The impact of vaccination on the number of COVID-19 cases

Matyáš Mattanelli & Vít Kulhánek

January 2022

Abstract.

Contents

1 Introduction **1**

2 Literature Review **1**

3 Data Description **2**

4 Methodology **5**

5 Interpretation of Results **5**

6 Conclusion **5**

1 Introduction

Covid-19 is one of the most frequently used words in the last three years. This disease has had an effect on virtually everybody on Earth. Consequently, there are ongoing attempts to offset and eventually put an end to its negative impact. One possible way how to return to the "normal" living circumstances is a vaccine. It does not completely protect people against the disease, however, it is expected to reduce the quantity of hospitalizations and deaths. The aim of this project is to determine whether vaccination also decreases the number of new cases, and thus constitutes an essential tool for the elimination of the pandemic. The rest of the work is structured as follows. The second section provides a brief overview of the current literature, the third section describes the utilized data, and the fourth section is concerned with the methodology. The last two sections contain the discussion of the results and the conclusion, respectively.

2 Literature Review

There are not many studies with similar topic because the situation is still evolving and vaccination has only been increasing rapidly since the beginning of the second quarter of 2021. Velasco *et al.* (2021) performed a cross-country examination of which factors have a significant effect on the number of new Covid-19 cases. The authors utilized data from the beginning of the outbreak till the end of 2020 to estimate an OLS regression. Since the vaccination data were unavailable at that time, it was not included. The main findings of the study were that the number of tests and the average temperature have a significant and positive effect on the number of cases. However, we have identified some issues that may weaken the reliability of the results. Since the estimation was concerned with panel data and the individuals were represented by countries, we expect for the unobserved heterogeneity to be present and also correlated with our independent variables. This is due to the fact that it is nearly impossible

to control for all country-specific factors. For example, each country may have enacted different policies which could have a direct impact on the number of test as well as on the number of new cases. The failure to account for the unobserved heterogeneity brings endogeneity into the model which renders OLS estimation inconsistent. We intend to improve upon this study by performing corresponding tests and employing conventional panel data estimation methods.

Toharudin *et al.* (2021) used Bayesian structural time series models including variables such as new cases, recovery cases, and number of deaths of Covid-19. They found that the vaccination program that took place in Jakarta only had a significant effect on the number of recovered cases.

Li *et al.* (2020) were testing Covid-19 in the USA. The authors found that temperature has a significant effect on the number of Covid-19 cases. Their results showed that higher temperature reduces the number of cases, however it does not affect the death rate.

To the best of our knowledge, there is still quite some unexplored area in the proposed topic. By utilizing the most recent data available, we will attempt to shed alight on the problematic whether vaccination plays a key role in the reduction of the number of Covid-19 cases.

3 Data Description

For the purpose of our analysis, we downloaded a large data set from Ritchie *et al.* (2020) which contains various information regarding Covid-19. More specifically, it includes daily observations for 207 countries. The main variables are the number of new cases of Covid-19, the number of tests determining the presence of the virus, the quantity of newly vaccinated people, total population, and many others. We aggregated the daily observations to monthly intervals in an attempt to reduce the measurement error and also to eliminate multidimensionality since most of the conventional panel data estimators are appropriate for "short" panels. As a result, we have data beginning in February 2020 and ending in December 2021.[1] Furthermore, we scaled the data by total population to provide higher cross-country comparability, thus we have variables in a form of new cases per thousand people or number of tests per thousand people.

Our additional source of the data is Weatherbase (2022). This website contains information about monthly average temperatures in over 260 countries. Based on the current literature, we consider average temperature as an important factor influencing the number of new Covid-19 cases, and therefore we include it in our model as a control variable. Since there are measurements provided for several cities in each country, we scraped the available data from the website and then computed average monthly temperatures for each country in each month.[2] As a next step, we merged the data with the previous data set.

[1] We did not include January 2022 since in the time of writing it has not ended yet.

[2] We utilized the Python programming language, especially the requests module. The code is available in the appendix

As it was already mentioned, we are examining the effect of vaccination on the number of new cases of Covid-19. Thus, the dependent variable in our model is the number of new cases per thousand people which encompasses the quantity of positively tested people every month. Based on the histogram provided below, we transformed our dependent variable into logarithm because its distribution is extremely right-skewed.[3] As desired, the logarithmic transformation adjusts the distribution to be closer to Gaussian normal distribution.

Our key independent variable is the number of vaccinations per thousand people. It shows how many people were vaccinated in a particular month. This variable was also transformed into logarithm because its distribution was heavily skewed. Based on our hypothesis, we expect negative relation between the number of new cases and the number of vaccinations. We define two additional control variables to ensure the robustness of our results. Li *et al.* (2020) found out that temperature is a significant factor in Covid-19 development, and therefore we include average temperature which is expected to have a negative impact on the spread of coronavirus. Lastly, the number of tests per thousand people was added as an additional control variable. The skewness of its distribution visible in the histogram forces us to use logarithm for this variable as well. We expect a positive effect of tests because without tests we cannot detect the presence of Covid-19.

