Identifying When Effect Restoration Will Improve Estimates of Causal Effect

Hüseyin Oktay* Akanksha Atrey* David Jensen*

Abstract

Several methods have been developed that combine multiple models learned on different data sets and then use that combination to reach conclusions that would not have been possible with any one of the models alone. We examine one such method—*effect restoration*—which was originally developed to mitigate the effects of poorly measured confounding variables in a causal model. We show how effect restoration can be used to combine results from different machine learning models and how the combined model can be used to estimate causal effects that are not identifiable from either of the original studies alone. We characterize the performance of effect restoration by using both theoretical analysis and simulation studies. Specifically, we show how conditional independence tests and common assumptions can help distinguish when effect restoration should and should not be applied, and we use empirical analysis to show the limited range of conditions under which effect restoration should be applied in practical situations.

1 Introduction

Growing use of machine learning has led to an interest in combining models learned on different data sets and using those models to make inferences that would not have been possible with any one model. This is particularly valuable when the goal is causal inference [17, 20], one of the most challenging tasks in machine learning. For example, researchers in statistics have long used meta-analysis to produce causal estimates with statistical power that exceeds any individual study [9]. Researchers in causal graphical models have taken a very different approach and studied how to learn a single model from multiple data sets with overlapping sets of variables [23].

In this paper, we use models learned on multiple data sets to correct for *confounding* variables, perhaps the most common threat to the validity of causal inferences. Ignoring confounding variables introduces bias in estimates of treatment effect [17, 20], and conditioning on confounding variables is a common approach to correct this bias. We study how to accurately condition on confounding variables even when those variables are not accurately measured in a data set.

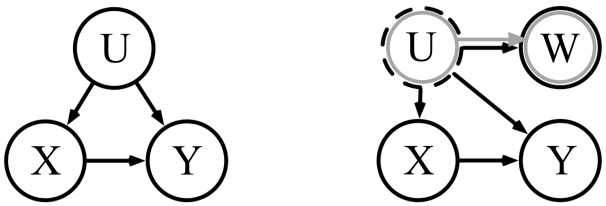
Specifically, we examine how models learned from a second data set can be used to recover or *restore* a confounding variable and thus reduce the bias of the first study's estimate of causal effect. To do this, we apply *effect restoration*, a technique originally proposed by Kuroki & Pearl [8] to adjust for measurement error in confounding variables. We show that the models necessary to implement this method can be drawn from different data sets and that, under specific conditions, such models can greatly reduce the bias of causal estimates.

For example, consider the graphical model in Fig. 1b. One data set contains a treatment X, an outcome Y, and another variable W. Both X and Y are caused by an unmeasured confounder U that also causes W. A second data set contains both U and W. Effect restoration can be used to combine the studies, *restoring* the effect of U on X and Y by using knowledge of $P(U, W)$ from the second study. When the underlying graphical structure in Fig. 1b is assumed and the two studies draw data from similar distributions, effect restoration will reduce the bias due to the unmeasured confounder in the first study.

This application of effect restoration is relatively straightforward, and the theoretical properties of effect restoration established by Kuroki & Pearl also hold in this particular application. However, much remains unknown about the practical utility of the proposed methods. First, it is far from straightforward to conclude that the assumed structure holds for a specific combination of treatment, outcome, and potential confounders. Second, it is unclear whether the adjustment made by effect restoration still provides a bias-free estimate when the underlying generative process does not correspond to the structure in Fig. 1b, a question Kuroki & Pearl considered out-of-the scope in their study. Finally, the range of potential benefits and the conditions under those benefits occur have not been previously examined.

In this paper, we extend the use cases of effect

*University of Massachusetts Amherst, Amherst, MA, USA {hoktay, aatrey, jensen}@cs.umass.edu



(a) (b)

Figure 1: Graphical models with confounding. In (a), U is a confounder for treatment (X) and outcome (Y). In (b), one study (shown in black) contains measured variables X, Y, and W, and an unmeasured confounder U, and a second study (shown in gray) contains measured variables U and W.

restoration and characterize its behaviour under various realistic scenarios. Specifically:

RQ1. How does the actual causal structure affect the accuracy of effect restoration?

RQ2. Practically, what conditions are necessary for effect restoration to substantially improve estimates of causal effect?

RQ3. For a given set of observations, X, Y, and W, what are the sufficient conditions to identify the underlying graphical structure?

Our answers to these questions make important contributions to the understanding of effect restoration. First, we show that most variants of the structure assumed in the original work on effect restoration lead the method to either have no effect or to harm estimates of the causal effect of treatment on outcome. Second, we use empirical analysis to show that the relative benefit of effect restoration is highest for estimating small treatment effects with large confounding bias. By using empirical data from a real-world randomized experiment, we show how effect restoration removes bias more than its natural alternatives. Third, by leveraging graphical models and d-separation, we show for the first time that simple rules and typical temporal ordering assumptions are sufficient to identify whether an empirical system has the causal structure necessary for effect restoration to improve causal estimates.

