JMIR Publications
The Leading eHealth Publisher

Infodemiological study to understand the community risk perceptions of COVID-19 outbreak in South Korea

Atina Husnayain, Eunha Shim, Anis Fuad, Emily Chia-Yu Su

Submitted to: Journal of Medical Internet Research
on: May 02, 2020

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Infodemiological study to understand the community risk perceptions of COVID-19 outbreak in South Korea

Atina Husnayain[1, 2] MPH; Eunha Shim[3] PhD; Anis Fuad[1] DEA; Emily Chia-Yu Su[2, 4, 5] PhD

[1]Department of Biostatistics, Epidemiology, and Population Health Faculty of Medicine, Public Health and Nursing Universitas Gadjah Mada Yogyakarta ID
[2]Graduate Institute of Biomedical Informatics College of Medical Science and Technology Taipei Medical University Taipei TW
[3]Department of Mathematics Soongsil University Seoul KR
[4]Clinical Big Data Research Center Taipei Medical University Hospital Taipei TW
[5]Research Center for Artificial Intelligence in Medicine Taipei Medical University Taipei TW

Corresponding Author:
Emily Chia-Yu Su PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
172-1 Keelung Rd, Sec 2
Taipei
TW

Abstract

Background: South Korea is among the best-performing countries to tackle the coronavirus pandemic utilizing massive drive-through tests and facemasks, as well as extensive social distancing. However, understanding the patterns of risk perception could also facilitate effective risk communication to minimize the impact of disease spread during crisis.

Objective: We aimed to explore the patterns of community health risk perception of coronavirus disease 2019 (COVID-19) in South Korea using Internet search data.

Methods: Google and NAVER relative search volume (RSV) data were collected using COVID-19-related terms in Korean language and retrieved according to time, gender, age groups, type of devices, and location. Online queries were compared with the number of new COVID-19 cases and tests on a daily basis recorded in the Kaggle open access data set by Joong Kun Lee and colleagues. Spearman's rank correlation coefficients were employed to assess whether correlations between new COVID-19 cases and Internet searches were affected by time.

Results: The number of COVID-19-related queries in South Korea increased during the local events including the local transmission, approval of coronavirus test kits, implementation of coronavirus drive-through tests, facemask shortage, and widespread campaign for social distancing as well as during international events such as the announcement of Public Health Emergency of International Concern (PHEIC). Online queries were also higher in women (r-0.763?0.823; p<0.05), age groups of ?29 (r-0.726?0.821; p<0.05), 30–44 (0.701?0.826; p<0.05), and ?50 (0.706?0.725; p<0.05). In terms of spatial distribution, Google and NAVER RSV were higher in affected areas. Moreover, greater correlations were found in mobile searches (0.704?0.804; p<0.05) compared to that of desktop searches (0.705?0.717; p<0.05), indicating the changing behavior in searching health online information during outbreak. Those varied Internet searches related to COVID-19 represented the community health risk perception. In addition, as country with high number of coronavirus tests, results showed that adults perceived the coronavirus test-related information as more important than disease-related knowledge. Meanwhile, the younger and older age groups have a different perception, making the infection-related information among the essential searches.

Conclusions: The use of both Google and NAVER RSV to explore the patterns of community health risk perception could be beneficial for targeting risk communication in several perspectives including time, population characteristics, and location.

(JMIR Preprints 02/05/2020:19788)
DOI: https://doi.org/10.2196/preprints.19788

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Infodemiological study to understand the community risk perceptions of COVID-19 outbreak in South Korea

Abstract

Background: South Korea is among the best-performing countries in tackling the coronavirus pandemic by utilizing mass drive-through testing, facemasks use, and extensive social distancing. However, understanding the patterns of risk perception could also facilitate effective risk communication to minimize the impacts of disease spread during this crisis.
Objective: We attempted to explore patterns of community health risk perceptions of COVID-19 in South Korea using Internet search data.
Methods: Google Trends (GT) and NAVER relative search volumes (RSVs) data were collected using COVID-19-related terms in the Korean language and were retrieved according to time, gender, age groups, types of device, and location. Online queries were compared to the number of daily new COVID-19 cases and tests reported in the Kaggle open-access dataset for time period of December 5, 2019 to May 31, 2020. Spearman's rank correlation coefficients were employed to assess whether correlations between new COVID-19 cases and Internet searches were affected by time. We also constructed a prediction model of new COVID-19 cases using the number of COVID-19 cases, tests, GT, and NAVER RSVs in lag periods (of 3 to 1 days). Single and multiple regressions were employed using backward elimination and a variance inflation factor (VIF) of <5.
Results: Numbers of COVID-19-related queries in South Korea increased during local events including local transmission, approval of coronavirus test kits, implementation of coronavirus drive-through tests, a facemask shortage, and a widespread campaign for social distancing as well as during international events such as the announcement of a Public Health Emergency of International Concern by the World Health Organization. Online queries were also stronger in women (r=0.763~0.823; p<0.05), and age groups of ≤29 (r=0.726~0.821; p<0.05), 30~44 (r=0.701~0.826; p<0.05), and ≥50 years (r=0.706~0.725; p<0.05). In terms of spatial distribution, GT and NAVER RSVs were higher in affected areas. Moreover, greater correlations were found in mobile searches (r=0.704~0.804; p<0.05) compared to those of desktop searches (r=0.705~0.717; p<0.05), indicating changing behaviors in searching for online health information during the outbreak. Those varied Internet searches related to COVID-19 represented community health risk perceptions. In addition, as a country with a high number of coronavirus tests, results showed that adults perceived coronavirus test-related information as being more important than disease-related knowledge. Meanwhile, younger and older age groups had different perceptions. Moreover, NAVER RSVs can potentially be used for health risk perception assessments and disease predictions. Adding COVID-19-related searches provided by NAVER could increase the performance of the model compared to that of the COVID-19 case-based model and potentially be used to predict epidemic curves.
Conclusions: The use of both GT and NAVER RSVs to explore patterns of community health risk perceptions could be beneficial for targeting risk communication from several perspectives, including time, population characteristics, and location.

Keywords: Google Trends; risk perception; risk communication; COVID-19; South Korea

Introduction

The World Health Organization (WHO) declared the coronavirus disease 2019 (COVID-19) outbreak a pandemic on March 11, 2020 (1). By May 31, 2020, the disease had infected 5,934,936 individuals worldwide (2) including 11,468 individuals in South Korea. The first COVID-19 case in South Korea was confirmed on January 20, 2020 (3). Slow upturns in disease transmission were reported before February 19, 2020; the huge local clusters observed in Daegu led to daily increases in the number of new cases (4). Numerous approaches were undertaken to prevent disease transmission, including coronavirus drive-through testing and social distancing (5, 6). Coronavirus drive-through tests were identified as a safe and efficient screening approach, with each test taking approximately 10 min, thus minimizing cross-infection among people being tested (6). To date, the average number of daily new cases is lower by ten-fold or more compared to those during the peak of the epidemic (from February 19 to March 15, 2020) (3). Consequently, South Korea is considered among the best-performing countries in tackling the pandemic.

On the contrary, adequate risk communication could also have helped minimize the impacts of disease transmission (7). Thus, in the pandemic period, the WHO suggests regular risk communication of updating any changes in the status of the pandemic to the public and stakeholders (8). This action might be challenging because proper risk communication needs a robust understanding of risk perceptions which helps to identify what knowledge the public needs (7). However, studies exploring risk perception are often conducted using survey methods or content analyses (7, 9-11), which require huge resources and longer time. In particular, when investigating an emerging disease, those approaches might be less affordable since the health system will be overburdened with the surge of health-care utilizations, thus resulting in more barriers to assessing community health risk perceptions.

Therefore, this study aimed to explore patterns of community health risk perceptions towards COVID-19 in South Korea using Internet search data. This study is part of infodemiological research that was first introduced in 1996 (12) and explores the distribution of information on the Internet (13) for public health and policy about the ground situation in the population. Infodemiology commonly deals with disease-related topics as well as outbreaks and epidemics (14). This approach can potentially be used since Internet query data can be provided easily, promptly, (15), and in a cost-effective manner compared to survey methods (16), and also it can potentially capture anomalous patterns in near-real-time (17).

In this analysis, we utilized COVID-19-related Internet search data provided by Google Trends (GT) and NAVER to represent online queries from the world's largest search engine and Korean local search engine which has a higher market share than Google in South Korea (18). This study explored patterns of public health risk perceptions towards the ongoing outbreak from several different perspectives, including time, population characteristics, and location as used in epidemiological studies. We also constructed a prediction model of new COVID-19 cases using the number of COVID-19 cases, tests, GT, and NAVER relative search volumes (RSVs) in lag period (of 3 to 1 days). Future studies are warranted to define the best lag period to perform effective risk communication in the early stages of a disease outbreak.

Methods

Datasets

The daily numbers of new COVID-19 cases and coronavirus tests from January 20 to May 31, 2020 were collected from the Kaggle open-access dataset by Jihoo Kim and colleagues (3). We used the Time.csv dataset to retrieve the number of new daily COVID-19 cases and daily tests and the TimeProvince.csv dataset to collect cumulative coronavirus cases by region. Those datasets covered all cities in South Korea. In addition, Internet search data related to COVID-19 were retrieved from GT (https://trends.google.com/) (19) and NAVER websites (https://datalab.naver.com/) (20) in the same collocation. The information searched was collected 6 weeks earlier from December 5, 2019, to explore patterns before the occurrence of the first COVID-19 case in South Korea. Data were collected using COVID-19-related terms, including coronavirus (□□□ □□□□), coronavirus test (□□□ □□□□ □□□), MERS (Middle Eastern respiratory syndrome) (□□ □), facemask (□□□), social distancing (□□□ □□□□), and Shinchoenji (□□□) in the Korean language, and data were retrieved according to time, gender, age groups, types of device, and location. These keywords were used to represent online information searches for COVID-19-related information, personal protective measures, and preventive approaches. Specific keywords for MERS (□□□) were used to assess whether there was an increase of information searches in the early stage of the outbreak using specific terms related to MERS as reported in previous research (21). In addition, the Shinchoenji (□□□) keyword was also used to collect online information searches following a huge cluster in the Shinchoenji church and to define whether this cluster induced a surge of online information searches. For more than one-word terms, quotes were used to increase the accuracy of data in both GT and NAVER as suggested in an earlier GT research framework (22). Health category and web search option for GT queries were also utilized.

Online search data retrieved from GT and NAVER are presented as a relative number called the RSVs that ranges from 0 to 100. The RSVs represents search requests made to those search engines. For GT, the RSVs for a specific term are normalized according to the corresponding time and location (23). GT RSVs can be downloaded for different times and locations (19) while NAVER provides queries for various times, genders, ages, and types of device categories (20).

Statistical analysis

Analyses of health risk perceptions toward COVID-19 were performed using data from January 20 to March 22, 2020. This time frame was selected since this study aimed to explore patterns of Internet searches representing health risk perceptions in the initial weeks of the outbreak. Data were analyzed in a single graphical form to explore trends in new COVID-19 cases, numbers of tests, and Internet searches on a daily basis. Time-lag correlations calculated by Spearman's rank correlation coefficients were employed to assess whether correlations of new COVID-19 cases with GT and NAVER RSVs were affected by time within 3 days of a lag or lead period. Statistical analyses were performed using STATA13, and strong correlations were defined as correlation coefficients (*r*) of >0.7. Moreover, multilayer maps created using Tableau Public 2020 were generated to define the distributions of new COVID-19 cases and Internet searches.