The resulting data set contains quite a lot of missing observations. This is due to the fact that the first vaccination data are available in December 2020. Therefore, we disregard earlier observations. In addition, there are some missing values even after this date. We assume that they are distributed randomly and opt for listwise deletion. This is a strong assumption, however, even if it is incorrect and the missing values occur due to some country-specific characteristic, we will deal with it in the analysis.

[3]We had to add a small constant to each observation since the presence of zeros renders the logarithmic transformation unfeasible

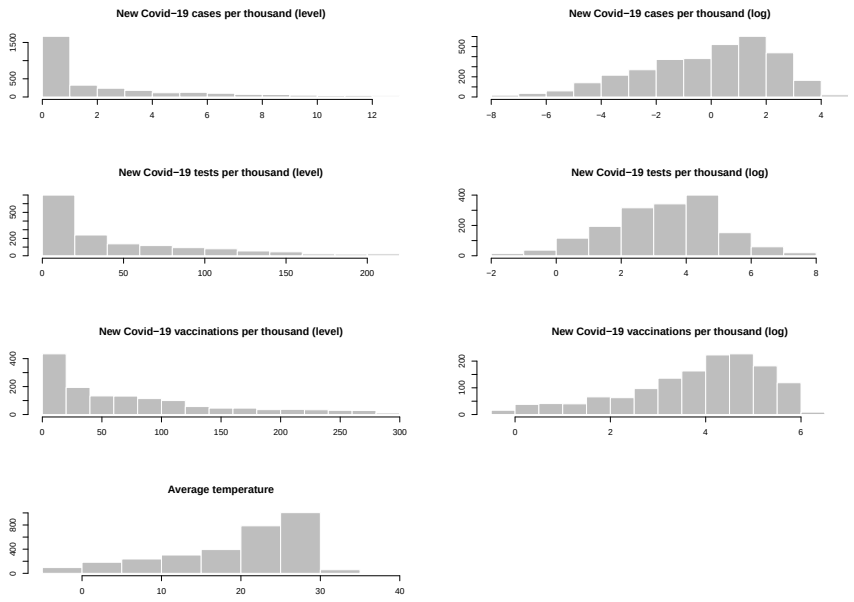

Figure 1: Histograms

Additionally, we provide the correlation matrix below in order to inspect the issue of multicollinearity. As expected, the correlation between the number of new cases and the number of tests is quite high which is caused by the nature of the variables meaning that nobody is considered as Covid-19 positive until they receive a positive test. Curiously, the number of vaccinations does to appear to be highly correlated with our dependent variable. However, we expect the relationship to be more complex and thus it cannot be fully captured by a mere bivariate correlation. The sign on the correlation of average temperature matches our predictions. There is no high correlation between our independent variables so we should not be facing the issue of multicollinearity.

Table 1: Correlation matrix

	new_cases_per_1000	new_tests_per_1000	new_vaccinations_per_1000	avg_temp
new_cases_per_1000	1	0.706	0.008	−0.201
new_tests_per_1000	0.706	1	0.019	−0.296
new_vaccinations_per_1000	0.008	0.019	1	−0.016
avg_temp	−0.201	−0.296	−0.016	1

4 Methodology

Since our aim is to investigate the impact of the number of newly vaccinated on the number of new Covid-19 cases using cross-country data over a certain period of time, we are essentially facing panel data analysis. As a result, as mentioned in the introduction, we need to consider the unobserved heterogeneity. Since we are dealing with countries, we expect them to be heterogeneous. For example, we do not control for geographical position, institutional design, culture, and many other factors that could have a direct effect on our dependent variable and which may also be correlated with our explanatory variables. Consequently, to verify our hypothesis, we perform several tests for the present of individual and time effects, namely the F test and LM test. In addition, we utilize the Hausman test to decide whether the unobserved heterogeneity is correlated with our independent variables. Given the results, we employ the within estimator which removes the fixed-effects which is one of the ways to achieve consistency. To evaluate the validity of our results we further test for heteroskedasticity, serial correlation, and cross-sectional dependence. We acknowledge that these issues render the statistical inference invalid, and therefore we treat them with the usage of robust standard errors.

5 Interpretation of Results

6 Conclusion

References

Li, A. Y., T. C. Hannah, J. R. Durbin, N. Dreher, F. M. McAuley, N. F. Marayati, Z. Spiera, M. Ali, A. Gometz, J. Kostman, & T. F. Choudhri (2020): "Multivariate analysis of factors affecting covid-19 case and death rate in u.s. counties: The significant effects of black race and temperature." *medRxiv* .

Ritchie, H., E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, & M. Roser (2020): "Coronavirus pandemic (covid-19)." *Our World in Data* Available at `https://ourworldindata.org/coronavirus`. [Accessed 2022-01-16].

Toharudin, T., R. S. Pontoh, R. E. Caraka, S. Zahroh, P. Kendogo, N. Sijabat, M. D. P. Sari, P. U. Gio, M. Basyuni, & B. Pardamean (2021): "National vaccination and local intervention impacts on covid-19 cases." *Sustainability* **13(15)**.

Velasco, J. M., W.-C. Tseng, & C.-L. Chang (2021): "Factors affecting the cases and deaths of COVID-19 victims." *Int J Environ Res Public Health* **18(2)**.

Weatherbase (2022): "Travel weather averages (weatherbase)." Available at `https://www.weatherbase.com/`. [Accessed 2021-01-16].