Original Article

J Prev Med Public Health 2015;48:203-215 • http://dx.doi.org/10.3961/jpmph.14.057

pISSN 1975-8375 eISSN 2233-4521

Journal of Preventive Medicine & Public Health

Correlations Between the Incidence of National Notifiable Infectious Diseases and Public Open Data, Including Meteorological Factors and Medical Facility Resources

Jin-Hwa Jang[1,2], Ji-Hae Lee[2,3], Mi-Kyung Je[2,3], Myeong-Ji Cho[2], Young Mee Bae[4,5], Hyeon Seok Son[2,3], Insung Ahn[1]

[1]Biomedical Prediction Technology Laboratory, Convergence Technology Research Division, Korea Institute of Science and Technology Information, Daejeon; [2]Laboratory of Computational Biology and Bioinformatics, Institute of Public Health and Environment, Graduate School of Public Health, Seoul National University, Seoul; [3]Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, Seoul; [4]Department of Parasitology and Tropical Medicine, Seoul National University College of Medicine, Seoul; [5]Institute of Endemic Diseases, Seoul National University Medical Research Center, Seoul, Korea

Objectives: This study was performed to investigate the relationship between the incidence of national notifiable infectious diseases (NNIDs) and meteorological factors, air pollution levels, and hospital resources in Korea.

Methods: We collected and stored 660 000 pieces of publicly available data associated with infectious diseases from public data portals and the Diseases Web Statistics System of Korea. We analyzed correlations between the monthly incidence of these diseases and monthly average temperatures and monthly average relative humidity, as well as vaccination rates, number of hospitals, and number of hospital beds by district in Seoul.

Results: Of the 34 NNIDs, malaria showed the most significant correlation with temperature ($r=0.949$, $p<0.01$) and concentration of nitrogen dioxide ($r=-0.884$, $p<0.01$). We also found a strong correlation between the incidence of NNIDs and the number of hospital beds in 25 districts in Seoul ($r=0.606$, $p<0.01$). In particular, Geumcheon-gu was found to have the lowest incidence rate of NNIDs and the highest number of hospital beds per patient.

Conclusions: In this study, we conducted a correlational analysis of public data from Korean government portals that can be used as parameters to forecast the spread of outbreaks.

Key words: Infectious disease, Correlation coefficient, Incidence, Meteorology

INTRODUCTION

Outbreaks of infectious diseases are shaped by ecological and environmental factors, as well as social, economic, and demographic structures and their interactions [1]. According to the World Health Organization, the ongoing 2014 to 2015 Ebola outbreak in West Africa has claimed the death of 11 274 individuals among 27 724 cases in eight countries, as of July 24, 2015 [2]. The spread of Ebola in the current outbreak is affected by its pathogenic attributes and is expedited by the climate of West Africa and behavioral factors, such as the state of public hygiene in the region and the ever-increasing number of international travelers [3]. In the Republic of Korea (hereafter Korea) 75 741 patients were reported to have acute infectious diseases in 2013, representing an increase of 24 251 from the corresponding figure of 51 490 patients in 2012, according

Received: December 29, 2014 Accepted: July 22, 2015

Corresponding author: Insung Ahn, PhD
245 Daehak-ro, Yuseong-gu, Daejeon 305-806, Korea
Tel: +82-42-869-1053 Fax: +82-42-869-1687
E-mail: isahn@kisti.re.kr

to the Infectious Diseases Surveillance Yearbook published by the Korea Centers for Disease Control and Prevention (KCDC). Of particular note, 252 cases of imported dengue fever were reported in 2013, which was a 69% increase from the figure of 149 cases in 2012 [4]. As stated above, the spread of infectious diseases is affected by a wide range of factors, meaning that it is necessary for studies to examine the epidemiological factors affecting the spread of each disease.

Domestic data must be collected in order to identify and analyze the factors associated with infectious diseases in Korea. Currently, a wide range of public data managed by government agencies are available for public use, supported by the Public Data Declassification Plan (March 2010) and the establishment of Government 3.0 in 2013. As of October 2014, the public has access to approximately 11 711 databases from approximately 706 Korean governmental agencies, including the Bureau of Statistics, the National Weather Service, and the KCDC. The public utilization of these data is rapidly increasing each year, with 110 000 instances of utilization in 2012, 160 000 instances of utilization in 2013, and 22 billion instances of utilization between January and October 2014 [5]. These data, made available by Government 3.0, can be used to predict trends in infectious diseases in Korea. For example, it has been reported that the occurrence of infectious diseases is affected by changes in weather patterns [6]. Most recently, it has been reported that various vectors are proliferating more rapidly as a result of rising temperatures, expediting the spread of infectious diseases [6]. Air pollutants, such as fine particle particulate matter ($PM_{2.5}$) and carbon monoxide (CO), also contribute to an increased number of cases of communicable respiratory illnesses [7]. Using simulation tools (such as CommunityFlu, FluSurge, and FluAid, developed by the US Centers for Disease Control and Prevention), designed to predict the spread of infectious diseases, we set population density, immunization rate, number of hospitals, and number of hospital beds as parameters [8-10]. Public data from Korea were plugged into these parameters to develop more accurate predictions of trends in infectious diseases in Korea. We also performed a correlational analysis between the incidence rates of national notifiable infection diseases (NNIDs) and hospital resources among the districts of Seoul, in order to confirm the importance of preparedness in terms of hospital and vaccination rates.

This study had three primary goals: to conduct a survey of the availability of relevant public data (domestic and international); to perform a correlational analysis of the incidence of NNIDs according to parameters reflected in governmental datasets, including climactic and air pollution-related factors; and to perform a correlational analysis between the incidence of NNIDs and population density, immunization rate, number of hospitals, and number of hospital beds by district in Seoul. Seoul was selected for our study because it has the most public data. Public data were extracted from the Public Data Portal (www.data.go.kr), the Diseases Web Statistics System (http://is.cdc.go.kr/nstat/index.jsp), the Korean Statistical Information Service (http://kosis.kr), the Korea Meteorological Administration (http://sts.kma.go.kr), the Ministry of the Environment (http://stat.me.go.kr), and Seoul City Open Data Square (http://data.seoul.go.kr).

