is only given to the patients with the highest risk, a naive correlational analysis may suggest that the medication itself increases the risk of mortality. Causal inference methods establish causal claims by controlling for confounding and other sources of bias [Pearl, 2009].

All causal inference methods rely on assumptions about the underlying data-generating process; a common but untestable assumption is the absence of unobserved confounding [Groenwold et al., 2009]. If a confounding variable is unobserved, it is in general impossible to estimate the desired causal effect without bias [Shpitser and Pearl, 2006]. Even with the wide availability of electronic health record (EHR) data, it may be difficult to rule out the possibility of unobserved confounding [Groenwold et al., 2008]. Many causal methods attempt to detect and mitigate unobserved confounding [Flanders et al., 2011, Baiocchi et al., 2014, Uddin et al., 2016, Tchetgen et al., 2020].

One such methodological approach considers the case where a necessary confounder is unobserved, but a proxy for that variable is instead observed. A canonical example is Medicaid enrollment as a proxy for socioeconomic status (SES), because Medicaid eligibility is tied to household income [Simpson, 2020]. While a naive approach of treating SES and Medicaid enrollment as equivalent variables could introduce bias [Greenland and Robins, 1985, Ogburn and VanderWeele, 2012], methods developed by Pearl [2010] and Kuroki and Pearl [2014] can recover unbiased causal estimates through a measurement error correction. When such a proxy (e.g., Medicaid enrollment) is unavailable, one option may be to train a machine learning (ML) classifier to predict the values of the unobserved confounder and then apply this measurement error correction to account for the classifier's errors [Wood-Doughty et al., 2020].

This paper builds upon this line of work, extending prior methods from simple classifiers and synthetic data to large language models (LLMs) and two real-world datasets. Following Feng et al. [2018], we analyze the causal effect of transthoracic echocardiography (TTE) on 28-day mortality among ICU patients diagnosed with sepsis. A limitation of prior work is its inability to account for patient smoking status; this variable is not explicitly recorded in the EHR data, but tobacco use is considered a risk factor for mortality among patients with sepsis [Alroumi et al., 2018, Zhang et al., 2022]. Smoking status could also plausibly affect a clinician's decision to administer TTE, which would make it an unobserved confounder. We train and apply an ML classifier to predict patients' smoking status and estimate the model's error rate [Mulyar et al., 2019]. We extend the measurement error correction to categorical variables and use it to debias our causal estimates. Our method produces estimates with modest variance and our results provide evidence that TTE prevents mortality, with or without accounting for patient smoking status.

This primary contribution of this paper is a methodology for imputing unmeasured variables with large language models and then estimating unbiased causal effects with a measurement error correction. We release our code to enable future research.[1]

2 Related Work

2.1 ML and LLMs for Clinical Prediction

Machine learning (ML) methods have advanced over the past decade to provide impressive real-world performance in a wide variety of domains. For text data, large language models (LLMs) have become the dominant paradigm for text generation and classification tasks [Zhou et al., 2023]. The most common approach for training LLMs uses generative pretraining followed by supervised task-specific fine-tuning [Radford et al., 2018]. In the clinical domain, LLMs have been applied to a wide variety of tasks, such as 30-day all-cause readmission prediction, in-hospital mortality prediction, comorbidity index prediction, length of stay prediction, and insurance denial prediction [Jiang et al., 2023]. While such methods can demonstrate state-of-the-art performance on a variety of datasets, their applications to real-world clinical practice have been somewhat limited [Chen and Asch, 2017]. Such models are complicated and often uninterpretable, which may prevent clinicians from trusting the reasoning underlying their predictions [Tonekaboni et al., 2019]. Additionally, test set evaluations may not be fully indicative of real-world performance in the presence of domain shift and other sources of bias [Subbaswamy and Saria, 2020, Finlayson et al., 2021].

We focus on the specific task of predicting patient smoking status from clinical notes, using a labeled dataset released by Uzuner et al. [2008]. Variations on this task have been a sustained focus of ML applications because of the relevance of tobacco use to a wide variety of health outcomes [Sohn and Savova, 2009, Palmer et al., 2019].

2.2 Causal Inference

Whereas practical applications of ML methods to the clinical domain are a relatively recent phenomenon [Jiang et al., 2023], many of the foundational methods in causal inference were developed to enable epidemiological studies [Pearl, 2009]. In this work, we use the potential outcomes framework that explores the causal effect of a treatment X on an outcome Y: specifically, following Feng et al. [2018], the causal effect of TTE on 28-day mortality among patients with sepsis. We use

---

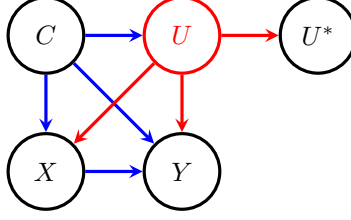[1]https://github.com/controllingunobservedpaper/NEURIPS_submission_2024

Figure 1: DAG with treatment X, outcome Y, and observed confounders C. U^* is the observed but noisy proxy for the unobserved confounder U.

causal do-notation $p(Y \mid \mathrm{do}(x))$ to denote the distribution of a patient's hypothetical mortality had they be *randomly assigned* TTE value x [Pearl, 2012]. In an RCT, $E[Y \mid X = x]$ is equivalent to $E[Y \mid \mathrm{do}(x)]$ because correlation between the randomly-assigned treatment and outcome implies (in probability) a causal effect [Pearl, 2009]. Because in our dataset TTE was not randomly assigned, we need additional assumptions to *identify* the counterfactual distribution $E[Y \mid \mathrm{do}(x)]$ from the observed data.