2 Related Work

A wide variety of methods for estimating causal effects assume that all confounders are observed. This assumption, sometimes referred to as *causal sufficiency* [20], implies that all variables that are causes of two or more observed variables in a given data set are also observed. Causal sufficiency is typically required to guarantee consistent and/or unbiased causal estimates [5, 18, 22].

We explore the use of effect restoration to account for bias from unobserved confounder variables. Apart from randomized experiments, several methods have been proposed to account for unobserved confounders in non-experimental contexts. These include instrumental variable designs [1], which can produce unbiased estimates of causal effect under some very restrictive assumptions, as well as sensitivity analysis techniques, which are used to assess the likelihood of the existence of a potential unobserved confounder by estimating the required confounding bias to reverse the estimated effect [10]. In contrast, our work accounts for unobserved confounders by using models of confounders derived from a separate data set.

Our approach resembles transfer learning approaches in machine learning where a target learning task is performed by using knowledge obtained from previous related tasks [11, 12, 15, 25]. Our study departs from the standard transfer learning paradigm because we first obtain knowledge from a predictive learning task and then use such knowledge for a subsequent causal estimation task.

3 Background

To formalize the problem of effect restoration, we use graphical models and Pearl's *do*-calculus [17]. For example, in the graphical model shown in Fig. 1a, we denote the direct effect of X on Y as $P(Y \,|\, do(X))$. In observational studies, this quantity is different than simply conditioning on X (i.e., $P(Y \,|\, X)$). Conditioning denotes the probability distribution of Y in each possible world defined by a specific X value. However, the *do* operator denotes the probability distribution of Y after actively setting the value of X (i.e., intervention). Hence $P(Y \mid do(X))$ represents the effect of manipulating the values of X and $P(Y|X)$ represents passive observation. Readers should consult Pearl [17] for additional discussion of the *do*-calculus.

For the graphical model in Fig. 1a, when all variables are observed and under identifiability conditions (again, see Pearl [17] for details), the probability $P(Y \mid do(X))$ can be estimated [8]:

$$P(Y \mid do(X = x)) = \sum_{U} \frac{P(X, Y, U)}{P(X \mid U)} \tag{3.1}$$

Given this *interventional* distribution, for a binary X, the *treatment effect* (TE) of X on Y can be calculated (note that Y_1 is equal to $Y = 1$ and the former format is chosen for all variables for brevity):

$$TE = \log\left(\frac{P(Y_1 \mid do(X_1))}{P(Y_0 \mid do(X_1))}\right) - \log\left(\frac{P(Y_1 \mid do(X_0))}{P(Y_0 \mid do(X_0))}\right)$$

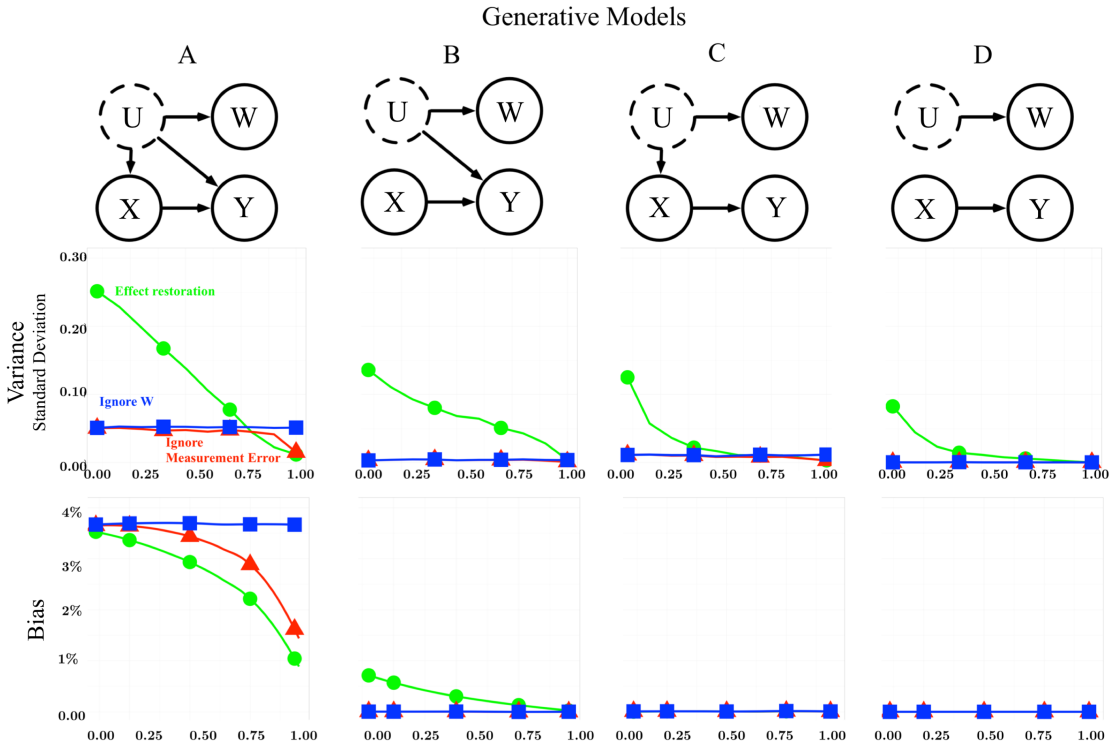


Figure 2: Bias and Variance Results for Different Underlying Structures.

