Classifiers for Measurement Error Correction in Causal Inference

Sayuj Choudhari, California Institute of Technology
Mentor: Zach Wood-Doughty Ph.D., Northwestern University
Associate Mentor: Chris Umans Ph.D., California Institute of Technology

August 2023

Abstract

Causal inference methods are relevant for identifying causal relationships across many fields such as medicine. When a causal problem contains unobserved confounding, classification of this confounding must be adjusted using measurement error correction. Current methods for measurement error correction for classification of unobserved confounding variables in causal inference involve a manual calculation of error rates for different examples of data for a training set that includes the true unobserved confounding. This method becomes vulnerable to miscalculated error rates with increased data dimensionality. A larger variation in examples of data over a set training size would result in less data per example and risk of miscalculation. We propose an alternative method of training a classifier to predict the error rate of a given example of data from which the predicted error rate on each example can be used to correct the distribution of the unobserved confounding variable. This classification method should perform better in higher dimensional cases where the model can learn the impact of each dimensional variable on the error rate and better learn the true error rate of examples with little training data. Our results show that in higher dimensional cases of classification on simple synthetic text generation, that the classification model performs significantly better than the original method. On artificially classified data, the model method outperforms the original method in cases where a large enough amount of data is seen per example.

1 Background

Causal inference is the study of determining whether a causal relationship between a treatment and outcome can be established given an analysis of relevant data. Researching these causal relationships and our ability to identify them holds various applications in "statistics, computer science, education, public policy, and economics" [1]. A main immediate application of causal inference

is in healthcare analyses for the causal relationships of a treatment on an outcome given past patient data. A canonical example question of causal inference in healthcare is whether smoking causes cancer. While data may show that a higher proportion of smokers tend to develop cancer, this correlation by itself does not fully establish a causal relationship. It fails to establish a cause since it does not consider other confounding variables that may be impacting the treatment and outcome. For example, a given smoker with a strong family history of some type of cancer is very likely to develop cancer due to that family history, not the fact that they smoke. Thus, to more accurately establish a causal relationship, we must aim to consider all possible confounding variables of the treatment and outcome in question (smokers, cancer) [2]. In the example of smoking and cancer, possible confounding variables include the background medical information and other lifestyle choices of the patients. The effects of confounding variables in causal inference is further described by Greenland et al. [3].

An issue that raises when trying to consider confounding variables is that these variables are not necessarily explicitly measured in the data. Causal methods generally rely on an assumption that no unobserved confounding is present, as they do not account for the effect it would have and thus lose all theoretical properties. One promising approach to handling unobserved confounding is to find a noisy proxy that is similar to the true unobserved variable. A canonical example of this appears when studying the causal relationship between some drug and cardiovascular disease with patient obesity as a confounding variable, since the patient's obesity may impact whether they are prescribed the drug and can also impact risk of cardiovascular disease. However, if patient obesity is not observed, we might instead use Body Mass Index (BMI), which is often linked with obesity. Mahadevan and Ali [4] study the relationship between BMI and obesity, and show that BMI can serve as a noisy proxy for the obesity variable. Naively replacing obesity with BMI would result in errors in a causal analysis as the BMI indicator may falsely classify whether a subject in the data set is obese.

The effects of using a noisy proxy for an unobserved confounding variable and the error levels it produces in our analysis creates a line of research within causal inference for measurement error correction. If we are able to calculate these error levels we can then look to correct the error produced by the noisy proxy and determine the true effect of the real confounding variable on the causal relationship (e.g. if we can calculate the error produced by using BMI as a measure of obesity, we can aim to correct that error to more accurately represent obesity in our analysis). Thus, through measurement error correction, we can look to measure the true effect of our unobserved confounding variable instead of the noisy proxy, producing more accurate results for causal relationships and expanding the amount of real-world data/ scenarios we are able to use for causal inference [5].

