Received: 12 March 2020 | Accepted: 6 April 2020

DOI: 10.1002/jmv.25850

RESEARCH ARTICLE

JOURNAL OF MEDICAL VIROLOGY WILEY

Data analysis of coronavirus COVID-19 epidemic in South Korea based on recovered and death cases

Nadia AL-Rousan | Hazem AL-Najjar

Department of Computer Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, Istanbul, Turkey

Correspondence
Nadia AL-Rousan, Department of Computer Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, Istanbul 34310, Turkey.
Email: nadia.rousan@yahoo.com

Abstract

Coronavirus epidemic caused an emergency in South Korea. The first infected case came to light on 20 January 2020 followed by 9583 more cases that were reported by 29 March 2020. This indicates that the number of confirmed cases is increasing rapidly, which can cause a nationwide crisis for the country. The aim of this study is to fill a gap between previous studies and the current rate of spreading of COVID-19 by extracting a relationship between independent variables and the dependent ones. This study statistically analyzed the effect of factors such as sex, region, infection reasons, birth year, and released or diseased date on the reported number of recovered and deceased cases. The results found that sex, region, and infection reasons affected both recovered and deceased cases, while birth year affected only the deceased cases. Besides, no deceased cases are reported for released cases, while 11.3% of deceased cases positive confirmed after their deceased. Unknown reason of infection is the main variable that detected in South Korea with more than 33% of total infected cases.

KEYWORDS

engineering and technology, epidemiology, infection, South Korea

1 | INTRODUCTION

The first case of coronavirus COVID-19 disease in South Korea was announced on 20 January 2020.[1] The distance of South Korea (located on 37° North and 127° East) from China is 2123 km. After China and Italy, South Korea is the third affected country by the epidemic of coronavirus.[2] Coronavirus infected more than 10 156 people by 5 April 2020 in South Korea while the global number of infected cases is 1 201 943.[3] According to the reports of World Health Organization, there are 249 127 recovered cases and 64 781 deceased cases globally. South Korea has 183 deceased cases and 6463 recovered cases.[4,5] By this time, the number of confirmed infected cases has been rapidly increasing in South Korea. The growth rate of confirmed cases was rapid until 11 March 2020; however, there was a slow increase after March 11, 2020 until the current time.

The first deceased case was reported on 20 February which climbed to 183 cases by the end of 5 April 2020, the first recovered case was reported on 7 February that reached 6463 cases by the end of 5 April 2020. All of the infected cases suffered from several symptoms before their infection to the coronavirus was confirmed[6,7]; these symptoms were feeling cold, flu, and pneumonia. In total, 10 206 people were tested for coronavirus by the end of 8 March in South Korea. It was reported that most of the infected cases in South Korea had, at some point of time, visited local cites before confirming positive to the virus(ie, isolation hospital, airport, restaurant, market, café, clinic, company, movie theater, etc).[8]

Evidently, several probable reasons for the spread of coronavirus in South Korea are summarized by many researchers in the field.[9,10] Several research works were conducted to find the probable reasons for the spread of the virus in South Korea rather than other countries. Researchers have started to extract information about the infected cases and have analyzed the biomedical information and their medical histories to extract the main parameters that could cause coronavirus spreading. They suggested that the spreading of coronavirus could be linked to sex, birth year, or the region they come from.

The aim of this study is to study the effect of several attributes on the spreading of Coronavirus COVID-19 in South Korea based on real collected data and published reports. The main target is to study the effect of sex, birth year, the region they come from, and the place they visited on the number of deceased and recovered cases in South Korea. The χ^2 test is used to find the impact of the previous attributes on the number of recovered and deceased cases. The study would give an overview about the current situation in South Korea, besides, it may show the main parameters that can be used to build a forecasting model.

2 | METHODOLOGY

As explained earlier, this study analyzes the effect of sex, age, region, and transportation on susceptibility to COVID-19 in South Korea. Official time-series data from Korean Centers for Disease Control and Prevention for coronavirus disease 2019 (COVID-19) cases in South Korea from 20 January to 29 March are used.[11] The obtained data contains several information about 2771 infected cases (where the rest of data is missing and not reported) in South Korea that includes their sex, birth year, the original country they come from, the region that they live in, whether they have any pre-existing condition, infection reason and order, confirmed date, deceased or released date, and their current state. The data contains several missing variables that are excluded from the analysis to give a clear overview about the coronavirus epidemic in South Korea.

