UNIVERSIDADE CATÓLICA PORTUGUESA

Explaining mortality rates from COVID-19

An application of Business Analytics

Daniela Ribeiro Silva

Católica Porto Business School

2020

UNIVERSIDADE CATÓLICA PORTUGUESA

Explaining mortality rates from COVID-19

An application of Business Analytics

Thesis presented to Católica Porto Business School
to fulfill the requirements of Master in Management

by

Daniela Ribeiro Silva

Advisors:

PhD Maria da Conceição Andrade e Silva
PhD António Pedro de Pinho de Brito Duarte Silva

Católica Porto Business School
June 2020

Abstract

The COVID-19 pandemic has generated a lot of demand for responses to prevention, treatment, how to control, how to predict evolutions, among others. This thesis aims to answer the question about what affects mortality. Thus, through the use of Analytics, 26 different variables were studied for 37 duly selected countries. The results showed that the country's economic structure has no impact on mortality, while vaccination policy for BCG, changes in mobility within the country, such as “stay at home”, and the prevalence of diabetes have an impact on mortality.

Keywords: COVID-19, mortality, business analytics.

Resumo

A pandemia COVID-19 tem gerado muita procura por respostas para prevenção, tratamento, como controlar, como prever evoluções, entre outras. Esta tese pretende responder à pergunta sobre o que afeta a mortalidade. Assim, através do uso do Analytics foram estudadas 26 diferentes variáveis para 37 países devidamente selecionados. Os resultados permitiram concluir que a estrutura económica do país não tem impacto na mortalidade, enquanto que a política de vacinação para a BCG, as alterações da mobilidade dentro do país, tais como o "stay at home", e a prevalência de diabetes têm impacto para a mortalidade.

Palavras-chave: COVID-19, mortalidade, business analytics.

Acknowledgments

To my supervisor, Professor Conceição Silva, for all the feedback and input from beginning to end, fundamental to the realization of this thesis. For having accompanied me attentively, ensuring that the best conditions were always met to better conduct my work.

To my supervisor, Professor Pedro Silva, for the tireless help and availability at the right time. For all the support, all the assertive inputs and comments that have become fundamental to the finalization of my thesis.

To my family, my parents and my brother, who have always believed in me, often giving up the best for them, to provide the best for me. For all the times I was allowed my only concern on this journey was this thesis, and for all the patience shown in this long way.

To my friends, T4+1, who were part of this journey in the best possible way. Not only for being my company on long study days, but also for being the best company to do all the other stuff. For all the moments of leisure and for all the fundamental presence in the realization of this work, my thanks to Carlos, Dânia, Luís and Seara.

To my friends, marega branco, for being my oldest friends and the best friends I could have. Because they were part of my growth both academically and personally and because they believed in me and always motivated me on this long journey, always making me believe that I was really doing a good job and that everything was going to be okay.

Lastly, to Luis, for being there, always. Cheering me up when I most needed, being supportive and always wanting the best for me.

Contents

Table Contents

Figure Contents

Abbreviations

BA- Business Analytics

BCG- Bacilo Calmette–Guérin

GDP- Gross Domestic Product

IBM- International Business Machines

ICU- Intensive care units

LRI- Low respiratory infections

NLR - Neutrophil-lymphocyte ratio

OECD- Organisation for Economic Co-operation and Development

RD- Respiratory Disease

RFID – Radio-Frequency identification

US - United States of America

WHO- World Health Organization

1.Introduction

Nowadays, the COVID-19 pandemic is strongly present worldwide, and has been affecting the way of living, economies, and most important, is being the cause of a lot of deaths. There is, on a global scale, a huge concern for seeking solutions to this pandemic, including trying to find a vaccine, sometimes trying or find the best medication, and also trying to explain how the pandemic evolves, among others.

In this thesis, we seek, through an application of analytics, to explain how mortality varies between countries. Thus, we take not only the daily data regarding updates of COVID-19 but also seek to introduce variables that effectively differentiate the country, both in terms of demographics, economic structure, time of response to the pandemic, among others, in order to understand which characteristics or even decisions directly affect mortality in countries.

Thus, this thesis is divided into three fundamental parts. In the first part we try to define business analytics (BA) and its main tools, then we talk about BA applications in healthcare so far. Also, in the first part, we started to focus more on the theme of this thesis, an application of business analytics to explain mortality by COVID-19, and so we begin to briefly describe the pandemic and then present in what ways analytics is being used in this context. Finally, we present some studies conducted up to the moment where some seek to predict the evolution of the pandemic curve and others seek to find the reasons for the differences between results for different countries.

Then, in section 4, we put forth the method used to estimate the model, how the variables to be studied were collected, the choice of the countries studied that were in total 37, the optimization process of the model and finally, the process of selecting variables to be included in the model.

Finally, we present the results of the estimated model, with deaths from COVID-19 as a dependent variable. After explaining the results of this model, two other models are estimated to see if it is possible to go further to explain differences between countries on severity of the disease. The results of these two estimated models are discussed and the importance of some variables is adressed.

2. Literature Review

2.1 Business Analytics

According to Davenport et al. (2007), Business Analytics (BA) is the use of data, statistical analysis, quantitative methods and mathematical or computer-based models to help managers gain improved insight about their operations so they can make better decisions.

Later, Delen and Demirkan (2013)defined Analytics as a facilitator to draw business objectives through the report of data, the analysis of trends and the creation of predictive models to better understand future problems and opportunities and optimizing processes to increase organizational performance. Beyond this definition they were able to notice the rapid evolution of Business Analytics and pointed the utility of the data to the decision makers as the main reason. With the use of business analytics, it was possible to provide the information the decision makers needed and with the guarantee of quality Delen and Demirkan (2013).

Davenport and Patil (2012) went even further declaring data scientist as the sexiest job of the 21^{st} century.

In 2014, data were deemed the new oil (Acito and Khatri 2014), by recognizing the importance of leveraging value from them, which should require an alignment between strategy and desirable business performance with analytic tasks and capabilities.

Finally, Griffin and Wright (2015) pointed business analytics as the influencer of almost every aspect of major companies' decisions, strategies and forecasting.

Business Analytics is a recent trend but has been growing over the last years. In 2011, on a survey performed by IBM Tech Trends Report of over 4000 information technology professionals from 93 countries and 25 industries, business analytics was identified as one of the four major technology trends in the 2010s.

Still in 2011, on another survey, performed by Bloomberg Businessweek,
97% of companies with revenues above $100 million were found to use business analytics.

This rapid evolution and growth of business analytic led Hal Varian, Chief Economist at Google comment the following: "So what's getting ubiquitous and cheap? Data. And what is complementary to data? Analysis. So, my recommendation is to take lots of courses about how to manipulate and analyze data: databases, machine learning, econometrics, statistics, visualization, and so on."

On 2014, a search on google scholar suggested 19400 articles published on business analytics since 2012, which is about one article per hour over two years. (Acito and Khatri 2014). Doing the research nowadays (April, 2020) we can see that there are a total of 736000 articles published in the field of business analytics where 34800 were posted since 2019.

So, it is clear that Business Analytics is getting increasingly useful and used on these days, but how does it work?

Having the big data extracted from both internal and external data sources, it is possible for managers to utilize data analytics techniques in order to answer questions like, what has happened, what will happen and what are the optimized solutions. These questions are usually grouped into three dimensions of business analytics, descriptive analytics, predictive analytics and prescriptive analytics (Appelbaum et al. 2017).

2.1.1Descriptive Analytics

Descriptive analytics is the most used approach to business analytics and it usually uses descriptive statistics, dashboards and other types of visualizations (Dilla, Janvrin, and Raschke 2010).

The main output of a descriptive analysis is to identify business opportunities and problems (Delen and Demirkan 2013) and also to find an answer for what happened/is happening (Souza 2014) by analyzing the data collected.

2.1.2Predictive Analytics

This analytics approach is the next step after the knowledge provided by descriptive analytics.

Predictive analytics is used to answer the question of what will be happening and why it will be happening (Souza 2014) using data and mathematical techniques, data mining and statistical timeseries forecasting to project the future (Delen and Demirkan 2013).

Predictive models calculate probability of future events using historical data collected over time Appelbaum et al. (2017).

2.1.3 Prescriptive Analytics

Prescriptive analytics is about making decisions based on both descriptive and predictive models and mathematical optimization models (Souza, 2014).

The main output of this analysis is a course of action given a specific situation or an amount of information that is provided to a decision maker (Delen and Demirkan 2013).

Despite the techniques and tools for predictive and prescriptive analytics may seem similar ,"The main difference between prescriptive and predictive analytics is not one of required data types, but one of orientation – that is, is this an optimization query or a trend-based analysis?"(Appelbaum et al. 2017)

The three dimensions of Business Analytics can be summarized as follows:

Table 1- Taxonomy of Business Analytics

	Descriptive Analytics	**Predictive Analytics**	**Prescriptive Analytics**
Question	What happened in the past? / What is happening now?	What will happen in the future? Why it will happen in the future?	How should we act in the future?
Process/Focus	Reporting Measuring Monitoring KPI	Forecasting Probability Assessment Risk management Prediction Data mining Text mining	Scenario based planning Strategy formulation Strategy simulation Optimization Decision modeling Expert systems
Tools/Techniques	Static and interactive reports Dashboards Performance scorecards Data warehousing	What-if Analysis Machine Learning Predictive modeling Neural Networks Data visualization	Discrete choice modeling Linear and non-linear programming Value analysis
Outcome	Well defined business problems and opportunities	Accurate projections of the future states and conditions	Best possible business decisions and transactions

2.2. Business Analytics in Healthcare

Nowadays we can easily collect and analyze data, and the role of Business Analytics in Healthcare should be to transform that data in order to improve healthcare delivery. There are many sectors that benefit from the use of analytics, but healthcare is probably one of the sectors where analytics can have a bigger impact. The analytics in healthcare is of utmost importance given the existence of large amounts of data.

Ward, Marsolo, and Froehle (2014), mentioned seven different platforms for data generation in the healthcare context:

Table 2- Example of data sources within a healthcare delivery system (Ward, Marsolo, and Froehle 2014)

Data source	Data Generated
EHRs (electronic health record)	Clinical documentation, patient history, results reporting, patient orders.
LIMS (Laboratory information management systems)	Laboratory results. Typically interfaced with EHRs.
Diagnostic or monitoring instruments	Range from images (e.g., magnetic resonance imaging) to numbers (e.g., vital signs) to text report (result interpretation). May or may not be interfaced with EHRs.
Insurance claims/billing	Information on what was done to the patient during a visit, the cost of those services and the expected payment. The level of service is often determined from data in EHRs.
Pharmacy	Information on the fulfillment of medication orders. Not typically part of EHRs.
Human resources and supply chain	Lists of employees and their roles in the institution and the location and utilization of medical supplies. Not typically interfaced with EHRs.
Real- time locating systems	Positions and interactions of assets and people.

Meanwhile, on a systematic review regarding big data analytics and healthcare, and after selecting 58 articles out of 12390, Mehta and Pandit (2018) pointed to the following data sources.

Table 3- Sources of Healthcare (Mehta and Pandit 2018)

Data Sourced	Data Generated
EMRs	Detailed patient-related information (physician prescriptions, medications, medical history)
Diagnostic	Diagnostic Results (imaging results, laboratory reports)
Biomarkers	Molecular data (genomic, proteomic, transcriptomic, metabolomic)
Ancillary	Administrative data (admission, discharge, transfer) & financial data (claims)
Medical Claims	Medical reimbursement data (procedures, hospital stay, insurance policy details)
Prescription Claims	Prescription reimbursement data (drugs, dose, duration)
Clinical Trails	Design parameters (compound, size, end points)
Social Media	Community discussions
Wearable & Sensors	Wellness & lifestyle data (smartphones, fitness monitors)

Having the data sources set, Mehta and Pandit (2018) and Ward, Marsolo, and Froehle (2014) found the same applications for business.

Ward et al. (2014) distinguish the different uses that Analytics have on the field of Healthcare, which we highlight below:

- Dashboards and control charts
- Genetics
- Cost of treatment and guiding investments
- Chronic disease databases
- Disaster planning
- Patient flow
- RFID (Radio-Frequency identification)

Charts allow the visualization of data in real-time which helps on making decisions faster. Charts are a widget type on a dashboard that display trends and metrics in live data.

In the case of Healthcare, control charts and dashboards are used to monitor outcomes so the Hospital can have information on patient waiting times, rooms utilization, operational metrics, etc.

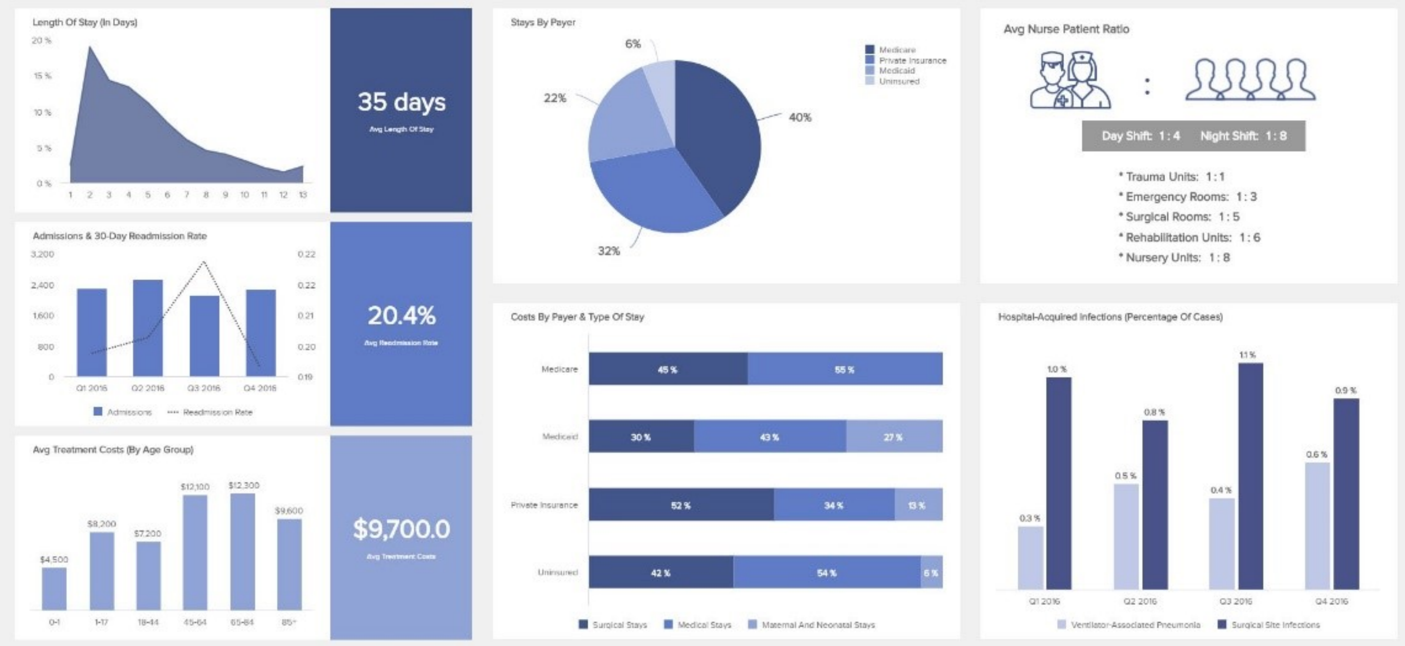


Figure 1-Open Saint Martins Clinical Performance Dashboard in 2016 from: [*https://www.datapine.com/dashboard-examples-and-templates/healthcare*](https://www.datapine.com/dashboard-examples-and-templates/healthcare)

In Figure 1 it is possible to see a Dashboard from Saint Martins Clinic the KPI's of the organization are shown. Another type of dashboard could have been Patient Satisfaction Dashboard, which allows the managers of the Hospital to know how their service is being evaluated by the patients in what aspects they should improve.

