

# Covid-19 Epidemiological Factor Analysis: Identifying Principal Factors with Machine

Serge Dolgikh<sup>a,b</sup> and Oksana Mulesa<sup>c</sup>

<sup>a</sup> Solana Networks, 301 Moodie Dr., Ottawa, K2H9C4, Canada

<sup>b</sup> National Aviation University, 1 Liubomyra Huzara Ave, 1, Kyiv, 03058, Ukraine

<sup>c</sup> Uzhgorod National University, Narodna sq., 3, Uzhhorod, 88000, Ukraine

## Abstract

Based on a set of Covid-19 statistical data of national and subnational jurisdictions at the time point of approximately two months after the local onset of the pandemics (early April, 2020), an analysis of the factors with strong influence on the reported local outcomes was performed with several different statistical methods. The consistent conclusion of the analysis with the available statistical data confirms epidemiological policy and management as the dominant factors in the outcome. Other factors with significant influence on the development of epidemiological scenarios among the considered were current or recent universal Bacille Calmette-Guérin (BCG) immunization record and the prevalence of smoking in the population. The methods proposed in the study can be used to evaluate principal factors at a number of future time points to reach a confident conclusion.

## Keywords

Infectious diseases, epidemiology, Covid-19, machine learning, statistical analysis

## 1. Introduction

A possible link between the effects of Covid-19 pandemics such as the rate of incidence and the severity of cases on one hand; and a universal immunization program against tuberculosis with Bacille Calmette-Guérin (BCG vaccine and universal BCG immunization program or UBIP, hereinafter) was suggested in [1] and further investigated in a number of works, offering a novel and interesting perspective on a possibility of relations between certain characteristics of jurisdictions and development of epidemics. A number of factors with potential influence on the epidemiological outcome have been discussed at length, such as population density, age demographics and other. Identification of factors of significance for the development of epidemics, and methods allowing such identifications can provide important inputs to development of effective policy.

## 2. Problem Statement

A common challenge in the analysis of statistical data related to a developing situation, such as in this work, the developing epidemiological scenario related to a dangerous infection with potentially high impact on health and safety of population, economy and the society as a whole is evaluation of methods and models with the objective of identifying the approaches that could be most effective in describing the process that is being studied. Such a choice may itself depend on the problem and the data. For one time series the best approach can be an autoregression model, for another, Brown model or Winters models and so on.

To avoid or reduce the possible ambiguity related to the selection of the method of analysis of statistical data, in this work we used several common methods of statistical analysis specifically,

---


*IT&I-2020 Information Technology and Interactions, December 02–03, 2020, KNU Taras Shevchenko, Kyiv, Ukraine*

EMAIL: sdolgikh@nau.edu.ua (A. 1); Oksana.mulesa@uzhnu.edu.ua (A. 2)

ORCID: 0000-0001-5929-8954 (A. 1); 0000-0002-6117-5846 (A. 2)

© 2020 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

evaluation and ranking of factor influence with an expectation that if consistency between the results of different methods can be achieved, it would enhance the confidence in the result that can be essential for the development of reliable and effective policies based on the conclusions of factor analysis.

With a variety of statistical methods and techniques used to evaluate the correlation hypothesis as discussed above, we set out to provide an analysis of principal factors influencing the development of the epidemics in the national and subnational jurisdictions based on the available data for the first group of countries that were exposed to Covid-19 pandemics in late January – beginning of February, 2020. This objective is approached by applying several commonly used methods of factor analysis and ranking, looking for consistency of results between different methods. A consistency between the results of different methods would improve confidence in the findings, providing a grounded and reliable statement of their influence on the epidemiological outcome, and providing a confident and informative input to the situation analysis and development of policy.

