# 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:

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

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:

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

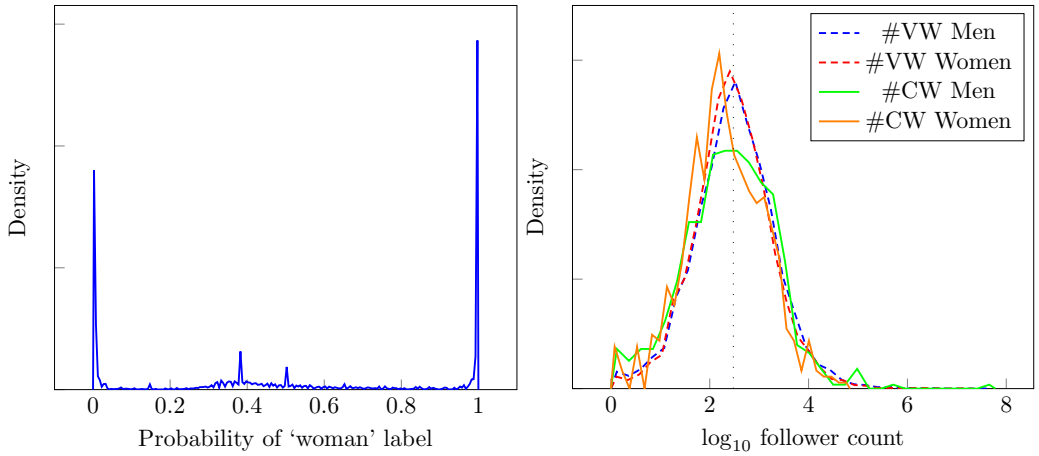
### Measurement Error

The derivation becomes more complicated with measurement error. We can follow Equation (7) from [35] 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 $\epsilon_0+\delta_0 = \epsilon_1+\delta_1$.

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

Then,

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




(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

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

Plugging in Equations (5) and (6) into Equation (3) gives us $\tau_{U^*}$ 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.[6] 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 [47] to remove tweets that are not from individuals. [7] This removes 94k tweets, 20% of the #VaccinesWork and 13% of the #CDCWhistleBlower. Next, we remove accounts dedicated solely to 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*.

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

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

| | 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.

Finally, we obtain perceived author gender using a gender classifier [22], 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. 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.

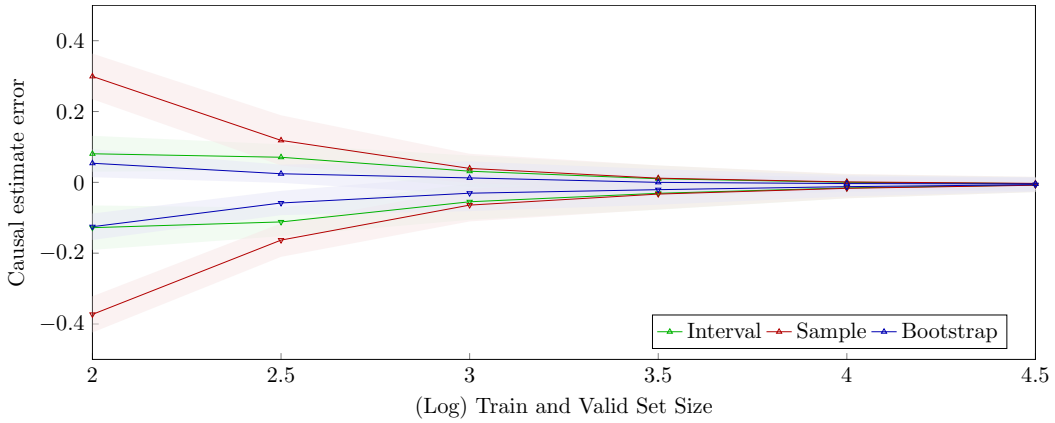



Figure 8: Our sensitivity analyses with a classifier trained on synthetic data, where our methods assume non-differential error. Our methods are over-confident and converge so as to not contain the true causal effect. Each line and its bounds represent the mean and standard deviation calculated from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

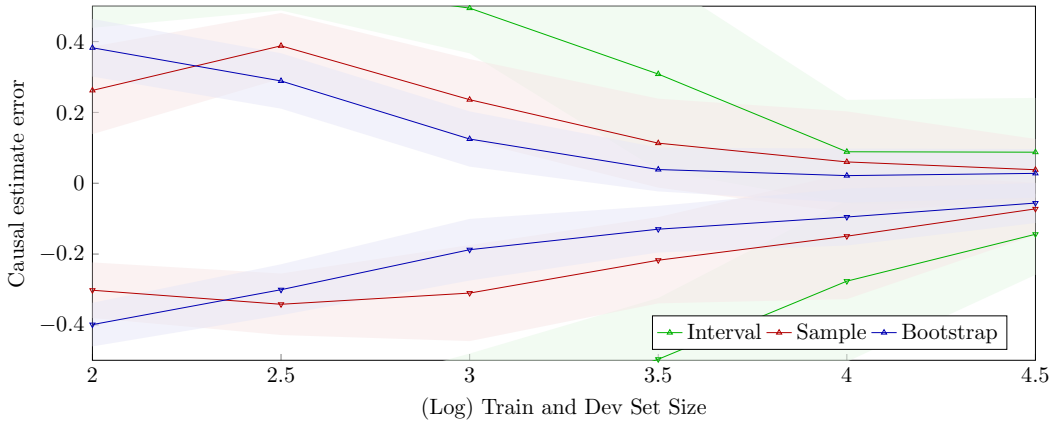



Figure 9: Our sensitivity analyses with a classifier trained on synthetic data, where our methods assume differential error. Our methods tend to be under-confident, with the interval method providing uninformative bounds for many validation set sizes. Each line and its bounds represent the mean and standard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

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.