We use directed acyclic graphs (DAGs) to represent our assumptions about the data [Pearl, 1995]. Figure 1 shows the causal DAG for our problem. In addition to our treatment X and outcome Y, C represents our vector of observed confounders, including patient age, gender, and several clinical markers. U is smoking status, which is hypothesized as a possible confounder but is unobserved in our data. U^*, introduced in §2.3, is our noisy proxy.

Both our treatment and outcomes are binary variables, and our causal estimands of interest are the risk ratio and odds ratio, equations (1) and (2) below, respectively [Hernan and Robins, 2020]:

$$\mathrm{RR}: \frac{p(Y = 1 \mid \mathrm{do}(x = 1))}{p(Y = 1 \mid \mathrm{do}(x = 0))} \tag{1}$$

$$\mathrm{OR}: \frac{p(Y = 1 \mid \mathrm{do}(x = 1))p(Y = 0 \mid \mathrm{do}(x = 0))}{p(Y = 0 \mid \mathrm{do}(x = 1))p(Y = 1 \mid \mathrm{do}(x = 0))} \tag{2}$$

2.3 Measurement Error

If all confounders C and U were observed, then $p(Y \mid \mathrm{do}(x))$ is identified – as we show in Equation 10 below – and can be estimated in a number of ways [Glynn and Quinn, 2010]. If U is unobserved, however, then the causal effect of X on Y is *unidentified* and the problem becomes impossible without additional assumptions [Shpitser and Pearl, 2006]. We assume access to U^*, a noisy proxy with the same cardinality as U that is governed by some noisy relationship $p(U^* \mid U)$. If a matrix representation of this probability $p(U^* \mid U)$ can be inverted, then the work of Pearl [2010] and Kuroki and Pearl [2014] show that we can recover $p(Y \mid \mathrm{do}(x))$. This method is known as and alternatively referred to as 'matrix adjustment' or 'effect restoration' [Kuroki and Pearl, 2014]; we demonstrate this derivation in §3.2. Throughout this paper, following §3.1, we assume that U^* is predicted by an LLM classifier and that $p(U^* \mid U)$ is the error rate of the classification. Our method assumes no *additional* unobserved confounding other than U.

An alternative to matrix adjustment methods is simulation-extrapolation (SIMEX), a widely-used method for measurement error [Cook and Stefanski, 1994]. SIMEX generally requires strong parametric assumptions about the data [Sevilimedu and Yu, 2022], prior knowledge of the known measurement error variance [Cook and Stefanski, 1994], and the unknown confounder to be continuous [Lederer and Küchenhoff, 2006]. To overcome these limitations, Küchenhoff et al. [2005] proposed MC-SIMEX which avoids parametric assumptions but has weaker theoretical foundations. While we focus on developing matrix correction methods and leave a comprehensive survey of SIMEX methods for other work [Sevilimedu and Yu, 2022], we compare our proposed methodology against MC-SIMEX in our experiments.
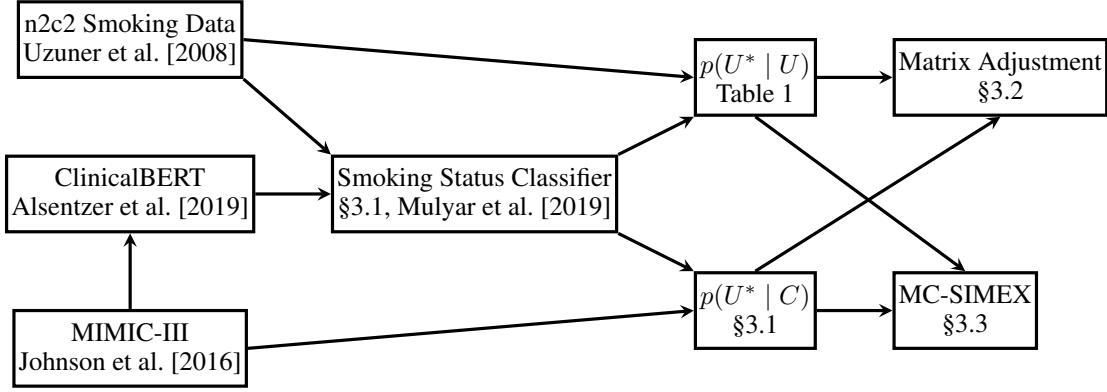
Figure 2: Overview of training and estimation methodology

3 Methodology

We divide our methodological approach into two sections. First, we discuss the classifier we use to predict smoking status for patients in the MIMIC dataset. We discuss the model's training and validation procedure and its classification of our data. Then, we incorporate our model's predictions and our estimate of its error rate into the measurement error formulation to estimate causal effects. We discuss our extensions of past work to categorical mismeasured variables and high-dimensional observed confounders. Figure 2 shows the high-level overview of our entire methodology, including the pretrained LLMs and most important data sources.

