**Title:** A methodological blueprint to identify COVID-19 vulnerable locales by socioeconomic factors, developed using South Korean data.

**Authors:** Bayarmagnai Weinstein, Alan R. da Silva, Dimitrios E. Kouzoukas, Tanima Bose, Gwang-Jin Kim, Paola A. Correa, Santhi Pondugula, Jihoo Kim, David O. Carpenter*

**Author Affiliations:** University at Albany, Rensselaer, New York, USA (B. Weinstein, D. Carpenter); University of Brasília, Brasília, Brazil (A. da Silva); Edward Hines Jr. VA Hospital, Hines, Illinois, USA; Loyola University Chicago, Maywood, Illinois, USA (D. Kouzoukas); Ludwig-Maximilian University of Munich, Planegg-Martinsried, Germany (T. Bose); University of Freiburg, Freiburg, Germany (G. Kim); Howard Hughes Medical Institute, Ashburn, Virginia, USA (P. Correa); University of Florida, Florida, USA (S. Pondugula); Hanyang University, Seoul, South Korea (J Kim); Harvard University, Boston, Massachusetts, USA (B. Weinstein)

*Correspondence to: David O. Carpenter, Institute for Health and the Environment, University at Albany, 5 University Place, Rensselaer, NY 12144, USA; email: dcarpenter@albany.edu**;** Tel: +1-518-252-2660

**Word Count:** 2999 words

## Abstract

COVID-19 has more severely impacted socioeconomically (SES) disadvantaged populations. Lack of SES measurements and inaccurately identifying high-risk locales can hamper COVID-19 mitigation efforts. Using South Korean COVID-19 incidence data (January 20 through July 1, 2020) and established social theoretical approaches, we identified COVID-19-specific SES factors. Principal component analysis created composite indexes for each SES factor, while Geographically Weighted Negative Binomial Regressions mapped a continuous surface of COVID-19 risk for South Korea. High area morbidity, risky health behaviors, crowding, and population mobility elevated area risk for COVID-19, while improved social distancing, healthcare access, and education decreased it. Our results indicated that falling SES-related COVID-19 risks and spatial shift patterns over three consecutive time periods reflected the implementation of reportedly effective public health interventions. While validating earlier studies, this study introduced a methodological blueprint for precision targeting of high-risk locales that is globally applicable for COVID-19 and future pandemics.

**Keywords:** COVID-19, Pandemics, Socioeconomic Factors, Spatial Regression, South Korea

## Introduction

COVID-19 corroborated insights gained from SARS, H1N1 influenza, and MERS pandemics that more severely affect socioeconomically (SES)-disadvantaged populations[1]. Area-health and SES lead to differential incidence and mortality as they define the disease agent exposure extent, susceptibility to contracting disease once exposed, and disease severity[2]. Although existing studies provided COVID-19 risk factors, none identified COVID-19-vulnerable

locales associated with SES factors with acceptable generalizability and methodological capacity.

Current studies use isolated arbitrary SES variables based on researchers' preferences. These lack conceptual clarity, and thereby fail to truly uncover how SES affects health outcomes. Consequently, a lack of harmony (reduced external validity) exists among the studies conducted across national and sub-national settings. This lack impedes rapid response and effective control measures, especially for developing countries relying on resource-rich countries for guidance. To address this, we integrated Coleman's Social Theory and Blumenshine's mechanistic framework (Figure 1), which permits a universal SES definition and SES indicator selection mechanistically/causally relevant to the health outcome.

SES data span an expanse of highly intercorrelated variables such as *healthcare utility rate*, *insurance coverage rate*, and *unmet healthcare needs*, each used interchangeably as a healthcare access measure. A composite measure approach can summarize the multi-dimensional information of individual SES variables, while addressing the inadequacy of traditional statistics in simultaneously handling multiple highly intercorrelated variables. A commonly-used index, area-deprivation index (ADI)[3], can adjust models to SES as a confounder, but not when SES is the exposure. Since disease risk is determined by multiple SES factors, understanding SES-related disease risk requires SES concept-specific data. Therefore, we recommend causally-important multiple composite indexes, as these allow researchers to delineate the mechanism through which SES theme(s) increase the COVID-19 risk[4].

Thus far, only one study explored geographic methodology with a multi-scale Geographically Weighted Regression (GWR) accounting for selected SES variables (median household income, income inequality, percentage of nurse practitioners, and black female population) in the US[5]. Since multi-scale GWR doesn't fit a beta distribution typical for infectious

disease rates[6], we recommend Geographically Weighted Negative Binomial Regression (GWNBR) to improve accuracy. This directly takes discrete count data without further transformation, and is robust to overdispersion, spatial/temporal clustering and false-positives[7, 8].

Globally, the COVID-19 pandemic emerged in waves with country-specific mitigation strategies producing sharp declines. We chose South Korean COVID-19 incidence data because it presented extremely high overdispersion and spatial clustering, being more complex than typical infectious disease data. South Korea, as a real extreme case scenario, allowed us to check our framework's functionality. Our study's goals were to 1) provide methodological guidance for identifying COVID-19-vulnerable locales associated with SES factors; and 2) operationalize a framework using South Korean data to demonstrate its value and in the interpretation of the results.

## Methods

### Study design and population

We used COVID-19 incidence data from January 20 through July 1, 2020, released from the Korea Centers for Disease Control and Prevention (KCDC)[9] and prepared by the DS4C project[10]. Analytical data consisted of 11 811 COVID-19 cases aggregated by 250 districts (Table S1) aligned to SAS's South Korean geographic matrix. Since data was unavailable for Daegu's subparts, we estimated the incidence from KCDC's press release cluster reports.

### Conceptual model

Figure 1 shows the Coleman-Blumenshine Framework (CBF) refined approach, based on Coleman's Social Theory and Blumenshine's mechanistic framework[2, 11]. The model defines SES as a function of social and human capitals[11, 12] and emphasizes pathways by how SES

indicators differentially increase SARS-CoV-2 exposure and susceptibility to developing COVID-19[2]. Based on the CBF model and COVID-19 risk factors literature[13-17], we identified seven area-level health and SES factors that determined the SARS-CoV-2 exposure level and the likelihood of developing COVID-19 after exposure.

