Original research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–18

DOI: 10.1177/20552076231207573
journals.sagepub.com/home/dhj

Sage

Investigation of factors associated with mental health during the early part of the COVID-19 pandemic in South Korea based on machine learning algorithms: A cohort study

Junggu Choi[1] and Sanghoon Han[1,2]

Abstract

Objective: The coronavirus disease 2019 (COVID-19) pandemic is among the most critical public health problems worldwide in the last three years. We tried to investigate changes in factors between pre- and early stages of the COVID-19 pandemic.

Methods: The data of 457,309 participants from the 2019 and 2020 Community Health Survey were examined. Four mental health-related variables were selected for examination as a dependent variable (patient health questionnaire-9, depression, stress, and sleep time). Other variables without the aforementioned four variables were split into three groups based on the coefficient values of lasso and ridge regression models. The importance of each variable was calculated and compared using feature importance values obtained from three machine learning algorithms.

Results: Psychiatric and sociodemographic variables were identified, both during the pre- and early pandemic periods. In contrast, during the early pandemic period, average sleep time variables ranked the highest with the dependent variables regarding the experience of depression. The difference in sleep time before and after the pandemic was validated by the results of paired *t*-tests, which were statistically significant (p-value < 0.05).

Conclusions: Changes in the importance of mental health factors in the early pandemic period in South Korea were identified. For each mental health-dependent variable, average sleep time, experience of depression, and experience of accidents or addictions were found to be the most important factors. House type and type of residence were also found in regions with larger populations and a higher number of confirmed cases.

Keywords

COVID-19, pre- and early pandemic periods, mental health, Community Health Survey, machine learning

Submission date: 21 March 2023; Acceptance date: 28 September 2023

Introduction

In the last three years, the coronavirus disease 2019 (COVID-19) pandemic has impacted global public health with rapid changes in both medical practices and society's daily life.[1–3] To investigate the impacts of the pandemic, many researchers have attempted to analyze variations in related phenomena in diverse domains. Olszewska-Guizzo et al.[4] compared hemodynamic responses in urban spaces with lockdowns during the pandemic to verify the effects of changes in the environment. Almeida et al.[5] examined the effect of the pandemic on women's mental health. Furthermore, Xiong et al.[6] systematically reviewed relevant

[1]Yonsei Graduate Program in Cognitive Science, Yonsei University, Seoul, Republic of Korea
[2]Department of Psychology, Yonsei University, Seoul, Republic of Korea

Corresponding author:
Sanghoon Han, Department of Psychology and Yonsei Graduate Program in Cognitive Science, Yonsei University, Seoul 03722, Republic of Korea.
Email: sanghoon.han@yonsei.ac.kr

literature on the effects of COVID-19 on psychological outcomes and associated risk factors.

In various studies, including those mentioned above, researchers focused on mental health-related issues resulting from COVID-19 to identify the relationship between the pandemic and mental health among diverse research topics. Bojdani et al.[7] proposed novel guidelines for psychiatric care in the United States. They suggested a role for policymakers in the healthcare system that goes beyond care guidelines. Wang et al.[8] found the potential for negative psychological and social effects resulting from quarantines and isolation in China. Furthermore, researchers have investigated the multifactorial impacts of COVID-19 in terms of biological, environmental, and social aspects.[9] Particularly, researchers have found that as a stressor, the COVID-19 pandemic can trigger neuropsychiatric outcomes (e.g. neuroinflammation and behavioral impairment) in adults. In addition, O'Connor et al.[10] traced adult groups living in the United Kingdom to determine the trajectory of mental health and well-being in the first six weeks of lockdown; it was confirmed that the rate of suicidal thoughts increased over time.

Various factors associated with mental health in a pandemic have already been proven in previous studies. De Figueiredo et al.[11] examined a multifactorial influence of COVID-19 on children and adolescent populations. Possibilities of stressors in a pandemic to neuroinflammation and behavioral impairments were proven in their analysis results. Wu et al.[12] verified that good marital relationships and social support should be checked to investigate the mental health status in the students' parents population during the COVID-19 pandemic. Magson et al.[13] found that policies concerning government restrictions and concerns about the spread of the virus were associated with increased anxiety, symptoms of depression, and decreased life satisfaction. Furthermore, Blix et al.[14] confirmed the relationship between higher levels of COVID-related worry and higher psychological stress in the general Norwegian population.

To validate the effectiveness of factors, the periods of the pandemic were considered as important criteria in the study design of previous studies. Kwong et al.[15] separated the duration of follow-up for comparison between the pre- and post-pandemic periods. Based on their experimental results, the authors provided initial indications of anxiety and depression in the younger population. Ravens-Sieberer et al.[16] conducted a nationwide survey involving two successive waves during the pandemic to investigate longitudinal changes in mental health and identify risk and resource factors.

Multiple variables from large-scale datasets have been analyzed to identify complex associations between mental health and factors from the pandemic. Diverse methodologies, including the use of statistical models and tests, have also been applied to analyze large-scale and longitudinal datasets. Wang et al.[17] used Bayesian generalized compartmental models to determine the effects of public health interventions on hospitals during the pandemic. Recently, machine learning (ML) and deep learning models have been widely used to examine the latent patterns in datasets. According to these trends, some researchers have applied ML algorithms to datasets to validate their research hypotheses. Khattar et al.[18] selected an ML approach to analyze the effects of the pandemic on the mental health of young Indian students. During a period of lockdown, feelings of frustration, crushing boredom, and anxiousness were highlighted, which were identified through topic modeling. In addition, Rezapour and Hansen[19] examined important factors for predicting mental health decline in frontline worker groups using ML algorithms. They found that the amount of sleep individuals had and the amount of COVID-19-related news they were exposed to were related to individuals' mental health status.

Based on the aforementioned studies, in this study, we attempted to investigate variations of associated factors between the pre- and early pandemic periods using ML algorithms. To confirm the effects of the pandemic on the importance of these factors, 2019 and 2020 datasets were used from a longitudinal survey dataset collected from 2008 to 2020 in South Korea. Moreover, we included as many variables as possible, including psychiatric and socioeconomic variables, to analyze from various perspectives. The datasets were divided by the 16 regions of South Korea to reflect additional information such as population density or the number of confirmed cases. Furthermore, we split variables into three groups based on the magnitude of the coefficient from the lasso and ridge regression model to interpret the relative importance of each variable in the group in detail. Finally, the feature importance of variables from the trained ML algorithms was checked to compare relative importance. The overall scheme of the study design is shown in Figure 1.

Methods

Data source

We used the open-source Community Health Survey (CHS) dataset in this study. This dataset was released by the Korea Disease Control and Prevention Agency (KDCA)[20,21] to compare mental health factors during the pre- and early pandemic periods in South Korea. The CHS investigates the health status of people living in various regions of Korea through longitudinal surveys. It has been conducted annually by the KDCA since 2008. In total, 228,303 respondents participated in the survey from 2008 to 2020. The CHS dataset was composed by using 49 categories of survey variables. Detailed categories of variables are listed in Table 1.