According to Ward, Marsolo, and Froehle (2014) integrating genetic data into healthcare practices provides a lot of interesting possibilities for the field of analytics, like testing on risk factors for certain genetic conditions.

Furthermore, by using patient's genomic information it is possible to build individualized strategies for each patient since big data, trough the examination of large data sets uncovers unknow correlations, hidden patterns which make it possible to get some other insights (He, Ge, and He 2017).

Analytics can be used in healthcare in a more managerial way, in order to compare the costs and effectiveness of medical devices, treatments, interventions, etc. One example is that an hospital can choose to stop utilizing some replacement hips by comparing its cost and performance and exclude the ones that cost more and provide less outcomes for patients (Ward, Marsolo, and Froehle 2014)

The use of analytics in chronic diseases consists in having a large database that allow the creation of tools and processes to better support and supervise the patient along its chronic disease. This support and supervision is done by using the data to improve the quality of information during the patient encounter, develop personalized software applications, estimate comparative treatment effects of biologic agents and to develop governance and data-sharing processes (Ward, Marsolo, and Froehle 2014)

Another example where analytics can play an important role is in situations of disaster which can occur in different forms like hurricanes, terrorist attacks, airplane crashes, infectious diseases and so on. Given the current situation, the spread of COVID-19 worldwide, an infectious disease that cause respiratory failure, the use of ventilator to assist with breathing until a patient recover can be really required and vital.

In these cases, having a real-time data on the availability of ventilators and other equipment and resources within a location can result in a more efficient way to organize resources and so resulting on improved outcomes and the avoidance of deleterious outcomes in consequence of delayed treatments (Ward, Marsolo, and Froehle 2014) which can save lives.

For decades, the management of resources in healthcare facilities has been affected by the lack of knowledge of the managers on the resources needed and staff. (Ward, Marsolo, and Froehle 2014)

The improvement of patient flow, the reduction of waiting and the potential improvement of patient outcomes have been becoming a reality with the use of analytical tools like closed-form mathematical modelling and empirical and statistical analyses (Froehle and Magazine 2013). Another application of analytics is the use of predictive analytics to forecast and plan for excessive patient waiting in emergency departments, in order to build and observation unit for patients requiring abbreviated admissions (Hoot et al. 2008).

Beyond asset management, Radio Frequency Identification (RFIC) is being integrated into healthcare in order to provide real-time identification and tracking of patients and staff.

Applying this real-time tracking to analytics, it can be used for example to evaluate potential retained surgical equipment (Rogers, Jones, and Oleynikov 2007) and for patients with chronic diseases like diabetes and hypertension that can be monitored for achieving specific numeric targets in their blood sugar and blood pressure (Moore 2009).

In these devices, the main issues are the accuracy of data and privacy issues for both patients and staff (Hawrylak et al. 2012).

3. COVID-19 Case

Coronavirus Disease 2019 (COVID-19) recognized in December 2019, is the latest threat to global health and it seems to be caused by a novel coronavirus that is structurally related to the virus that causes severe acute respiratory syndrome (Fauci, Lane, and Redfield 2020).

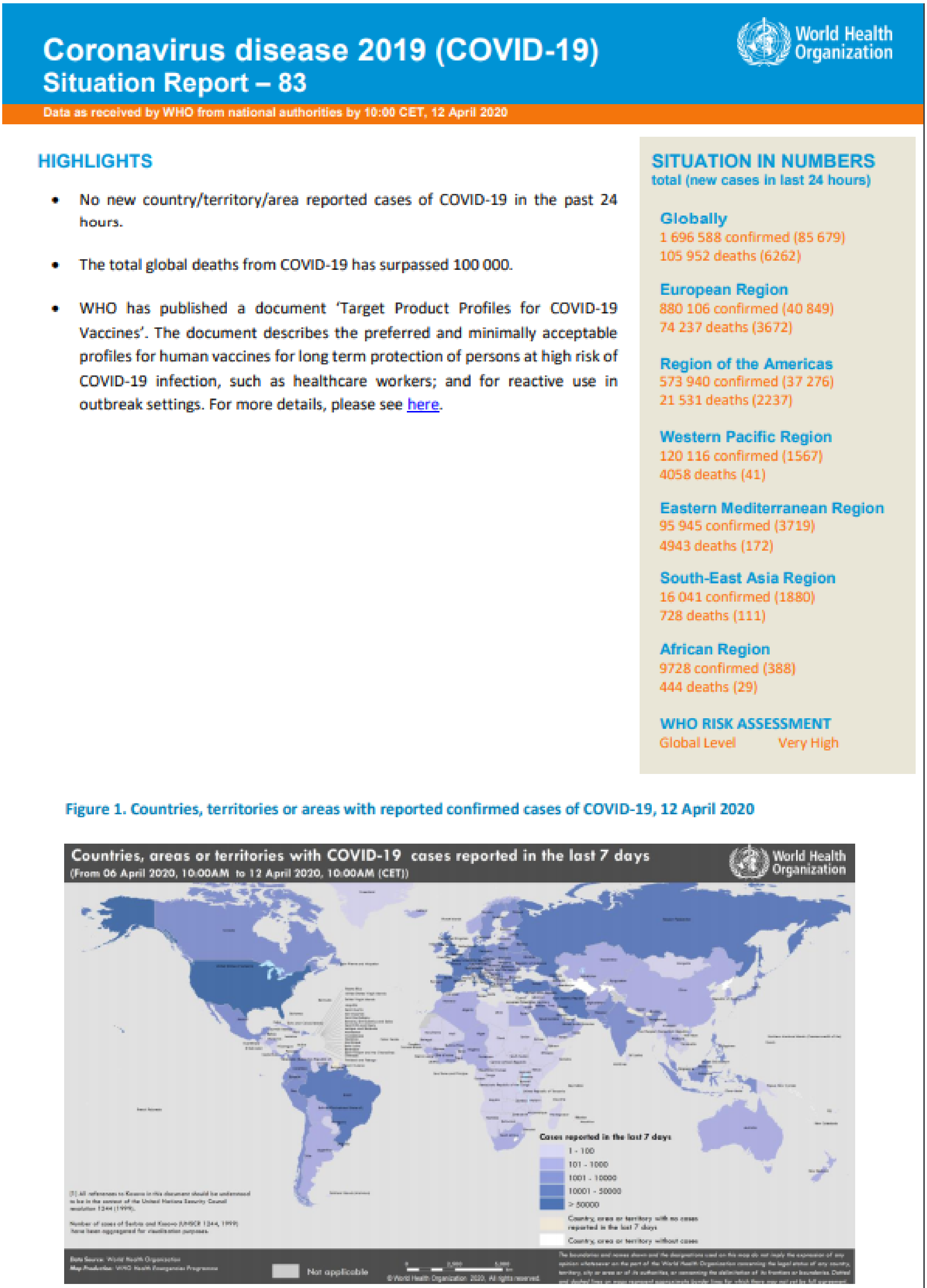
There are strike similarities between COVID-19 and others infectious respiratory diseases like SARS (severe acute respiratory syndrome) but its differences determined whether to use the same measures used for SARS or not. COVID-19 has a bigger infectious period, transmissibility, clinical severity and extent of community spread (Wilder-Smith, Chiew, and Lee 2020)

The novelty of the epidemic situation that the world is living is demanding a response from health services and institutions (to treat patients and to study and research vaccines and treatments), but also from other researchers outside the health system. In particular, data scientists have been very active during this period, because the enormous amount of data that has been generated needs to be analyzed, displayed and treated to inform the governments, managers, healthcare decision makers and the population. In the next sections we will outline the main analytic tools that have been provided in the various areas of analytics within the pandemic.

3.1The use of analytics in COVID-19

The World Health Organization (WHO) provides daily situation reports since 21 January 2020. In all reports the first page highlights the most important topics and provides a global situation in numbers for the world and for regions like the total number of cases confirmed per region, total number of deaths confirmed per region, new cases in last 24 hours confirmed per region and new cases in last 24 hours of deaths per region. It also shows a map where it is possible to note the most affected areas in the last seven days.

Figure 2 shows the first page of a daily report, on 12 April 2020.



Coronavirus disease 2019 (COVID-19)
Situation Report – 83

World Health Organization

Data as received by WHO from national authorities by 10:00 CET, 12 April 2020

HIGHLIGHTS

- No new country/territory/area reported cases of COVID-19 in the past 24 hours.
- The total global deaths from COVID-19 has surpassed 100 000.
- WHO has published a document 'Target Product Profiles for COVID-19 Vaccines'. The document describes the preferred and minimally acceptable profiles for human vaccines for long term protection of persons at high risk of COVID-19 infection, such as healthcare workers; and for reactive use in outbreak settings. For more details, please see here.

SITUATION IN NUMBERS
total (new cases in last 24 hours)

Globally
1 696 588 confirmed (85 679)
105 952 deaths (6262)

European Region
880 106 confirmed (40 849)
74 237 deaths (3672)

Region of the Americas
573 940 confirmed (37 276)
21 531 deaths (2237)

Western Pacific Region
120 116 confirmed (1567)
4058 deaths (41)

Eastern Mediterranean Region
95 945 confirmed (3719)
4943 deaths (172)

South-East Asia Region
16 041 confirmed (1880)
728 deaths (111)

African Region
9728 confirmed (388)
444 deaths (29)

WHO RISK ASSESSMENT
Global Level Very High

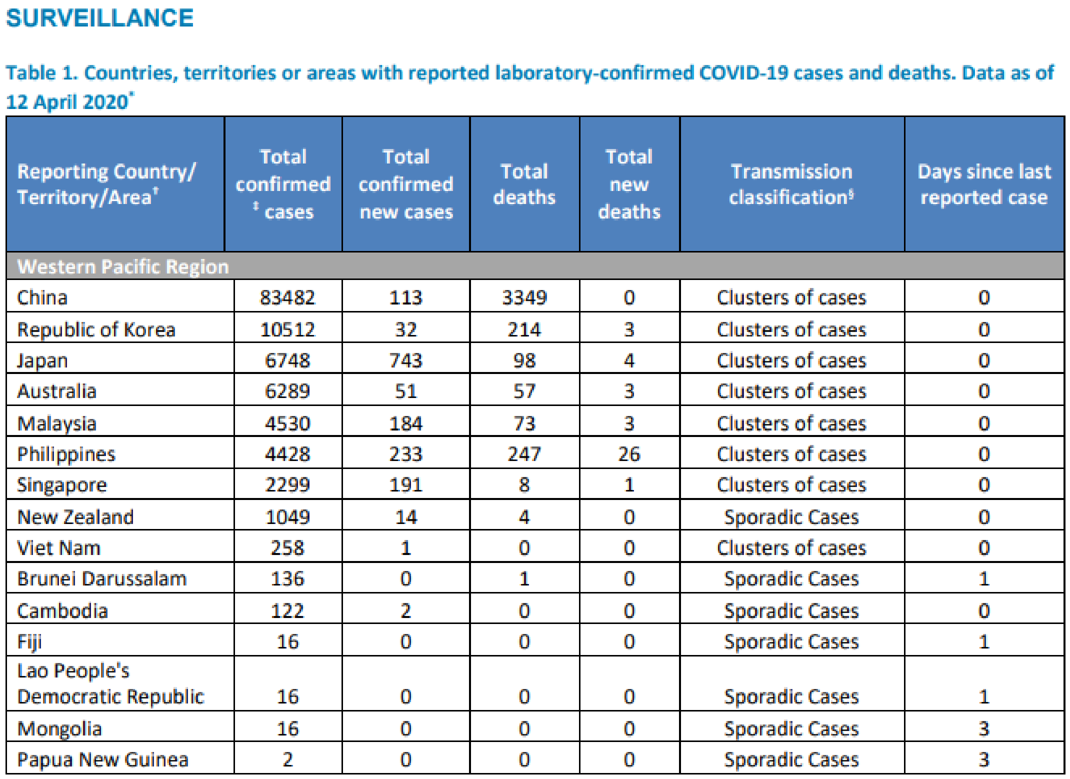
Figure 1. Countries, territories or areas with reported confirmed cases of COVID-19, 12 April 2020

Countries, areas or territories with COVID-19 cases reported in the last 7 days
(From 06 April 2020, 10:00AM to 12 April 2020, 10:00AM (CET))
World Health Organization
Cases reported in the last 7 days
1 - 100
101 - 1000
1001 - 10000
10001 - 50000
> 50000
Country, area or territory with no cases reported in the last 7 days
Country, area or territory without cases
[1] All references to Kosovo in this document should be understood to be in the context of the United Nations Security Council resolution 1244 (1999).
Number of cases of Serbia and Kosovo (UNSCR 1244, 1999) have been aggregated for visualization purposes.
Data Source: World Health Organization
Map Production: WHO Health Emergencies Programme
Not applicable
©World Health Organization 2020, All rights reserved.
The boundaries and names shown and the designations used on this map do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

Figure 2 - WHO daily report

On the following pages, the report has a table with Countries, territories or areas with reported laboratory-confirmed COVID-19 cases and deaths for 12 April 2020. The table provides for all countries, territories and areas affected with COVID-19 information regarding total cases confirmed including both domestic and repatriated cases, total confirmed new cases, total deaths, total new deaths, transmission classification which can be *no cases, sporadic cases, clusters of cases* or *community transmission* and days since last reported case.

