Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/29/20240825

Machine learning drug discovery based on graph neural network and large language model

Tianqi Huang

Cranbrook schools, Bloomfield Hills, USA

martinhuang2007@163.com

Abstract. The COVID-19 pandemic has presented an urgent need to understand the long-term health implications faced by survivors. Post-COVID-19 complications, such as acute kidney injury, arrhythmia, and stroke, pose significant challenges to public health. Despite extensive research on COVID-19 complications, a comprehensive understanding of the risk factors remains elusive due to the potential confounding variables present in the data. Traditional statistical models, while insightful, may not fully capture the causal relationships between these risk factors and post-COVID-19 complications. Motivated by this gap in the literature, we propose a novel approach using causal inference models to predict the likelihood of post-COVID-19 complications based on patient demographics and pre-existing conditions. Our model, trained on a dataset of COVID-19 inpatients in Wuhan Province, China, estimates the causal effect of these factors on the likelihood of patients experiencing post-COVID-19 complications. This approach allows us to isolate the causal impact of each factor while accounting for potential confounders, providing a more accurate understanding of the underlying mechanisms driving these relationships. Unlike traditional models that predict the probability of certain outcomes, our model provides insights into the causal relationships between risk factors and complications, offering a more reliable and comprehensive understanding of the underlying mechanisms. This approach can help identify at-risk patients, inform targeted interventions, and contribute to the development of effective prevention and treatment strategies. Our work aims to contribute to the current understanding of the virus and inform public health policies and interventions.

Keywords: Causal Inference, Post-COVID Complication, Risk Factors, Public Health

1. Introduction

Post Covid-19 sequelae [1, 2] such as acute kidney injury [3], arrhythmia [4] and stroke [5] have been a formidable challenge for the society throughout the pandemic, which have a consistent impact on public health. Over 30% of COVID-19 patients have reported persistent syndromes up to 9 months after recovery [6]. The increasing concern in the public, reflected by the many reports on the Post Covid-19 syndrome and complications [7], make us wonder what factors are causally responsible for post-COVID sequelae. Although it has been argued by some people that the COVID-19 pandemic is now over, considering the long-term effects of sequelae, we would still benefit a lot from a better understanding of the risk factors.

DOI: 10.54254/2753-8818/29/20240825

The existing body of literature on COVID-19 complications has extensively explored the statistical associations between risk factors and complications, providing insights by identifying highly correlated risk factors for certain sequelae [8]. However, relying solely on statistical dependencies may not provide a comprehensive understanding of these risk factors. When dealing with COVID-19 datasets that include extensive patient demographics, there may be numerous confounding variables present that can produce spurious relationships between variables that are not actually causally related. For instance, the age of inpatients can be highly correlated with other pre-existing conditions such as diabetes and heart disease. For the sake of argument, suppose previous heart disease is identified as a primary risk factor for COVID-19 cardiovascular complications. We may observe a statistical association between diabetes and cardiovascular complications in the data, but it is not necessarily true that diabetes causes cardiovascular complications. Instead, the observed association may be the result of confounding variables such as age, which can affect the likelihood of developing diabetes, the likelihood of heart disease, and the risk of cardiovascular complications.

Therefore, it is essential to carefully consider the potential confounding variables when analyzing the relationship between patient demographics and COVID-19 complications. By addressing the impact of confounding factors, we can gain a more accurate understanding of the underlying mechanisms that drive these relationships, which can lead to more effective prevention and treatment strategies. Establishing a causal relationship between risk factors and COVID-19 complications is also essential for clinicians to accurately determine whether a patient is at a high risk of developing complications. However, understanding the pathology of sequelae can be challenging from the perspective of medicine. The mechanisms underlying the development of sequelae can be complex and multifactorial, and conducting the necessary longitudinal studies can be both costly and time-consuming.

Despite these challenges, innovative approaches in data science may offer new breakthroughs in understanding the causal relationships between risk factors and complications. In our research, we collect the dataset of COVID-19 inpatients in the Wuhan Province, China. The dataset indicates the presence of Post COVID conditions for each patient respectively. While the virus affects individuals differently, evidence suggests that certain patient demographics and past medicine conditions may be more susceptible to severe complications. Taking this into consideration, we train a causal inference model that gives a causal effect estimation of the probability that COVID-19 inpatients will experience sequelae, according to their age groups, gender, clinical symptoms, severity of COVID-19, and prior comorbidities and habits of patients, etc. Unlike machine learning models or statistical models that only predict the probability of certain sequelae for a patient, our causal model renders understanding the cause of COVID-19 complications possible by highlighting the responsible demographics and/or medical conditions for the clinicians and researchers. This causation provided by our model is more stable since it considers the effect of confounders and can provide a comprehensive evaluation on the relevance of Pre-COVID symptoms or conditions against sequelae that patients may potentially experience. In all, this model can help identify at-risk patients and inform targeted interventions to prevent and manage long-term COVID-19 complications.