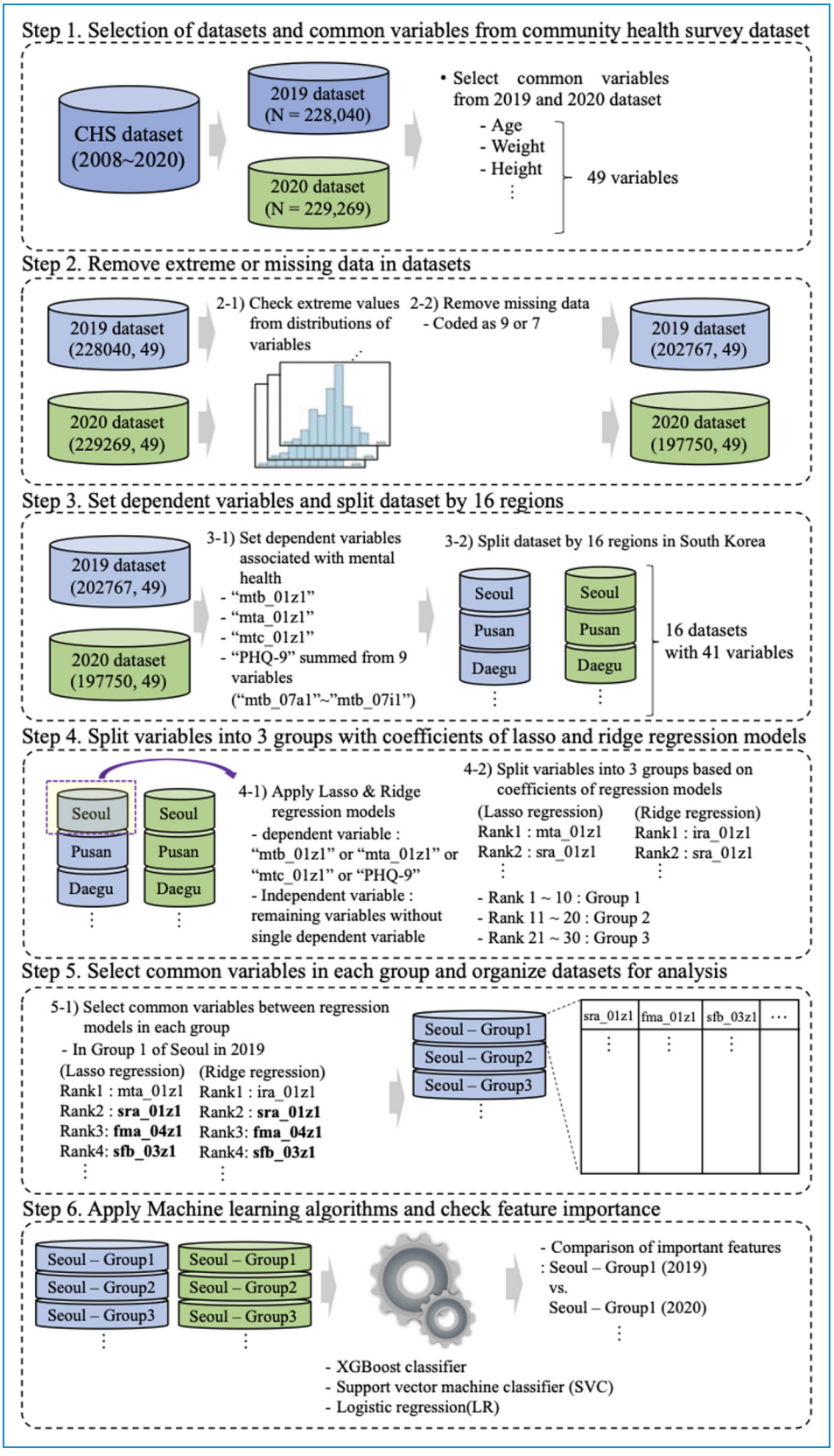

Figure 1. Study design.

Moreover, we used additional public datasets on population density and the number of confirmed cases of COVID-19 released by the Korean Statistical Information Service to investigate the influence of population and COVID-19 severity in the region.

Preprocessing

Selection of associated variables from datasets. We used two datasets (2019 and 2020) to check for changes in important factors after the pandemic. The 2019 dataset was utilized to identify variations in the pre-pandemic period. Similarly, the 2020 dataset was used for a comparison between the early and pre-pandemic periods. To consider as many variables as possible in our analysis, we selected common variables across the two datasets (without excluding variables in specific categories). A total of 49 common variables distributed across 17 categories were selected. After the selection of the common variables, the dimension of the 2019 dataset was (228,040, 49) and that of the 2020 dataset was (229,269, 49). Detailed descriptions of the selected 49 variables are presented in Table 2.

Exclusion of missing and extreme data. The distribution of each variable was checked to remove extreme and missing values from the dataset. Missing or non-response values were coded as 9 or 7 in the CHS dataset. To reflect the exact responses of the participants in our analysis, the distributions of the variables were investigated using histograms. After removing data with outliers or missing values, the dimension of the 2019 dataset was (202,767, 49) and that of the 2020 dataset was (197,750, 49). The distributions of the variables in this study are shown in Figure 2.

Setting dependent variables and splitting datasets into 16 regions. To compare the important factors associated with mental health between the pre- and early pandemic periods, we set four dependent variables. First, the "mtb_01z1" variable indicates the experience of depression in the last year. The survey question for "mtb_01z1" was "During the past year, have you felt so sad or hopeless that it interfered with your daily life for more than two weeks in a row?" with binary answers. Second, "mta_01z1" denotes the subjective stress level. For this variable, answers in four category levels for the question "How much stress do you feel in your daily life?" were provided. Third, "mtc_01z1" indicates the average sleep time based on continuous answers. Answers for "mtc_01z1" were collected using the question, "How many hours per day do you usually sleep?" Finally, nine patient health questionnaire-9 (PHQ-9) variables ("mtb_07a1," "mtb_07b1," "mtb_07c1," "mtb_07d1," "mtb_07e1," "mtb_07f1," "mtb_07g1," "mtb_07h1," and "mtb_07i1") denote depression levels with four-category answers. These nine variables were converted to single variables by summing the values of the nine variables. After converting the PHQ-9 variable, 41 columns remained.

To compare factors between regional conditions, we split the single dataset including converted dependent variables to 16 datasets based on the 16 regions of South Korea (Seoul, Pusan, Daegu, Incheon, Gwangju, Daejeon, Ulsan, Gyeonggi, Kangwon, Chungbuk, Chungnam, Jeonbuk, Jeonnam, Gyeongbuk, Gyeongnam, and Jeju). The detailed dimensions of the 16 regions in 2019 and 2020 are listed in Appendix 1.

Feature selection and separation of variables into three groups. Before applying the dataset to ML classifiers, we utilized lasso and ridge regression models to select variables associated with four dependent variables ("mtb_01z1," "mta_01z1," "mtc_01z1," and "PHQ-9"). The coefficients of the variables from the lasso and ridge regression models were arranged in ascending order. Only the top 30 common variables were selected based on the magnitude of the coefficients.

Furthermore, to compare changes in the relative importance of each variable before and after the pandemic in detail, we divided the selected top 30 variables into three groups according to their rank of coefficients. Common variables between the arranged variable lists were selected for the lasso and ridge regression models. Variables included in each group were applied to ML algorithms to compare the influences of variables on dependent variables. The detailed coefficient values of the lasso and ridge regression models are listed in Appendix 2.

Evaluation of ML classification algorithms. To validate the importance changes in variables from trained ML algorithms in the simplest task condition, class labels of each dependent variable ("mtb_01z1," "mta_01z1," "mtc_01z1," and "PHQ-9") were converted to binary classes for matching class label conditions in dependent variables. In the case of "mta_01z1," with four categories, the first- and second-level answers were changed to the label "1." Similarly, third- and fourth-level answers were changed to the label "2." Furthermore, the values of the "mtc_01z1" variable were converted to labels "1" and "2" based on the median value of the variable. Moreover, "PHQ-9" variables were also converted to labels "1" and "2" through their value size.

According to these class label conditions (i.e. binary class), we applied three ML classification algorithms (support vector machine classifier (SVC), logistic regression (LR) model, and XGBoost classifier) in our study. Furthermore, to evaluate the performance of each classifier in a rigorous setting, 10-fold cross-validation (10-fold CV) was used in the training and evaluation steps. Moreover, to improve the imbalance in the number of class labels, we used weights for insufficient class labels during a 10-fold CV.

Table 1. Variable categories of the CHS dataset.