3.1 Smoking Status Classifier

Our modeling approach most closely follows that of Mulyar et al. [2019]. We start with a ClinicalBERT LLM architecture using the weights that were released by Alsentzer et al. [2019] after pretraining on PubMed abstracts, PMC articles, and all MIMIC-III notes [Alsentzer et al., 2019, Lee et al., 2019]. We then adapt this LLM by training an LSTM on top of its representation to predict categorical smoking status in the 2006 n2c2 smoking dataset [Uzuner et al., 2008]. The training dataset contains 398 clinical notes and corresponding labels of the patient as either past smoker, current smoker, non-smoker, and unknown. These labels were manually annotated by pulmonologists. We follow the hyperparameter choices of Mulyar et al. [2019] and train for 1,000 epochs.

After training, we evaluate our model on the test set of 101 additional patient records. Our model achieves 86.1% accuracy with the confusion matrix shown in Table 1.[2] This confusion matrix, after dividing by the row sums to produce a stochastic matrix, provides the error-rate matrix $p(U^* \mid U)$ that will be used in our measurement error correction in the following section.

To apply this model to patients in the MIMIC dataset, we first preprocess that data using the MIMIC-EXTRACT approach [Wang et al., 2020]. We then follow the preprocessing steps of Feng et al. [2018] to enable a direct comparison against their analysis. Our trained classifier is then used to predict smoking status for the patients in the evaluation set of MIMIC patients who met the criteria for sepsis [Angus et al., 2016]. Of those, 2,058 were classified as past smoker, 93 as current smoker, 1,413 as non-smoker, and 1,171 as unknown. We discard 64 patients for whom the classifier did not make a prediction due to a confidence threshold set by Mulyar et al. [2019]. Specifically, the model makes no prediction if its final `tanh` activation outputs a negative logit value for all four classes.

---

[2]As accuracy here is equivalent to micro-averaged F1, our model performs lower than the 92.8% accuracy reported by Mulyar et al. [2019]. We hypothesize this is simply due to random variation or changes in the Pytorch libraries.

Table 1: Confusion matrix for model predictions of smoking status on the n2c2 test set [Uzuner et al., 2008]. True labels are on the rows, predicted values on the columns.

	Past	Current	Never	Unknown
Past	8	0	2	1
Current	4	4	3	0
Never	1	0	14	1
Unknown	1	0	1	61

3.2 Matrix Adjustment

We can now incorporate our smoking status predictions into our estimate of the causal effect of TTE on 28-day mortality. Figure 1 portrays our assumptions about the underlying data distribution. Our treatment X is TTE, our outcome Y is 28-day mortality, and C is a vector of 39 observed covariates.[3] Patient smoking status U is our unobserved confounder, and U^* is our classifier's prediction. Given this notation, we have data on the joint probability $p(X, Y, C, U^*)$ from patients in MIMIC. We additionally have the error rate $P(U^* \mid U)$ from our classifiers validation on the n2c2 dataset. We assume that the classifier's error rate remains constant between the two datasets; future work could relax this assumption by collecting additional ground-truth annotations on the MIMIC dataset. We further make a *non-differential error* assumption, that:

$$p(U^* \mid U, X, Y, C) = P(U^* \mid U) \tag{3}$$

Because our smoking status is a categorical variable with four values, our error rate distribution $p(U^* \mid U)$ can be represented as a 4x4 matrix $M(U^*, U)$ computed from the confusion matrix in Table 1 by dividing each cell by its corresponding row sum. In general, using u_i to denote $U = i$ and u_j^* to denote $U^* = j$, this can be written as:

$$\begin{bmatrix} P(u_0^* \mid u_0) & P(u_0^* \mid u_1) & P(u_0^* \mid u_2) & P(u_0^* \mid u_3) \\ P(u_1^* \mid u_0) & P(u_1^* \mid u_1) & P(u_1^* \mid u_2) & P(u_1^* \mid u_3) \\ P(u_2^* \mid u_0) & P(u_2^* \mid u_1) & P(u_2^* \mid u_2) & P(u_2^* \mid u_3) \\ P(u_3^* \mid u_0) & P(u_3^* \mid u_1) & P(u_3^* \mid u_2) & P(u_3^* \mid u_3) \end{bmatrix}$$

We can use $M(U^*, U)$ to write our observed data distribution as a function of the underlying distribution and our error rate $p(U^* \mid U)$.

$$p(Y, X, C, U^*) = \sum_U p(U^* \mid Y, X, C, U) p(Y, X, C, U) \tag{4}$$

$$= \sum_U p(U^* \mid U) p(Y, X, C, U) \tag{5}$$

$$= M(U^*, U) \times p(Y, X, C, U) \tag{6}$$

Equation 4 holds by rules of probability, (5) holds by (3) above, and (6) is rewritten to formulate this summation over probability distributions as a matrix multiplication. This follows the approach of Pearl [2010]; for fixed values of Y, X, C, $p(Y, X, C, U)$ is a vector of four values. We then multiply that vector by our 4x4 error matrix $M(U^*, U)$ to get a different vector of four values.