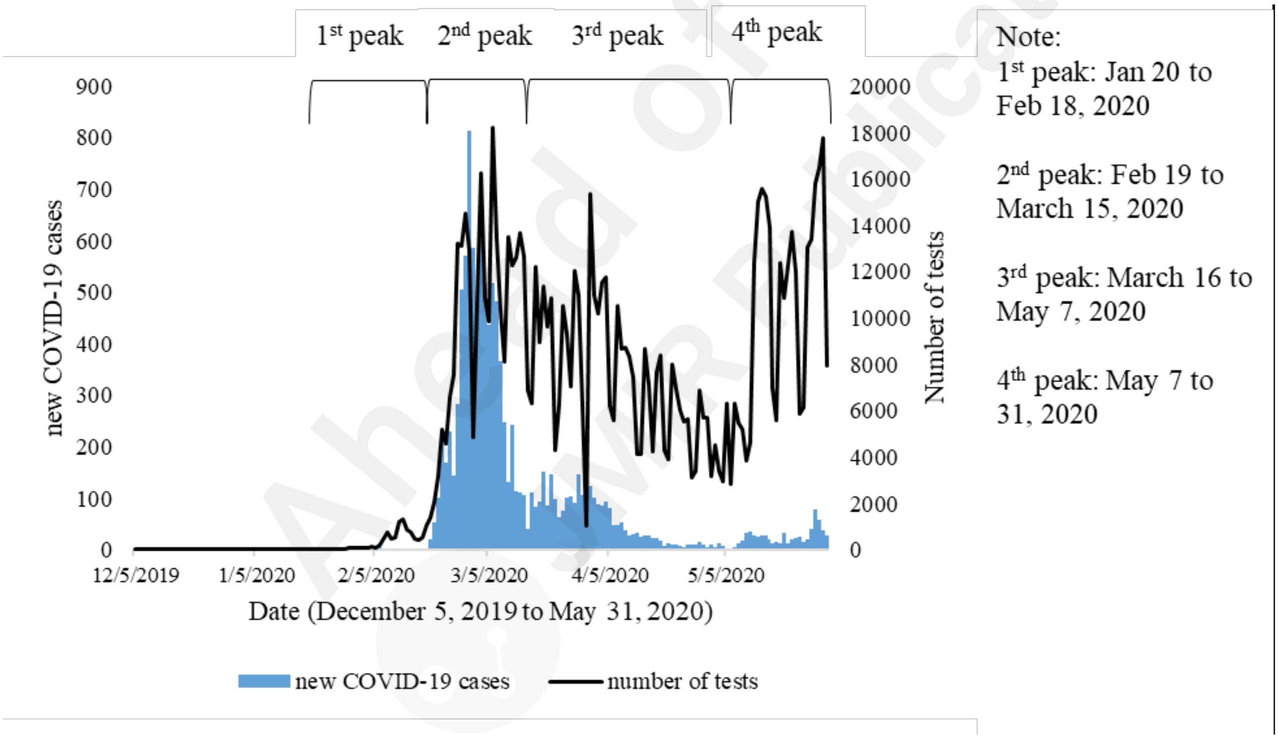
This study also undertook the task of predicting new COVID-19 cases. Several predictors, including the number of COVID-19 cases, tests, and GT and NAVER RSVs in lag periods (of 3 to 1 days) were used to predict the target variable which was the number of new COVID-19 cases. The prediction value was calculated using single and multiple regressions employing backward elimination and a variance inflation factor (VIF) of <5 in STATA13. A lower VIF level was considered to minimize the presence of multicollinearity in the model, particularly in epidemiologic studies (24). Models were constructed using the development dataset (January 20 to March 22, 2020) as used in health risk perception analyses and validated using the future validation dataset (March 23 to May 31, 2020). The root mean squared error (RMSE) was assessed for evaluating the models' performances, as well as Akaike information criterion (AIC) for selecting a correct model and Bayesian information criterion (BIC) for finding the best model for future predictions (25).

Results

Community health risk perceptions captured by GT and NAVER RSVs were divided into several parts, including patterns by time, population characteristics, and location.

Trends in new COVID-19 cases, number of tests, and Internet searches on a daily basis

South Korea reported the first case of COVID-19 on January 20, 2020, (Figure 1) with four peaks of disease transmissions as of May 31, 2020. The first peak occurred until February 18, 2020. The average new cases increased to 311 per day and dramatically decreased to 50 cases per day since March 16, 2020. The fourth peak was observed on May 8, 2020, which corresponded with implementation of a new normal starting on May 6, 2020 (26). Furthermore, as of May 31, 2020, South Korea had reported 11,468 cases of COVID-19. Large numbers of tests were also performed during the outbreak. South Korea performed 6,848 tests, on average per day from January 20 to May 31, 2020, and 910,822 tests in total, making South Korea as one of the countries with the highest number of tests performed.



Note:
1st peak: Jan 20 to Feb 18, 2020

2nd peak: Feb 19 to March 15, 2020

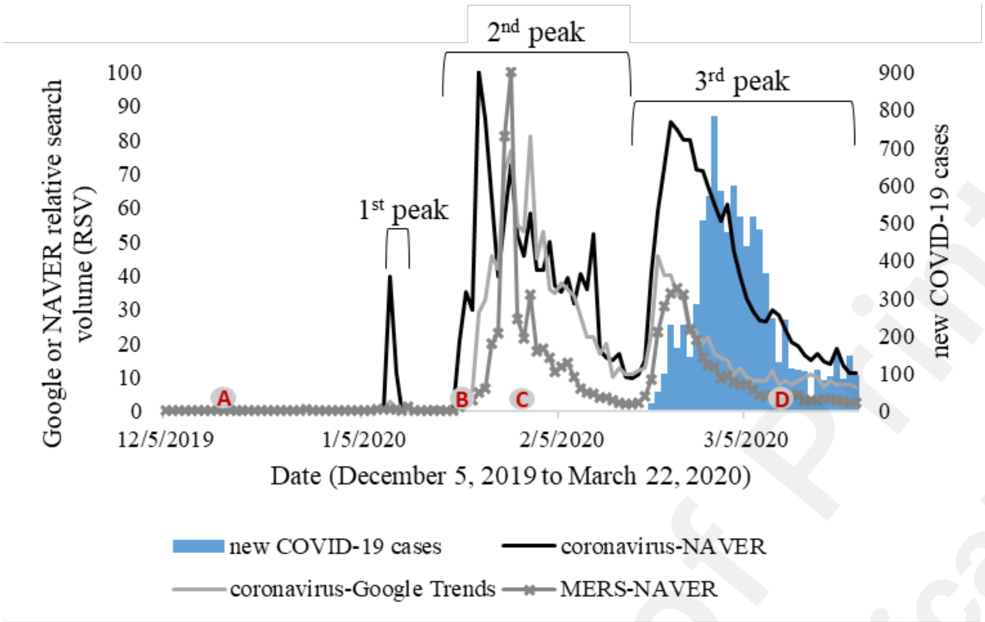
3rd peak: March 16 to May 7, 2020

4th peak: May 7 to 31, 2020

Figure 1. Time series of new COVID-19 cases and number of tests in South Korea.

During the outbreak, trends of information searches for coronavirus (□□□ □□□□) captured by GT and NAVER were similar (Figure 2). Three huge peaks of Internet searches were observed in the second and fifth weeks of January and in the fourth week of February 2020. Coronavirus-related searches remained high for several days after the first COVID-19 case was reported in Wuhan on December 12, 2019, along with MERS (□□ □)-related queries, which were also elevated in the last two peaks. However, massive surges of information searches occurred along with the identification of the first COVID-19 case in South Korea on January 20 and with the WHO's declaration of the Public Health Emergency of International Concern (PHEIC) on

January 30, 2020. Compared to the daily data on new COVID-19 cases, information searches provided by GT and NAVER peaked 7~9 days earlier. The third peak of coronavirus searches possibly corresponded to the immense increase in the number of new COVID-19 cases due to local transmission. Searches gradually decreased even after the outbreak was declared a pandemic by the WHO on March 11, 2020 (1).



Note:
A: 1st case is reported in Wuhan [Dec 12, 2019]

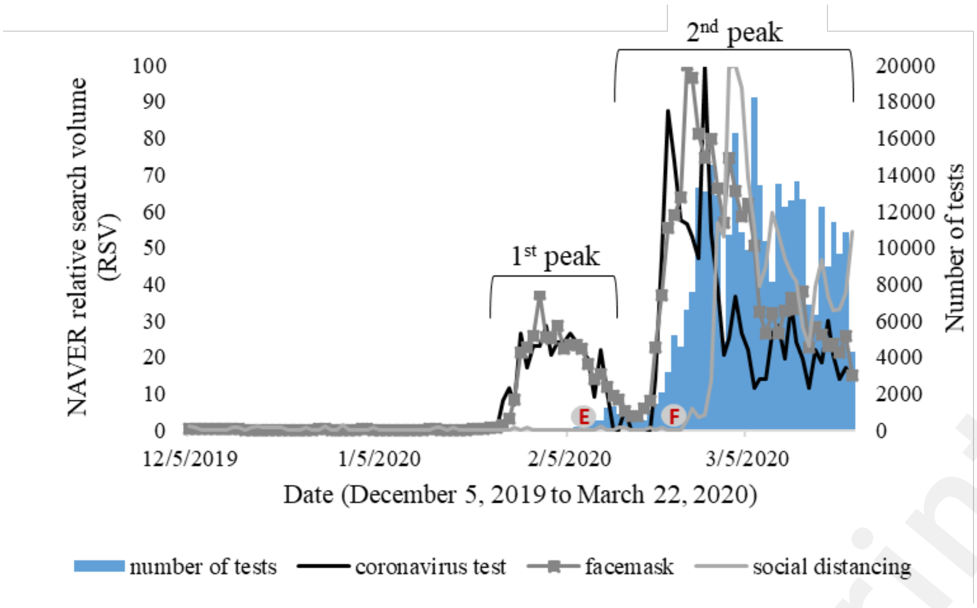
B: 1st case is reported in South Korea [Jan 20, 2020]

C: WHO declares the outbreak as Public Health Emergency of International Concern [Jan 30, 2020]

D: WHO declares the outbreak as pandemic [Mar 11, 2020]

Figure 2. Time series of new COVID-19 cases, Google Trends, and NAVER relative search volumes (RSVs) related to the coronavirus and Middle Eastern respiratory syndrome (MERS) in South Korea.

Furthermore, coronavirus test-related (□□□ □□□□ □□□) searches were not captured in GT; hence, Figure 3 only illustrates NAVER RSVs related to coronavirus tests, facemasks, and social distancing. Increases in Internet searches were observed weeks after the COVID-19 cases were reported and before a coronavirus test kit was approved on February 7, 2020 (27). The second wave of information searches was found in the third week of February 2020, which might have been caused by an increase in the number of new COVID-19 cases and the implementation of coronavirus drive-through tests on February 23, 2020 (6). However, patterns of coronavirus test-related searches seemed more similar to trends of new COVID-19 cases compared to the daily numbers of tests.