Table 1: Conditional (In)dependence Relationships for All Simple Graphical Structures.

	U is temporally prior to X and Y				U is temporally posterior to X or Y				Cycle
	A	B	C	D	E	F	G	H	I
$X \perp\!\!\!\perp Y$	✖	✖	✖	✖	✖	✖	✖	✖	N/A
$X \perp\!\!\!\perp Y \mid W$	✖	✖	✖	✖	✖	✖	✖	✖	N/A
$X \perp\!\!\!\perp W$	✖	✔	✖	✔	✖	✖	✖	✖	N/A
$X \perp\!\!\!\perp W \mid Y$	✖	✖	✖	✔	✖	✔	✖	✖	N/A
$Y \perp\!\!\!\perp W$	✖	✖	✖	✔	✖	✖	✖	✖	N/A
$Y \perp\!\!\!\perp W \mid X$	✖	✖	✔	✔	✖	✖	✔	✖	N/A

Table 2: Effect of Activity Level on Weight Gain

Activity Level	Weight Gain	
	0	1
0	0.20	0.80
1	0.60	0.40

From perfect observations of X, Y, and U, unbiased estimates of TE can be obtained using various modeling methods. However, these methods fail to provide unbiased estimates if U is measured with error. Kuroki & Pearl used the *do*-calculus to adjust for confounding variables with measurement error [8]. Specifically, the adjustment uses knowledge of the error distribution to correct the causal estimate made using the observed values. Fig. 1b shows the graphical model assumed in their extended framework. They propose that, under certain conditions, the TE of X on Y can be restored bias-free given an observed surrogate variable for the confounder U (i.e., W) and a known error distribution (i.e., $P(W \mid U)$).

Our experiments use a specific formulation of the Kuroki & Pearl TE estimator with the *do*-calculus framework. For example, when estimating the effect of activity level (X) on weight gain (Y), age (U) is a potential confounding variable because

it affects both activity level and weight gain. Our goal is to estimate the interventional distribution of $P(WeightGain \mid do(ActivityLevel))$.

Assume the distribution shown in Table 2. According to this distribution, when a person has low activity level, the odds of weight gain is $\frac{0.80}{0.20} = 4$. Whereas, when a person has high activity level, the odds of weight gain is $\frac{0.40}{0.60} = 0.66$. We define the difference in the log-odds ratio for different treatment values as the causal effect of treatment (e.g., activity level) on outcome (e.g., weight gain).

4 Effects of Underlying Structure on Bias and Variance

The effect restoration method proposed by Kuroki & Pearl assumes U is a confounder as shown in Fig. 1b. Here, we relax this assumption and characterize the performance of effect restoration under all possible modifications of the simple graphical structure suggested in the original paper.

Fig. 2 shows all possible modified structures between X, Y, W, and U with the assumptions that are typical in many social science and medical studies [5, 17, 22]: (1) X is temporally prior to Y; (2) W is a noisy measurement of U; and (3) U is temporally prior to X and Y. We employ simulation studies to characterize the performance of effect restoration for each graphical model structure.

4.1 Data Generation We generated continuous and discrete synthetic data consistent with the graphical structures in Fig. 2. We explain the discrete data generation process for the graphical structure shown in column A of the figure. Other graphical structures follow the same steps except that the added, removed, or reversed dependencies in the structure entail adding, removing, or changing the order of steps in the data generation process, respectively.

In the generative processes below, italic letters denote scalar values (e.g., N); upper-case bold characters denote vectors (e.g., $\mathbf{W}$); each element of a vector is accessed by an index subscript (e.g., w_i); correlations between variables are referred with subscripts such as ρ_{uw} denoting the correlation between $\mathbf{U}$ and $\mathbf{W}$; marginal and conditional probabilities are denoted by the upper-case letter P. Following is the pseudocode to generate binary data for the structure in column A in Fig. 2.