In this project, we will explore new methods for measurement error correction for confounding variables that are estimated through Machine Learning (ML) Classifiers (e.g. our proxy for obesity is the estimate produced by a clas-

sifier), by using a second classifier to estimate the error of these imperfect proxy classifiers. We will apply these methods to synthetically generated text data that is classified through a true classifier as well as artificially.

2 Related Work

Current methods for measurement error correction are based off the effect restoration by matrix adjustment method calculation developed by Pearl [5], in which the estimation of true classifier error rates is plugged into a matrix that is then used to transform the initial distribution of the unobserved proxy variable to a predicted true distribution. This method manually calculates the error rate for each combination of possible data values on a test set of data, assuming these values are the true error rate for the proxy classifier. These values are then used for matrix adjustment. The viability of the effect restoration through matrix adjustment method is studied by Kuroki and Pearl [6]. They conclude that effect restoration is useful in conditions where the unobserved variable "U is a sufficient confounder relative to (X, Y)" (treatment and effect variables), meaning U influences both X and Y to a statistically significant degree. If this condition is met subject to additional technical conditions (Conditions 1-4 in [6]), we can conclude that effect restoration by matrix adjustment would likely improve the predicted distribution of the unobserved confounder and accuracy of the overall causal estimation. The use of effect restoration in experimental cases with synthetic data has been more recently studied by Oktay et al. [7], where these methods were more effective in correcting classification bias than an alternative propensity score matching method. The study also furthered the conclusion by Kuroki and Pearl [6] that as the influence of U on X, Y increases the benefit of effect restoration increases as well.

Methods for causal inference problems with confounding include propensity score matching and inverse propensity weighting (IPW). Propensity score matching, detailed by Rosenbaum and Rubin [8] is the study of causal effects through pairing test subjects by matching their propensity scores of how likely they are to receive a certain treatment. If two data points or experimental subjects are statistically very similar to each other (e.g. two patients in a medical study who have nearly all the same vitals data and medical history) and they receive two different treatments, we assume that any difference in the outcome is caused by the treatment, and not any other confounding variable as both patients share the same confounding traits. This similarity in patients would result in a similar propensity score in how likely they are too receive a treatment, making propensity score a good indicator of how similar two examples are in an experiment. For problems with unobserved confounding, this translates to similar values of a variable influenced by the unobserved confounding assumed to be the same confounding effect. The IPW method, recently reviewed by Guo et al. [9], uses propensity scores as well for weighting the effect measured by each example. This allows for the overall causal analysis to treat examples with unlikely features (e.g. an example with a low propensity score still receiv-

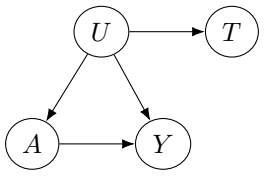

Figure 1: Simple graph showing the measurement error setting. A is the treatment, Y the effect, U the unobserved confounding, and T observed text data that is influenced by unobserved U.

ing the treatment) with less priority than more likely examples, strengthening the overall causal estimate by heavily weighting data that is expected to be observed. While both the propensity score matching and IPW methods have shown effective results in causal inference problems, both methods theoretically will run into larger errors once unobserved confounding is introduced. Consider treatment A, effect Y, unobserved confounder U, and observed text T influenced by U as seen in Figure 1. For causal estimation, both methods require $P(A \mid U)$, but with U unobserved use $P(A \mid T)$ instead. However, U is not T nor is perfectly correlated with T and thus $P(A \mid T)$ will vary from $P(A \mid U)$ risking error in the overall causal estimation since no correction is being made to transform $P(A \mid T)$ to converge to $P(A \mid U)$.

Results from Wood-Doughty et al. [10] suggest that measurement error correction is the most effective method for problems with unobserved confounding, specifically with classification of synthetically generated text as a proxy variable. The experiments run on all three methods show that overall causal estimation error is only correlated with classification accuracy for the matrix adjustment method while propensity score and IPW show no correlation between the accuracy of the text classifier and the overall causal estimate making it difficult for either of the methods to confidently yield accurate estimates. The negative correlation between overall causal estimation error and classifier accuracy for the measurement error correction with matrix adjustment, gives reasoning to continue studying measurement error correction methods and developing more accurate classification for improved overall causal estimate. A potential path for improving measurement error correction is through implementation of machine learning into the correction methods. Yang et al. [11] implement a random forest classifier to mitigate measurement bias that effectively recovers a distribution similar to the true the original unbiased distribution of a binary variable.