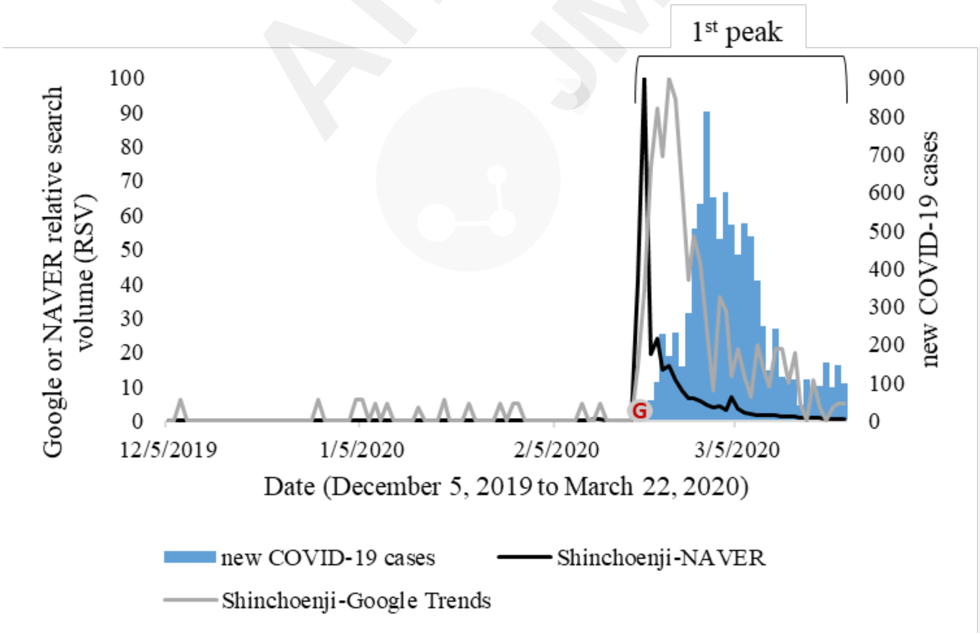


Note:
E: 1st coronavirus test kit is approved by South Korean CDC [Feb 7, 2020]

F: 1st coronavirus drive-through test is opened [Feb 23, 2020]

Figure 3. Time series of new COVID-19 cases and NAVER relative search volumes (RSVs) related to coronavirus tests, facemasks, and social distancing in South Korea.

Similar patterns of online queries about coronavirus tests were also identified for facemasks (□□□). From the perspective of personal protective measures, the number of facemask-related queries were increased in the same period when people began to search for coronavirus tests and facemask shortages in early February (28) and gradually declined in late February as a regular supply of facemasks was provided by the federal government (29). Moreover, the massive increase in locally acquired cases also induced huge internet searches related to social distancing (□□□ □□□□) as one of the preventive approaches. Those searches reached a peak as a widespread campaign for social distancing was commended in the first week of March 2020 in South Korea (5). In contrast, the number of Shinchoenji (□□□)-related searches increased as the Shinchoenji cluster had discovered on February 18, 2020 (30), and gradually decreased thereafter, even before the surge in new COVID-19 cases peaked on February 29, 2020 (Figure 4).



Note:
G: 31st case is reported and later confirmed to be related to the Shinchoenji cluster [Feb 18, 2020]

Figure 4. Time series of new COVID-19 cases, Google Trends, and NAVER relative search

volumes (RSVs) related to the Shinchoenji cluster in South Korea.

Time-lag correlations between new COVID-19 cases and Internet searches in different gender and age groups

The results in Table 1, demonstrated a moderate correlation (*r*=0.631) between new COVID-19 cases and GT RSVs related to coronavirus with a lag of 3 days. On the contrary, strong correlation (*r*=0.718) of coronavirus information searches counting for both men and women with a lag of 3 days showed no differences for NAVER RSVs. However, the correlations varied across different age groups and lag periods. Strong correlations were observed with a lag of 3 days for all ages (*r*=0.729), and those aged ≤18 years (*r*=0.821), 19~24 years (*r*=0.784), 25~29 years (*r*=0.726), 50~54 years (*r*=0.706), and ≥50 years (*r*=0.725). Meanwhile, the weakest correlation was found in the age group of 35~39 years (*r*=0.622). The ≤18 year and 19~24 year age groups for NAVER RSVs had strong correlations in almost all lag and lead periods. Moreover, the strength of the correlations decreased in the lead period or a few days after the number of new COVID-19 cases increased, for both GT and NAVER RSVs. Compared to NAVER RSVs, GT RSVs for coronavirus had weaker correlations with new COVID-19 cases.

Different patterns were noted in coronavirus test-related searches. No correlation could be calculated for GT RSVs due to the insufficient number of queries recorded. Strong correlations were found with a lag of 1 day for men (*r*=0.795) and a lead of 1 day for women (*r*=0.823) for NAVER RSVs, as well as for all age groups with a lead of 1 day (*r*=0.828). Moreover, weak to strong correlations were reported in different age groups. The 19~24-year age group had a strong correlation (*r*=0.725) with a lag of 1 day followed by the 30~34-year age group (*r*=0.786 with a lead of 1 day), 35~39-year age group (*r*=0.826 with a lag of 1 day), and 40~44-year age group (*r*=0.755 with a lag of 0 days).

Table 1. Time-lag correlation coefficients between new COVID-19 cases, Google Trends, and NAVER relative search volumes (RSVs) related to the coronavirus and coronavirus tests in South Korea.

		Coronavirus (□□□ □□□□)											
Day	Google Trends	NAVER											
		Gender			Age groups (years)								
		Men	Women	Overall	≤18	19~24	25~29	30~34	35~39	40~44	45~49	50~54	≥55
-3	**0.631**	**0.718**	**0.718**	**0.729**	**0.821**	**0.784**	**0.726**	**0.661**	**0.622**	**0.648**	**0.685**	**0.706**	**0.725**
-2	0.614	0.684	0.684	0.694	0.805	0.759	0.696	0.621	0.581	0.607	0.655	0.680	0.693
-1	0.594	0.670	0.670	0.681	0.812	0.759	0.678	0.601	0.561	0.593	0.638	0.662	0.682
0	0.588	0.654	0.654	0.663	0.803	0.737	0.659	0.578	0.538	0.565	0.606	0.634	0.655
1	0.567	0.647	0.647	0.661	0.794	0.736	0.660	0.579	0.536	0.560	0.606	0.633	0.658
2	0.531	0.591	0.591	0.606	0.759	0.688	0.600	0.513	0.477	0.508	0.554	0.580	0.606
3	0.511	0.579	0.579	0.597	0.749	0.682	0.587	0.500	0.468	0.498	0.537	0.565	0.592

		Coronavirus test (□□□ □□□□ □□□)											
Day	Google Trends	NAVER											
		Gender			Age groups (years)								
		Men	Women	Overall	≤18	19~24	25~29	30~34	35~39	40~44	45~49	50~54	≥55

-3	N/A	0.739	0.769	0.770	**0.595**	0.681	0.654	0.701	0.734	0.696	0.624	**0.612**	0.441
-2	N/A	0.769	0.790	0.797	0.505	0.650	0.687	0.752	0.786	0.692	**0.673**	0.581	0.445
-1	N/A	**0.795**	0.799	0.824	0.500	**0.725**	0.645	0.775	**0.826**	0.704	0.630	0.532	0.434
0	N/A	0.778	0.799	0.812	0.542	0.720	0.653	0.746	0.783	**0.755**	0.559	0.551	0.358
1	N/A	0.775	**0.823**	**0.828**	0.508	0.682	**0.688**	**0.786**	0.814	0.718	0.586	0.557	**0.450**
2	N/A	0.756	0.802	0.805	0.549	0.620	0.623	0.774	0.762	0.731	0.586	0.537	0.433
3	N/A	0.744	0.763	0.781	0.465	0.572	0.606	0.694	0.756	0.633	0.633	0.518	0.424

Note: All correlations were statistically significant at *p*≤0.05. Shaded text, strong correlation with *r*>0.7. Text in bolds, the strongest correlation.

Trends in online information searches based on the type of device used for accessing the Internet

Figure 5 and 6 show trends of online information searches for coronavirus and coronavirus tests using mobile devices and desktops. Mobile search queries for coronavirus were higher in all peaks of information searches. For coronavirus test-related searches, mobile searches seemed to be more frequent and stable than those of desktop searches, in all peaks.

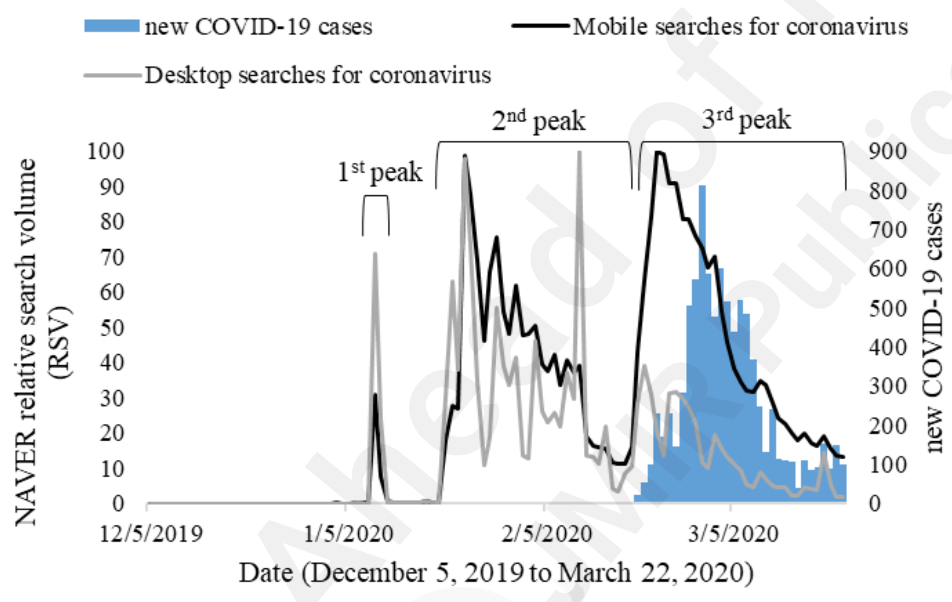


Figure 5. Time series of new COVID-19 cases and NAVER relative search volumes (RSVs) related to the coronavirus in South Korea.