Figure 3 shows part of a table included on the 12 April 2020 situation report of WHO.



SURVEILLANCE

Table 1. Countries, territories or areas with reported laboratory-confirmed COVID-19 cases and deaths. Data as of 12 April 2020[*]

Reporting Country/ Territory/Area[†]	Total confirmed [‡] cases	Total confirmed new cases	Total deaths	Total new deaths	Transmission classification[§]	Days since last reported case
Western Pacific Region						
China	83482	113	3349	0	Clusters of cases	0
Republic of Korea	10512	32	214	3	Clusters of cases	0
Japan	6748	743	98	4	Clusters of cases	0
Australia	6289	51	57	3	Clusters of cases	0
Malaysia	4530	184	73	3	Clusters of cases	0
Philippines	4428	233	247	26	Clusters of cases	0
Singapore	2299	191	8	1	Clusters of cases	0
New Zealand	1049	14	4	0	Sporadic Cases	0
Viet Nam	258	1	0	0	Clusters of cases	0
Brunei Darussalam	136	0	1	0	Sporadic Cases	1
Cambodia	122	2	0	0	Sporadic Cases	0
Fiji	16	0	0	0	Sporadic Cases	1
Lao People's Democratic Republic	16	0	0	0	Sporadic Cases	1
Mongolia	16	0	0	0	Sporadic Cases	3
Papua New Guinea	2	0	0	0	Sporadic Cases	3

Figure 3- Information provided on 12 April 2020 situation report of WHO.

And finally, WHO provides a Graph that shows the epidemic curve of confirmed COVID-19, by date of report an WHO region through, in this case, 12 April 2020 as in Figure 4.

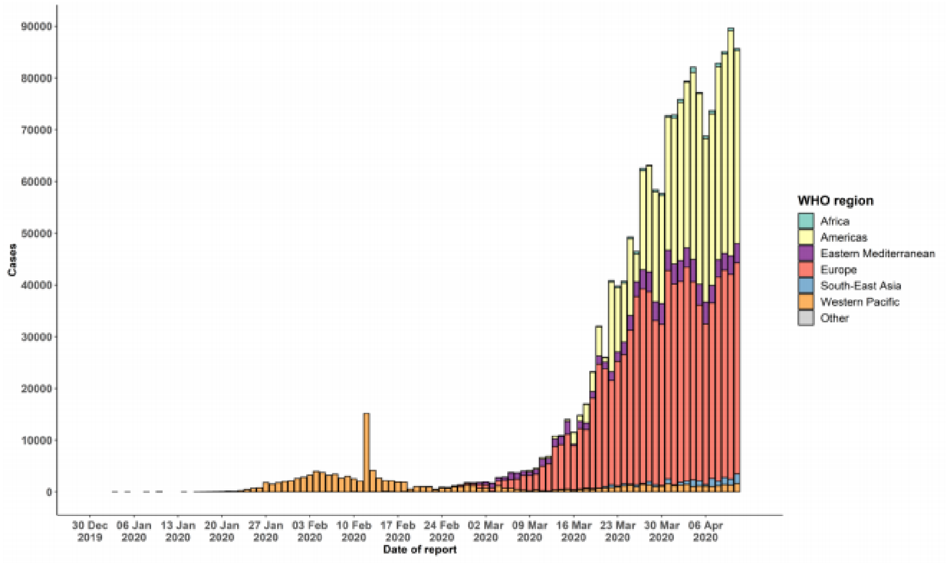
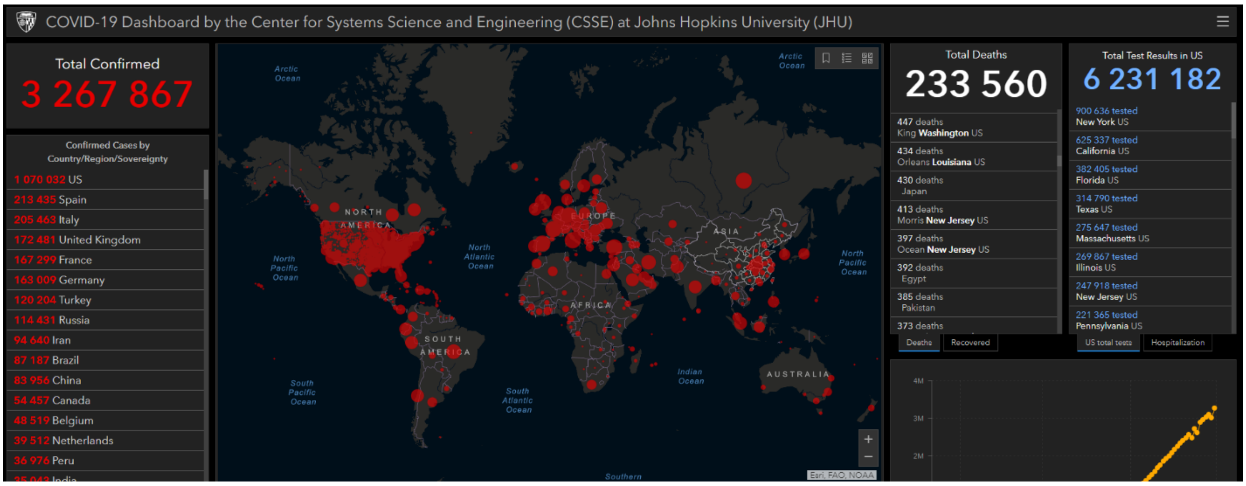


Figure 4-Epidemic curve of confirmed COVID-19 by date

Dashboards have been very utilized to show the data from coronavirus worldwide. They are simple to read and have all the information gathered with a dynamic form of analysis and interaction.

Figure 5 shows another dashboard, created by the Johns Hopkins University, of the worldwide situation of COVID-13 on April 30.


COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

Total Confirmed
3 267 867

Confirmed Cases by Country/Region/Sovereignty

1 070 032 US
213 435 Spain
205 463 Italy
172 481 United Kingdom
167 299 France
163 009 Germany
120 204 Turkey
114 431 Russia
94 640 Iran
87 187 Brazil
83 956 China
54 457 Canada
48 519 Belgium
39 512 Netherlands
36 976 Peru
35 043 India

Total Deaths
233 560

447 deaths
King Washington US

434 deaths
Orleans Louisiana US

430 deaths
Japan

413 deaths
Morris New Jersey US

397 deaths
Ocean New Jersey US

392 deaths
Egypt

385 deaths
Pakistan

373 deaths

Total Test Results in US
6 231 182

900 636 tested
New York US

625 337 tested
California US

382 405 tested
Florida US

314 790 tested
Texas US

275 647 tested
Massachusetts US

269 867 tested
Illinois US

247 918 tested
New Jersey US

221 365 tested
Pennsylvania US

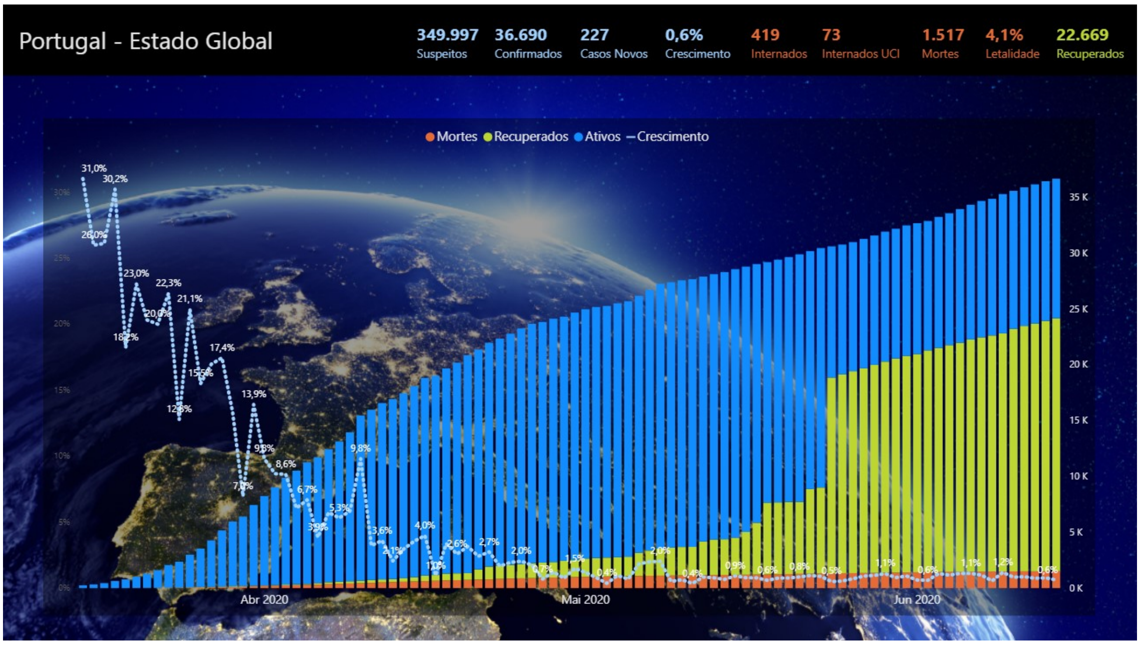
Figure 5-COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)

In this dashboard it is possible to find information on total confirmed cases worldwide, total confirmed cases in every country, total deaths and recovered worldwide, total deaths and recovered by country, total testes result in US and hospitalization by states in US.

In the map in red we can see the most affected areas, and below the map we can choose different tabs that give us information of Cumulative confirmed cases, active cases, incidence rate, case-fatality ratio, testing rate and hospitalization rate, where all these are for each country except the last two, testing and hospitalizations which are only for the US.

Dashboards have been widely used in COVID-19 and not only global dashboards are available but also country specific ones. For example, in Portugal, COTEC provides 4 different kinds of dashboards.

The first one is figure 6, where COTEC presents the evolution in time, since March 15, of deaths, recoveries, active cases and growth of COVID-19 cases nationally.



Portugal - Estado Global

349.997	36.690	227	0,6%	419	73	1.517	4,1%	22.669
Suspeitos	Confirmados	Casos Novos	Crescimento	Internados	Internados UCI	Mortes	Letalidade	Recuperados

Figure 6- Global state of COVID-19 in Portugal

Furthermore, in the header they provided the real time data of total suspects, total confirmed, new cases, growth of cases of COVID-19, hospitalizations, hospitalizations in ICU, total deaths, lethality and total recoveries.

By clicking in any of the days on the below graph, the header changes for the data of that specific day.

The second one, in figure 7, provides the evolution of all the data in the header with six different graphs representing the evolution over time of:

(1) Confirmed cases and new cases;

(2) Suspected cases, confirmed cases and the ratio between both confirmed cases and suspected cases;

(3) Recoveries and in recover;

(4) Hospitalizations and the ratio of hospitalizations by confirmed cases;

(5) Deaths and lethality;

(6) Hospitalizations, hospitalizations in ICU and the ratio between both.

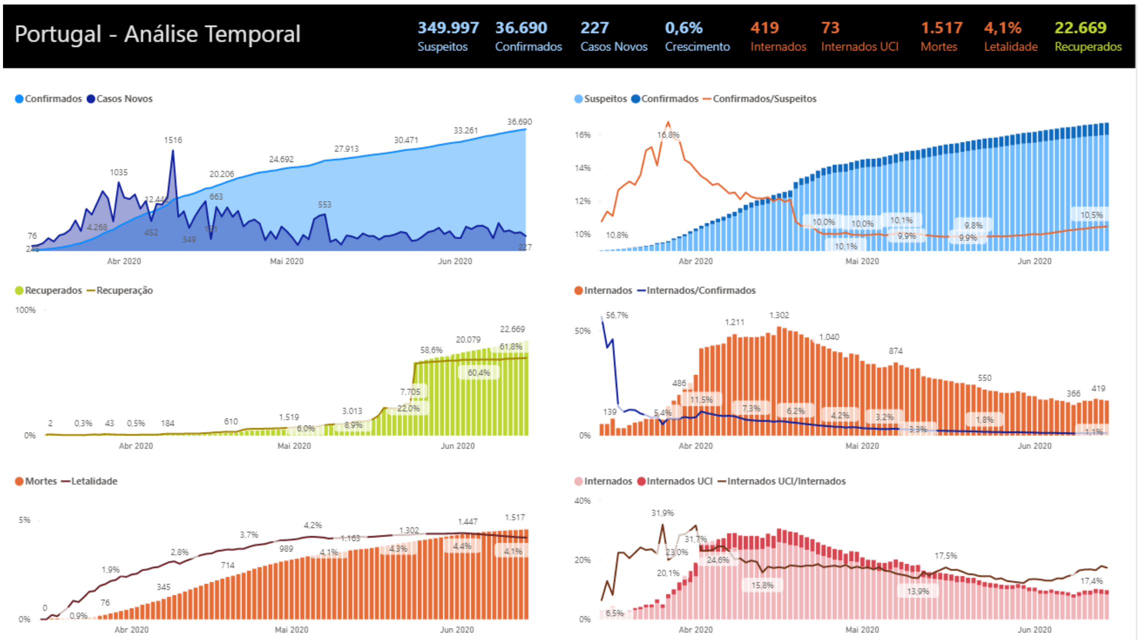


Figure 7- COVID-19 temporal analysis in Portugal

Beyond the COVID-19 global state and temporal analysis in Portugal presented above, in the third dashboard COTEC provides age and gender analysis.