Developing effective methods for causal inference problems with unobserved confounding is important for expanding the application of causal inference methods to more types of real-world problems and data. Methods to classify unobserved confounding can be used to analyze problems with influence from qualitative data such as text, allowing for causal inference to apply to many quantitative problems where qualitative data may be a factor. Medicine is a field where causal inference methods for unobserved confounding can be par-

ticularly useful. Rajkomar et al. [12] review machine learning applications to healthcare, detailing the effectiveness of deep learning methods in predicting many types of medical events and trends. The benefits of applying machine learning to healthcare given large-scale medical data sets are studied by Chen and Asch [13], highlighting the combination of human analysis and machine learning analysis on large-scale data to outperform either of the methods independently. These conclusions support studying machine learning applications for causal inference to potentially be effective for medical analysis, specifically for medical questions on causal relationships between treatment and outcome on patients. Existing methods for causal inference have shown meaningful results in healthcare studies such as the causal effect of transthoracic echocardiography (TTE) on all-cause mortality in patients with sepsis by Feng et al. [14], where causal inference methods applied to the MIMIC-III database for anonymized clinical data [15] were able to identify the benefit of transthoracic echocardiography (TTE) for ICU patients. Methods for unobserved confounding would be able to extend this application to medical causal relationships where data such as Socio-economic status (SES), obesity, or physician's notes may be an influencing variable. In this study, we propose a measurement error correction method that mathematically follows the matrix adjustment process, except using a classifier to predict the true error rates of the proxy classifier rather than a manual calculation of error rates on test data that are assumed to be the true errors. This regression model method for correction is expected to outperform the matrix adjustment method for overall causal estimate in cases where less data is seen per assignment in the probability distribution of data $P(A, Y, C, U)$. The regression should be able to group data points as well as identify trends or effects caused by certain confounding, to better analyze lesser observed data points in problems with smaller test data available, or more complex problems with many influencing factors. Thus, the regression method could potentially expand the use of causal inference to medical relationships with less observed data (rarer medicines, medical situations) or with many influencing factors.

3 Results

This study compares the use of a classifier to construct an error matrix for the matrix adjustment method in measurement error correction to the current method of matrix adjustment that uses manual calculation of error rates per example. The classifier takes in all variables (e.g. other confounders, treatment, and outcome) influencing the classification of the unobserved confounding (including the proxy itself), and outputs the probability that the proxy classification was an error. For further technical details on this classification method, refer to Section 4. These predicted errors from the classifier represent the same errors that are required for matrix adjustment and can therefore be applied to adjust and correct the initial distribution of the unobserved confounding towards the truth.

Two main complications exist for accurately measuring true classification

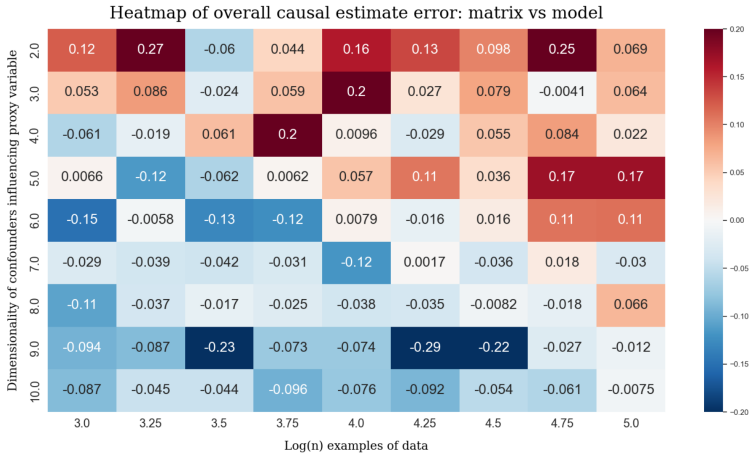

Figure 2: Heat map of difference in overall causal estimate error of matrix manual error calculation vs. model error classification. Positive values highlighted in red represent experiments where the model approach outperformed the manual calculation.