SES measurement

All SES related data sources were retrieved from the Korean Statistical Information Service's (KOSIS) online data archive[18]. Table S3 presents the reviewed data sources used for SES measurement. Table 1 shows 24 data items out of 124 candidates relevant to the seven health/SE areas. We used an independent variable proxy for *education*, and by Principal Component Analysis (PCA) created six thematic composite indices: *healthcare access*, *health behavior*, *crowding*, *area morbidity*, *education*, *difficulty to social distancing*, and *population mobility*. Factors were computed as linear combinations of the original variables selected for each health/SE theme. We used the first component scores[19] in calculating the composite scores since they explained the largest data variation. Then we computed each variable's weight by dividing each factor score by the sum of all variable factor scores as:

$$Weight_i = Score_i / \sum_{i=1}^{p} Score_i \quad (1)$$

where *i* relates to each theme's variable and *p* is the number of each theme's variables. Each thematic composite index was computed as the weighted average for all 250 district values. For example, the composite index for *health behavior* was calculated as:

$$\textit{Health behavior}_k = 0.438 \times \textit{obesity by measurement}_k + 0.429 \times \textit{alcohol drinking}_k + 0.100 \times \textit{current smoking}_k + 0.033 \times \textit{self-reported obesity}_k \quad (2)$$

where *k* is the original variable's value for district *k*. Note that weights sum to 1 (0.438 + 0.429 + 0.100 + 0.033 = 1). Six thematic composite indices and an individual proxy for education

(percentage of high school educated people) were used in the final models as independent variables.

The ratio of the 25th percentile and each theme's maximum value was for *healthcare access*: 2.5; *health behavior*: 1.2; *crowding*: 1.9; *area morbidity*: 1.3; *education*: 1.5; *difficulty to social distancing*: 3.0; and *population mobility*: 10.3.

Statistics

Our model outcome was the confirmed case counts of COVID-19 aggregated by 250 districts. Global negative binomial regression (GNBR) and GWNBR[7] computed relative risk of COVID-19 associated seven area health/SE themes.

Global models

GNBR models calculated relative COVID-19 risk for the entire study period and each pandemic phases. The global model was set as:

$$\text{COVID-19} = \exp(\beta_0 + \beta_1\, \textit{healthcare access} + \beta_2\, \textit{health behavior} + \beta_3\, \textit{crowding} + \beta_4\, \textit{area morbidity} + \beta_5\, \textit{education} + \beta_6\, \textit{difficulty to social distancing} + \beta_7\, \textit{population mobility} + \varepsilon \quad (3)$$

where, $\beta_0, \ldots, \beta_n$ were the intercept and regression coefficients, whereas $\varepsilon$ was the model random error.

Local spatial models

We used Gaussian GWNBR to model a discrete count data and handle overdispersion issue. GWNBR computed parameter estimates for all districts following:

$$y_j \sim NB\left[t_j \exp\left(\sum_k \beta_k(u_j, v_j)\, x_{jk}\right), \alpha(u_j, v_j)\right] \quad (4)$$

where $(u_j, v_j)$ are the locations (coordinates) of the data points *j*, for *j* = 1, . . . , *n*. The models empirically computed bandwidth, via the cross-validation criterion, and achieved minimal Akaike information criterion (AIC) as:

$$CV = \sum_{j=1}^{n} \left[ y_j - \hat{y}_{\neq j(b)} \right]^2 \quad (5)$$

where $\hat{y}_{\neq j(b)}$ is the estimated value for point *j*, omitting the observation *j,* and *b* is the bandwidth. The likelihood of false-positives was corrected by the method of da Silva and Fotheringham[20].

All statistical analyses including specific macro programs for spatial weight matrices and GWNBR models were implemented using SAS (version 9.4). Missing data (2%) were excluded from the analyses.

## Results

Figure 2 compared the spatial COVID-19 distribution across pandemic phases. The initial outbreak wave occurred in Daegu which then spread to Gyeongsangbuk-do and surrounding provinces in the early phase[16]. The second wave occurred in Seoul and its surrounding metropolises, Ulsan and Busan, and Gyeonggi-do province in the late phase of the pandemic.

### SES measurement

Table 1 presents the PCA results including the contributing variables for each of the six composite indexes. The factor scores and weights of each contributing variable associated with the first PCA-identified component and the quartiles of the health/SE themes are shown.

### Global and local spatial models

In the entire study period model using GNBR, the COVID-19 risk increased risky *health behavior, area morbidity,* and *difficulty to social distancing* (Table 2). Inverse associations indicate an increased COVID-19 risk with reduced *healthcare access*, lower *education*, and increased outflux in *population mobility*. The *crowding-associated* risk was not statistically significant.

We implemented separate global and local spatial models for the early, middle, and late pandemic phases. Figure 3 presents the relative COVID-19 risk with its 95% CI from GNBR models, and Figure 4, the relative risk spatial distribution from GWNBR associated with seven thematic areas by pandemic phases. Supplementary Table S4 provides more details on the stratified GNBR models. GWNBR fit data better than the global model given smaller AIC for the middle and late phases, respectively, ($AIC_{gwnbr}$ ~1034 vs $AIC_{gnbr}$~1044, $AIC_{gwnbr}$ ~1038 vs. $AIC_{gnbr}$~1074) except for the early phase of the pandemic ($AIC_{gwnbr}$ ~3533 vs $AIC_{gnbr}$~1527). This reflects the large spatial case cluster emerging from Daegu during the early phase that subsequently spread to its neighboring districts. The GNBR and GWNBR model results agreed across all pandemic phases. In the early phase, lower *healthcare access* and *education,* and increased risky *health behavior, area morbidity, difficulty to social distancing,* and *population mobility* associated with higher COVID-19 risk. *Crowding-associated* risk was not significant in GNBR. In the middle phase, *healthcare access, area morbidity, education, and difficulty to social distancing* remained to be significant factors. In the late phase, only *healthcare access, health behavior,* and *increased crowding* remained significant.

Early phase maps showed elevated risk in noncontiguous districts, which may reflect virus transmission in the initially affected districts before spreading over expanded areas. Relative risks associated with healthcare access, health behavior, and crowding indices were significant in the early phase where each index varied in spatial coverage (Figure 4). In the middle phase, only healthcare access remained significant with a northwest (capital region) and

southwest spatial shift from the south and east. In the late phase, the risk associated with these three factors became significant and concentrated around the capital region, Chungcheongbuk-do, and Jeollabuk-do provinces. The period and observed spatial pattern here are consistent with the second wave that emerged in Seoul and its surrounding areas.

Other health/SE themes were not significant risk factors in the late phase. *Difficulty to social distancing* increased COVID-19 risk in the capital and middle regions in the early phase which then shifted to the country's southeast part in the middle phase. *Area morbidity*-associated risk was concentrated in the western part which then gradually shifted north in the middle phase. *Education*-associated risk was higher in the west in the early phase until it shifted southwest in the middle phase. *Population mobility* elevated COVID-19 risk only in the early phase for South Korea's northern, easter, and western parts.