We vary $\rho_{uw}, \rho_{ux}, \rho_{uy}$ and ρ_{xy} between (0,1). We model $P(Y \mid U, X)$ as a *noisy-OR* conditional probability distribution [6] where:

$$P(Y = 0 \mid \mathbf{U}, \mathbf{X}) = (1 - \lambda_0) * (1 - \rho_{uy})^{\mathbf{U}} * (1 - \rho_{xy})^{\mathbf{X}}$$
$$P(Y = 1 \mid \mathbf{U}, \mathbf{X}) = 1 - (1 - \lambda_0) * (1 - \rho_{uy})^{\mathbf{U}} * (1 - \rho_{xy})^{\mathbf{X}}$$

Algorithm 1 Binary Data Generator

```
Initialize correlation values ρ_uw, ρ_uy, ρ_ux, ρ_xy
Draw prior P_u ~ Uniform(0, 1)
Draw N values for U from Bernoulli(P_u)
for u_i in U do
    Draw p' ~ Uniform(0, 1)
    if p' < ρ_uw then
        w_i = u_i
    else
        Draw a value for w_i ~ Bernoulli(p = 0.5)
for u_i in U do
    Draw p' ~ Uniform(0, 1)
    if p' < ρ_ux then
        x_i = u_i
    else
        Draw a value for x_i ~ Bernoulli(p = 0.5)
Draw Y ~ Binomial(N, P(Y = 1 | U, X))
```

We set the value of $\lambda_0 = 0.01$ as the noise parameter of the noisy-OR model. The noisy-OR model is a popular choice to represent discrete conditional probability distributions [16] due to its compact representation with few parameters and its ability to approximate many learned distributions [26].

For each parameter setting $\{\rho_{uw}, \rho_{uy}, \rho_{ux}, \rho_{xy}\}$, we generate 50 data sets, each containing 10,000 instances. As noted, for other graphical structures we revise the data generation process for the corresponding removed dependence. For example, in Fig. 2 column B, the dependence between U and X is removed. To account for this, we sample a prior for P_X from a uniform distribution; we sample values for X using the prior instead of using $P(X \mid U)$.

In our experiments, we calculate the true treatment effect (i.e., TE) as our ground truth by using the values of U. We estimate the treatment effect (i.e., TE') with the following three approaches: (1) **Ignore W**: Simply ignoring the measurements of W, (2) **Ignore measurement error:** Using W and ignoring the measurement error, and (3) **Effect restoration:** Using W and adjusting for the measurement error. Note that these three methods do not use values of U, only use values of X, Y, and W. We measure the standard deviation of error values across experiments and the lines correspond to the mean standard deviation for varying strength of effect values between U and W. We measure the bias for each approach in each experiment by normalized error, as shown in equation 4.2 and the lines correspond to the *local weighted regression* of bias values for varying strength of effect between U and W.

$$\epsilon = \frac{TE - TE'}{TE} \tag{4.2}$$

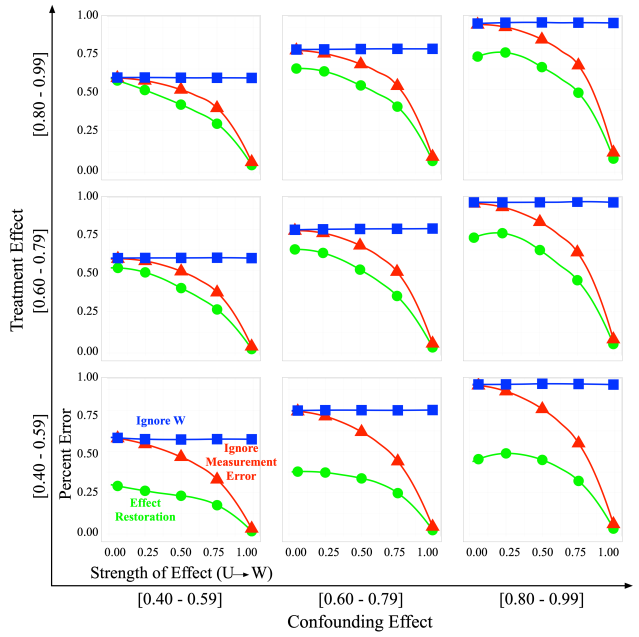


Figure 3: Bias as Treatment and Confounding Effect Change.

4.2 Bias and Variance for Different Underlying Structures We perform simulation analysis for each graphical structure in the first row of Fig. 2. We plot both the variance and the bias in estimating the treatment effect in the second and third row, respectively. In each plot, we show the behaviour for each of the three different approaches as the measurement error changes along the x-axis. We calculate the measurement error by the strength of dependence between U and W using the Cramér's V coefficient. The stronger the dependence, the weaker the measurement error. On the y-axis we plot locally smoothed normalized absolute error for bias and locally smoothed standard deviation for the variance for the three different approaches. Fig. 2 shows the results for a fixed treatment and confounding effect for all graphical structures considered.