errors of unobserved confounding. First is the dimensionality of the variables influencing the proxy and therefore the classification of the true unobserved variable. In a given causal inference problem with unobserved confounding, observed confounding may also exist that directly affects the proxy variable making it more complex to identify the relationship between the unobserved confounder and the proxy. The treatment and outcome variables can also directly affect the proxy, making identifying the unobserved confounding increasingly complex as the amount of influencing variables increases [6]. This problem is further complicated by the size of the training set of data that the errors are calculated or predicted on. The less data available per assignment within the distribution of data, the more likely manual calculations or the classifier training will deviate from the truth. Thus, it is important to study the performance of the model classification method versus the manual calculation method under various degrees of complexity in both training set size as well as dimensionality of influencing variables.

The heat map in Figure 2 shows the difference in overall causal estimate error rate on data sets with varying training set sizes and dimensionality of influencing variables. The difference as calculated as *manual calculation error - model classification error*, so stronger positive values signify model classification better outperforming the manual method. The model classification outperforms the manual calculation for experimental settings in the upper diagonal region of the cell grid, suggesting there is an upper limit of complexity under which the model outperforms the original method. While the model performs better in many complex cases such as $\log(n) = 3.75$ examples of data and 4 influencing variables, the upper limit on this performance contradicts expectations that the

model would further outperform the original method as complexity increased. Looking at the data output of estimated errors from the original method, it was noticed that in cases of high complexity (high amount of influencing variables on a small training set, lower diagonal region), the manual calculation would often not see an assignment within the training set and default the calculated error to zero, meaning no adjustment would be made. This result shows that the upper limit of model performance was not compared to the original method, but zero adjustment, which is more understandable as not adjusting the initial distribution may be advantageous to any adjustment in scenarios where the error risk due to misadjustment outweighs the lack of (e.g. higher complexity cases where training sets and classification may vary far from the truth). Comparisons of measurement error correction against uncorrected classifiers are studied in detail by Wood-Doughty et al. [16] who also produce results of uncorrected classifiers outperforming measurement error correction methods in small $\log(n)$ validation set settings where little to zero observed assignments are likely to exist within the data set.

To further define and understand this upper limit, the methods were compared against the expected amount of data points seen per assignment. This calculated expectation will serve as the estimated complexity of the data set to be analyzed. Plotting the difference of the model and original method performance calculated as in the heat map in Figure 2, general trends in the comparative performance of both methods should be identifiable.

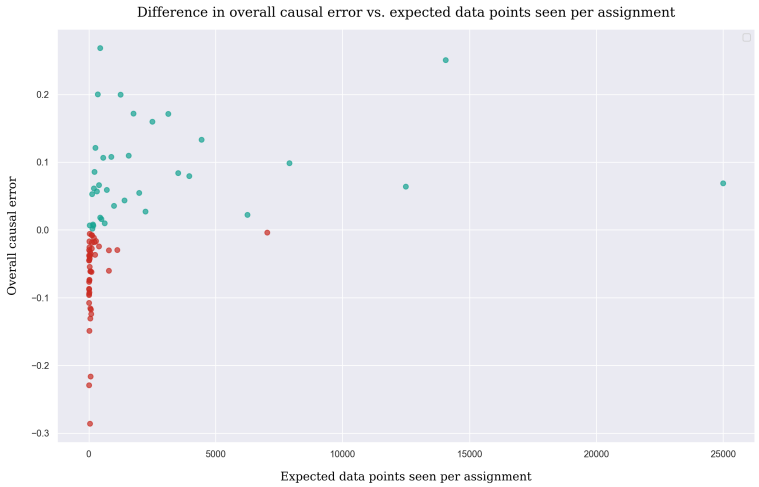

Figure 3: Scatter plot of difference in overall causal estimate error of matrix manual error calculation vs. model error classification against expected data points seen per assignment (representative of complexity of analysis). Blue data points represent experiments where the model outperformed the original method.

