COVID-19 mortality and BCG vaccination: defining the link using machine learning

Nathan A. Brooks
The University of Texas MD Anderson Cancer Center, Department of Urology
Ankur Puri
McKinsey & Company
Sanya Garg (sanya_garg@mckinsey.com)
McKinsey & Company
Swapnika Nag (Swapnika_Nag@mckinsey.com)
McKinsey & Company
Jacomo Corbo
McKinsey & Company
Anas El Turabi
McKinsey & Company
Noshir Kaka
McKinsey & Company
Rodney W. Zemmel
McKinsey & Company
Paul K. Hegarty
Mater Private Hospital, Department of Urology, Cork, Ireland
Ashish M. Kamat
The University of Texas MD Anderson Cancer Center, Department of Urology

Method Article

Keywords: SARS-CoV-2, COVID-19, bacillus Calmette-Guerin, BCG, machine learning, clustering, factor analysis

Posted Date: June 18th, 2020

DOI: https://doi.org/10.21203/rs.3.pex-976/v1

Abstract

Population-level data have suggested that bacille Calmette-Guerin (BCG) vaccination may lessen the severity of COVID-19; prior reports have demonstrated conflicting results. We leveraged publicly available databases and unsupervised machine learning, adjusting for established confounders designated a priori, to assign countries into similar clusters. The primary outcome was the association of deaths per million related to COVID-19 (CSM) 30 days after each included country reported 100 cases with several factors including vaccination. Validation was performed using linear regression and country-specific modeling. This protocol details the statistical analyses used to establish an association between BCG vaccination and CSM, which includes : Definition of the target function, data processing, exploratory factor analysis for variable selection, k-means clustering and step wise linear regression for validation. This protocol is differentiated from previous works on the same subject by its' comprehensive nature which considers the effect of several confounding variables while studying the association between BCG vaccination and CSM. There are still several potential measured and unmeasured confounding variables which could not be included in this study. It is also unclear if the protection from neonatal vaccination with BCG is transferable to those receiving vaccination as an adult and how long such protection lasts. The authors advise caution against routine BCG vaccination for the prevention of COVID-19 until prospective trials are completed.

Introduction

The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) and the resulting clinical condition coronavirus disease (COVID-19) have caused a worldwide pandemic. There have been 4.8 million confirmed infections and 318,000 deaths worldwide as of May 19, 2020[1] resulting in significant global and personal insecurity[2,3]. Mitigation of the pandemic requires a multifaceted strategy to reduce clinical morbidity/mortality, disease spread, and ultimately vaccination. The bacille Calmette-Guerin (BCG) vaccine has been administered to almost 4 billion people worldwide for over 100 years for the prevention of tuberculosis (TB)[4]. When given in conjunction with anti-viral vaccinations including yellow fever and influenza, patients pre-treated with BCG have demonstrated reduced viremia, decreased levels of circulating cytokines associated with cytokine storms, and no difference in, or an improved, anti-viral antibody response[5,6]. These observations may be associated with a shift in the T-cell mediated response to pathogens, enhanced trained innate immunity, and/or an as yet undiscovered pathway[7]. However, they provide an immunologic foundation which suggests BCG vaccination is associated with clinically meaningful immunomodulatory function. Several studies have observed reduced CSM/CFR in countires with active BCG vaccination programs, suggesting that there is some degree of protection from severe COVID-19 infection, especially in elderly populations[8,9,10]. Employing unsupervised machine learning methods with adjustment for numerous variables and potential established confounders associated with mortality, we evaluated the association between covariates designated *a priori* including BCG vaccination programs and mortality associated with COVID-19 at a country level utilizing pre-specified inclusion criteria.

Reagents

None were used

Equipment

To execute the proposed protocol, researchers need access to data and computational resources to run the software packages required for the implementation of this protocol. There are several software packages available which help in processing large datasets and running ML algorithms for pattern recognition. Proposed below is one of the possible methods of implementing this protocol which corresponds to the methodology detailed in the manuscript.

Software

R for statistical computing V 4.0.1 (https://cran.r-project.org/bin/windows/base/)

R Studio for statistical computing V 1.3.959 (https://rstudio.com/products/rstudio/download/)

Package "cluster"

Package "factoextra"

Package "dplyr"

Package "purrr"

Package "ggplot2"

Package "randomForest"

Package "caret"

Package "glmnet"

Package "psych"

Package "nFactors"

Package "fpc"

Package "NbClust"

Package "corrplot"

Package "regclass"

Package "leaps"

Package "bestglm"

Package "MASS"

*For detailed information on all these packages, please refer to "https://cran.r-project.org/web/packages"

Procedure

This procedure description, consists of the following : 1) Selection of countries for inclusion in analysis, 2) Data collection and consolidation, 3) Data treatment, 4) Feature selection using exploratory factor analysis, 5) Grouping of countries based on k-means clustering, 6) Validation using step wise linear regression.