We investigated the correlations between all pairs of composite indices (Table S5). The largest Pearson's correlation coefficient was 0.603 between healthcare access and area morbidity. We verified no multicollinearity given that the model standard error of area morbidity was the smallest (0.0064) and healthcare access standard error was small (0.025) compared to the largest standard error, 0.09, associated with crowding. For comparison, GNBR and GWNBR models were carried out with the same variables and stratified by the same periods (NBR: Figure 3, and GWNBR: Figure 4). AIC and dispersion coefficients were used to compare the models' performances in computing COVID-19 risk associated with multiple health/SE themes across the country.

## Discussion

GWNBR created a continuous surface of relative COVID-19 risk for all 250 districts associated with area-health and socioeconomic determinants by the pandemic phases (Figure

4). Our findings are consistent with individual and population-level studies that reported elevated COVID-19 risk associated with less *healthcare access*[21], and *education*[22, 23], and more *risky health behavior*, *crowding*, *specific comorbidities*[13, 14, 17], *difficulty to social distancing* [15, 24] and *population mobility*[25]. Our study's high internal validity was shown since the GNBR and GWNBR results agreed except for *crowding* in the early phase.

Our approach captured significant spatial variation by pandemic phases for all themes, consistent with the reported pattern of COVID-19 distribution in the country. Since its first confirmed case on January 20th, 2020, South Korea experienced two major outbreak waves in Daegu and Seoul, and the surrounding Gyeonggi-do province, respectively, in February (early phase) and May 2020 (late phase). The country responded to the first wave with nationwide directives that included mass testing based on contact tracing, self-quarantine/isolation, strengthening medical centers for rapid diagnostics, emergency medical responses, and treatment aids[16]. These specific measures combined with high public adherence to the school and business closures, personal hygiene, and social distancing significantly dropped the case counts until the second wave erupted in May when businesses reopened[26].

SES-related risk declined over the pandemic phases. Analysis stratified by periodic phases found that the initially high risk in the early period gradually decreased in strength and spatial coverage in the middle and late phases (Figure 3 & Figure 4). The exceptions were health behavior and crowding- associated risk, which increased in the late phase. Reductions could be explained by the impact of effective control measures that lowered the risk associated with these determinants and/or a drop in an effective reproduction number (Re), as the number of infection-susceptible people decreased over time[27].

*Types of significant SES risk factors varied over the pandemic phases.* In the early phase, all health/SE themes were significant risk factors. In the middle phase, all of the previous risk factors except for risky *health behaviors, population mobility*, and *crowding* were significant.

In the late phase, only increased risky *health behavior*, increased *crowding* and reduced *healthcare access* remained significantly associated with COVID-19 incidence. The decrease in risk associated with risky *health behaviors*, *population mobility* and *crowding* in the middle phase could reflect the impact of the Prime Minister's declaration implementing active interventions for social distancing, community health education, testing with local contact tracing and screening center followed ups with quarantine and isolation during the first weeks of March. Notably, the late phase findings are consistent with the risk factors reported associated with the second wave in early May. During the late phase, South Korea experienced the culmination of strain in its struggling health system[26] which may have led to reduced *healthcare access* as indicated by our findings. In addition, our findings may have captured behavioral fatigue at the population-scale as a response to the country's multiple quarantine period extensions[28]. This was observed in other countries that likely were exacerbated by business re-openings in early May. The elevated risk associated with increased *crowding* occurred during the second wave of outbreak and happened in Seoul and its surroundings, the most crowded region of the country. In contrast, the elevated risk associated with decreased crowding in the early pandemic reflects the population density of the associated locales. The health behavior-associated increase in risk is consistent with reports showing increased risk with poor *emotional health*[22], *smoking*[29], and *obesity*[14, 30]. Lusignan et al.,2020[14] reported smoking was a protective factor; however, the authors warned that the proportion of current smokers in their study sample was low (11.4%), resulting in a wide confidence interval of the reported odds ratio, 0.59 (0.42-0.83). This increases the uncertainty of this result. Increased COVID-19 risk among the people with lower education has been explained as a reduced *awareness of disease risk* and *low-income*. Thus, not being able to afford losing employment resulted in reduced exercise of protective measures[22, 23].

*Spatial variation in the SES-related risk factors across the pandemic phases potentially reflect the geography-specific control measures and/or the differential public response to the*

*measures.* GWNBR models revealed the pandemic phase-specific spatial variation for all health/SE themes except for population mobility which was not significant beyond the early phase. This may indicate that the effectiveness of the control measures varied over time potentially due to the differential interventions or the public response across the municipal districts. Our findings may also indicate a dynamic change in population vulnerability throughout the pandemic "a person not considered vulnerable at the outset of a pandemic can become vulnerable depending on the policy response" as a *Lancet* editorial stated[31].

The factors increasing our recommended framework's robustness include: 1) SES measurement and relationship conceptualization of the exposure (health/SE themes) and outcome (COVID-19 incidence) based on the refined conceptual framework; 2) joint use of conceptual and statistical modeling; 3) complementary use of global and local spatial statistics; and 4) stratified analysis by pandemic phases that enable us to capture the spatial variation over pandemic phases. However, this methodological framework relies on carefully collected country-specific data.

Biases due to differential testing rates country-wide are low since the testing was based on contact tracing, and government-supported (free). Our study is subject to ecological fallacy inherent to the study design. However, our empty hierarchical mixed model accounting for the individual and district-level data shows that 61% of the COVID-19 incidence distribution variation was explained by the district-level factors, leaving 39% of the variability for explanation by individual factors. We verified that the data estimation for Daegu city subparts did not affect the study results significantly. The comparison of the intercept, standard error, relative risk and P-value between the models with and without the estimated data showed that the intercept and standard error were diminished by 2.2% and by 9.2%, respectively in the models including estimated data [each calculated by 100 × (-12,29 -(-12.56))/-12.29 and 100 × (1.88-2.95)/1.88, respectively]. A significance level change was observed for none of the model estimates, except

for *crowding.* The P-value changed from ~0.06 to ~0.04 when the estimated data were excluded. However, the *crowding-associated* risk remains significant at P = 0.1. Model details are provided in Table S2. To assess the periodic trend in the relative COVID-19 risk associated with SES factors, we conducted stratified analyses by the early, middle, and late phases corresponding with January 20-March 20, March 21-April 15, and April 16-July 1, 2020.

We intended this framework to improve international knowledge exchange and enable rapid pandemic responses in high-risk populations. To illustrate practical application of this framework in South Korea, decision-makers could have prioritized improved healthcare access and promoted protective health behaviors focusing more on the crowded areas in the capital and surrounding regions in the second wave of the pandemic. Such precision targeting can bolster preventive measures to reduce the healthcare burden and economic damage.