### 3. Literature Review

Miller et al. [1] provided one of the first indications of the possible link between BCG immunization and milder course of the epidemics in the national jurisdiction. This link was further investigated in a number of works with a consistent, with varying level of confidence, conclusion of the significance of the correlation hypothesis. In [2,3], a strong correlation between the BCG immunization record and Covid-19 mortality in a number of culturally and socially similar European countries was observed ( $R^2 = 0.88$ ;  $P = 8 \times 10^{-7}$ ), indicating that every 10% increase in the BCG index was associated with a 10.4% reduction in Covid-19 mortality. The results imposed strong constraints on the null hypothesis (that is, of no correlation between a current or previous UBIP in the jurisdiction and Covid-19 impact, suggesting that BCG may have a certain broad protective effect resulting in a milder epidemiological scenario.

A similar conclusion is supported by the results in [4,5] establishing a strong correlation between a current and previous record of a consistent UBIP, and lower values of Covid-19 outcomes in the reporting jurisdictions, measured by infection incidence and the resulting mortality “The results ... show that countries without a universal BCG policy (such as Belgium, Italy, the United States, and the Netherlands) have increased incidence of COVID-19 ( $2810.9 \pm 497.1$  (mean  $\pm$  SEM) per million) compared with countries with ongoing national BCG policy ( $570.9 \pm 155.6$  (mean  $\pm$  SEM) per million)” (Sharma et al., [4]).

In [2] qualitative and quantitative analysis of distribution of Covid-19 impacts among national and subnational jurisdictions in Europe, North America and Middle East was performed with a number of observations consistently pointing to a possibility of a correlation between UBIP and a milder type of the epidemiological scenario, while in [6] a quantitative statistical analysis of the significance of the correlation at two time points imposed strong constraints on the null hypothesis excluded with a P-value below 0.0001.

The task of modeling and forecasting time-series processes of different nature is essential and arises in different fields such as planning [7], the study of the dynamics of climate change [8] and importantly in the current situation, health science and epidemiology. It involves the stages of identification of parameters that can be measured; collection of representative sets of data; and application of methods of analysis of data allowing to identify the factors with the highest influence on the observed outcome.

Known models and methods of factor analysis are based on using integrated information about the background of the predicted processes [7, 9]. Among the tasks of forecasting an important place is occupied by the methods of factor estimation and time-series analysis that includes a variety of methods and approaches including fuzzy sets [10], expert models and methods [11], genetic and neural network methods [12, 13] and other.

In application and analysis of the results with a wide array of methods of factor correlation, significance and ranking consistency between the observed results is of primary importance as it allows to distinguish and differentiate between spurious effects and / or artifacts of the particular

method or dataset, a genuine effect representing a reliable relation between a set of influencing factors and the outcome of interest.

## 4. Methodology

The development of the Covid-19 epidemics in the national and subnational jurisdictions up to the present point clearly shows that timing considerations can play major role in the epidemiological scenario observed in any given case, and for that reason can be crucial in an accurate analysis of the corresponding statistical data. To ensure the validity of the analysis from this perspective, in this work two approaches were used: 1) the data was synchronized, or aligned with respect to the duration of the development of the epidemics in a given jurisdiction, that is, the cases in the dataset were selected on the basis of having similar time of the exposure to the pandemics. And in the cases where it was not the case, 2) the statistical data of the case was resynchronized with respect to the reporting time point to the same or similar time of development in the local jurisdiction. To simplify the synchronization, the starting time point (time zero) of the global Covid-19 pandemics was defined in [2] as: *December 31, 2019 (31.12.2019)*. The period of local exposure to the epidemics is shown in the format  $TZ + y$  months is relative to start of the pandemics.

The analysis of scientific publications, in particular [1-3], showed that the following factors have a strong influence on the development of the epidemic including but not limited to the following: the time of the local development of the epidemics; traditions, social and lifestyle factors; demographics including gender and age; the level of the economic and social development; quality standard and epidemiological efficiency of the public healthcare system, and not in the least, the quality of public health policy making and execution.

In order to reduce the number of factors, national and subnational statistics from countries and regions with similar social and economic situations were selected. The aim of the study was to develop methods of reliable factor analysis and ranking by influence and verify that they can be effective in identifying principal factors in the development of epidemiological scenarios. The data and the methods are described in detail in this section.