METHODS

Materials

For the survey of the availability of relevant public data worldwide, public data websites established in each country were identified. The categories of data, data formats, and the amount of data offered were then examined, and were in turn classified by country, continent, category, and format. In order to utilize the relevant subsets of these data in our correlational analysis, we obtained data on infectious disease outbreaks from the Diseases Web Statistics System implemented by the Ministry of Health and Welfare and the KCDC. The Diseases Web Statistics System contains statistical data pertaining to NNIDs collected from medical establishments across the nation. It offers various statistics by disease type, area, year, epidemiological characteristics, and samples. For each NNID, we extracted the number of patients by year and area from 2001 to 2013. For 2013, the year with the most recent data, we organized the number of patients newly diagnosed with each infectious disease by month, week, area, age, gender, and occupation. Parameters involving population, health, traffic, and weather were obtained as data on the average age and household size (as of December 2013) organized by lower level administrative districts (such as eub, myun, or dong), from the public data portal run by the Ministry of Government Administration and Home Affairs, the number of hospitals by type and location (2009 to 2013) from the Korean Statistical Information Service, the population by district and age group, the population density by administrative district from the 2010 Census, and influenza immunization trends by age group from the Korea National

Health and Nutrition Examination Survey. Meteorological factors Korea evaluated by obtaining Examination data regarding the average temperatures (℃), high and low temperatures (℃), precipitation (mm), and relative humidity (%) from 10 major cities in 2013 from the National Weather Service's Korean peninsula weather statistics. Monthly air pollution was assessed by obtaining data regarding $PM_{2.5}$ levels (μg/m^3), sulfur dioxide (SO_2) gas levels (parts per million [ppm]), ozone (O_3) levels (ppm), nitrogen dioxide (NO_2) levels (ppm), and CO levels (ppm) by city and province in 2013 from the environmental statistics portal of the Ministry of the Environment. Population data and immunization rates by disease, as reported by public health care centers and private medical establishments nationwide, were obtained for the 25 administrative districts in Seoul from Seoul Open Data Square. Finally, the number of hospitals and hospital beds available in Seoul were obtained from medical center statistics.

Methods

In order to construct a local database using available public data, we collected 660 000 pieces of public data belonging to 26 categories, and converted them to the input file format of the MySQL database management system. JAVA and JSP were used to manipulate the collected data sets, including open API data. This database is publicly available at http://147.47.72.71:8080/opendata/information.html. Figure 1 provides a schematic view of the data collection and analysis procedure implemented in this study.

Our study concerned cases of NNIDs reported in 2013. These diseases are classified into four legally reportable communicable disease groups (1-4). We performed a correlational analysis to identify factors associated with the total number of newly diagnosed patients reported each month from January to December 2013, excluding infectious diseases that affected fewer than five patients in 2013. The environmental variables used in our analysis were monthly weather data reported in 2013 and average monthly $PM_{2.5}$, SO_2, O_3, NO_2, and CO levels in 10 major cities in Korea. Supplemental Table 1 presents descriptive statistics regarding these variables. The monthly air pollution level is potentially affected by meteorological factors; therefore, the association between these two categories was also used in our data interpretation. Overall population data and immunization records from districts in Seoul for major infectious diseases, such as diphtheria, tetanus, typhus, tuberculosis, and influenza, were used for our analysis of epidemiological dynamics in Seoul. The total numbers of hospitals and

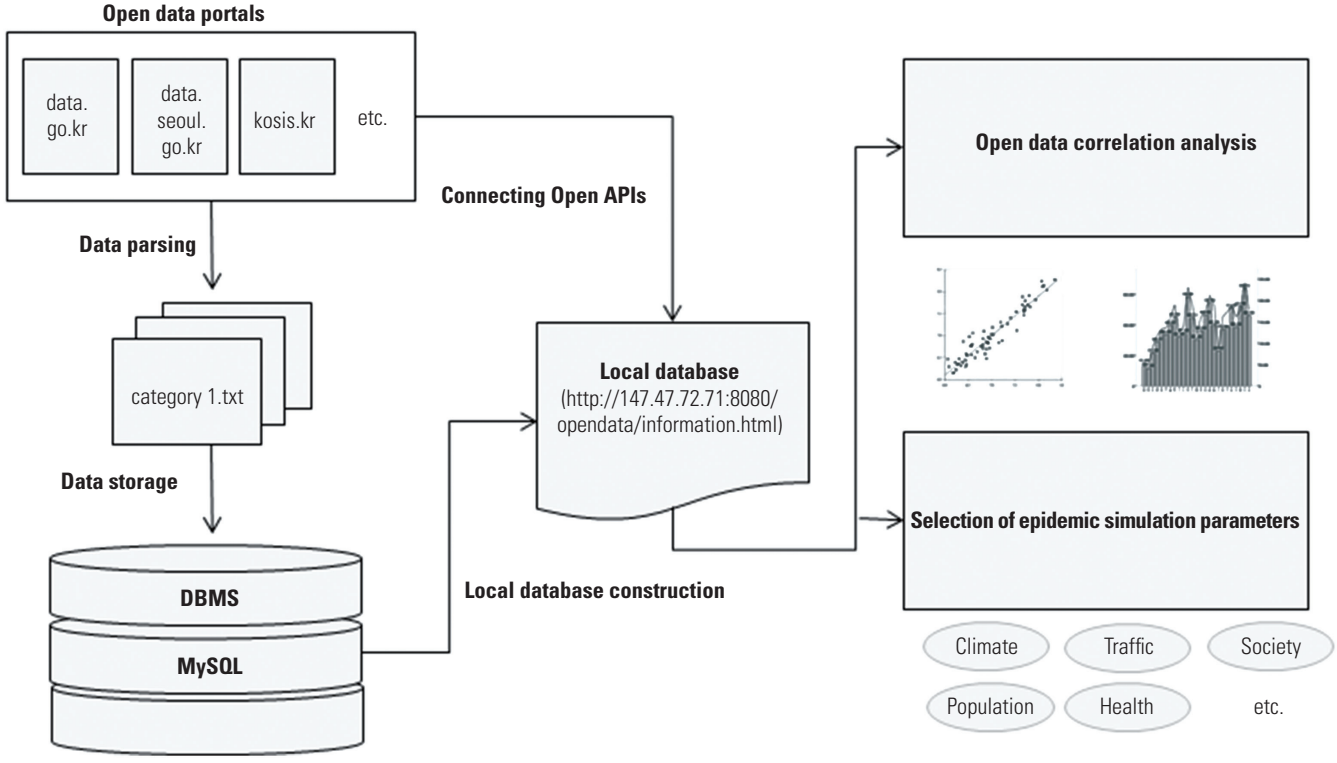


Figure 1. Schematic view of data collection, database construction, and data analysis.