**Acknowledgments**

The work described here was not funded by any source. Dr. Kouzoukas receives research grant support from the US Department of Veterans Affairs (I21 RX003170 to DEK). The views expressed here are those of the authors and do not necessarily reflect the of any government agency or institution.

**Conflict of interest:** None declared.

**Data availability:** The data and SAS code underlying this article will be shared on reasonable request to the corresponding author.

## References


1. Delaere PR, Hermans R. Clinical transplantation of a tissue-engineered airway. *Lancet* 2009; **373**: 717-8; author reply 8-9.

2. Blumenshine P, Reingold A, Egerter S, Mockenhaupt R, Braveman P, Marks J. Pandemic influenza planning in the United States from a health disparities perspective. *Emerg Infect Dis* 2008; **14**: 709-15.

3. Jarman B, Townsend P, Carstairs V. Deprivation indices. *BMJ* 1991; **303**: 523.

4. Surgo Foundation. The COVID-19 Community Vulnerability Index (CCVI). https://precisionforcovid.org/ccvi; 2020.

5. Mollalo A, Vahedi B, Rivera KM. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Sci Total Environ* 2020; **728**: 138884.

6. Zhu Y, Chen YQ. On a Statistical Transmission Model in Analysis of the Early Phase of COVID-19 Outbreak. *Stat Biosci* 2020: 1-17.

7. da Silva AR, Rodrigues TCV. Geographically Weighted Negative Binomial Regression—incorporating overdispersion. *Statistics and Computing* 2013; **24**: 769-83.

8. Ma J. Estimating epidemic exponential growth rate and basic reproduction number. *Infect Dis Model* 2020; **5**: 129-41.

9. Korea Centers for Disease Control and Prevention. Coronavirus Infectious Disease-19 Outbreak in Korea (Regular Briefing on July 1). http://ncov.mohw.go.kr/; 2020.

10. Kim J. DS4C: Data Science for COVID-19 in South Korea. https://www.kaggle.com/kimjihoo/coronavirusdataset; 2020.

11. Coleman JS. *Foundations of Social Theory*: Harward University Press; 1998.

12. Oakes JM, Rossi PH. The measurement of SES in health research: current practice and steps toward a new approach. *Soc Sci Med* 2003; **56**: 769-84.

13. Covid-19 National Emergency Response Center E, Case Management Team KCfDC, Prevention. Coronavirus Disease-19: The First 7,755 Cases in the Republic of Korea. *Osong Public Health Res Perspect* 2020; **11**: 85-90.

14. de Lusignan S, Dorward J, Correa A, et al. Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. *The Lancet Infectious Diseases* 2020; **20**: 1034-42.

15. Koh D. Occupational risks for COVID-19 infection. *Occup Med (Lond)* 2020; **70**: 3-5.

16. Oh J, Lee JK, Schwarz D, Ratcliffe HL, Markuns JF, Hirschhorn LR. National Response to COVID-19 in the Republic of Korea and Lessons Learned for Other Countries. *Health Syst Reform* 2020; **6**: e1753464.

17. Richardson S, Hirsch JS, Narasimhan M, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* 2020; **323**: 2052-9.

18. Korea S. *KOSIS. KOrean Statistical Information System* 2020 [cited; Available from: http://kosis.kr/index/index.do

19. Wichern RAJDW. *Applied Multivariate Statistical Analysis*. 4 ed: Pearson; 1998.

20. da Silva AR, Fotheringham AS. The Multiple Testing Issue in Geographically Weighted Regression. *Geographical Analysis* 2016; **48**: 233-47.

21. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health* 2020; **8**: e480.

22. Bavel JJV, Baicker K, Boggio PS, et al. Using social and behavioural science to support COVID-19 pandemic response. *Nat Hum Behav* 2020; **4**: 460-71.

23. Bruine de Bruin W, Bennett D. Relationships Between Initial COVID-19 Risk Perceptions and Protective Health Behaviors: A National Survey. *Am J Prev Med* 2020; **59**: 157-67.

24. Quinn SC, Kumar S, Freimuth VS, Musa D, Casteneda-Angarita N, Kidwell K. Racial disparities in exposure, susceptibility, and access to health care in the US H1N1 influenza pandemic. *Am J Public Health* 2011; **101**: 285-93.

25. Jia JS, Lu X, Yuan Y, Xu G, Jia J, Christakis NA. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* 2020; **582**: 389-94.

26. Her M. How Is COVID-19 Affecting South Korea? What Is Our Current Strategy? *Disaster Med Public Health Prep* 2020: 1-3.

27. Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet* 2020; **395**: 931-4.

28. Brooks SK, Webster RK, Smith LE, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet* 2020; **395**: 912-20.

29. Vardavas CI, Nikitara K. COVID-19 and smoking: A systematic review of the evidence. *Tob Induc Dis* 2020; **18**: 20.

30. Lighter J, Phillips M, Hochman S, et al. Obesity in Patients Younger Than 60 Years Is a Risk Factor for COVID-19 Hospital Admission. *Clin Infect Dis* 2020; **71**: 896-7.

31. Lancet T. Redefining vulnerability in the era of COVID-19. *Lancet* 2020; **395**: 1089.

32. Quinn SC, Kumar S. Health inequalities and infectious disease epidemics: a challenge for global health security. *Biosecur Bioterror* 2014; **12**: 263-73.

## Tables

**Table 1.** PCA details showing factor scores with their weights, and selected area health and SES variables and thematic composite indices