No.	Categories	No.	Categories	No.	Categories	No.	Categories
1	Smoking	14	Stroke	27	Thyroid lesion	40	Overactive bladder Dry
2	Alcohol	15	Cardiac infarction	28	AIDS	41	Hepatitis C
3	Sense of safety	16	Angina	29	Backache	42	Tuberculosis
4	Physical activity	17	Arthritis	30	Glaucoma	43	Medical use
5	Food life	18	Osteoporosis	31	Anemia	44	Prostate
6	Female health	19	Asthma	32	Hemorrhoids	45	Use health agency
7	Oral health	20	Work loss and quality of life	33	Gastroduodenal ulceration	46	Cardiopulmonary resuscitation
8	Vaccination and check-up	21	Education and economic activity	34	Obesity and weight control	47	Accident and poisoning
9	Illness and poisoning accident	22	Metabolic syndrome	35	Inflammation in the middle ear	48	Anthropometric survey and blood pressure
10	Hyperlipidemia	23	Hepatitis B	36	Personal hygiene	49	Basic information
11	Depression	24	Cataract	37	Urinary incontinence		
12	Mental health	25	Allergic rhinitis	38	Cancer		
13	Diabetes	26	Atopic dermatitis	39	Social environment		

CHS: Community Health Survey; AIDS: acquired immune deficiency syndrome.

Calculation of variable importance scores from feature importance results. We obtained 10 sets of feature importance score results (e.g. F-score for the XGBoost classifier, coefficient value for LR, and SVC) from the evaluated ML algorithms for each experimental condition. To compare the importance levels of the variables, 10 sets of scores were required to convert a single set of scores. We proposed a calculation formula for converting multiple sets of feature importance into a single set based on a previous study.[11] Based on this formula, we calculated single-importance scores for each variable. The detailed formula was as follows:

$$\text{Variable importance score} = \frac{\sum_{i=0}^{n}(1-\alpha_i)\times\beta_i}{n}$$

where α represents the normalized rank of variables ranging between 0 and 1, and β indicates the normalized importance score within the same range (i.e. from 0 to 1). As 10 score sets from the 10-fold CV were applied for the calculation, n was 10 in our cases.

ML classification algorithms. To identify important factors related to mental health from the datasets, three ML classification algorithms were utilized. The first algorithm was the XGBoost classifiers. This algorithm is an ensemble version of multiple decision-tree classification algorithms. This algorithm minimizes the errors between the predicted and target values through objective functions in the training process. The objective functions consist of differentiable convex loss functions and penalized terms for regularization.

The second classification algorithm was an SVC with linear kernels. This algorithm classifies the feature space in the datasets by hyperplanes to classify class labels. To verify the importance of the coefficient values from the algorithms, we used linear kernels in our cases. The third algorithm was LR. The coefficients of the regression model were estimated using maximum likelihood estimation methods. This models calculated the likelihood value $L(x)$ in ranges from 0 to 1 (i.e. $0 \leq L(x) \leq 1$). The likelihood value indicates the association between input data and class in the data. We utilized a random search to determine the optimal hyperparameters for the aforementioned three ML algorithms. Applied hyperparameters are listed in Table 3.

Evaluation criteria. The classification performance of each algorithm was evaluated using five evaluation criteria.

Table 2. Detailed descriptions of the selected 49 variables.

No.	Variable	Description of variable	Variable type	Variable category
1	id	Participant id	Continuous	Basic information
2	city_cd	Region number	Categorical	Basic information
3	town_t	Type of residence	Categorical	Basic information
4	apt_t	House type	Categorical	Basic information
5	sex	Gender of participant	Categorical	Basic information
6	age	Age of participant	Categorical	Basic information
7	fma_01z1	Number of whole household members	Continuous	Social environment
8	fma_02z1	Number of household members over the age of 19	Continuous	Social environment
9	fma_19z1	Generation type	Continuous	Social environment
10	fma_04z1	Whether or not to receive basic livelihood support	Categorical	Social environment
11	fma_13z1	Household income (year)	Continuous	Social environment
12	fma_14z1	Household income (year)	Continuous	Social environment
13	qoa_01z1	Subjective health levels	Categorical	Depression
14	sma_01z1	Lifetime smoking	Categorical	Smoking
15	sma_03z2	Currently smoking	Categorical	Smoking
16	dra_01z1	Lifetime drinking	Categorical	Alcohol
17	drb_03z1	Amount of alcohol consumed at one time	Categorical	Alcohol
18	drb_01z2	Annual drinking frequency	Categorical	Alcohol
19	sfb_05z2	Car drunk driving experience	Categorical	Sense of safety
20	sfb_03z2	Motorcycle drunk driving experience	Categorical	Sense of safety
21	phb_01z1	Number of walking days	Continuous	Physical activity
22	phb_02z1	Walking time (hour)	Continuous	Physical activity
23	phb_03z1	Walking time (minutes)	Continuous	Physical activity
24	oba_02z1	Height	Continuous	Basic information
25	oba_03z1	Weight	Continuous	Basic information
26	obb_01z1	Weight control experience	Categorical	Obesity and weight control
27	ora_01z1	Subjective oral health level	Categorical	Oral health

(continued)

Table 2. Continued.

No.	Variable	Description of variable	Variable type	Variable category
28	orb_01z1	Chewing discomfort experience	Categorical	Oral health
29	mtc_01z1	Average sleep time	Continuous	Mental health
30	mtb_01z1	Depression experience	Categorical	Depression
31	mta_01z1	Subjective stress levels	Categorical	Mental health
32	sca_01z1	Influenza vaccination	Categorical	Vaccination and check-up
33	hya_04z1	High blood pressure diagnosis experience	Categorical	Anthropometric survey and blood pressure
34	dia_04z1	Diabetes diagnosis experience	Categorical	Diabetes
35	sra_01z3	Failure to essential medical services	Categorical	Use health agency
36	ira_01z1	Accident or addiction experiences	Categorical	Accident and poisoning
37	ira_02z1	Number of accidents or addictions	Continuous	Accident and poisoning
38	soa_06z2	Job type	Categorical	Education and economic activity
39	sob_01z1	Education level	Categorical	Education and economic activity
40	sob_02z1	Graduation from an educational institution	Categorical	Education and economic activity
41	mtb_07a1	(PHQ-9: Question 1) No interest in work	Categorical	Work loss and quality of life
42	mtb_07b1	(PHQ-9: Question 2) Feeling depressed or hopeless	Categorical	Work loss and quality of life
43	mtb_07c1	(PHQ-9: Question 3) Difficulties falling asleep or sleeping too much	Categorical	Work loss and quality of life
44	mtb_07d1	(PHQ-9: Question 4) Tiredness or low energy	Categorical	Work loss and quality of life
45	mtb_07e1	(PHQ-9: Question 5) Loss of appetite or overeating	Categorical	Work loss and quality of life
46	mtb_07f1	(PHQ-9: Question 6) Feeling that I'm a failure or that my family is unhappy because of me	Categorical	Work loss and quality of life
47	mtb_07g1	(PHQ-9: Question 7) Difficulty concentrating while reading the newspaper or watching television	Categorical	Work loss and quality of life
48	mtb_07h1	(PHQ-9: Question 8) Slowing down enough to be noticed by others or wandering around because of irritability or anxiety	Categorical	Work loss and quality of life
49	mtb_07i1	(PHQ-9: Question 9) Thoughts of hurting myself	Categorical	Work loss and quality of life

We calculated the true positive (TP), false positive (FP), true negative, and false negative (FN) scores using a confusion matrix from the experimental results to evaluate the performances of the classifiers. The ratio of incorrectly classified instances was calculated using FN and FP. In contrast, correctly classified samples were represented by TP

