Sensitivity Analyses for Incorporating Machine Learning Predictions into Causal Estimates

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze
{zach,ilyas,mdredze}@cs.jhu.edu
Johns Hopkins University, Baltimore, MD 21211

Abstract

Causal inference methods can yield insights into causation from observational datasets. When some necessary variables are unavailable for a causal analysis, machine learning systems may be able to infer those variables based on unstructured data such as images and text. However, if these inferred variables are to be incorporated into causal analyses, the error rate of the underlying classifier should affect the uncertainty in the causal conclusions. Past work has framed classifier accuracy as measurement error to incorporate predictions into consistently-estimated causal effects. However, this estimator is sensitive to small errors in its estimation of the classifier's error rate, leading to erratic outputs that are uninterpretable for any given analysis. In this paper we introduce three sensitivity analyses that capture the uncertainty from using a machine learning classifier in causal estimation, and show that our methods enable more robust analyses.

1 Introduction

Causal inference methods applied to large datasets have potential to help improve healthcare treatments [11] and inform policy decisions [45]. Such analyses are often limited by the available data; if an important variable is unobserved, unbiased estimation of causal effects may be impossible. Suppose we want to measure the causal effect of smoking on lung cancer. Such an analysis should adjust for variables like socioeconomic status (SES), which may be an important covariate [42]. However, suppose a dataset omits SES, but clinical notes are predictive of it [51, 37]. Rather than rely on humans to read clinical notes and infer SES labels in an entire dataset, we could infer these labels using supervised machine learning (ML).

ML methods are widely studied, and often demonstrate exceptional predictive accuracy, but such performance does not provide guarantees on the consistency of downstream analyses [33]. For a causal analysis that we hope can inform clinical decision-making, how accurate does the ML classifier need to be to produce a result we can trust? [20, 8]. We cannot expect a simple answer such as, "doctors should only trust analyses that use a classifier with greater than 95% accuracy." Instead, the error rate dictates what analyses are possible. We seek to connect classifier error to uncertainty in downstream causal analyses.

Our work follows [50], which proposed estimating causal effects using classifier outputs. The estimator is a function of the ML classifier's predictions and an estimate of its error rates, drawing on measurement error literature [36, 29]. We adopt this framework but use simulation studies to show this estimator performs poorly in many finite-sample cases. We introduce three sensitivity analyses to quantify the uncertainty of this estimator. We evaluate the coverage properties of these methods on our synthetic datasets and show they enable more robust analyses. We demonstrate our methods and discuss how their assumptions map onto a specific Twitter dataset and ML classifier.

Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020

2 Background and motivation

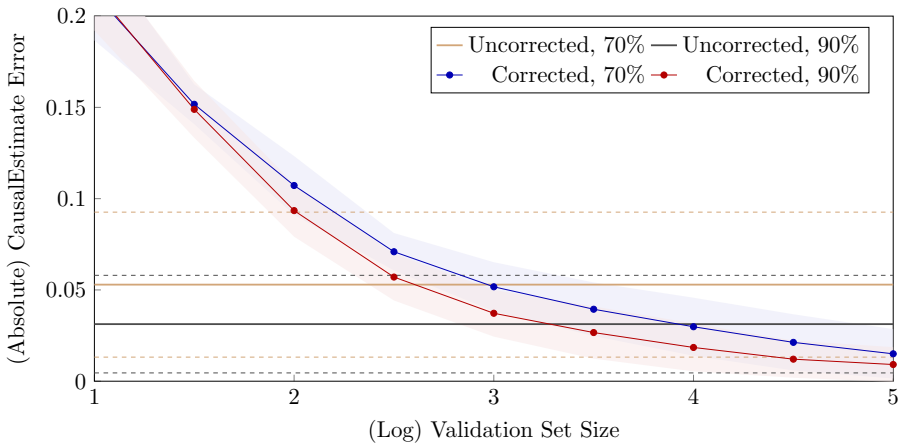


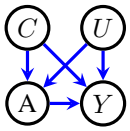
Figure 1: The empirical motivation for this work. Ground truth causal effect is at $y{=}0$. X-axis shows log size of the validation set used to estimate $p(U^*|U)$, the error rate of the classifier. Each line shows the causal error for an estimator that relies upon a classifier with fixed accuracy (either 70% or 90% accurate). For the corrected estimators which rely upon $p(U^*|U)$, causal error decreases as the validation size increases. When the validation set is too small, a naive uncorrected estimator outperforms a theoretically-sound corrected estimator. Experimental details are in §4

As motivation, consider a retrospective analysis of electronic health records (EHR) that seeks to estimate the causal effect of smoking on lung cancer. Variables representing possible common causes of smoking and cancer, such as SES, should be included in the statistical analysis. However, in many EHR datasets SES may not be explicitly recorded, but rather only indirectly noted in free-text clinical notes [28, 51]. To avoid the cost and privacy concerns of human annotators reading clinical notes and inferring structured variables, we need methods that harmonize with a valid causal analysis.

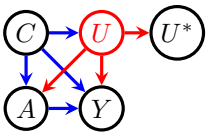
Causal inference uses observational data to reason about hypothetical interventions; in our example, "did choosing not to smoke prevent cancer?" Causal models and their requisite assumptions are often represented in directed acyclic graphs (DAGs) [35]. In such a model, we can connect counterfactual random variables to observed random variables with assumptions. Assuming all relevant variables are observed, we can use the *g-formula* to write a causal effect as a function of observed data [39]. In our example, suppose the only common causes of smoking (A) and lung cancer (Y) are genetics (C) and SES (U). Then Figure 2a represents a causal DAG, and the counterfactual $Y(a)$, meaning "cancer status if smoking, possibly contrary to fact, had been set to a," is identified via the g-formula as $E[Y(a)]{=}\sum_{c,u} E[Y\,|\,A{=}a,c,u]p(c,u)$. If U is never observed and we cannot infer it, then $E[Y(a)]$ is *not identified* and we cannot proceed with a causal analysis. We are interested in cases in which we do not observe U but an ML classifier can produce a noisy proxy variable, U^*, for U. Figure 2b shows a causal DAG in which the U is unobserved but we have access to a proxy U^*. While the g-formula cannot be used in this model, we can identify the counterfactual $Y(a)$ from Figure 3 [36].

Questions of mismeasured data and measurement error have been widely studied in diverse fields. Measurement error has been a central concern in epidemiology research for decades [15, 48, 7, 26]. The statistics literature has also considered questions of measurement error, from parametric models [44] to semiparametric models [43, 47]. Most relevant to our work is [54], which considers causal inference under unobserved confounding when that confounder is observed in a validation set[1]. Measurement error concerns arise in the use of structural equation models across disciplines [17, 3]. Information systems (IS) and management science has modeling noisy measurements, e.g. in studying consumer preferences [14]. Recent IS work has also considered a setting similar to ours, where measurement error occurs due to machine learning model predictions [52, 53]. In this work,

[1] A primary difference in our methods is that our approach is applicable (under a non-differential error assumption) when our validation set only contains U and U^*.