Both statistical analysis and the χ^2 test are used to analyze the collected data and to ensure about the impact of sex, region, infection reason, released and deceased cases, and the birth year on the number of recovered and deceased cases in South Korea. The χ^2 test is heavily recommended to be used in survey research, business intelligence, engineering, and scientific research as well. The χ^2 test is a common mathematical test that is used to check the relationship between two variables in a contingency table that presents (multivariate) frequency distribution of the variables. The χ^2 test, which is normally known as independence test is used to test the hypotheses for categorical variables and to test whether these variables are independent population variable. The χ^2 test can be calculated by finding the summation of dividing the square difference between the observed (O) and the expected (E) values by the expected value for each category in data shown as follows.[12]

$$\sum X_{i-j}^2 = \frac{(O-E)^2}{E} \tag{1}$$

where χ^2 denotes the χ^2 test value, O and E denote the observed and the expected values, respectively. A significant test should be done to determine whether the data is significant or not. The variables are significant if the probability of chance of association occurrence (P) is not more than one out of 1000 cases (.0001), otherwise, it will be considered as nonsignificant. Thus, the association between variables will be rejected. Moreover, to understand the relationship between independent variables and recovered and deceased cases, a cross-tabulation method is used. The cross-tabulation method is used between one independent and one dependent variable, to understand how the dependent variable is moving based on the movement of independent variables. Besides, the multinomial logistic regression classification method is used to check the validity and the robustness of using the χ^2 test to trace the association of each variable namely, sex, confirmation date, birth year, region, and infected reason with both recovered and deceased cases.

3 | RESULTS

The study of the effect of sex variable on the recovered cases found that out of the total number of 2765 cases, 1547 were females and 1218 were males. The number of recovered female cases is 511 while 366 recovered male cases exist, as shown in Figure 1.

By using the χ^2 test to find the impact of sex variable on the number of recovered cases, the results found that χ^2 function is χ^2 (2, 2771) = 14.44, $P = .006$ which indicates that sex variable is

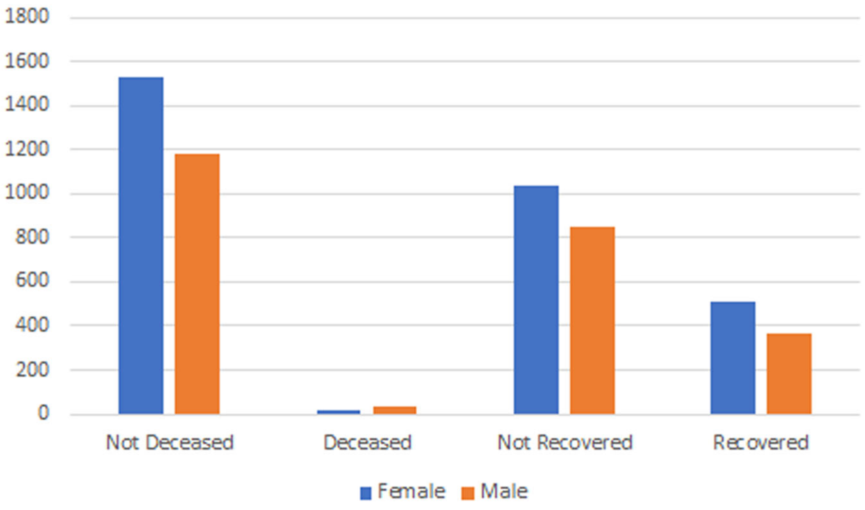

FIGURE 1 Distribution of recovered, unrecovered, deceased, and undeceased cases based on sex

statistically significant in the number of recovered cases. The same test was used to find the effect of sex variable on the number of deceased cases. The results found that χ^2 function is χ^2 (2, 2771) = 12.64, $P = .002$. The results indicate that the sex variable is significant in the number of deceased cases as well. The number of deceased male cases is greater than the number of deceased female cases with 36 and 17 cases, respectively, as shown in Figure 1.

Moreover, the study of the effect of region on the number of recovered and deceased cases found that 879 cases were classified as recovered and the rest of patients either deceased or were isolated. The analysis found that Gyeongsangbuk-do registered the highest number of recovered cases (ie, 430 cases). Besides, the maximum number of infected cases were reported in Gyeongsangbuk-do region, while the range ratio of recovered cases to infected cases is from 3% to 70%, as shown in Figure 2.