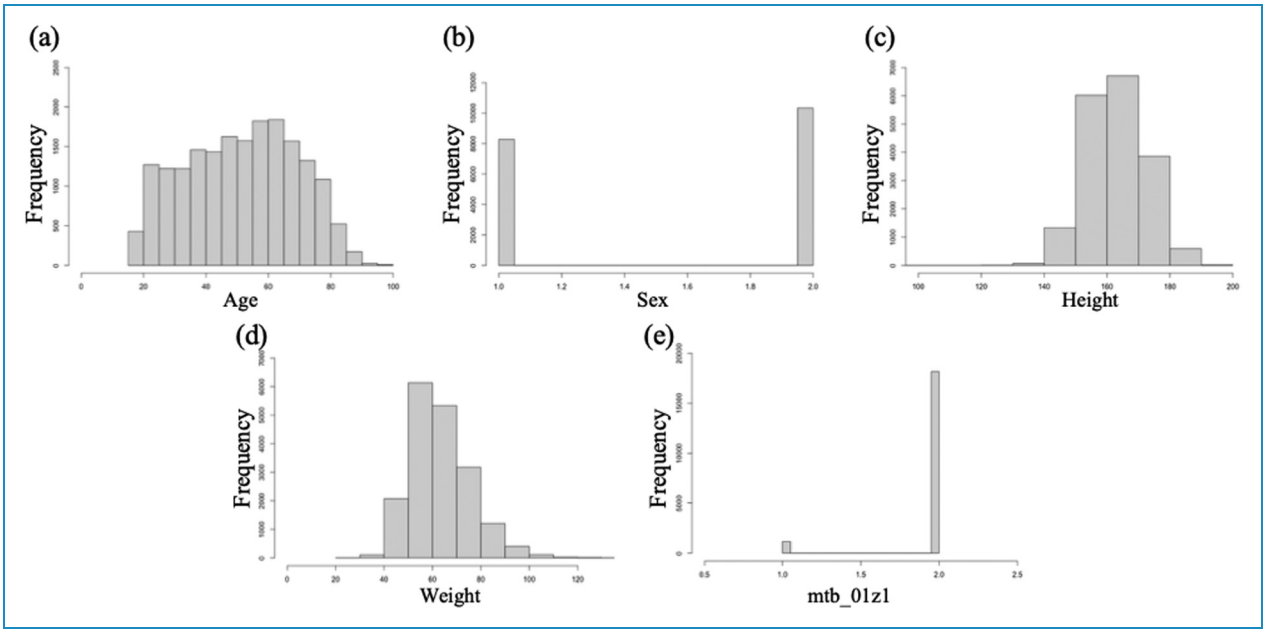

Figure 2. Examples of variable distributions in the dataset (Seoul, 2019):
(a) distribution of the "age" variable; (b) distribution of the "sex" variable ("1.0": male/"2.0": female); (c) distribution of the "height" variable; (d) distribution of the "weight" variable; (e) distribution of the "mtb_01z1" (experience of depression) variable ("1.0": yes/"2.0": no).

Table 3. Hyperparameters of the three ML classification algorithms.

	Hyperparameter	Argument
XGBoost classifier	Eta	0.3
	Gamma	0
	max_depth	6
	min_child_weight	1
SVC	Kernel	Linear
	Gamma	Auto
LR	Penalty	L2
	Solver	Newton-CG

ML: machine learning; SVC: support vector machine classifier; LR: logistic regression.

and FP. Consequently, we obtained four indices: accuracy, recall, precision, and f1-score. Moreover, to organize the receiver operating characteristic (ROC) curve, the TP rate and FP rate values were determined. The area under the ROC curve was calculated using the ROC curve. The formula for calculating these indices was as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{True positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{False positive rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Tools. Python (version 3.7.1; scikit-learn, version 2.4.1) and R (version 4.0.3) programming languages were used to write all codes for ML classifiers, data preprocessing, and visualization.

Results

Baseline characteristics

After preprocessing and removing outliers or missing values in the datasets, the characteristics of the 2019 and 2020 datasets in the CHS dataset were examined to investigate their demographic information. The preprocessed dataset included survey results collected from 400,517

participants. The mean age of the participants at baseline was 55.25 years (SD = 17.86). In addition, the mean sleep time in the last week was 6.62 h (SD = 1.39), and the mean walking Dtime in the last week was 20.33 h (SD = 36.44). The detailed characteristics are listed in Table 4.

Classification performance of ML classifiers. To identify the optimal ML algorithms for our research topic (i.e. comparing the importance of variables with regard to mental health variables), we compared the classification performances of three ML algorithms (XGBoost classifier, SVC, and LR). Among these, the XGBoost classifier showed the best performance (average of over 92% in five indices) under all experimental conditions. The detailed averaged experimental results are presented in Table 5.

Feature importance of ML algorithms with regard to mental health variables. Among the three ML algorithms for classification, the feature importance results of the XGBoost classifier were validated based on the classification performance results. In addition, the important variables in the three groups between 2019 and 2020 were compared to examine the influences of the COVID-19 pandemic on mental health-related dependent variables. Moreover, population densities and the number of COVID-19 confirmed cases in each of the 16 regions were considered to reflect external conditions for comparison. This information is presented in Table 6 and Figure 3.[22]

Some similarities and differences existed in the overall results. Regarding similarities, psychiatric (e.g. stress level) and sociodemographic variables (e.g. income) were found together simultaneously. Second, in Group 1, the variables with the most importance, that is, "fma_01z1" (number of household members), "fma_02z1" (number of household members over the age of 19 years), "ora_01z1" (subjective oral health level), "orb_01z1" (chewing discomfort), and "sra_01z3" (failure to receive essential medical services), were commonly checked. Third, demographic variables such as "age" or "weight" were found mainly in Group 2.

In contrast, there were some differences in the ranking of variables according to the dependent variable. First, in the case of "mtb_01z1" (experience of depression in the last year), the "mtc_01z1" (average sleep time) variable ranked the highest in 2020 (i.e., early pandemic) for the overall region. Second, in the results for "mta_01z1" (subjective stress levels), depression-related variables ("mtb_01z1") were included in Group 1. This trend was especially evident in the data collected from four regions (Seoul, Gyeonggi, Incheon, and Daegu) with many confirmed cases. Furthermore, "ira_01z1" and "ira_02z1" (accident or addiction experience) variables were included regardless of year. Third, in the case of "PHQ-9," oral health-related variables disappeared in Group 1 during the early pandemic period (i.e. 2020). Furthermore, accident or addiction experience-related variables ("ira_01z1" and "ira_02z1") were newly added in 2020. Considering only the regions based on the number of confirmed cases, "apt_t" (house type) and "town_t" (type of residence) variables were included in four regions (i.e. Seoul, Gyeonggi, Incheon, and Daegu) with many confirmed cases and high population densities. However, we could not identify any significant differences in the results for "mtc_01z1" (average sleep time). The detailed importance of the variables for each dependent variable in the four regions is listed in Table 7 and Appendix 3. Other results for the remaining regions are included in Appendix 4.

Based on the above results, we additionally validated the changes in variable values between 2019 and 2020 using statistical tests. Among the several variables to be verified, only continuous variables ("mtc_01z1") were used for validation with paired *t*-tests. The null hypothesis was set as "No difference exists in average sleep time before and after the pandemic." The results of paired *t*-tests confirmed that the difference in sleep time before and after the pandemic was statistically significant (p-value < 0.05). The detailed results of the paired *t*-tests are presented in Table 8.

Table 4. Baseline characteristics of the CHS dataset (2019 and 2020).

Characteristic		CHS
Age (years), mean (SD)		55.25 (17.86)
No. of participants (*n*)		400,517
Gender, *n* (%)	Male	180,233 (44.42%)
	Female	220,284 (55.58%)
Height (cm), mean (SD)		162.76 (7.29)
Weight (kg), mean (SD)		62.67 (8.92)
BMI, mean (SD)		22.36 (0.46)
Sleep time in last week (hour), mean (SD)		6.62 (1.39)
Walking time in last week (hour), mean (SD)		20.33 (36.44)

CHS: Community Health Survey; BMI: body mass index.

Discussion

Principal results. In this study, we attempted to identify changes in the factors associated with mental health, including depression and stress levels, during the early pandemic period in South Korea. To compare the variables between the pre- and early pandemic periods, we utilized the 2019 and 2020 datasets from the CHS dataset based on the

time the World Health Organization declared COVID-19 to be a pandemic (12 March 2020). To consider diverse variable categories for analysis, sociodemographic and psychiatric variables were included in the datasets during our preprocessing steps. In addition, the population and number of confirmed cases of COVID-19 in the 16 regions of South Korea were considered simultaneously to reflect external conditions.

Among 49 selected common variables across the 2019 and 2020 datasets, we selected 12 mental health-related

Table 5. Averaged classification performance results for ML classifiers with four dependent variables ("mtb_01z1," "mta_01z1," "mtc_01z1," and "PHQ-9").