(a) Simple Confounding



(b) Measurement Error

Figure 2: Causal DAGs. In (b), missing arrows to U^* from C, A, Y represents a non-differential error assumption (see Appendix D). Throughout this paper, we assume for simplicity of presentation that all variables are binary, though the measurement error correction only relies on U being discrete. Recent work has explored recovering from measurement error more generally [29, 41].

$$\tau_{U^*} = \sum_{c,u} \left[\frac{p(c,u^*) - \epsilon_u P(c)}{1 - \epsilon_u - \delta_u} \cdot \left(\frac{p(Y=1,u^*|A=1,c) - \epsilon_u p(Y=1|A=1,c)}{p(u^*|A=1,c) - \epsilon_u} - \frac{p(Y=1,u^*|A=0,c) - \epsilon_u p(Y=1|A=0,c)}{p(u^*|A=0,c) - \epsilon_u} \right) \right]$$

Figure 3: Estimand (τ_{U^*}) for the causal effect of A on Y in the DAG given in Figure 2b. Define $\epsilon_u = p(U^* = u | U \neq u)$ and $\delta_u = p(U^* \neq u | U = u)$. All variables are assumed binary for simplicity of presentation. Derivation is in Appendix A.

we draw from recent work on non-parametric identification of causal effects under measurement error [36, 25, 34].

Our work most closely follows that of [50], which considered measurement error to account for errors[2] produced by natural language processing (NLP) classifiers. Throughout, we will assume that we have access to a classifier f that produces U^* with some (unknown) error distribution $p(U^*|U)$. We will also assume we have a small validation dataset with full data on $p(U)$ which we can use to estimate $p(U^*|U)$. Then, we have a large dataset without U on which we can apply our classifier f to produce a dataset $p(C, A, Y, U^*)$. Using the 'effect restoration' approach proposed by [36], we can then estimate our desired causal effects.

The theoretically-sound estimator used by [50], however, has a counter-intuitive empirical trend. It can be outperformed in practice, even at large finite samples, by a naive estimator that assumes $U^* = U$ [34]. To understand this limitation, we conceptualize their method as a two step approach: the classifier which produces a $p(U^*, C, A, Y)$ distribution and a 'corrector' which estimates $p(U^*|U)$ and adjusts the causal estimate. Each of these steps is imperfect, e.g. due to finite sample variability. The classifier error is how often the predicted U^* differs from the true U, and depends on the size of the training data. The corrector error is the difference between the true error rate $p(U^*|U)$ and our estimate of that error rate from the examples in our validation dataset.

Using simulation studies that we will discuss in §4, we show in Figure 1 that the method fails if and only if the corrector step fails. If we have low corrector error, our causal estimate will be accurate; and regardless of classifier accuracy, high corrector error will bias our estimates. Thus, we should seek to quantify the uncertainty of the corrector step. Rather than trust an uncertain point estimate, we want reliable bounds on the causal effect. We introduce three sensitivity analyses that propagate uncertainty from the corrector step to our final causal estimate.

3 Sensitivity analyses for the error estimate

A causal analysis typically outputs a parameter estimate that reflects some real-world phenomenon which may be impossible to further validate. If we know our estimator can be unreliable under certain conditions, how do we know when to trust that a causal estimate is accurate?

Consider a plug-in estimator for τ_{U^*} identified via the functional in Figure 3 (with the derivation following [36] shown in Appendix A). The 'corrector' step of this estimator involves dividing by an

[2]While [50] considered a mismeasured treatment, our sensitivity analyses apply generally.

estimate of the error term; small changes to that estimate may result in large changes to the overall causal estimate. A sensitivity analysis for the estimate of the error rate allows us to explore how the final estimate would change as the error rate estimate changes. Rather than accepting a point estimate as our 'best guess,' our uncertainty in $p(U^*|U)$ should be represented in an interpretable manner. To explore ways to make these methods more robust and interpretable, we introduce three sensitivity analyses that can capture the uncertainty in the $p(U^*|U)$ estimate.

Each of our sensitivity analyses will introduce a sensitivity parameter, γ, which controls the trade-off between the width and coverage of our intervals[3]. Our methods are designed such that $\gamma=0$ returns a point estimate (no interval) and as we increase γ, the interval widens. As our outcome Y is binary, a maximally-wide interval for τ_{U^*} spans from -1 to 1.

An alternative to the sensitivity analysis approach taken here is to obtain confidence intervals for τ_{U^*} using ideas from the post-selection inference literature [2, 38, 27]. Popular existing methods of this type have often been applied in parametric settings, and do not translate in a straightforward way to the setting we consider here, where estimation of functionals corresponding to causal effects does not employ parametric nuisance models. The next three subsections discuss theoretical and empirical trade-offs of each sensitivity analysis; we evaluate each on synthetic data in § 4.

Bootstrap resampling Our first approach to a sensitivity analysis draws from non-parametric bootstrap [18]. A classical statistical approach would bootstrap the entire estimator, retraining ML methods many times on many resampled datasets for the estimate of the causal estimand. Given that many ML models can take from hours to months to train, bootstrapping our entire analysis from training data to causal estimate is unrealistic [24, 12]. But it is computationally easy to calculate our estimate of the error rate $p(U^*|U)$ and use it to compute our estimand.

Our bootstrap sensitivity analysis involves resampling k validation datasets of the same size as the original validation dataset. On each bootstrapped validation set, we calculate our error distribution $p(U^*|U)$ and use it to compute τ_{U^*}. This method gives us k τ_{U^*} estimates. To build the curve in Figure 4, we plot the intervals given by the middle $\gamma_{\text{Bootstrap}}\%$ of these k estimates; when $\gamma_{\text{Bootstrap}}$ = 0, our interval has width 0, and when $\gamma_{\text{Bootstrap}}=100$ our interval endpoints are the min and max of these k estimates. For our experiments, we also set $k=100$ and find that this allows for good coverage of our causal effect in simulated studies. Depending on the practical setting, good coverage may be possible with a lower k or may require a larger k. A disadvantage of this method is that it requires full access to the validation set. If we are using a classifier validated on data that is proprietary, private or otherwise inaccessible, we would need to collect a new validation set.

Binomial sampling of error rates Our second sensitivity analysis again relies on sampling, but instead samples synthetic error rates from a binomial distribution. Our point estimate of the error rate $p(U^*=u'|U=u)$, assuming U is binary, consists of counting which validation set examples our classifier got right. We can model this as a binomial distribution $B(n,p)$, where 'success probability' p is our point estimate of $p(U^*=c'|U=c)$ and the 'number of trials' n is the number of validation set examples where $p(U=c)$. Our sensitivity analysis samples $n\tilde{p}\sim B(n,p)$ and constructs a new error rate $p(U^*|U)=\tilde{p}$, which we use to calculate our estimand τ_{U^*}.