1. Selection of countries for inclusion in analysis

Inclusion criteria included: more than 2,000 cases as of May 5, 2020, population greater than 5 million, and land area greater than 1,000 km^2 (to exclude city-states with the potential for non-representative population densities). Exclusion criteria included countries where BCG program start year could not be ascertained.

2. Data collection and consolidation

All data leveraged originated from publicly available data sources (Supplementary Table 1). A set of potential disease related mortality drivers spanning seven domains - socio-economic, health system readiness, environmental, existing disease burden, demographics, vaccination programs, and response to the pandemic were selected a priori (Supplementary Table 2). COVID-19 specific mortality (CSM) was the primary outcome, defined as deaths related to COVID-19 per million population assessed 30 days after 100 reported cases.

3. Data Treatment

All variables were uniformly capped at the 97'th percentile and floored at the 1'st percentile.

Missing value treatment is detailed in Supplementary Table 5.

4. Feature selection using exploratory factor analysis

We sought to group countries into comparable clusters based on previously described CSM drivers. To do this, we first assessed the correlation amongst pre-determined variables related to CSM (Supplementary

Figure 1) which demonstrated substantial correlation between several explanatory variables. Therefore, exploratory factor analysis, an unsupervised machine learning method to reduce the original set of explanatory variables, was performed. The optimum number of factors were chosen using the scree plot (Supplementary Figure 2). An elbow was observed between 7 and 8 factors (Supplementary Tables 3a and 3b)[11]. Varimax rotation was used to maximize the loading of each variable on a single factor. From each factor group, variables were chosen as inputs for subsequent clustering and multiple regression analysis based on loading characteristics and expert consensus where loading values were similar. Given the large size of the first factor group, three variables were selected from the group. Population density was considered as a distinct group given low loading (below 0.3) value and included in addition to one other variable from group 6. There was low variation of values for factors in group 7 thus no variables were included from this group. The variables selected included GDP per capita, population, population density, temperature (Celsius), percentage of the population above 65 years of age, and stringency index (SI) (a measure of country level interventions in response to COVID-19)[12].

5. Grouping of countries based on k-means clustering

Countries were clustered utilizing the k-means algorithm[13]. The optimal number of clusters was determined using the average silhouette coefficient and Dunn Index (Supplementary Table 4, Supplementary Figure 3). Countries within a cluster were further segmented based on a categorical metrics related to BCG vaccination programs including if the country's BCG vaccination program was active and at least 40 years old or 15 years old based on prior works indicating a reduction of vaccination efficacy after a period of 15-40 years[14,15]. Deaths per million from COVID-19 thirty days after each country crossed 100 reported cases was compared for countries with currently active universal BCG vaccination programs and for either the preceding 40 or 15 years and those without within a cluster. Countries within each cluster demonstrated lower coefficients of variation in testing rates compared to the whole population, and therefore normalization of testing rates was not performed.

6. Validation using step wise linear regression

To explore whether the findings were robust compared to alternate analytical approaches, we performed sensitivity analyses using linear regression models analyzing variables from each of the factor groups and CSM as the dependent variable. Step wise linear regression was used to retain variables with a statistically significant impact on CSM.

Detailed R code can be found in the Supplementary Files.

Troubleshooting

The code has been uploaded in supplementary files. Any issues with loading of packages can be resolved usually by checking the Java version and reinstalling R and RStudio if needed.

Time Taken

1) Selection of countries for inclusion in analysis - 1 day

2) Collection of data - 2 weeks

3) Data treatment - 3 to 4 days

4) Feature selection using exploratory factor analysis - 2 to 3 days

5) Grouping of countries based on k-means clustering - 1 to 1.5 weeks

6) Validation using step wise linear regression - 2 to 3 days

Anticipated Results

Of 212 countries/territories, 57 countries were included in analysis (Figure 1). Nine city states with insufficient land area or population and 141 countries with insufficient cases were excluded. Four countries met inclusion criteria but start dates for BCG vaccination programs were not available. China was excluded from the analysis as it was the first country to report widespread cases of the virus and therefore might have introduced a lead time bias.