Dependent variable	mtb_01z1 (experience of depression in the last year)											
Group (year)	Group 1 (2019)			Group 1 (2020)			Group 2 (2019)			Group 2 (2020)		
Model	XGBoost	SVC	LR	XGBoost	SVC	LR	XGBoost	SVC	LR	XGBoost	SVC	LR
Precision	0.9169	0.7499	0.5384	0.9229	0.7446	0.5375	0.9026	0.7294	0.5230	0.9210	0.7482	0.5211
Recall	0.9100	0.6445	0.6462	0.9176	0.6559	0.6484	0.9030	0.6038	0.5897	0.9205	0.6543	0.5914
F1-score	0.9015	0.6462	0.4833	0.9142	0.7080	0.4888	0.8930	0.6853	0.4392	0.9144	0.7139	0.4371
Accuracy	0.9347	0.7629	0.6544	0.9397	0.7698	0.6699	0.9343	0.7596	0.5852	0.9390	0.7740	0.5935
AUC	0.9473	0.7853	0.6300	0.9453	0.7747	0.6580	0.9420	0.7813	0.6180	0.9473	0.7773	0.6360
Dependent variable	**mtb_01z1 (experience of depression in the last year)**						**mta_01z1 (the subjective stress level)**					
Group (year)	**Group 3 (2019)**			**Group 3 (2020)**			**Group 1 (2019)**			**Group 1 (2020)**		
Model	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**
Precision	0.9148	0.7328	0.5244	0.8951	0.7477	0.5238	0.9537	0.8521	0.6651	0.9536	0.8606	0.6643
Recall	0.9139	0.6329	0.5930	0.8927	0.6633	0.5998	0.9363	0.8043	0.7284	0.9294	0.7890	0.7376
F1-score	0.9096	0.7142	0.4410	0.9174	0.7174	0.4507	0.9244	0.8262	0.5855	0.9191	0.8540	0.5916
Accuracy	0.9376	0.7647	0.5867	0.9421	0.7737	0.6214	0.9387	0.8758	0.7349	0.9426	0.8881	0.7615
AUC	0.9467	0.7720	0.6180	0.9487	0.7747	0.6400	0.9527	0.8927	0.7387	0.9507	0.8933	0.7600
Dependent variable	**mta_01z1 (the subjective stress level)**											
Group (year)	**Group 2 (2019)**			**Group 2 (2020)**			**Group 3 (2019)**			**Group 3 (2020)**		
Model	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**
Precision	0.9648	0.8570	0.6547	0.9480	0.8525	0.6577	0.9584	0.8581	0.6562	0.9343	0.8740	0.6585
Recall	0.9417	0.8099	0.7005	0.9294	0.7782	0.7081	0.9395	0.8065	0.7036	0.9121	0.8018	0.7174
F1-score	0.9339	0.8727	0.5673	0.9111	0.8339	0.5729	0.9217	0.8456	0.5749	0.9245	0.8593	0.5791
Accuracy	0.9414	0.8939	0.7279	0.9410	0.8803	0.7269	0.9367	0.8955	0.7385	0.9429	0.8811	0.7494
AUC	0.9540	0.8853	0.7587	0.9473	0.8787	0.7587	0.9540	0.9080	0.7733	0.9533	0.8727	0.7393

(continued)

Table 5. Continued.

Dependent variable	mtc_01z1 (average sleep time)											
Group (year)	Group 1 (2019)			Group 1 (2020)			Group 2 (2019)			Group 2 (2020)		
Model	XGBoost	SVC	LR	XGBoost	SVC	LR	XGBoost	SVC	LR	XGBoost	SVC	LR
Precision	0.9561	0.8530	0.6700	0.9556	0.8615	0.6628	0.9658	0.8575	0.6552	0.9507	0.8539	0.6593
Recall	0.9381	0.8122	0.7357	0.9314	0.7940	0.7323	0.9429	0.8145	0.6994	0.9314	0.7848	0.7151
F1-score	0.9256	0.8295	0.5947	0.9211	0.8550	0.5868	0.9342	0.8723	0.5688	0.9137	0.8364	0.5774
Accuracy	0.9359	0.8774	0.7427	0.9431	0.8857	0.7546	0.9377	0.9002	0.7260	0.9417	0.8784	0.7347
AUC	0.9520	0.8938	0.7456	0.9508	0.8900	0.7513	0.9539	0.8888	0.7600	0.9487	0.8838	0.7581
Dependent variable	**mtc_01z1 (average sleep time)**						**PHQ-9**					
Group (year)	**Group 3 (2019)**			**Group 3 (2020)**			**Group 1 (2019)**			**Group 1 (2020)**		
Model	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**
Precision	0.9599	0.8586	0.6569	0.9372	0.8735	0.6583	0.9561	0.8530	0.6700	0.9556	0.8615	0.6628
Recall	0.9409	0.8114	0.7030	0.9152	0.8053	0.7163	0.9381	0.8122	0.7357	0.9314	0.7940	0.7323
F1-score	0.9236	0.8470	0.5764	0.9261	0.8598	0.5780	0.9256	0.8295	0.5947	0.9211	0.8550	0.5868
Accuracy	0.9347	0.9017	0.7366	0.9424	0.8819	0.7462	0.9359	0.8774	0.7427	0.9431	0.8857	0.7546
AUC	0.9547	0.9100	0.7781	0.9545	0.8769	0.7400	0.9520	0.8938	0.7581	0.9508	0.8900	0.7575
Dependent variable	**PHQ-9**											
Group (year)	**Group 2 (2019)**			**Group 2 (2020)**			**Group 3 (2019)**			**Group 3 (2020)**		
Model	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**	**XGBoost**	**SVC**	**LR**
Precision	0.9658	0.8575	0.6552	0.9507	0.8539	0.6593	0.9599	0.8586	0.6569	0.9372	0.8735	0.6583
Recall	0.9429	0.8145	0.6994	0.9314	0.7848	0.7151	0.9409	0.8114	0.7030	0.9152	0.8053	0.7163
F1-score	0.9342	0.8723	0.5688	0.9137	0.8364	0.5774	0.9236	0.8470	0.5764	0.9261	0.8598	0.5780
Accuracy	0.9377	0.9002	0.7260	0.9417	0.8784	0.7347	0.9347	0.9017	0.7366	0.9424	0.8819	0.7462
AUC	0.9539	0.8888	0.7725	0.9487	0.8838	0.7644	0.9545	0.9100	0.7844	0.9454	0.8769	0.7463

ML: machine learning; PHQ-9: patient health questionnaire-9; SVC: support vector machine classifier; LR: logistic regression; AUC: area under the receiver operating characteristic curve.

variables ("mtb_01z1," "mta_01z1," "mtc_01z1," and nine "PHQ-9" variables) as dependent variables. Finally, we set four variables, including a single variable for "PHQ-9," by summing the values of nine variables ("mtb_01z1," "mta_01z1," "mtc_01z1," and "PHQ-9"). Analyses were then conducted according to the four dependent variables. To compare the detailed importance of each variable, we split the variables into three groups (Groups 1, 2, and 3) based on the magnitude of coefficients from the lasso ridge regression models (i.e. variables contained in Group 1 indicated that variables showed the highest rank of coefficient of two regression models).

ML classification algorithms were applied to identify latent patterns between independent variables in various

categories and dependent variables. The class labels of the dataset in all experimental conditions for applying the algorithms were matched to binary class conditions to compare the experimental results in rigorous settings. Feature importance results of the best classification performances (i.e. XGBoost classifier) were used to identify differences in variables during the early pandemic period. As a result, we validated alterations in variable importance between 2019 (pre-pandemic) and 2020 (early pandemic) in three groups and 16 regions in South Korea.

Table 6. Population of the 16 regions in South Korea (2020).