By providing a comprehensive analysis of the causal relationship between patient demographics and COVID-19 complications, this paper aims to contribute to the current understanding of the virus and inform public health policies and interventions.

2. Background Description

Traditionally, machine learning and statistical models have been employed to identify associations between patient characteristics and COVID-19 complications. These models have been useful in detecting patterns and correlations in the data. However, they primarily focus on associations rather than causation, which can limit their ability to inform targeted interventions or provide insights into the underlying mechanisms driving the observed relationships.

Causal inference, on the other hand, aims to elucidate cause-and-effect relationships in observational data. By employing causal inference techniques, researchers can obtain a more profound understanding of the causal mechanisms that may contribute to the development of COVID-19 complications. This

DOI: 10.54254/2753-8818/29/20240825

approach can help isolate the direct impact of specific patient demographics and pre-existing medical conditions on the likelihood of severe outcomes, which is critical for developing effective prevention and treatment strategies.

Causal inference models differ from traditional machine learning and statistical models in several ways. Firstly, they explicitly model the causal relationships between variables, allowing for the estimation of causal effects. Secondly, they account for confounding factors, which may bias the observed associations between variables. Thirdly, causal inference models can provide insights into potential interventions, enabling researchers to predict the impact of changes in specific patient characteristics on the risk of COVID-19 complications.

3. Experimental

3.1. Problem Description

The COVID-19 complication tabular dataset can be identified by the patients' demographics, symptoms and prior diseases before hospital admissions $A \in Rn \times r$, and the COVID-19 complications $B \in Rn \times h$. Here, n denotes the number of patients in the dataset. r and h denote the number of features for A and B, respectively. Our main goal in this paper is to use causal inference to analyze the average treatment effect (ATE) of each feature $T \in Rn,\ T \in A$ for each feature $Y \in Rn,\ Y \in B$ among the populations, while removing the confounding factors from other features $X = A - T$. In other words, $T \rightarrow Y$ with $T \perp\!\!\!\perp X$, where "→" denotes the causal relationship and "⫫" denotes the decorrelation.

To achieve this goal, we need to make several assumptions for our causal inference setup:

1. SUTVA (Stable Unit Treatment Value Assumption), which asserts that the treatment effect *T* for a given individual is independent of the treatment assignment or outcome of any other individual in the study. In other words, SUTVA posits that there is no interference or spillover effect between people in the treatment and control groups, and that the treatment effect is consistent and stable for a particular unit.
2. Ignorability (or Unconfoundedness): This assumption posits that, conditional on the observed confounders X, the treatment assignment *T* is independent of the potential outcomes *Y*. In other words, once we control for *X*, we think there are no unmeasured confounders that could affect the relationship between *T* and *Y*.
3. Positivity (or Overlap): This assumption ensures that there is a non-zero probability of observing each level of treatment *T* for every combination of confounders *X*. This is crucial for estimating causal effects, as it ensures that we have sufficient data to compare the outcomes under different treatment conditions.
4. Consistency: This assumption requires that the observed outcome *Y* for an individual is equal to the potential outcome under the treatment they received. In our context, this means that the observed COVID-19 complication for each patient corresponds to the potential complication under their actual demographic and pre-existing medical conditions.

The above assumptions are very important in the literature of causal inference. The question, however, is whether these assumptions align with the real-world situation. We do not want to analyze results based on unreliable assumptions. Let us examine the assumptions:

1. SUTVA (Stable Unit Treatment Value Assumption): Given that we're working with medical data, SUTVA is reasonable because medical treatments and outcomes for one patient generally do not affect another patient's treatment or outcome. For example, the demographics, symptoms, or pre-existing diseases (the treatment *T*) of one patient have no influence on the same factors of another patient. Each patient's case can be considered independently of others, which is in line with the SUTVA assumption.
2. Ignorability (or Unconfoundedness): The data set includes comprehensive information about each patient (demographics, symptoms, and prior diseases), which are all factors that could influence their treatment and outcome. By controlling for these variables, we can reasonably assume that these cover most of the confounders in the world that could affect the relationship

DOI: 10.54254/2753-8818/29/20240825

between *T* and *Y*. This is however a strong assumption, since there might still be variables not included in our data that could affect the treatment (like genetic factors etc.).

3. Positivity (or Overlap): With a large dataset of patients, we're likely to have a wide variety of demographics, symptoms, and pre-existing diseases represented. This diversity in the data ensures a non-zero probability of observing each level of treatment *T* for every combination of confounders *X*. However, positivity could be violated if there are rare combinations of patient characteristics that we might not have in our data, thus making it hard to infer the treatment effect for these combinations.
4. Consistency: Given that the data set records the observed complications for each patient, it can be reasonably assumed that these outcomes are the ones that would be observed under their actual demographics and pre-existing conditions. This assumption is reliant on accurate data recording and collection. It also assumes that the complications are entirely attributable to COVID-19, which may not always be the case if patients have other concurrent health issues.