| Health/SE themes | Selected variables from national surveys | PCA, Factor[a] | Weights | Quartiles of composite indices | | | |
|---|---|---|---|---|---|---|---|
| | | | | 25th pctl | Median | 75th pctl | Max |
| Healthcare access | Healthcare utility rate | 0.892 | 0.448 | 18.0 | 21.0 | 29.6 | 44.2 |
| | Insurance coverage rate | 0.869 | 0.437 | | | | |
| | Unmet healthcare needs | 0.226 | 0.115 | | | | |
| Health behavior | % people with obesity, by measurement | 0.955 | 0.438 | 41.6 | 44.0 | 45.7 | 49.9 |
| | % people drunk alcohol <1 per month | 0.936 | 0.429 | | | | |
| | % people who currently smoke | 0.220 | 0.100 | | | | |
| | % people obese, self-reported | 0.069 | 0.033 | | | | |
| Crowding | Area per person | 0.900 | 0.402 | 11.6 | 12.9 | 14.6 | 22.6 |
| | Urbanicity | 0.734 | 0.329 | | | | |
| | N of people per household | 0.602 | 0.269 | | | | |
| Area morbidity | Overall, AAMR☐ | 0.919 | 0.190 | 88.1 | 96.7 | 102.7 | 116.9 |
| | Respiratory, AAMR | 0.838 | 0.173 | | | | |
| | Circulatory, AAMR | 0.743 | 0.153 | | | | |
| | Infectious and parasitic diseases, AAMR | 0.702 | 0.145 | | | | |
| | % people with severe disability | 0.689 | 0.142 | | | | |
| | % people on diabetes treatment | 0.483 | 0.099 | | | | |
| | % people with stroke symptoms | 0.436 | 0.090 | | | | |
| | % people with mental health diseases | 0.020 | 0.008 | | | | |
| Difficulty to social distancing | % people living in apartment buildings | 0.794 | 0.286 | 9.2 | 13.3 | 15.4 | 27.6 |
| | % workers in retail services | 0.749 | 0.269 | | | | |
| | N of students per class, high school | 0.677 | 0.244 | | | | |
| | % workers in health and social services | 0.558 | 0.201 | | | | |
| Population mobility | Net migration between districts | 0.732 | 0.500 | 0.7 | 1.1 | 1.9 | 7.2 |
| | % foreign residents in the area | 0.732 | 0.500 | | | | |

[a]Age-adjusted mortality rate

**Table 2.** Parameter estimates and 95% CI of the relative risk of COVID-19 associated with health and SES determinants during the entire study period

| Health/SE themes[a] | Estimates | Relative Risk (95% CI) | P-value |
|---|---|---|---|
| Healthcare access | -0.13 | 0 .88 (0 .84 - 0 .92) | < .0001 |
| Health behavior | 0.04 | 1 .04 (1 .01 - 1 .07) | 0.019 |
| Crowding | 0.05 | 1 .05 (0 .89 - 1 .25) | 0.545 |
| Area morbidity | 0.04 | 1 .04 (1 .03 - 1 .06) | < .0001 |
| Education | -0.09 | 0 .91 (0 .86 - 0 .97) | 0.002 |
| Difficulty to social distancing | 0.06 | 1 .06 (1 .01 - 1 .12) | 0.017 |
| Population mobility | -0.22 | 0 .80 (0 .69 - 0 .93) | 0.003 |
| Dispersion[b] | 2.49 | | |
| AIC[c] | 1850 | | |

[a]Data for the entire study period (January 20 - July 1, 2020)
[b]The variance of a negative binomial distribution
[c]Akaike information criterion (AIC), a measure of goodness of model fit

**Figure Legends**

**Figure 1. Conceptual model of the causal relationship between the SARS-CoV-2 and area health/SE determinants.** Subscripts (i, j, k): Number of variables used from the data sources. Data sources: Korean Community Health Survey by the KCDC, Health Insurance Statistics by the National Health Insurance Services, Disability Status by the Ministry of Health and Welfare, Death Cause Statistics by the National Statistics Agency, Korean Census Bureau, Internal Migration Statistics by the Statistics Korea, and the State of Urban Planning Report by the Ministry of Land, Infrastructure, Transport, and Tourism. *Material, Human and Social Capital refers to latent structural components of the area health and SES determinants Health and SES connected by arrows indicate the integration of area health and SES as a composite exposure in this study, hereinafter, denoted as a health/SE. ☐ Area-health/SE themes identified relevant to COVID-19 based on the person and population-level literature. Modified images from source[2, 32].

**Figure 2. The spatial distribution of COVID-19 cases across pandemic phases.** Early phase: from January 20 to March 20, 2020. Middle phase: March 21 to April 15, 2020. Late phase: April 16 to July 1, 2020. The background map in blue shades indicates the spatial variation of the population density. A blue color gradient corresponds with a higher (darker) to lower (lighter) population density.

**Figure 3. Relative Risk of COVID-19 associated with area health and SES determinants.** Each panel shows the relative risk and its 95% confidence intervals associated with each thematic area. Colors represent the pandemic's early (January 20 to March 20, 2020), middle (March 21 to April 15, 2020), and late phases (April 16 to July 1, 2020). The dashed line shows the reference level (1). Values over or below the reference line indicate statistically significant results at $\alpha = 0.05$.

**Figure 4. Spatial variation in the relative risk of COVID-19 associated with area-health and SES themes in the early, middle, and late phases of the pandemic, GWNBR models.** In the maps, the blue color gradient corresponds with larger (darker) to lower (lighter) relative risk. Areas in white indicate the relative risks statistically not significant ($\alpha = 0.05$). Columns titled Early, Middle, and Late phases refer to the pandemic's early (January 20 to March 20, 2020), middle (March 21 to April 15, 2020), and late phases (April 16 to July 1, 2020).

## Supplementary tables.

**Table S1.** Confirmed cases of COVID-19 diagnosed during January 20 - July 1, 2020 in South Korea by pandemic phases

| Province | Phase 1 | Phase 2 | Phase 3 |
|---|---|---|---|
| Busan | 99 | 18 | 24 |
| Chungcheongbuk | 34 | 11 | 12 |
| Chungcheongnam | 85 | 17 | 28 |
| Daegu | 6358 | 529 | 35 |
| Daeguon | 21 | 17 | 80 |
| Gangwon | 30 | 20 | 13 |
| Gwangju | 48 | 2 | 3 |
| Gyeonggi | 324 | 325 | 567 |
| Gyeongsangbuk | 1098 | 111 | 45 |
| Gyeongsangnam | 86 | 26 | 17 |
| Incheon | 38 | 47 | 253 |
| Jeju | 3 | 9 | 6 |
| Jeollabuk | 8 | 9 | 4 |
| Jeollanam | 4 | 10 | 0 |
| Sejong | 34 | 10 | 7 |
| Seoul | 298 | 278 | 656 |
| Ulsan | 32 | 8 | 15 |