In conclusion it needs to be noted that the intent of the work at this stage in the development of the situation was not to offer definitive answers as to the importance and ranking of certain factors of influence but rather to establish an approach and a platform for repeated and continuous analysis at different points in the time series of the cases as the situation develops that would allow to make a confident conclusion about the epidemiological and social factors with strong influence on the course of the epidemics.

### 4.1. Data

In the first stage of the analysis we are going to use only the cases of the first wave of the pandemics with the local arrival at approximately  $TZ + 1$  month (i.e., end of January, 2020). Those cases had sufficient time to develop by the time of collection of statistical data for the analysis. To ensure consistency of the data for the analysis and reduce the number of potential factors of influence of the group of identified Wave 1 cases, a subset of national and subnational cases satisfying the following consistency criteria was selected:

1. The countries in the dataset were at the similar level of development, thus excluding the influence of the factors such as the level of prosperity and development.
2. Sufficient level of confidence in the timeliness and accuracy of the statistical data provided by reporting jurisdictions.
3. A certain minimal level of local exposure to the developing epidemics identified by a minimum threshold number of cases.

Based on these selection criteria and publicly available epidemiological information from a number of trusted sources as indicated below, the dataset of 18 cases (Table 1) was constructed. The data included one provincial jurisdiction in Canada (Ontario), one state (California) and one municipal jurisdiction in the USA (New York City) and given the high geographical variation of the impacts, data with more detailed geographical breakdown is expected in the future studies. The time

point at which the data was collected was  $TZ + 3 \text{ months}$ , i.e. approximately two months of the local development of the epidemics in the selected group of jurisdictions.

The selection of national and subnational cases in the dataset allowed to exclude from consideration several common factors. Among of them were the time of arrival of the epidemics to the jurisdiction and local exposure; the level of prosperity and development; to a considerable degree, demographics (although one related factor, the median age was used in the analysis) thus helping to narrow down the number of potential factors with higher influence on the epidemiological scenario developing in the jurisdiction of the case.

In the preliminary analysis of the potential factors we found no obvious solutions to eliminating the influence of the policy management including quality of the policy making and execution of epidemics control policies and decisions; that in its turn includes a number of subfactors such as: general preparedness, effective deployment and management plans, sufficient resources, informed and trained personnel, effective and evidence-based policy making and execution and others. Due to time and resource constraints at the time of preparation of the analysis, the only available solution was found to model these parameters with a combined rating-type factor intended to reflect the overall efficiency of the public health policy.

The value of the factor was assigned manually based on the available information. An essential caveat here is that such an assignment could potentially and implicitly include some level of correlation with the observed outcomes, however at the short time of preparation of this analysis it was the only option available. We expect that future works should be able to develop more precise approaches and methods for evaluation of policy effectiveness.

#### 4.1.1. Influencing Factors

The following set of factors was considered in the analysis that follows:

1. **Policy:** a ranking parameter measuring the effectiveness of the epidemiological policy in the jurisdiction, range: 0 – 0.5, from most to less effective. The factors in evaluation of this parameter were: timeliness of response; clarity and consistency of the policy; and epidemiological preparedness of the public healthcare system to handle the onset of the epidemics. Given the challenges described earlier, an objective evaluation of this parameter will require further work.

2. **UBIP level:** defined in the range 0 – 0.5, with 0 representing band A [14] (i.e. a current universal BCG immunization program) and 0.5 – no UBIP (band C). The values in between were assigned in proportion to the time lag between the cessation of UIP and the time of the analysis. Some corrections were made for the cases where immunization was administered at an older age or only within a short time for example Spain (16 years).

3. **Smoking prevalence:** range 0 – 0.5, defined as the rate of smoking in per-cent in the population. Where a significant gender difference existed in the population with respect to this factor, the higher value was taken as it's expected to have a greater influence on the outcome.

4. **Population density:** the total population in the jurisdiction per 1 sq.km of the total area, divided by 100; we recognize that in some cases such as of very large area, averaging population over the area may lead to less consistent results; a more detailed analysis with more precisely defined geographic boundaries of the cases is intended for a future study.