Assuming $M(U^*, U)$ has an inverse $I(U^*, U)$, we then have:

$$p(Y, X, C, U) = I(U^*, U) \times p(Y, X, C, U^*) \tag{7}$$

This inverse matrix $I(U^*, U)$ allows us, given $p(Y, X, C, U^*)$ and $M(U^*, U)$, to recover $p(Y, X, C, U)$. This derivation follows those presented in prior work by Pearl [2010] and Kuroki and Pearl [2014]. To move from $p(Y, X, C, U)$ to our counterfactual $p(Y \mid \text{do}(x))$, we can follow a standard do-calculus derivation:

---

[3]See Table 1 of Feng et al. [2018] for a complete list.

$$p(Y \mid \mathrm{do}(x)) = \sum_{C,U} p(Y \mid \mathrm{do}(x), C, U) p(C, U \mid \mathrm{do}(x)) \tag{8}$$

$$= \sum_{C,U} p(Y \mid X = x, C, U) p(C, U \mid \mathrm{do}(x)) \tag{9}$$

$$= \sum_{C,U} p(Y \mid X = x, C, U) p(C, U) \tag{10}$$

Equation 8 hold by marginalization and chain rule, (9) holds by Rule 2 of do-calculus, and (10) holds by Rule 3 of do-calculus [Pearl, 2012].[4] However, the relative simplicity of the proof that the counterfactual is *identified* obscures some of the practical challenges of *estimating* causal effects in our specific domain application. To implement estimators for the risk and odds ratios requires fitting models to these probability distributions. Unlike for example in Wood-Doughty et al. [2020] where there is a single binary observed confounder, we have 39 covariates.

As Equation 10 is a function of $p(Y \mid X, C, U)$ and $p(C, U)$, we can first use marginalization and conditioning to derive these from $p(Y, X, C, U)$ as provided to us by Equation 7. To actually compute Equation 7, we fit three models: logistic regressions for $p(Y \mid X, C, U^*)$ and $p(X \mid C)$, and a multinomial logit model $p(U^* \mid X, C)$.[5] All models are fit using the *predicted* smoking statuses (i.e., using U^*, not on U); we perform the measurement error correction *after* fitting models.

$$p(Y, X, U \mid C) = I(U^*, U) p(Y, X, U^* \mid C)$$
$$= I(U^*, U) p(Y \mid X, U^*, C) p(U^* \mid X, C) p(X \mid C) \tag{11}$$

Because C is high-dimensional, we avoid modeling it and use the empirical approximation $1/N$.

$$p(U \mid C) = \sum_X p(U, X \mid C)$$

$$= \sum_X I(U^*, U) p(U^*, X \mid C)$$

$$= \sum_X I(U^*, U) p(U^* \mid X, C) p(X \mid C) \tag{12}$$

$$p(U, C) \approx \frac{1}{N} \sum_{i=1}^{N} p(U \mid C_i) \tag{13}$$

While Equation 11 defines $p(Y, X, U \mid C)$, we can easily transform it to calculate the conditional probability we need for our counterfactual.

$$p(Y \mid X, C, U) = \frac{p(Y, X, U \mid C)}{\sum_y p(Y = y, X, U \mid C)} \tag{14}$$

Then, plugging Equations 14 and 13 into (10), we have written our counterfactual $p(Y \mid \mathrm{do}(x))$ as a function of our observed data distribution. We can plug that counterfactual into the definitions of Equations 1 and 2 to estimate our causal effects.

3.3 MC-SIMEX

Following our discussion in §2.3, we compare our proposed matrix adjustment method against MC-SIMEX, another common method for handling measurement error. While the original SIMEX

---

[4]For a gentle and thorough derivation of this, see `https://www.andrewheiss.com/blog/2021/09/07/do-calculus-backdoors/`.

[5]Note that replacing $p(Y \mid \cdot)$ with a linear regression $E[Y \mid \cdot]$ would trivially extend this to applications with continuous outcomes where the risk difference is the estimand of choice.

Table 2: Comparison of our method's causal estimates against those of the five methods from Table 2 of Feng et al. [2018]. DR: Doubly Robust; PS: Propensity Score; IPW: Inverse Propensity Weighting. All results indicate a protective effect of TTE against mortality.

	Odds Ratio	Confidence Interval
Matrix §3.2	0.89	[0.60, 0.98]
MC-SIMEX §3.3	0.90	[0.83, 1.00]
DR Unbalanced	0.78	[0.68, 0.90]
DR All	0.64	[0.52, 0.78]
PS IPW	0.84	[0.78, 0.92]
PS Matching	0.78	[0.66, 0.92]
Multivariate	0.64	[0.53, 0.78]

Table 3: Boostrapped 95% confidence intervals for Risk Ratio and Odds Ratio estimates. The 'None' row shows the point estimate without bootstrapping. Each other row is computed across 100 total resampled datasets, with bootstrapping applied to the MIMIC dataset used to fit our models, the n2c2 test set used to estimate $M(U^*, U)$, or both datasets.