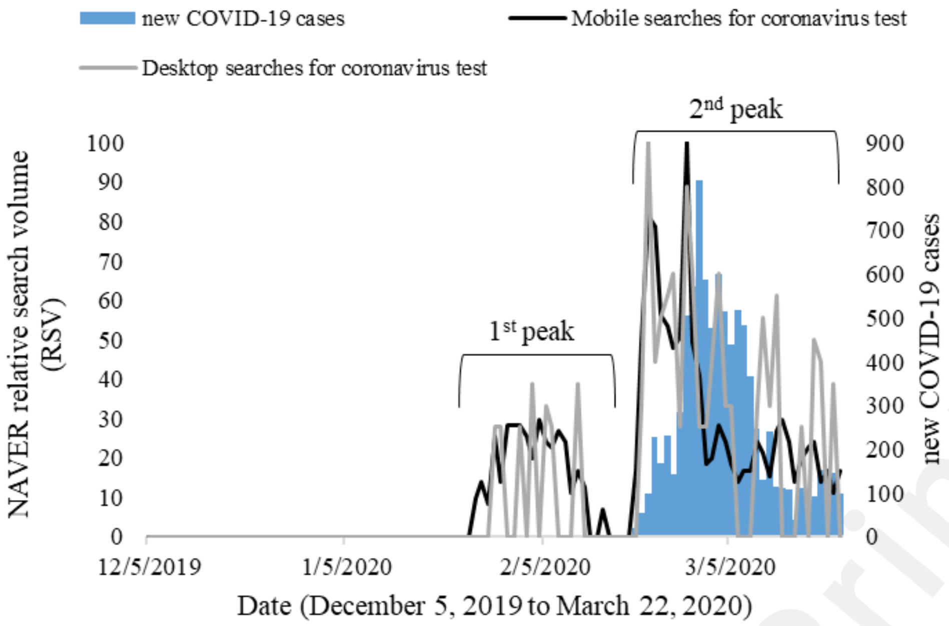
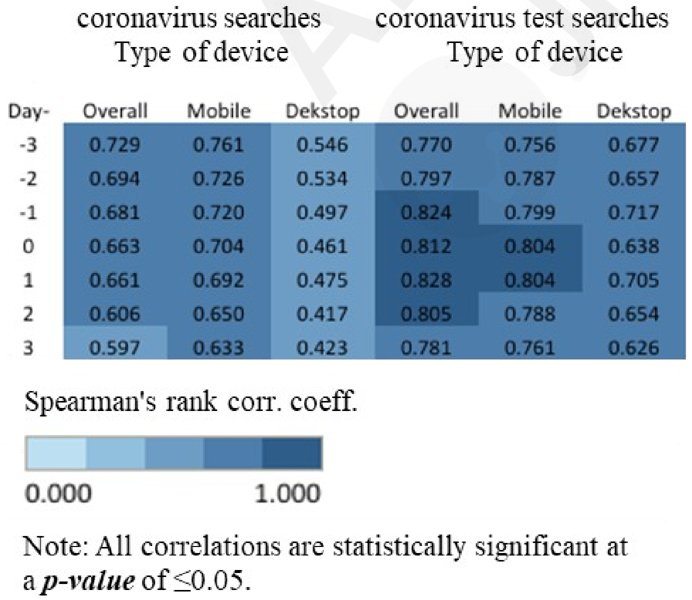


Figure 6. Time series of new COVID-19 cases and NAVER relative search volumes (RSVs) related to the coronavirus test in South Korea.

Spearman’s rank correlation coefficients in Figure 7 demonstrated strong correlations for the overall dataset (mobile and desktop searches) of coronavirus searches with a lag of 3 days (*r*=0.729), as well as mobile searches (*r*=0.761). Interestingly, mobile searches had stronger correlation coefficients for all lag and lead periods than did overall searches. However, weak to moderate correlations (*r*=0.417~0.546) were observed for coronavirus-related searches through desktop devices. For coronavirus test online searches, strong correlations (*r*=0.770~0.828) were reported for all lag and lead days. Still, mobile searches were observed to have a stronger correlation coefficient than desktop searches. The strongest correlations were found with a lag of 0 days for mobile searches (*r*=0.804) and with a lag of 1 day for desktop searches (*r*=0.717).



	coronavirus searches Type of device			coronavirus test searches Type of device		
Day-	Overall	Mobile	Dekstop	Overall	Mobile	Dekstop
-3	0.729	0.761	0.546	0.770	0.756	0.677
-2	0.694	0.726	0.534	0.797	0.787	0.657
-1	0.681	0.720	0.497	0.824	0.799	0.717
0	0.663	0.704	0.461	0.812	0.804	0.638
1	0.661	0.692	0.475	0.828	0.804	0.705
2	0.606	0.650	0.417	0.805	0.788	0.654
3	0.597	0.633	0.423	0.781	0.761	0.626

Spearman's rank corr. coeff.
0.000
1.000

Note: All correlations are statistically significant at a ***p-value*** of ≤0.05.

Figure 7. Time-lag correlation coefficients between new COVID-19 cases and NAVER relative

search volumes (RSVs) related to the coronavirus and coronavirus test in South Korea.

Distributions of new COVID-19 cases and Internet searches

Spatial distributions of new COVID-19 cases and GT RSVs are illustrated in Figure 8. Results showed that 9 days before confirmed cases were reported in South Korea, the numbers of GT RSVs related to the coronavirus captured in Gyeonggi-do, Seoul, Chungcheongnam-do, Daegu, and Ulsan Provinces increased. Thereafter, the aforementioned provinces reported COVID-19-confirmed cases. During the early weeks of disease transmission (as of February 15, 2020), COVID-19 had spread in Seoul, Incheon, Gwangju, Gyeonggi-do, and Jeollabuk-do (Figure 8). Similar patterns were also captured for GT RSVs which seemed to be elevated in those periods in the western part of South Korea where confirmed cases were reported.

Furthermore, a huge surge in new COVID-19 cases began on February 19, 2020. GT RSVs gradually increased during that period in the eastern part of South Korea, including Daegu, the epicenter of local transmission. Daegu contributed 71.79% of confirmed cases or 262.14 cases per 100,000 population as of March 22, 2020 (31) and had a higher estimated death rate than the national rate (32). Interestingly, increases in the number of online searches were observed a week before those massively expanding cases in provinces surrounding Daegu. The large numbers of locally acquired cases were reported from February 25 to March 4, 2020, and swiftly declined in mid-March. When the number of new cases decreased, the number of Internet searches in the western part of South Korea began to increase, which indicated an elevation in the number of COVID-19 cases in the latter part of the study period.

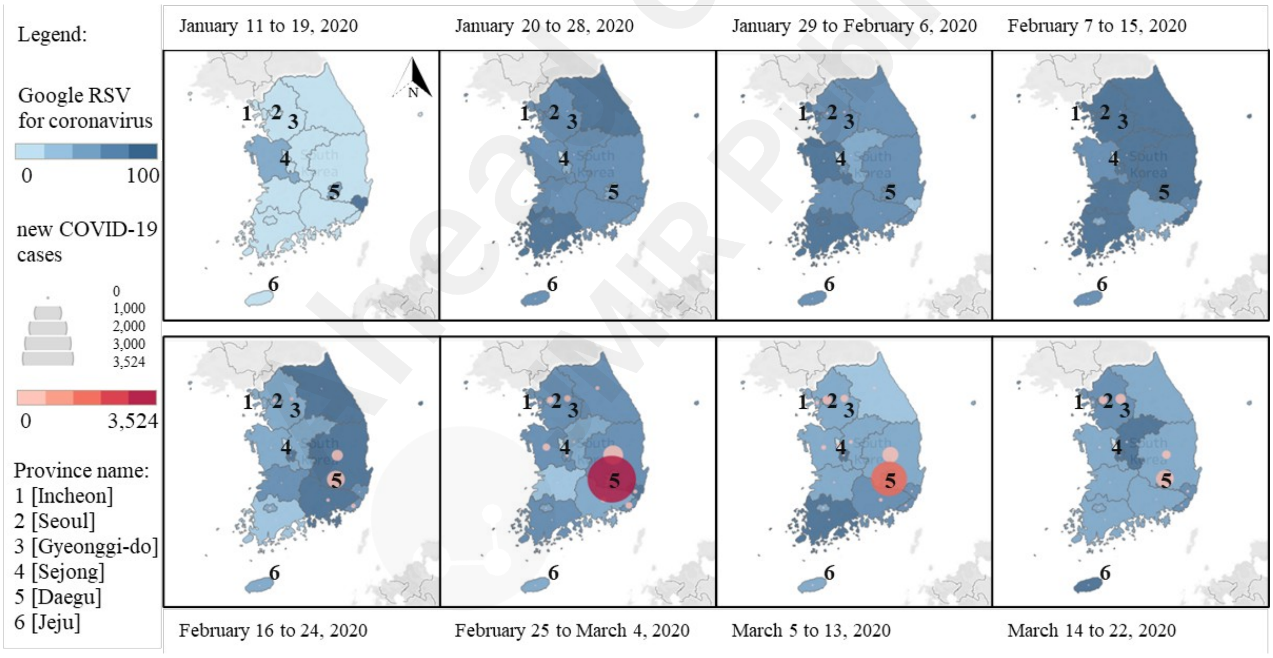


Figure 8. Distribution of new COVID-19 cases and Google Trends relative search volumes (RSVs) in South Korea.

Predicting new COVID-19 cases

Three different models for predicting new COVID-19 cases were established in this study (Table 2). New COVID-19 cases with a lag of 1 day, number of COVID-19 tests with lags of 2 and 1 days, GT coronavirus searches with a lag of 1 day, and NAVER coronavirus searches with a lag of 3 days were selected as important predictors for the models. Model 1 showed high performance, which indicates that this model represented 89% of new COVID-19 cases in contrast with model 2 which only represented 35% of cases as shown in the adjusted r^2 values.

By combining those two models (a case-based model and Internet search data-based model), the model's performance seemed to have slightly increased to nearly 90%, resulting in the lowest root mean squared error (RMSE) as observed in model 3.

Table 2. Prediction model of new-COVID-19 cases in South Korea.

Predictors	Coef. (95% CI)	*p-value* for *F* test	Adjusted r^2	RMSE	AIC	BIC
Model 1 (predictors included new COVID-19 cases and number of COVID-19 tests)						
New COVID-19 cases lag 1 day	0.942 (0.883 to 1.001)	0.000	0.891	54.348	1851.326	1864.03
Number of tests lag 2 days	-0.004 (-0.007 to -0.001)					
Number of tests lag 1 day	0.004 (0.001 to 0.007)					
Cons	3.957 (-5.415 to 13.329)					
Model 2 (predictors included GT and NAVER RSVs related to coronavirus)						
GT RSVs lag 1 day	-0.964 (-1.604 to -0.324)	0.000	0.354	133.802	2153.293	2162.805
NAVER RSVs lag 3 days	3.583 (2.859 to 4.308)					
Cons	28.920 (4.338 to 53.503)					
Model 3 (predictors included new COVID-19 cases, number of tests, GT and NAVER RSVs related to coronavirus)						
New COVID-19 cases lag 1 day	0.880 (0.809 to 0.951)	0.000	0.895	53.177	1835.169	1851.022
Number of tests lag 2 days	-0.004 (-0.006 to -0.001)					
Number of tests lag 1 day	0.004 (0.002 to 0.007)					
NAVER RSVs lag 3 days	0.536 (0.177 to 0.894)					
Cons	-4.334 (-15.136 to 6.467)					

Cons, Constant; Coef., Coefficient; CI, Confidence interval; RMSE, Root mean squared error; AIC, Akaike information criterion; BIC, Bayesian information criterion; GT, Google Trends; RSVs, Relative search volumes.

Models were then plotted in Figure 9 for both the development and validation sets. Model 3 performed better compared to the two other models in the development set as assessed by the value of the adjusted r^2 as well as RMSE, AIC, and BIC. In the validation set, this model also performed well, and this was indicated by the RMSE decreasing to 18.320.