Region	Population	Region	Population
Seoul	9,618,000	Chungnam	2,177,000
Pusan	3,356,000	Jeonbuk	1,806,000
Daegu	2,414,000	Jeonnam	1,793,000
Daejeon	1,492,000	Gyeongbuk	2,652,000
Incheon	2,951,000	Gyeongnam	3,340,000
Ulsan	1,139,000	Gyeonggi	13,452,000
Gwangju	1,480,000	Kangwon	1,519,000
Chungbuk	1,631,000	Jeju	669,000

To interpret our experimental results, we divided the results into four dependent variables and verified the results with those of related previous studies. Furthermore, before analyzing the results by dependent variables, three common points in the variables were found in the overall results. First, both socioeconomic (e.g. the number of household members and town type) and psychiatric variables (e.g. depression and stress level) were checked in Group 1. In this regard, De Figueiredo et al.[11] found that social life and stress-associated factors were highly correlated with mental health. Moreover, relationships between mental health-related variables and suicide attempts were also observed in their research.

Second, health variables not related to mental health (e.g. subjective oral health level, chewing discomfort, and

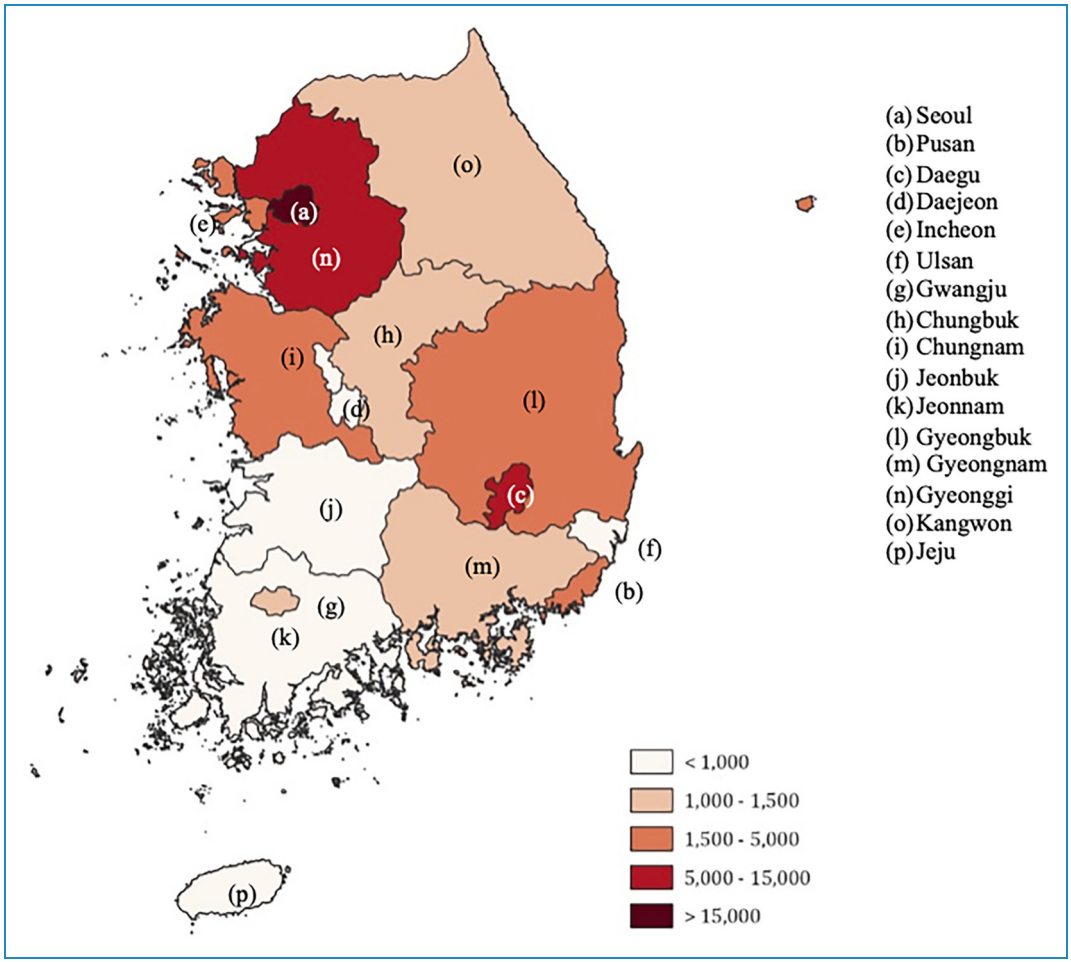

Figure 3. Coronavirus disease 2019 (COVID-19) confirmed cases in South Korea in 2020 by region.

failure to receive essential medical services) were commonly assessed in Group 1. Ciardo et al.[23] investigated the associations between oral health-related quality of life and depression and anxiety levels. Seo et al.[24] found that daily visits to mental health services reduced during the COVID-19 pandemic. They also suggested that patients with anxiety or depressive disorders may have concerns regarding the spread of COVID-19 and may be more reluctant to visit psychiatric outpatient clinics. Third, in Groups 2 and 3 (i.e. variables with relatively lower importance than variables in Group 1), variables related to demographic characteristics such as age, height, and weight were found. Xiong et al.[25] investigated the importance of socioeconomic variables on mental health through systematic

Table 7. Important variables and importance scores for XGBoost classifiers with "mtb_01z1" (experience of depression in the last year) in four regions (Seoul, Gyeonggi, Incheon, and Daegu) with many confirmed cases of COVID-19.

Dependent variable	mtb_01z1 (experience of depression in the last year)									
Seoul (2019)	Group 1					Group 2				
Variable (importance score)	fma_01z1 (1.000)	orb_01z1 (0.797)	mtb_01z1 (0.331)	sra_01z3 (0.087)	ira_01z1 (0.026)	oba_01z1 (1.000)	fma_20z1 (0.369)	phb_01z1 (0.121)	drb_01z2 (0.067)	sob_01z1 (0.030)
Seoul (2020)			Group1					Group2		
Variable (importance score)	mtc_01z1 (1.000)	mbhld_co (0.652)	orb_01z1 (0.469)	oob_01z1 (0.255)	mta_01z1 (0.132)	Age (1.000)	phb_03z1 (0.319)	phb_01z1 (0.230)	sob_02z1 (0.048)	sca_01z1 (0.010)
Seoul (2019)			Group3							
Variable (importance score)	oba_02z1 (1.000)	Age (0.516)	phb_03z1 (0.105)	phb_02z1 (0.035)	sfb_03z2 (0.000)					
Seoul (2020)			Group3							
Variable (Importance score)	drb_01z3 (1.000)	sfb_03z2 (0.000)								
Gyeonggi (2019)			Group1					Group2		
Variable (importance score)	obb_01z1 (1.000)	orb_01z1 (0.664)	fma_02z1 (0.494)	mtc_01z1 (0.369)	sma_03z2 (0.243)	oba_03z1 (1.000)	phb_03z1 (0.074)	drb_01z2 (0.051)	fma_19z1 (0.031)	phb_01z1 (0.014)
Gyeonggi (2020)			Group1					Group2		
Variable (importance score)	mtc_01z1 (1.000)	orb_01z1 (0.591)	fma_02z1 (0.456)	mta_02z1 (0.204)	sma_01z1 (0.112)	drb_01z3 (1.000)	sob_01z1 (0.641)	phb_03z1 (0.387)	obb_01z1 (0.155)	sfb_05z2 (0.000)

(continued)

Table 7. Continued.