Bootstrap	Matrix Adjustment		MC-SIMEX	
	RR	OR	RR	OR
None	0.88	0.89	0.93	0.90
MIMIC	[0.85, 0.90]	[0.87, 0.90]	[0.89, 0.96]	[0.76, 0.95]
n2c2	[0.77, 0.99]	[0.77, 0.98]	[0.93, 0.94]	[0.90, 0.91]
Both	[0.65, 0.92]	[0.60, 0.98]	[0.85, 1.01]	[0.83, 1.00]

method was designed for continuous unobserved variables, MC-SIMEX allows for categorical variables [Lederer and Küchenhoff, 2006]. We use the R package released by the authors.[6] As indicated in Figure 2, SIMEX also requires access to both the noisy imputed U^* values and our estimate of the error rate $p(U^* \mid U)$. Despite the similarities of these methods, our matrix adjustment approach has the benefit of more established theoretical foundations, and makes no assumptions about the parametric form of the estimator. It can also be easily applied to more complicated counterfactual estimands where the identifying functional of the observed data is not simply a linear model of the treatment and covariates, e.g., a front-door or proximal estimator [Bellemare et al., 2020, Cui et al., 2023].

4 Results

Our experimental analysis follows Figure 2 in applying both matrix adjustment and MC-SIMEX to the MIMIC-III dataset as preprocessed and analyzed by Feng et al. [2018]. We use our trained LLM classifier to impute patients' smoking status, fit our models for $p(Y \mid X, U^*, C)$, $p(U^* \mid X, C)$, and $p(X \mid C)$, and then incorporate those models and our estimated error rates $p(U^* \mid U)$ to produce our causal estimates. Our full experimental code is provided as an appendix.

Table 2 shows a direct comparison of our odds ratio estimates against the five separate estimation methods presented by Feng et al. [2018]. These numbers being less than 1 indicates our treatment reduces the probability of the outcome – that is, TTE prevents mortality among the patient population. Our results match those of past work, but produce more conservative estimates of the causal effect. This comparison cannot prove or refute the presence of unobserved confounding via smoking status; it merely shows that when taking smoking status into account, we produce a more conservative estimate of the protective effect of TTE.

If our assumptions from §2 and 3 are valid, our methods should produce unbiased estimates of the causal effects. However, with any finite sample data sample, we must worry about the variance of our methodology. Feng et al. [2018] also provide (Wald-type) confidence intervals for their analyses. To produce the confidence intervals for our methods shown in Table 3, we use the non-parametric

---

[6] https://github.com/cran/simex

Table 4: Comparison of risk ratios overall and within smoker subgroups. The left column shows our point estimates; the other columns show confidence intervals from bootstrapping either the MIMIC evaluation set or the n2c2 test set.

	Matrix Adjustment			MC-SIMEX		
	RR	MIMIC CI	n2c2 CI	RR	MIMIC CI	n2c2 CI
Overall	0.88	[0.85, 0.90]	[0.77, 0.99]	0.93	[0.86, 1.05]	[0.93, 0.94]
Past Smoker	0.86	[0.76, 0.93]	[0.32, 1.37]	0.93	[0.89, 0.96]	[0.93, 0.93]
Current Smoker	0.92	[0.92, 0.92]	[0.92, 0.92]	0.97	[0.91, 0.97]	[0.98, 0.98]
Non-Smoker	1.13	[0.65, 1.60]	[0.06, 2.39]	0.91	[0.84, 0.95]	[0.91, 0.91]
Unknown	0.88	[0.70, 1.00]	[0.54, 1.16]	0.94	[0.91, 0.97]	[0.94, 0.94]

bootstrap [DiCiccio and Efron, 1996]. Bootstrapping our entire analysis from LLM pretraining to effect estimation is unfortunately prohibitively expensive; we instead use two bootstrap approaches that can give us a full picture of our method's uncertainty.

First, we consider resampling our evaluation subset of MIMIC-III to produce 100 new datasets of 4,735 patients, with each patient in each new dataset chosen independently with replacement from the original dataset. For each of these datasets, we fit new models for $p(Y \mid X, U^*, C)$, $p(U^* \mid X, C)$, and $p(X \mid C)$. For each of these analysis, we hold fixed the estimate for $M(U^*, U)$ computed with the n2c2 test dataset. The 95% confidence interval from this bootstrap alone is shown in the 'MIMIC' row of Table 3.

Our second, orthogonal bootstrapping method follows Wood-Doughty et al. [2020]. Rather than resampling our analysis dataset of MIMIC patients, we resample the test set of 101 patients in the n2c2 data. For each of 100 such resamplings, we recalculate our error matrix $M(U^*, U)$, invert it to produce $I(U^*, U)$, and then use that in Equation 11 and (13) to compute our causal effects. Each recalculated $M(U^*, U)$ is used alongside the original MIMIC dataset of 4,735 patients. The 95% confidence intervals from these 100 resamplings is shown in the 'n2c2' row of Table 3.