Factor analysis resulted in the identification of six, distinct variables including GDP per capita, population, population density, temperature, percent population above 65 years, and stringency index (Table 1). Variables related to BCG administration were part of a distinct factor group. Countries within clusters had lower variation of both COVID-19 testing rates and Global Health Security Agenda (GHSA) scores, compared to the overall population. Two cluster solutions, with 6 and 9 clusters, demonstrated the highest scores (Dunn Index and Silhouette Score). Since findings were similar between the 6 and 9 cluster groups and cluster 9 only included 1 country in the 9-cluster solution (Supplementary Table 6), data for the remainder of the manuscript is presented from the six-cluster solution.

Deaths per million related to COVID-19 (CSM) was assessed 30 days after each included country reported 100 cases. Five of 6 clusters allowed division and comparison of CSM by the presence or absence of BCG vaccination programs for the preceding 15 years (BCG15) (Figure 2a). The remaining cluster composed exclusively countries with BCG vaccination programs (no comparison group-cluster 2). All 6 clusters allowed division and comparison of CSM by the presence or absence of BCG vaccination programs in the preceding 40 years (BCG40) (Figure 2b). Four of 5 clusters demonstrated lower mortality when they had BCG15 and 4 of 6 clusters demonstrated the same association with BCG40. For BCG40, specificity, clusters 1, 3, 5, and 6 demonstrated improved CSM with hazard ratios of 0.03, 0.01, 0.17, and

0.47, respectively. Cluster 2 and 4 demonstrated worse CSM with hazard ratios of 2.43 and 2.24, respectively. The results from the 9-cluster analysis were similar (supplementary table 7). Granular data regarding clustering is presented in supplemental tables 8a/b.

Using multivariate regression analysis, the presence of BCG15 (reduction of CSM by 71% (95% CI: 53 to 89%), total population (for every 1 million person increase there was a 1% decrease in CSM (95% CI: 0.53 to 1.47%), and share of the population above 65 years (CSM increased by 10% for each percent increase in population over 65 (95% CI: 2 to 18%) were shown to be significantly associated with CSM. Percent coverage metrics for vaccinations including RCV1 (Rubella), MCV1 (Measles) and OPV (Polio) were forced into the model and were not significantly associated with CSM.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases 2020; 20: 533-534.

2. Wang H-y, Xia Q, Xiong Z-z, et al. The psychological distress and coping styles in the early stages of the 2019 coronavirus disease (COVID-19) epidemic in the general mainland Chinese population: a web-based survey. PLoS One 2020; 15(5):e0233410.

3. McKibbin WJ, Fernando R. The global macroeconomic impacts of COVID-19: Seven scenarios. 2020: 19.

4. Oettinger T, Jørgensen M, Ladefoged A, Hasløv K, Andersen P. Development of the Mycobacterium bovis BCG vaccine: review of the historical and biochemical evidence for a genealogical tree. Tubercle and lung disease 1999;79:243-50.

5. Arts RJ, Moorlag SJ, Novakovic B, et al. BCG vaccination protects against experimental viral infection in humans through the induction of cytokines associated with trained immunity. Cell host & microbe 2018;23:89-100. e5.

6. Leentjens J, Kox M, Stokman R, et al. BCG vaccination enhances the immunogenicity of subsequent influenza vaccination in healthy volunteers: a randomized, placebo-controlled pilot study. The Journal of infectious diseases 2015;212:1930-8.

7. Kleinnijenhuis J, Quintin J, Preijers F, et al. Long-lasting effects of BCG vaccination on both heterologous Th1/Th17 responses and innate trained immunity. J Innate Immun 2014;6:152-8.

8. Hegarty PK, Sfakianos JP, Giannarini G, DiNardo AR, Kamat AM. COVID-19 and Bacillus Calmette-Guérin: What is the link? European Urology Oncology 2020;S2588-9311(20)30049-3.

9. Miller A, Reandelar MJ, Fasciglione K, Roumenova V, Li Y, Otazu GH. Correlation between universal BCG vaccination policy and reduced morbidity and mortality for COVID-19: an epidemiological study. MedRxiv

2020.

10. Sala G, Miyakawa T. Association of BCG vaccination policy with prevalence and mortality of COVID-19. medRxiv 2020.

11. UCLA: Statistical Consulting Group. 2006. (Accessed May2020, at https://stats.idre.ucla.edu/stata/ado/analysis/.)

12. Hale T, Petherick A, Phillips T, Webster S. Variation in government responses to COVID-19. Blavatnik School of Government Working Paper 2020;31.

