Patterns, Volume 5

### Supplemental information

## A latent transfer learning method for estimating

#### hospital-specific post-acute healthcare

#### demands following SARS-CoV-2 infection

Qiong Wu, Nathan M. Pajor, Yiwen Lu, Charles J. Wolock, Jiayi Tong, Vitaly Lorman, Kevin B. Johnson, Jason H. Moore, Christopher B. Forrest, David A. Asch, and Yong Chen

# **Supplemental Information**

### **Table of Contents**



# <span id="page-2-1"></span><span id="page-2-0"></span>**Supplemental Experimental Procedure A. Benchmarks**

Traditionally, the estimation of the average treatment effect can be assessed by following the causal inference framework in a conventional single-site setting where only target samples were available. For simplicity, we let  $i = 1, ..., n_0$  indicate subjects in the target hospital. In practical applications, the correct specification of an outcome model or a propensity score model (the model predicting the treatment assignment based on patient covariates) needs adequate background knowledge which is typically lacking. The Augmented Inverse Propensity Weighted (AIPW) estimator is a combination of two models, offering improved robustness. The AIPW estimator is doubly robust in the sense that when either the outcome model or the propensity score model is correctly specified, it produces a consistent estimator of average treatment effect<sup>1</sup>. Specifically, we consider a parametric model  $g_c(a, x_i; \theta_{c,a})$  for the conditional expectation of outcome in each treatment group of c-th subpopulation. The probability of treatment assignment in the cth subpopulation follows  $P(A_i = 1 | x_i) = e_c(x_i; \beta_c)$ .  $\hat{\theta}_{c,a}$  and  $\hat{\beta}_c$  denote corresponding estimated parameters. Then, the AIPW estimator is defined as

$$
\hat{\Delta}_{c} = \frac{1}{n_{0c}} \sum_{i=1}^{n_{0}} \left[ g_{c}(1, x_{i}; \hat{\theta}_{c,1}) - g_{c}(0, x_{i}; \hat{\theta}_{c,0}) + \frac{A_{i} \{ Y_{i} - g_{c}(1, x_{i}; \hat{\theta}_{c,1}) \}}{e_{c}(x_{i}; \hat{\beta}_{c})} - \frac{(1 - A_{i}) \{ Y_{i} - g_{c}(0, x_{i}; \hat{\theta}_{c,0}) \}}{1 - e_{c}(x_{i}; \hat{\beta}_{c})} \right] I(C_{i} = c),
$$

where  $n_{0c} = \sum_{i=1}^{n_0} I(C_i = c)$  be the number of samples from c-th subpopulation. To construct a confidence interval for the AIPW estimator, either the sandwich estimator<sup>2</sup> or non-parametric bootstrap can be utilized for the sampling variance of  $\hat{\Delta}_c$ .

#### <span id="page-2-2"></span>**B. Definitions of study variables**

**Table S1: Variables used in estimating the subpopulation-specific impacts of COVID-19 on healthcare utility in each hospital of PEDSnet network through the Latent-TL pipeline.**







Note: All of the domains in the table above are based on the PEDSnet common data model (CDM). More details are available through this link: [https://data-models-service.research.chop.edu.](https://data-models-service.research.chop.edu/)

## <span id="page-5-0"></span>**C. Population selection workflow**

**Figure S1: Selection workflow of the study population based on PEDSnet COVID-19 Database Version 2024-03-14. Note that the population selection is based on a more recent version of the database than the one used in the original manuscript. Hence, the numbers may differ slightly due to updates in the COVID-19 data submitted by each hospital.**



# <span id="page-6-0"></span>**Note S1 Additional Population Statistics**

**Table S2. Distribution of pre-existing chronic conditions of patients < 21 years of age in the COVID-19 infection, no COVID-19 infection and overall cohorts.**









### <span id="page-10-0"></span>**Note S2 Characterization of Patient Subpopulations**

In **Figure 3** of the manuscript, we use a heatmap to depict the identification of patient subpopulations derived from Multi-Site Latent Class Analysis (MLCA). This heatmap illustrates the incidence of chronic conditions across various subpopulations: each column represents a subpopulation, while rows correspond to distinct clusters of chronic conditions. While **Figure 3** emphasizes the top 50 most common chronic conditions, **Figure S2** offers a broader view, showcasing the incidence rates for all chronic conditions across each subpopulation.

**Figure S2. Chronic condition incidence across identified subpopulations via Multi-Site Latent Class Analysis (MLCA). This heatmap displays the prevalence of chronic conditions across four distinct subpopulations (or latent classes). Columns represent individual subpopulations, while rows detail the incidence rates of specific chronic condition clusters.**



#### <span id="page-12-0"></span>**Note S3 Causal Estimation of Baseline Healthcare Utilizations**

To further elucidate the connection between identified patient subpopulations and healthcare utilization trends, we projected baseline healthcare usage patterns under the theoretical premise where all patients remained unaffected by COVID-19. As delineated in **Figure S3**, the incidence of inpatient visits exhibited significant variation across the four discerned subpopulations, averaging across all hospitals at rates of 3.5%, 1.5%, 1.8%, and 12.4% for classes 1 through 4, respectively. On the other hand, the ED visit rates sustained a level of uniformity, spanning an average range of 9.9% to 13.1% across the quartet of subpopulations. Remarkably, Class 4, which encapsulates children grappling with complex chronic conditions, marked the peak frequency in baseline inpatient and ED visits during a period unaffected by the COVID-19 pandemic.

**Figure S3. Hospital- and subpopulation-specific baseline healthcare utilization patterns. This figure illustrates the baseline patterns of inpatient (panel A) and ED (panel B) visits occurring between 28 and 179 days post-index date under the hypothetical scenario where all patients remained unaffected by SARS-CoV-2 infection.**



### <span id="page-13-0"></span>**Note S4 Negative Control Experiments**

To mitigate potential biases originating from unmeasured confounders or systematic influences, we compiled a list of negative control outcomes. These outcomes, identified based on the ICD hierarchy<sup>3</sup>, are presumed not to be caused by COVID-19 infection (for a comprehensive list please refer to **Table S3**). After excluding exceedingly rare events in our database, 33 negative control outcome variables were incorporated in the analysis. By applying the same analytical strategy to this collection of negative control outcomes, we establish a baseline or "null" distribution, aiding in the accurate identification and correction of potential biases in estimating our primary outcomes of interest. The recalibrated estimates showcased in **Figure S4** indicate a minimal degree of systematic error, alongside a slight increase in uncertainty, as characterized by the distribution estimated from the negative control outcomes.



#### **Table S3. Negative control outcomes.**



**Figure S4: Hospital- and subpopulation-specific COVID-19 effects on inpatient (panel A) and ED (panel B) visits after calibrating negative control outcomes.** 



#### **Reference**:

- 1. Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. J Am Stat Assoc *89*, 846–866. https://doi.org/10.1080/01621459.1994.10476818.
- 2. Stefanski, L.A., and Boos, D.D. (2002). The Calculus of M-Estimation. Am Stat 56, 29–38. https://doi.org/10.1198/000313002753631330.
- 3. Hirsch, J.A., Nicola, G., McGinty, G., Liu, R.W., Barr, R.M., Chittle, M.D., and Manchikanti, L. (2016). ICD-10: History and Context. American Journal of Neuroradiology *37*, 596. https://doi.org/10.3174/ajnr.A4696.