The results of the χ^2 test are defined as χ^2 (16, 2771) = 516.49, $P < .0001$, which indicates that the region variable is statistically significant in the number of recovered cases, whereas the χ^2 result of deceased cases is defined as χ^2 (16, 2771) = 326.20, $P < .0001$. In addition, it was found that none of the unspecified cases were reported as deceased, while 2.6%, 31.7%, 2.7%, 0.21%, and 2.4% of cases in Busan, Daegu, Gangwon-do, Gyeonggi-do, Gyeongsangbuk-do regions were reported as deceased cases, respectively. Figure 3 shows the ratios of deceased cases to infected cases in all regions.

Moreover, to understand the impact of the reasons of infection on both studied variables including deceased and recovered cases, the study analyzed the infection sources types using cross-tabulation and the χ^2 test. The collected dataset from the South Korean hospital classified the infection reasons into several groups, namely Bonghwa Pureun Nursing Home (F1), Changnyeong Coin Karaoke (F2), Cheongdo Daenam Hospital (F3), contact with patient (F4), Dongan Church (F5), ETC (F6), Eunpyeong St. Mary's Hospital (F7), Geochang Church (F8), Guro-gu Call Center (F9), Gyeongsan Cham Joeun Community Center (F10), Gyeongsan Jeil Silver Town (F11), Gyeongsan Seorin Nursing Home (F12), gym facility in Cheonan (F13), gym facility in Sejong (F14), Ministry of Oceans and Fisheries (F15), Onchun Church (F16), overseas inflow (F17), Pilgrimage to Israel (F18), River of Grace Community Church (F19), Seongdong-gu APT (F20), Shincheonji Church (F21), Suyeong-gu Kindergarten (F22), and Unknown (F23). While studying the infection reasons in both recovered and deceased cases, it was found that 928 cases were recorded without determining the reason of infection. The analysis found that no reason of infection was found for 409 recovered cases and 38 deceased cases in South Korean hospitals. Besides, it was found that direct contact with patient, ETC, overseas inflow, Guro-gu Call Center are the main four causatives of the transmission of the virus. The results found that all the infected cases who visited to Changnyeong Coin Karaoke, Pilgrimage to Israel, River of Grace Community Church, and Suyeong-gu Kindergarten recovered. The range of recovered percentages of patients is between 3.57% and 66.67% based on infection case where Guro-gu Call Center, Bonghwa Pureun Nursing Home, Dongan Church, Gyeongsan Jeil Silver Town, Gyeongsan Seorin Nursing Home, and Gyeongsan Cham Joeun Community Center showed no recovery in the infected patients, as shown in Figure 4. In addition, 72% and 28% of deceased cases were infected because of unknown reasons or one of the suggested cases from F1 to F22. Moreover, no deceased case until 29 March 2020 is reported because of visiting Wuhan. The ratio of deceased cases is drawn in Figure 5.

The results of χ^2 of infection reasons on the number of recovered and deceased cases are χ^2 (22, 2771) = 383.5, $P < .0001$, and χ^2 (22, 2771) = 158.23, $P < .0001$, respectively, which indicates that the infection reasons are statistically significant with the number of recovered and deceased cases.

In addition, to validate the speed of South Korean hospitals in testing the coronavirus cases, a relationship between released and deceased cases was studied. The results found that no deceased