5. **Age demographics:** the median age in the reporting national or subnational jurisdiction, divided by 100.

##### Epidemiological Outcome

Given considerable differences in testing practices between the reporting jurisdictions, particularly in the early phase of the epidemics, mortality per capita was chosen as a more stable and reliable indicator of the impact of the epidemics per case. Given the large spread in the range of epidemiological outcomes between the cases in the dataset, a logarithmic scale was used in the evaluation of the impact of the epidemics represented by Measured Value parameter (MV) as the logarithm of mortality per capita (in cases per 1M of population in the jurisdiction):

$$MV(locality, t) = \log\left(\frac{Mortality, cases (t)}{Population}\right) \quad (1)$$

The dataset of cases used in the study is provided in Table 1.

**Table 1**  
Covid-19 Case Dataset

Case	Policy	UBIP	Smoking prevalence	Population density	Median age	Outcome (MV)
Taiwan	0	0	0.17	6.73	0.425	-2.252
Japan	0.1	0	0.337	3.47	0.484	-0.889
Singapore	0	0	0.165	83.58	0.422	0.138
Australia	0.1	0.15	0.149	0.032	0.379	0.289
South Korea	0.1	0	0.498	5.12	0.418	1.766
Finland	0.15	0.15	0.209	0.18	0.431	2.201
Canada	0.25	0.45	0.177	0.04	0.411	2.609
Ontario (Can.)	0.25	0.5	0.129	0.14	0.398	2.678
Sweden	0.2	0.3	0.304	2.4	0.457	3.776
Germany	0.15	0.1	0.206	0.25	0.411	5.274
UK	0.5	0.35	0.199	2.81	0.405	6.022
France	0.45	0.3	0.298	1.19	0.423	6.601
Italy	0.5	0.5	0.292	2.06	0.449	7.983
Belgium	0.35	0.5	0.16	3.76	0.413	6.814
California	0.25	0.5	0.116	2.51	0.36	3.407
NYC (USA)	0.5	0.5	0.125	10.19	0.358	8.034
USA	0.5	0.5	0.195	0.36	0.383	4.638

**Sources:**

- Epidemiological outcome (incidence and mortality) [15]
- World BCG atlas [16]
- World data: smoking [17], world population data [18]
- National and subnational jurisdictions Covid-19 information [19-22].

**Reservations and qualifications:**

1. Consistency and reliability of data: the statistics on the current epidemiological outcomes reported by the national, regional and local health administrations can be affected by specific practices and policies of reporting jurisdictions.

2. An exact alignment in the time of reported data could not be confidently ascertained due to differences in the reporting practices between the jurisdictions.

Finally, it is essential to note that the analysis that follows provides a statement for a single point in the time series and that the dataset would be updated in the future at a number of points in the course of development of the epidemics. Repeating the analysis at a number of time points in the series should be able to provide more confident statement about the influence of specific factors on the development of the epidemiological scenario.

## 4.2. Factor Analysis Methods

Several statistical methods were used to evaluate the influence of the selected factors to measure the consistency of obtained results.

1. Calculation of correlation between the resulting effect (MV) and specific factor;
2. Linear regression by single factor and a combination of factors [23]
3. Evaluation of factor importance with Random Forest regression [24]
4. Evaluation of factor influence or rank with SelectKBest, a feature ranking method in sklearn machine learning and data analysis library [25].

Method 1 calculates the correlation coefficient between the outcome variable (MV) and the factor of interest. An absolute value closer to 1 indicates stronger correlation between the resulting effect and the factor.

Method 2 produces the best fit linear approximation of the selected factors on the series of the recorded outcome (MV) and the total deviation from the trend. Comparing the error for different combinations of influencing factors can show which of the factors were most effective in approximating the resulting outcome.

Methods 3 and 4 produce ranking of factors with the highest influence on the value of the outcome variable.

## 5. Results

In this section we present the results of individual and multi-factor analysis as well as a brief discussion of the findings.