Table 1. Correlational analysis between the monthly incidence of nationally notifiable infectious diseases and monthly average temperature, humidity, and air pollution-related factors in Korea during 2013

Nationally notifiable infectious diseases	Monthly Tem factors (Avg)			Monthly humidity factors (Avg)		Monthly air pollution-related factors (Avg)				
	Tem (°C)	High Tem (°C)	Low Tem (°C)	RF (mm)	RH (%)	$PM_{2.5}$ (μg/m^3)	SO_2 (ppm)	O_3 (ppm)	NO_2 (ppm)	CO (ppm)
Group 1										
Typhoid fever	-0.163	-0.166	-0.155	-0.081	-0.113	0.670*	0.596*	0.301	0.376	0.388
Paratyphoid fever	0.484	0.480	0.485	0.135	0.287	-0.131	-0.263	0.421	-0.365	-0.303
Bacillary dysentery	-0.332	-0.355	-0.314	-0.099	0.073	0.163	0.204	-0.604*	0.354	0.464
Hepatitis A	0.117	0.131	0.114	0.481	0.019	0.204	0.025	0.527	-0.177	-0.245
EHEC	0.696*	0.679*	0.705*	0.208	0.548	-0.553	-0.574	0.320	-0.617*	-0.499
Group 2										
Pertussis	-0.008	-0.003	-0.018	-0.226	-0.008	-0.209	-0.225	-0.398	0.008	0.035
Tetanus	0.390	0.362	0.411	0.720**	0.670*	-0.447	-0.482	-0.001	-0.518	-0.330
Measles	0.549	0.577*	0.524	-0.040	0.111	-0.261	-0.275	0.559	-0.306	-0.448
Mumps	0.207	0.202	0.203	0.032	0.300	-0.268	-0.255	-0.225	-0.105	-0.035
Rubella	0.590*	0.570	0.605*	0.634*	0.637*	-0.513	-0.516	0.266	-0.589*	-0.454
Japanese encephalitis	0.268	0.271	0.250	-0.078	0.103	-0.638*	-0.538	-0.158	-0.366	-0.318
Chicken pox	-0.401	-0.405	-0.396	-0.233	-0.131	0.533	0.561	-0.335	0.600*	0.610*
Acute hepatitis B	-0.040	-0.076	-0.009	0.434	0.445	-0.125	-0.141	-0.456	-0.049	0.115
Mother hepatitis B	-0.340	-0.342	-0.341	-0.185	-0.113	0.063	0.049	-0.560	0.286	0.314
Pp hepatitis B	-0.426	-0.444	-0.414	-0.052	-0.030	0.059	0.085	-0.655*	0.305	0.372
Group 3										
Malaria	0.949**	0.934**	0.957**	0.636*	0.767**	-0.673*	-0.749**	0.591*	-0.884**	-0.753**
Scarlet fever	-0.397	-0.391	-0.401	-0.264	-0.198	0.421	0.461	-0.311	0.546	0.485
Legionellosis	0.401	0.394	0.399	0.403	0.319	-0.493	-0.443	0.384	-0.468	-0.464
Endemic typhus	-0.161	-0.143	-0.181	-0.251	-0.091	0.024	0.005	-0.460	0.230	0.234
Scrub typhus	-0.125	-0.108	-0.148	-0.212	-0.107	-0.264	-0.257	-0.458	0.046	0.003
Leptospirosis	0.248	0.270	0.214	-0.213	0.029	-0.647*	-0.579*	-0.235	-0.307	-0.346
Brucellosis	0.453	0.457	0.443	0.531	0.374	-0.383	-0.495	0.414	-0.535	-0.524
Leprosy	-0.008	0.022	-0.022	-0.266	-0.257	0.518	0.444	0.333	0.346	0.183
CJD	0.171	0.190	0.153	-0.143	-0.201	-0.088	-0.008	0.504	-0.124	-0.235
Primary syphilis	0.236	0.230	0.231	0.115	0.378	-0.365	-0.382	-0.308	-0.205	-0.085
Secondary syphilis	0.630*	0.614*	0.634*	0.436	0.663*	-0.611*	-0.650*	0.062	-0.646*	-0.467
Congenital syphilis	-0.194	-0.211	-0.183	0.097	0.093	0.013	0.052	-0.464	0.133	0.277
V. vulnificus sepsis	0.529	0.521	0.522	0.131	0.369	-0.783**	-0.700*	0.018	-0.628*	-0.510
M meningitis	-0.022	0.015	-0.049	-0.130	-0.304	0.388	0.257	0.359	0.188	0.000
HFRS	-0.056	-0.041	-0.079	-0.215	-0.030	-0.364	-0.319	-0.506	-0.008	-0.029
Group 4										
Dengue fever	0.690*	0.654*	0.711**	0.437	0.726**	-0.739**	-0.738**	0.109	-0.776**	-0.518
Q fever	0.682*	0.657*	0.703*	0.420	0.662*	-0.608*	-0.663*	0.148	-0.682*	-0.526
Lyme disease	0.317	0.301	0.319	0.128	0.323	-0.425	-0.421	-0.090	-0.312	-0.184
SFTS	0.797**	0.782**	0.808**	0.723**	0.777**	-0.505	-0.635*	0.438	-0.737**	-0.595*

Data were obtained from the Public Data Portal (www.data.go.kr), Diseases Web Statistics System (http://is.cdc.go.kr/nstat/index.jsp), the Korean Statistical Information Service (http://kosis.kr), the Korea Meteorological Administration (http://sts.kma.go.kr) and the Ministry of the Environment (http://stat.me.go.kr). Each value is the Pearson's correlation coefficient (r).

Avg, average; Tem, temperature; RF, rainfall; RH, relative humidity; $PM_{2.5}$, fine particulate matter; SO_2, sulfur dioxide; O_3, ozone; NO_2, nitrogen dioxide; CO, carbon monoxide; ppm, parts per million; EHEC, enterohemorrhagic *Escherichia coli* infection; CJD, Creutzfeldt-Jakob disease; Pp hepatitis B, perinatal hepatitis B; *V. vulnificus* sepsis, *Vibrio vulnificus* sepsis; M meningitis, meningococcal meningitis; HFRS, hemorrhagic fever with renal syndrome; SFTS, severe fever with thrombocytopenia syndrome.