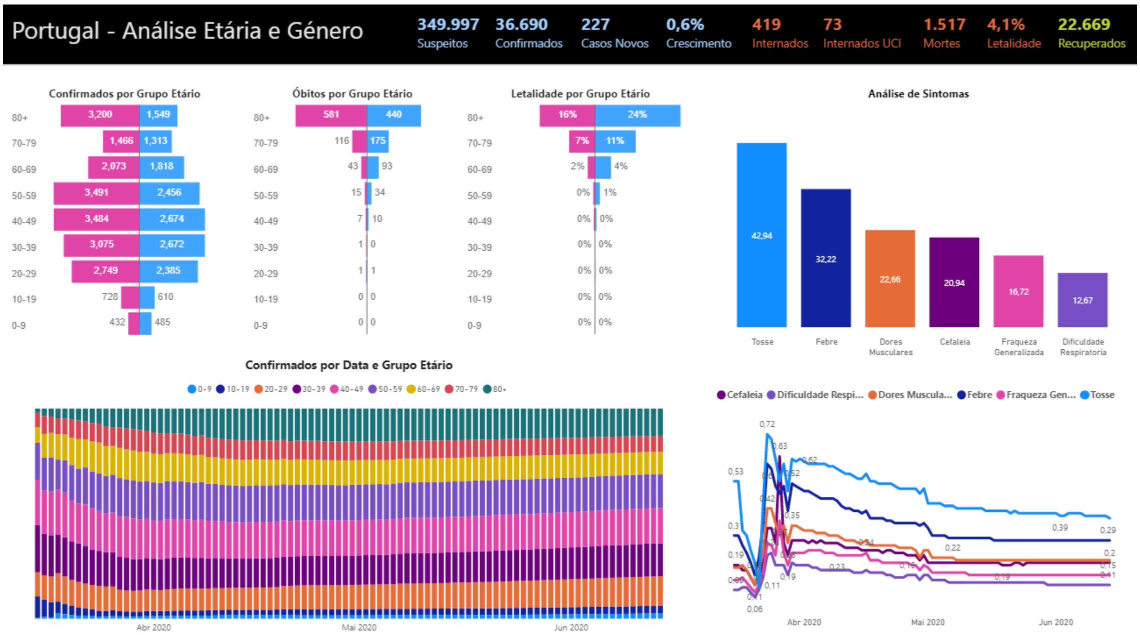


Figure 8- Age and gender analysis

In this dashboard it is possible to find confirmed cases, deaths and lethality per age group, the percentage of symptoms felt, the confirmed cases per date and age group and finally the percentage of symptoms per date.

Finally, in the last dashboard COTEC shows the geographic distribution of cases in Portugal.

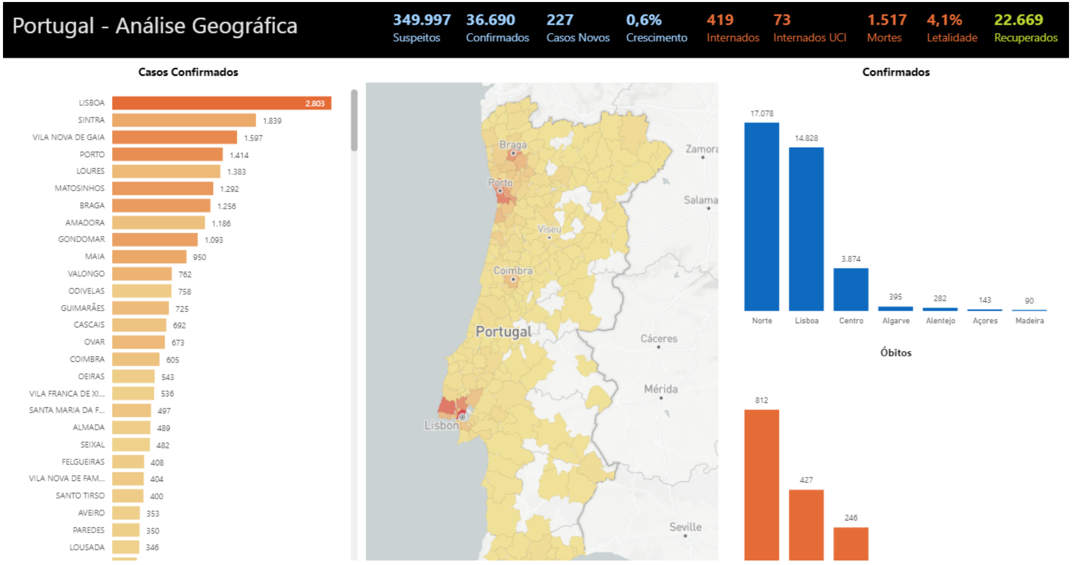


Figure 9- Geographic analysis

In this dashboard we can find the number of cases per city, a Portugal map with graduation colours showing the most affected areas (in dark orange and red), and lastly the confirmed cases and deaths per region.

3.2 Previous studies on the explanation of differences between countries

As this thesis will be an application of Analytics tools to explain differences observed between countries on their mortality and severity cases, in this point we review evidence from previous studies, to better conduct the thesis.

The previous studies are grouped according to two different criteria:

Studies that use econometric methods and prediction algorithms to predict the evolution of curves.

Studies that explain the constraints and differences between results for different countries.

3.2.1 Applying econometric method and prediction algorithms to predict the evolution of curves.

Shojaee et al. (2020) predict the mortality due to COVID-19 for the next month for Italy, Iran and South Korea using their current situation.

The method that the authors use is a simple linear regression model to predict the slope of the regression lines in the upcoming days.

Their analysis was performed from March 15 to April 15 using Rami Krispin dataset where they arranged results for the upcoming 10, 20 and 30 days regarding the current scenario in each country.

Given the situation it is possible to compare the predicted scenarios with the real ones by using the information provided by WHO and their situation reports that are daily published. To do the comparison, for 10 day it was used the situation report of 25 March of WHO, for 20 days it was used the situation report of 4 April and for 30 days it was used the situation report of 15 April.

Table 4 and 5 illustrate the results of Shojaee et al. (2020) and the situation report of WHO for both confirmed cases and death cases.

Table 4- Comparison of total cases between the prediction of Shojaee et al. (2020) and the situation report of WHO

	Total cases predicted (10 days)	Total WHO (10 days)	Total cases predicted (20 days)	Total WHO (20 days)	Total cases predicted (30 days)	Total WHO (30 days)
Italy	61725(357)	69176	114803(865)	119827	183979(1580)	162488
South Korea	11419(98)	9137	14807(221)	10156	18327(379)	10591
Iran	28788(303)	24811	48226(823)	53183	72251(1605)	74877

Table 5- Comparison of total deaths between the prediction of Shojaee, Pourhoseingholi et al. (2020) and the situation report of WHO

	Total deaths predicted (10 days)	Total WHO (10 days)	Total deaths predicted (20 days)	Total WHO (20 days)	Total deaths predicted (30 days)	Total WHO (30 days)
Italy	4840(99)	6820	9249(240)	14681	18395(437)	21069
South Korea	130(13)	126	199(29)	177	283(51)	225
Iran	2215(91)	1934	4562(255)	3294	7770(502)	4683

It is possible to see that using their model, the confirmed cases in South Korea were overestimate for all phases. Despite that, for death cases the prediction was very closed to the reality.

The opposite happened for Italy were they underestimate with huge differences both new and death cases for almost every phases, just overestimating the number of cases, again for huge difference, on the third phase, the 30 days prediction.

For Iran they overestimated death cases in all phases and in the first phase of new cases, and then they underestimate the results for both second and third phases of new cases. These predictions were somehow a little disparate from the reality which shows the difficulty of predicting the situation of a pandemic.

On the contrary, Luo (2020), since April 18, has been predicting the end of COVID-19 for several countries updating predictions daily, with the latest data.

Having in mind that the spread of coronavirus is dependent on several factors, like government measures, individuals' behaviors and even natural limitations of the virus, Luo (2020) decided to adapt his prediction regarding daily changes and new data worldwide.

For predicting the situation, Luo (2020) used SIR (susceptible-infected-recovered) model which incorporates two different parameters which are the responsible for the shape of a specific life cycle curve. The parameters can be regressed based on the available and actual data of a country, where, in this case, the author only reported regressions with satisfactory goodness-of-fit and with statistical significance. The result is the determination of a full pandemic life cycle where the initial side of the curve is the data collected up to the moment, and the rest of the curve is the predicted data.

By analyzing the curve, it is possible to get information on the date to reach the last case, the date to reach 97% of the total expected cases and the date to reach 99% of the total expected cases. In Figure 10 it is shown the result of the prediction for Italy on April 30.

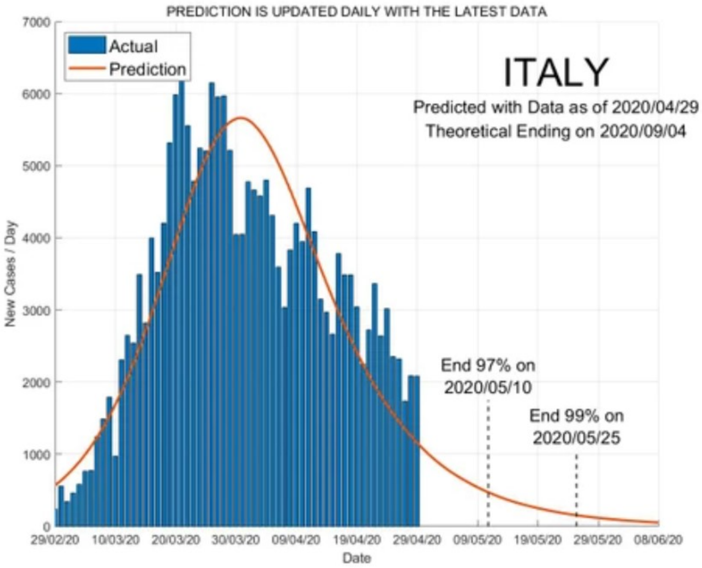


Figure 10- Lou's prediction for Italy on April 30

3.2.2 Explaining the constraints and differences between results for different countries.

When comparing SARS with the situation of COVID19 in China, based on observation of the confirmed cases and deaths, by January 30, 2020 the case numbers of COVID-19 had already surpassed those of SARS and according to Wilder-Smith, Chiew, and Lee (2020) it could be explained for several reasons:

- The epicenter of COVID-19 was in Wuhan, the largest city in central China with more than 11 million people. This city is a center for industry and commerce and has the largest train station, biggest airport and largest deep-water port in central China having triplicated its urban population in the past decade.
- As hospital were overwhelmed, and as a result of lack of beds, many patients weren't hospitalized and contributed to seeding in community.
- Another reason is the infectious period. While in SARS, despite there were asymptomatic patients, the transmission occurred from symptomatic ones who could easily be identified, in the case of COVID-19 the situation is quite different, where transmission during the early phase of illness also contributes to the overall transmission which means that isolation of symptomatic patients might be too late.
- Also, the transmissibility might be higher for COVID-19 than for SARS, where R_0, which is a central concept in infectious disease epidemiology and indicates the risk of an infectious agent with respect to its epidemical potential was found to be 3,28 for COVID-19 while it was 3 for SARS.
- Finally, SARS was mainly an outbreak propagated within hospitals while COVID-19 is already spread in community.

Wilder-Smith, Chiew, and Lee (2020) conclude that the lockdown in China was resulting in a daily decline of new cases since mid-February showing that China was on the right path and that the other countries should be aware and reduce the spread of COVID-19 by implementing early case detection, prompt isolation of ill people, comprehensive contact tracing and immediate quarantine of all contacts.

So, the focus for countries should be the containment of COVID-19 since the short-term costs of containment will be far lower than long-costs of non-containment. This means the closure of public places, institutions and restrictions in travel and trade, but governments should be aware that these measures can't last too much. So, in an initial phase it is important to contain as much as possible the outbreak of COVID-19 which will give time to health systems to scale up and be more prepared to respond to every needs, and will minimize the size of outbreak reducing the peak incidence which can reduce global deaths (Wilder-Smith, Chiew, and Lee 2020).

Majumder et al. (2015), on a sample of 159 patients with MERS in South Korea conclude that towards the middle east respiratory syndrome (MERS) outbreak in South Korea in 2015, infected people with pre-existing concurrent health conditions and with older age were risk factors for death.

Mizumoto et al. (2015), when comprising the outcomes of 185 patients with MERS in South Korea conclude that senior persons aged 60 or over were 9,3-fold more likely to die compared to younger cases and people under treatment were 7,8-fold more likely to die than others.

Fang, Karakiulakis, and Roth (2020) by analysing the results of three studies for patients in Wuhan, noted that diabetes was the most common comorbidity associated with patients in intensive care due to COVID-19. In the first study, Yang, Yu, et al. (2020), observed that in a group of 52 intensive care unit patients with COVID-19, where 32 patients didn't survived, 22% of those had diabetes and another 22% of those had cerebrovascular diseases. In the second study, Guan et al. (2020), observed that in a group of 1099 patients with COVID-19, where 173 of them had severe disease, 23,7% of those had hypertension, 16,2% had diabetes and 5,8% had coronary heart diseases. Finally, in the third study, Zhang et al. (2020) observed that 30% of 140 patients admitted to hospital with COVID-19 had hypertension and 12% of those patients had diabetes.

Shet et al. (2020), based on evidence that Bacille Calmette Guérin (BCG) vaccine has had some sort of protective effects against infections in the past, as well as against tuberculosis, the authors built a simple log-linear regression model to understand the association of BCG use and the mortality of COVID-19 for different countries' economic status (GDP per capita), different countries' proportion of elderly among the population and different status of epidemiological timeline.

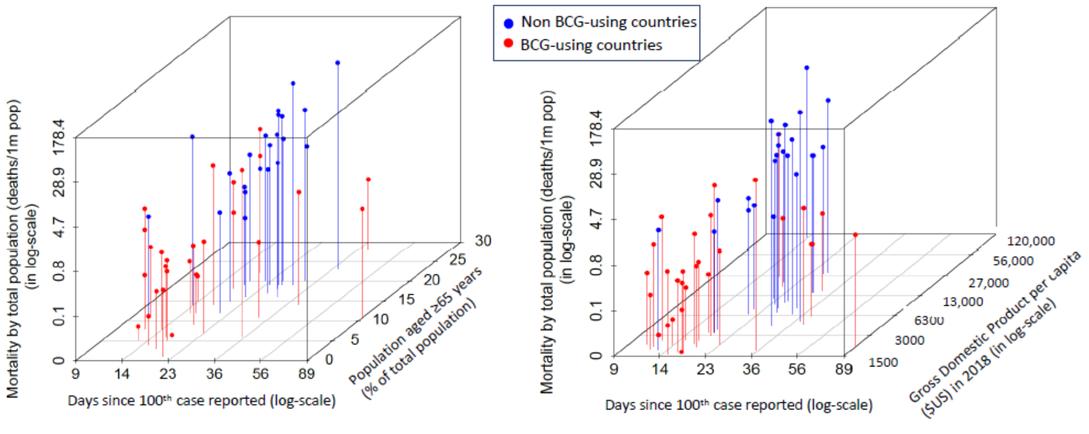


Figure 11 - Taken from Shet et al. (2020)

> – "Association between COVID-19-attributable mortality and BCG use in national immunization schedules, propotion of population aged ≥65years, time trajectory on the epidemiological curve and country-specific GDP per capita. Each dot is representative of a country. Red dots=BCG-using countries; Blue dots=Non BCG-using countries."*(Shet, Ray et al. 2020)*

In the case of GDP per capita , Shet et al. (2020), had counterintuitive results in the analysis of the association between mortality and country economic status since higher mortality rates happened in countries with higher GDP (see Figure 11 - right side).

