RESEARCH ARTICLE

A causal inference approach for estimating effects of non-pharmaceutical interventions during Covid-19 pandemic

Vesna Barros [1,2] *, **Itay Manes** [1], **Victor Akinwande** [3], **Celia Cintas** [3], **Osnat Bar-Shira** [1], **Michal Ozery-Flato** [1], **Yishai Shimoni** [1], **Michal Rosen-Zvi** [1,2]

1 AI for Accelerated Healthcare and Life Sciences Discovery, IBM Research Labs, Haifa, Israel, **2** Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel, **3** IBM Research Africa, Nairobi, Kenya

* vesna.barros@ibm.com

Abstract

In response to the outbreak of the coronavirus disease 2019 (Covid-19), governments worldwide have introduced multiple restriction policies, known as non-pharmaceutical interventions (NPIs). However, the relative impact of control measures and the long-term causal contribution of each NPI are still a topic of debate. We present a method to rigorously study the effectiveness of interventions on the rate of the time-varying reproduction number R_t and on human mobility, considered here as a proxy measure of policy adherence and social distancing. We frame our model using a causal inference approach to quantify the impact of five governmental interventions introduced until June 2020 to control the outbreak in 113 countries: confinement, school closure, mask wearing, cultural closure, and work restrictions. Our results indicate that mobility changes are more accurately predicted when compared to reproduction number. All NPIs, except for mask wearing, significantly affected human mobility trends. From these, schools and cultural closure mandates showed the largest effect on social distancing. We also found that closing schools, issuing face mask usage, and work-from-home mandates also caused a persistent reduction on R_t after their initiation, which was not observed with the other social distancing measures. Our results are robust and consistent across different model specifications and can shed more light on the impact of individual NPIs.

Introduction

The coronavirus disease 2019 (Covid-19) pandemic has caused an enormous impact on the economy and on global public health. As of January 1st 2022, the disease had over 290 million cases and more than 5 million deaths recorded in over 200 countries and territories [1]. In response to the state of emergency declared by the World Health Organization (WHO) in January 2020, governments worldwide have introduced multiple restriction policies, known as non-pharmaceutical interventions (NPI), to mitigate the spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Today, two years from the first outbreak registered in

Wuhan China, the relative impact of control measures and the long-term causal contribution of each NPI are still a topic of debate.

In June 2020, Flaxman and colleagues [2] pioneered the challenge of estimating the effectiveness of major interventions on the transmission of SARS-CoV-2. On the basis of mortality data and a Bayesian hierarchical model, they concluded that lockdowns had been effective in most European countries that were studied. Since then, due to the increased widespread adoption of restrictions, several studies have attempted to disentangle the effect of individual NPIs in several countries [3–10] or in the US alone [11–17]. Overall, most previous mathematical modeling suggest that public health interventions were associated with a reduction of Covid-19 incidence. Nonetheless, the conclusions regarding the effect of each specific intervention are not unequivocal. For example, [4, 5, 7, 8, 18] point out that school closures had a significant effect on the transmission of new infections, while a review conducted by [19] suggested that closing schools did not contribute to controlling the pandemic in countries like China, Hong Kong, and Singapore. In fact, existing evidence for the impact of policies is not consistent in the literature, as NPI effectiveness may vary across regions depending on the local context [16]. However, even with heterogeneous effects across countries, we found that estimating an overall effectiveness of interventions was possible due to generalization power encountered in cross-country data, which also contained a large enough sample size to detect a causal effect.

In this study, we provide a comprehensive analysis of the employment of five NPIs—confinement, school closure, mask wearing, cultural closure, and work restrictions—in 113 countries during the first 5 months of 2020. We observed that during the first-wave period there was a great deal of consistency in the set of restriction measures imposed throughout the world, which was substantially reduced in the subsequent waves. Thus, we focus on the initial outbreak period, when the long-term consequences of the virus were still poorly understood, vaccines were still not available, and policy-makers were not certain which control measures would be effective. With this approach, we hope to suggest a method to infer the efficiency of restriction measures and to inform future urgent preparedness response plans in the time-critical phase of a pandemic.

We estimate the impact of individual NPIs on social distancing and Covid-19 spread using causal analysis methodology, taking into account confounding factors such as concurrent NPIs, Covid-19 morbidity measures, and country-level socio-economic and demographics factors. We tested two different variables as outcomes in the causal effect estimation. First, we analysed how NPIs influence the mobility trends across different categories of places such as residential and retail/recreation areas. In the second approach, we evaluated the effect on the growth rate of the reproduction number R_t, i.e., the rate by which the pandemic spreads. To the best of our knowledge, this is the first study to quantify NPI effectiveness using a causal inference framework in such a wide geography coverage, while accounting for confounding biases and performing sensitivity analyses to assess the robustness of our findings.

Materials and methods

Data collection and pre-processing

Non-pharmaceutical interventions. We extracted the data on the restriction policies in 113 countries using the Worldwide Non-pharmaceutical Interventions Tracker for Covid-19 (WNTRAC) [20], a comprehensive database consisting of over 8, 000 Covid-19-related NPIs implemented worldwide. WNTRAC was updated periodically until October 5th 2021, and we use the latest release for the analyses described hereby, which covers the period until the day before. For our experiments, we selected a subset of the five most well-defined and frequently imposed NPIs in the data: confinement, entertainment/cultural sector closure, work

Table 1. Description of the five NPIs used in this study.

NPI	Description
School closure	Closing of any school category: kindergartens/daycares, primary/secondary schools or universities.
Cultural closure	Closing of any cultural establishment: bars, restaurants, night clubs, museums, theaters, cinema, libraries, festivities, parks and public gardens, gyms and pools, or churches.
Confinement	Physical distancing and universal lockdown measures. Considered as imposed when it is mandatory for the entire population.
Work-from-home order	Work restrictions and mandatory work-from-home orders for all nonessential workers.
Mask wearing	Facial coverings or mask wearing. Considered as imposed if it was mandatory or recommended in public spaces.

restrictions, mask-wearing, and school closure. Table 1 shows each NPI with its respective description.