To construct an interval from this approach, we sample k such error rates[4] and use the 2.5 and 97.5 percentiles as the bounds of our interval. To trade off between width and coverage in Figure 4, we redo this method several times, replacing the true validation dataset size n with a smaller, "synthetic sample size," n'. As this n' decreases, the variability in the sampled error rates increases, which widens the resulting interval. We let $n'=n^{\,1/(1+\gamma_{\text{Binomial}})}$; as γ_{Binomial} increases from zero, the "synthetic sample size" decreases from n. This approach has the advantage of not requiring access to the full validation set; in addition to the $p(U^*=u'|U=u)$ point estimate, we only need to know n, the number of validation set examples that estimate was computed on. However, this method makes the assumption that each error rate is binomially-distributed around its mean.

Clopper-pearson confidence interval Our third approach also makes a binomial assumption about the error rates. We can replace each $p(U^*|U)$ estimate with an interval, and then propagate the uncertainty of each interval into the final causal estimate. Because our error rates are proportions,

[3]Each method also uses, but is empirically robust to changes in, a hyperparameter k.

[4]We find that $k=100$ allows for intervals with good coverage in our experiments.

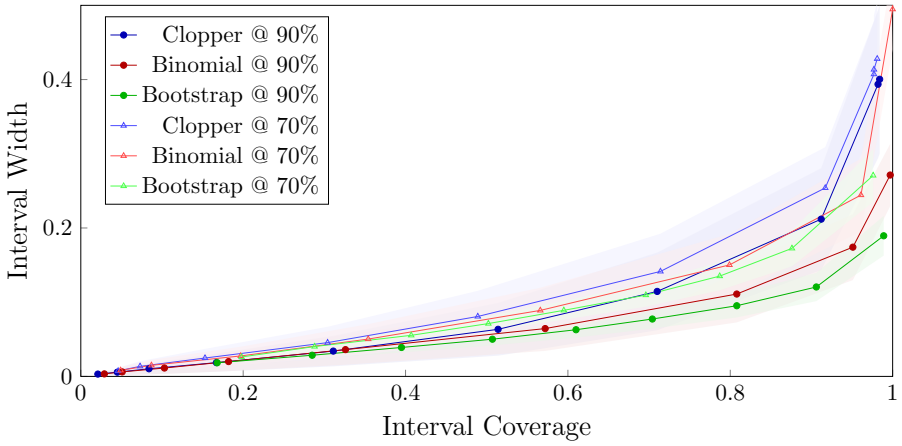


Figure 4: Comparison of proposed sensitivity analyses for the corrector step. Clopper refers to the Clopper-Pearson confidence interval, Binomial to the binomial sampling of error rates, and Bootstrap to a non-parametric bootstrap resampling. Noisy-oracle classifier has either 70% or 90% accuracy. As the sensitivity analyses' hyperparameter increases, both width and coverage increase. Bootstrap provides the best trade-off between coverage and width.

one reasonable choice of interval is the Clopper-Pearson method [10]. If we have computed a point estimate of $p(U^*{=}c'|U{=}c){=}p$ using n examples in our validation set, we compute a 95% Clopper-Pearson interval as the 2.5 and 97.5 percentiles of a Beta distribution with parameters $(np, n-np+1)$ and $(np+1, n-np)$. To propagate this interval for our error rate to an interval for final our causal estimand, we chosen $k{=}20$ evenly-spaced values along the interval for each error rate, and then use all combinations of those error-rate values to estimate τ_{U^*}. Finally, we use the min and max of these resulting estimates as the bounds of our interval. To trade off between width and coverage in Figure 4, we take the same approach as for the binomial sampling method, replacing the true n with a smaller $n'{=}n^{1/(1+\gamma_{\text{Clopper}})}$. As γ_{Clopper} increases from zero, the resulting interval widens. The method has the same advantage as the binomial approach in that it only requires knowing the size of the validation set.

4 Synthetic experiments

We conduct several simulation studies on synthetic datasets to explore the behavior of the measurement error estimator with and without our sensitivity analyses. The goal of these experiments is to understand where past work fails, and how our proposed analyses demonstrate improvements in the robustness and interpretability of the causal estimates. We first need to parameterize the distributions from which our synthetic datasets are drawn. We build on top of the code released by [50], but allow for much broader evaluations. Rather than limiting ourselves to a single data-generating distribution, we can sample arbitrarily-many $p(C, A, Y, U)$ distributions and then sample data from each. In our experiments, we evaluate each method on ten different distributions. To make comparisons more consistent, we restrict $p(C, A, Y, U)$ such that the true causal effect of A on Y is equal to 0.1. We release our code[5] to enable future analyses.

Synthetic evaluation of τ_{U^*} We now return to the estimator proposed in past work, and re-examine Figure 1 to show how it fails in certain settings. We split the estimator into two steps, the classifier which produces $p(U^*, C, A, Y)$ and the corrector which estimates $p(U^*|U)$ and adjusts the causal estimate. To highlight the sensitivity of the corrector, we replace the classifier with a 'noisy oracle' with a fixed classification accuracy. This means that in $p(U^*, C, A, Y)$, U^* takes the same value as U with a fixed probability, regardless of all other variables. We then sample a validation dataset from

[5] `https://github.com/zachwooddoughty/cdml20_sensitivity`

$p(U, C, A, Y)$ and use it to estimate the error rate $p(U^*|U)$. As our validation dataset grows, we should expect our error rate estimate to converge to the true accuracy.

Figure 1 shows overall causal estimate error as we increase the size of the validation set. We compare the τ_{U^*} (Corrected) estimator against a naive (Uncorrected) estimator that assumes $U=U^*$. We plot both such estimators for a noisy oracle classifier with two accuracies: 70% and 90%. The Uncorrected estimator ignores the validation set entirely, so has constant error. The Corrected estimator improves with additional validation data, but can perform worse than the Uncorrected estimator.