**Table S2.** GNBR model estimates with and without the estimated data for Daegu's subparts

| | Included estimated data | | | | Excluded estimated data | | | |
|---|---|---|---|---|---|---|---|---|
| **Health/SE themes[a]** | **Estimate[b]** | **SE** | **RR (LCL - UCL)** | **P-value** | **Estimate** | **SE** | **RR (LCL - UCL)** | **P-value** |
| Area-morbidity | 0.05 | 0.01 | 1.05 (1.03 - 1.06) | <.0001 | 0.04 | 0.01 | 1.04 (1.03 - 1.06) | <.0001 |
| Education | -0.11 | 0.04 | 0.90 (0.83 - 0.97) | 0.005 | -0.12 | 0.04 | 0.89 (0.82 - 0.96) | 0.003 |
| Crowding | 0.22 | 0.12 | 1.25 (0.99 - 1.57) | 0.06 | 0.27 | 0.13 | 1.30 (1.01 - 1.69) | 0.04 |
| Difficulty to social distancing | 0.07 | 0.03 | 1.07 (1.01 - 1.14) | 0.02 | 0.08 | 0.03 | 1.08 (1.02 - 1.15) | 0.01 |
| Population mobility | -0.38 | 0.09 | 0.69 (0.57 - 0.82) | <.0001 | -0.34 | 0.09 | 0.71 (0.59 - 0.85) | 0.0002 |
| Healthcare access | -0.14 | 0.03 | 0.87 (0.82 - 0.93) | <.0001 | -0.13 | 0.03 | 0.88 (0.82 - 0.93) | <.0001 |
| Health behavior | 0.04 | 0.02 | 1.04 (1.00 - 1.08) | 0.03 | 0.04 | 0.02 | 1.04 (1.00 – 1.08) | 0.05 |
| Dispersion | 3.68 | | | | 3.69 | | | |
| AIC | 1527 | | | | 1425 | | | |

[a]Study period: from January 20 through July 1, 2020
[b]Parameter estimates; SE: Standard error; RR: Relative Risk; LCL: lower boundary of 95% confidence interval; UCL: upper boundary of 95% confidence internal.

**Table S3.** Data sources reviewed and used for SES measurement

| Data sources | Data items |
|---|---|
| Korean Community Health Survey 2018, Korea Centers for Disease Control and Prevention | (%) people with obesity, by measurement<br>(%) people drunk alcohol <1 per month<br>(%) people who currently smoke<br>(%) people obese, self-reported<br>(%) of people who used health care last year<br>(%) people who could not use healthcare when needed last year<br>(%) people with depression based of PHQ-9[a] screening<br>(%) of the person who is all aligned for the early symptoms of stroke<br>(%) of the person on insulin and other treatment specific to diabetes mellitus |
| Health Insurance Statistics 2018, National Health Insurance Corporation | (%) people with health insurance |
| Disability Status 2018, Ministry of Health and Welfare | (%) people with severe disability |
| Death Cause Statistics 2018, National Statistics Agency | Age adjusted mortality rate due to neoplasm<br>Age adjusted mortality rate due to circulatory system disease<br>Age-adjusted mortality rate from infectious parasitic diseases<br>Overall, age-adjusted mortality rate<br>Age adjusted mortality rate due to respiratory diseases |
| Korean Census Bureau 2015 | (%) people with high school education<br>(%) of foreign registered people<br>Number of people per household |
| Office of Statistics 2015, Regional Statistics | GDP per capita in million won |
| Internal Migration Statistics 2018, Statistics Korea | Internal net migration between regions |
| State of Urban Planning 2018, Ministry of Land, Infrastructure, Transport and Tourism | Area per capita<br>Urban area per capita |

[a]PHQ-9: Patient Health Questionnaire -9, standard survey tool

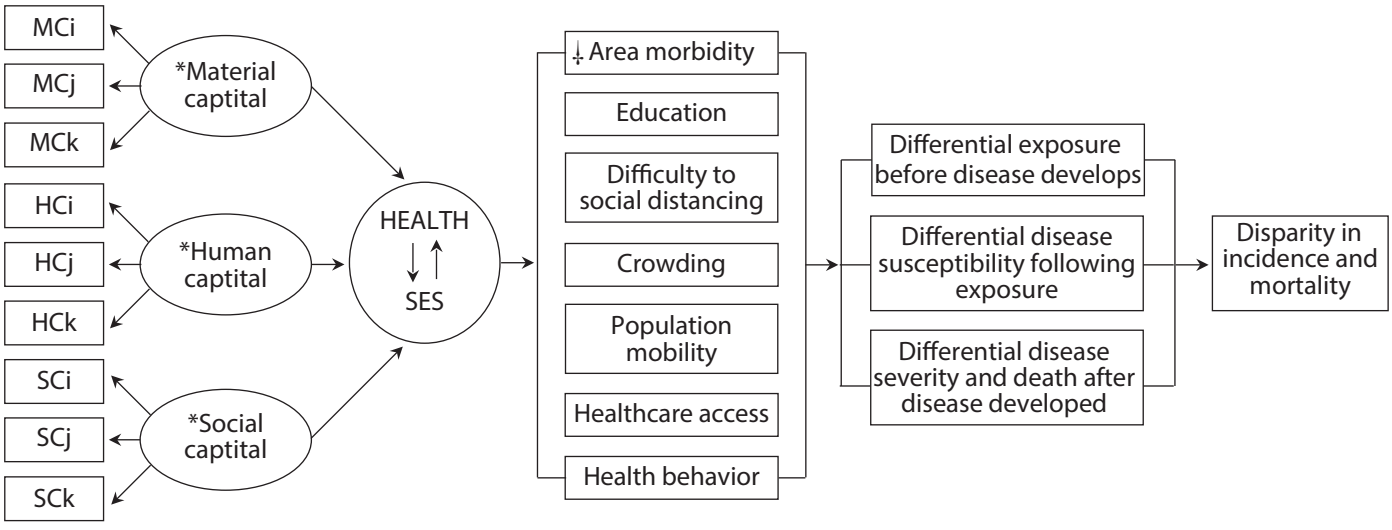
**Table S4.** Parameter estimates and the Relative Risk of the COVID-19 incidence associated with health and SES determinants by three time periods corresponding with the early, middle and late phases