As seen in Figure 3, the original method only outperforms the model in ex-

perimental settings where the complexity of the problem approaches zero data points seen per assignment. Furthermore, beyond the range of roughly 1500 data points expected per assignment, the model classification consistently outperforms the original method. It is noticed that data points near zero expected data points seen per method consistently favor the performance of the original method. These results support the conclusion from Figure 2 that when the original method sees little to no observed data for an assignment, it keeps the non-adjusted distribution which yields better results than with adjustment. For causal inference problems on synthetically generated data, from simpler problems with very few confounding variables and more data available per assignment to a complexity of roughly 1500 data points seen per assignment, the model classification is consistently favorable over the original method.

4 Methods

The model classification method for predicting error rates is motivated by the understanding that error rates for classification of an unobserved proxy would display consistent impacts and trends for given influencing variables, making the classification function fairly smooth and easy to learn. For notation purposes, A represents treatment, Y for outcome, C for confounding variables, U for unobserved confounding variables, T for text data, and U^* for the proxy classification for U.

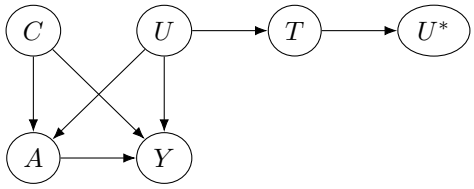

Figure 4: Graph of the measurement error setting where T is only influenced by unobserved U

Figure 4 displays the simplest causal graph for estimating unobserved confounding. Unobserved confounder U is the only variable influencing text data T, and thus the effect of U on T is more clearly observable and U^* can be more easily classified. For matrix adjustment with a binary unobserved confounding, there are 2 errors that need to be calculated, $P(U^* = 0|U = 1) = \epsilon$ and $P(U^* = 1|U = 0) = \delta$. The case of the causal graph above is the simplest possible problem since T is only influenced by U and therefore less conditional assignment's exist for δ and ϵ to be calculated on. The more difficult case is where T is not only influenced by U but by many other variables in the problem.

Figure 5 represents the causal graph of the more complex scenario, possibly with additional observed confounding $C_1, C_2, C_3, ..., C_n$ are added with an influence on T. Let S be the set of all possible assignments for all variables

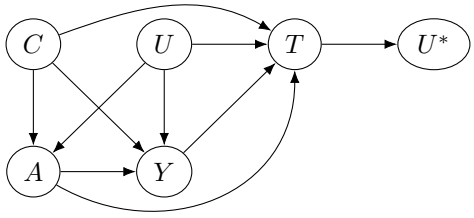

Figure 5: Graph of measurement error setting where T is influenced by all variables present confounding, treatment, and effect variables

that influence T. Since U is no longer the sole influence on T, U^* the distribution used for adjustment becomes more complicated as we cannot assume that $P(U^* \mid U) = P(U^* \mid U, s)$. Instead the error rate of the U^* classification must be calculated uniquely for each assignment s. Thus, the adjustment of the U^* distribution must instead calculate for each assignment $s \in S$: $P(U^* = 0 \mid U = 1, s) = \epsilon_s$ and $P(U^* = 1 \mid U = 0, s) = \delta_s$, adjusting the distribution of $U*$ uniquely for each assignment with the respective ϵ_s, δ_s, as the masking for $U \rightarrow T$ is different for each assignment requiring its own adjustments for the distribution $P(U^*)$ to converge to $P(U)$.