Mobility trends. To infer people's dynamic behavioral response to restriction policies, we obtained Google mobility data [21]. These reports include the per-day change in movement across different categories of places compared to a baseline day before the pandemic outbreak. Similar to [17], we chose the change in duration of time spent in residential areas as a primary metric to measure social distancing and policy adherence. We additionally considered the effect of NPIs on the changes of movement in retail and recreation areas, generally seen as nonessential visits. For each category, we smoothed weekly patterns by using the seven-day rolling averaged mobility. When values were missing, we performed a linear interpolation. We used country-level information for all countries in our analysis, except for the US, which contained state-level data for both NPI and mobility data.

Socio-economical and health indicators. We used development indicators compiled from officially recognized international sources to account for heterogeneity in terms of socio-economic and health factors within individual countries. From the World Bank's World Development Indicators (WDI) [22], we selected a subset of variables, including access to electricity, outdoor air pollution, and forest area. We also included country-level health information, ranging from life expectancy at birth, smoking rate, and prevalence of undernourishment. As we have indicators per year, we take the most recent metrics available per country. Similar covariates were used in previous works [3, 4].

We also included variables describing population distribution, age-structure, and human development index from Our World in Data (OWID) [23]. To avoid producing highly biased estimates with missing values imputation, we only analysed countries with data available for at least 70% of the WDI and OWID variables and imputed the remaining missing features with the mean. We characterize each country with a cluster ID obtained through DBSCAN clustering [24] (scikit-learn, version 1.0.1) performed on the vector of socio-economical and health features (see S1 Table in S1 Appendix with the final set of features, S1 Fig in Appendix E1 of S1 Appendix for a detailed explanation on the clustering method and S2 Fig in Appendix E2 of S1 Appendix for analysis on the robustness of missing values imputation).

Covid-19-related variables. Our model used daily cumulative confirmed cases and deaths from the WHO reports [25]. Before estimating R_t, we smoothed the number of new cases over time with a local polynomial regression using a window size of 14 days and a polynomial degree of 1 to minimize the impact on the edges. The 14 day span (7 days forward and backward) was used to ensure that an equal number of weekdays was used for smoothing and to account for a time lag between exposure, testing, and documenting case reports. We used the R package EpiEstim developed by Cori and colleagues [26] and extended by Thompson et al.

[27]. We chose a gamma distributed serial interval (mean, 3.96 days [SD, 4.75 days]; this interval was constant across periods and derived from previous epidemiological surveys on Covid-19 [28]), to estimate R_t and its 95% credible interval on each day via a 7-day moving average. EpiEstim was chosen due to its statistically robust analytical estimates and extensive use in the disease epidemiology literature.

Because estimating the serial interval distribution may not be possible in the early phase of an outbreak, or may be associated with significant uncertainty as countries still had to establish their documentation practices, we exclude the period before 100 cumulative cases of each country from our dataset. Therefore, the window of analysis for each country starts on the day the cumulative cases exceeded 100 and ends on June 1st 2020. When concatenating all data sources, we found an intersection of 113 countries that were used in our final analysis. Countries like China, Iran and part of central African countries were excluded due to missing mobility data (see geography selection in Appendix E3 of S1 Appendix).

Statistical analysis

Outcome prediction.   To investigate whether the observed variables (Covid-19 morbidity measures, socio-economical factors and status of NPIs) contained the predictive power for the outcomes of interest, we built prediction models to estimate R_t and the change in mobility. We split the dataset into two non-overlapping sets: 70% of samples used for training and 30% for validation. We used the gradient boosted trees algorithm with the XGBoost [29] package (version 1.2.0; https://xgboost.ai), with hyperparameters tuned using 5-fold cross validation on the train dataset, via a randomized grid search. The TreeExplainer from the SHapley Additive exPlanations (SHAP) [30] package was fit on the validation set to estimate the association and contribution of each feature to the XGBoost model.

We evaluated the accuracy of the predictions using the mean-squared error (MSE). To compare between the performance of different models at different scales, we normalized the estimates and the observed values in the test set before computing the MSE. The 95% confidence intervals (CI) were obtained with 100 bootstrapping iterations.

Potential outcome framework.   We formulated causal effects in terms of the potential outcome framework [31].

Each country i at each point in time t is characterized by a feature vector $X_{i,t}$, consisting of dynamic variables—the cumulative and new number of cases/deaths per million—and a static variable—a socioeconomic cluster ID, which is country-specific and constant over time. In addition, we included information regarding the status of the restrictions. This is a feature vector containing a set of binary features corresponding to whether other NPIs are in place at that time point. We denote $Y_{i,t}^{(a)}$ as the potential outcome in a given country i on day t for binary treatment a.

We were interested in two possible outcomes: (i) $M_{i,t}^{(a)}$, the difference in the country-level mobility defined as in [32] and (ii) $R_{i,t}^{(a)}$, the change in SARS-CoV-2 transmission represented by a ratio of the reproduction number R_t, defined as in [9]. We evaluated the outcomes with respect to a time-lag w (in days), as follows:

$$M_{i,t}^{(a)} = M_{i,t+w} - M_{i,t} \tag{1}$$

$$R_{i,t}^{(a)} = \frac{R_{i,t+w}}{R_{i,t}} \tag{2}$$

In this study, we considered the effect on the outcomes 2 weeks after the NPIs were enacted ± 1 week (i.e., w = 7, 14 and 21). Our goal was to estimate the Average Treatment Effect on the Treated (ATT), defined as

$$ATT_M = \mathbb{E}_{x \sim X|A=1}[Y|A=1, X] - \mathbb{E}_{x \sim X|A=1}[Y|A=0, X]$$

and

$$ATT_R = \frac{\mathbb{E}_{x \sim X|A=1}[Y|A=1, X]}{\mathbb{E}_{x \sim X|A=1}[Y|A=0, X]}$$

where ATT_M and ATT_R denote the causal effects of human mobility trends and R_t, respectively. Thus, a null effect of NPIs on mobility would be equivalent to $ATT_M = 0$. Similarly, a null effect on R_t corresponds to $ATT_R = 1$.

To identify such a causal effect, we make several assumptions: (i) conditional exchangeability (ii) Stable Unit Treatment Value Assumption (SUTVA) and (iii) positivity. Identifying the ATT requires a weaker version of the assumptions above, as we are only estimating potential outcomes for the treated part of the population. A detailed discussion about the assumptions can be found in Appendix E5 of S1 Appendix.