When the confounding variable is simply ignored for the graphical model in the first column in Fig. 2, unsurprisingly, we see a constant bias in our estimate of the treatment effect. However, when values of W are used as if they are the perfect observations of U (i.e., when the measurement error is ignored), the bias in treatment effect estimate is reduced. This reduction is particularly significant when W is highly correlated with U (i.e., measurement error is small). Finally, when values of W are used with the effect restoration adjustment, the bias is consistently reduced more than the other approaches.

Furthermore, the smaller the measurement error between W and U, the smaller the bias in estimation. Also, the larger the measurement error, the more the relative benefit of explicitly adjusting for it using effect restoration versus simply ignoring it. However, when U and W are poorly correlated (i.e., measurement error is high), applying effect restoration comes with the cost of increased variance, as shown in the second row in Fig. 2.

For the other graphical models in Fig. 2 columns B, C, D, we observe that ignoring or using W directly provide consistent estimates for the treatment effect. However, simply applying effect restoration might increase bias as well as variance. Hence, applying effect restoration regardless of the underlying structure can result in an incorrect estimate.

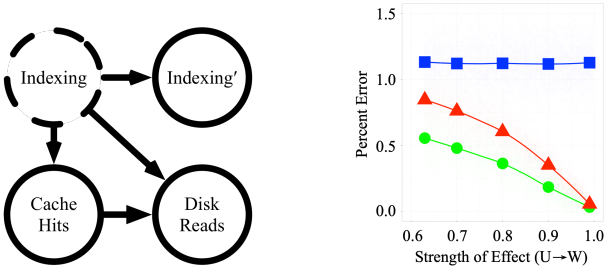
We perform additional experiments for the graphical model in column A by changing the values of treatment and confounding effect via modifying the correlation coefficient. Fig. 3 shows the results of our experiments. In these plots, the treatment effect increases along the big y-axis and confounding effect increases along the big x-axis. Along the x-axis in each plot, the strength of effect between U and W increases. These results suggest that effect restoration is most effective when the treatment effect is small and the confounding effect is high.

5 Detecting the Underlying Structure

In the previous section, we presented empirical evidence that when the underlying structure deviates from the confounding variable case, the adjustment provided by effect restoration can increase bias and variance. This raises a natural question: Can we detect when to apply effect restoration? Instead of assuming that the confounding bias exists, we propose to verify if it exists by using d-separation and typical temporal ordering constraints on the variables.

Note that U may or may not be a confounding variable for X and Y. Our goal is to identify sufficient conditions to determine if U is a confounding variable and only apply effect restoration when it is.

In Table 1, we list all possible underlying structures with variables X, Y, W, and U that satisfy the stated assumptions 1 and 2 in Section 4. There are nine possible graphical structures. We individually account for dependence and independence relationships in each of them. One of the possible structures contains a cycle (i.e., the structure in the last column of Table 1 is not a DAG) and hence out of scope for our discussion. In four of these structures, U is temporally prior to X and Y (i.e., columns A through D). In the remaining four structures, U is temporally posterior to X and Y (i.e., columns E to H.) In the rows of Table 1, for each graphical model, we list



(a) Relationships Between Indexing, Disk Reads, and Cache Hits. System.

(b) Bias in the Estimated Effect of Cache Hits on Disk Reads.

Figure 4: Effect Restoration on PostgreSQL Data.

all the marginal and conditional independence relations between X, Y, and W. Columns of the table correspond to possible underlying graphical models. Each column vector corresponds to the expected conditional dependence and independence relations for the corresponding graphical model.

First, we note that many of these graphical structures are identifiable based on empirically testable conditional independence relations (i.e., structures in columns B, D, and F in Table 1) assuming accurate conditional independence tests. However, some structures are indistinguishable with the given set of conditional independence relations (e.g., structures in A, E, and H share the same set of conditional independence relations).

Second, if we also make a common assumption that covariates are measured *pre-treatment* [24], then only the structures in columns of A, B, C, and D will be possible. Furthermore, conditional independence relations would be sufficient to distinguish among each of these graphical models.

Thus, a combination of common temporal assumptions and conditional independence relations are sufficient to determine the underlying graphical structure from X, Y, and W.

Finally, we note that two of the structures possible when those common temporal assumptions are *not* made introduce significant bias to causal estimates if effect restoration is applied. Applying effect restoration to structure E will remove one pathway for causal effect and thus (typically) underestimate the total effect of $do(X)$. Applying effect restoration to structure H will induce dependence between X and Y even if there exists no direct causal dependence between these variables. This classic example of conditioning on the descendant of a collider would (typically) overestimate the total effect of $do(X)$.

6 Empirical Evaluation of Effect Restoration

In this section, we empirically evaluate the performance of effect restoration in real-world settings as well as demonstrate how effect restoration can be used to combine results from different machine learning models to estimate causal effects.

6.1 Effect Restoration in Real Data We first evaluate the performance of effect restoration on experimental data obtained from real-world settings. We use the experimental data compiled by Garant and Jensen [4] about the effects of interventions on large-scale software systems.