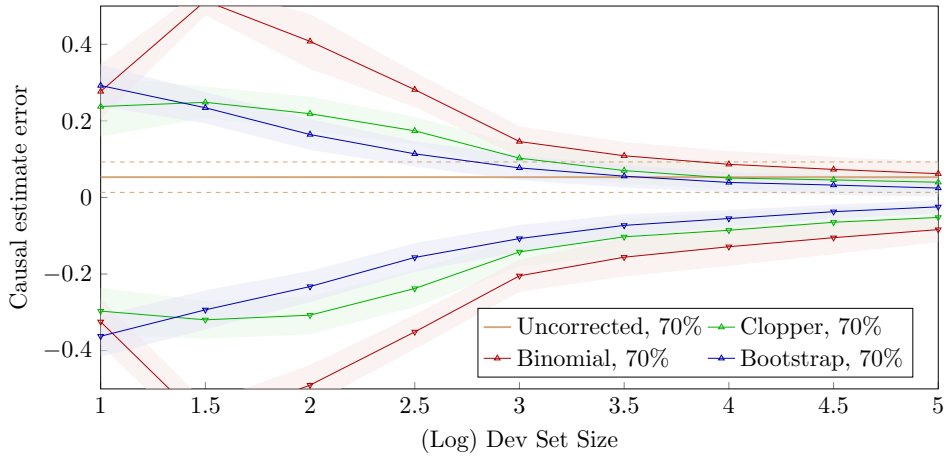


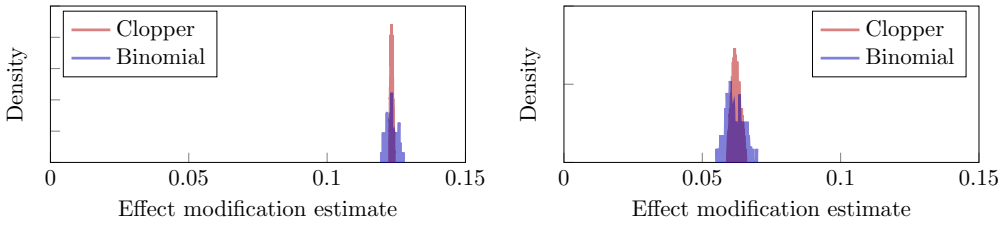
Figure 5: Our sensitivity analyses provide upper/lower bounds that capture the correction step's uncertainty. The true causal effect is at $y=0$. All experiments consider data with a known classifier accuracy (70%); the validation set is only used to estimate classifier error rates. Each line and shaded bounds represent the mean/stdev calculated from 100 simulations on 10 distributions. Uncorrected estimator bounds are represented by the dashed line. In the limit our bounds converge to the truth.

This motivates our need for a sensitivity analysis. If we use such an estimator and wish to draw real-world conclusions from a causal estimate, we need the ability to trust the robustness and reliability of that estimate. If the theoretically-consistent Corrected estimator and the naive Uncorrected estimator disagree, what can we do? We need our sensitivity analyses to inform our degree of uncertainty in the final causal effect.

We now provide a synthetic evaluation of our three proposed sensitivity analyses. Each of the proposed methods have advantages and disadvantages, which may change how they perform. Each method produces bounds on the final causal effect; we want to understand the width of those bounds and whether those bounds contain the true effect. Figure 4 investigates the trade-off between interval width and coverage of the true causal effect. Each analysis is run 100 times on each of ten different distributions, and we calculate the mean and standard deviation across those ten distributions.

For the bootstrap method, we take $k=100$ bootstrap resamples and sweep over $\gamma_{\text{Bootstrap}}$ to the intervals given by truncating the empirical bootstrap distribution at different percentiles. Each percentile produces an upper and lower bound, which corresponds to a single dot in Figure 4, highlighting the width and coverage of those bounds. For our sampling experiments, we sweep over γ_{Binomial}, sampling $k=100$ error rates to produce an interval for each effective validation size n'. For our interval method, we sweep over γ_{Clopper} and calculate $k^2=400$ causal effects for each n' to produce our interval.

It is trivial get 100% coverage with an interval that covers all possible causal effects. It is also trivial to have a perfectly narrow interval but that has 0% coverage. We see two expected trends: as we increase the oracle classifier accuracy from 70% to 90%, all methods improved dramatically; as we increase the hyperparameter for each method, we see coverage increase but interval width also increases. With a 90% accurate classifier, we relatively small differences between the different sensitivity analyses; at 70% accuracy, Bootstrap demonstrates the best performance. For real without known causal effects, we cannot know how interval coverage varies with its width. A robust analysis should use domain knowledge to inform such sensitivity analyses; see §6 for more discussion.



(a) Plot of our estimates of the gender effect modification on #VaccinesWork tweets. All estimates vary between 0.119 and 0.128.

(b) Plot of our estimates of the gender effect modification on #CDCWhistleBlower tweets. All estimates vary between 0.055 and 0.070.

Figure 6: Estimating τ_{U*} using our approach on our Twitter datasets.

We now combine our proposed sensitivity analyses into the overall estimator to show how our methods can improve downstream analyses. Figure 5 combines our proposed methods into the analyses previously shown in Figure 1. Using a noisy oracle estimator with 70% accuracy we estimate the error rate on a validation set. We use our analyses to produce bounds for our causal estimate. For all validation set sizes, our upper and lower bounds contain the truth, and the interval width provides an interpretable quantification of our uncertainty. As the validation set size increases, our bounds tighten around the truth.

When conducting a downstream analysis, our hyperparameters let us control the trade-off between width and coverage. If we examine a simulation study that attempts to mirror a real-world dataset, a sensitivity analysis on the synthetic data can inform our interpretations with real data.

5 Causal effect modification of gender in Twitter data

We now use a real-data analysis to demonstrate how our methods can inform the reliability of a causal effect analysis in a complex domain. Social media, like EHR notes, is noisy but a valuable source of data. We focus in particular on public perception of vaccines, as such opinions are critically important to public health and have been influentially studied in social media [13]. Looking at users who tweet about vaccines, we ask: Does the perceived gender of a tweet's author affect the relationship between popularity and engagement?

From a Twitter stream spanning 2014 to 2019, we collect 21k tweets using the pro-vaccine #VaccinesWork hashtag, and 1.2k tweets with the anti-vaccine #CDCWhistleBlower. We detail the pre-processing steps to Appendix B, including robustness checks. Our outcome Y is the whether a tweet receives at least one like, our treatment A is if the author has $300+$ followers[6]. The author's verification status is an always-observed confounder, C. Our U^* is the binary gender prediction of a demographics classifier [23], which is a noisy proxy for the perceived gender of the author U. If the effect of A on Y varies with the value of U, we say U 'modifies the effect' of A on Y [22].

Tables 1(a-b) summarizes the data. For #VaccinesWork tweets, popular users clearly receive more engagement than unpopular users, and women receive more engagement than men. For #CDCWhistleBlower tweets, both such statements are true marginally, but tweets by unpopular women have 3% *more* engagement than those by popular women. However, these tables do not factor in the uncertainty of our gender classifier, which we now consider with our sensitivity analysis.

6 Sensitivity analysis of gender effect modification

We conduct a sensitivity analysis of the Twitter data that accounts for the error rate of our gender classifier to estimate the effect modification of gender. The estimand for effect modification is simply a rearrangement of terms of τ_{U*}, giving the difference between two conditional causal effects: does high-popularity increase engagement more for men than it does for women?

[6]These thresholds are picked in Appendix B and their impact is discussed in Appendix C.

	Women	Men	Total
Unpopular	56.75	41.99	50.84
Popular	53.77	54.31	54.02
Total	55.47	48.05	52.29

(a) $E[Y=1|A,U^*]$; probability that #CDCWhistleBlower users receive 1+ likes.

	Women	Men	Total
Unpopular	41.96	35.87	39.17
Popular	53.79	49.09	51.49
Total	47.70	42.72	45.34

(b) $E[Y=1|A,U^*]$; probability that #VaccinesWork users receive 1+ likes.

Table 1: The percent of tweets that receive at least one like, stratified by classified gender and popularity. Table 2 (Appendix) shows the joint distribution of gender and popularity (cutoff set at 300 followers.) Each user is assigned the maximum likelihood gender label, and probabilities shown are marginalized over verification status. Datasets are created after all preprocessing steps listed in § 5.