### 5.1. Single Factor Analysis

The influence of selected individual factors as defined in Section 2.1.1 is shown in Table 2:

**Table 2**  
Covid-19 Factor Influence

Factor	Correlation, MV	Linear trend error, MV	RandomForest significance	SelectKBest Importance
Policy	0.893	1.528	0.838	56.25
UBIP	0.805	2.015	0.124	7.657
Smoking	0.320	3.393	0.017	0.670
Age demographics	-0.045	3.392	0.006	0.164

As can be seen from the results in the table, all methods produced consistent results with the same rating of the evaluated factors. Apart from the policy factor for which as already discussed, a strong correlation can be expected, the strongest influence factor for the data in the analysis were universal BCG immunization (UBIP), with a strong positive correlation value of *0.81*, and the smoking prevalence, at *0.32*.

The latter can be expected to be a factor of significance in the epidemics due to already established link with a number of conditions, including respiratory [26]; as a standalone factor it did not show a strong influence on the recorded outcome, however it can have noticeable influence as a secondary factor as discussed in the next section.

In the light of the information about generally less severe outcomes for younger population [27] a stronger negative correlation of the epidemiological outcome with the age demographics could have been hypothesized and expected; however the results of the single factor analysis with linear regression can be explained by a competition of factors, such as: 1) higher susceptibility of the older population group favoring the negative correlation of the recorded outcome with the median age in the jurisdiction of the case, versus higher social contact and mobility of the younger population, that can and was shown in a number of cases, to stimulate the spread of the epidemics and thus, driving the trend in the opposite direction.

The opposing trends would be more likely to produce a less pronounced overall influence of the age demographics on the epidemiological outcome in the jurisdiction, and correspondingly, a lower than expected value of the significance for this factor. A more detailed and specific study will be needed to investigate the interaction of these factors in sufficient detail.

### 5.2. Multiple Factor Analysis

In this section the cumulative effect of the combination of factors with the highest significance of the correlation with the measured epidemiological outcome as established in the previous section, namely: the epidemiological policy in the jurisdiction; the record of universal immunization (UBIP);

and smoking prevalence on the epidemiological outcome, measured as discussed previously, by logarithmic mortality per capita of the overall population in the jurisdiction, was evaluated with multiple factor linear regression.

The combination of factors was calculated as a weighed sum of factor values. In the first iteration of the analysis the weights of the factors were assigned uniformly due to insufficient historical data for more precise evaluation of weights.

**Notes**

In the cases with very large geographic area and correspondingly, low population density, a correction offset was added to account for a slower rate of development of the epidemics as follows: Canada, Australia: 0.2; Finland, Ontario, USA: 0.1; adding this correction did not change the outcome of the analysis essentially.

The results of the multi-factor analysis are presented in Table 3.

**Table 3**  
Covid-19 Multiple Factor Linear Regression Analysis

Factor	Correlation, MV	Linear trend error, MV
Policy, UBIP	0.867	1.226
Policy, Smoking	0.802	1.637
UBIP, Smoking	0.751	1.439
Policy, UBIP, Smoking	0.883	0.971

As can be observed immediately from the results, the combination of three factors with the highest single-factor influence: the policy, BCG immunization and smoking prevalence had the highest correlation, and the lowest linear regression error with the recorded epidemiological outcome.

The results also confirm UBIP as the second most influential factor among the considered, with the data available at the time. Indeed, the highest decrease in the correlation coefficient value after removing a factor from the cumulative sum was seen for the policy (11.6%) confirming it as the most influential factor among the considered, and the lowest, smoking prevalence (1.6%). Removing UBIP from the cumulative sum of the factors resulted in the correlation decrease of 8.1%, noticeably higher than other secondary factors among the considered.

The findings of this analysis can be illustrated by plotting the dependency of the epidemiological outcome (Y-axis) on the cumulative value of the dominant factors identified in the single-factor analysis (X-axis).