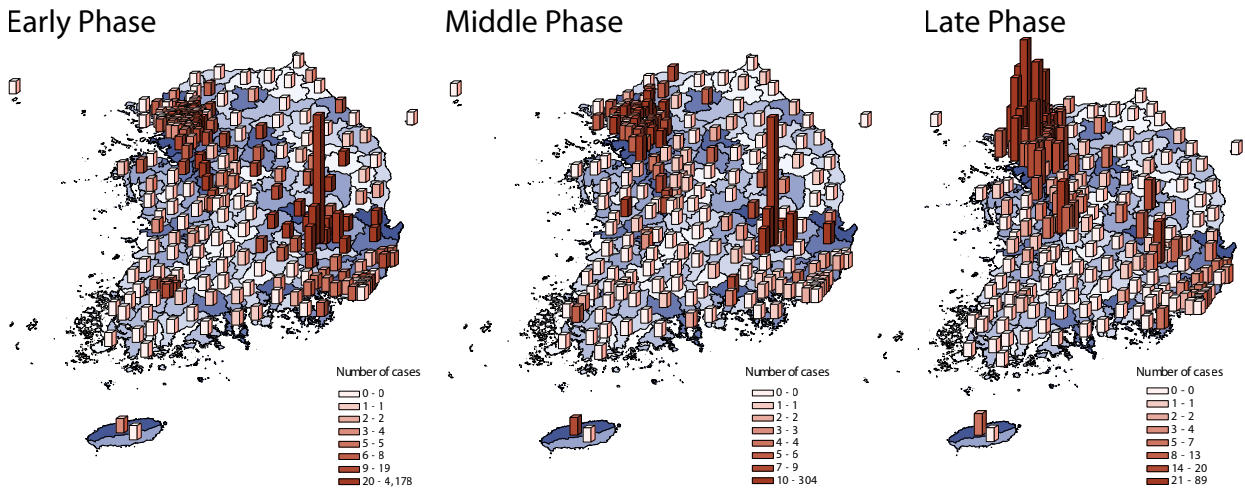
| | Early phase | | | Middle phase | | | Late phase | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Estimates** | **RR (LCL - UCL)** | **P-value** | **Estimates** | **RR (LCL - UCL)** | **P-value** | **Estimates** | **RR (LCL - UCL)** | **P-value** |
| Healthcare access | -0.14 | 0.87 (0.82 - 0.93) | <.0001 | -0.13 | 0.88 (0.84 - 0.93) | <.0001 | -0.09 | 0.92 (0.87 - 0.96) | 0.001 |
| Health behavior | 0.04 | 1.04 (1.00 - 1.08) | 0.028 | 0.03 | 1.03 (1.00 - 1.06) | 0.080 | 0.05 | 1.05 (1.02 - 1.08) | 0.001 |
| Crowding | 0.22 | 1.25 (0.99 - 1.57) | 0.058 | -0.05 | 0.96 (0.83 - 1.10) | 0.512 | -0.23 | 0.79 (0.70 - 0.89) | 0.000 |
| Area morbidity | 0.05 | 1.05 (1.03 - 1.06) | <.0001 | 0.04 | 1.04 (1.02 - 1.06) | <.0001 | 0.003 | 1.00 (0.99 - 1.02) | 0.729 |
| Education | -0.11 | 0.90 (0.83 - 0.97) | 0.005 | -0.08 | 0.92 (0.88 - 0.97) | 0.003 | -0.03 | 0.97 (0.93 - 1.02) | 0.223 |
| Difficulty to social distancing | 0.07 | 1.07 (1.01 - 1.14) | 0.021 | 0.06 | 1.06 (1.02 - 1.11) | 0.006 | 0.005 | 1.00 (0.96 - 1.05) | 0.843 |
| Population mobility | -0.38 | 0.69 (0.57 - 0.82) | <.0001 | -0.02 | 0.98 (0.84 - 1.14) | 0.791 | 0.14 | 1.15 (0.99 - 1.33) | 0.061 |

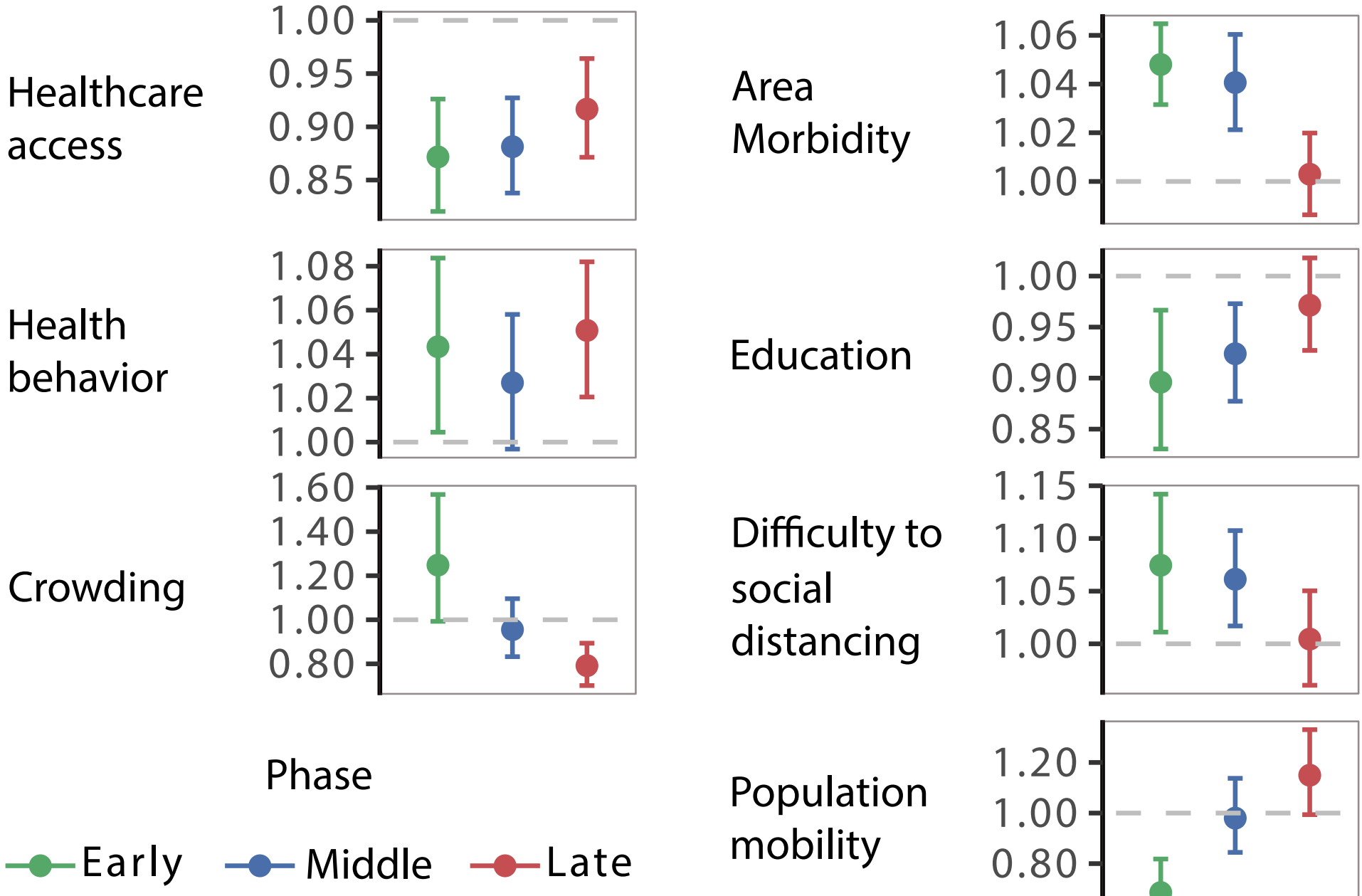
Early phase: January 20 to March 20, 2020; Middle phase: March 21 to April 15, 2020; Late phase: April 16 to July 1, 2020
RR: Relative Risk; LCL: lower boundary of 95% confidence interval; UCL: upper boundary of 95% confidence internal.