The perceived gender classifier released by [23] was estimated to have error rates of $p(U^*=1|U=0)=17.0\%$ and $p(U^*=0|U=1)=19.3\%$ on a validation set of 52k users. As discussed in §3, our sensitivity analysis requires a hyperparameter choice that encodes some prior uncertainty. We choose $\gamma_{\text{Bootstrap}}$ and γ_{Clopper} following the synthetic results in § 4, which equates to an effective validation size of 900 users. While the public nature of Twitter data means we could in fact re-collect this dataset and use our bootstrap method, we do not for the purposes of this analysis. In many real-world cases, we could not have full access to the validation set.

A histogram of the outputs of both methods are shown in Figures 6a and 6b. Interpreting these plots relies upon connecting our domain knowledge to our methodology. How should an analyst choose γ for these sensitivity analyses? If our binomial sampling method produces an outlier estimate of -0.5, can we safely disregard it? Answering such questions in any applied analysis relies on domain knowledge. Suppose previous work suggests that a particular classifier's error rate varies greatly by domains, then we may want to choose a conservatively large hyperparameters for our analyses. In contrast, if we are confident that past work has established the true classifier accuracy, we may be willing to trust tighter bounds. Our Twitter data and sensitivity analyses give very tight bounds, but we could further validate our approach by collecting another validation set on our vaccine-specific data.

While parametrizing uncertainty in our analyses is certainly helpful, it does not obviate the need to draw on domain expertise for interpretability. Past public health research has found conflicting results on whether gender plays a significant role in vaccine skepticism or decision-making [31]. However, the existence of vaccination gender disparities [1] and the need to effectively communicate to diverse audiences [9, 32] necessitates further study of gender differences in vaccination trends.

Analysis limitations Our Twitter case study demonstrated the efficacy of our sensitivity analysis, but we caution against drawing conclusions about users' behavior. We estimate that the effect modification is robust to bias from the gender classifier, but not other assumptions in our analysis: gender is not binary and we do not differentiate between perceived and self-identified gender [19, 16, 5]. While conceptualizing the inherent uncertainty of gender prediction in a measurement error framework is better than taking its predictions as truth, but could still cause harm if used to misgender individual users [21]. Second, data processing details can change the outcome, e.g. removing retweets and prolific users or how we model follower-counts and like-responses. For example, follower-count and like-response are only noisy proxies of an underlying concept of user and tweet status. We may have added new bias by artificially binning these two variables into binary categories. While we could bin these into a larger number of discrete variables, our matrix adjustment approach needs additional assumptions for real-valued distributions [29]. Third, we restrict our analysis to a three-variable causal model, an over-simplification of social media behaviors. There are likely other unobserved confounders, such as malicious foreign actors [4]. Additionally, if the classifier error rate is correlated with user popularity, our correction step may introduce new biases. Finally, we treat tweets as independent samples, ignoring network effects of the platform which may correlate user behaviors.

Many of these implicit assumptions are unrealistic and likely introduce some bias into our conclusions, though such assumptions are ubiquitous in social media analyses. We accept those assumptions to focus on explicitly addressing measurement error induced by an imperfect classifier.

7 Conclusions

We have presented a new measurement error formulation that provides a means for incorporating estimated errors of ML classifiers into a causal analysis framework. Our formulation provides a more robust framework for reaching causal conclusions using classifier predictions. This work creates new opportunities for the analysis of causal factors in a variety of domains, including in computational social sciences that rely on analyses of high-dimensional data such as text. We highlight our methods on synthetic and real that highlights the interpretability provided by our sensitivity analyses.

There are several directions along which further work can extend our framework. The sensitivity analysis could be extended to work with non-binary classifications, high-dimensional C or U vectors, differential measurement error, or interference. Each of these research directions would expand the real-world applications of our methods, and would benefit from our contributions.

8 Broader impact

We have introduced three sensitivity analyses for understanding the uncertainty of methods that rely on machine learning model predictions to estimate causal effects.

Many existing ML applications need causal reasoning to inform possible interventions [45]. There are many examples of predictive models performing well in an experimental setting but poorly in real-world settings [46]. Our methods provide a generally-applicable method for quantifying the uncertainty that is introduced into causal analyses that depend upon a trained machine learning model.

A primary implication of our methods is that they enable causal analyses that rely on trained machine learning models. ML methods trained on high-dimensional features in large datasets can extract information at a scale that cannot be matched by human experts. If such models, e.g. in medical image analysis [40] or in processing EHR notes [37], can be incorporated into a principled causal framework, many new analyses will be possible.

Despite this promise, we must be wary of overconfidence in ML methods that perform well in experiments but poorly in practice. Our methods make assumptions about the relationship between a machine learning classifier's models and the variables relevant to the ultimate causal analysis. If a causal analysis informs healthcare policy, it must address broader concerns of fairness and transparency [30]. If a classifier's errors are disproportionately affect demographic groups, the resulting biases may go unnoticed unless an effort is made to look for them.

There are several initiatives that could guide this line of work towards better societal impacts. From a technical side, we could benefit from new connections between theoretical convergence rates of estimators and the empirical bounds we focus on in this work. In practical settings, we need ways to calibrate methods to optimize for societal goals. If a deployed model informs medical interventions, we need clearly-established guidelines for evaluating the costs of over- or under-confidence of bounds on a causal estimate. Such initiatives require the collaboration of ML researchers and the domain experts who may hope to apply their models.

9 Acknowledgements

The authors thank Noam Finkelstein, Katie Keith, Jaron Lee, Dan Malinsky, Betsy Ogburn, Eli Sherman, Adarsh Subbaswamy, and several sets of anonymous reviewers for their insights and helpful comments. This work was sponsored in part by Office of Naval Research grant N00014-18-1-2760, by National Institute of Allergy and Infectious Diseases grant R01 AI127271–01A1 and by National Institute of General Medical Sciences grant 5R01GM114771.

References

[1] C. Abat and D. Raoult. Human papillomavirus vaccine: Urgent need to promote gender parity. *European journal of epidemiology*, 33(3):259–261, 2018.

[2] R. Berk, L. Brown, A. Buja, K. Zhang, L. Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[3] K. A. Bollen, K. M. Gates, and Z. Fisher. Robustness conditions for miiv-2sls when the latent variable or measurement model is structurally misspecified. *Structural equation modeling: a multidisciplinary journal*, 25(6):848–859, 2018.

[4] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *AJPH*, 108(10):1378–1384, 2018.

[5] J. Butler. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre journal*, 40(4):519–531, 1988.

[6] J. S. Butler, R. V. Burkhauser, J. M. Mitchell, and T. P. Pincus. Measurement error in self-reported health variables. *Review of Economics and Statistics*, 69(4):644–650, 1987.

[7] R. J. Carroll, D. Ruppert, C. M. Crainiceanu, and L. A. Stefanski. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[8] J. H. Chen and S. M. Asch. Machine learning and prediction in medicine – beyond the peak of inflated expectations. *NEJM*, 376(26):2507, 2017.