Study design. A crucial component for estimating causal effect from observational studies is the ability to conceptualize and to emulate randomized experiments. To this end, we designed a cohort study exclusively separating samples (in this case, days) from the treatment and control groups. For each NPI $k \in K$, we defined an *event* as the day NPI k was enacted in a certain country, given that this intervention was not in action the day before. We denote this day t_k^*. We defined the treatment group for NPI k as the set of t_k^* in the 113 countries during their respective period of analysis (Fig 1). The control group includes all days in the cohort except $t_k^* \pm w \,\forall k$ where w is the number of days before and after the event, and it is also the same time-lag period defined in Eqs 1 and 2. The control cohort represents a period where none of the five NPIs were enacted (an event did not occur), but they might still be in place during that period. Unlike the treatment groups, which are NPI-specific, the control groups are the same for all NPIs used as treatment.

Covariate balancing. We used Adversarial Balancing (AdvBal) [33] to estimate the mean potential outcome of the control group. In brief, AdvBal borrows principles from generative adversarial networks to assign weight to each sample in some source data, such that the resulting weighted data becomes similar to a given target data. Unlike the original paper, we defined only the treated group as the target population, and not the entire population, i.e., we fixed the weights of the treated samples and adjusted the weights of the control in order to balance the groups. The implementation of AdvBal and the resulting balancing evaluation of the models were done with the Python package *causallib* [34] (version 0.7.1; https://github.com/IBM/causallib). The final effects and the associated 95% confidence intervals were computed with 1,000 bootstrap samples.

Sensitivity analysis. We performed sensitivity analyses to test our conclusions under different scenarios. First, we reran our framework with a different model to estimate treatment effect from observational data. We used inverse probability weighting (IPW) [35], a longstanding popular method that overcomes confounding by weighting samples by the inverse of their probability of being assigned to their treatment, conditioned on their covariates. These probability parameters were estimated with a logistic regression model. Once confounding was adjusted and the weights were computed, we estimated the causal effect on the treated sample as a weighted average of the outcome.
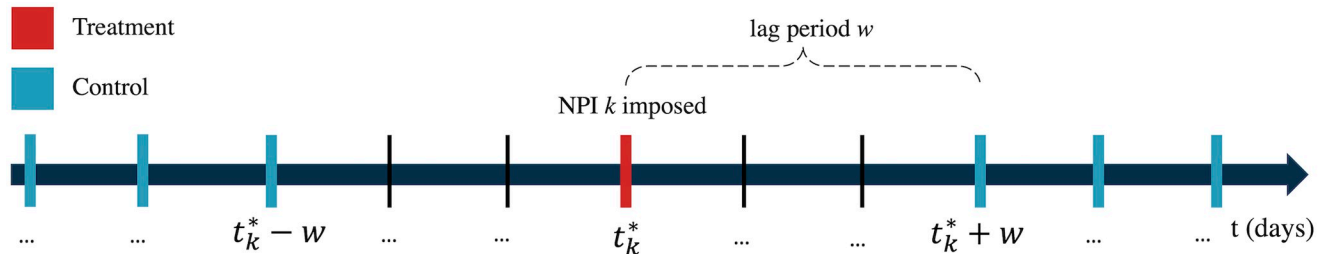
Fig 1. Schematic presentation of the study design. Vertical lines represent days. We denote day t_k^* (in red) as the day when NPI k was enacted, considering that this NPI was not active the day before (i.e., event date). The treatment group for NPI k consists of t_k^* across all countries. The control group are the remaining days (in blue) outside the time interval $t_k^* \pm w$, where w is a lag period.

Second, we performed a complementary analysis to test whether the estimated effects were consistent when using an alternative study design. In this approach, the treatment groups are the same as in the original study (Fig 1). On the other hand, the control group consists of events from the remaining NPIs (apart from the NPI used as treatment). For example, if we were to estimate the effects of NPI $k \in K$, the treated group would contain only the events for this specific NPI, t_k^*, whereas the control group would be the events of all other NPIs K_{-k}, as long as these events do not overlap with t_k^*.

The entire workflow of our study is simplified in Fig 2.

Results

NPI employment statistics

As a preliminary step, we examined the frequency at which each NPI was imposed in the 113 countries (Fig 3). Since governments introduced and lifted lockdowns and social distancing measures consistently throughout time, we observed a continuous growth in the number of confinement events (Fig 3A). This behavior was the opposite for mask wearing events, which had a swift and sustained increase until October 2020, and then stayed constant for the rest of the period. By the end of March 2020, over 60% of all documented NPI events until October 2021 had already happened. We then analysed the distribution of NPIs across different countries before and after June 1st 2020 (S6 Fig in Appendix E9 of S1 Appendix). We discovered that in the first wave, most countries imposed a large and diverse set of NPIs, which was greatly reduced in the following waves. For example, European countries like Spain and Italy introduced very similar government policies during their first waves, but adopted different strategies in the subsequent period: while Italy imposed similar preventive measures with less frequency, Spain placed more emphasis on lockdowns. Overall, throughout 2021, countries had to choose the appropriate NPIs that best fit their socio-economic circumstances and gradually lifted restrictions to avoid negative impacts in the economy [36].