*$p<0.05$, **$p<0.01$.

hospital beds by district were used for the correlational analysis of population density, immunization rates, and medical resources available. In addition to climate and air quality, age, gender, occupation, and various other factors may also affect the incidence of infectious diseases. However, our study incorporated only the data currently available, thus excluding factors that were not available for analysis. Additionally, only factors relevant for districts of Seoul were incorporated into the correlational analysis of population density with the infectious disease incidence rate, the immunization rate, and the availability of medical resources. SPSS version 22.0 (IBM Corp., Armonk, NY, USA) was used for Pearson correlation analysis. A p-value <0.05 was considered to indicate statistical significance.

RESULTS

Status of Domestic and International Public Data

Many countries around the world are pursuing open government policies and declassifying public data, with the hope of facilitating economic growth. Major countries continue to expand the public availability of data through policies and web portals, including Canada, the US, the UK, Singapore, Australia, New Zealand, and Germany. Supplemental Table 2 shows the number of public databases available by data category, nation, and continent. Canada provides the most databases as of October 13, 2014 (209 250), followed by the US (157 586), the UK (19 628), Japan (12 347), Korea (11 433), and Singapore (8275). North America (the US and Canada) is the continent that accounts for the majority of available databases (85%), followed by Asia (Japan, Korea, and Singapore) and Europe (the UK, Germany, and Spain), which provides 7000 fewer datasets than Asia, with a total of 25 000. Economic data account for most of the public data currently available, followed by data relating to the environment, education, society, industry, demographics, health, culture, and energy. Data on animals, geology, and soil science were less frequently available. Public data are provided in the XML, JSON, HTML, KML, PDF, CSV, XLS, and DOC file formats. Of these, the XML and JSON formats are the most frequently available. Figure 2 shows the availability of open data, classified by country, categories, and data format, in fifteen countries.

Correlations Between Communicable Diseases and Categories of Public Data

The correlation coefficients of the monthly numbers of pa-

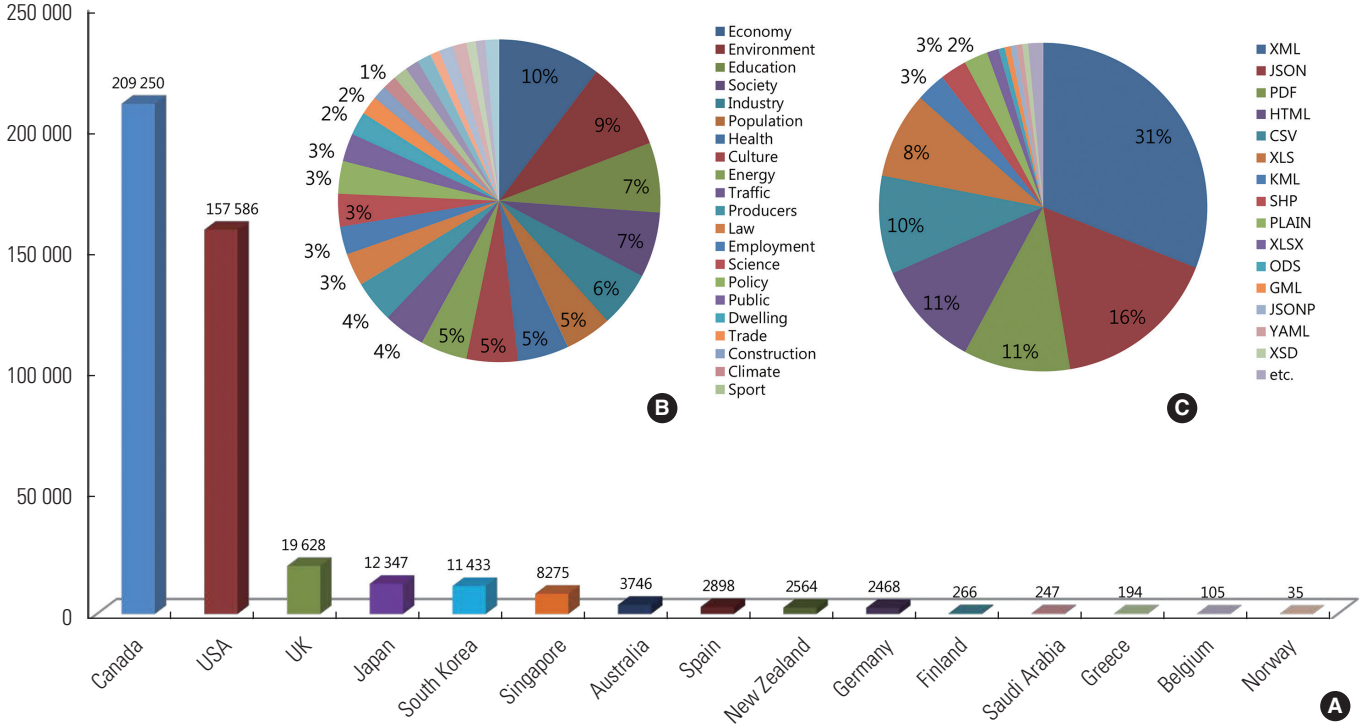


Figure 2. Sources of open data classified by country, categories, and data format. (A) Open data sources classified by country, (B) percentages of various categories of open data, and (C) percentages of various formats of open data. Source from open data portal site in each country [cited 2014 Oct 13].

tients of various NNIDs with average values of temperature, humidity, and air pollution parameters are shown in Table 1. The incidence rates of the following NNIDs showed a strong positive correlation, with a high confidence interval, between the average monthly temperature and the monthly number of patients: enterohemorrhagic *Escherichia coli* infections (r=0.696, p=0.012), malaria (r=0.949, p=0.000), secondary syphilis (r=0.630, p=0.028), dengue fever (r=0.690, p=0.013), Q fever (r=0.682, p=0.015), and severe fever with thrombocytopenia syndrome (r=0.797, p=0.002). Similar results were obtained when assessing these correlations in terms of average high and low temperatures. In particular, the number of malaria patients had the highest correlation with average monthly low temperatures (r=0.957, p=0.000) and average monthly high temperatures (r=0.934, p=0.000). The incidence rates of the following NNIDs showed a significant positive relationship with the average monthly rainfall: tetanus (r=0.720, p=0.008), rubella (r=0.634, p=0.027), malaria (r=0.636, p=0.026), and severe fever with thrombocytopenia syndrome (r=0.723, p=0.008). Positive correlations were also found between the average monthly relative humidity and the incidence of the following NNIDs: tetanus (r=0.670, p=0.017), rubella (r=0.637, p=0.026), malaria (r=0.767, p=0.004), secondary syphilis (r=0.663, p=0.019), dengue fever (r=0.726, p=0.008), Q fever