[9] T. Chen and M. Dredze. Vaccine images on twitter: Analysis of what images are shared. *JMIR*, 20(4), 2018.

[10] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, pages 404–413, 1934.

[11] I. J. Dahabreh, J. M. Robins, S. J. Haneuse, and M. A. Hernán. Generalizing causal inferences from randomized trials: counterfactual and graphical identification. *arXiv*, 2019.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018.

[13] M. Dredze, D. A. Broniatowski, M. C. Smith, and K. M. Hilyard. Understanding vaccine refusal: why we need social media now. *American journal of preventive medicine*, 50(4):550–552, 2016.

[14] J. Eliashberg and J. R. Hauser. A measurement error approach for modeling consumer risk preference. *Management Science*, 31(1):1–25, 1985.

[15] J. Fleiss and P. Shrout. The effects of measurement errors on some multivariate procedures. *American journal of public health*, 67(12):1188–1191, 1977.

[16] H. Frohard-Dourlent, S. Dobson, B. A. Clark, M. Doull, and E. M. Saewyc. "i would have preferred more options": accounting for non-binary youth in health research. *Nursing inquiry*, 24(1):e12150, 2017.

[17] R. Grewal, J. A. Cote, and H. Baumgartner. Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing science*, 23(4):519–529, 2004.

[18] P. Hall. Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, pages 927–953, 1988.

[19] F. Hamidi, M. K. Scheuerman, and S. M. Branham. Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems. In *CHI*, page 8. ACM, 2018.

[20] H. Jiang, B. Kim, M. Guan, and M. Gupta. To trust or not to trust a classifier. In *NeurIPS*, pages 5541–5552, 2018.

[21] O. Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

[22] M. J. Knol and T. J. VanderWeele. Recommendations for presenting analyses of effect modification and interaction. *International journal of epidemiology*, 41(2):514–520, 2012.

[23] R. Knowles, J. Carroll, and M. Dredze. Demographer: Extremely simple name demographics. In *NLP+CSS*, pages 108–113, 2016.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[25] M. Kuroki and J. Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

[26] S. le Cessie, J. Debeij, F. R. Rosendaal, S. C. Cannegieter, and J. P. Vandenbrouckea. Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*, pages 551–560, 2012.

[27] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[28] K. H. McVeigh, R. Newton-Dame, P. Y. Chan, L. E. Thorpe, L. Schreibstein, K. S. Tatem, C. Chernov, E. Lurie-Moroni, and S. E. Perlman. Can electronic health records be used for population health surveillance? validating population health metrics against established survey data. *eGEMs*, 4(1), 2016.

[29] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[30] R. Nabi, D. Malinsky, and I. Shpitser. Optimal training of fair predictive models. *arXiv preprint arXiv:1910.04109*, 2019.

[31] J. M. Nagata, I. Hernández-Ramos, A. S. Kurup, D. Albrecht, C. Vivas-Torrealba, and C. Franco-Paredes. Social determinants of health and seasonal influenza vaccination in adults over 65 years: a systematic review of qualitative and quantitative data. *BMC Public Health*, 13(1):388, 2013.

[32] X. Nan. Communicating to young adults about hpv vaccination: Consideration of message framing, motivation, and gender. *Health Communication*, 27(1):10–18, 2012.

[33] Z. Obermeyer and E. J. Emanuel. Predicting the future – data, machine learning, and clinical medicine. *NEJM*, 375(13):1216, 2016.

[34] H. Oktay, A. Atrey, and D. Jensen. Identifying when effect restoration will improve estimates of causal effect. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 190–198. SIAM, 2019.

[35] J. Pearl. *Causality*. Cambridge university press, 2009.

[36] J. Pearl. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432. AUAI Press, 2010.

[37] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

[38] S. Reid, J. Taylor, and R. Tibshirani. Post-selection point and interval estimation of signal sizes in gaussian samples. *Canadian Journal of Statistics*, 45(2):128–148, 2017.

[39] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period with application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[40] D. Shen, G. Wu, and H.-I. Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[41] X. Shi, W. Miao, J. C. Nelson, and E. J. Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[42] J. Siemiatycki, S. Wacholder, R. Dewar, E. Cardis, C. Greenwood, and L. Richardson. Degree of confounding bias related to smoking, ethnic group, and socioeconomic status in estimates of the associations between occupation and cancer. *Journal of occupational medicine*, 30(8):617–625, 1988.

[43] S. Sinha and Y. Ma. Semiparametric analysis of linear transformation models with covariate measurement errors. *Biometrics*, 70(1):21–32, 2014.

[44] L. A. Stefanski and R. J. Carroll. Covariate measurement error in logistic regression. *The Annals of Statistics*, pages 1335–1351, 1985.

[45] A. D. Stern and W. N. Price. Regulatory oversight, causal inference, and safe and effective health care machine learning. *Biostatistics*, 2019.

[46] A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127, 2019.

[47] X. Wang and Q. Wang. Semiparametric linear transformation model with differential measurement error and validation sampling. *Journal of Multivariate Analysis*, 141:67–80, 2015.

[48] W. Willett. An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Statistics in Medicine*, 8(9):1031–1040, 1989.

[49] Z. Wood-Doughty, P. Mahajan, and M. Dredze. Johns hopkins or johnny-hopkins: Classifying individuals versus organizations on twitter. In *PEOPLES*, pages 56–61, 2018.

[50] Z. Wood-Doughty, I. Shpitser, and M. Dredze. Challenges of using text classifiers for causal inference. In *EMNLP*, pages 4586–4598, 2018.

[51] C.-Y. Wu, C.-K. Chang, D. Robson, R. Jackson, S.-J. Chen, R. D. Hayes, and R. Stewart. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PloS one*, 8(9), 2013.

[52] M. Yang, G. Adomavicius, G. Burtch, and Y. Ren. Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, 29(1):4–24, 2018.

[53] M. Yang, E. McFowland, G. Burtch, and G. Adomavicius. Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *Kelley School of Business Research Paper*, (19-20), 2019.

[54] S. Yang and P. Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, pages 1–33, 2019.

A Derivations of estimand

Simple Confounding

Assume A is a treatment, C and U are confounders, and Y is an outcome. The causal effect of A on Y can be thought of a difference in *counterfactual* probabilities. In the canonical case of smoking and cancer, if our 'treatment' A is whether you smoke, our outcome Y is cancer, and our confounders C and U are genetics and socioeconomic status, then we define the counterfactual random variable $Y(a)$ as an individual's hypothetical cancer outcome had they randomly been *assigned* to smoke without regard to their socioeconomic status.

The causal effect of smoking on cancer, then, is $E(Y(1))-E(Y(0))$; the population-level increased risk of cancer if everyone in the population had been assigned smoking as the treatment in a randomized control trial (RCT).