The opposite happened for Sonego et al. (2015) that have found evidence that the risk of death from acute respiratory illness in children in low-and middle-income countries is higher and for Khaltaev and Axelrod (2019) whose found that low-and middle-income countries have a higher risk of mortality of chronic respiratory diseases due to life style, socio-demographic and economic risk factors.

However, on the left side of Figure 10, regarding age, the authors find age over 65 years to be an important factor. And finally, after all the adjustments the authors conclude that it could be found a significant association between countries using BCG and lower COVID-19 mortality.

When studying the drivers of COVID-19 progression, to a sample of 93 patients in Wuhan, Yang, Liu, et al. (2020) concluded that NLR (Neutrophil-lymphocyte ratio) and age were good predictors to prognosis and evaluate the severity of clinical symptoms in COVID-19 patients, concluding that these factors may be related to the severity of the infection and the outcome of the condition. Logistic regression was applied in this study.

In addition, Al-Najjar and Al-Rousan (2020), collected data between February 20 and March 9, in order to get into a classifier prediction model to predict the status of COVID-19 patients in South Korea.

They used different independent variables such as country, infection reason (transmission way), sex, group, confirmation date, birth year and region, and, the dependent variable was one of the following, death or recovered.

The authors split the data into a training and a test data sets, and then used the training set to build one hidden layer neural network classifier for predicting death and recovered cases. The great accuracy of both training and testing data showed that the proposed models for both death and recovered cases had the ability to classify the death and recovered cases based on the variables selected.

By analyzing the results it was found that the main variables for death cases were infection reason, confirmation date and region while for recovered cases the main effective variables were region, birth year and confirmation date (Al-Najjar and Al-Rousan 2020).The authors conclude that by choosing the most effective categorical variables and numerical variables could enhance the prediction model but if they had some variables with less importance, they could stabilize a neural network predictor and avoid overfitting which would result on an improved prediction output.

4 Methodology

4.1Case study purpose

The purpose of this thesis is to find evidence to explain mortality by COVID-19 and the factors that affect it globally. Having the knowledge that each country has its specific characteristics, the purpose of this thesis is to answer these questions:

- Does economic structure affect the mortality by COVID-19 of the country?
- Do healthcare resources of each country affect its COVID-19 mortality?
- Is there any comorbidity associated with mortality by COVID-19?
- Does government response time affect mortality by COVID-19?
- Does the “stay at home” measure affects mortality?

4.2 Model

COV19 mortality rates will be explained by panel data regression models since they model information on individual behaviour, both across individuals and over time.

The package used in R to estimate the model was plm, a package that automates some of basic data management tasks as lagging, summing and others, (Croissant and Millo 2008).

The model can be described as following:

(1) $Y_{c,w} = \beta_0 + \sum_{j=1}^{p} \beta_j \, X_{j,c,w-l_j} + \sum_{j=p+1}^{q} \beta_j \, Z_{j,c} + u_{c,w}$

(2) $u_{c,w} = \mu_c + v_{c,w}$

Table 6- The one-way error component model adapted from Clower, (2019)

β_0	Parameter which measures an intercept that is constant across all coutries and weeks.
β_j	Parameter that measures the impact of the time-varying variable Xj, or the time-invariante variable Z_j, on Y. It is constant across all countries and weeks.
μ_c	Country-specific variation in Y which stays constant across time for each country. In the random effects model this follows a random distribution with parameters that must be estimated.
$v_{c,w}$	Usual stochastic regression disturbance which varies across weeks and countries.

Where c is the individual, in this case, the *country*, w(1,...,T) is the time index, in this case, *week_1st* and $u_{c,w}$ is a random disturbance term of mean 0.

Regarding time index, *week_1st* , this variable was considered in order to estimate the model to be on an equal footing for all countries. Thus, instead of the model being estimated with the effective weeks (from 1 to 22 weeks from the first case in the world in December 30, 2019, until the end of week 22 in May 31, 2020), for each country week 0 equals the week in which there was the first death by COVID-19 in that country, and the other weeks were counted from then on. In addition, since to explain deaths in week 0 we needed to use past data, we considered 6 weeks of data before week 0 for all countries, and as a result our data set starts in week -6 for all countries considered.

To better explain this variable let's look at the situation of Portugal in table 7:

Table 7- Variable week_1st for Portugal

Country	week_1st	week	cases	deaths
Portugal	na	0	0	0
Portugal	na	1	0	0
Portugal	na	2	0	0
Portugal	na	3	0	0
Portugal	na	4	0	0
Portugal	na	5	0	0
Portugal	-6	6	0	0
Portugal	-5	7	0	0
Portugal	-4	8	0	0
Portugal	-3	9	0	0
Portugal	-2	10	21	0
Portugal	-1	11	148	0
Portugal	0	12	1111	12
Portugal	1	13	3890	88
Portugal	2	14	5354	166
Portugal	3	15	5463	204
Portugal	4	16	3698	217
Portugal	5	17	3998	216
Portugal	6	18	1507	120
Portugal	7	19	2216	103
Portugal	8	20	1404	77
Portugal	9	21	1661	99
Portugal	10	22	1732	94

In Table 7, the week that corresponds to 0 in the variable week_1st corresponds to week 12 of COVID-19 worldwide. If we used the variable week for all countries and not the variable week_1st we would be comparing countries in completely unequal states of disease.

Panel data models can be estimated using fixed effects or random effects. In this case it was used random effects since is the only one which considers the effect of constant variables which are many in this study. Furthermore, using random effects models the country effect (u_c) is modelled by a normal distribution where $u_{c,w}$ are not correlated with the explanatory variables, which will be verified by the Hausman test.

4.2 Collected variables

Taking into account that in a pandemic situation, especially in the case of COVID-19 and its rapid spread, there are many factors that seem to affect the country's reaction to the situation and its mortality rate, for the construction of the model for the explanation of mortality rates by COVID-19, were collected not only the data of COVID-19 tracking like *cases, deaths, tests and hospitalizations* but also data regarding demographic characteristics of the country, health system capacity, country structure and other variables mentioned in other studies as possible causes of the evolution or mortality rate of COVID-19 in certain countries.

Variables were collected according to at least one of the following criteria:

(1) Data regarding the current situation of COVID-19 ;

(2) Socio-economic and public structure of the country;

(3) Demographic characteristics of the country;

(4) Changes in mobility as a result of isolation caused by COVID-19;

(5) Variables pointed out in other studies as influencing the country's performance against the pandemic, COVID-19.

So, the variables collected are these indicated in table 8 with their respective source.

Table 8- Variables and their sources

Variable code	Variable description	Source
cases	Number of weekly COVID-19 cases	European Centre for Disease Prevention and Control
deaths	Number of weekly COVID-19 deaths	European Centre for Disease Prevention and Control
tests	Number of weekly COVID-19 tests performed	Institute for Health Metrics and Evaluation

admin	Number of patients weekly admitted to hospital with COVID-19 diagnosis	Institute for Health Metrics and Evaluation
bed capacity	Number of hospital beds	Institute for Health Metrics and Evaluation
uci_capacity	Number of UCI beds	Institute for Health Metrics and Evaluation
uci_bed	Number of patients weekly in intensive care units with COVID-19	Institute for Health Metrics and Evaluation
Days since the first case until lockdown	Number of days that each government took to decree a national lockdown after the first case in its country.	BBC News
retail	Variation in percentage of mobility in retail and leisure	Cotec
grocery	Variation in percentage of mobility in supermarkets and pharmacies	Cotec
parks	Variation in percentage of mobility in parks	Cotec
transit	Variation in percentage of mobility in public transport	Cotec
workplace	Variation in percentage of mobility in workplaces	Cotec
residential	Variation in percentage of mobility in residential areas.	Cotec
Diabetes_prevalence	Percentage of population with diabetes.	Our World in Data
Percentage older 65	Percentage of population aged 65 or over.	Our World in Data
Percentage older 70	Percentage of population aged 70 or over.	Our World in Data
population	Population	Our World in Data
population density	Population density	Our World in Data

Percentage RD LRI deaths	Percentage of deaths from respiratory diseases and low respiratory infections	Our World in Data
GDP per capita	Gross Domestic Product per capita	World Bank Population Data
inv.healthcare in percentage of GDP	Investment in healthcare per percentage of GDP	OECD
inv.healthcare per capita	Investment in healthcare per capita	OECD
physicians per 1000 people	Number of physicians per 1000 people.	Our World in Data
nurses per 1000 people	Number of nurses per 1000 people.	Our World in Data
female smokers	Percentage of female smokers	Our World in Data
male smokers	Percentage of male smokers.	Our World in Data
BCG	Number of years each country stopped having a mandatory national BCG vaccination policy for all	BCG World Atlas

The data was collected for the period ranging from 30 December 2019 until 31 of May, 2020. In this period some variables remain constant within country, while others change as time passes. In the latter case, the data was aggregated by weekly sums.

Looking at the criteria mentioned above, variables in table 6 can be grouped as:

(1) Data regarding the current situation of COVID-19 are captured through the variables cases, deaths, tests, admin and uci_bed.

(2) Socio-economic and public structure of the country is shown by the variables bed_capacity, uci_capacity, Physicians per 1000 people, Nurses per 1000 people, GDP, inv healthcare %GDP and inv healthcare per capita.

Regarding bed_capacity and uci_capacity it is important to note that even taking into account that the response of some countries to the pandemic was to increase the available bed capacity in both hospitals and ICUs, the data collected for the model are only corresponding to the pre-pandemic period so these variations are not accounted for.

(3) Demographic characteristics of the country can be found in variables

Percentage older 65, percentage older 70, diabetes_prevalence, percentage of RD LRI deaths, female smokers, male smokers, population and population density.

(4) Changes in mobility as a result of isolation caused by COVID-19 are the variables retail, grocery, parks, transit, workplace and residential.

(5) Variables pointed out in other studies as influencing the country's performance against the pandemic, COVID-19 are captured trough the variables Days since the first case until lockdown and BCG. Although they are already grouped in demographic characteristics, the main criterion for the insertion of the variables percentage older 65, percentage older 70 and diabetes_prevalence was the fact that they were mentioned in other studies.

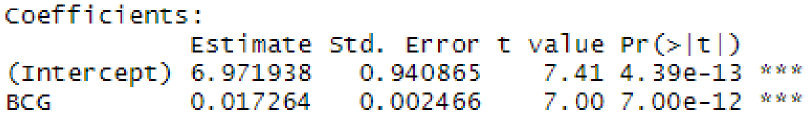
Regarding variable *Days since the first case until lockdown*, as mentioned in table 6 it determines the number of days that each government took to decree a national lockdown after the first case in its country but it is important to take note that in the case of the countries where the lockdown was not decreed nationally, it was assumed the value of 154 days corresponding to the total of days in the sample of 22 weeks. The same for BCG, where in countries where there has never been a national policy for all, it was assumed a value of 100 years.

Furthermore, regarding BCG, before taking this variable into consideration it was done a linear regression to test the effect of the vaccine in total deaths per COVID-19.

The regression was the following:

$deaths_c = \beta_0 + \beta_1 BCG_c + u_c$

Where c represents the country.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.971938   0.940865    7.41 4.39e-13 ***
BCG         0.017264   0.002466    7.00 7.00e-12 ***
```

The results showed that BCG was a very significative variable explaining mortality for COVID-19, where an additional year of a country without policy of vaccination for all people implies an expected increase of 0,017 deaths per million in habitants. The R-squared of this regression is 10%, which indicates that 10% of mortality can be explained by BCG vaccination.

4.3 Sampled Countries

The countries used for the sample were selected by 2 criteria.

In a first stage only countries that by the date of 10 May 2020 had more than 3000 cases were collected to make it a significant sample.

Considering this criterion, 66 countries were selected, and after this first selection, the data mentioned above for each of the 66 countries was searched. Since it was not possible to collect the test variable for the 66 countries, and it is a very important variable for the model, the countries where it wasn't possible to collect the tests were eliminated, leaving 37 countries.

On table 9 we can find on the left side the countries selected for the model, and on the right side, the countries discharged due to unavailability of tests' data.

Table 9- Countries Selection

Selected Country	Discharged country
USA	Russia
Spain	Iran
Italy	China
UK	India
Germany	Saudi Arabia
Brazil	Pakistan
France	Singapore
Turkey	Belarus
Canada	Qatar
Peru	United-Arab Emirates
Belgium	Japan
Netherlands	Ukraine
Mexico	Bangladesh
Switzerland	Indonesia
Ecuador	Serbia
Portugal	South Africa
Chile	Kuwait
Sweden	Australia
Ireland	Morocco
Israel	Algeria
Austria	Kazakhstan
Poland	Bahrain
Romania	Ghana
Republic of Korea	Nigeria
Philippines	Luxembourg
Colombia	Afghanistan
Denmark	Oman
Dominican Republic	Armenia
Egypt	Thailand
Panama	
Czechia	
Norway	
Malaysia	
Finland	
Argentina	
Republic of Moldova	
Hungary	

4.4 Descriptive Statistics

In table 10 we present the descriptive statistics for the used variables.

Table 10- Descriptive statistic of variables

Variable	Min	Max	Mean	Median	Standard deviation
cases	0	217714	6925	697,50	24405,61
deaths	0	18302	500	22,5	1662,53
tests	0	3029158	89030	18871	261776
admin	0	57206	2237	143	6954,84
uci_bed	0	18357	669	31	2071,82
bed capacity	8663	938974	152641	71296	201403,2
uci_capacity	303	70771	7284	2096	13485,17
Days since the first case until lockdown	9	154	52	52	63,18
retail	-91,43	17	-33,99	-31,72	31,63
grocery	-84,52	20	-13,96	-5,71	20,2
parks	-92,57	144,21	-8,92	0,00	40,83
transit	-87,43	17	-34,62	-39,57	30,37
workplace	-80,66	14,29	-27,99	-30,31	25,5
residential	-2,29	41,71	12,43	11,31	11,51
Diabetes_prevalence	3,28	17,31	7,43	6,80	3,08
Percentage older 65	4,8	23,02	14,39	16,76	5,63
Percentage older 70	2,66	16,24	9,39	10,20	4,05
population	4033963	331002647	49880306	32365998	65681923
population density	4,08	529,65	148,54	106,960	139,83
Percentage RD LRI deaths	4,46	17,84	9,62	9,15	3,3
GDP	1.144e+10	2.054e+13	1.445e+12	4.553e+11	3.483427e+12
inv.healthcare in percentage of GDP	3,9	16,94	8,45	8,86	2,69
inv.helathcare per capita	131	10586	3120	2780	2455,3
physicians per 1000 people	0,8	5,4	2,95	3,1	1,17
nurses per 1000 people	0,2	18,1	6,99	6,4	4,79

female smokers	0,2	34,2	16,65	18,80	9,33
male smokers	9,9	50,1	28,84	28,90	9,98
BCG	0	100	21,95	0,00	34,35

For variables *cases, deaths, tests, admin* and *uci_bed,* their minimum value is 0 once the count started at 30.12.2020, after the first known case in China, where none of the selected countries was infected by COVID-19. In addition, as countries were affected at different times, while in some countries there were already several numbers for these variables, in others the number remained at 0.

Regarding the variables above, the maximum value reached in a week was always by the United States of America where the maximum for cases was at week 15, for both deaths and hospitalizations was at week 16, tests was at week 22 and hospitalizations in intensive care units was at week 17.

Looking at the standard deviation for each of the variables, it is possible to notice that all of them are quite high which can be explained by the discrepancy of countries both in demographic and socio-economic terms.

The variables related to mobility also have a large standard deviation evidencing the changes in mobility, compared to when the country was not affected by COVID-19 and when it becomes.

Looking at the median, the variables where the difference between the mean and the median is starker are those that concern the capacity of beds in both hospitals and an ICU's, which indicates the presence of outliers like the USA that have an enormous capacity of beds.

The variable Days since the first case until lockdown has a minimum of 9 days, which corresponds to the days that Peru took to enter national lockdown after the first case. The maximum value (154) was an assumed value for the study for countries that did not have lockdown and was calculated by (22*7) corresponding to the duration of the study, 22 weeks multiplying by 7 days. In this case the average is very biased by the countries that have not practiced national lockdown (United States of America, Turkey, Brazil, Egypt, Republic of Korea, Mexico, Canada, Philippines, Hungary, Sweden, Chile, Dominican Republic, Norway, Israel, Finland, Republic of Moldova). Therefore, looking only at the countries that practiced national lockdown, the average of days that they took to take the measure was 28 days.

Finally, the variable BCG that has a minimum value of 0 for countries that maintain the national vaccination policy (Egypt, Republic of Moldova, Malaysia, Chile, Turkey, Republic of Korea, Philippines, Romania, Hungary, Poland, Portugal, Argentina, Ireland, Mexico, Dominican Republic, Brazil, Colombia, Peru and Panama) and a maximum value of 100 for countries that have never had a national vaccination policy (Belgium, Italy Netherlands, United States of America and Canada). Finally, countries that have had a national vaccination policy but have stopped having it for a few years where the minimum of 10 years is in Czechia and the maximum is 45 in Sweden. As the average considering the values 0 and 100 would be biased (21.95) was calculated the average of years considering only the countries that had national vaccination but no longer have, and the same is 23.85 years without vaccination of BCG.

4.5 Choosing the right lag

The first step before estimating the model was choosing the right lag to use for the estimation. The lag will be used on explanatory variables to better understand their effect on the final output, in this case, deaths.

For example, it is not expected that the cases confirmed in one week, will affect the deaths of that current week, so, variable-specific lags were chosen by trial and error trying to maximize the goodness of fit for the selected model.

4.6 Multicollinearity

When trying to apply the variable *population* to the model, we faced a problem of multicollinearity since it had linear relations with variables like *cases* and *tests*.

To solve this problem, it was decided to change the dependent variable and all variables that were related with the population.

So, instead of having a dependent variable *deaths* it as changed for *deathspm (=(deaths/population)*1000000))*meaning the total deaths per million in habitants.

The same calculation happened for the following variables:

casespm=(cases/population)*1000000

testspm=(tests/population) *1000000

adminpm=(admin/population) *1000000

beds_usagepm=(beds_usage/population) *1000000

bedcappm=(bed.capacity/population) *1000000

ucicappm=(uci_capacity/population) *1000000

4.7 Variables selection process

To estimate the model, first all the variables were considered, and despite the R-Squared was 98%, most of the independent variables were meaningless. However, when analyzing carefully the variables it was possible to find a reason for the insignificance of some of them, which will be enumerated.

4.7.1 Population Density

The reason behind choosing variable population density in the model was the thought that when dealing with an infectious disease, countries with more aggregation of people would tend to have more risk of transmission and so, more cases and more deaths. What happened is that the variable didn't have any influence in the model, which can be explained by the fact that population density doesn't consider large urban centers. For example, the population density of Portugal, doesn't explain the difference between the aggregation of people in the centre of Lisbon and Aljustrel. Another example is Egypt that the population density is underestimated by the existence of deserts.

4.7.2 Percentage of RD LRI deaths

This variable was considered to understand whether countries that already have high percentages of deaths from RD and LRI respond better or worse to this pandemic.

This variable doesn't have significance to the model and when analysing it with more attention it doesn't allow to take any further conclusions.

Looking for the relation between percentage of deaths in every country for respiratory diseases and low respiratory infections and total deaths, total cases, investment in healthcare, age of the population and healthcare resources in every country, this variable doesn't show any evidence of relation with any of the variables.

It might be associated with country's specific characteristics, but it doesn't allows any interpretation for the specific case of deaths per capita of the pandemic COVID-19.

4.7.3 Bed_capacity, uci_capacity, Nurses and Physicians per 1000 people

Healthcare resources doesn't have significance in any model, and it can be justified in the same way it was for population density. For the case of Italy, for example, the capacity and resources available for all country doesn't represent the lack of resources that were felt in Lombardia due to an enormous amount of COVID-19 hospitalizations.

Anyway, to better corroborate the exclusion of these variables, there were created another two, to calculate hospitals and ICU's occupation rate.

The variables were as following:

hospital_occupation=admin/bed.capacity,

icu_ocupation= uci_bed/uci_capacity

These variables not only came to be insignificant for the model but also showed that for the countries selected, the national mean occupation of hospitals regarding COVID-19 patients was 1% and the national mean occupation of ICU's regarding COVID-19 patients was 12% which can't explain any impact in the mortality since the national resources were always more than enough.

These averages for occupation might be overestimated since changes in countries' capacity of hospitals and ICU's to better respond to COVID-19 pandemic weren't considered.

4.7.4 Female smokers and male smokers and country economic structure

There wasn't any result showing evidence of variable smokers in the output of COVID-19 per country. It might be the result of having the percentage of male smokers and female smokers per country but not the data regarding the percentage of men and women per country.

For the economic structure of the country, nothing seems to have significance both in hospitalizations and in deaths per COVID-19. For example, the case of United States of America, that are by far the country with greatest GDP and investment in healthcare both per capita and per percentage of GDP and have a lot of hospitalizations and their ratio deaths/cases is way far from being the best.

5. Results

After excluding all the variables above, the next step was estimating the model.

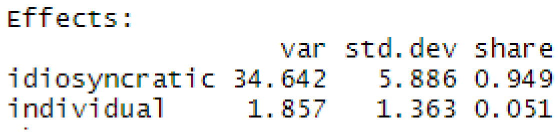
Deaths per million

The model used for estimation is shown below where we can find variable *testspm* with a lag of 2 week, *casespm, adminpm* and *residential* with a lag of one week and finally, *uci_beppm* with a lag of 0 weeks. The other variables are constant so no lag was applied.

(3) $deathspm_{c,w} = \beta_0 + \beta_1 tests_{c,w-2} + \beta_2 casespm_{c,w-1} + \beta_3 adminpm_{c,w-1} +$
$\beta_4 uci\_bedpm_{c,w} + \beta_5 X..older.70\ _c + \beta_6 Diabetes\_prevalence\ _c +$
$\beta_7 residential_{c,w-1} + u_{c,w}$

(4) $u_{c,w} = \mu_c + v_{c,w}$

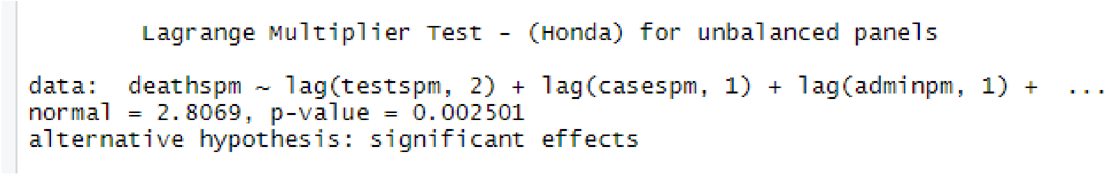
The results were as following:



```
Effects:
                  var std.dev share
idiosyncratic 34.642   5.886 0.949
individual     1.857   1.363 0.051
```

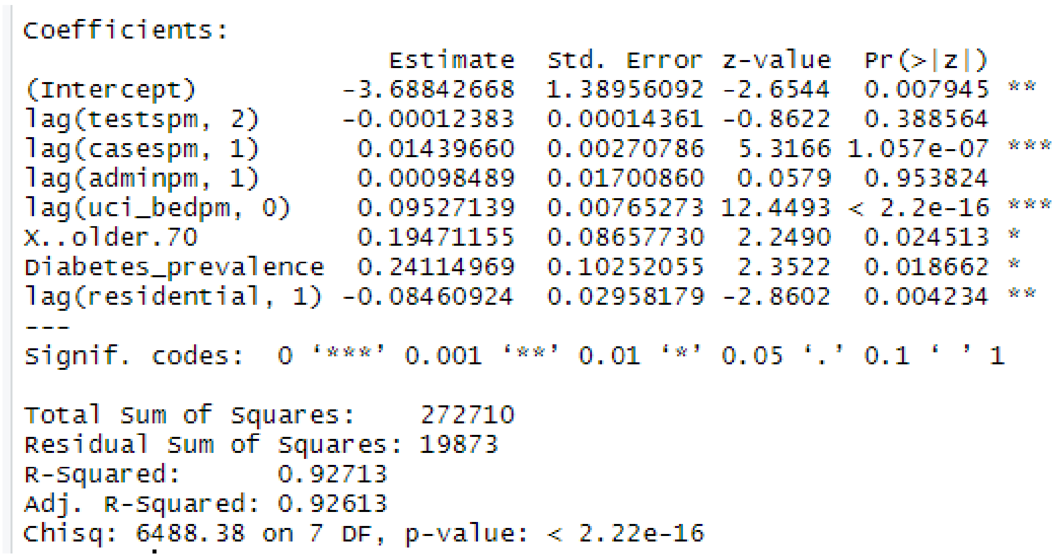
Where we can see that the total variance is 36,5 but the individual variance (country effect) is only 1,86. It shows that there is evidence of variability between countries, although it is in a very small percentage, 5,1% (last column).

So, the country effect exists but it barely affects error variance.

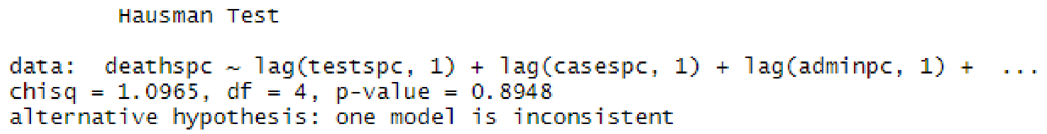


```
        Lagrange Multiplier Test - (Honda) for unbalanced panels

data:  deathspm ~ lag(testspm, 2) + lag(casespm, 1) + lag(adminpm, 1) +  ...
normal = 2.8069, p-value = 0.002501
alternative hypothesis: significant effects
```

Applying Lagrange Multiplier Test, we could confirm that effectively, there is variability between countries, proving the country effect.



```
Coefficients:
                       Estimate  Std. Error z-value  Pr(>|z|)
(Intercept)         -3.68842668  1.38956092 -2.6544  0.007945 **
lag(testspm, 2)     -0.00012383  0.00014361 -0.8622  0.388564
lag(casespm, 1)      0.01439660  0.00270786  5.3166 1.057e-07 ***
lag(adminpm, 1)      0.00098489  0.01700860  0.0579  0.953824
lag(uci_bedpm, 0)    0.09527139  0.00765273 12.4493 < 2.2e-16 ***
X..older.70          0.19471155  0.08657730  2.2490  0.024513 *
Diabetes_prevalence  0.24114969  0.10252055  2.3522  0.018662 *
lag(residential, 1) -0.08460924  0.02958179 -2.8602  0.004234 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    272710
Residual Sum of Squares: 19873
R-Squared:      0.92713
Adj. R-Squared: 0.92613
Chisq: 6488.38 on 7 DF, p-value: < 2.22e-16
```