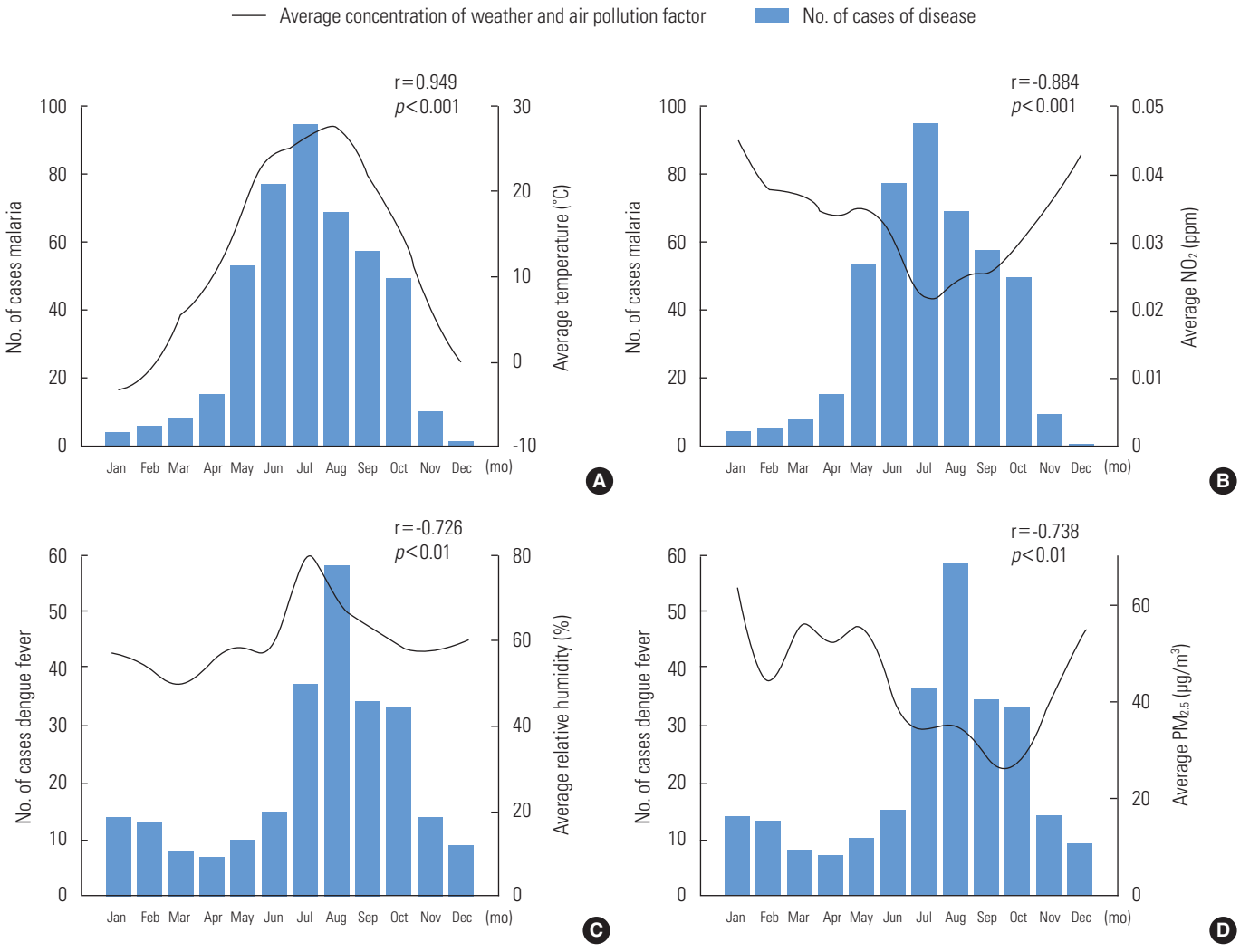


Figure 3. Strong correlations between the monthly incidence of selected infectious diseases and monthly average values of climactic and air pollution-related factors. (A) The monthly number of cases of malaria and average temperature. (B) The monthly number of cases of malaria and average nitrogen dioxide (NO_2) levels. (C) The monthly number of cases of dengue fever and average relative humidity levels. (D) The monthly number of cases of malaria and the average fine particulate matter ($PM_{2.5}$) level. Data were extracted from the Public Data Portal (www.data.go.kr) and Seoul City Open Data Square (http://data.seoul.go.kr).

(r=0.662, p=0.019), and severe fever with thrombocytopenia syndrome (r=0.777, p=0.003).

The incidence of several NNIDs was found to be related to the monthly average concentration of several air pollution parameters. Significant negative correlations were found between the incidence of the following NNIDs and the monthly average concentration of particulate matter: Japanese encephalitis (r=-0.638, p=0.026), malaria (r=-0.673, p=0.017), *Vibrio vulnificus* sepsis (r=-0.783, p=0.003), leptospirosis (r=-0.647, p=0.023), dengue fever (r=-0.738, p=0.006), and Q fever (r=-0.608, p=0.036). The incidence of malaria (r=-0.749, p=0.005), *V. vulnificus* sepsis (r=-0.700, p=0.011), secondary syphilis (r=-0.650, p=0.022), dengue fever (r=-0.739, p=0.006), Q fever (r=-0.663, p=0.019), and severe fever with thrombocytopenia syndrome (r=-0.635, p=0.027) showed a strong negative correlation coefficient with the monthly average concentration of SO_2 gas. The monthly average concentration of O_3 was positively correlated with the incidence of malaria (r=0.591, p=0.043) and negatively correlated with the incidence of bacillary dysentery (r=-0.604, p=0.037) and perinatal hepatitis B (r=-0.655, p=0.021). The incidence of the following NNIDs was significantly negatively correlated with the monthly average concentration of NO_2: enterohemorrhagic *E. coli* infections (r=-0.617, p=0.033), chickenpox (r=-0.600, p=0.039), malaria (r=-0.884, p=0.000), *V. vulnificus* sepsis (r=-0.628, p=0.029), secondary syphilis (r=-0.646, p=0.023), dengue fever (r=-0.776, p=0.003), Q fever (r=-0.682, p=0.015), and severe fever with thrombocytopenia syndrome (r=-0.737, p=0.006).