Specifically, we use their experimental data set about *PostgreSQL*, a large open-source relational database management system. The authors ran over 11,000 queries under each of eight different system configurations. This creates a nearly ideal data set for causal estimation because each subject (i.e., query) is observed in each treatment condition (i.e., configuration setting). This approach allows direct interventional estimates of the effect of a treatment on outcome variables in the context of other variables representing characteristics of queries and intermediate states of the database server.

The authors then use *GES*, a score-based algorithm for structure learning [2], to recover the underlying structure for the PostgreSQL domain. From their partially learned graphical structure, we focus on specific regions in which the sub-structures satisfy the effect restoration conditions as shown in Table 1.

We identify two cases in which we can apply effect restoration. One is shown in Fig. 4a and represents the relationships between *indexing* (whether indexes are available), *disk reads* (the level of disk reads), and *cache hits* (the level of cache hits). Generally speaking, indexing affects both disk reads and cache hits. Indexing reduces the disk reads (i.e., a result of a query can be retrieved with fewer block reads) and increases the cache hits. Fig. 4a shows the cache hits as a cause of disk reads. The authors note that the direction of this edge between disk reads and cache hits might also be in the opposite direction. In our analysis, we separately analyzed graphical structures in both directions and the results for effect restoration were similar in both.

For simplicity, we converted each variable to a binary variable by using the median value of each variable as a threshold. Our goal is to estimate the effect of cache hits on disk reads under noisy measurements of indexing. To obtain such noisy measurements, we manually added noise to the values of indexing to obtain *indexing'* by sometimes ignoring the conditional dependence between its true values and

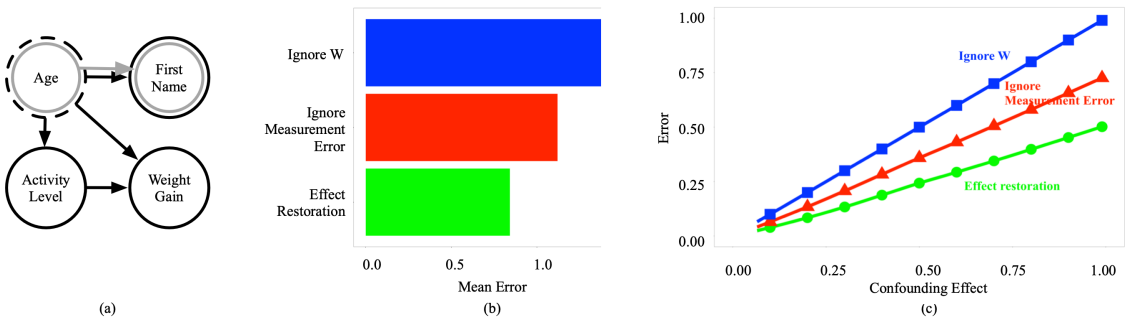


Figure 5: Effect Restoration for Unobserved Variables in Synthetic Data.

randomly assigning values as controlled by a noise parameter. Then we use these noisy measurements to adjust for the confounding bias of indexing. As in our prior experiments, we compare three approaches: (1) Ignoring the values of noisy measurements for indexing, (2) Using the values of indexing, ignoring that they are noisy, (3) Using the values by correcting for the noise in their measurements. We calculate the true effect by using the true values of indexing.

In Fig. 4b, we plot the normalized error in the estimated treatment effect with respect to measurement error in indexing. As with our results on simulated data, estimation with effect restoration provides significantly smaller bias than alternatives. In addition, as the measurement error of the confounding variable increases (i.e., the strength of effect between U and W decreases), the relative benefit of applying effect restoration increases over simply ignoring the measurement error.

6.2 Effect Restoration with Predictive Models Conditioning on a confounding variable by using noisy measurements can also be thought of conditioning on an unobserved variable when both its predictions and the error distribution of estimations can be inferred using an independent estimation process. For example, assume we want to estimate the effect of *activity level* on *weight gain*. Assume *age* is a confounding variable (i.e., age causes both activity level and weight gain) and that it is unobserved for the population of interest. Clearly, in this example, we need to adjust for the effects of age to get an unbiased estimate. One idea is to estimate age by using other observables for the subjects under study. For example, such observables can be based on users' social media content [13], links on social networks [27], first names [14], or names with personal images [3].

We can use the predictions of the model as noisy measurements of age and, given its error distribution, adjust for its confounding bias. This idea assumes that the population under study is similar to the population from which the predictive model was estimated so that we can transfer the knowledge of that model to obtain the corresponding predictions and error distribution. This assumption, referred to as external validity or transportability, is common to nearly all statistical modeling.

6.2.1 Empirical Results on Synthetic Data We evaluate the idea of using predictive models for effect restoration by constructing a scenario in which we observe activity levels, and weight gain recordings of a population along with their first names. Although one can use more complicated models for predicting age, here we use the model described by Oktay et al. [14] because of its simplicity.