Dependent variable	mtb_01z1 (experience of depression in the last year)									
Seoul (2019)	Group 1					Group 2				
Gyeonggi (2019)			Group3							
Variable										
(importance score)	oba_02z1 (1.000)	Age (0.331)	phb_02z1 (0.000)							
Gyeonggi (2020)			Group3							
Variable										
(Importance score)	phb_01z1 (1.000)	mtc_01z1 (0.165)	mbhld_co (0.000)							
Incheon (2019)			Group1					Group2		
Variable										
(importance score)	orb_01z1 (1.000)	mta_01z1 (0.700)	fma_02z1 (0.535)	ora_01z1 (0.407)	sra_01z3 (0.077)	drb_01z2 (1.000)	sob_01z1 (0.532)	sma_03z2 (0.171)	sob_02z1 (0.009)	hya_04z1 (0.000)
Incheon (2020)			Group1					Group2		
Variable										
(importance score)	orb_01z1 (1.000)	mta_01z1 (0.545)	sra_01z3 (0.294)	sma_01z1 (0.191)	fma_04z1 (0.078)	Age (1.000)	mtc_01z1 (0.241)	phb_03z1 (0.173)	sob_01z1 (0.041)	fma_19z3 (0.008)
Incheon (2019)			Group3							
Variable										
(importance score)	oba_02z1 (1.000)	phb_01z1 (0.069)	phb_03z1 (0.020)	obb_01z1 (0.006)	phb_02z1 (0.000)					
Incheon (2020)			Group3							
Variable										
(Importance score)	phb_02z1 (1.000)	soa_06z2 (0.328)	Sob_02z1 (0.000)							
Daegu (2019)			Group1					Group2		
Variable										
(importance score)	fma_02z1 (1.000)	orb_01z1 (0.698)	obb_01z1 (0.499)	mtc_01z1 (0.342)	mta_01z1 (0.185)	oba_03z1 (1.000)	Age (0.469)	phb_03z1 (0.118)	sob_01z1 (0.040)	sfb_03z2 (0.000)

(continued)

Table 7. Continued.

Dependent variable	mtb_01z1 (experience of depression in the last year)									
Seoul (2019)	Group 1					Group 2				
Daegu (2020)	Group1					Group2				
Variable										
(importance score)	mtc_01z1 (1.000)	orb_01z1 (0.699)	mbhld_co (0.568)	obb_01z1 (0.401)	mta_01z1 (0.192)	fma_20z1 (1.000)	mtc_01z1 (0.212)	phb_01z1 (0.107)	phb_03z1 (0.036)	fma_19z3 (0.000)
Daegu (2019)	Group3									
Variable										
(importance score)	oba_02z1 (1.000)	fma_19z1 (0.137)	sca_01z1 (0.007)	sfb_05z2 (0.000)						
Daegu (2020)	Group3									
Variable										
(importance score)	phb_02z1 (1.000)	soa_06z2 (0.740)	drb_03z1 (0.492)	drb_01z3 (0.311)	fma_02z1 (0.140)					

reviews of previous studies compared to demographic factors, including age and body mass index.

In the case of our results with four dependent variables, first, regarding the "mtb_01z1" (experience of depression) variable, unlike during the pre-pandemic period (2019 dataset), average sleep time variables ("mtc_01z1") were newly included in Group 1, and they showed the highest ranking during the early pandemic period (2020 dataset). We found the same trends for all 16 regions. Franceschini et al.[26] showed that sleep-related factors were associated with mental health during the COVID-19 lockdown in the general population living in Italy. They found that changing the sleep–wake rhythm (i.e. habitual bedtime and awakening) could affect psychological distress during a pandemic. Kocevska et al.[27] found that changes in sleep quality throughout the pandemic were associated with negative affect and worry in the Netherlands. Moreover, they found that pre-pandemic good sleepers frequently experienced sleep complaints during the pandemic.

Second, in our results for the dependent variables concerning subjective stress level ("mta_01z1"), the experience of depression variables ("mtb_01z1") was added to Group 1 of the early pandemic period (2020). Khademian et al.[28] identified a relationship between mental health factors (e.g. anxiety and depression) and stress-related factors (e.g. living with high-risk family members and social capital) during the pandemic in Iran. In addition, they verified that age did not have significant associations with depression, anxiety, and stress in their research. Furthermore, Othman[29] investigated the relationships between high levels of mental health disorders during the pandemic and increasing depression in Iraq. Among the 16 regions in South Korea in particular, four regions (Seoul, Gyeonggi, Incheon, and Daegu) with higher confirmed cases and populations showed clear trends compared with other regions. Henning-Smith et al.[30] compared rural and urban locations in the United States to identify differences in the prevalence of mental health and social well-being outcomes. The authors found that the number of COVID-19 concerns of urban residents is higher than that of rural residents. These results suggest that these trends could be affected by the spreading situation in urban areas that occurred during the early pandemic.

Third, with respect to the results in which the "PHQ-9"(depression level) variable was the dependent variable, variables related to experiences of accident or addiction ("ira_01z1" and "ira_02z1") were included in Group 1 during the early pandemic period. Davis et al.[31] investigated pre-COVID-19 posttraumatic stress disorder with greater alcohol use and binge drinking and found that these were associated with a higher risk of mental health, including loneliness, in American veteran groups. They also examined the associations between economic hardship and negative reactions to COVID-19 and alcohol use and drinking. Moreover, the results for the four regions with higher populations and confirmed cases showed that house type and type of residence variables ("apt_t" and "town_t") were included in Group 1 with

Table 8. The results of paired *t*-tests regarding sleep time differences before and after the pandemic.

Before the pandemic	After the pandemic	t-statistics	*P-value*
Seoul in 2019	Seoul in 2020	−7.181	2.2×10^{-12}
Pusan in 2019	Pusan in 2020	−3.911	0.00011
Daegu in 2019	Daegu in 2020	−1.894	0.00600
Daejeon in 2019	Daejeon in 2020	−4.134	0.00009
Ulan in 2019	Ulsan in 2020	−2.969	0.00382
Gwangju in 2019	Gwangju in 2020	−3.561	0.00052
Incheon in 2019	Incheon in 2020	−3.040	0.00262
Chungbuk in 2019	Chungbuk in 2020	−2.606	0.00970
Chungnam in 2019	Chungnam in 2020	−3.337	0.00096
Gyeongbuk in 2019	Gyeongbuk in 2020	−4.847	0.00001
Gyeongnam in 2019	Gyeongnam in 2020	−5.751	1.9×10^{-12}
Jeonbuk in 2019	Jeonbuk in 2020	−2.148	0.03266
Jeonnam in 2019	Jeonnam in 2020	−3.372	0.00086
Gyeonggi in 2019	Gyeonggi in 2020	−11.251	7.52×10^{-13}
Kangwon in 2019	Kangwon in 2020	−4.034	0.00007
Jeju in 2019	Jeju in 2020	−2.552	0.00458

higher ranks. We considered the possibility of the influence of location type (i.e. urban or rural cities) to interpret these results, similar to previous interpretations.[30]

In summary, based on the aforementioned studies, we verified whether similar trends in mental health were shown in the early pandemic period in South Korea with other countries. Moreover, the aforementioned trends were checked more clearly in the regions with higher populations or the number of confirmed cases.[32–34] Based on these results, we validated our analytical methodologies through comparisons with previous studies. The impact of the COVID-19 pandemic on the mental health of the general population living in South Korea was also confirmed by our experimental results. In addition, we identified other factors that were related to the population and number of confirmed cases in each region.

Strength and limitations. This study has several strengths. First, longitudinal datasets collected from 16 regions in South Korea were used to compare mental health-related factors during the pre- and early pandemic periods. Second, variables in diverse categories without any exclusion of categories were applied to identify associations with mental health during the COVID-19 pandemic. Third, ML algorithms were used to identify the relationships between candidate factors and mental health-related factors. Fourth, an additional analysis with regional characteristics (population and number of confirmed cases) was conducted to verify the influence of external conditions on mental health factors. However, our study had some limitations. First, we did not reflect on all characteristics that could affect the mental health of groups living in South Korea, such as specific diseases. Nevertheless, all common variables, including oral health and diabetes, were used for the analysis. Second, diverse deep-learning algorithms could be used to identify related factors in our research. Although deep learning algorithms have been applied in many previous studies that have explored similar research topics, ML algorithms have an advantage in terms of convenience for feature importance compared with deep learning algorithms. Third, we conducted this research based on the populations collected in South Korea. To generalize our experimental results, we need to compare analysis results from other countries. Finally, other mental health-related factors could be checked in a more detailed analysis including causal analysis. Moreover, the influences of each related factor can change during the pandemic (i.e. from the pandemic stage to the endemic stage). We plan to conduct further analyses with detailed topics and additional multivariable analysis methods (e.g. deep learning algorithms) in future studies.

Conclusions

Our findings show that the importance of mental health factors changed during the early pandemic period in South Korea. The most important factors for each mental health dependent variable were average sleep time, experience of depression, and experience of accidents or addictions. In regions with a higher population and confirmed cases, house type and type of residence were identified, together with the aforementioned factors.