Because a large number of restriction measures were initiated simultaneously in multiple countries within a short period of time, we analysed the events' coincidence per country until end of May 2020. The co-occurrence matrix in Fig 3B shows that the NPIs are co-linear in this period, i.e., they frequently co-occurred. For example, out of 57 confinement events, 12 of them occurred within a period of two weeks (seven days earlier or later) of a work restriction event in the same country. When high co-linearity exists, individual effect estimates are more challenging, as it is harder to identify what in fact drives the change in SARS-CoV-2 transmission. To overcome this problem and to control for the influence of individual NPIs on the
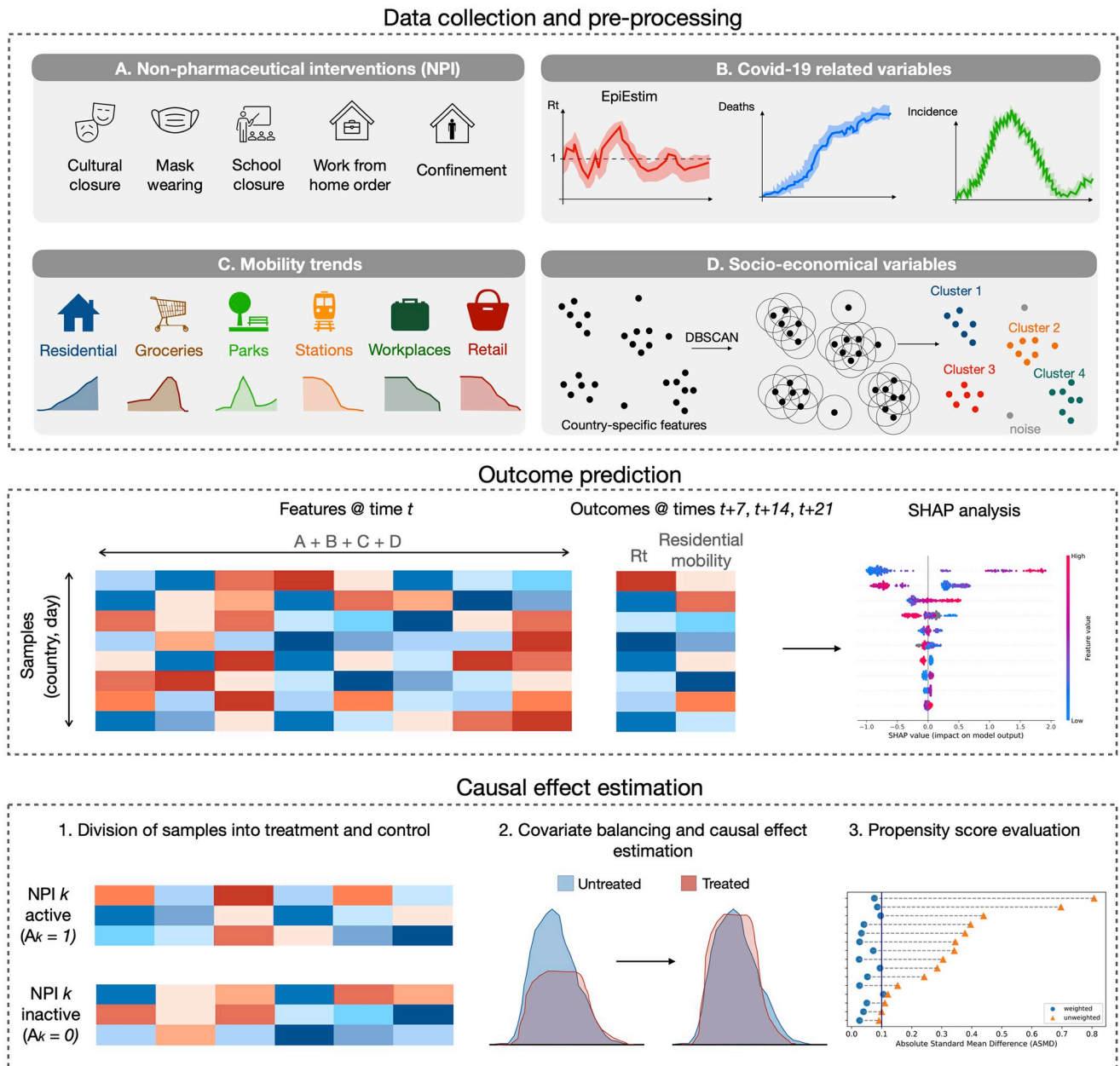
Fig 2. Study workflow. The study can be divided into three main parts: data collection and pre-processing, outcome prediction and causal effect estimation.

estimated effects, we opted to add the status of NPIs as confounders to our causal model (recall section Potential outcome framework).

Predictors of mobility change and Covid-19 R_t

The prediction performance on the validation dataset of both outcomes and their feature contribution plots are summarized in Fig 4 (see S3 Table in Appendix E4 of S1 Appendix for technical specifications of the XGBoost algorithms). To evaluate the performance of the XGBoost models, we compared their accuracy to the one of a model based on permutation tests [37].
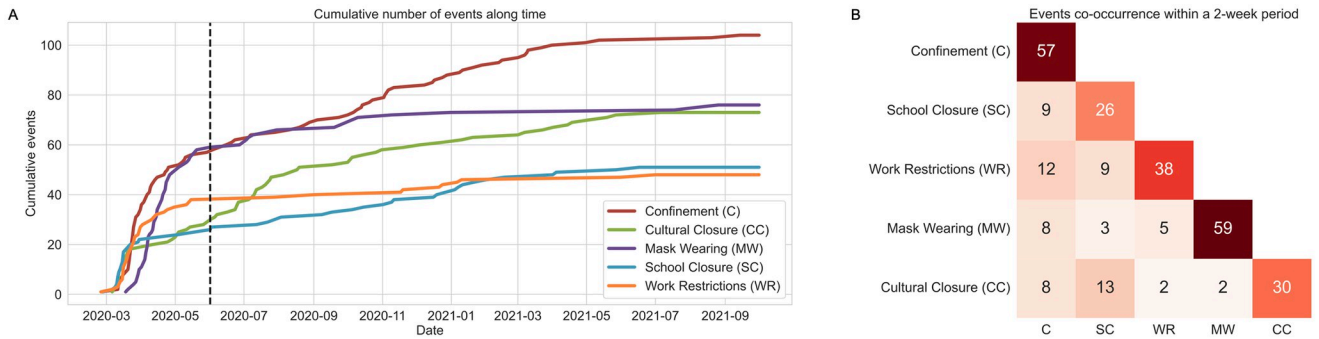
Fig 3. Distribution of NPIs and frequency of their co-existence. (A) Cumulative number of events over time. An event is defined as the day when an intervention was enacted in a country, considering that this NPI was not active the day before. Until June 1st 2020 (dashed vertical line), 60% of all events had already happened. (B) Co-occurrence matrix of events in the limited period until June 1st 2020. The diagonal values represent the number of events. Cell values indicate the sum of times that NPI *i* (x-axis) and NPI *j* (y-axis) were enacted together within a period of 2 weeks in the same country or US state.