The diagram on the left side shows the functional relationship of epidemiological outcome in mortality per capita in the cases in the studied dataset with the weighted sum of the principal factors identified in the single factor analysis; whereas the one on the right shows the dependency of the logarithmic impact value (1) on the combined value of principal factors. A clear exponential trend can be seen in the left-side diagram vs. a linear one on the right, confirming the conclusions of Sections 3.1 and 3.2 on the significance of the identified principal factors of influence established with the selected methods in the earlier sections.

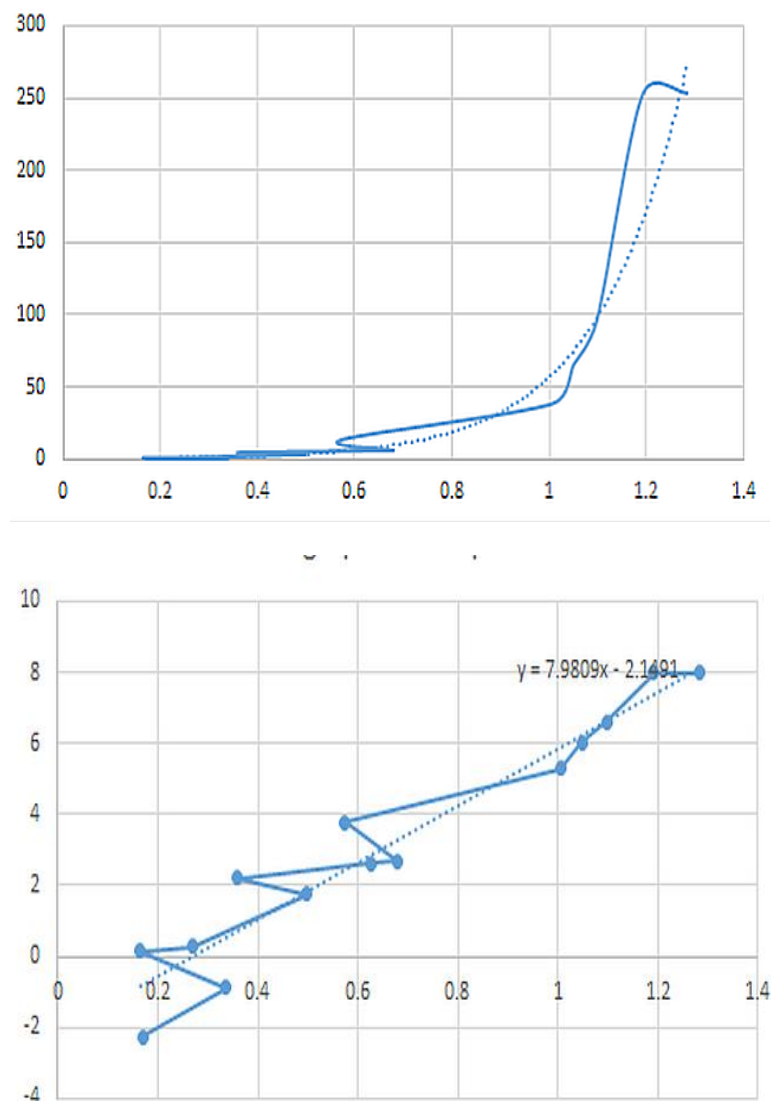
A number of outlier cases with higher than the trend impact can be seen clearly in the diagram on the right as well; these can be attributed to ternary and other factors as well as the possibility of statistical fluctuations that appear to be common occurrence with Covid-19; a detailed analysis of the other potential factors of influence will require further study.

**5.3. Specific Cases**

Some observations on the influence of specific influencing factors such as UBIP and smoking can be derived from comparison of specific cases in the dataset. While at the time of writing these cases were anecdotal and may not be sufficient for a statistically confident conclusion, they can provide some directions and rationale for further studies.

For example, comparing the incidence and the recorded epidemiological outcome between countries of Northern Europe, with similar development, cultural, demographics and some of the

other identified factors show strong correlation with the period after cessation of UBIP in the country [2]. A similar pattern can be seen by comparing national cases with differences in immunization programs in southern Europe.