Assuming that each assignment will have unique error rates caused by the assigned values of influencing variables, each combination of variables should have a specific effect on T and therefore U^*. For the proposed classification model, a classifier such as a Logistic Regression would be able to learn these trends easily if the classification surface is relatively smooth as each influencing variable value contributes a consistent effect. This makes the classification model advantageous to the original method of manually calculating errors from a matrix of training set data for problems with complexity to the degree that data size for each assignment varies largely and some assignments are not frequently seen. Complexity to the extent that a manual calculation on such little data may be far from the truth. For example, if it is expected to see 5 samples of data per assignment where the true error for an assignment is .33 it is unlikely that we will be able to manually calculate .33. However, the classification model can take into account conclusions made on the effect of each variable observed throughout the data set as well as results of more frequently observed similar assignments. Inferring the error rate of $P(U^* = 1|U = 0)|A, Y, C_{1_i}, ..., C_{n_i}, U_i^*$ observed 10 times based on the error rate of similar assignment $P(U^* = 1|U = 0)|A, Y, C_{1_i}, ...C_{10_j}, ..., C_{n_i}, U_i^*$ observed 1000 times is expected to be closer to the truth than manually calculating the error on the 10 observed data points.

For the purpose of testing the performance of the classifier model versus the original method, an artificial classifier is used to approximate U^* so both methods can be tested to correct a specified classification. From a distribution of

error rates that is constructed from a covariance matrix of all other influencing variables, each variable value contributes a certain weight to the error when present in an assignment. These error rates are then calculated for each assignment, creating a distribution of error rates for U^* that are sampled to construct the data set to be analyzed. Analysis of synthetic text classifiers showed that in more complex cases the model had an error rate of .5, so the distribution of error rates to construct the U^* proxy was normalized to ($\mu = .5$, $\sigma = .1$). The data set with specified size was split into a training set and a test set, with the training set having access to the true U. A logistic regression classifier would then be trained to predict $P(U^* \neq U | s \in S) \forall s \in S$, and the original distribution of U^* adjusted for each of those predictions. The overall causal estimate analysis would then follow the original method.

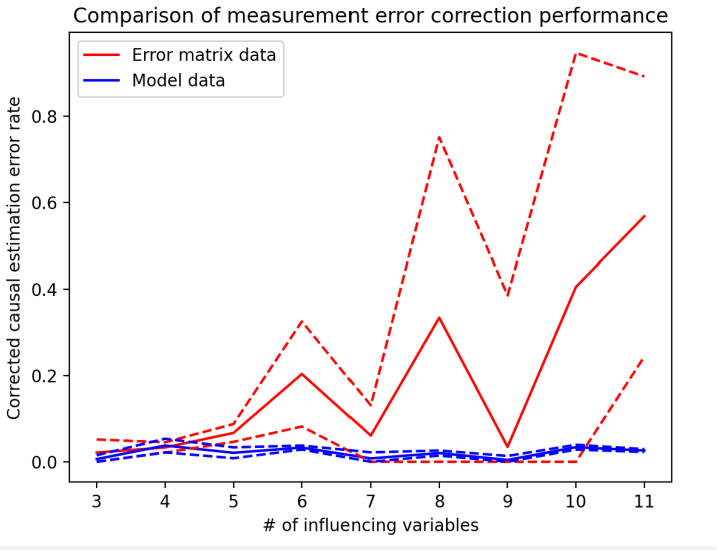

Figure 6: Plot of mean overall causal estimate errors for original method (Error matrix data) and the classification model over varying complexity of influencing variables, with single standard deviation bands. $\log(n) = 4$ examples of data analyzed.

For each combination of experimental settings ($\log(n)$ data set size and amount of influencing variables on T) tested, experiments would be run 4 times for each method and the average overall causal error used for comparison. See Figures 2 and 3 in Section 3 for visualizations of experiments. The results suggest that the model consistently outperforms the original method for these more controlled scenarios within a certain complexity level of above 1500 examples expected for each assignment in the data set. This provides motivation to continue evaluating the classification model on synthetically generated text, a less controlled problems with more difficult data to learn. The simplest ex-

ample of synthetic text classification to test was the bag-of-words synthetic text generation process developed by Wood-Doughty et al. [10] in which influencing variables add a certain weighted probability for a given word to be present within an example of text data, for a vocabulary size of roughly 4300 words.

