Challenges and prospects for data analytics: an attempt to estimate the number of deaths due to COVID-19 using publicly available secondary data

Hiroko Oe, Bournemouth University
Yasuyuki Yamaoka, The Open University of Japan

Abstract:
COVID-19 is a coronary-type virus that emerged in Wuhan, China, at the end of 2019. As it was a new virus never before encountered by humankind, it has been imperative to prevent the spread of the virus, and countries around the world implemented lockdown measures (city closures) for much of 2020. There already exist a number of sophisticated mathematical models in immunology, such as the susceptible–exposed–infectious–removed (SEIR) model. However, the data used in these models are only available to specialist organisations, such as medical research institutes, and therefore cannot be tested by outsiders. Thus, in this study we attempted to use open source data from the Ministry of Health, Labour and Welfare (MHLW) to explore the policy implications of their statistical conclusions.

1. Confirmation of data sources

We begin this study by reviewing the dataset required by the statistical model. The aim here is to determine whether it is possible for social scientists, such as myself, who are not experts in epidemiology, or for policy makers working in the public sector, to test certain simple estimation methods using general-purpose data and publicly available secondary data and whether it is possible to estimate the number of deaths due to COVID-19 using such a method. The first step in this study is to determine what data are accessible. Table 1 presents the data that could be attained from an open source and the data we developed for this research.

	Data	Ministry of Health, Labour and Welfare Open Data	Derived and added data
1	Number of PCR tests (single day)	○	
2	Number of PCR positive patients (single day)	○	
3	Number of people treated in hospital (cumulative)	○	
4	Number of discharged and released patients (cumulative)	○	
5	Number of deaths (cumulative)	○	
6	Number of PCR tests (cumulative)		○
7	Number of PCR positive patients (cumulative)		○
8	Number of patients treated in hospital (single day)		○
9	Number of discharged and decertified patients (single day)		○
10	Number of deaths (single day)		○
11	Number of deaths (14-day delay)		○
12,13	Tokyo maximum and minimum temperatures		○
14,15	Tokyo average humidity/lowest humidity		○

5

Table 1. Data sources and development for the study

Data from a total of 442 days from 5 February 2020 to 30 April 2021, downloaded from MHLW Open Data were used. SPSS version 26 was used for the analysis.

2. Analysis

2.1 Data overview

Figure 1 shows an overview of the data sorted based on the available data, as outlined in the Section 1.

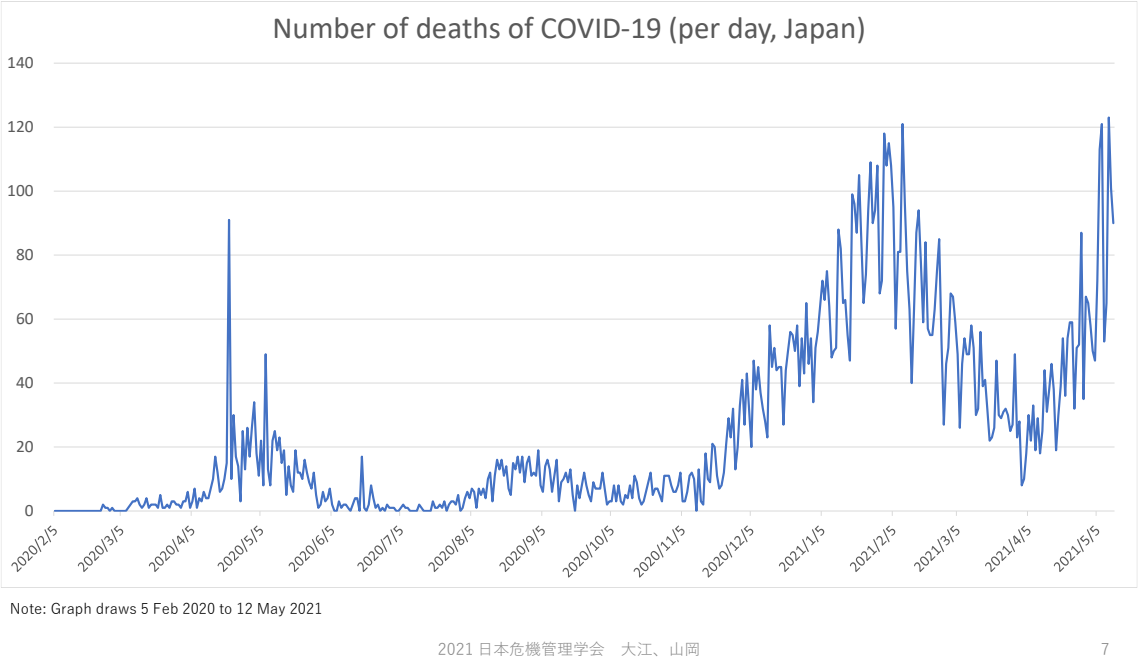

Figure 1. Number of deaths due to COVID-19 (daily basis)

Figure 1 illustrates the daily evolution of the number of deaths due to COVID-19 (the explained variable), which is the main theme of this study.

As already mentioned, 15 types of data were prepared in this study, but in the actual preparation of the estimation model, data with a risk of multiple correlations were deleted. The number of deaths was delayed by 14 days to take account for the delay associated with the illness (Harrison et al. 2021).

2.2 Data preparation

As a result of this data preparation work, we chose to use six variables in our estimation model. As the six variables obtained as candidates for estimating the dependent variables had no collinearity problems and were not too strongly correlated with one another, we decided to use them to conduct a multiple regression analysis for the estimation of the explained variables. These six variables were as follows:

- number of positive polymerase chain reaction (PCR) tests in a single day;
- people treated in hospital (cumulative);
- maximum temperature;
- minimum temperature;

- minimum humidity; and
- average humidity.