**Figure 1:** Epidemiological outcome vs. dominant factors, M.p.c. (above) and MV (below)

In another example, all countries in the Asia group have similar values of most known factors, including the alignment in the time of the epidemics onset, the policy and BCG immunization (all countries are in the UBIP group A). The analysis of the case data clearly shows that countries with higher smoking prevalence: South Korea and Japan have recorded higher impact of the epidemics than those with lower smoking rates (Taiwan, Singapore).

A similar pattern can be seen with some cases in South America. Neighboring countries, with similar values of the other considered factors but with significantly different smoking rates such as: Ecuador – Peru, Chile – Argentina also show significant difference in Covid-19 impact.

Understandably, statistical fluctuations are certainly possible with a relatively small dataset used in the study and a confident conclusion can be reached with monitoring and repeated analysis of this trend over an extended period at a number of different time points.

These observations may point at a possibility of significance of some of the secondary factors such as smoking prevalence in the population in the earlier example for the overall epidemiological outcome, however a confident conclusion would require an analysis with more data and will be attempted in a future work.



## 6. Conclusion

The approaches in epidemiological factor analysis demonstrated in this work with an early Covid-19 epidemiological dataset of selected national and subnational jurisdictions and based on a number of well-known methods of data and factor analysis can be used in identification of factors with the strong influence on the development of the epidemics. This information can be instrumental in development of effective responses and policies in public health care system to minimize the impact of the epidemics and protect the population.

The findings confirm the importance of clear, timely and evidence-based epidemiological policy [28] as the factors with the highest influence on the development of the epidemiological scenario. This finding is consistently produced by all methods of analysis used in the study.

The results reported in this work offer additional arguments in support of the hypothesis of some form of general population-wide protection effect against Covid-19 as an effect of previous universal immunization program with Bacillus Calmette–Guérin vaccine (BCG), that has been reported in a number of earlier results [1-3], adding arguments to the rationale for further studies of the possible correlation and the mechanisms of such general protection with potential benefits that may extend beyond Covid-19 pandemics.

Additionally, the analysis pointed at significance of secondary factors such as smoking prevalence consistently confirmed by several independent methods. The findings of this study can be instrumental in development of epidemiological models, forecasting epidemiological scenarios and as an input to development of effective policy to control and contain the spread of the infection, with potential applications beyond Covid-19.

In conclusion, the authors would like to emphasize that the results reported in this study should not be taken as a definitive statement of a correlation between the investigated factors and the resulting effect as they relate to a single point in the time series of epidemiological scenarios in the considered cases. Rather, they are relevant as an evaluation of methods and demonstration of an approach that can be applied repeatedly over a time series of epidemiological data, allowing to reach confident conclusions by establishing and analyzing the trend over an extended period of time.

## 7. Acknowledgements

The authors are grateful to the colleagues at the Information Technology Department, National Aviation University and Uzhhorod National University for valuable discussion of the methods and findings of this study.

This work received no specific funding.

## 8. References

- [1] A. Miller, M-J. Reandelar, K. Fasciglione, V. Roumenova, Y. Li, G.H. Otazu. Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study, medRxiv 2020.03.24.20042937.
- [2] S. Dolgikh. Further evidence of a Possible Correlation Between the Severity of Covid-19 and BCG Immunization, MedRxiv doi: 10.1101/2020.04.07.20056994v1 April 2020.
- [3] L. E. Escobar, A. Molina-Cruz, C. Barillas-Mury BCG vaccine protection from severe coronavirus disease 2019 (COVID-19). Proceedings of the National Academy of Sciences, 117(30), 2020, 17720-17726.
- [4] A. Sharma, S. K. Sharma, Y. Shi, E. Bucci, E. Carafoli, G. Melino, G. Das. BCG vaccination policy and preventive chloroquine usage: do they have an impact on COVID-19 pandemic? Cell death & disease, 11(7), 2020, pp. 1-10.