The plot in Figure 6 supports the use of the classification model for the more difficult problem of correcting a synthetic text classifier. Interestingly, the model consistently outperforms the matrix even at very high amounts of influencing variables where less than 10 examples are expected per assignment, while the original method diverges to error rates and bounds within the .5 range where random prediction may be equally effective. It will be useful to test the classification model on more difficult synthetic text generation processes as well as real-world data such as the MIMIC-IV database of anonymized Intensive Care Unit patient data (Johnson et al. [17]), to validate against known causal relationships. In the medical setting, a common example of classification of unobserved confounding is classifying some patient characteristic based off of physicians notes which are generated by an extremely complicated process, as the text produced by the physician can be influenced by hundreds of factors that may be confounding the causal relationship. Thus, we would expect the distribution of U to be extremely complex as well, and thus exploring various measurement error correction methods under various levels of complexity is relevant for future applications of causal inference methods to real-world scenarios.

5 Acknowledgements

Thank you to my mentor Professor Zach Wood-Doughty for his guidance and support throughout the project. I am grateful for my associate mentor Professor Chris Umans, for his support in my project proposal. Thank you to the coordinators of the SURF program for providing funding and the opportunity to conduct this project.

References

[1] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15:1–46, 05 2021. doi: 10.1145/3444944.

[2] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

[3] Sander Greenland, James M. Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, 1999. ISSN 08834237. URL `http://www.jstor.org/stable/2676645`.

[4] Shriraam Mahadevan and Iftikhar Ali. Is body mass index a good indicator of obesity? *International Journal of Diabetes in Developing Countries*, 36, 06 2016. doi: 10.1007/s13410-016-0506-5.

[5] Judea Pearl. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 425–432, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

[6] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 03 2014. ISSN 0006-3444. doi: 10.1093/biomet/ast066. URL `https://doi.org/10.1093/biomet/ast066`.

[7] Hüseyin Oktay, Akanksha Atrey, and David Jensen. *Identifying When Effect Restoration Will Improve Estimates of Causal Effect*, pages 190–198. 05 2019. ISBN 978-1-61197-567-3. doi: 10.1137/1.9781611975673.22.

[8] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL `https://doi.org/10.1093/biomet/70.1.41`.

[9] Shenyang Guo, Mark Fraser, and Qi Chen. Propensity score analysis: Recent debate and discussion. *Journal of the Society for Social Work and Research*, 11(3):463–482, 2020. doi: 10.1086/711393. URL `https://doi.org/10.1086/711393`.

[10] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Generating synthetic text data to evaluate causal inference methods, 2021.

[11] Mochen Yang, Edward McFowland, Gordon Burtch, and Gediminas Adomavicius. Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *SSRN Electronic Journal*, 01 2019. doi: 10.2139/ssrn.3339983.

[12] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew Dai, Nissan Hajaj, Peter Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, and Jeff Dean. Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 1, 01 2018. doi: 10.1038/s41746-018-0029-1.

[13] Jonathan Chen and Steven Asch. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *New England Journal of Medicine*, 376:2507–2509, 06 2017. doi: 10.1056/NEJMp1702071.

[14] 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(6): 884—892, 06 2018. ISSN 0342-4642. doi: 10.1007/s00134-018-5208-7. URL `https://doi.org/10.1007/s00134-018-5208-7`.

[15] Alistair Johnson and Tom Pollard. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3(160035), 05 2016. ISSN 0006-3444. doi: 10.1038/sdata.2016.35. URL `https://doi.org/10.1038/sdata.2016.35`.

[16] Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Sensitivity analyses for incorporating machine learning predictions into causal estimates. In *NeurIPS 2020 Workshop on Causal Discovery and Causality-Inspired Machine Learning*, 2020.

[17] Alistair Johnson, Lucas Bulgarelli, and Tom Pollard. Mimic-iv, a freely accessible electronic health record dataset. *Sci Data*, 10(1), 01 2023. doi: 10.1038/s41597-022-01899-x. URL `https://doi.org/10.1038/s41597-022-01899-x`.