Therefore, we can generally trust that these assumptions hold in real practice. It's important to note that while these assumptions are reasonable, unmeasured confounding variables or unexpected interferences can still occur. These assumptions provide a basis for the analysis, but should be considered with caution. Especially, the ignorability assumption, as raised by many previous works, is often hard to satisfy. Even though the data we use has as many as 38 variables, it may still ignore some significant variables that can have huge impact on a certain complication. To deal with this, we will develop an unbiased model in the experiment to alleviate the error of violating this assumption in the real practice.

Given these assumptions, we can proceed with the estimation of the average treatment effect (ATE) for each *T* on each *Y*. This estimates the counterfactual outcomes for each patient under different treatment conditions and computing the average difference in outcomes between treated and untreated patients. By doing so, we can isolate the causal effect of each feature *T* on the COVID-19 complications B while accounting for potential confounders *X*.

3.2. Methodology

The major difference between causal inference analysis and statistical analysis is whether the effect of confounding factors is removed or had something changed in the system we are observing. ATE analyzes causality of each individual to each complication, which reveals the pathological mechanisms rather than clinical risk factors. ATE is not easily affected by any changes in the data distribution, because it eliminates the influence of other variables in the estimation, so the difference in risk caused by other variables in different populations will not show up. Statistical analysis provides us with the probability (risk factor) in the population that a symptom will find specific sequelae, while it can be influenced by other correlated variable, and is thus vulnerable to distribution shifts and does not reveal the pathological mechanisms. Note that using a single variable for predicting the risk factor cannot remove the influence of other variables, even if other variables are not inputs to the estimation model.

Specifically, we want to estimate the ATE as accurately as possible for providing the insights of the cause of a certain COVID-19 complication. To provide an accurate estimation, here we develop two lines of models for cross-validation: the propensity score estimation line and the structural equation model line. Propensity Score Estimation (abbreviated as ATE directly) is commonly used in observational studies to estimate the effect of a treatment or intervention on an outcome, while accounting for confounding variables. The primary goal is to create a situation similar to a randomized control trial where treatment assignment is independent of potential confounders. Propensity scores denote the probability of a patient receiving the treatment given the observed characteristics. These scores are then used to balance the treated and control groups on these observed characteristics to calculate the causal treatment effect. Structural Equation Modeling (SEM) is a multivariate technique used to analyze structural relationships between measured variables and latent constructs. This technique is used to analyze the structural relationship between measured variables, and provide a cross validation whether it aligns with the ATE obtained from our propensity score estimation models.

DOI: 10.54254/2753-8818/29/20240825

Propensity score estimation is primarily used to address confounding and estimate causal effects in observational studies where random assignment is not feasible. It focuses on the treatment effect on an outcome while controlling for observed covariates. Propensity score estimation relies heavily on the assumption of ignorability, meaning that given the propensity score (the estimated probability of treatment), the assignment to the treatment group is independent of the potential outcomes. If this assumption holds, then propensity score methods (such as matching, weighting, or stratification) can provide unbiased estimates of the average treatment effect. However, if there are unobserved confounders—variables that affect both the treatment assignment and the outcome, but are not included in the propensity score model—then this assumption will be violated, leading to biased estimates. To tackle this, we leverage an unbiased ATE method, i.e., the doubly robust estimation.

Let us denote Y as the outcome variable, T as the binary treatment assignment variable (1 if treated, 0 if control), X as a vector of covariates, p(X) as the true propensity score (i.e., the probability of treatment given the covariates E[T|X]), p'(X) as the estimated propensity score, $\mu_1(X)$ as the conditional expectation function for the treated group (E[Y|T=1, X]), $\mu_0(X)$ as the conditional expectation function for the control group (E[Y|T=0, X]). Let us also denote $\mu_1'(X)$ and $\mu_0'(X)$ as the estimated outcome regression functions. A naive propensity score estimation that is misspecified would be

$$ATE_{ps} = \frac{1}{N}\sum\left[\frac{T_i Y_i}{p'(X_i)} - (T_i - p'(X_i))\,\mu_1'(X_i)\right] + \frac{1}{N}\sum\left[\frac{(1-T_i)\,Y_i}{1-p'(X_i)} + (T_i - p'(X_i))\,\mu_0'(X_i)\right]$$

while the unbiased doubly robust estimation would be

$$ATE_{DR} = \frac{1}{N}\sum\left[\frac{T_i(Y_i - \mu_1'(X_i))}{p'(X_i)} + \mu_1'(X_i)\right] - \frac{1}{N}\sum\left[\frac{(1-T_i)(Y_i - \mu_0'(X_i))}{1-p'(X_i)} + \mu_0'(X_i)\right]$$

This doubly robust estimator will be unbiased if either $p(X) = p'(X)$ (the propensity score model is correctly specified), or $\mu_1(X) = \mu_1'(X)$ and $\mu_0(X) = \mu_0'(X)$ (the outcome model is correctly specified), which alleviates the estimation errors of violating the ignorability assumption.