We define $E[Y(a)]$ as follows:

$$\begin{aligned} E[Y(a)] &= \sum_{c,u} E[Y(a)|c,u]p(c,u) \\ &= \sum_{c,u} E[Y(a)|A{=}a,c,u]p(c,u) \qquad (1) \\ &= \sum_{c,u} E[Y|a,c,u]p(c,u) \qquad (2) \end{aligned}$$

Equation 1 is true because of 'conditional ignorability.' That is, for a counterfactual variable $Y(a)$, it is independent of the factual A given the treatment was assigned interventionally. In graphical terms, that $Y(a) \perp A|C$.

Equation 2 is true due to 'consistency.' We assume that for individuals for whom $A{=}a$ in observational data, if we assign treatment a in a hypothetical RCT, then the distribution over Y is the same in the observational data and in the hypothetical data.

Our causal effect is then defined as:

$$\tau_U = \sum_{c,u} p(c,u) \Big(E[Y{=}1|A{=}1,c,u] - E[Y{=}1|A{=}0,c,u] \Big) \qquad (3)$$

Measurement Error

The derivation becomes more complicated with measurement error. We can follow Equation (7) from [36] to see how to define $p(U,C,A,Y)$ in terms of our observed $p(U^*,C,A,Y)$ and our error rates. For the two error rates our classifier might make, define $\epsilon_u=p(u^*|U\neq u)$ and $\delta_u=p(U^*\neq u|U=u)$, and note $\epsilon_0+\delta_0=\epsilon_1+\delta_1$.

$$\begin{aligned} &p(Y=y,A=a,c,u) \\ &\quad=\frac{p(Y=y,A=a,c,u^*)(1-\epsilon_u)}{1-\epsilon_u-\delta_u}-\frac{p(Y=y,A=a,c,U^*\neq u)\epsilon_u}{1-\epsilon_u-\delta_u} \\ &\quad=\frac{p(Y=y,A=a,c,u^*)-\epsilon_c p(Y=y,A=a,c)}{1-\epsilon_u-\delta_u} \end{aligned} \tag{4}$$

Then,

$$\begin{aligned} p(Y=y|A=a,c,u)&=\frac{p(Y=y,A=a,c,u)}{\sum_y p(Y=y,A=a,c,u)} \\ &= \frac{p(Y=y,A=a,c,u^*)-\epsilon_u p(Y=y,A=a,c)}{\sum_y p(Y=y,A=a,c,u^*)-\epsilon_u p(Y=y,A=a,c)} \\ &= \frac{p(Y=y,u^*|A=a,c)-\epsilon_u p(Y=y|A=a,c)}{p(u^*|A=a,c)-\epsilon_u} \end{aligned} \tag{5}$$

$$\begin{aligned} p(c,u)&=\sum_{a,y} p(Y=y,A=a,c,u) \\ &=\sum_{a,y}\frac{p(Y=y,A=a,c,u^*)-\epsilon_u p(Y=y,A=a,c)}{1-\epsilon_u-\delta_u} \\ &=\frac{p(c,u^*)-\epsilon_u P(c)}{1-\epsilon_u-\delta_u} \end{aligned} \tag{6}$$

As Y is binary, $E[Y|a,c,u]=p(Y|a,c,u)$ so if we plug (5) and (6) into (3) we get τ_{U^*} shown in Figure 3.

Gender Confounding Effect

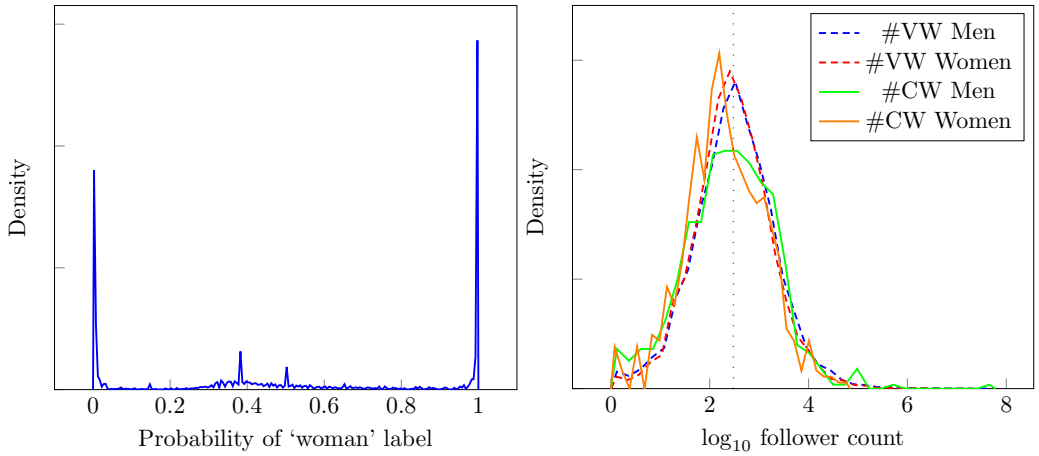
The gender confounding effect is the difference of conditional causal effects. We start with $E[Y(1)|U]$ as the counterfactual expectation. We define the gender confounding as $E[Y(1)-Y(0)|U=1]-E[Y(1)-Y(0)|U=0]$, which can be easily deried as above via Equations (5) and (6).

B Twitter Dataset Collection and Pre-processing

We consider tweets collected from the Twitter streaming API that mention vaccine-relevant keywords (e.g. "vaccine," "immunization") from November 2014 to April 2019. We select tweets containing two hashtags strongly associated with pro-vaccine (#VaccinesWork) and anti-vaccine (#CDCWhistleBlower) tweets. For the 1.8M tweets and retweets containing these hashtags, we re-download them using the Twitter API and remove tweets that have been deleted from the platform.[7] We recursively extract tweets from the `retweeted_status` and `quoted_status` fields and keep all unique tweets which contained one of the two hashtags. This produced 404k unique tweets for #VaccinesWork and 236k for #CDCWhistleBlower. To study general individuals on Twitter sharing vaccine information, we further filter the dataset given several criteria. First, we only consider accounts representing individuals (not organizations). We use an individual vs. organization classifier [49] to remove tweets that are not from individuals. [8] This removes 94k tweets, 20% of the #VaccinesWork and 13% of the #CDCWhistleBlower. Next, we remove accounts dedicated solely to

[7] Tweets and accounts can be deleted; this is most common with spam or bot removal.

[8] Future work could also model the measurement error in this step of our analysis.



(a) Distribution of continuous gender label.

(b) Distribution of follower-count by gender and hashtag. The vertical dotted line is at 300 followers.

Figure 7: Gender distributions

the promotion of vaccination information since we are interested in general users, not vaccine-specific accounts. We remove users who posted more than ten such tweets. This yielded tweets from 21k #VaccinesWork and 1.2k #CDCWhistleBlower *users*.