- [5] K. Yitbarek, G. Abraham, T. Girma, T. Tilahun, M. Woldie. The effect of Bacillus Calmette–Guérin (BCG) vaccination in preventing severe infectious respiratory diseases other than TB: implications for the COVID-19 pandemic. *Vaccine* 2020 38(41), 2020, 6374–6380.
- [6] S. Dolgikh S. Covid-19 vs BCG: Statistical Significance Analysis, MedRxiv, doi: 10.1101/2020.06.08.20125542v2.
- [7] Kuharev, V.N., Sally, V.N., Erpert A.M. Economic-mathematical methods and models in the planning and management. Kiev: Vishcha School, 328 (1991).
- [8] Kozadaev, A.S., Arzamasians, A.A. Prediction of time series with the apparatus of artificial neural networks. The short-term forecast of air temperature. *Bulletin of the University of Tambov. Series: Natural and Technical Sciences*, №3, is 11, 299-304 (2006).
- [9] Snytiuk, V. Ye. Forecasting. Models. Methods. Algorithms: Tutorial. K. Maklout, 364 (2008).
- [10] Mulesa, O. Information Technology for time series forecasting with considering fuzzy expert evaluations, 12th international scientific and technical conference “Computer Science and Information Technologies – CSIT 2017” (Lviv, Ukraine), 105–108 (2017).
- [11] Mendel, A.S. Method counterparts in predicting short time series: expert-statistical approach. *Machine Telemechanics*, 4, 143-152 (2004).
- [12] Zaichenko, Y.P., Mohammed, Shapovalenko N.V. Fuzzy neural networks and genetic algorithms in problems of macroeconomic forecasting. *Scientific news*, 4, 20-30 (2002).
- [13] Kasabov, N. K., Song, Q. DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction. *IEEE Transactions in Fuzzy Systems*,10(2), 144-154 (2002).
- [14] Zwerling A., Behr M.A., Verma A., Brewer T.F., Menzies D., Pai M., The BCG World Atlas: a database of global BCG vaccination policies and practices. *PLOS Medicine*, doi: 10.1371/journal.pmed.1001012, 2011.
- [15] BCG World Atlas, URL: <http://www.bcgatlas.org/>
- [16] Coronavirus data and map, URL: <https://www.google.com/covid19-map/> (4.04.2020).
- [17] Our World in Data: World smoking prevalence, URL: <https://ourworldindata.org/smoking> (4.04.2020).
- [18] Worldometers: Population data, URL: <https://www.worldometers.info/world-population/> (4.04.2020).
- [19] Canada Covid-19 Situation Update, URL: <https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html?topic=tilelink> (4.04.2020).
- [20] Taiwan Center for Disease Control Covid-19 information, URL: [https://www.cdc.gov.tw/En/Category/ListContent/bg0g\\_VU\\_Ysrgkes\\_KRUDgQ](https://www.cdc.gov.tw/En/Category/ListContent/bg0g_VU_Ysrgkes_KRUDgQ) (30.03.2020).
- [21] CDC Covid-19 Advice, URL: <https://www.cdc.gov/coronavirus/2019-ncov/index.html> (2020).
- [22] NHS Covid-19 Advice, URL: <https://www.nhs.uk/conditions/coronavirus-covid-19/> (2020).
- [23] Freedman D., *Statistical Models: Theory and Practice*. Cambridge University Press (2005).
- [24] Random Forest regression, sklearn-kit, URL: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html?highlight=random%20forest#sklearn.ensemble.RandomForestRegressor>
- [25] SelectKBest feature ranking and selection, URL: sklearn-kit [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.SelectKBest.html](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html)
- [26] Johns Hopkins Medicine: Smoking and respiratory diseases, URL: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-respiratory-diseases>
- [27] Levin A.T., Cochran K.B., Walsh S.P., Assessing the age specificity of infection fatality rates for COVID-19: meta-analysis & public policy implications, National Bureau of Economic Research, working paper No. 27597, July 2020.
- [28] Zeka A., Tobias A., Leonardi G., et al. Responding to COVID-19 requires strong epidemiological evidence of environmental and societal determining factors. *The Lancet*, 4(9), 375-376 (2020).