SEM, on the other hand, is a comprehensive approach that allows for the examination of complex relationships among variables, both observed and unobserved (latent variables). The SEM consists of two parts: (1) the measurement model and (2) the structural model. The measurement model describes how the observed variables (indicators) measure the latent variables (factors). If we have M latent variables and each latent variable η is measured by J observed variables Y, then we could write: $Y_j = \lambda_j * \eta + \delta_j$, where Y_j represents the j-th observed variable within the total J variables, η is the latent variable, λ_j is the factor loading of the j-th observed variable on the latent variable. It describes the relationship between the latent variable and its indicators, and δ_j is the error term associated with the j-th observed variable. The structural model describes the relationships among the latent variables, which are similar to the relationships examined in multiple regression. If we have M latent variables (η_m, for $m = 1, \ldots, M$), we can write the structural model as:

$$\eta_m = \beta_{m1}\eta_1 + \beta_{m2}\eta_2 + \ldots + \beta_{mM}\eta_M + \zeta_m$$

where η_m is the m-th latent variable, β_{mi} is the regression coefficient of the i-th latent variable in the equation of the m-th latent variable. This matrix describes the relationships among the latent variables, and ζ_m is the error term associated with the m-th latent variable. The model is estimated using the generalized least squares (https://github.com/uber/causalml/tree/master).

3.3. Data analysis (setup)

In our code, we have used the CausalML package[1] developed by Uber to create a machine learning model which provides an accurate analysis of the tendency of COVID-19 inpatients to experience post COVID-19 sequelae. However, it is noteworthy that a number of data in the dataset is continuous, such

DOI: 10.54254/2753-8818/29/20240825

as Age and Duration of Death; due to the fact that the ATE estimation is not compatible with processing continuous data such as age, we have discretized such data into several categories in the dataset. This was done by separating the portion of data into several reasonable intervals which can be processed. As for age, the data has been separated into the following intervals: [0, 18], [18, 30], [30, 60], [60, 80], [80, 100]. This has been done considering the different impact of prior comorbidities of different age groups on post COVID-19 sequelae, where [0, 18] represents adolescents, [18, 30] represents young adults, [30, 60] represents adults, [60, 80] represents the elderly, and [80, 100] represents very old adults. And as for days of hospital stay, the data has been separated into the intervals as follows: [0, 7], [7, 28], [28, 200].

Then, we created a causal inference model using the LRSRegressor (linear regression) function, an OLS S-Learner model from the CausalML package. This model is designed to output the causal relevance between prior comorbidities and post COVID sequelae. Average Treatment Effect (ATE) is used to represent the degree of causal relevance between prior comorbidities and post COVID sequelae among the entire populations, in a range between [-1, 1], where -1 represents a negative impact of prior comorbidities on the sequelae, and 1 representing a positive impact of prior comorbidities on the post COVID sequelae outcome.

4. Results and discussion

4.1. Dataset description

In our research, we have used two datasets[1] (Tables 1&2) of inpatients from the COVID-19 pandemic in Wuhan. The primer contains data of 9 demographic variables and 29 different types of prior comorbidities and habits from a range of 36359 patients. These pre-COVID conditions are as follows: (0) 'severity of COVID-19', (1) 'age', (2) 'sex', (3) 'smoking', (4) 'drinking', (5) 'Diabetes', (6) 'Hypertension', (7) 'Hyperlipidemia', (8) 'Heart disease', (9) 'Cancer', (10) 'COPD', (11) 'Tuberculosis', (12) 'Chronic kidney disease', (13) 'Liver disease', (14) 'Intracerebral hemorrhage', (15) 'Asthma', (16) 'Fever', (17) 'Cough', (18) 'Hemoptysis', (19) 'Muscle soreness', (20) 'Fatigue', (21) 'Running nose', (22) 'Pharyngalgia', (23) 'Shortness of breath', (24) 'Chest pain', (25) 'Anorexia', (26) 'Nausea and vomiting', (27) 'Diarrhea', (28) 'Stomachache', (29) 'Deadache', (30) 'Dizziness', (31) 'Disturbance of consciousness', (32) 'prior stroke', (33) 'Duration of hospital stay, days', (34) 'Platelets', (35) 'Albumin', (36) 'ALT', (37) 'Prothrombin time'.

The second dataset contains information of 12 different post COVID-19 conditions from a range of 2493 patients. These post COVID-19 conditions are as follows: (0) 'ARDS', (1) 'Respiratory_failure', (2) 'Acute_myocardial_infarction', (3) 'Acute_myocardial_injury', (4) 'Arrhythmia', (5) 'Acute_cardiac_insufficiency', (6) 'Acute_kidney_injury', (7) 'Acute_liver_injury', (8) 'Gastrointestinal_bleeding', (9) 'Acute_ischemic_stroke', (10) 'Sepsis', (11) 'Coagulopathy'.