Contributorship: JGC and SHH contributed to the conception and design of the study. JGC and SHH contributed to the analysis and interpretation of the data. JGC contributed to the drafting of the manuscript.

Declaration of conflicts of interest: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded in part by the Yonsei Signature Research Cluster Program of 2023 (2023-22-0013) and in part by the National Research Foundation of Korea (NRF) grant through the Korean Government (Ministry of Science and ICT [MSIT]) under Grant 2019R1A2C1007399.

Ethical approval: Not applicable.

Consent statement: Not applicable. (The CHS dataset used in this study was collected by the KDCA and released for research purposes and were approved for data analysis by the institute.)

Copyright information: Not applicable. (Figure 3 was used in a previously published paper and we received confirmation from the journal that the research paper[22] is an open-access journal and there are no restrictions on the reuse of the figure.)

Guarantor: SHH

ORCID iD: Junggu Choi https://orcid.org/0000-0003-2412-2822

References

1. Kaden U. COVID-19 school closure-related changes to the professional life of a K-12 teacher. *Educ Sci* 2020; 10: 165.
2. Oreskovic NM, Kinane TB, Aryee E, et al. The unexpected risks of COVID-19 on asthma control in children. *J Allergy Clin Immunol Pract* 2020; 8: 2489–2491.
3. Al-Tawfiq JA, Al-Yami SS and Rigamonti D. Changes in healthcare managing COVID and non-COVID-19 patients during the pandemic: striking the balance. *Diagn Microbiol Infect Dis* 2020; 98: 115147.
4. Olszewska-Guizzo A, Mukoyama A, Naganawa S, et al. Hemodynamic response to three types of urban spaces before and after lockdown during the COVID-19 pandemic. *Int J Environ Res Public Health* 2021; 18: 6118.
5. Almeida M, Shrestha AD, Stojanac D, et al. The impact of the COVID-19 pandemic on women's mental health. *Arch Womens Ment Health* 2020; 23: 741–748.
6. Xiong J, Lipsitz O, Nasri F, et al. Impact of COVID-19 pandemic on mental health in the general population: A systematic review. *J Affect Disord* 2020; 277: 55–64.
7. Bojdani E, Rajagopalan A, Chen A, et al. COVID-19 pandemic: impact on psychiatric care in the United States. *Psychiatry Res* 2020; 113069: 289.
8. Wang Y, Shi L, Que J, et al. The impact of quarantine on mental health status among general population in China during the COVID-19 pandemic. *Mol Psychiatry* 2021; 26: 4813–4822.
9. De Figueiredo CS, Sandre PC, Portugal LCL, et al. COVID-19 pandemic impact on children and adolescents' mental health: biological, environmental, and social factors. *Prog Neuro-Psychopharmacol Biol Psychiatry* 2021; 110171: 106.
10. O'Connor RC, Wetherall K, Cleare S, et al. Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 mental health & well-being study. *Br J Psychiatry* 2021; 218: 326–333.
11. De Figueiredo CS, Sandre PC, Portugal LCL, et al. COVID-19 pandemic impact on children and adolescents' mental health: biological, environmental, and social factors. *Prog Neuro-Psychopharmacol Biol Psychiatry* 2021; 106: 110171.
12. Wu M, Xu W, Yao Y, et al. Mental health status of students' parents during COVID-19 pandemic and its influence factors. *Gen Psychiatry* 2020; 33: 4.
13. Magson NR, Freeman JY, Rapee RM, et al. Risk and protective factors for prospective changes in adolescent mental health during the COVID-19 pandemic. *J Youth Adolesc* 2021; 50: 44–57.
14. Blix I, Birkeland MS and Thoresen S. Worry and mental health in the COVID-19 pandemic: vulnerability factors in the general Norwegian population. *BMC Public Health* 2021; 21: 1–10.
15. Kwong AS, Pearson RM, Adams MJ, et al. Mental health before and during the COVID-19 pandemic in two longitudinal UK population cohorts. *Br J Psychiatry* 2021; 218: 334–343.
16. Ravens-Sieberer U, Kaman A, Erhart M, et al. Quality of life and mental health in children and adolescents during the first year of the COVID-19 pandemic: results of a two-wave nationwide population-based study. *Eur Child Adolesc Psychiatry* 2021; 32: 575–588.
17. Wang X, Ren R, Kattan MW, et al. Public health Interventions' effect on hospital use in patients with COVID-19: comparative study. *JMIR Public Health Surveill* 2020; 6: e25174.
18. Khattar A, Jain PR and Quadri SMK. Effects of the disastrous pandemic COVID 19 on learning styles, activities and mental health of young Indian students-a machine learning approach. In: *2020 4th international conference on intelligent computing and control systems (ICICCS)*, Madurai, India, 13–15 May 2020, pp. 1190–1195. Madurai, India: IEEE.
19. Rezapour M and Hansen L. A machine learning analysis of COVID-19 mental health data. arXiv preprint arXiv:2112.00227; 2021.
20. Community Health Survey, 2019, Korea Centers for Disease Control and Prevention, https://chs.kdca.go.kr/chs/rawDta/rawDtaProvdMain.do (2019, accessed 12 March 2022)
21. Community Health Survey, 2020, Korea Centers for Disease Control and Prevention, https://chs.kdca.go.kr/chs/rawDta/rawDtaProvdMain.do (2020, accessed 12 March 2022)
22. Lee M and Finerman R. COVID-19, commuting flows, and air quality. *J Asian Econ* 2021; 77: 101374.
23. Ciardo A, Simon MM, Sonnenschein SK, et al. Impact of the COVID-19 pandemic on oral health and psychosocial factors. *Sci Rep* 2022; 12: 1–12.
24. Seo JH, Kim SJ, Lee M, et al. Impact of the COVID-19 pandemic on mental health service use among psychiatric outpatients in a tertiary hospital. *J Affect Disord* 2021; 290: 279–283.
25. Xiong J, Lipsitz O, Nasri F, et al. Impact of COVID-19 pandemic on mental health in the general population: a systematic review. *J Affect Disord* 2020; 277: 55–64.
26. Franceschini C, Musetti A, Zenesini C, et al. Poor sleep quality and its consequences on mental health during the COVID-19 lockdown in Italy. *Front Psychol* 2020; 11: 574475.

27. Kocevska D, Blanken TF, Van Someren EJ, et al. Sleep quality during the COVID-19 pandemic: not one size fits all. *Sleep Med* 2020; 76: 86–88.
28. Khademian F, Delavari S, Koohjani Z, et al. An investigation of depression, anxiety, and stress and its relating factors during COVID-19 pandemic in Iran. *BMC Public Health* 2021; 21: 1–7.
29. Othman N. Depression, anxiety, and stress in the time of COVID-19 pandemic in Kurdistan region, Iraq. *Kurdistan J Appl Res* 2020; 5: 37–44.
30. Henning-Smith C, Meltzer G, Kobayashi LC, et al. Rural/urban differences in mental health and social well-being among older US adults in the early months of the COVID-19 pandemic. *Aging Ment Health* 2022; 27: 505–511.
31. Davis JP, Prindle J, Castro CC, et al. Changes in alcohol use during the COVID-19 pandemic among American veterans. *Addict Behav* 2021; 122: 107052.
32. Vigo D, Patten S, Pajer K, et al. Mental health of communities during the COVID-19 pandemic. *Can J Psychiatry* 2020; 65: 681–687.
33. Caqueo-Urízar A, Urzúa A, Aragón-Caqueo D, et al. Mental health and the COVID-19 pandemic in Chile. *Psychol Trauma* 2020; 12: 521–523.
34. Shah K, Kamrai D, Mekala H, et al. Focus on mental health during the coronavirus (COVID-19) pandemic: applying learnings from the past outbreaks. *Cureus* 2020; 12: 3.

Appendix 1

The dimensions of the datasets of 16 regions in 2019 and 2020.

Appendix 2

The coefficient values of the lasso and ridge regression models.

Appendix 3

Remaining experimental results for the 12 regions.

Appendix 4

The importance of variable results for remaining 8 regions.