Figure 3 illustrates the correlations between climate and air pollution factors with the monthly incidence of NNIDs (p<0.01). Among the meteorological variables, the monthly average temperature (line graph) showed the highest correlation coefficient (r=0.949) to the monthly number of malaria patients (Figure 3A). The monthly average concentration of NO_2 (line graph), which showed a negative correlation with malaria (r=-0.884), had its lowest levels in July (0.022 ppm), corresponding to the highest number of patients (Figure 3B). The number of patients with dengue fever (bar graph) was elevated from July to September, as was the average relative humidity (Figure 3C). The highest concentration of $PM_{2.5}$ was 63.88 μg/m^3 in January, and this parameter was negatively correlated with malaria (Figure 3D).

Table 2 shows Pearson's correlation coefficients between the incidence rates of NNIDs and hospital resources in 25 districts of Seoul. A positive correlation was found between the incidence rates of NNIDs per 1000 person-years and the number of hospitals across 25 districts of Seoul (r=0.606, p=0.001). Moreover, the number of hospital beds in these districts showed a positive correlation (r=0.456, p=0.022) with hospital resources. In contrast, the incidence rate of NNIDs was negatively correlated with vaccination rates (r=-0.049). The number of hospital

Table 2. Pearson's correlation coefficients between the incidence rate of nationally notifiable infectious diseases and hospital resources among the 25 districts of Seoul during 2013

Variables	I	II	III	IV
Incidence rate per 1000 person-years (I)				
Pearson correlation	1	-0.049	0.606**	0.456*
Significance (two-tailed)		0.816	0.001	0.022
Vaccination rate (II)				
Pearson correlation	-0.049	1	-0.220	-0.079
Significance (two-tailed)	0.816		0.292	0.707
No. of hospitals (III)				
Pearson correlation	0.606**	-0.220	1	0.599**
Significance (two-tailed)	0.001	0.292		0.002
No. of beds (IV)				
Pearson correlation	0.456*	-0.079	0.599**	1
Significance (two-tailed)	0.022	0.707	0.002	

Data were obtained from the Public Data Portal (www.data.go.kr) and Seoul City Open Data Square (http://data.seoul.go.kr). Each value is the Pearson's correlation coefficient (r). Incidence rate per 1000 person-years=(Number of cases of nationally notifiable infectious diseases that occur in a population observed in Seoul during the year of 2013/Sum of all persons observed among those at risk during that period of time)×1000; Vaccination rate=(Number of cases of vaccinations for infectious diseases by district in Seoul during 2013)/Sum of all persons by district in Seoul during 2013).
*p<0.05, **p<0.01.

beds by patient in each district is shown in Supplemental Figure 1. Geumcheon-gu was found to have the highest number of hospital beds per capita (8.66) and a low incidence rate of NNIDs (23rd of the 25 districts). In contrast, Jongno-gu had a low number of hospital beds per capita (4.81) and the highest incidence rate of NNIDs (first of the 25 districts). Dongjak-gu had the lowest number of hospital beds per capita (2.47) and a high incidence rate of NNIDs (fourth of the 25 districts)

DISCUSSION

This study examined the accessibility of public data, both domestic and foreign, and the correlation between the following factors that affect the number of patients with infectious diseases in Korea: climate, air pollution, population density, vaccination rates, and healthcare resources. In accordance with the implementation of Government 3.0 policies, more Korean government agencies, such as the Ministry of Education; the Ministry of Health and Welfare; the Ministry of Environment; the Ministry of Land, Infrastructure, and Transport; and the Ministry of Agriculture, Food, and Rural Affairs, are making an increasing amount of data publicly accessible. According to a comparison of the categories of data available from each country, the Korean government has the fifth largest amount of public data and is the second largest provider of public data in Asia after Japan. However, this only applies to the number of data categories, because public data portal sites only offer information about the categories of data, not the size of the datasets. Public data that can be leveraged for infectious disease predictions in Korea include patient statistics, population density, climactic and air conditions, and medical resources. Infectious diseases pose an ever-increasing risk, as the number of Ebola victims continues to rise without effective treatment; thus, epidemiological forecasting and the ability to estimate the spread rate are paramount in establishing public health policies. Parameters that form the basis of infectious disease forecasting simulation tools include, but are not limited to, local population by age, contact rate, attack rate, vaccination rate by age, local-level healthcare resources, and the number of school vacation days. Once accurate baseline data are entered into such tools, more precise and effective predictions of the spread of infectious diseases can be made. In addition, reverse forecasts can be made using patient numbers by entering coefficients that are related to the baseline parameters. For example, mosquitoes have a faster multiplication rate at higher temperatures, and the incidence rate of malaria is connected to changes in temperature. The incidence of malaria increases by 3.4% for every 1°C increase in the mean national temperature in Korea, which has four distinct seasons [11]. Similar studies have recently been published, such as one reporting a link between air pollution and increased mortality due to lung cancer and other cardiopulmonary diseases [12].

In the present study, the monthly patient statistics for NNIDs were examined to determine if the incidence of NNIDs was related to the following factors in Korea during 2013, based on the following available public data: mean temperature, mean high and low temperatures, precipitation, mean humidity, and the concentrations of $PM_{2.5}$, SO_2, O_3, NO_2, and CO. As has been reported previously, malaria showed a strong correlation with average temperature and humidity. Enterohemorrhagic *E. coli*, dengue fever, Q fever, and severe fever with thrombocytopenia syndrome had a higher incidence during the summer and a lower incidence during the winter, which is likewise due to changes in temperature and humidity. Therefore, it is reasonable to conclude that the aforementioned arthropod-borne diseases are also affected by climactic factors, since their carriers, such as mosquitoes and mites, have higher reproductive rates during the summer. This conclusion is in accordance with Hunter's finding that vector-borne diseases, such as malaria and dengue fever, have a higher incidence rate when the breeding conditions of the vectors, such as mosquitoes and mites, are favorable [13]. Mumps, a viral disease, showed higher incidence rates than other infectious diseases in December, which is a period of greater pathogen replication due to low temperature and humidity and because human immune systems are weakened. Tetanus was found to be linked more closely with mean precipitation than temperature, and showed the highest incidence during July. The incidence of typhoid was linked with higher concentrations of $PM_{2.5}$. The increased concentration of suspended solids in water after days with a high concentration of $PM_{2.5}$ can lead to an increased incidence of typhoid [14]. However, these results must be subjected to additional analysis, because this study did not adjust for confounding in the temperature data. *V. vulnificus* sepsis and malaria showed the highest incidence in July and August; during these two months, $PM_{2.5}$ concentrations were relatively low due to high amounts of precipitation, and the replication rates of those pathogens increased. The mean concentration of SO_2 was higher in the winter than in the summer, and was neg-

atively correlated with the incidence of malaria, *V. vulnificus* sepsis, and dengue fever. However, since the mean difference in SO_2 concentrations was at most 0.004 ppm, this correlation was not very meaningful from an analytical perspective. The concentrations of NO_2 and CO were highest between December and February due to increased fuel consumption, and were lower during the summer months, resulting in a negative correlation with the diseases that have the highest incidence during summer, such as malaria, dengue, and Q fever. With this in mind, the incidence rates of many NNIDs showed a relationship with air pollution factors because this study did not adjust for confounding factors (temperature, humidity, and season). However, the incidence of typhoid fever and bacillary dysentery showed a correlation with air pollution factors due to the effect of particulate matter on the contamination of water [14]. Therefore, we can predict that air pollution and the incidence rates of NNIDs are indirectly correlated with a number of external factors, such as climactic and environmental variables.