4.2. Model description

For comparison, we first develop a Pearson statistical analysis based on linear regression on the combination of each pre-COVID condition and each complication, which represents the risk factor as done by many previous works. Then, we develop the doubly robust propensity score estimator and unbiased Structural Equation Model (SEM) for ATE. We also provide a naive propensity score estimator for ATE estimation, which could very well contain biases introduced by inevitably violating the ignorability assumption. We visualize the results of all three models in heat maps and compare the results.

DOI: 10.54254/2753-8818/29/20240825

4.3. *Result analysis*

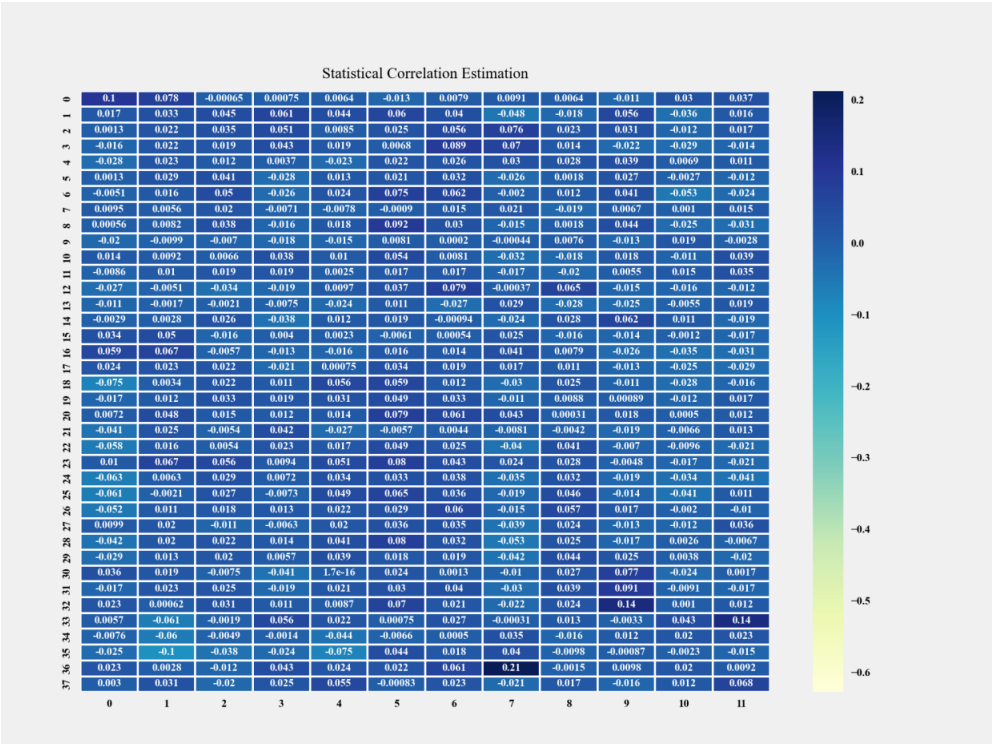

Figure 1. Heatmaps of Pearson statistical correlation.

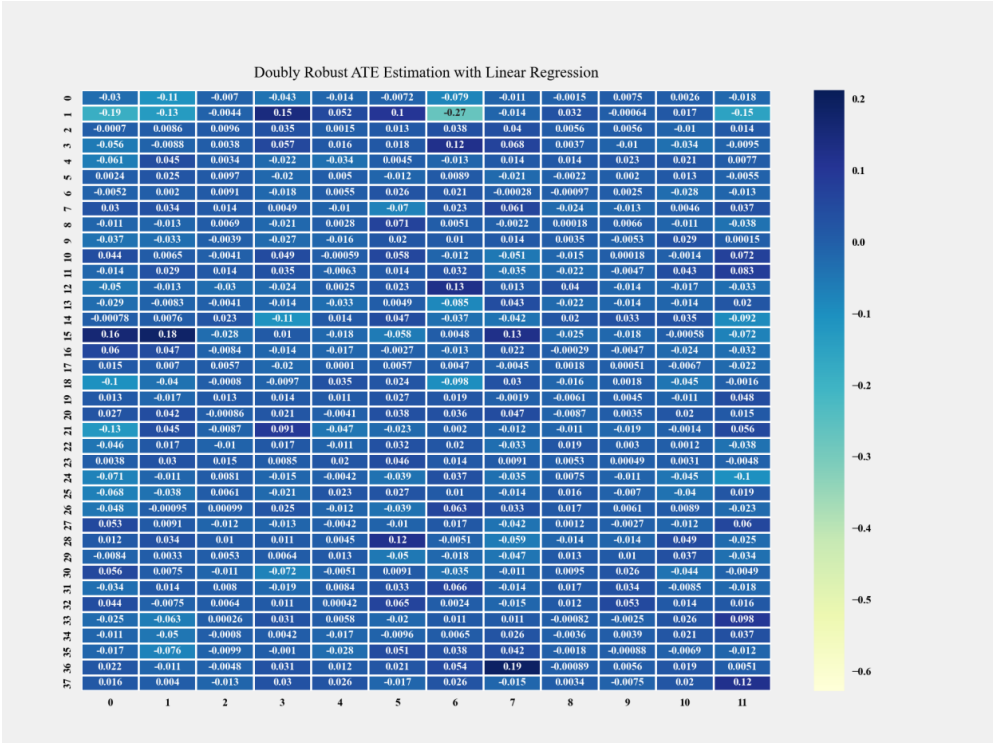

Figure 2. Heatmaps of Doubly Robust ATE estimation with linear regression learner.