Finally, we obtain perceived author gender using a gender classifier [23], which infers a probability that a user is a 'man' or 'woman.' We use the probability that that user is labeled as a woman by the classifier, which gives us a gender label between 0 and 1 to use in our analysis. Figure 7a in Appendix C shows the distribution over the gender label probability for all users. The raw data for our treatment (follower-count) and outcome variables (likes received) are both discrete (integers) variables. Since our analysis framework assumes fitting the joint density of binary variables, we convert follower-count (A) and likes received (Y) into binary variables by binning. We remove users with fewer than 10 or more than 10k followers, and split the remainder at a cutoff of 300 followers; $A{=}1$ if a tweet's author has more than 300 followers, and $A{=}0$ otherwise. As the majority of tweets in our dataset receive no likes, we define $Y{=}1$ if a tweet receives at least one like, and $Y{=}0$ otherwise. A possible concern with any binning process is that it assumes homogeneity within each group. For example, if it were the case that all men with more than 300 followers actually had 3,000 followers but all women with more than 300 followers only had 400 followers, then any differences we attributed to gender might actually be attributable to the heterogeneity within our 'popular' treatment category. However, Figure 7b (Appendix C) shows the distribution of follower-count and likes-received is fairly similar for men and women.

C Robustness to twitter preprocessing

In §B, we listed several preprocessing steps that narrow the focus of our analysis. Choices such as the cutoff between popular and unpopular users may unduly influence the results of our analysis. To consider the influence such choices may have, we provide additional details of our data.

Figure 7a demonstrates that treating the gender classifier output as a binary label for the purposes of Figures 6a and 6b does not throw away too much information.

Figure 7b demonstrates that the follower distributions conditional on (binarized) gender and hashtags are quite similar, and we should not expect any particular cutoff to conflate popularity with gender.

D Differential measurement error

All of our analyses as presented rely upon an assumption that the measurement error is *non-differential*, meaning that the error rate is independent of A,C,Y. If this is not the case, our measurement error correction becomes more difficult; we must model the error rate as it depends on those variables.

	Women	Men	Total
Unpopular	32.63	21.77	54.40
Popular	24.53	21.07	45.60
Total	57.16	42.84	100.00

(a) $p(A,U^*)$ distribution of 1,092 #CDCWhistleBlower users.

	Women	Men	Total
Unpopular	27.07	22.85	49.91
Popular	25.53	24.56	50.09
Total	52.60	47.40	100.00

(b) $p(A,U^*)$ distribution of 19,890 #VaccinesWork users.

Table 2: Complement to Table 1. Tables (c) and (d) show the joint distribution of popularity and inferred gender.

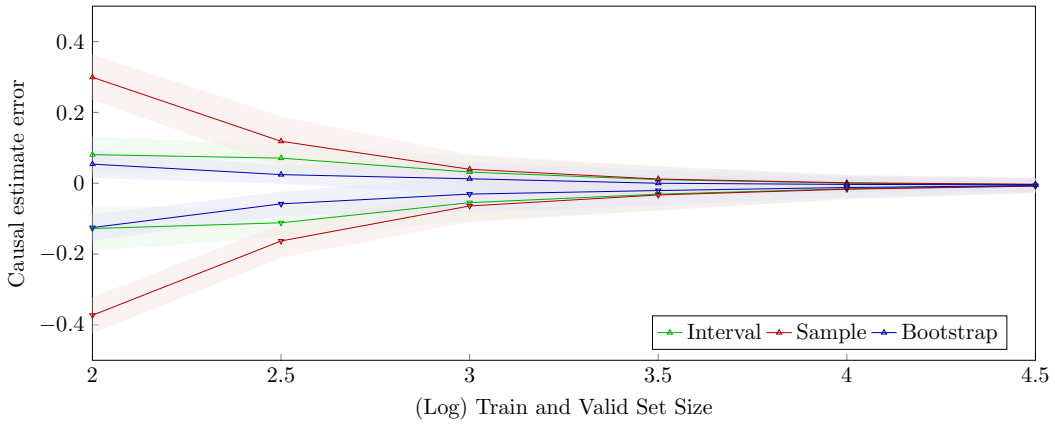


Figure 8: Our sensitivity analyses with a classifier trained on synthetic data, where our methods assume non-differential error. Our methods are over-confident and converge so as to not contain the true causal effect. Each line and its bounds represent the mean and standard deviation calculated from 100 simulations on ten different distributions.

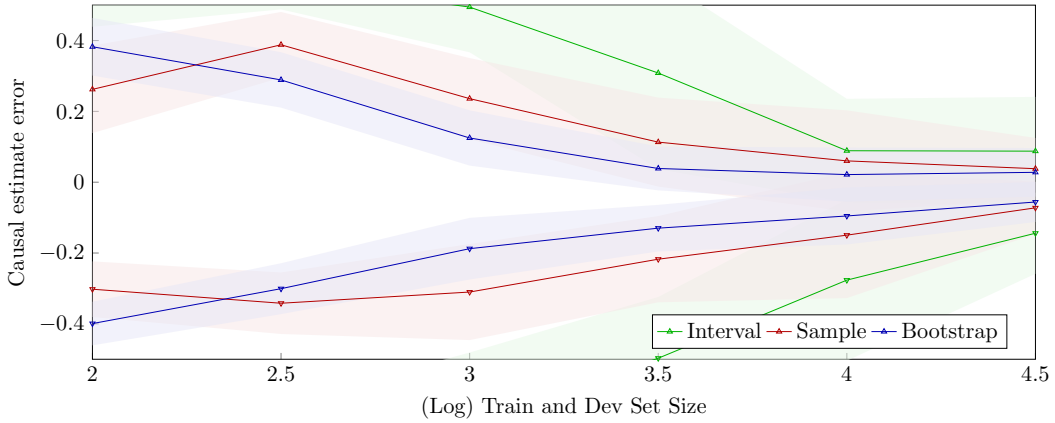


Figure 9: Our sensitivity analyses with a classifier trained on synthetic data, where our methods assume differential error. Our methods tend to be under-confident, with the interval method providing uninformative bounds for many validation set sizes. Each line and its bounds represent the mean and standard deviation calculated from 100 simulations on ten different distributions.

Much applied work on measurement error assumes non-differential error, as causal effects can be unidentified without such an assumption [6, 7].

Each of our sensitivity analyses need to adapt to differential measurement error in different ways. Full differential error means that for a C confounder of dimension k, we need to estimate 2^{k+2} error rates even in the fully-binary case. The Bootstrap analysis is identical, but may require many more samples to cover the variability of differential error. The Binomial sampling approach can simply sample

these many error rates, but again it may take many samples to cover the space of possible causal effects. Our Clopper-Pearson approach becomes quickly intractable to compute as the dimension of the causal DAG increases. If we want to consider all combinations of interval endpoints for a DAG with k variables, we must calculate 2^{2^k} endpoint combinations. This is 65k calculations for 4 variables, and many billions for 5 variables. Future work could explore better ways to balance coverage, interval width, and computational tractability in the differential error setting.

Comparing Figures 8 and 9 shows the how our estimates change when we assume or do not assume differential error for a trained classifier. When our methods try to account for the need to estimate additional error rates, they converge more slowly, with the Clopper-Pearson interval approach providing uninformative bounds when the validation set is small.