13. Charrad M, Ghazzali N, Boiteau V, Niknafs A. Determining the number of clusters using NbClust package. MSDM 2014 2014:1.

14. Nguipdop-Djomo P, Heldal E, Rodrigues LC, Abubakar I, Mangtani P. Duration of BCG protection against tuberculosis and change in effectiveness with time since vaccination in Norway: a retrospective population-based cohort study. The Lancet infectious diseases 2016;16:219-26.

15. Lord J, Willis S, Eatock J, et al. Economic modelling of diagnostic and treatment pathways in National Institute for Health and Care Excellence clinical guidelines: the Modelling Algorithm Pathways in Guidelines (MAPGuide) project. Health technology assessment (Winchester, England) 2013;17:1-192.

Acknowledgements

Figures

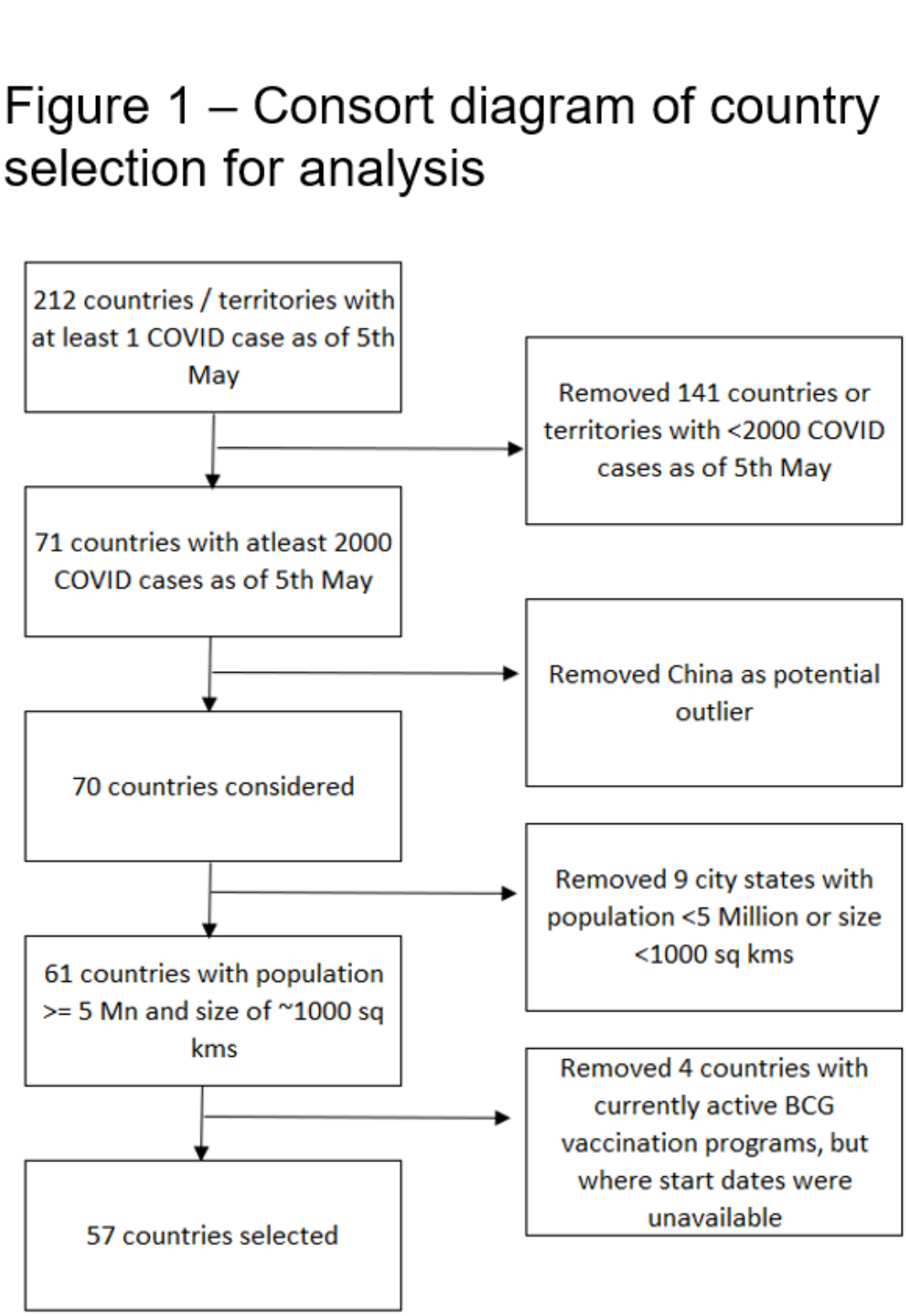

Figure 1

Consort diagram for country selection

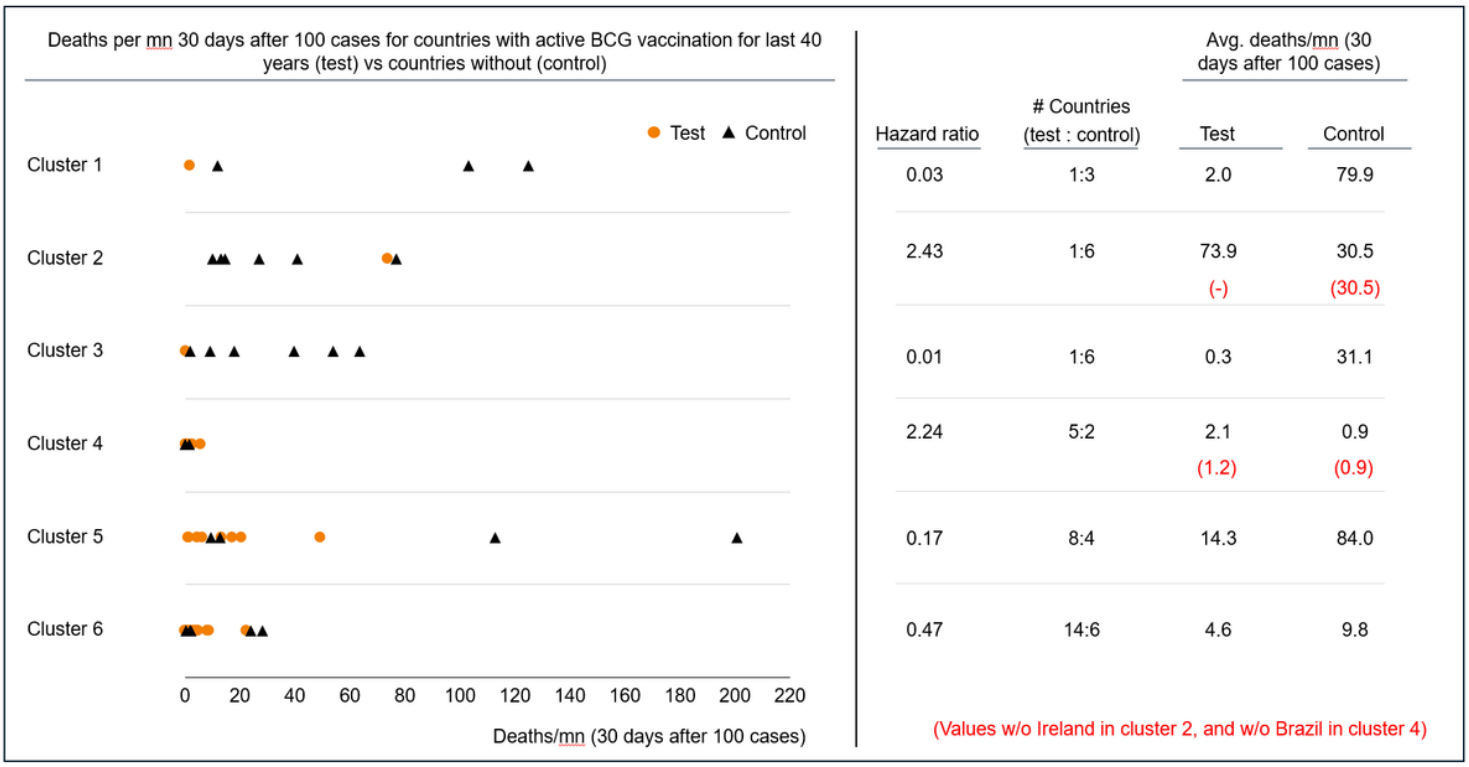

Figure 2

Cluster wise CSM comparisons based on BCG vaccination present in last 40 years

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Code1Rmarkdownscript.html
- Figure2ScreePlotforFatorAnalysis.docx
- Table1DataSources.docx
- Figure1CorrelationMatrix.docx
- Table2DataDictionary.docx
- Figure3WSSCurveforclustering.docx
- Table3FactorAnalysisSummary.docx
- Table4ClusteringStatistics.docx
- Table5MissingValueTreatment.docx
- Table69clusterdescription.docx
- Table7AgeStratifiedMortalityAnalysis.docx
- Table8Granularclusterdescriptions.docx
- Table9Linearregressionresults.docx