DOI: 10.54254/2753-8818/29/20240825

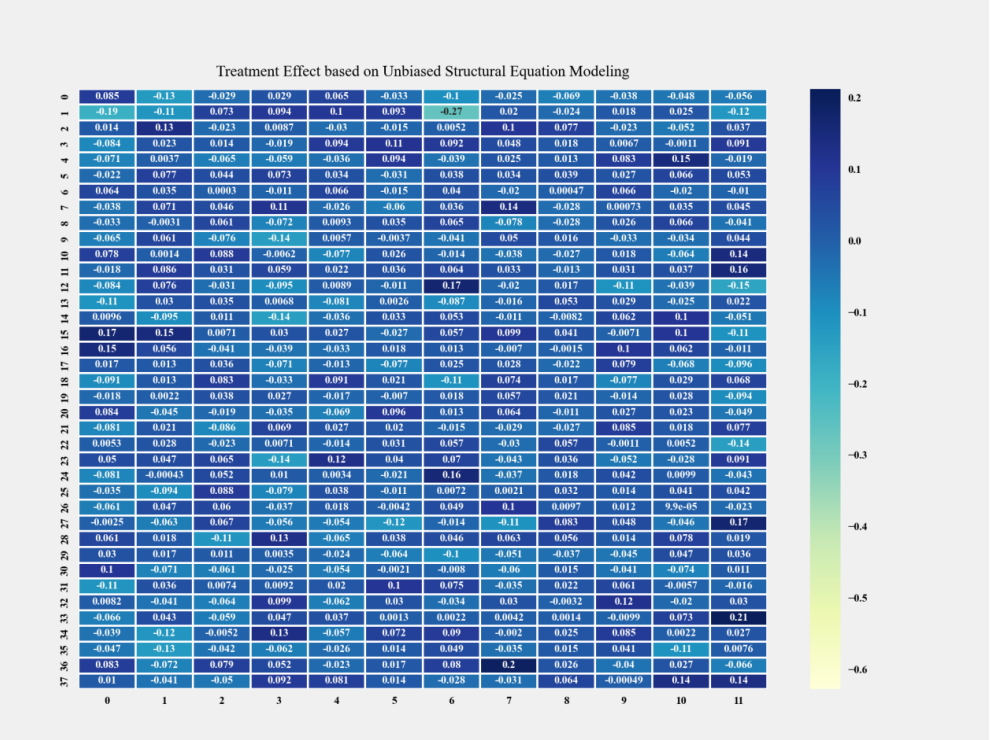

Figure 3. Heatmaps of treatment effect based on unbiased structural equation model.

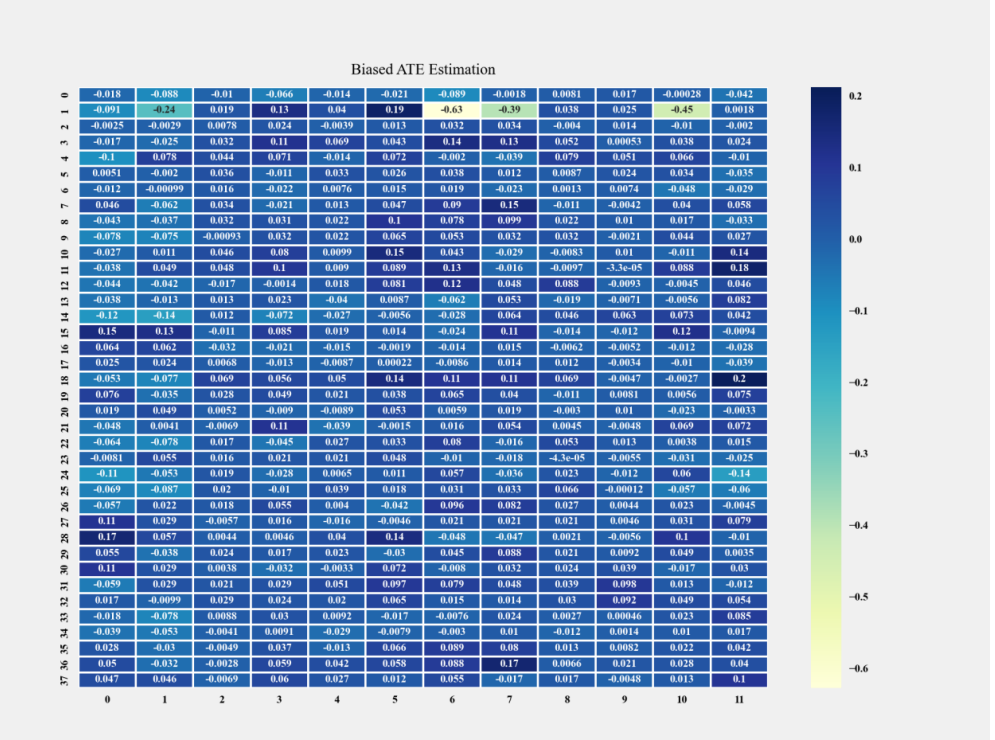

Figure 4. Heatmaps of ATE estimation without bias correction.