In our experiments, we generate synthetic data for each subject with activity levels, weight gain, and a first name. We then use the first names to infer an age value for each subject. Finally, we use the inferred age values along with the corresponding error distribution as reported by Oktay et al. [14] to adjust for the confounding effect of age.

One might suggest that instead of using a model to estimate age, we could simply condition on the values of first names. There are three reasons to avoid this approach. First, the number of possible first names is large and using first names directly can lead to a high-variance estimator of causal effect for most data sets. Second, conditioning on first names leaves the back-door path unblocked between activity levels and weight gain, as shown in Fig. 5 (see Pearl [17] for a detailed discussion of back-door paths). This implies that the confounding bias would still exist. Third, our proposal is to plug in *any* predictive model for the unobserved confounder, and such models can use many independent variables rather than just one. Again, conditioning on many independent variables can lead to a high-variance estimator. In fact, from the perspective of effect

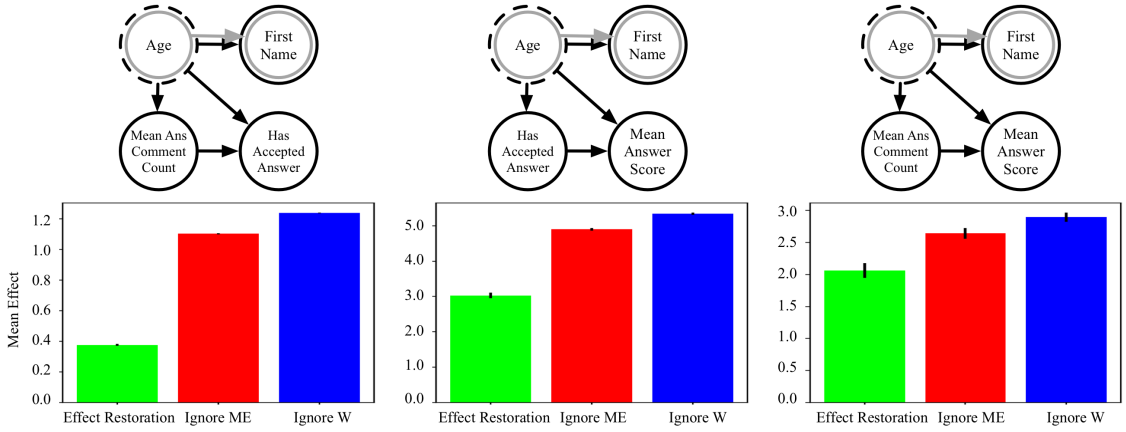


Figure 6: Effects of Unobserved Variables on StackOverflow Data.

restoration, it may be desirable to use high-capacity models in causal estimation process to drive down prediction error and subsequently reduce bias.

In this way, our proposed approach—using a predictive model for effect restoration in pursuit of a low-variance estimator of causal effect—is similar in spirit to the use of predictive models for propensity score matching [19]. Both approaches use a predictive model to summarize the effect of a potentially large number of variables.

Fig. 5a shows the graphical model representation of the experiment. As before, we compare estimating the effect when a confounding variable is ignored, when the error in the confounding variable is ignored, and finally when effect restoration is used. Our results with an independent estimation process are similar to those obtained earlier from both the simulation and real data experiments. We observe the most reduction in bias when correction based on measurement error is used. We also show that as the influence of confounding variable increases, the relative benefit of effect restoration increases.

6.2.2 Empirical Results on Stack Overflow Data We demonstrate an application of effect restoration using predictive models on real data from a popular programming questions and answers site, *Stack Overflow*. Our data consists of *questions*, *answers* and *users* in Stack Overflow from 2008 to early 2018. Users interact by posting questions, answering existing questions, and voting. For each question and answer, an associated score is accumulated as users vote on them. Furthermore, user reputations are built based on the scores of their posts.

Recent studies suggest that different demographic groups behave differently on Stack Overflow [7, 21]. Hence, practitioners might want to condition on demographic variables while estimating causal effects, which are often unobserved. For example, age can be an important variable to consider. However, despite the users' having the ability to self-report their age, the data is often missing and potentially incorrect. One way to adjust for the effect of age is to apply predictive models to infer users' ages and then use those estimated values and error distribution to perform effect restoration. Again, we employ the age model described by Oktay et al. [14].

We use 51 treatment-outcome pairs obtained from *question* and *answer* variables. Using conditional independence tests on the variable pairs, as described in Section 5, we identify causal structures where age can be a confounding variable. Specifically, we use chi-squared independence tests to get the marginal and conditional independence between all pairs of variables. Due to a large sample size ($> 400k$), the p-values are significant even for low effect sizes. Thus, we focus on marginal independence effect sizes that are greater than 0.1. This narrows down the set of possible variable pairs to 14, out of which we consider three structures with the highest confounding effects. We use continuous treatment and outcome variables, and binary confounder variables for the three structures.