```
        Hausman Test

data:  deathspc ~ lag(testspc, 1) + lag(casespc, 1) + lag(adminpc, 1) +  ...
chisq = 1.0965, df = 4, p-value = 0.8948
alternative hypothesis: one model is inconsistent
```

Applying Hausman test, we accepted the null hypothesis (the hypothesis of non-correlated errors), and the assumptions of the random effects model seem to be tenable for this data.

Testspm (tests per million) and adminpm (hospitalizations per million) don’t seem to have significance in the model. However, as there is a very strong evidence of the significance of these variables to the mortality, it was decided to keep them and analyze their coefficients.

Looking at the coefficients, in the cases of testspm we can observe that for one additional unit of tests per million, reduces mortality in 0,0001 per million two weeks after. Regarding casespm, one additional unit of cases per million, increases deaths per million in 0,014 one week later. In the case of hospitalizations per million (adminpm), an additional unit of hospitalizations per million will increase the mortality per million in 0,0009 one week later.

Hospitalizations in ICU, a very significant variable seems to strongly affect mortality, where in this case, an additional unit of hospitalizations in ICU per million (uci_bedpm) generates an increase of death per million in 0,095.

Regarding elderly population, when the percentage of people with age above 70 increases 1%, the mortality per million increases in 0,194. The same for diabetes, when the percentage of diabetes prevalence increases 1%, deaths per million increase 0,241.

Finally, the variable residential shows that a positive variation of 1% in the mobility on residential zones decreases deaths per million in 0,085 one week later. This model has a R-squared of 0,927 which means that 92,7% of COVID-19 deaths per million can be explained by this model.

Therefore, as expected the more people are tested, the more people are identified and in turn, isolated, reducing the risk of transmission. Thus, testing the population as much as possible has shown to be a solution to contain the transmission of COVID-19 and thus reduce the number of cases which will reduce hospitalizations both in hospital and ICU which are strongly correlated with deaths by COVID-19.

In addition, age above 70 also revealed to be a contributing factor to mortality, something that would be expected to the extent that the symptoms caused by COVID-19 are mostly difficulty in breathing, which is a factor that is usually already a major obstacle for older people.

The variable residential shows that effectively, confinement is a very important measure, since, when the greater the mobility in the residential area, the lower the mortality by COVID-19. Since there is a general belief that the virus can be transmitted by asymptomatic people, and with the impossibility to test everyone, the more contact between people is avoided, the lower the risk of transmission for all.

Intensive care units hospitalizations

Looking again at the results of uci_bedpm, it can be interpreted that one out of ten people hospitalized in ICU is expected to die from COVID-19. So, we can interpret how hospitalized people in ICU affect mortality, but how is the hospitalization in ICU affected?

To better answer this question, it was estimated another model where the dependent variable was hospitalizations per million in ICU (uci_bedpm).

The model was estimated using the same variables as the model before but all the other variables were again tested.

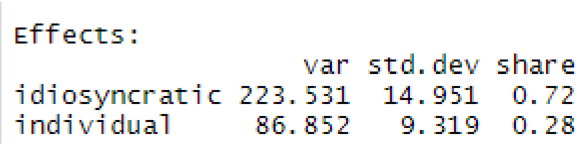
Intensive care units hospitalizations per million

(5) $uci\_bedpm_{c,w} = \beta_0 + \beta_1 tests_{c,w-2} + \beta_2 casespm_{c,w-1} + \beta_3 BCG\ _c\ +$

$\beta_4 X..older.70\ _c ++\ \beta_5 residential_{c,w-1} +\ u_c$

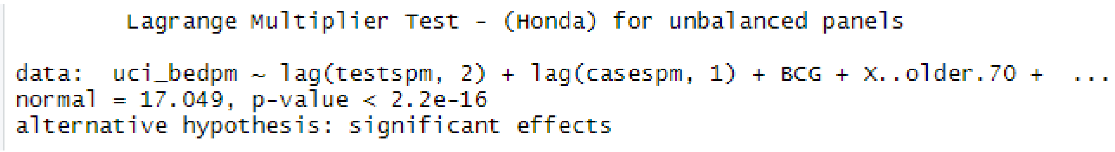
(6) $u_{c,w} = \mu_c + v_{c,w}$

And the results were:



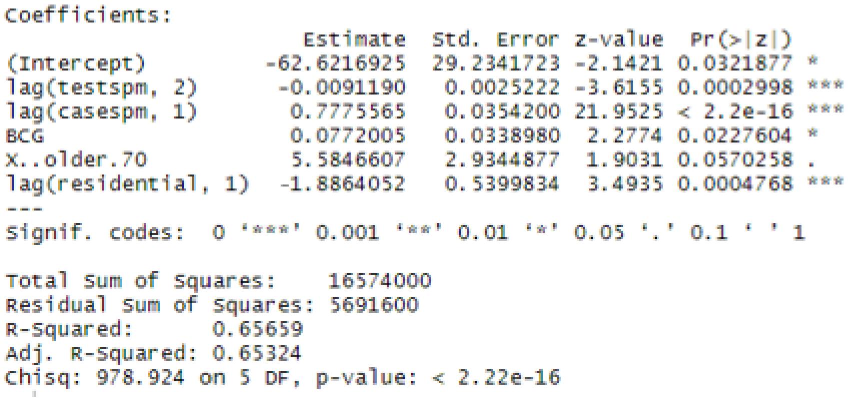
```
Effects:
                  var std.dev share
idiosyncratic 223.531  14.951  0.72
individual     86.852   9.319  0.28
```

Where we can see that the total variance is 310,38 and the individual variance is 86,85. It shows that there is evidence of variability between countries, where 28% of the error's variance are affected by the country effect. Interesting that variance between countries is much bigger and that the country has an effect here. This reveals that the country is not a very important factor in explaining deaths, but it is very important in explaining ICU hospitalizations, which may reveal different treatment practices between countries – some more effective than others.

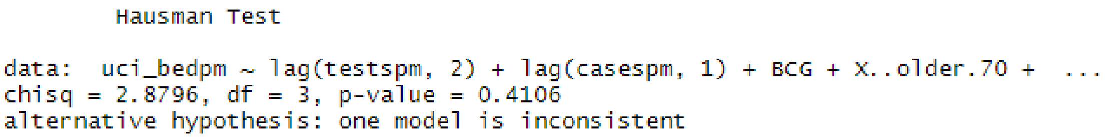


```
        Lagrange Multiplier Test - (Honda) for unbalanced panels

data:  uci_bedpm ~ lag(testspm, 2) + lag(casespm, 1) + BCG + X..older.70 +  ...
normal = 17.049, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Applying once again the Lagrange Multiplier Test we could prove the country effect that shows some variability between countries.



```
Coefficients:
                        Estimate  Std. Error z-value  Pr(>|z|)    
(Intercept)          -62.6216925  29.2341723 -2.1421 0.0321877 *  
lag(testspm, 2)       -0.0091190   0.0025222 -3.6155 0.0002998 ***
lag(casespm, 1)        0.7775565   0.0354200 21.9525 < 2.2e-16 ***
BCG                    0.0772005   0.0338980  2.2774 0.0227604 *  
X..older.70            5.5846607   2.9344877  1.9031 0.0570258 .  
lag(residential, 1)   -1.8864052   0.5399834  3.4935 0.0004768 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    16574000
Residual Sum of Squares: 5691600
R-Squared:      0.65659
Adj. R-Squared: 0.65324
Chisq: 978.924 on 5 DF, p-value: < 2.22e-16
```



```
        Hausman Test

data:  uci_bedpm ~ lag(testspm, 2) + lag(casespm, 1) + BCG + X..older.70 +  ...
chisq = 2.8796, df = 3, p-value = 0.4106
alternative hypothesis: one model is inconsistent
```

With Hausman Test we could again accept the null hypothesis and use the estimation of random effects panel models.

All the variables excluded from the previous model were again excluded in this one except the variable BCG.

Regarding tests per million, an additional unit of tests per million, reduces hospitalizations per million in intensive care units in 0,0091 two weeks later.

Variable cases per million, as in the previous model affects positively hospitalizations in ICU, where an additional unit of cases per million increases 0,776 hospitalizations per million in ICU one week later.

Despite the previous model, BCG vaccination has significance to hospitalizations in ICU where an additional year of a country without policy of vaccination for all people increases 0,073 hospitalizations per million in ICU.

As expected, age affects hospitalizations in ICU, where an additional 1% of people above 70 years in a country, might increase 6,76 hospitalizations per million in ICU.

The variable residential shows that a positive variation of 1% in the mobility on residential zones decreases hospitalizations per million in ICU in 1,87 one week later. This model has a R-squared of 0,657 which means that 65,7% of COVID-19 hospitalizations per million in ICU can be explained by this model.

Finally, it was estimated the same model for hospitalizations per million to understand if there is any variable that differentiate hospitalizations and hospitalizations in ICU.

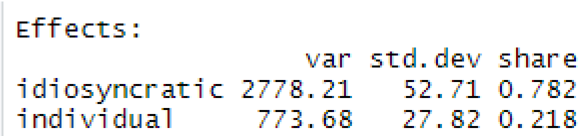
Hospitalizations per million

(7) $adminpm_{c,w} = \beta_0 + \beta_1 tests_{c,w-2} + \beta_2 casespm_{c,w-1} + \beta_3 BCG_{\ c} +$

$\beta_4 X..older.70_{\ c} + \beta_5 residential_{c,w-1} + u_c$

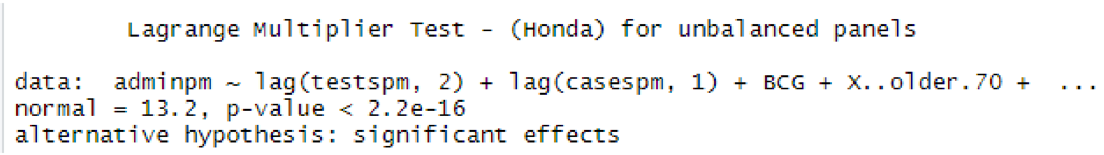
(8) $u_{c,w} = \mu_c + v_{c,w}$

And the results were as following:



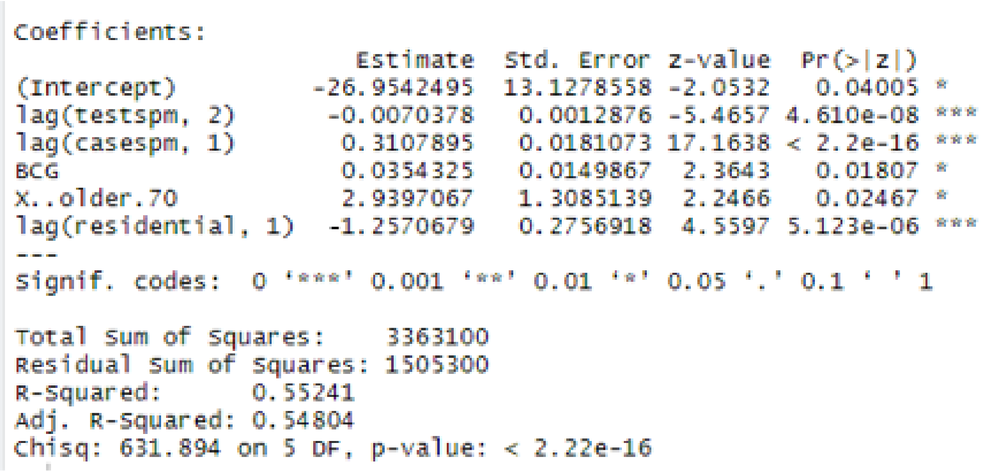
```
Effects:
                  var std.dev share
idiosyncratic 2778.21   52.71 0.782
individual     773.68   27.82 0.218
```

Where we can see that the total variance is 3551,89 and the individual variance is 773,68. It shows that there is evidence of variability between countries, where 21,8% of the error's variance are affected by the country effect.

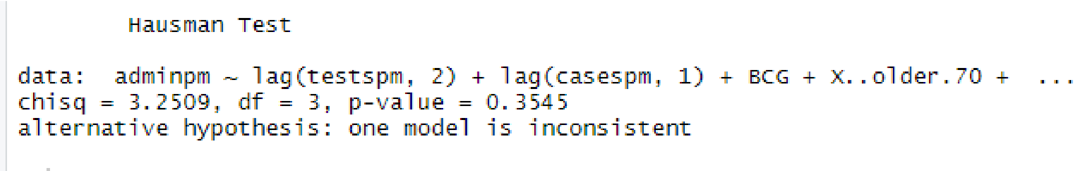


```
        Lagrange Multiplier Test - (Honda) for unbalanced panels

data:  adminpm ~ lag(testspm, 2) + lag(casespm, 1) + BCG + X..older.70 +  ...
normal = 13.2, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Finally, in this model we could again observe the existence of variability between countries and prove the country effect.



```
Coefficients:
                        Estimate  Std. Error z-value  Pr(>|z|)    
(Intercept)          -26.9542495  13.1278558 -2.0532   0.04005 *  
lag(testspm, 2)       -0.0070378   0.0012876 -5.4657 4.610e-08 ***
lag(casespm, 1)        0.3107895   0.0181073 17.1638 < 2.2e-16 ***
BCG                    0.0354325   0.0149867  2.3643   0.01807 *  
X..older.70            2.9397067   1.3085139  2.2466   0.02467 *  
lag(residential, 1)   -1.2570679   0.2756918  4.5597 5.123e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    3363100
Residual Sum of Squares: 1505300
R-Squared:      0.55241
Adj. R-Squared: 0.54804
Chisq: 631.894 on 5 DF, p-value: < 2.22e-16
```