The 'Both' row of the table shows a combination of both bootstrap methods. We resample 10 new $M(U^*, U)$ matrices and 10 new MIMIC datasets on which we fit our models. Then, for all $10 \times 10 = 100$ comparisons, we compute our final estimates of the risk and odds ratios. This gives us 100 resamplings, matching the MIMIC-only and n2c2-only bootstrap intervals. The confidence intervals we showed previously in Table 2 use this combined bootstrap method, as it has the widest interval of the three.

Table 3 allows us to highlight and compare two sources of uncertainty in our method. For the matrix adjustment method, it is unsurprising that the n2c2 bootstrap introduces more uncertainty into our final estimates, as the test set contains only 101 patients. The matrix inverse we compute in Equation 7 and use in (11) and (12) implicitly requires us to divide by values in our estimated $M(U^*, U)$ matrix. As with inverse propensity weighting, this can introduce high variance when dividing by small numbers [Ma and Wang, 2020]. We could reduce some of this uncertainty by using a larger dataset to validate the smoking status classifier. Interestingly, our MC-SIMEX estimates have much less variance when bootstrapping the n2c2 test set. This may be related to the lack of theoretical foundation to explain valid estimates for standard errors in simulations [Lederer and Küchenhoff, 2006]. None of our current methods are able to quantify the uncertainty introduced from the BERT pretraining or classifier fine-tuning. While there are some methods that can capture uncertainty in neural networks (e.g., Gal and Ghahramani [2016]), we leave further sensitivity analyses and uncertainty quantification for future work.

Table 4 shows our risk ratio broken down across our four subgroups of smoking statuses. While the risk and odds ratios are closely related, the former is *collapsible*, so the risk ratio of the entire population is the weighted average of the risk ratios within each subgroup [Hernan and Robins, 2020]. This allows us to easily ask whether TTE is more or less effective among patients with different smoking statuses.[7] For the matrix adjustment method, for all categories except current smokers, the

---

[7]In §3.1 we discarded 64 patients for whom the classifier made no prediction. If we instead label those patients as "Unknown," then using Matrix Adjustment and no bootstrapping, the overall risk ratio is 0.90, the odds ratio is 0.82, and the subgroup risk ratios are [0.91, 0.92, 1.15, 0.97].

confidence intervals in this subgroup analysis widen and several intervals contain 1.0, suggesting we lack the requisite data for such focused analyses. For non-smokers the point estimate of our risk ratio is above 1.0, which indicates that TTE is in fact harmful; while this could warrant further study, the wide intervals suggest this may be due to random variability within the small sample size. Across both methods, an analysis of the coefficients in our $p(Y \mid X, U^*, C)$ models for the smoking status variable indicates that non-smokers have the best overall odds of survival, which aligns with past investigations of the connection between tobacco use and mortality among sepsis patients [Alroumi et al., 2018, Zhang et al., 2022]. An unlikely but alternative explanation for the risk ratio of 1.13 could additional unobserved confounding if, hypothetically, non-smoker patients are *only* administered TTE if they are at higher risk of mortality due to a factor not captured by our observed covariates X.

In our matrix adjustment subgroup analysis for current smokers, our methods produce a risk ratio confidence interval that has almost zero width. This is somewhat unintuitive, especially as our MIMIC classifier predicts only 93 current smokers. We believe this is a consequence of our test set confusion matrix from Table 1; there are no examples in which the model falsely predicts a current smoker. Because we do not smooth this confusion matrix, $p(U^* = \text{Current} \mid U \neq \text{Current}) = 0$ which means that our matrix adjustment represented by Equation 7 has relatively little effect. This helps explain the zero-width interval for the n2c2 bootstrap. Homogeneity amongst the patients classified as current smokers could possibly explain the lack of variability from the MIMIC bootstrap.

The subgroup analyses for MC-SIMEX show an overall narrowing of confidence intervals, despite the reduced sample sizes. As in Table 3, we see very narrow intervals especially for the n2c2 bootstrap. The average width of MC-SIMEX n2c2 bootstrap intervals is 0.05, which is rounded down to 0 with only two significant digits. We leave for future work simulation studies and theoretical analyses that could better explore the unexpectedly low variance of this method.

5 Conclusions and Future Work

Our method can control for unobserved confounding using an LLM classifier and a measurement error correction. We have applied this method to estimate a causal effect with implications for clinical care, accounting for a possible violation of the assumptions of past work. Our overall results – that TTE protects against mortality – comport with prior work and do not suggest any particular changes to the standards of clinical care. However, our method more broadly offers a straightforward approach to predicting unobserved confounders from clinical notes, correcting for measurement error, and controlling for confounding. As unobserved confounding is an untestable assumption on which all non-randomized causal analyses rely, our method demonstrates a widely-applicable method for exploring the impact of possible confounding. For any analysis of clinical data where physicians' notes and corresponding classifiers are available, our method can be used to control unobserved confounding or as a sensitivity analysis to test the robustness of the results.