This model generates a null distribution by calculating the normalized MSE under the null hypothesis, where in each bootstrap replicate the features are kept the same but the outcomes undergo different permutations. We found that for all scenarios considered in this study (different time lags and different outcome variables), the features we utilized showed enough statistical power to allow a good prediction (Fig 4C). Associations between the outcome variables
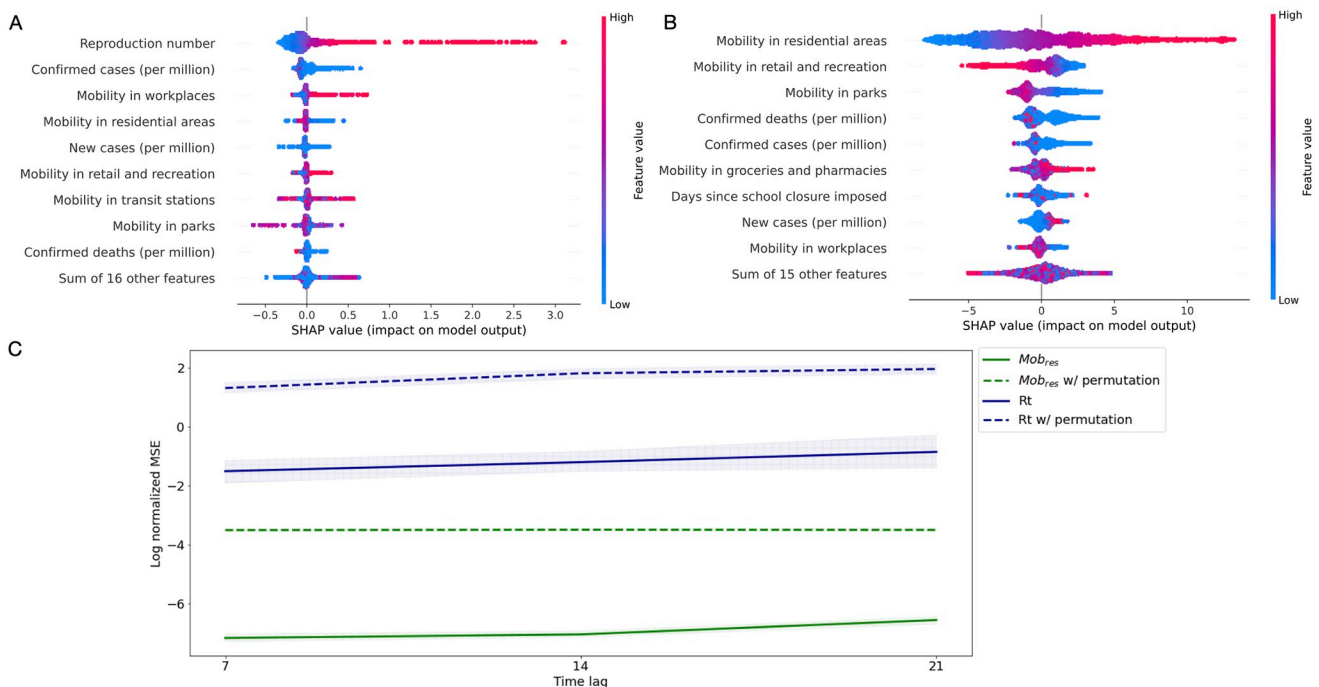

Fig 4. Important predictors of outcome and accuracy of prediction for different time-lags. (A) Identification of top-10 predictive variables affecting R_t estimation. (B) Top-10 most contributing features for residential mobility estimation. SHAP analysis in A and B is based on predictions using a time-lag of 7 days. Each dot represents a single data sample in the validation set (i.e., country at a date). The dot color represents the feature value (red = high, blue = low). The farther a dot is from 0 on the x-axis, the more effect (positive or negative) this feature had on the prediction model for this particular sample. (C) Model performance (log of normalized MSE) of R_t (solid dark blue line) and residential mobility (solid green lines) in the validation set for different time lags. The dashed lines of the same color correspond to random prediction derived by permutation tests with their respective models. Both models significantly outperformed random prediction.

and baseline features can be visualized using SHAP "beeswarm" plots, that show the top-10 contributing features for the outcomes prediction using a time-lag of 7 days (Fig 4A and 4B). The strongest predictor of R_t was its own value 7 days before, implying a clear positive association between R_t measures within a one-week period (Fig 4A). Such strong a association was not observed with other important features, such as confirmed cases and deaths. Changes in mobility within several categories also impacted the R_t prediction. In particular, an increase in mobility in workplaces and retail contributed to a higher SARS-CoV-2 transmission. The feature importance analysis of residential mobility prediction revealed that mobility categories are highly correlated: as the time spent at home increases, the movement outside residence decreases (Fig 4B).

Estimated effect of interventions on policy adherence and SARS-CoV-2 transmission

We derived causal effect estimates using balancing weights that minimized the confounding biases. The absolute standard mean difference (ASMD) was used as a measure to compare the distribution of observed baseline covariates between treated and untreated groups. The AdvBal algorithm could significantly decrease differences between covariate distributions, bringing the ASMD to less than 0.25 for all covariates (S3 Fig in Appendix E6 of S1 Appendix). This value is considered a reasonable cutoff for acceptable standardized biases, indicating that the effect estimates are reliable and robust to confounding [38].

We report the estimated causal effects $\widehat{ATT}_M$ and $\widehat{ATT}_R$ without balancing weights (unadjusted) and with balancing weights generated by the AdvBal algorithm and IPW (Fig 5). The estimated effects were compared with different models to show how the results change under different scenarios, and to strengthen our analysis by providing strong corroboration
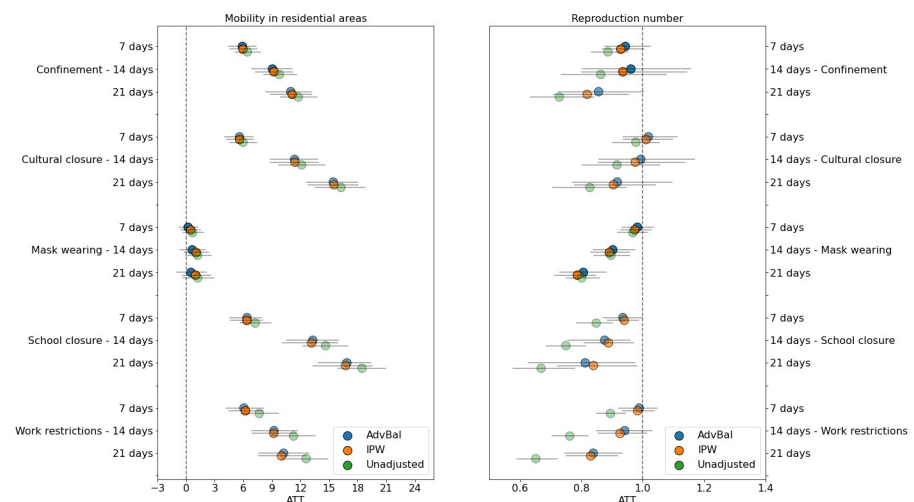