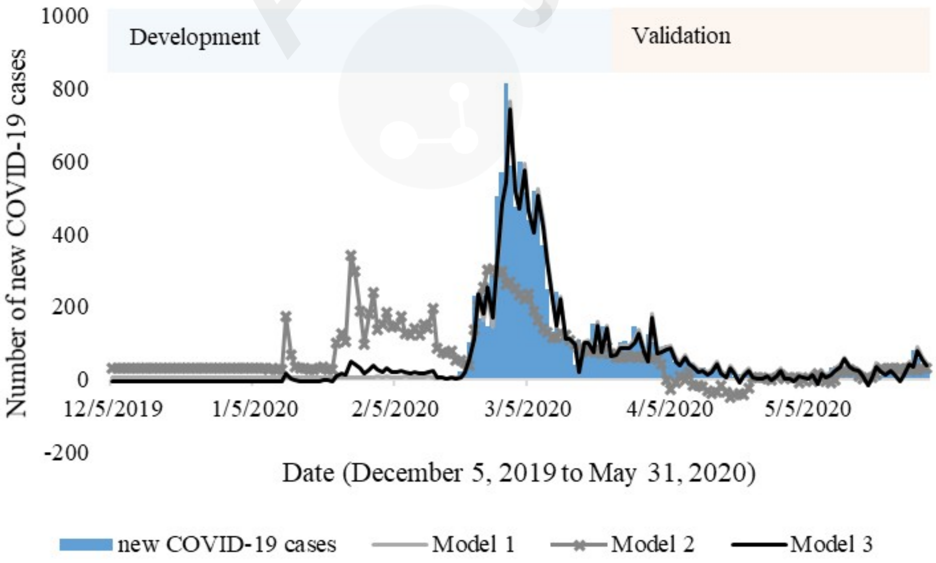


Figure 9. Prediction of new COVID-19 cases in South Korea.

Discussion

Risk perception is defined as a person's subjective judgment toward the likelihood of negative occurrences including diseases or illnesses (33). In terms of disease outbreaks, understanding community health risk perceptions are urgently needed in the early phase of an outbreak, particularly in the case of an emerging disease. This is because in the initial period, there will be limited treatments, few numbers of resources, and delays in active interventions (34). Therefore, exploring the perception of risk is a necessary step in managing the risk of an outbreak. Since a robust public risk perception assessment could help in divining effective risk communication, this step should be taken immediately to reduce the impact of the COVID-19 outbreak. Consequently, it is more affordable to conduct the community health risk perception assessment using Internet search data, since it can be provided more easily, promptly, and cost-effectively compared to a survey methods (16) and also can potentially capture anomalous patterns in real time (17). With the widespread use of mobile devices and the Internet, Internet search data can be more accurate in representing the community health risk perceptions (35) as information-seeking intentions are directly affected by risk perceptions (9).

Principal results

In this study, we found various correlations, which ranged from weak to strong, among GT and NAVER RSVs, new COVID-19 cases, and the number of tests. Previous studies also reported strong correlations between GT and NAVER RSVs compared to surveillance data (16, 36). Therefore, increased searches for COVID-19-related information might represent community health risk perceptions during local and international events. NAVER RSVs, as a local search engine that has the largest market share in South Korea (57.31% for all search categories in 2020 as of June 14) (18), seemed to be more sensitive to local issues such as coronavirus tests as shown in Figure 3. A similar result was also reported in a previous study which demonstrated that Baidu (in China) has better predictive performance for disease prediction than GT RSVs (36). These findings suggest that NAVER RSVs could also potentially complement the use of GT RSVs, which are excessively utilized in the fields of infodemiology.

Patterns of community risk perceptions retrieved from information searches in this analysis were explained by examining different aspects: time, gender, age groups, types of device used for accessing the Internet, and spatial distributions. Patterns according to time revealed that the number of online queries related to COVID-19 increased during local events, including local transmission, approval of coronavirus test kits, implementation of coronavirus drive-through tests, a facemask shortage, a widespread campaign for social distancing, and transmission of the Shinchoenji cluster, as well as during international events such as the announcement of the PHEIC. Yet, South Korea was also one of the countries affected by the MERS epidemic (37). That experience might have also contributed to the increased number of searches for coronavirus information even though cases had not yet been detected till then. Moreover, MERS-related searches also remind high during study period. These findings indicated that public health risk perceptions increased following both local and international crises. Hence, risk communication should promptly be conducted, considering that health risk perceptions might change over time as the outbreak progresses.

Patterns according to time also revealed decreased numbers of GT and NAVER RSVs in the middle of the epidemic curve, which might have been caused by the extensive availability of online news and health expert reports during that period (38). It might also have been provoked by decreased risk perceptions as the epidemic progressed (7). Thus, utilizing Internet query data to analyze community risk perceptions could be useful in the early stage of an outbreak.

Moreover, patterns categorized by different age groups revealed that younger (≤29 years) and older age groups (≥50 years) had strong correlations of Internet searches for coronavirus information with new COVID-19 cases. This finding demonstrated the high-risk perceptions of those age groups, even 3 days before an increase in the number of new COVID-19 cases locally. High-risk perceptions in younger age groups might have been induced by massive Internet access for acquiring information and high numbers of confirmed cases in that age group (33.24%) in South Korea (31, 39). Meanwhile, perceived vulnerability might be common in older age groups, since an older age is one of the prominent risk factors for COVID-19 mortality (40), and 98.08% of fatal cases in South Korea occurred in older adults (31). Additionally, a previous study showed that the older age group had higher risk perceptions (7).

In contrast, the age group of 30~49 years only showed weak to moderate correlations even 3 days before the event. This might have been due to the lower percentage of confirmed cases (23.94%) in that age group compared to that in the younger age group (≤29 years), which could also have influenced health risk perceptions. Meanwhile, online queries concerning coronavirus tests showed high-risk perceptions in the 35~44-year age group. These findings illustrate that adults perceived the coronavirus test-related information to be more important than disease-related knowledge. It might also have been influenced by the massive numbers of coronavirus tests conducted so far. Meanwhile, younger (aged ≤29 years) and older age groups (aged ≥50 years old) had a different perception, thereby making infection-related information an essential search. In terms of gender, both men and women perceived the coronavirus as having similar levels of risk, but risk perception for coronavirus test was higher among women. This result is similar to that reported in a previous study which showed a higher risk perception in the women's group (7). Hence, health risk communication should target both men and women as well as vulnerable age groups.

As to device utilization, patterns demonstrated that mobile device searches had stronger correlations with COVID-19-related searches compared to desktop queries. Strong correlations for mobile device searches were even observed 3 days before the outbreak. However, desktop searches showed a strong correlation with a lag of 1 day, which was 2 days later, compared to mobile searches. This finding implies that high-risk perceptions stimulated an enormous number of mobile searches during the outbreak period. Identical results were also illustrated in a previous study by Shin and colleagues (16). The widespread use of mobile devices in the digital era (35), has promoted changes in behavior, from desktop to mobile device users. Therefore, the government should ensure that risk communication can be easily accessed through mobile platforms for rapid dissemination. Research findings also demonstrated that the spatial distributions of Internet searches were higher in locations with new COVID-19 cases. This finding was similar to that in previous studies which indicated that individuals in affected areas have higher risk perceptions (7, 11).

Later in the analysis, we also addressed the prediction of new COVID-19 cases using 3 different models. Results showed that adding COVID-19-related searches provided by NAVER could increase the performance of the model compared to that of the COVID-19 case-based model. This result resembled an earlier study (17) which also found that a model's performance increased with use of Internet search data from local search engines. Furthermore, in the validation set, this model performed better, which might have been caused by a long-enough period for querying NAVER data, therefore trends could be adjusted better and affect the model's performance in the validation set. Hence, considering NAVER RSVs data for case prediction could be important, although employing the same dataset to better understand health risk perceptions is also of the utmost importance particularly in the early stage of an outbreak.

Briefly, this study provides a depiction of community health risk perceptions toward COVID-19 in South Korea, which tended to be higher in the period of local and international events, also for women, certain age groups, and people in affected areas. During the outbreak, people were more likely to access the Internet through mobile devices, which are potential channels where health risk communication can be effectively and densely disseminated. Moreover, NAVER RSVs can potentially be used for health risk perception assessments and disease prediction. This method demonstrated an easy and low-cost approach for estimating health risk perceptions during a pandemic. Since providing a rapid risk perception assessment is urgently needed in the early stage of an outbreak, combining GT and NAVER RSVs could be beneficial for targeting risk communication in terms of time, population characteristics, and location. GT RSVs alone only revealed patterns according to time and location (41). However, this study only explored the positive risk perceptions toward COVID-19 rather than negative risk perceptions such as psychological impacts. As multiple studies also reported increases incidence of anxiety, depression, anger, insomnia, distress, and suicidality during the initial phase of the epidemic (42), exploring the negative risk perceptions of COVID-19 pandemic would be important for future works.

Limitations

As online search queries might change over time, identifying the best lag time for conducting risk communication is challenging. However, using either GT or NAVER RSVs allowed flexibility in defining the time range of data queries. Thus, we can collect adequate retrospective datasets for identifying the best lag time. In addition, this analysis might be limited to specific time frames, included only 2 popular search engines, and certain keywords, as well as limited for positive risk perceptions. Therefore, further research that considers those aspects to improve results of the risk perception analysis is required.

Conclusions

Community health risk perceptions toward the COVID-19 outbreak in South Korea observed from GT and NAVER RSVs increased during local and international events and were higher in women, certain age groups, and in affected areas. While NAVER RSVs tended to be more sensitive in terms of local issues, integrating GT and NAVER RSVs could potentially provide varied search patterns in terms of time, population characteristics, and location. Moreover, online searches also identified as important variables in predicting epidemic curves in the initial stage of an outbreak.

Acknowledgments

AH designed the study, performed the experiments, analyzed the data, and drafted the manuscript. ES contributed analytical suggestions and revised the manuscript. AF analyzed the data and revised the manuscript. ECYS conceived the study, designed the experiments, and revised the manuscript. All authors approved the final version. Moreover, we gratefully thank Jihoo Kim and colleagues for providing open-access data on COVID-19-related information in

South Korea on kaggle.com.

This study was funded in part by the Ministry of Science and Technology (MOST) in Taiwan (grant no. MOST108-2221-E-038-018) and the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan (grant no. DP2-108-21121-01-A-01-04) to Emily Chia-Yu Su. This work also was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. 2018R1C1B6001723) to Eunha Shim. The sponsor had no role in the research design or contents of the manuscript for publication.