Our method relies on a number of important assumptions, and future work could explore relaxing them. First, our choice of datasets and smoking status requires us to validate the test accuracy of our LLM classifier on the n2c2 dataset and to assume that the classifier's performance is the same on the MIMIC dataset we analyze. In general, we would not expect the classifier to perform identically on two separate datasets, though our bootstrapping analysis suggests our results are robust to variations in our estimates of the model's performance. This assumption could be relaxed in future work by collecting a small test set of labeled smoking statuses amongst the MIMIC patient data. With unobserved confounders that could hypothetically be labeled by manual chart review, our method could be rendered unnecessary when sufficient time and money is available. In practice, however, the large quantity of data involved and possible concerns about manual review of clinical notes provides strong motivation for our method.

Our approach also relies on Equation (3), the assumption of non-differential error. While it seems plausible that the classifier we use should have an error rate independent of patients' covariates, this may not be true. In the most severe case where the error rate depends on X, Y, and C, this would require fitting an additional model $p(U^* \mid U, X, Y, C)$. We leave such an exploration to future work.

Our method demonstrates a general-purpose approach to adjusting for an unobserved confounder when that variable can be accurately predicted from an existing classifier with known error rate. Extensions of this work could explore the challenges of correcting for multiple unobserved confounders with multiple classifiers.

References

Fahad Alroumi, Ahmed Abdul Azim, Rachel Kergo, Yuxiu Lei, and James Dargin. The impact of smoking on patient outcomes in severe sepsis and septic shock. *Journal of intensive care*, 6(1): 1–11, 2018.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, 2019.

Derek C Angus, Christopher W Seymour, Craig M Coopersmith, Clifford S Deutschman, Michael Klompas, Mitchell M Levy, Gregory S Martin, Tiffany M Osborn, Chanu Rhee, and R Scott Watson. A framework for the development and interpretation of different sepsis definitions and clinical criteria. *Critical care medicine*, 44(3):e113–e121, 2016.

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

Marc F Bellemare, Jeffrey R Bloem, and Noah Wexler. The paper of how: Estimating treatment effects using the front-door criterion. *Oxford Bulletin of Economics and Statistics*, 2020.

Jonathan H Chen and Steven M Asch. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine*, 376(26):2507, 2017.

J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994. doi: 10.1080/01621459.1994.10476871. URL https://doi.org/10.1080/01621459.1994.10476871.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, pages 1–12, 2023.

Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, 11(3): 189–228, 1996.

Mengling Feng, Jakob I McSparron, Dang Trung Kien, David J Stone, David H Roberts, Richard M Schwartzstein, Antoine Vieillard-Baron, and Leo Anthony Celi. Transthoracic echocardiography and mortality in sepsis: analysis of the mimic-iii database. *Intensive care medicine*, 44:884–892, 2018.

Samuel G Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

W Dana Flanders, Mitchel Klein, Lyndsey A Darrow, Matthew J Strickland, Stefanie E Sarnat, Jeremy A Sarnat, Lance A Waller, Andrea Winquist, and Paige E Tolbert. A method for detection of residual confounding in time-series and other observational studies. *Epidemiology (Cambridge, Mass.)*, 22(1):59, 2011.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Adam N Glynn and Kevin M Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56, 2010.

Sander Greenland and James M Robins. Confounding and misclassification. *American journal of epidemiology*, 122(3):495–506, 1985.

RHH Groenwold, AW Hoes, KL Nichol, and E Hak. Quantifying the potential role of unmeasured confounders: the example of influenza vaccination. *International journal of epidemiology*, 37(6): 1422–1429, 2008.

RHH Groenwold, E Hak, and AW Hoes. Quantitative assessment of unobserved confounding is mandatory in nonrandomized intervention studies. *Journal of clinical epidemiology*, 62(1):22–28, 2009.

MA Hernan and J Robins. *Causal inference: What if*. Boca Raton: Chapman & Hall/CRC, 2020.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, et al. Health system-scale language models are all-purpose prediction engines. *Nature*, pages 1–6, 2023.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016. doi: https://doi.org/10.1038/sdata.2016.35.

Helmut Küchenhoff, Samuel M. Mwalili, and Emmanuel Lesaffre. A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX. *Biometrics*, 62(1):85–96, 07 2005. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2005.00396.x. URL `https://doi.org/10.1111/j.1541-0420.2005.00396.x`.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Wolfgang Lederer and Helmut Küchenhoff. A short introduction to the simex and mcsimex. *R News*, 6:26–31, 2006. ISSN 1609-3631. https://journal.r-project.org/articles/RN-2006-031/.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz682. URL `http://dx.doi.org/10.1093/bioinformatics/btz682`.

Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

Izet Masic, Milan Miokovic, and Belma Muhamedagic. Evidence based medicine–new approaches and challenges. *Acta Informatica Medica*, 16(4):219, 2008.

Andriy Mulyar, Elliot Schumacher, Masoud Rouhizadeh, and Mark Dredze. Phenotyping of clinical notes with improved document classification models using contextualized neural language models. *arXiv preprint arXiv:1910.13664*, 2019.

Elizabeth L Ogburn and Tyler J VanderWeele. On the nondifferential misclassification of a binary confounder. *Epidemiology (Cambridge, Mass.)*, 23(3):433, 2012.

Ellen L Palmer, Saeed Hassanpour, John Higgins, Jennifer A Doherty, and Tracy Onega. Building a tobacco user registry by extracting multiple smoking behaviors from clinical notes. *BMC medical informatics and decision making*, 19(1):1–10, 2019.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432, 2010.

Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 3–11, 2012.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't, 1996.

Robert William Sanson-Fisher, Billie Bonevski, Lawrence W Green, and Cate D'Este. Limitations of the randomized controlled trial in evaluating population-based health interventions. *American journal of preventive medicine*, 33(2):155–161, 2007.

Varadan Sevilimedu and Lili Yu. Simulation extrapolation method for measurement error: A review. *Statistical Methods in Medical Research*, 31(8):1617–1636, 2022. doi: 10.1177/09622802221102619. URL https://doi.org/10.1177/09622802221102619. PMID: 35607297.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

Michael Simpson. The implications of medicaid expansion in the remaining states: 2020 update. *Washington, DC: Urban Institute*, 500, 2020.

Sunghwan Sohn and Guergana K Savova. Mayo clinic smoking status classification system: extensions and improvements. In *AMIA Annual Symposium Proceedings*, volume 2009, page 619. American Medical Informatics Association, 2009.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*, pages 359–380. PMLR, 2019.

Md Jamal Uddin, Rolf HH Groenwold, Mohammed Sanni Ali, Anthonius de Boer, Kit CB Roes, Muhammad AB Chowdhury, and Olaf H Klungel. Methods to control for unmeasured confounding in pharmacoepidemiology: an overview. *International journal of clinical pharmacy*, 38:714–723, 2016.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 01 2008. ISSN 1067-5027. doi: 10.1197/jamia.M2408. URL https://doi.org/10.1197/jamia.M2408.

Tyler J VanderWeele and Ilya Shpitser. On the definition of a confounder. *Annals of statistics*, 41(1): 196, 2013.

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. Mimic-extract: a data extraction, preprocessing, and representation pipeline for mimic-iii. *Proceedings of the ACM Conference on Health, Inference, and Learning*, page 222–235, 2020. doi: 10.1145/3368555.3384469. URL https://doi.org/10.1145/3368555.3384469.

Wood-Doughty, Shpitser, and Dredze. Sensitivity analyses for incorporating machine learning predictions into causal estimates. *NeurIPS Workshop on Causal Discovery & Causality-Inspired Machine Learning*, 2020.

Nai Zhang, Yujuan Liu, Chuang Yang, Peng Zeng, Tao Gong, Lu Tao, and Xinai Li. Association between smoking and risk of death in patients with sepsis: A systematic review and meta-analysis. *Tobacco Induced Diseases*, 20, 2022.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We include the main claims in the abstract and introduction: proposing a methodology to take into account the measurement error caused by an unknown confounder, smoking status, and providing experiments to show the effects of measurement bias.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include discussion of our assumptions and their limitations throughout the paper from §3 to 5. These include the assumptions common to the matrix adjustment method and some specific to our data analysis, such as assuming our LLM classifier performs identically on the MIMIC-III data as it does on the n2c2 testing set.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide derivations (as well as assumptions made) on the effects of measurement error when considering an unobserved confounder when measuring causal effects in 3.2. We also provide experiments to demonstrate the practical aspect of such derivations in 4.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: High-level explanations of how experiments were done are included in 4 as well as the derivation to implement in 3.2. We also provide an anonymized Github that contains the implementation of the derivations in 3 and results in 4. We cannot release the n2c2 or MIMIC-III datasets due to data usage agreements, but they are publicly available to anyone who accepts those agreements.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide an anonymized Github that contains the implementation of the derivations in 3 and results in 4. We cannot release the n2c2 or MIMIC-III datasets due to data usage agreements, but they are publicly available to anyone who accepts those agreements.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain our hyperparameter choices (following Mulyar et al. [2019]) in §3 and the details of our experiments (e.g., bootstrap confidence intervals, subgroup analyses) in §4.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We use nonparametric bootstrap resampling to generate 95% confidence intervals for measuring causal effects to provide additional context when explaining the effects of measurement error as well as quantify the significance our results in 4. However, our main contribution is the demonstration of our causal methodology applied to a real-world dataset, not a claim of improved model performance on a given benchmark.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The scale of experiments done in this paper do not require large numbers of CPUs and GPUs. Specifically, one 3070TI GPU was used to fine-tune the LLM and run experiments (as most relied on simple linear models) for the proposed methodology. The longest experiment (excluding model training) can be done in under an hour.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have adhered to the NeurIPS Code of Ethics (i.e. not releasing the n2c2 and MIMIC-III datasets) and preserved anonymity in both the paper and released code.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Although the practical application of imputing smoking status as unobserved confounder and seeing the measurement bias is an important contribution, the main focus of this paper is to propose a methodology that can be used for general causal methodologies, specifically seeing the effects of measurement bias when estimating unbiased casual effects in the presence of an unobserved confounder in a pre-existing relationship between two variables.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our proposed methodology does not pose such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have included the proper citations of the specific packages utilized throughout the paper and in our list of references.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not provide new assets.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing. We follow the data usage agreement for the MIMIC-III and n2c2 data.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing. We follow the data usage agreement for the MIMIC-III and n2c2 data, and do not require additional IRB approval.