Fig 5. Estimated causal effects of NPIs over time on residential mobility (left) and reproduction number rate (right). In both plots, the opacity of the markers represents the ability of the balancing weights method to balance the treatment groups: the smaller the ASMD is, the more opaque the markers are. Fully opaque markers indicate an ASMD <0.1, half-transparent markers indicate $0.1 \leq$ ASMD ≤ 0.25 and most transparent ones represent ASMD > 0.25. Apart from mask wearing mandates, all NPIs caused a significant increase in time spent at home. Of those, school and cultural closures were the most effective. Closing schools, issuing face mask usage, and work-from-home mandates also caused a persistent reduction in R_t after their initiation, which was not observed with the other social distancing measures. Code used for generating figure is available at https://github.com/barakm-ki/symptoms-dynamics-of-COVID-19-infection/blob/master.

https://doi.org/10.1371/journal.pone.0265289.g005

for the correction of bias. All estimates are presented with 95% confidence intervals and were obtained after assessing whether the balancing weights approaches were able to reduce biases.

Our results suggest that the full extent of NPIs, except mask wearing, significantly affected human mobility change as early as 7 days after their initiation. However, changes in mobility varied between these 4 restrictions. School and cultural closures caused a quick and sustained increase in the time spent inside the home, whereas confinement and work-from-home orders had a slower and plateauing effect over time. Estimates from the AdvBal algorithm suggest that, following the introduction of NPIs, the time spent at home 14 days later was estimated to increase by 9.06% [95% CI: 6.86%, 11.10%], 9.22% [95% CI: 6.90%, 11.66%], 13.29% [95%CI: 10.57%, 15.95%] and 11.38% [95%CI: 8.83%, 13.80%], in response to confinement, work restrictions, school and cultural closure, respectively. Since people already spent a lot of time in their residence before the pandemic, movement changes in residential areas were likely to be smaller compared to outside locations, such as recreation and retail areas (S4 Fig in Appendix E7 of S1 Appendix).

From the five interventions considered, all except cultural closure and confinement caused a significant decline in the transmission of SARS-CoV-2, but within different time lags. It took an estimated 7 days after initiation of school closures to bring $\widehat{ATT}_R$ below 1, as opposed to 14 days for face mask mandates, and 21 days for work restrictions. By the end of 3 weeks, school closures had continuously reduced R_t until 0.81 [95% CI: 0.63, 0.98], mask wearing caused a decrease until 0.81 [95% CI: 0.73, 0.88], and work-from-home orders led to R_t reduction of 0.84 [95% CI: 0.75, 0.93]. Overall, the confidence intervals for the individual effects of all restrictions overlapped, suggesting a comparable effect between each other. IPW and the AdvBal algorithm showed similar trends in their resulting effects, but the balancing of covariates was marginally better on the latter (S3 Fig in Appendix E6 of S1 Appendix).

Results of a complementary analysis based on an alternative study design are described in Appendix E8. We found that under a stricter cohort design, where we compare the effect of NPIs with respect to the other NPIs, school closure had the greatest impact on increasing mobility in residences and in reducing R_t 14 days after its initiation (S5 Fig in S1 Appendix).

Discussion

We constructed a dataset that combines rich information about countries and their reaction to the urgent need to control the pandemic spread. The data include information on social-economics and health benefits, NPIs, and mobility data from more than 100 countries. We showed that the data have predictive power, and that the prediction of changes in mobility after imposing NPIs is more accurate than the prediction of the reproduction number. We employed a causal inference approach to quantify the effect of NPIs on the rate of R_t (i.e., transmission of SARS-CoV-2) and on the change of human mobility, which is considered a proxy measure of population adherence and social distancing. The purpose of this study was to help infer the efficiency of interventions in the early months of a pandemic, when a number of control measures had already been imposed by multiple countries in the absence of vaccines.

The most common use of causal inference seeks to estimate the average treatment effect (ATE). Such an analysis would answer questions such as "what would have happened if every country applied the NPI?". However, in this case an analysis of this type proved to be unfeasible, since it was not possible to balance the confounding differences between the treated and untreated countries. We therefore opted to measuring the impact of NPIs in the countries that

chose to impose the NPIs, which is known as the average effect on treated (ATT). This answers the question of "what was the effect of applying the NPIs in the countries that applied it?". This approach allowed covariate balancing, which provided more reliable estimates of the effect of NPIs.

Our findings showed that mask wearing did not significantly impact mobility patterns in the first wave. Although a number of countries favored face mask usage early in their outbreaks [39], the main reason people changed their behavior was social distancing policies. We found that issuing confinement, work-from-home orders, or school/cultural closure mandates resulted in high levels of policy compliance even one week after their initiation, as measured by changes in movement in residential areas. This result was consistent with recent findings [32, 40].

The estimated effect on R_t showed that not all NPIs significantly contributed to a decrease in SARS-CoV-2 transmission. In particular, school closure achieved a sustained decline on the rate of R_t, similar to what was found in observational studies of the first wave [4, 5, 7–9] and second European wave [41]. Because infected children can experience mild or no symptoms more frequently than older individuals [42] and tend to have more social contact than adults [43], it is expected that closing schools would considerably contribute to reduce the transmission. In contrast, on its own, closing cultural establishments does not seem to have an effect on the reproduction number, nor do work restrictions in the first 14 days after imposing the NPIs. Notably, we did not find substantial differences in the results when performing sensitivity analysis.