Conflicts of Interest

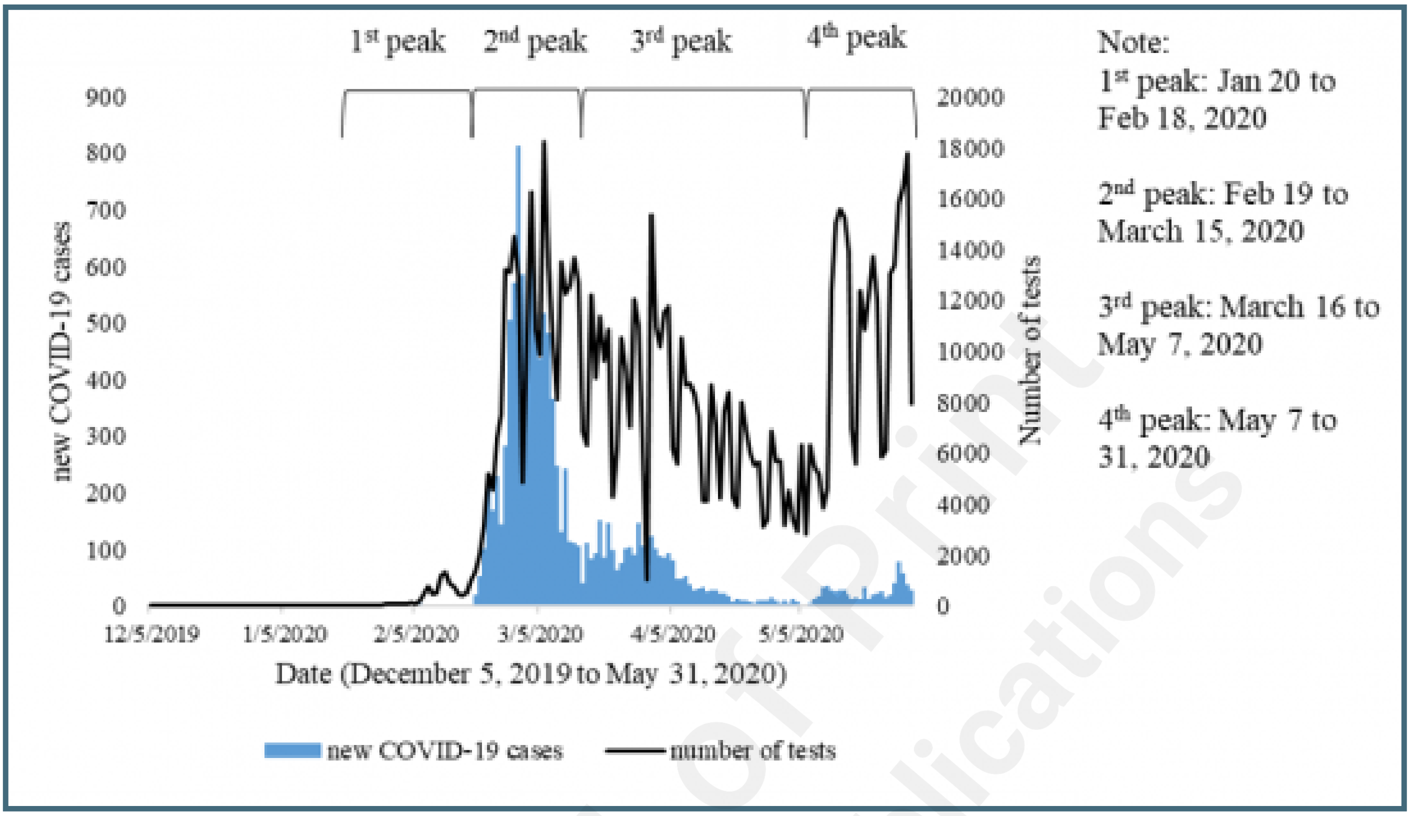
None declared.

References

Supplementary Files

Figures

Time series of new COVID-19 cases and number of tests in South Korea.



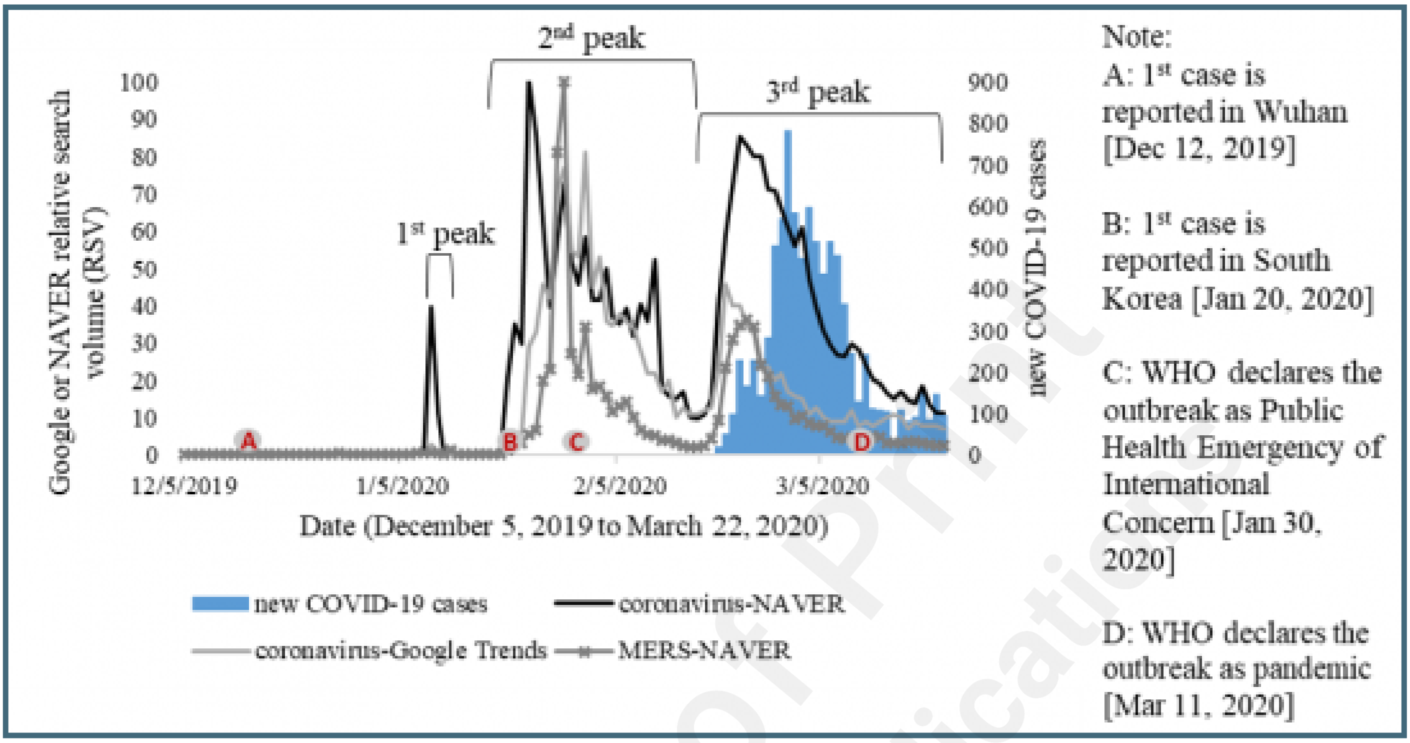
Note:
1st peak: Jan 20 to Feb 18, 2020

2nd peak: Feb 19 to March 15, 2020

3rd peak: March 16 to May 7, 2020

4th peak: May 7 to 31, 2020

Time series of new COVID-19 cases, Google Trends, and NAVER relative search volumes (RSVs) related to the coronavirus and Middle Eastern respiratory syndrome (MERS) in South Korea.



Note:

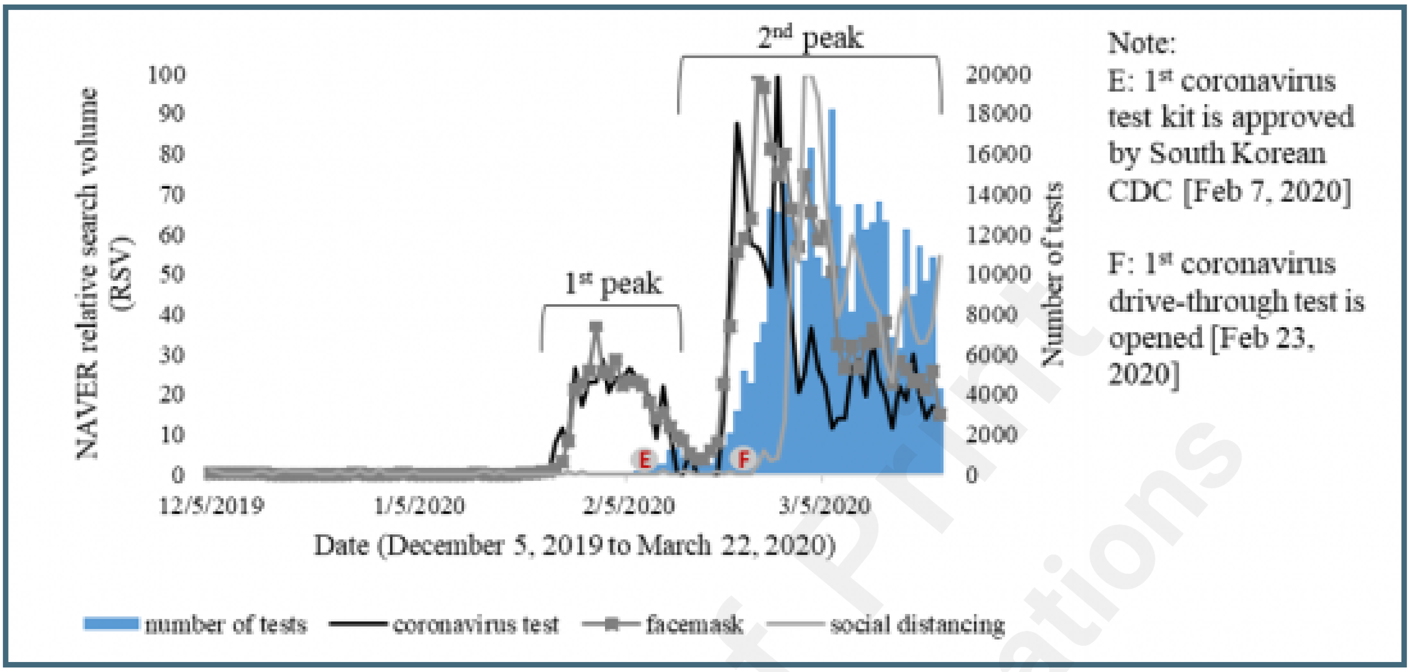
A: 1st case is reported in Wuhan [Dec 12, 2019]

B: 1st case is reported in South Korea [Jan 20, 2020]

C: WHO declares the outbreak as Public Health Emergency of International Concern [Jan 30, 2020]

D: WHO declares the outbreak as pandemic [Mar 11, 2020]