4.3.1. Similarity in the heatmaps of ATE and statistical correlation. When comparing the two Heatmaps of correlation and causation (Figure 1 and Figure 2), there are various aspects that we considered despite

DOI: 10.54254/2753-8818/29/20240825

the absence of clear clusters. First, the two heatmaps display a prevalent similarity in their overall patterns and distribution, while the ATE heatmap has a higher overall intensity than the correlation heatmap. Considering the fact that these two separate models are coming from the same data, this consistency indicates an appreciable degree of reliability of the methodology of our ATE estimation, providing confidence that the results are not a model specific artifact nor a product of chance. To exemplify, row 36 column 7 (Alanine transaminase & Acute liver injury) shows a consistent similarity between the two models where ATE values of 0.17 and 0.19 have been outputted. The value of 0.17 and 0.19 indicates a strong and robust relationship between Alanine transaminase level and Acute liver injury. This also resembles the statistical correlation value of 0.21, which again proves the credibility of the method of ATE estimation. This is because, even though risk factors and causation reflect different aspects of understanding the COVID-19 complication, it is likely that pure statistical analysis is sufficient for most cases. Otherwise, all the previous works based on statistical analysis methods will all be invalid and lead to mismatch to clinical experiences of physicians, which should have been reported by the physicians long before our research and is thus highly improbable. However, due to the difference between the methodologies of ATE and correlation, there will be some interesting cases we would find insightful where ATE differentiates a lot from the correlation results, for which we will analyze in the following.

4.3.2. The differences in certain factors between ATE and statistical correlation. We observe mismatched results in some underlying factors in the data. For instance, row 18 and column 10 in Figs. 1 and 2 representing the correlation between Hemoptysis and Sepsis in the form of post COVID sequelae has generated rather contrary values of -0.016 from statistical correlation and 0.2 from ATE correlation estimation. Because it is impossible to know the "ground truth" causal treatment effect or correlation, as both results are solely based on analysis of the data, we look for medical literature for references and judgment.

To address this contrast, we took reference from the HOPE Sepsis score [9], which is a scale involving 9 parameters which effectively identifies the risk for COVID-19 patients to develop Sepsis. In the scale, Hemoptysis possesses the highest score amongst various other comorbidities such as smoking, creatinine, decreased BP and decreased SpO2. Considering that the different data sets involved in research will significantly influence the results obtained, statistical correlation will be prone to influences from distribution shift; this explains the inaccurate results generated by statistical correlation on our data in this example that contradicts the HOPE score. Since the research for HOPE Sepsis score was conducted exclusively to investigate Sepsis as a post COVID sequelae, we can reasonably assume that it generates more accurate results than the correlation value of -0.016, given that their dataset has a more balanced ratio of Septic/Non-septic patients and their study is specifically designed for understanding the complication of Sepsis. The similar results between the ATE obtained based on our data and the HOPE score have proven the accuracy of the ATE estimation in this paper. Because ATE removes the confounders of other correlated variables, the result is consistent across different data environments and distributions [10].

There are other examples, such as the pair of Asthma as the pre-COVID condition to ARDS as the complication in row 15 and column 0. Here, the Correlation value = 0.034 and ATE estimation = 0.14. We can also find references that support the validity of our ATE result. In a study for understanding ARDS clinically and statistically from COVID-19 hospitalized patients [11], the authors point out and I quote, "Other viral respiratory diseases, and in particular influenza, are well known for their ability to induce severe respiratory complications, including ARDS, mainly in frail elderly people and in patients with severe comorbidities (obesity, diabetes, or heart or respiratory failure)". This shows that the effect of respiratory problems such as Asthma should have more influence on the ARDS than the correlation value of 0.034. Moreover, it is found that in row 11 and column 11, the statistical correlation coefficient between Tuberculosis and Coagulopathy is only 0.035, while the ATE is as large as 0.18. This gap between ATE and Pearson correlation is again insightful and verifies the importance of ATE estimation. As pointed out by an existing medical literature [12], quoted "Tuberculosis (TB) affects the production

DOI: 10.54254/2753-8818/29/20240825

and life span of all hematologic cellular components" in the paper's abstract, TB indeed has a mechanism that might lead to complications like Coagulopathy, which is largely ignored in Pearson correlation in Fig. 1.