**Table S5.** Matrix table of Pearson's correlation coefficients (r)

| | **Healthcare access** | **Health behavior** | **Crowding** | **Area morbidity** | **Education** | **Difficulty to social distancing** | **Population mobility** |
|---|---|---|---|---|---|---|---|
| Healthcare access | 1 | | | | | | |
| Health behavior | 0.547 | 1 | | | | | |
| Crowding | 0.568 | 0.124 | 1 | | | | |
| Area morbidity | 0.604 | 0.534 | 0.415 | 1 | | | |
| Education | 0.127 | 0.390 | -0.080 | 0.510 | 1 | | |
| Difficulty to social distancing | -0.032 | 0.370 | -0.047 | 0.311 | 0.315 | 1 | |
| Population mobility | -0.096 | 0.086 | -0.019 | 0.018 | 0.163 | 0.129 | 1 |


MCi
MCj
MCk
*Material captital
HCi
HCj
HCk
*Human captital
SCi
SCj
SCk
*Social captital
HEALTH
SES
Area morbidity
Education
Difficulty to social distancing
Crowding
Population mobility
Healthcare access
Health behavior
Differential exposure before disease develops
Differential disease susceptibility following exposure
Differential disease severity and death after disease developed
Disparity in incidence and mortality


Early Phase
Number of cases
0 - 0
1 - 1
2 - 2
3 - 4
5 - 5
6 - 8
9 - 19
20 - 4,178
Middle Phase
Number of cases
0 - 0
1 - 1
2 - 2
3 - 3
4 - 4
5 - 6
7 - 9
10 - 304
Late Phase
Number of cases
0 - 0
1 - 1
2 - 2
3 - 4
5 - 7
8 - 13
14 - 20
21 - 89

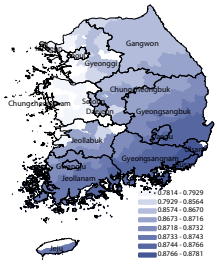

Healthcare access
1.00
0.95
0.90
0.85
Area Morbidity
1.06
1.04
1.02
1.00
Health behavior
1.08
1.06
1.04
1.02
1.00
Education
1.00
0.95
0.90
0.85
Crowding
1.60
1.40
1.20
1.00
0.80
Difficulty to social distancing
1.15
1.10
1.05
1.00
Population mobility
1.20
1.00
0.80
0.60
Phase
Early
Middle
Late

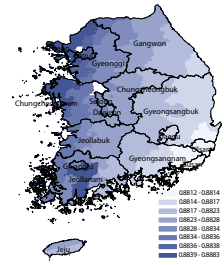
Phase 1
( Jan. 20- Mar. 20, 2020)

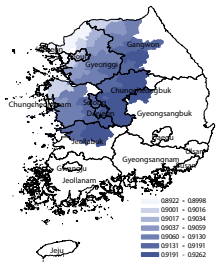
Phase 2
(Mar. 21 - Apr. 15, 2020)

Phase 3
(Apr. 16 - Jul. 1, 2020)

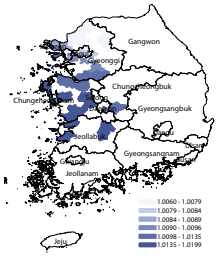
Healthcare Access

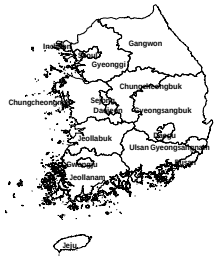


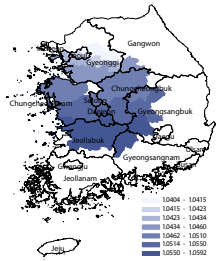



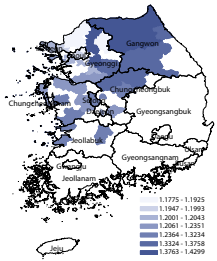

Health Behavior

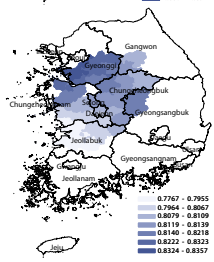


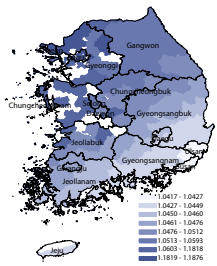



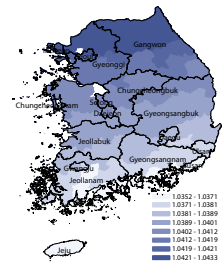

Crowding



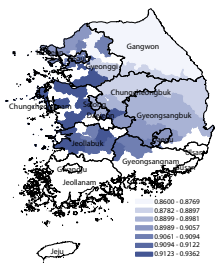


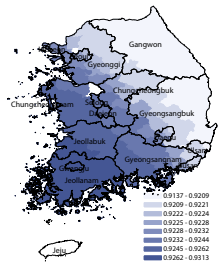

Area Morbidity

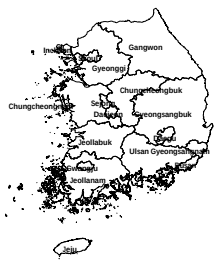


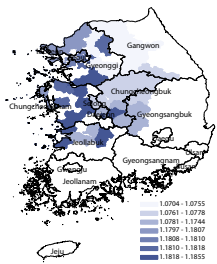
Education

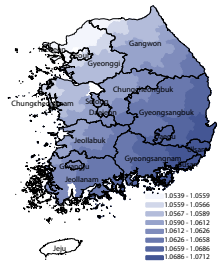
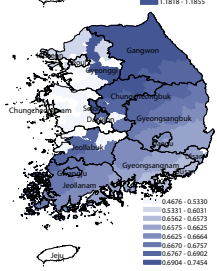
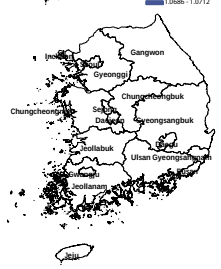
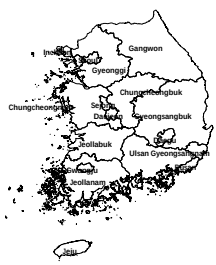



Difficulty to Social Distancing






Population Mobility