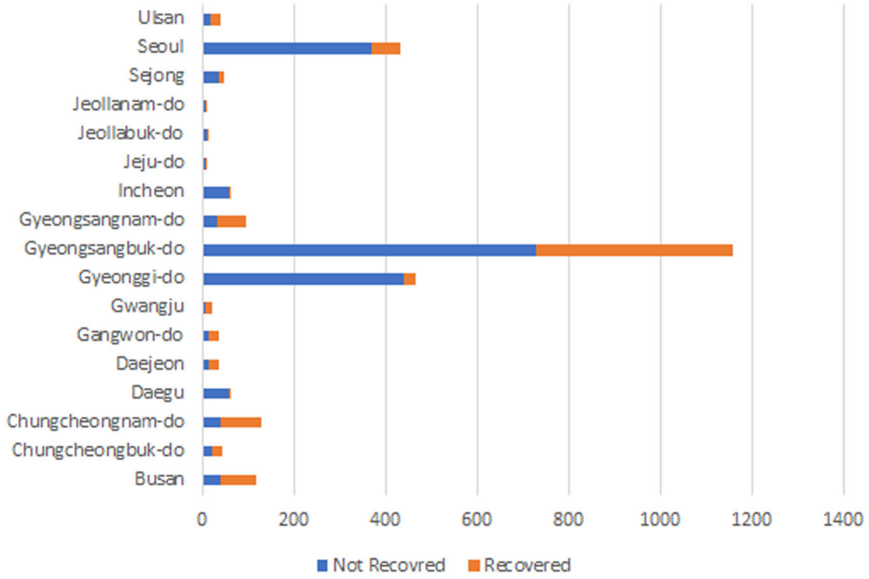

FIGURE 2 Ratios of recovered cases to infected cases based on region

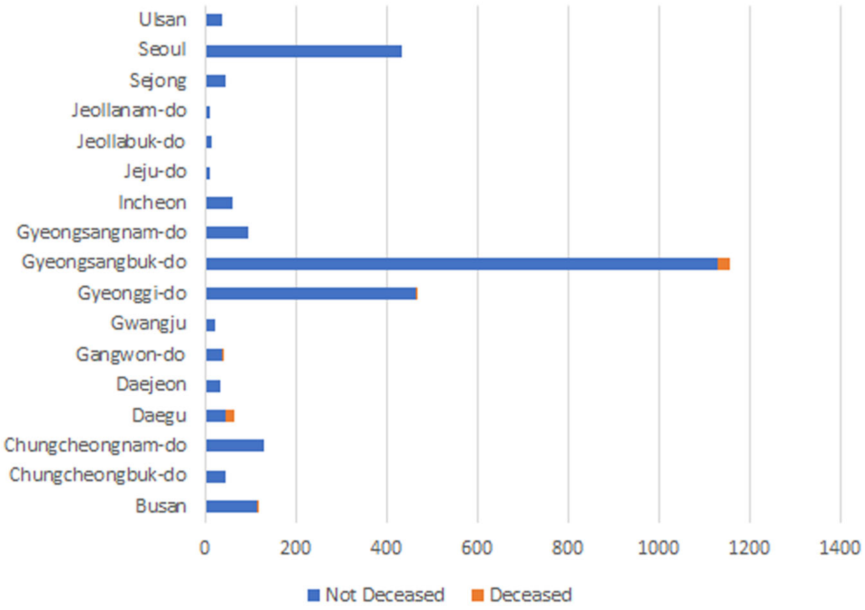

FIGURE 3 Ratios of deceased cases to infected cases based on region

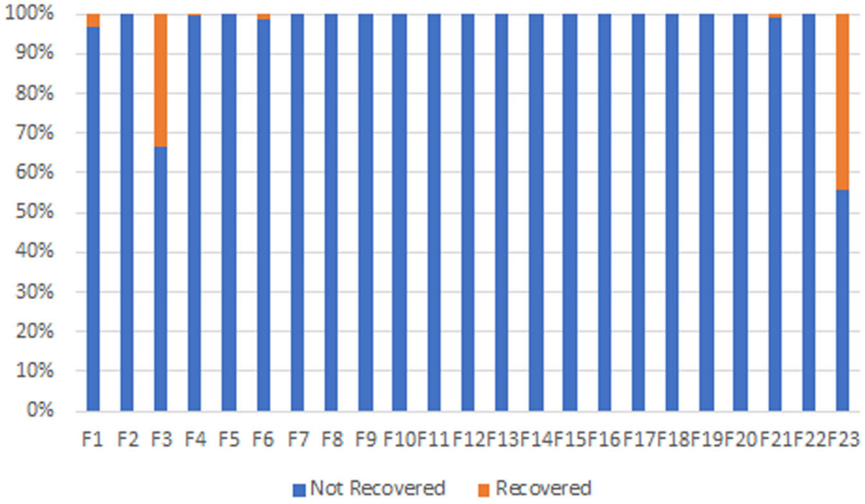

FIGURE 4 Ratio of recovered cases to infected cases based on the infection reasons