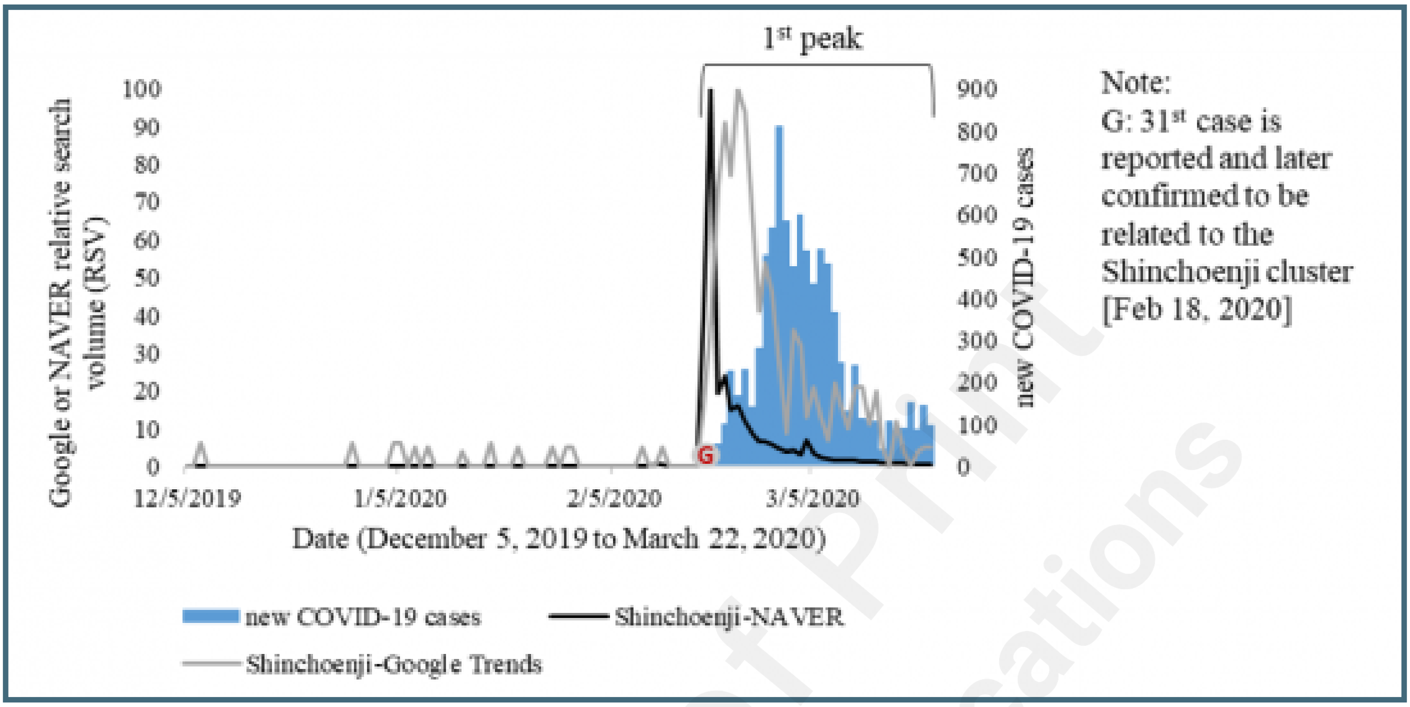
Time series of new COVID-19 cases and NAVER relative search volumes (RSVs) related to coronavirus tests, facemasks, and social distancing in South Korea.



Note:
E: 1st coronavirus test kit is approved by South Korean CDC [Feb 7, 2020]

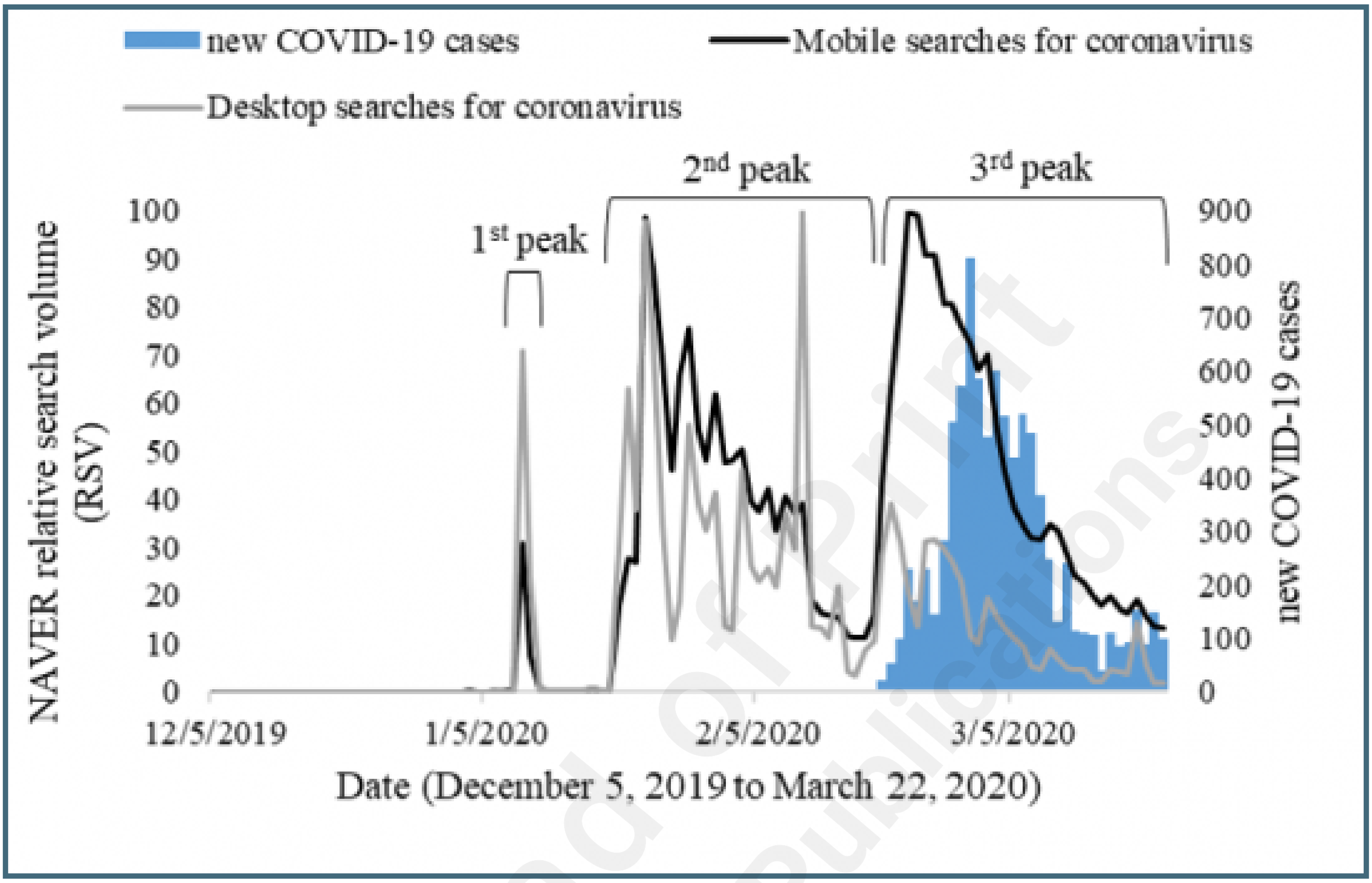
F: 1st coronavirus drive-through test is opened [Feb 23, 2020]

Time series of new COVID-19 cases, Google Trends, and NAVER relative search volumes (RSVs) related to the Shinchoenji cluster in South Korea.

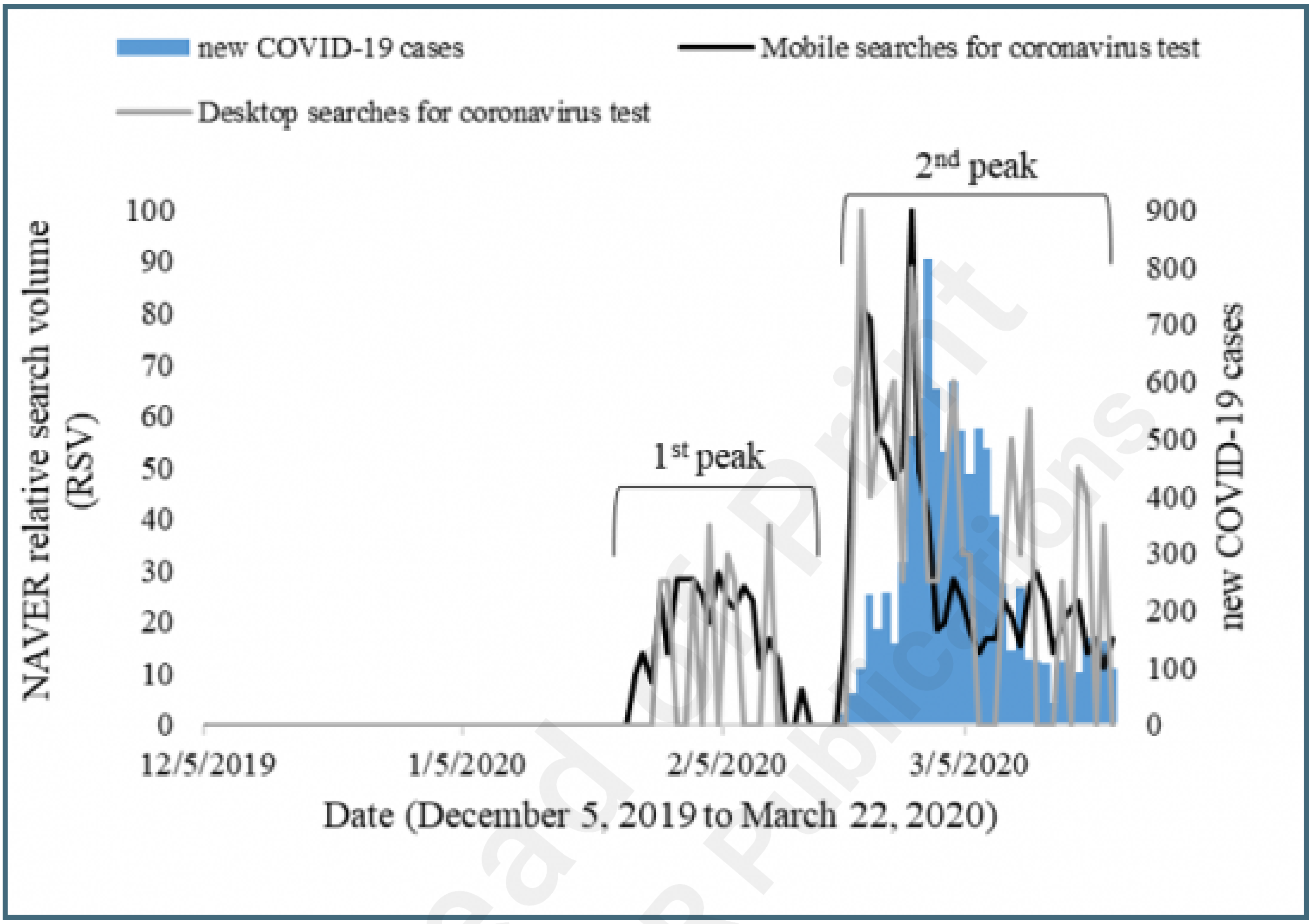


Note:
G: 31st case is reported and later confirmed to be related to the Shinchoenji cluster [Feb 18, 2020]