Our study extends previous first-wave estimate studies [2, 4, 5, 7–10] in a number of ways. First, we used the potential outcome framework to infer NPI effects. In its simplest form, our causal model made use of standard causal inference methods to correct observed biases and obtain valid effects with more transparent confidence intervals. Second, we addressed the issue of concurring NPIs by using their status as covariates in the causal model. By ensuring that all NPI-related covariates are well balanced between treatment groups, we enhanced the power to detect independent NPI effects. Third, we account for heterogeneity of countries by including a social economical cluster indicator in the dataset. Fourth, we conducted complementary analyses under alternative scenarios to test our conclusions not only with a different cohort study design, but also with another balancing weights generation approach.

We acknowledge several limitations in our analyses. Even with data from multiple countries that had diverse sets of interventions in place, inferring NPI effects still remained a challenging task. First, the R_t estimation was based on epidemiological parameters that are only known with uncertainty, due to many mild or asymptomatic cases that make it difficult to model the timing for the onset of symptoms and serial interval distributions. On top of that, R_t also relies on the data of confirmed cases, which were generally unreliable in the early days of the pandemic due to lack of testing availability and not-established documentation practices. To account for this, we began our analysis at each country's 100th case. Secondly, the data are retrospective and observational, meaning that unobserved factors could confound the results. Third, we were unable to assess the effect of lifting interventions. Since the NPI events in WNTRAC are automatically extracted from Wikipedia articles, which report the introduction of NPIs more frequently than their relaxation, the number of lifting events documented in the database did not have enough statistical power for causal inference. Yet, we believe we set the ground for a thorough analysis of NPIs and we were able to draw conclusions regarding the effect different NPIs had on the pandemic spread. Future work can assess the causal effect of the post-vaccine newly defined NPIs where health certificate notions were introduced in somewhat similar ways across different countries.

Code availability

The source code for the causal inference evaluation of NPIs is available in a public GitHub repository at https://github.com/IBM/causallib. Please refer to the README file in the repository for further instructions on using the code. Requests for the code used to generate the results and the plots should be directed to the corresponding author.

Supporting information

S1 Appendix.
(ZIP)

Author Contributions

Conceptualization: Michal Ozery-Flato, Yishai Shimoni, Michal Rosen-Zvi.

Data curation: Vesna Barros, Itay Manes, Victor Akinwande, Celia Cintas, Osnat Bar-Shira.

Formal analysis: Vesna Barros, Victor Akinwande, Osnat Bar-Shira, Michal Ozery-Flato.

Funding acquisition: Vesna Barros.

Investigation: Vesna Barros, Itay Manes, Celia Cintas, Osnat Bar-Shira.

Methodology: Vesna Barros, Itay Manes, Michal Ozery-Flato, Yishai Shimoni.

Project administration: Michal Ozery-Flato, Yishai Shimoni, Michal Rosen-Zvi.

Resources: Celia Cintas.

Software: Michal Ozery-Flato.

Supervision: Michal Ozery-Flato, Yishai Shimoni, Michal Rosen-Zvi.

Validation: Vesna Barros, Itay Manes, Victor Akinwande.

Visualization: Vesna Barros, Itay Manes, Victor Akinwande, Osnat Bar-Shira.

Writing – original draft: Vesna Barros.

Writing – review & editing: Vesna Barros, Itay Manes, Victor Akinwande, Celia Cintas, Osnat Bar-Shira, Michal Ozery-Flato, Yishai Shimoni, Michal Rosen-Zvi.

References

1. Coronavirus W. Dashboard| WHO Coronavirus (COVID-19) Dashboard With Vaccination Data; 2021.

2. Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature. 2020; 584(7820):257–261. https://doi.org/10.1038/s41586-020-2405-7 PMID: 32512579

3. Qian Z, Alaa AM, van der Schaar M. When and How to Lift the Lockdown? Global COVID-19 Scenario Analysis and Policy Assessment using Compartmental Gaussian Processes. Advances in Neural Information Processing Systems. 2020; 33.

4. Haug N, Geyrhofer L, Londei A, Dervic E, Desvars-Larrive A, Loreto V, et al. Ranking the effectiveness of worldwide COVID-19 government interventions. Nature human behaviour. 2020; 4(12):1303–1312. https://doi.org/10.1038/s41562-020-01009-0 PMID: 33199859

5. Banholzer N, Van Weenen E, Lison A, Cenedese A, Seeliger A, Kratzwald B, et al. Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave. PLoS one. 2021; 16(6):e0252827. https://doi.org/10.1371/journal.pone.0252827 PMID: 34077448

6. Bendavid E, Oh C, Bhattacharya J, Ioannidis JP. Assessing mandatory stay-at-home and business closure effects on the spread of COVID-19. European journal of clinical investigation. 2021; 51(4):e13484. https://doi.org/10.1111/eci.13484 PMID: 33400268

7. Brauner JM, Mindermann S, Sharma M, Johnston D, Salvatier J, Gavenčiak T, et al. Inferring the effectiveness of government interventions against COVID-19. Science. 2021; 371 (6531). https://doi.org/10.1126/science.abd9338 PMID: 33323424

8. Liu Y, Morgenstern C, Kelly J, Lowe R, Jit M, Group CCW, et al. The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. medRxiv. 2020;.

9. Li Y, Campbell H, Kulkarni D, Harpur A, Nundy M, Wang X, et al. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. The Lancet Infectious Diseases. 2021; 21(2): 193–202. https://doi.org/10.1016/S1473-3099(20)30785-4 PMID: 33729915

10. Hsiang S, Allen D, Annan-Phan S, Bell K, Bolliger I, Chong T, et al. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. Nature. 2020; 584(7820):262–267. https://doi.org/10.1038/s41586-020-2404-8 PMID: 32512578

11. Chernozhukov V, Kasahara H, Schrimpf P. Causal impact of masks, policies, behavior on early covid-19 pandemic in the US. Journal of econometrics. 2021; 220(1):23–62. https://doi.org/10.1016/j.jeconom.2020.09.003 PMID: 33100476

12. Singh S, Shaikh M, Hauck K, Miraldo M. Impacts of introducing and lifting nonpharmaceutical interventions on COVID-19 daily growth rate and compliance in the United States. Proceedings of the National Academy of Sciences. 2021; 118(12). https://doi.org/10.1073/pnas.2021359118 PMID: 33658331

13. forecasting team IC. Modeling COVID-19 scenarios for the United States. Nature medicine. 2020;.