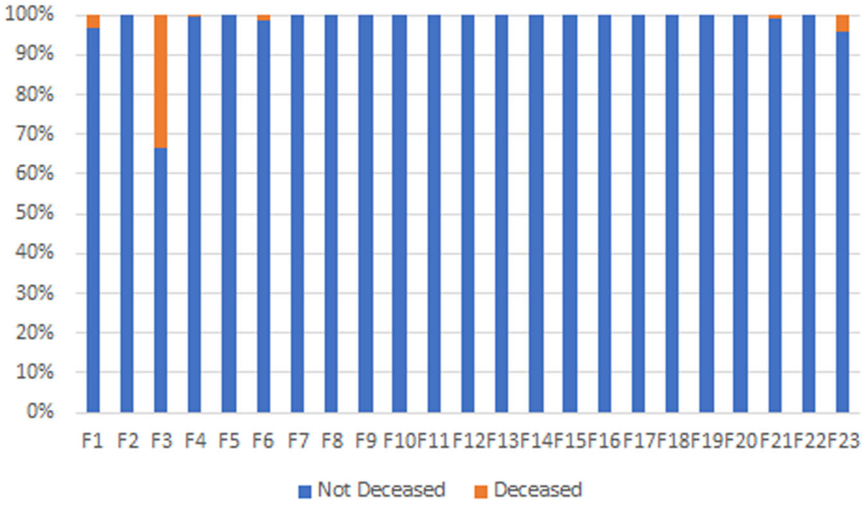

FIGURE 5 Ratio of deceased cases to infected cases based on the infection reasons

TABLE 1 Classification results based on multinomial logistic regression

	0.00	1.00	Percent correct
Death			
0.00	2361	3	99.9%
1.00	8	42	84.0%
Overall percentage	98.1%	1.9%	99.5%
Recovered			
0.00	1426	212	87.1%
1.00	78	698	89.9%
Overall percentage	62.3%	37.7%	88.0%

cases were reported for any of the recovered cases, where, 11.3% of 53 deceased cases were confirmed to be infected by the coronavirus either after they deceased or on the same day of their death, which indicates that the process to discover coronavirus symptoms is acceptable. Moreover, after studying the relationship between the birth year and the number of recovered and deceased cases, it was found that there was no relationship between the birth year and the number of recovered cases, while 83% of deceased cases are persons who are more than 60 years old. Birth date is an effective variable on the number of deceased and recovered cases. The χ^2 test found that the relationship is statistically significant with χ^2 function equal to χ^2 (99, 2771) = 327.89, $P < .0001$ and χ^2 (99, 2771) = 175.32, $P < .0001$, respectively.

To verify the relationship between selected independent variables and one of the dependent variables (ie, recovered and deceased), a multinomial logistic regression was used. The overall percentage results of percent correct for deceased and recovered cases are 99.5% and 88.0%, respectively, as shown in Table 1. In additon, to check the significance level of the independent variables, a likelihood test is adopted for both classifiers as shown in Table 2. The deceased results showed that Birth_Date, Sex, Country, Region, Infection_Reason, and confirmed_date are statistically significant factors to predict the deceased cases. The recovered results showed that Sex, Region, Infection_Reason, confirmed_date, and Birth_Dateare statistically significant factors to predict the recovered cases, where Country is not statistically significant. The results revealed that determining the infection reason and confirmed date are useful information to determine the deceased cases, besides the results found that determining the region of the patients, early detecting the COVID-19, the reason of infection, and the gender of the patient could increase the probability of treating the patients.

4 | DISCUSSION

Sharing the patients' information can help researchers and governments to understand the virus transmission. The sequence of COVID-19 and the virus are shared between different laboratories to study the virus and to find its characteristics. This study adopted the following variables including sex, region, infection reasons, birth date, confirmed-deceased date, and confirmed-recovered date. The study found that sex has a strong relationship with recovered and deceased cases with the majority of infected patients being male. This conclusion is in line with the conclusion in[13] which proved that the number of female smokers is less than the number of male smokers. After considering the region variable in the recovered and deceased cases, the results revealed that the number of deceased and recovered varies based on the region of the infected case. This result is in line with findings[9] that showed the number of cases is changed based on the weather variables, geographical area, populous density, people

TABLE 2 Likelihood ratio tests

	Model fitting criteria			Likelihood ratio tests		
Effect	AIC of reduced model	BIC of reduced model	−2 log likelihood of reduced model	χ^2	*df*	Sig.
Death						
Intercept	482	1703	60	0	0	
Birth_Date	574	1790	154	94	1	.000
Sex	487	1703	67	7	1	.006
Country	17 711	18 243	17 527	17 467	119	.000
Region	846	2027	438	379	7	.000
Infection_Reason	493	1593	113	54	21	.000
confirmed_date	449	1352	137	77	55	.027
Recovered						
Intercept	1606	2827	1184	0	0	
Sex	1632	2848	1212	28	1	.000
Country	1604	2820	1184	1	1	.443
Region	1738	2270	1554	370	119	.000
Infection_Reason	1639	2820	1231	47	7	.000
confirmed_date	1625	2725	1245	61	21	.000
Birth_Date	1917	2820	1605	422	55	.000

communication, transports, and the nature of Koreans' bodies that permitted the incubation of the disease. The infection reasons classification can give a hint to researchers in the field to trace the reason of the infection to classify the COVID-19 which may help the doctors to give extra treatment to special cases. The infection reasons variable gave an alert to all researchers and governments that the virus has a strong transmission ability between people, besides there is a limited information about how the virus can infect a new person without communicating with an infected one. The birth date variable proved that the majority of deceased cases is more than 60 years old, where no indicator was found between the number of recovered cases and birth date. The results are in line with the findings of the China Center for Disease Control which indicates that the fatality rate of infected cases less than 40 years is less than patients over 80 years.