This study also examined the relationship between the incidence rates of NNIDs, vaccination rates, and the number of clinics and hospital beds across 25 districts in Seoul, Korea, and also aimed to identify whether levels of preparedness regarding hospital resources affected the incidence rates. A meaningful positive correlation was found between the incidence rates of NNIDs and the number of clinics and hospital beds. However, no correlation was observed between the incidence rates of NNIDs and vaccination rates or the number of clinics and hospital beds, suggesting that districts with higher incidence rates and more clinics and hospital beds do not necessarily have higher vaccination rates. This means that each district should prepare hospital resources appropriately during outbreaks of NNIDs. Geumcheon-gu was found to have the lowest incidence rate of NNIDs and the highest number of hospital beds per patient, and Jongno-gu had the highest incidence rate and a low number of hospital beds per patient. These findings facilitate the estimation of vaccine supply requirements and the availability of beds per capita during outbreaks of NNIDs.

A limitation of this study is that it collected public data from 2013 only. The dataset for assessing the relationships of the incidence rates of NNIDs with vaccination rates and healthcare resources was confined to Seoul, as Seoul had the largest amount of public health data available. Moreover, we did not adjust for confounding factors (temperature, season, disease lag time), so it is difficult to interpret causal relationships in the correlations between the monthly number of NNID patients and the parameters related to average temperature, humidity, and air pollution.

This study surveyed the availability of public data (domestic and international), performed a correlational analysis of the incidence of NNIDs with climactic and air pollution-related parameters, and performed a correlational analysis of the incidence of NNIDs and population density, immunization rates, the number of hospitals, and the number of hospital beds by district in Seoul. According to the survey of the availability of public data from various countries, the Korean government provides the fifth largest amount of public data and is the second largest provider of public data in Asia after Japan. It is therefore possible to obtain analytic results in public health using these open data, as well as to identify correlations between incidence rates and related factors, such as environmental and social data, using a long period of observation of infectious outbreaks with access to more public data from related governments. Our correlational analysis, based on public data, may facilitate forecasts of the effects of temperature and humidity on the incidence rates of NNIDs. Moreover, some districts in Seoul have more hospital resources that can be mobilized during outbreaks of infectious diseases. The identification of other factors that can be used as forecast parameters in similar studies will be of value in developing highly reliable and customized infectious disease spread simulation models for Korea.

ACKNOWLEDGEMENTS

This study was conducted by using the open-source public data from Korean governmental agencies, so we give special thanks to all the governmental institutions that made this possible.

CONFLICT OF INTEREST

The authors have no conflicts of interest with the material presented in this paper.

REFERENCES

1. Morse SS. Factors in the emergence of infectious diseases. Emerg Infect Dis 1995;1(1):7-15.
2. World Health Organization. Ebola response roadmap situation report; 2014 [cited 2014 Nov 3]. Available from: http://www.who.int/csr/resources/publications/ebola/response-roadmap/en/.

3. Xu L, Stige LC, Kausrud KL, Ben Ari T, Wang S, Fang X, et al. Wet climate and transportation routes accelerate spread of human plague. Proc Biol Sci 2014;281(1780):20133159.
4. Korea Centers for Disease Control and Prevention. Infectious disease surveillance yearbook, 2013. Cheongju: Korea Centers for Disease Control and Prevention; 2014, p. 19-32 (Korean).
5. Ministry of Security and Public Administration. DATA.GO.KR [cited 2014 Nov 7]. Available from: https://www.data.go.kr/main.jsp#/L21haW4= (Korean).
6. Zhang Y, Bi P, Hiller JE. Climate change and the transmission of vector-borne diseases: a review. Asia Pac J Public Health 2008;20(1):64-76.
7. O'Neill MS, Hajat S, Zanobetti A, Ramirez-Aguilar M, Schwartz J. Impact of control for air pollution and respiratory epidemics on the estimated associations of temperature and daily mortality. Int J Biometeorol 2005;50(2):121-129.
8. Atkins C, Meltzer MI. Community Flu 2.0; 2012 [cited 2014 Nov 7]. Available from: http://www.cdc.gov/flu/pandemic-resources/tools/communityflu.htm.
9. Zhang X, Meltzer MI, Wortley PM. FluSurge--a tool to estimate demand for hospital services during the next pandemic influenza. Med Decis Making 2006;26(6):617-623.
10. Centers for Disease Control and Prevention. FluAid 2.0; 2000 [cited 2014 Nov 7]. Available from: http://www.cdc.gov/flu/pandemic-resources/tools/fluaid.htm.
11. Chae SM, Kim DJ, Yoon SJ, Shin HS. The impact of temperature rise and regional factors on malaria risk. Health Soc Welfare Rev 2014;34(1):436-455 (Korean).
12. Lee BJ, Kim B, Lee K. Air pollution exposure and cardiovascular disease. Toxicol Res 2014;30(2):71-75.
13. Hunter PR. Climate change and waterborne and vector-borne disease. J Appl Microbiol 2003;94 Suppl:37S-46S.
14. Dewan AM, Corner R, Hashizume M, Ongee ET. Typhoid Fever and its association with environmental factors in the Dhaka Metropolitan Area of Bangladesh: a spatial and time-series approach. PLoS Negl Trop Dis 2013;7(1):e1998.