With the number of deaths (difference of 14 days) as the dependent variable, the observed variables were narrowed down from the above seven candidate explanatory variables using a stepwise method, and the following four variables were judged to be valid for the estimation of deaths due to COVID-19:

- number of people treated in hospital (cumulative);
- minimum temperature;
- number of positive PCR tests in a single day; and
- minimum humidity.

2.3 Regression analysis to estimate the deaths due to COVID-19

Following the data preparation, a regression analysis was conducted, and the outcome was discussed. Table 2 shows that the R-squared/adjusted R-squared values are high (over 80%), suggesting that the model has sufficient explanatory power.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
Final	0.918	0.843	0.842	10.871	1.328
Predictors: (Constant), Number of patients (cummulative), Lowest temperature, PCR positive number (day), Lowest humidity%					
Dependent Variable: Number of deaths (14 days shift)					

Table 2. Regression analysis outcomes

Model		Sum of Squares	df	Mean Square	F	Sig.
Final	Regression	277310.491	4	69327.623	586.667	.000[e]
	Residual	51641.129	437	118.172		
	Total	328951.620	441			
Predictors: (Constant), Number of patients (cummulative), Lowest temperature, PCR positive number (day), Lowest humidity%						

Table 3. ANOVA outcomes

Table 3 shows the results of an analysis of variance (ANOVA), which also demonstrates

that the model is reliable and compatible with the dataset (p = 0.000). Therefore, we will continue to the stage of further discussion based on the regression outcome.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
Final	(Constant)	14.429	1.694		8.518	0.000		
	Number of patients (cummulative)	0.001	0.000	0.631	12.885	0.000	0.150	6.669
	Lowest temperature	-0.545	0.085	-0.160	-6.403	0.000	0.577	1.733
	PCR positive number (day)	0.004	0.001	0.193	4.029	0.000	0.157	6.389
	Lowest humidity%	-0.074	0.033	-0.052	-2.257	0.024	0.673	1.486
Dependent Variable: Number of deaths (14 days shift)								

Table 4. Development of estimation model

To determine the significance of the four explanatory variables in the estimation model, Table 4 shows the extent to which the four candidate explanatory variables selected by the stepwise method were effective in estimating the dependent variable. The statistical significance of all explanatory variables was $p < 0.05$, indicating that they are valid as variables in the estimation model. Also, for all four variables, the variance inflation factor (VIF) did not exceed 10, that is, there was no multiple correlation in this model.

This analysis yielded a multiple regression equation to estimate the number of deaths due to COVID-19:

Number of deaths (with 14-day delay) = 14.429 – 0.001 * (people treated in hospital (cumulative)) – 0.545 * (minimum temperature) + 0.004 * (positive PCR tests in a single day) – 0.074* (minimum humidity %) .... (Equation 1)

3. Discussion

3.1 Overall discussion

The number of deaths due to COVID-19 can be estimated from the following four observables as described, but of those, the one with the greatest impact on the estimation is the minimum temperature, followed by the minimum humidity. So far, the risk of death due to COVID-19 has been discussed mainly in terms of the number of

patients with positive PCR tests per single day, the number of people treated in hospital (cumulative), and the percentage of patients each local authority can accept.

However, our pilot analysis suggests that climatic factors, such as the minimum temperature and minimum humidity, may be more relevant to the number of deaths than previously thought. Although it would be inappropriate to immediately generalise the results of this trial analysis, policy-makers and public sector officials should at the very least consider how to respond in advance, assuming that mortality rates will vary with climate factors.

3.2 Findings from a study of the Spanish flu in 1918

The younger generation was most affected by the Spanish flu of 1918. There is a great need to consider the socioeconomic factors (including access to health care) behind the pandemic then and now. Analysis by a European research team shows that of the determinants of mortality, latitude was the main explanatory variable for the overall mortality rate, while the excess mortality rate varied most greatly according to occupation (e.g. miners). Further, population density was negatively correlated with mortality in lower income groups (based on the presence of an urban premium; Basco et al. 2021).

The economic situation in Spain in 1918 resembles the current economic situation in many developing countries, where budgetary constraints make social distancing difficult. This observation may help us to examine the impact of COVID-19 on developing countries.

It will continue to be important to reflect once again on the insights and implications garnered from the Spanish flu and to attempt to mitigate the socioeconomic impact of the pandemic as much as possible through an interdisciplinary approach (social sciences, medicine and public health). In particular, it is an urgent task to establish a system to ensure that data-based quantitative analysis is carried out and that the results are reflected in policy-making.

4. Conclusion

In this study, data on 442 days from February 2020 to April 2021 were used, but it is unclear whether the “mutant strains” under discussion after May 2021 can be adopted in

the analysis or not.

As climatic factors, we used the minimum temperature and humidity in Tokyo as a representative example, but in order to improve the accuracy of the regression equation, these factors should be broken down for each region.

One possible direction of data analytics on the theme of the pandemic is to compare the COVID-19 pandemic with the case of the Spanish flu in Europe in 1918 and to use the findings as a guide. In any case, it will become increasingly important to collaborate with experts in different fields, such as the social sciences, medicine, public health and statistics, to make detailed studies and forecasting models based on data and to create policies that reduce the impact of pandemics as much as possible.

References are available on request:
Corresponding author: Dr Hiroko Oe, Bournemouth University
(hoe@bournemouth.ac.uk)