Moreover, the results of the COVID-19 test of 11.3% deceased cases were reported either after they deceased or on the same day of their death. This revealed that the doctors were unable to treat these cases since no positive or negative information about the virus infection was reported, whereas all the recovered case did not show any deceased cases until 29 March 2020. Besides, the results found that multinomial logistic regression could give initial indicator about the possibility to survive or die based on the collected data. It is found that the results of multinomial logistic are in line with the results of the χ^2 test.

5 | CONCLUSION

This study highlighted the main variables that could be considered to understand the COVID-19. The main variables that are considered in this study are sex, region, infection reasons, birth date, confirmed-deceased date, and confirmed-recovered date. After discussing the obtained results, it is found that to mitigate the coronavirus disease in South Korea, several procedures should be followed. These procedures are related to prevent any direct contact with patients, especially those inside isolated hospitals and prevent any kind of community events (ie, visiting patients, going to restaurants, shopping at groups and big stores, etc). Besides, the processes of testing people against coronavirus should be faster. Keeping South Korea safe from coronavirus would affect all nearby countries.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

ORCID

Nadia AL-Rousan http://orcid.org/0000-0001-8451-898X
Hazem AL-Najjar https://orcid.org/0000-0002-6143-2734

REFERENCES

1. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *Int J Antimicrob Agents*. 2020;55(3):105924–105933. https://doi.org/10.1016/j.ijantimicag.2020.105924
2. 2019-nCoV Global Cases (by Johns Hopkins CSSE). https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6. Accessed March 11, 2020.
3. World Health Organization. Home Care for Patients With Suspected Novel Coronavirus (nCoV) Infection Presenting With Mild Symptoms and Management of Contacts: Interim Guidance. 2020.
4. Shim E, Tariq A, Choi W, Lee Y, Chowell G. Transmission potential of COVID-19 in South Korea. 2020.
5. Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019-nCoV in Italy: where they come from? *J Med Virol*. 2020;92(5):518–521. https://doi.org/10.1002/jmv.25699
6. European Centre for Disease Prevention and Control (ECDC). Risk Assessment: Outbreak of Acute Respiratory Syndrome Associated With a Novel Coronavirus, Wuhan, China; First Update. 2020. https://www.ecdc.europa.eu/en/publications-data/risk-assessment-outbrea. Accessed March 11, 2020.
7. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223):507-513. https://doi.org/10.1016/s0140-6736(20)30211-7
8. Khan N, Naushad M. Effects of corona virus on the world community. *SSRN Electronic Journal*. 2020. https://doi.org/10.2139/ssrn.3532001. Accessed April 13, 2020.
9. AL-Rousan N, Al-Najjar H. Nowcasting and forecasting the spreading of novel coronavirus 2019-nCoV and its association weather variables in 30 Chinese provinces: a case study (2/9/2020). *SSRN Electronic Journal*. 2020. https://doi.org/10.2139/ssrn.3537084. Accessed April 13, 2020.
10. Yoo JH, Hong ST. The outbreak cases with the novel coronavirus suggest upgraded quarantine and isolation in Korea. 2020;35(5):e62. https://doi.org/10.3346/jkms.2020.35.e62
11. Korea Centers for Disease Control and Prevention. 2020. http://ghdx.healthdata.org/organizations/korea-centers-disease-control-and-prevention-kcdc
12. Islam JY, Khatun F, Alam A, et al. Knowledge of cervical cancer and HPV vaccine in Bangladeshi women: a population based, cross-sectional study. 2018;18(1). https://doi.org/10.1186/s12905-018-0510-7
13. Hwang JE, Choi Y, Yang YS, Oh Y. Gender differences in the perceived effectiveness of female-focused graphic health warnings against smoking in South Korea. *Health Education Journal*. 2020;79(1):58–72.

How to cite this article: AL-Rousan N, AL-Najjar H. Data analysis of coronavirus COVID-19 epidemic in South Korea based on recovered and death cases. *J Med Virol*. 2020;92:1603–1608. https://doi.org/10.1002/jmv.25850