Supplemental Table 1. The number of domestic and international public datasets

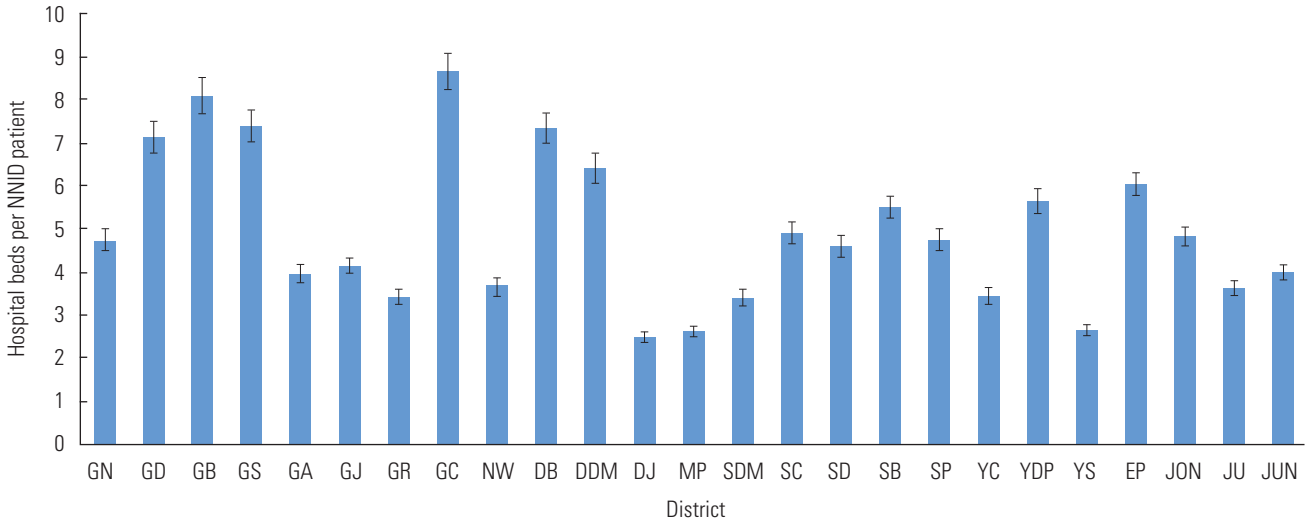
Ranking	Country	Site	No. of datasets	Format of data
1	Canada	data.gc.ca	209 250	SHP, XML, HTML, CSV, PDF, GEOTIF
2	USA	data.gov	157 586	CSV, HTML, XML, JSON, SOAP, XLS
3	United Kingdom	data.gov.uk	19 628	XML, JSON, KML, KMZ, HTML, CSV, VMS
4	Japan	data.go.jp	12 347	HTML, PDF, XLS, ZIP, CSV, DOC
5	Korea, Republic of	data.go.kr	11 433	XLS, HWP, REST, HTML, PDF, XML, JSON
6	Singapore	data.gov.sg	8275	CSV, TXT, URL, XLS, XML
7	Australia	data.gov.au	3746	ZIP, CSV, PLAIN, SHP, XLS, PDF, WMS, JSON
8	Spain	datos.gob.es	2898	PC-AXIS, CSV, XLS, HTML, JSON, SPARQL
9	New Zealand	data.govt.nz	2564	XLS, PDF, HTML, API, KML/SHP, XML
10	Germany	portalu.de	2468	XML, PDF, DOC, FLA, ZIP
11	Finland	suomi.fi	266	XML, JSON
12	Saudi Arabia	saudi.gov.sa	247	PDF, XLS, XLSX
13	Greece	geodata.gov.gr	194	XLS, ODS, SHP, GML, KML ODS, MAP
14	Belgium	data.gov.be	105	HTML, XHTML, XLS, XML, CSV
15	Norway	data.norge.no	35	HTML, CSV, PDF, JSON, JSONP, YAML

Obtained from each country's open data portal site [cited 2014 Oct 13].

Supplemental Table 2. Descriptive statistics of climactic and air pollution-related variables

Variables	Monthly average											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Average temperature (°C)	-3.4	-1.2	5.1	10	18.2	24.4	25.5	27.7	21.8	15.8	6.2	-0.2
Average high temperature (°C)	0.3	2.8	10.8	15	23.6	29.2	28.3	31.1	25.9	21.2	10.7	3.5
Average low temperature (°C)	-6.6	-4.9	0.7	5.7	13.7	20.5	23.4	24.8	18	10.9	2.1	-3.5
Rainfall (mm)	22.1	74.1	27.3	71.7	132	28.3	676.2	148.6	138.5	13.5	46.8	24.7
Relative humidity (%)	57	54	49	54	58	60	79	69	63	58	58	60
$PM_{2.5}$ ($\mu g/m^3$)	63.88	44.52	55.36	52.24	55.33	40.12	34.16	35	28.08	29.16	42.56	55.04
SO_2 (ppm)	0.008	0.006	0.006	0.006	0.006	0.005	0.004	0.004	0.004	0.004	0.005	0.007
O_3 (ppm)	0.012	0.018	0.023	0.031	0.034	0.033	0.024	0.027	0.023	0.017	0.012	0.009
NO_2 (ppm)	0.045	0.038	0.037	0.034	0.035	0.03	0.022	0.024	0.026	0.03	0.036	0.043
CO (ppm)	0.9	0.6	0.5	0.5	0.5	0.4	0.4	0.4	0.4	0.4	0.6	0.8

Data were obtained from the Korea Meteorological Administration (http://sts.kma.go.kr).
$PM_{2.5}$, fine particulate matter; SO_2, sulfur dioxide; O_3, ozone; NO_2, nitrogen dioxide; CO, carbon monoxide; ppm, parts per million.



Supplemental Figure 1. The number of hospital beds per NNID patient by district in Seoul. Data were extracted from the Public Data Portal (www.data.go.kr) and Seoul City Open Data Square (http://data.seoul.go.kr). GN, Gangnam-gu; GD, Gangdong-gu; GB, Gangbuk-gu; GS, Gangseo-gu; GA, Gwanak-gu; GJ, Gwangjin-gu; GR, Guro-gu; GC, Geumcheon-gu; NW, Nowon-gu; DB, Dobong-gu; DDM, Dongdaemun-gu; DJ, Dongjak-gu; MP, Mapo-gu; SDM, Seodaemun-gu; SC, Seocho-gu; SD, Seongdong-gu; SB, Seongbuk-gu; SP, Songpa-gu; YC, Yangcheon-gu; YDP, Yeongdeungpo-gu; YS: Youngsan-gu; EP, Eunpyeong-gu; JoN, Jongno-gu; Ju, Jung-gu; JuN, Jungnang-gu; NNID, nationally notifiable infectious disease.