Fig. 6 contains the graphical model representations of the top three causal structures and the observed effects using the three causal estimation methods. Similar to our results on the simulated data, conditioning on estimated values reduces effect size estimates more than simply ignoring the confounding variables. Furthermore, applying effect restoration reduces the effect size estimates even more. This suggests that we are able to remove the effect of age on our treatment and outcome variables.

7 Conclusions and Future Work

In this paper, we demonstrate the effectiveness of using effect restoration to combine results from different machine learning methods. First, we characterize the behaviour of effect restoration with measurement error under several plausible graphical structures. We show that it is desirable to use effect restoration only in one of the four possible graphical models. In our simulation analysis, effect restoration adjustment is most effective for small treatment and large confounding effects. Next, we show that the combination of common temporal assumptions and d-separation rules can identify if the underlying structure matches the conditions under which effect restoration is effective. We also provide empirical evidence that effect restoration can reduce bias on causal estimation tasks in real data. Finally, we show that this approach can be used to adjust for unobserved confounding when used with independent predictive models and their corresponding error distributions.

Several future research directions appear promising. First, our empirical study of effect restoration could be generalized for mixed-type data sets. Second, the use of high-capacity predictive models and their limitations could be studied. Third, the implications of effect restoration for estimating joint causal structures could be further explored.

References

[1] J. D. Angrist, G. W. Imbens, and D. B. Rubin, *Identification of causal effects using instrumental variables*, Journal of the American Statistical Association, 91 (1996), pp. 444–455.

[2] D. M. Chickering and C. Meek, *Finding optimal Bayesian networks*, in UAI, 2002, pp. 94–102.

[3] A. Gallagher and T. Chen, *Estimating age, gender and identity using first name priors*, in CVPR, 2008.

[4] D. Garant and D. Jensen, *Evaluating causal models by comparing interventional distributions*, in SIGKDD Workshop on Causal Discovery, 2016.

[5] K. Hirano, G. Imbens, and G. Ridder, *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica, 71 (2003), pp. 1161–1189.

[6] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press, 2009.

[7] G. Kowalik and R. Nielek, *Senior programmers: Characteristics of elderly users from Stack Overflow*, in SOCINFO, 2016, pp. 87–96.

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

[9] M. W. Lipsey and D. B. Wilson, *Practical Meta-Analysis.*, Sage Publications, Inc, 2001.

[10] W. Liu, S. J. Kuramoto, and E. A. Stuart, *An introduction to sensitivity analysis for unobserved confounding in non-experimental prevention research*, Prevention Science, 14 (2013).

[11] M. Long, J. Wang, G. Ding, D. Shen, and Q. Yang, *Transfer learning with graph co-regularization*, in AAAI, 2012.

[12] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, *Causal transfer learning*, CoRR, (2017).

[13] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, *"How old do you think I am?": A study of language and age in Twitter*, in ICWSM, 2013.

[14] H. Oktay, Z. Ertem, and A. Firat, *Demographic breakdown of Twitter users: An analysis based on names*, in ASE SocialCom, 2014.

[15] S. J. Pan and Q. Yang, *A survey on transfer learning*, IEEE Transactions on Knowledge and Data Engineering, 22 (2010), pp. 1345–1359.

[16] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., 1988.

[17] ———, *Causality: Models, Reasoning and Inference*, Cambridge University Press, 2nd ed., 2009.

[18] M. A. Pourhoseingholi, A. R. Baghestani, and M. Vahedi, *How to control confounding effects by statistical analysis*, Gastroenterology and Hepatology From Bed to Bench, 5 (2012), pp. 79–83.

[19] P. R. Rosenbaum and D. B. Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika, 70 (1983), pp. 41–55.

[20] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, The MIT Press, 2nd ed., 2000.

[21] Stack Overflow, *Developer survey results*, 2017. `https://insights.stackoverflow.com/survey/2017`, Last accessed on 2019-01-23.

[22] E. A. Stuart, *Matching methods for causal inference: A review and a look forward*, Statistical Science, 25 (2010), p. 1.

[23] S. Triantafillou, I. Tsamardinos, and I. Tollis, *Learning causal structure from overlapping variable sets*, in AISTATS, 2010, pp. 860–867.

[24] C. Winship and M. Sobel, *Causal Inference in Sociological Studies*, Sage Publications, 2004.

[25] Q. Yang, V. W. Zheng, B. Li, and H. H. Zhuo, *Transfer learning by reusing structured knowledge.*, AI Magazine, 32 (2011), pp. 95–106.

[26] A. Zagorecki and M. J. Druzdzel, *Knowledge engineering for Bayesian networks: How common are noisy-max distributions in practice?*, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43 (2013), pp. 186–195.

[27] F. A. Zamal, W. Liu, and D. Ruths, *Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors*, in ICWSM, 2012.