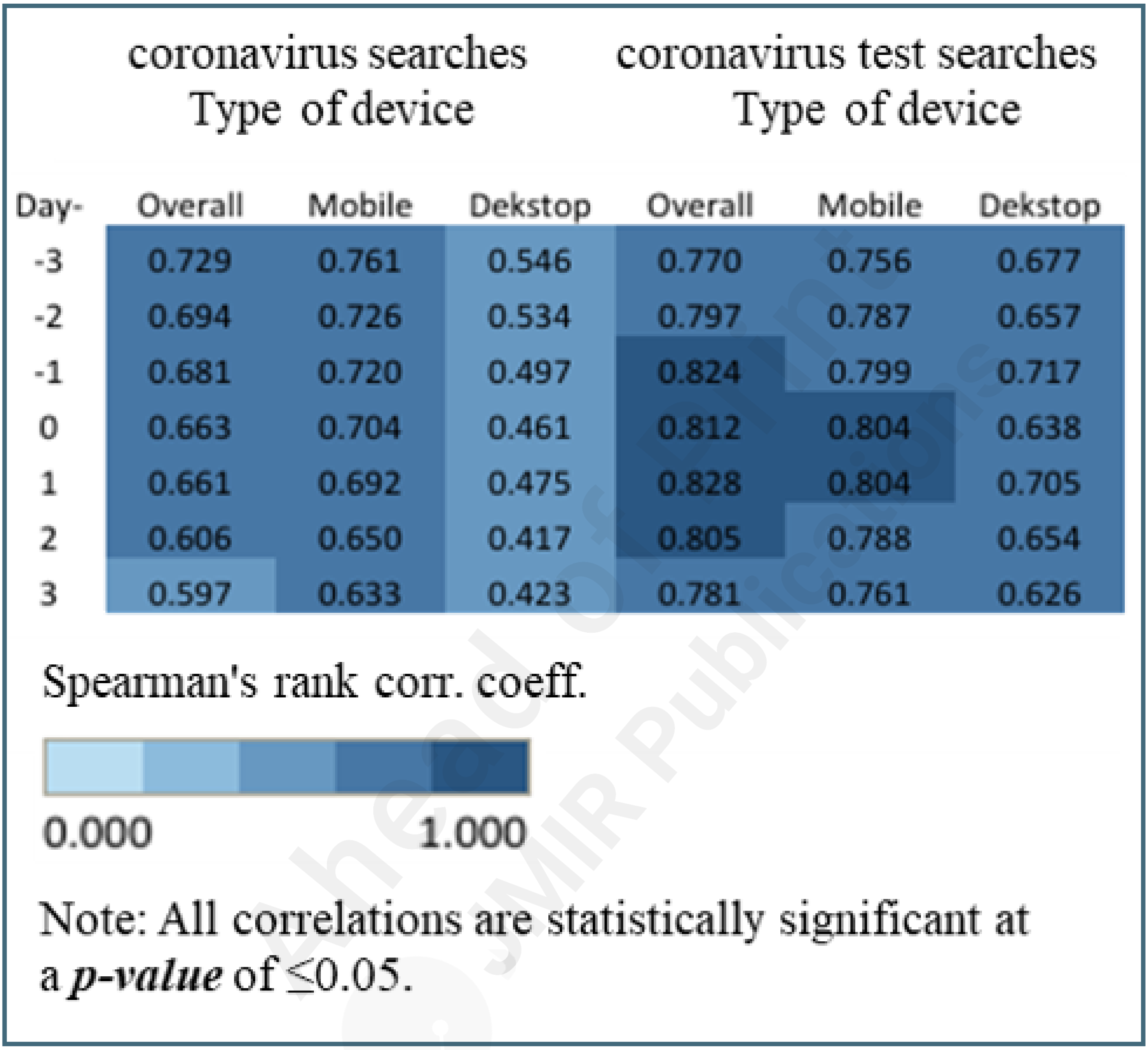
Time series of new COVID-19 cases and NAVER relative search volumes (RSVs) related to the coronavirus in South Korea.



Time series of new COVID-19 cases and NAVER relative search volumes (RSVs) related to the coronavirus test in South Korea.



Time-lag correlation coefficients between new COVID-19 cases and NAVER relative search volumes (RSVs) related to the coronavirus and coronavirus test in South Korea.



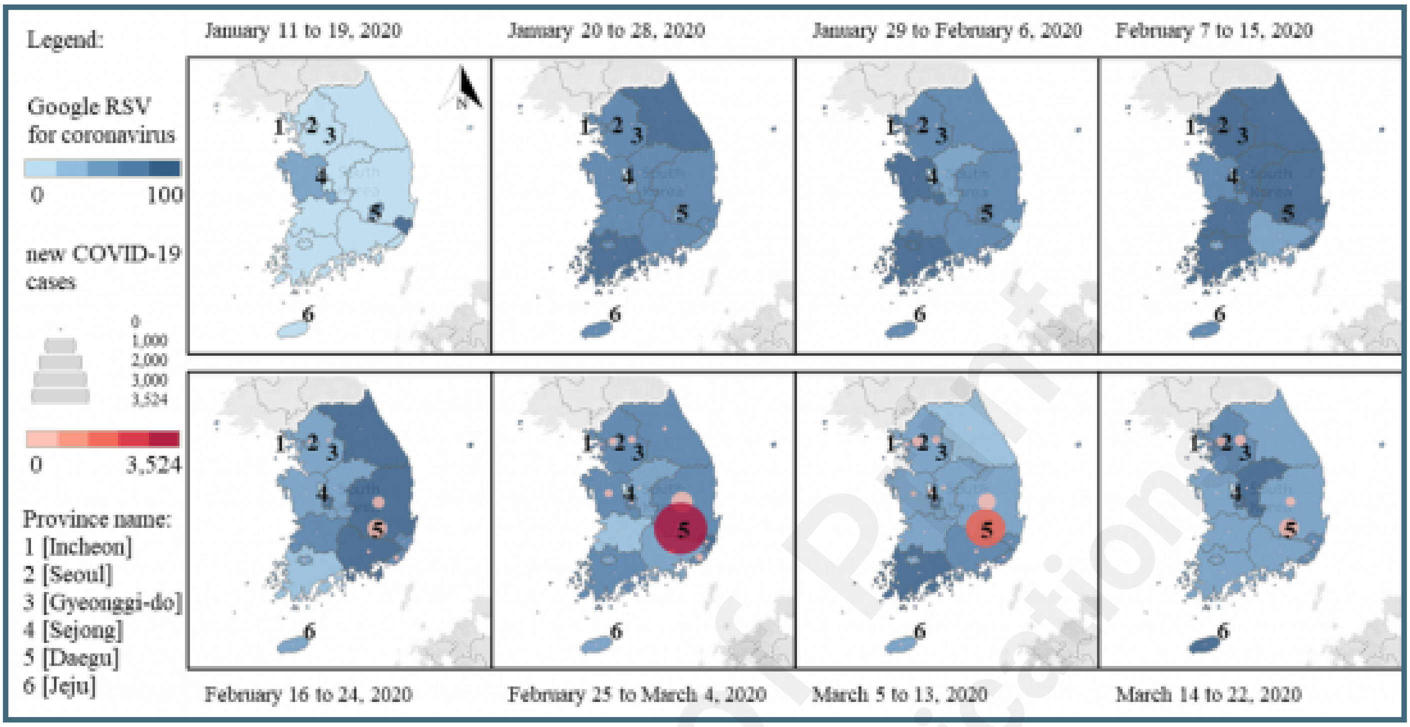
	coronavirus searches Type of device			coronavirus test searches Type of device		
Day-	Overall	Mobile	Dekstop	Overall	Mobile	Dekstop
-3	0.729	0.761	0.546	0.770	0.756	0.677
-2	0.694	0.726	0.534	0.797	0.787	0.657
-1	0.681	0.720	0.497	0.824	0.799	0.717
0	0.663	0.704	0.461	0.812	0.804	0.638
1	0.661	0.692	0.475	0.828	0.804	0.705
2	0.606	0.650	0.417	0.805	0.788	0.654
3	0.597	0.633	0.423	0.781	0.761	0.626

Spearman's rank corr. coeff.

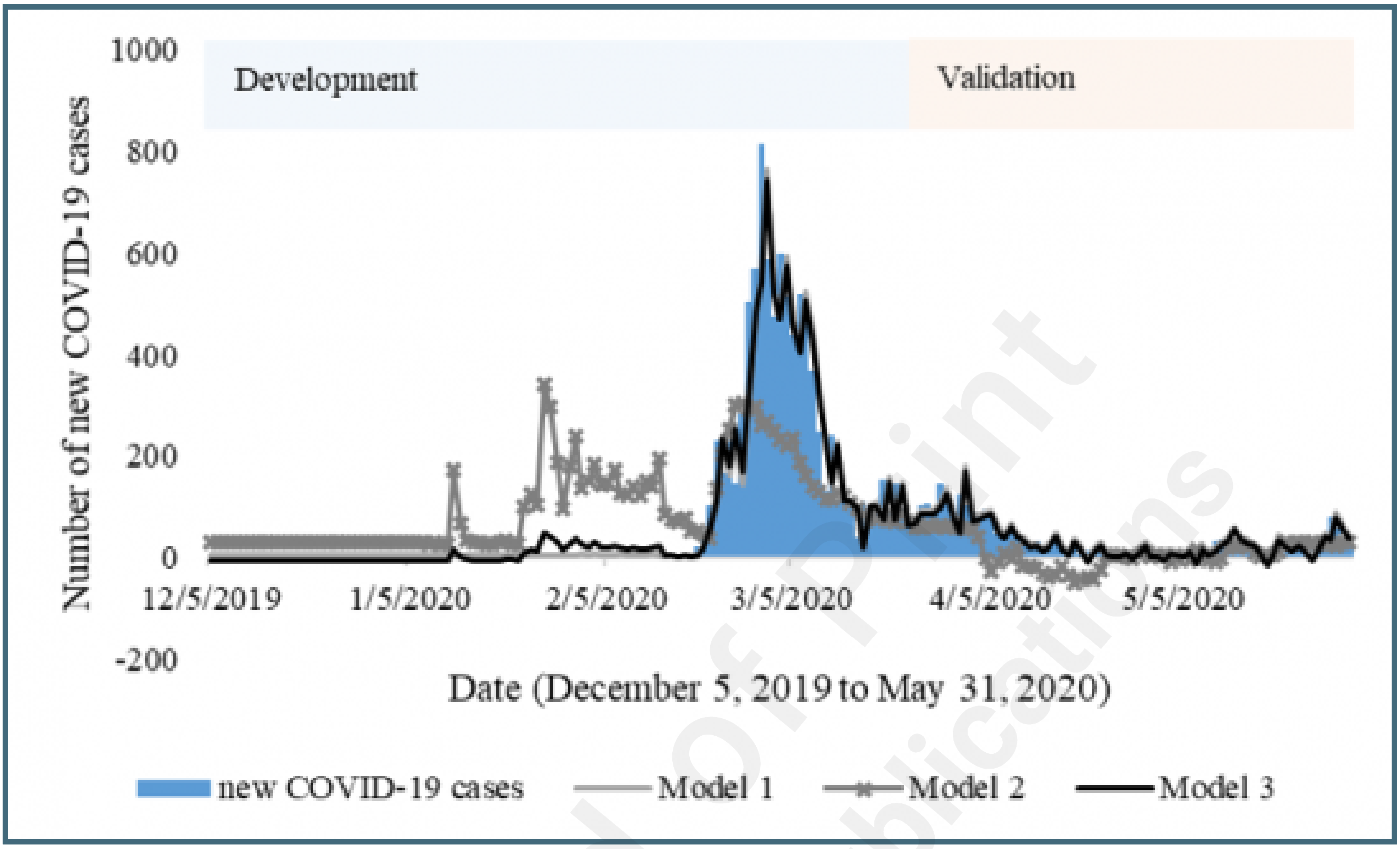
0.000 1.000

Note: All correlations are statistically significant at a *p-value* of ≤0.05.

Distribution of new COVID-19 cases and Google Trends relative search volumes (RSVs) in South Korea.



Prediction of new COVID-19 cases in South Korea.


Development
Validation
Number of new COVID-19 cases
12/5/2019
1/5/2020
2/5/2020
3/5/2020
4/5/2020
5/5/2020
Date (December 5, 2019 to May 31, 2020)
new COVID-19 cases
Model 1
Model 2
Model 3