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 **4**

5 Interpretation of Results **5**

6 Conclusion **5**

Bibliography **6**

Appendix **7**

1 Introduction

Covid-19 is one of the most frequently repeated terms 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 repercussions. 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 vaccinations also decrease the number of new cases, and thus constitute an essential tool for the elimination of the pandemic. The rest of the work is structured as follows. Section 2 provides a brief overview of the current literature, Section 3 describes the utilized data, and Section 4 is concerned with the methodology. The last two sections (5, 6) contain the discussion of the results and the conclusion, respectively.

2 Literature Review

There is only a handful of studies on similar topic since the situation is still evolving and vaccinations have 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 the 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 tests 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. It is important to note that the study focused only on the very specific case of Jakarta.

Li *et al.* (2020) were testing the effects on 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 a light 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

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

in each month.[2]

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. After inspecting the histogram available in Figure 1, we transformed our dependent variable into logarithms 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 logarithms because its distribution is heavily skewed. Based on our hypothesis, we expect negative relation between the number of new cases and the number of vaccinations since the vaccine is expected to enhance individual's immunity. 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 logarithms 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. As a result, our final data set is an unbalanced panel containing 103 countries with the number of periods available ranging from 1 to 13. The descriptive statistics are presented in Table 1. We can see that in our data set there were, on average, approximately 6 monthly cases per thousand people in a country. Since the median is equal to roughly 3 cases per thousand people, the mean is likely inflated by many outliers in the upper tail. This issue should be mitigated by the logarithmic transformation since it reduces the emphasis put on extreme observations. New vaccinations and tests show a similar trend. Lastly, the average temperature in our data set is approximately 16 degrees Celsius.

Additionally, we provide the correlation matrix in Table 1 to inspect the bivariate relationship among our variables. As expected, the correlation between the number of new cases and the number of tests is quite high which is caused

[2]We utilized the Python programming language, especially the requests module.

[3]We had to add a small constant to each observation since the presence of zeros renders the logarithmic transformation unfeasible. In addition, all outliers were trimmed in order to enhance the appearance of the figures

Table 1: Descriptive statistics

Statistic	Mean	St. Dev.	Min	Median	Max
New cases	5.753	7.642	0.003	2.978	63.576
New vaccinations	97.935	95.244	0.000	71.717	570.621
New tests	126.222	269.659	0.032	42.274	3,089.120
Average temperature	16.443	9.320	−17.858	17.354	34.660

Note: All variables apart from Average Temperature are in per thousand people terms

by the nature of the variables. Nobody is considered as Covid-19 positive until they receive a positive test. On the other hand, the number of vaccinations does not appear to be significantly 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 direction of the relationship between new cases and average temperature matches our predictions. Regarding multicollinearity, there is not a high pairwise correlation among our independent variables so it should not cause an issue in our analysis.

Table 2: Correlation matrix

	New cases	New tests	New vaccinations	Average temperature
New cases	1			
New tests	0.63*	1		
New vaccinations	−0.00027	0.019	1	
Average temperature	−0.32*	−0.39*	−0.016	1

Note: All variables apart from Average temperature are in logs per thousand people
* Significant at 95%

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 the number of new cases and which may also be correlated with our explanatory variables. In addition, we expect the number of new cases to depend on its past values. To support our hypothesis we test for a serial correlation in a static model estimated by fixed-effects. The inclusion of a lagged dependent variable moves us to the area of dynamic panel data estimators. In this context, the conventional estimators such as the within estimator or the first-difference estimator

are inconsistent due to the Nickell bias (Nickell 1981). A solution was proposed by (Arellano & Bond 1991) who devised the Difference Generalized Method of Moments estimator. The procedure is as follows. Initially, the data is transformed using first-differencing to remove the unobserved heterogeneity. Then, with the utilization of internal instruments the resulting equation is estimated using GMM.[4] In this way, the endogeneity caused by the lagged dependent variable and also any other endogenous variables is dealt with. This is very useful in our case since our variables may be subjected to measurement error. As Roodman (2009a) suggests, the Difference GMM estimator is appropriate for "short" panels with many individuals, models with possibly endogenous regressors, and the presence of fixed-effects. To verify the reliability of our results, we perform the Hansen-Sargan test and the Arellano-Bond test which are essential for the validity of internal instruments. The former is a test of overidentifying restrictions while the latter tests the second order serial correlation in the differenced errors.[5] We also make sure to restrict the number of instruments to avoid instrument proliferation as indicated by Roodman (2009b).

5 Interpretation of Results

6 Conclusion

[4]The instruments are the lagged levels of the included variables. These instruments are exogenous under the assumption of no serial correlation.

[5]The order of the serial correlation depends on the number of included lags. If two lags are included, the serial correlation of third order is tested.

Bibliography

Arellano, M. & S. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* **58(2)**: pp. 277–297.

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* .

Nickell, S. (1981): "Biases in dynamic models with fixed effects." *Econometrica* **49(6)**: pp. 1417–26.

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].

Roodman, D. (2009a): "How to do xtabond2: An introduction to difference and system gmm in stata." *The Stata Journal* **9(1)**: pp. 86–136.

Roodman, D. (2009b): "A note on the theme of too many instruments*." *Oxford Bulletin of Economics and Statistics* **71(1)**: pp. 135–158.

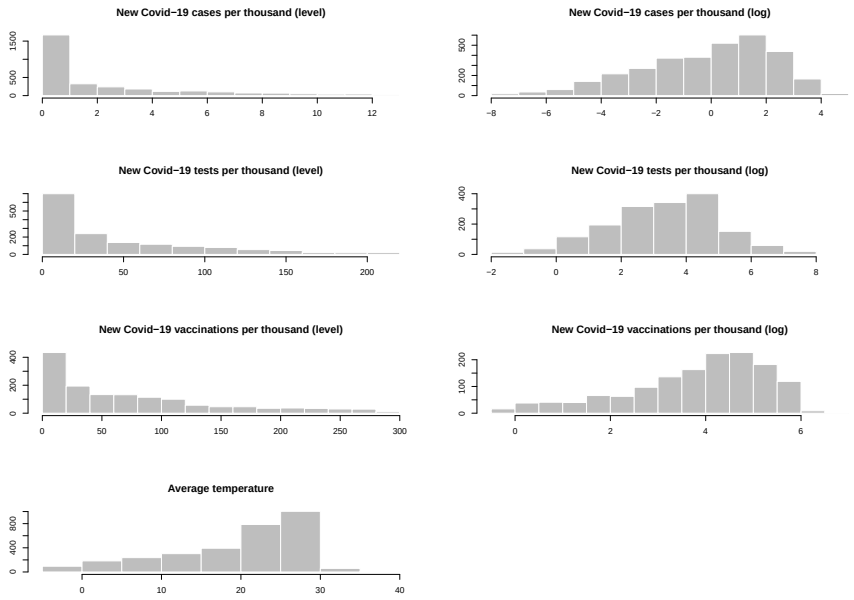
Toharudin, T., R. S. Pontoh, R. E. Caraka, S. Zahroh, P. Kendogo, N. Sijabat, M. D. P. Sari, P. U. Gio, M. Basyuni, & B. Pardamean (2021): "National vaccination and local intervention impacts on covid-19 cases." *Sustainability* **13(15)**.

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].

Appendix

Figure 1: Histograms



Source: Authors' computations based on the compiled data set