```
        Hausman Test

data:  adminpm ~ lag(testspm, 2) + lag(casespm, 1) + BCG + X..older.70 +  ...
chisq = 3.2509, df = 3, p-value = 0.3545
alternative hypothesis: one model is inconsistent
```

The estimation of random effects panel models was again used since we could accept the null hypothesis of the Hausman Test and so, accept the assumptions of random effects panel models.

The results were very similar to the model estimated for hospitalizations in ICU where all other variables remain insignificant and BCG continues to affect positively hospitalizations per million. This model has a R-squared of 0,553 which means that 55,3% of COVID-19 hospitalizations per million can be explained by this model.

Diabetes prevalence

This variable seems to affect mortality but when introduced on models to explain the two kind of hospitalizations it doesn’t have any significance.

In fact, it makes some sense, because diabetics don’t have higher probability of getting infected with COVID-19, but, once they get, their outcome is more likely to be the worst of the scenarios.

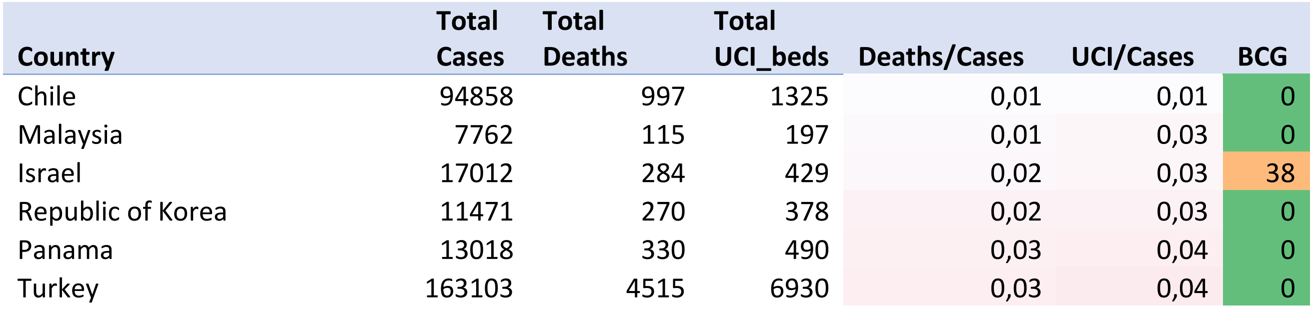
In fact, in a recent publication in MedScape, Tucker (2020) conclude that 10% of diabetics hospitalized for COVID-19, die within a week.

BCG

The variable BCG vaccination didn’t have any significance when trying to explain deaths per million of COVID-19, nevertheless, it helped explaining both normal hospitalizations of patients COVID-19 and hospitalizations in ICU of patients with COVID-19.

It was decided to analyze some evidence between BCG vaccination and the ratios deaths/cases and uci_bed/cases because of the evidence mentioned earlier in other studies that the BCG vaccination seems to have protective effects against respiratory infections, mitigating the symptoms and reducing mortality. And by looking at table 12 we can see that the worst ratios are mainly in countries that never had a policy vaccination for all or have left that policy some years ago.

Table 11-Relation between deaths/cases, ICU/cases and BCG vaccination



Country	Total Cases	Total Deaths	Total UCI_beds	Deaths/Cases	UCI/Cases	BCG
Chile	94858	997	1325	0,01	0,01	0
Malaysia	7762	115	197	0,01	0,03	0
Israel	17012	284	429	0,02	0,03	38
Republic of Korea	11471	270	378	0,02	0,03	0
Panama	13018	330	490	0,03	0,04	0
Turkey	163103	4515	6930	0,03	0,04	0

Norway	8411	236	334	0,03	0,04	11
Peru	155671	4371	10604	0,03	0,07	0
Dominican Republic	16908	498	790	0,03	0,05	0
Colombia	28236	890	1223	0,03	0,04	0
Argentina	16111	518	768	0,03	0,05	0
Czechia	9230	319	452	0,03	0,05	10
Republic of Moldova	8098	291	373	0,04	0,05	0
Egypt	23449	913	1560	0,04	0,07	0
Austria	16638	668	873	0,04	0,05	30
Portugal	32203	1396	1797	0,04	0,06	0
Poland	23571	1061	1464	0,05	0,06	0
Finland	6826	314	424	0,05	0,06	14
Germany	181482	8500	11238	0,05	0,06	22
Denmark	11633	581	793	0,05	0,07	34
United Kingdom	733669	38376	52592	0,05	0,07	15
Switzerland	30762	1655	2594	0,05	0,08	33
Philippines	17224	950	1637	0,06	0,10	0
Brazil	498440	28834	42628	0,06	0,09	0
United States of America	1770384	103781	126105	0,06	0,07	100
Romania	19133	1253	1675	0,07	0,09	0
Ireland	24929	1650	2337	0,07	0,09	0
Canada	90179	7073	9288	0,08	0,10	100
Ecuador	38571	3318	8889	0,09	0,23	6
Mexico	87512	9779	7519	0,11	0,09	0
Spain	239068	27127	37964	0,11	0,16	39
Sweden	37113	4395	5820	0,12	0,16	45
Netherlands	46257	5951	8173	0,13	0,18	100
Hungary	3867	524	687	0,14	0,18	0
Italy	232664	33340	43389	0,14	0,19	100
Belgium	58186	9453	12620	0,16	0,22	100
France	151496	28770	39075	0,19	0,26	13

It isn't an obvious relation, but it seems to be some evidence regarding the vaccination of BCG and the output of deaths/cases and UCI/cases.

So, as vaccination helps mitigating some symptoms it explains why countries with no vaccination tend to have more hospitalizations both normal and in ICU but nothing seems to explain why BCG vaccination doesn't have any effect on mortality.

One reason to explain this results might be the quality of the data since countries have different criteria when considering deaths per COVID-19, which might have influenced the result of the estimation of the model deaths per COVID-19 and so, affected the coefficients and significance of some variables.

Variables of mobility

As we can see in the final model, only the variable *residential* was significative in the model but it is important to note that there is a very strong correlation between this variable and the other variables of mobility.

	transit	workplace	retail	grocery	parks
residential	-0,96	-0,97	-0.96	-0,92	-0,65

There are very strong negative relations between residential and transit, workplace, retail and grocery which mean that they strongly grow in opposite directions. As more people stayed at their residential area, the mobility in public

transports, workplaces, retail and leisure and supermarkets and pharmacies had strongly decreased.

The relation between residential and parks is still negative, which means that they still grow in opposite directions, but it isn't as stronger as the relations before which makes a lot of sense when looking to what is happen around the world. With the shutdown of gyms, malls cafes and other areas of leisure, people tend to start using more and more parks to do exercise, pet their dogs, go for a walk, simply not to be always at home.

7. Conclusion

In conclusion, there is effectively no evidence that the economic structure of the country affects its mortality. Like the case of the United States of America which has one of the strongest economic structures in the world and in terms of the death-to-case ratio per COVID-19, are among the worst outcomes. In addition, Shet et al. (2020) had even found counterintuitive results between the economic structure and the output of the country, in which countries with higher GDP had worse results in terms of mortality.

Regarding healthcare resources, no evidence was found. However, this study presents some limitations so that it is not possible to draw conclusions about the importance of healthcare resources for the country's results in terms of mortality. This is because in our study we talked about national capacity, but mortality in countries was often affected by strong contagion in regional terms. For example, in Italy, in the south, the pandemic was hardly felt, healthcare resources were practically unused, but in the Lombardy area, resources were insufficient, we saw a lot of news that showed the stocking and chaos that were the hospitals and the ICU's in that area. Nevertheless, in this study, situations such as this were not taken into account, and, therefore, it is possible to say that the national capacity for resources does not affect mortality, but it is not possible to say that the resources in particular do not affect it.

Regarding comorbidities, it was possible to observe that countries with a higher percentage of diabetics have worse results in terms of mortality, and that countries that do not have a national vaccination policy for BCG were also more affected.

Regarding the response time of each country, there is no evidence that the time it took them, if they did, to enter lockdown, have contributed to mortality. This also has a lot to do with the attitude of the population, and therefore the variables of mobility ended up explaining this situation much better. For example, in

Portugal, even before the lockdown was enacted, people voluntarily started staying at home and many customer service establishments began to close on their own initiative. Moreover, in some countries the national lockdown was not enacted because the situation did not require it, since the contagion was very regional, and it was enough to isolate the region in question.

The measure of "stay at home" effectively has an impact on mortality. Given a pandemic in which the main objective is to have the minimum number of contagion to be more easily controllable in terms of resources, the more people stay at home, the less contact there is, so fewer cases, so fewer hospitalizations, so less burden in hospitals and health professionals, so fewer deaths.

It is important to note that this study has some limitations in terms of data, more specifically in terms of covid-19 mortality. This is because countries have different criteria for considering deaths from COVID-19.

For example, France only includes data on those who die in hospitals, and Spain does not record untested deaths in nursing homes for the elderly.

Finally, as a proposal for future investigations, it would be interesting to study the impact of variables of hospital resources at the regional level to understand whether there is effectively an impact in terms of bed resources and health professionals.

In addition, to find a criterion for standardizing deaths recorded by COVID-19 in all countries to be a more reliable comparison with more concrete results.

8. Bibliography

Acito, Frank, and Vijay Khatri. 2014. "Business analytics: Why now and what next?" In.: Elsevier.

Al-Najjar, H, and N Al-Rousan. 2020. 'A classifier prediction model to predict the status of Coronavirus COVID-19 patients in South Korea', *European Review for Medical*, 24: 3400-03.

Alice Zwerling, M., 2020. BCG Wold Atlas. [online] Bcgatlas.org. Available at: <http://www.bcgatlas.org/index.php> [Accessed 13 May 2020]

Appelbaum, Deniz, Alexander Kogan, Miklos Vasarhelyi, and Zhaokai Yan. 2017. 'Impact of business analytics and enterprise systems on managerial accounting', *International Journal of Accounting Information Systems*, 25: 29-44.

COVID 19. 2020. Mobilidade. [online] Available at: <https://insights.cotec.pt/index.php/component/sppagebuilder/?view=page&id=21> [Accessed 10 May 2020].

Croissant, Yves, and Giovanni Millo. 2008. 'Panel data econometrics in R: The plm package', *Journal of statistical software*, 27: 1-43.

Data.worldbank.org. 2020. Population, Total | Data. [online] Available at: <https://data.worldbank.org/indicator/sp.pop.totl> [Accessed 2 May 2020].

Davenport, Thomas H, Jeanne G Harris, George L Jones, Katherine N Lemon, David Norton, and Michael B McCallister. 2007. 'The dark side of customer analytics', Harvard business review, 85: 37.

Davenport, Thomas H, and DJ Patil. 2012. 'Data scientist', Harvard *business review*, 90: 70-76.

Delen, Dursun, and Haluk Demirkan. 2013. "Data, information and analytics as services." In.: Elsevier.

Dilla, William, Diane J Janvrin, and Robyn Raschke. 2010. 'Interactive data visualization: New directions for accounting information systems research', *Journal of Information Systems*, 24: 1-37.

European Centre for Disease Prevention and Control. 2020. Download Today'S Data On The Geographic Distribution Of COVID-19 Cases Worldwide. [online] Available at: <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>[Accessed 6 May 2020]

Fang, Lei, George Karakiulakis, and Michael Roth. 2020. 'Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection?', *The Lancet. Respiratory Medicine*, 8: e21.

Fauci, Anthony S, H Clifford Lane, and Robert R Redfield. 2020. "Covid-19—navigating the uncharted." In.: Mass Medical Soc.

Froehle, Craig M, and Michael J Magazine. 2013. 'Improving scheduling and flow in complex outpatient clinics.' in, *Handbook of healthcare operations management* (Springer).

Griffin, Paul A, and Arnold M Wright. 2015. 'Commentaries on Big Data's importance for accounting and auditing', *Accounting Horizons*, 29: 377-79.

Guan, Wei-jie, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, and David SC Hui. 2020. 'Clinical characteristics of coronavirus disease 2019 in China', *New England journal of medicine*, 382: 1708-20.

Hawrylak, Peter J, Nakeisha Schimke, John Hale, and Mauricio Papa. 2012. 'Security risks associated with radio frequency identification in medical environments', *Journal of medical systems*, 36: 3491-505.

He, Karen Y, Dongliang Ge, and Max M He. 2017. 'Big data analytics for genomic medicine', *International journal of molecular sciences*, 18: 412.

Hoot, Nathan R, Larry J LeBlanc, Ian Jones, Scott R Levin, Chuan Zhou, Cynthia S Gadd, and Dominik Aronsky. 2008. 'Forecasting emergency department crowding: a discrete event simulation', J Annals of emergency medicine, 52: 116-25.

Institute for Health Metrics and Evaluation. 2020. Institute For Health Metrics And Evaluation. [online] Available at: <http://www.healthdata.org/> [Accessed 4 May 2020].

Khaltaev, Nikolai, and Svetlana Axelrod. 2019. 'Chronic respiratory diseases global mortality trends, treatment guidelines, life style modifications, and air pollution: preliminary analysis', Journal of thoracic disease, 11: 2643.

Krispin, R., 2020. Ramikrispin - Overview. [online] GitHub. Available at: <https://github.com/RamiKrispin>[Accessed 15 April 2020].

Luo, Jianxi 2020. 'Predictive Monitoring of COVID-19', *SUTD Data-Driven Innovation Lab*.

Majumder, Maimuna S, Sheryl A Kluberg, Sumiko R Mekaru, and John S Brownstein. 2015. 'Mortality risk factors for Middle East respiratory syndrome outbreak, South Korea, 2015', *Emerging infectious diseases*, 21: 2088.

Mehta, Nishita, and Anil Pandit. 2018. 'Concurrence of big data analytics and healthcare: A systematic review', *International journal of medical informatics*, 114: 57-65.

Mizumoto, Kenji, Akira Endo, Gerardo Chowell, Yuichiro Miyamatsu, Masaya Saitoh, and Hiroshi Nishiura. 2015. 'Real-time characterization of risks of death associated with the Middle East respiratory syndrome (MERS) in the Republic of Korea, 2015', *BMC medicine*, 13: 228.

Moore, Bert 2009. 'The potential use of radio frequency identification devices for active monitoring of blood glucose levels', *Journal of diabetes science*, 3: 180-83.

Oecd.org. 2020. OECD.Org - OECD. [online] Available at: <https://www.oecd.org/> [Accessed 27 April 2020].

Rogers, A, E Jones, and Dmitry Oleynikov. 2007. 'Radio frequency identification (RFID) applied to surgical sponges', Surgical endoscopy, 21: 1235-37.

Roser, M., Ritchie, H., Ortiz-Ospina, E. and Hasell, J., 2020. Coronavirus Pandemic (COVID-19). [online] Our World in Data. Available at: <https://ourworldindata.org/coronavirus> [Accessed 29 April 2020].

Shet, Anita, Debashree Ray, Neelika Malavige, Mathuram Santosham, and Naor Bar-Zeev. 2020. 'Differential COVID-19-attributable mortality and BCG vaccine use in countries', *MedRxiv*.

Shojaee, Sajad, Mohamad Amin Pourhoseingholi, Sara Ashtari, Amir Vahedian-Azimi, Hamid Asadzadeh-Aghdaei, and Mohammad Reza Zali. 2020. 'Predicting the mortality due to Covid-19 by the next month for Italy, Iran and South Korea; a simulation study', *Gastroenterology*

Hepatology from Bed to Bench

13: 177.

Sonego, Michela, Maria Chiara Pellegrin, Genevieve Becker, and Marzia Lazzerini. 2015. 'Risk factors for mortality from acute lower respiratory infections (ALRI) in children under five years of age in low and middle-income countries: a systematic review and meta-analysis of observational studies', *PloS one*, 10.

Souza, Gilvan C 2014. 'Supply chain analytics', *Business Horizons*, 57: 595-605.

Ward, Michael J, Keith A Marsolo, and Craig M Froehle. 2014. 'Applications of business analytics in healthcare', *Business Horizons*, 57: 571-82.

Wilder-Smith, Annelies, Calvin J Chiew, and Vernon J Lee. 2020. 'Can we contain the COVID-19 outbreak with the same measures as for SARS?', *The Lancet Infectious Diseases*.

Yang, Ai-Ping, Jianping Liu, Wenqiang Tao, and Hui-ming Li. 2020. 'The diagnostic and predictive role of NLR, d-NLR and PLR in COVID-19 patients', *International Immunopharmacology*: 106504.

Yang, Xiaobo, Yuan Yu, Jiqian Xu, Huaqing Shu, Hong Liu, Yongran Wu, Lu Zhang, Zhui Yu, Minghao Fang, and Ting Yu. 2020. 'Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study', *The Lancet Respiratory Medicine*.

Zhang, Jin-jin, Xiang Dong, Yi-yuan Cao, Ya-dong Yuan, Yi-bin Yang, You-qin Yan, Cezmi A Akdis, and Ya-dong Gao. 2020. 'Clinical characteristics of 140 patients infected with SARS‐CoV‐2 in Wuhan, China', *Allergy*.