4.3.3. The accuracy of different treatment effect estimation method. We analyze the results from Figures 2, 3 and 4. In order to find the most accurate method of ATE estimation, we assumed that the average of values of all methods should not be biased. However, the nature of methods such as the naive ATE estimation without biased correction is extremely likely to generate inaccurate results, which makes it less reliable. Both SEM and doubly robust estimators attempts to remove biases from different perspectives, therefore, should they be both effective in removing the biases, their results should be similar. As per experimental result shown above, Figure 2 and Figure 3 are more similar to each other while Figure 4 has some anomaly values (colored as yellow in the heatmap) such as the effect of (1) age to complications like (6) Acute_kidney_injury, (7) Acute_liver_injury and (10) Sepsis. Through cross-validation using multiple models to generate correlation values, the overall similarity in patterns and individual values have validated the accuracy and credibility of ATE estimation when applying to post COVID-19 sequelae.

5. Conclusions

Our research has demonstrated the potential of causal inference models in predicting post-COVID-19 complications. By analyzing the causal relationships between patient demographics, pre-existing conditions and post-COVID complications, we have been able to provide a more comprehensive understanding of the risk factors involved. This approach has allowed us to isolate the causal impact of each risk factor while accounting for potential confounders, leading to a more accurate understanding of the underlying mechanisms driving these relationships. The experimental results have shown that, based on the unbiased average treatment effect estimation, we identify specific demographics and pre-existing conditions that are causally related to certain complications. For example, our model identified a strong causal relationship between Hemoptysis and Sepsis and between Asthma and ARDS, which align well with the findings of existing clinical studies. Such strong causal relationships are not able to be discovered by statistical correlation methods such as Pearson correlation in our study, which highlights the insights we get from the developed causal inference models.

However, it is important to note that while our model provides valuable insights, it is not without limitations. The assumptions made in our causal inference setup, while generally reasonable, should be considered with caution. Future work should aim to further validate these assumptions and explore the potential of incorporating additional variables that may impact post-COVID-19 complications. By continuing to refine and build upon this approach, we believe that causal inference models can play a crucial role in the development of effective prevention and treatment strategies for post-COVID-19 complications. As we continue to grapple with the impacts of the pandemic, such research will be vital in informing public health policies and interventions.

References

[1] Long, B., Brady, W. J., Koyfman, A., & Gottlieb, M. (2020). Cardiovascular complications in COVID-19. *The American journal of emergency medicine*, *38*(7), 1504-1507.

[2] SeyedAlinaghi, S., Afsahi, A. M., MohsseniPour, M., Behnezhad, F., Salehi, M. A., Barzegary, A., ... & Dadras, O. (2021). Late complications of COVID-19; a systematic review of current evidence. *Archives of academic emergency medicine*, *9*(1).

[3] Qian, J. Y., Wang, B., Lv, L. L., & Liu, B. C. (2021). Pathogenesis of acute kidney injury in coronavirus disease 2019. *Frontiers in physiology*, *12*, 586589.

[4] Desai, A. D., Lavelle, M., Boursiquot, B. C., & Wan, E. Y. (2022). Long-term complications of COVID-19. *American Journal of Physiology-Cell Physiology*, *322*(1), C1-C11.

[5] Gabriel, J. T. (2020). Stroke as a complication and prognostic factor of COVID-19. *Neurología (English Edition)*, *35*(5), 318-322.

DOI: 10.54254/2753-8818/29/20240825

[6] Rettner, R. (2021). 30% of people with COVID-19 experience symptoms up to 9 months later, https://www.livescience.com/long-covid-19-most-common-symptoms.html

[7] Nalbandian, A., Sehgal, K., Gupta, A., Madhavan, M. V., McGroder, C., Stevens, J. S., ... & Wan, E. Y. (2021). Post-acute COVID-19 syndrome. *Nature medicine*, *27*(4), 601-615.

[8] Zhang, H., Wu, Y., He, Y., Liu, X., Liu, M., Tang, Y., ... & Wang, W. (2022). Age-related risk factors and complications of patients with COVID-19: a population-based retrospective study. *Frontiers in medicine*, 2643..

[9] Abumayyaleh, M., Nunez-Gil, I. J., El-Battrawy, I., Estrada, V., Becerra-Muñoz, V. M., Uribarri, A., ... & Akin, I. (2021). Sepsis of patients infected by SARS-CoV-2: real-world experience from the international HOPE-COVID-19-registry and validation of HOPE sepsis score. *Frontiers in medicine*, *8*, 728102.

[10] Kuang, K., Cui, P., Athey, S., Xiong, R., & Li, B. (2018, July). Stable prediction across unknown environments. In *proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1617-1626).

[11] Beltramo, G., Cottenet, J., Mariet, A. S., Georges, M., Piroth, L., Tubert-Bitter, P., ... & Quantin, C. (2021). Chronic respiratory diseases are predictors of severe outcome in COVID-19 hospitalised patients: a nationwide study. *European Respiratory Journal*, *58*(6).

[12] Balepur, S. S., & Schlossberg, D. (2016). Hematologic Complications of Tuberculosis. *Microbiology Spectrum*, *4*(6), 10-1128.