14. Courtemanche C, Garuccio J, Le A, Pinkston J, Yelowitz A. Strong Social Distancing Measures In The United States Reduced The COVID-19 Growth Rate: Study evaluates the impact of social distancing measures on the growth rate of confirmed COVID-19 cases across the United States. Health Affairs. 2020; 39(7):1237–1246. https://doi.org/10.1377/hlthaff.2020.00608 PMID: 32407171

15. Pei S, Kandula S, Shaman J. Differential effects of intervention timing on COVID-19 spread in the United States. Science advances. 2020; 6(49):eabd6370. https://doi.org/10.1126/sciadv.abd6370 PMID: 33158911

16. Pan W, Fernández D, Tyrovolas S, Iago GV, Dasgupta RR, Zaitchik BF, et al. Heterogeneity in the Effectiveness of Non-Pharmaceutical Interventions during the first SARS-CoV2 wave in the United States. Frontiers in public health. 2021; p. 1857. https://doi.org/10.3389/fpubh.2021.754696 PMID: 34912768

17. Wellenius GA, Vispute S, Espinosa V, Fabrikant A, Tsai TC, Hennessy J, et al. Impacts of social distancing policies on mobility and COVID-19 case growth in the US. Nature communications. 2021; 12(1):1–7. https://doi.org/10.1038/s41467-021-23404-5 PMID: 34035295

18. Auger KA, Shah SS, Richardson T, Hartley D, Hall M, Warniment A, et al. Association between statewide school closure and COVID-19 incidence and mortality in the US. Jama. 2020; 324(9):859–870. https://doi.org/10.1001/jama.2020.14348 PMID: 32745200

19. Viner RM, Russell SJ, Croker H, Packer J, Ward J, Stansfield C, et al. School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. The Lancet Child & Adolescent Health. 2020; 4(5):397–404. https://doi.org/10.1016/S2352-4642(20)30095-X PMID: 32272089

20. Suryanarayanan P, Tsou CH, Poddar A, Mahajan D, Dandala B, Madan P, et al. AI-assisted tracking of worldwide non-pharmaceutical interventions for COVID-19. Scientific Data. 2021; 8(1):1–14. https://doi.org/10.1038/s41597-021-00878-y PMID: 33767205

21. Google, COVID-19 Community Mobility Reports.;. https://www.google.com/covid19/mobility, accessed on 2021-10-05.

22. Bank W. World Development Report 2021: Data for Better Lives. The World Bank; 2021.

23. Ritchie H, Roser M. Human development index, age structure, population distribution;. https://ourworldindata.org, accessed on 2021-10-05.

24. Ester M, Kriegel HP, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. vol. 96; 1996. p. 226–231.

25. World Health Organization. Coronavirus disease 2019 (COVID-19): situation reports;. https://covid19.who.int/info, accessed on 2021-10-05.

26. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. American journal of epidemiology. 2013; 178(9):1505–1512. https://doi.org/10.1093/aje/kwt133 PMID: 24043437

27. Thompson R, Stockwin J, van Gaalen RD, Polonsky J, Kamvar Z, Demarsh P, et al. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics. 2019; 29:100356. https://doi.org/10.1016/j.epidem.2019.100356 PMID: 31624039

28. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA. Serial interval of COVID-19 among publicly reported confirmed cases. Emerging infectious diseases. 2020; 26(6):1341. https://doi.org/10.3201/eid2606.200357 PMID: 32191173

29. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–794.

30. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv preprint arXiv:170507874. 2017;.

31. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology. 1974; 66(5):688. https://doi.org/10.1037/h0037350

32. Snoeijer BT, Burger M, Sun S, Dobson RJ, Folarin AA. Measuring the effect of Non-Pharmaceutical Interventions (NPIs) on mobility during the COVID-19 pandemic using global mobility data. NPJ digital medicine. 2021; 4(1):1–12. https://doi.org/10.1038/s41746-021-00451-2 PMID: 33986465

33. Ozery-Flato M, Thodoroff P, Ninio M, Rosen-Zvi M, El-Hay T. Adversarial balancing for causal inference. arXiv preprint arXiv:181007406. 2018;.

34. Shimoni Y, Karavani E, Ravid S, Bak P, Ng TH, Alford SH, et al. An Evaluation Toolkit to Guide Model Selection and Cohort Definition in Causal Inference. arXiv preprint arXiv:190600442. 2019;.

35. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41–55. https://doi.org/10.1093/biomet/70.1.41

36. McKee M, Stuckler D. If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future. Nature Medicine. 2020; 26(5):640–642. https://doi.org/10.1038/s41591-020-0863-y PMID: 32273610

37. Ojala M, Garriga GC. Permutation tests for studying classifier performance. Journal of Machine Learning Research. 2010; 11(6).

38. Stuart EA, Lee BK, Leacy FP. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. Journal of clinical epidemiology. 2013; 66(8):S84–S90. https://doi.org/10.1016/j.jclinepi.2013.01.013 PMID: 23849158

39. Howard J, Huang A, Li Z, Tufekci Z, Zdimal V, van der Westhuizen HM, et al. An evidence review of face masks against COVID-19. Proceedings of the National Academy of Sciences. 2021; 118(4). https://doi.org/10.1073/pnas.2014564118 PMID: 33431650

40. Oh J, Lee HY, Khuong QL, Markuns JF, Bullen C, Barrios OEA, et al. Mobility restrictions were associated with reductions in COVID-19 incidence early in the pandemic: evidence from a real-time evaluation in 34 countries. Scientific reports. 2021; 11(1):1–17. https://doi.org/10.1038/s41598-021-92766-z PMID: 34215764

41. Sharma M, Mindermann S, Rogers-Smith C, Leech G, Snodin B, Ahuja J, et al. Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. Nature communications. 2021; 12(1):1–13. https://doi.org/10.1038/s41467-021-26013-4 PMID: 34611158

42. Davies NG, Klepac P, Liu Y, Prem K, Jit M, Eggo RM. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nature medicine. 2020; 26(8):1205–1211. https://doi.org/10.1038/s41591-020-0962-9 PMID: 32546824

43. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS medicine. 2008; 5(3):e74. https://doi.org/10.1371/journal.pmed.0050074 PMID: 18366252