DS4C Patient Policy Province Dataset: a Comprehensive COVID-19 Dataset for Causal and Epidemiological Analysis

Jimi Kim[*]
Data Science for COVID-19
kjm0623v@gmail.com

Seojin Jang[*]
Data Science for COVID-19
sarah539884@khu.ac.kr

Woncheol Lee[*]
Data Science for COVID-19
willthor@hanyang.ac.kr

Joong Kun Lee[*]
Mind's Lab, Data Science for COVID-19
joongkul@andrew.cmu.edu

Dong-Hwan Jang[*]
Mind's Lab, Seoul National University
jh01120@snu.ac.kr

Abstract

We present DS4C South Korea Patient, Policy, and Provincial data (DS4C-PPP dataset). The dataset contains comprehensive data that could be used for causal analysis, such as per-patient symptom onset and confirmed date, travel frequency, hospital accessibility, and 61 preventative policies enacted in South Korea. Furthermore, to highlight the research value of the DS4C-PPP dataset, we perform causal analysis with vanilla PC algorithm and gradient boosting algorithm to show: 1) the causality between the diagnosis interval and mortality rate, 2) the effectiveness of COVID-19 intervention policies.

1 Introduction

COVID-19, first identified in Wuhan City, Hubei Province, China, has spread worldwide with over 40 million infected and 1 million deceased[1]. Amongst governments worldwide fighting COVID-19, South Korea has been remarkably effective at containing the disease with less than 100 cases per day [2]. Naturally, researchers and politicians are investigating Korea's success in an attempt to uncover the key methods behind its success.

The DS4C-PPP dataset offers a unique and comprehensive view of COVID-19 in Korea. The dataset offers insights into 3 categories of data: 1) Patient data: per-patient symptom onset and confirmed date along along with travel frequency, 2) Policy data: detailed descriptions and dates of the 61 COVID-19 intervention policies enacted, 3) Provincial data: the number of COVID-19 screening centers, area, and population density of each municipal district.

- **Patient data**: DS4C-PPP dataset contains patient details unforseen in any other COVID-19 dataset. The patient data includes: symptom onset dates, confirmed dates, and travel frequency of every patient. Symptom onset and confirmed dates have played a critical role in understanding epidemiology and infection pattern of different sectors of population [3], effectiveness suppression policies[4], impact of delayed diagnosis on transmission [5], and critical variables behind the success of infection suppression in South Korea [6]. However, until now, there has never been a COVID-19 dataset with symptom onset and confirmed dates for each patient. The per-patient symptom onset and confirmed dates will allow a

[*] Equally Contributed

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

more precise and targeted analysis of infection spread pattern and behaviours. Furthermore, we publish travel frequency – the number of locations each infected person has visited after infection but before diagnosis and isolation. It's a crucial factor in behavioural analysis that has also never been available to academia due to difficulty in tracking the location of each patient.

- **Policy data**: We provide brief descriptions of 61 COVID-19 suppression policies enacted by the South Korean government. This is the most detailed and fine-grained record of South Korean policies [7].
- **Provincial data**: DS4C-PPP provides provincial data necessary to estimate the per-province accessibility to screening centers. Accessibility to COVID-19 screening centers is a critical variable that is closely tied to the infection and mortality rates and effectively encapsulates the more relevant sociodemographic features[8, 9, 10, 11, 12, 13]. As many countries are struggling to suppress the spread of the infection, there has been an ever-increasing effort to identify specific socioeconomic and behavioral factors that are closely associated with increased risk of infection [10]. And the accessibility to screening centers has been identified as one of the most crucial factors with both direct and indirect associations to the spread [9].

Different from the DS4C-PPP, the complete DS4C South Korea dataset is a superset of the DS4C-PPP dataset that contains an even wider breadth of data along with extremely fine-grained patient details, sycg as patient-level details such as age, sex, infection route, symptom onset date, confirmed date, recovery date, travel routes, and more. This level of detail is unseen in any other COVID-19 dataset, and the research value of the dataset is attested by the publications from third-party research organizations that extracted key insights by analyzing our dataset [6, 14, 15, 16, 17, 18, 19, 20, 21, 22]. However, due to the extremely detailed data, the main dataset is currently undergoing anonymization process. As a preamble to the publication of the main DS4C dataset, we publish DS4C-PPP dataset which only includes information that is anonymized but yet retains the rich research potential.

In order to demonstrate the research value of the dataset, we perform two causal analyses to: 1) analyze the effectiveness of South Korean policies, and 2) identify the causality between diagnostic interval and mortality. First, we identify causal relationships amongst the multitude of variables with PC-algorithm [23], then we estimate the strength of the causal relations with gradient boosting [24].

2 Data setup

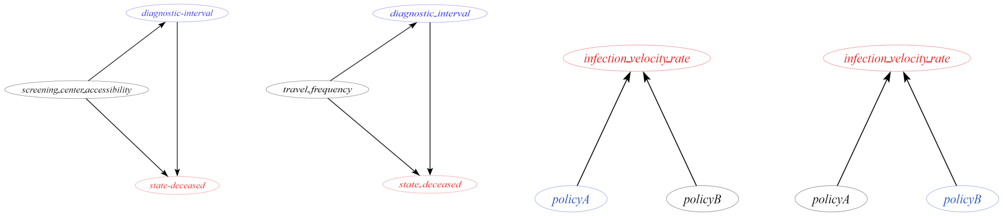
2.1 Data record

Our dataset consists of the following three categories: patient data, policy data, and provincial data. Along with the aforementioned raw data, we also publish the post-processed data used in our experiment. The process of data processing and feature selection will be described in 3.

In the first category, we introduce per-patient: *symptom_onset_date*, *confirmed_date*, *travel_frequency*, and *state_deceased*. Symptom onset date is the date that each patient first displayed symptoms of COVID-19, and confirmed date is the date that the patient was diagnosed. Note that patient-wise symptom onset and confirmed dates can be aggregated to determine the number of newly diagnosed and un-diagnosed patients per day. *travel_frequency* indicates the number of locations that an infected but un-diagnosed patient visited before isolation.

Second, policy data contains descriptions and start and end dates of 61 preventitive policies enacted by the Korean government. 61 policies include a wide type of policies such as social, education, and immigration. In order to help understand the effectiveness of the policies, the cumulative number of confirmed patients is appended as part of the dataset.

Lastly,in provincial data category, we provide: *the number of screening centers*, *land size*, and *population* of each province. The number of screening centers along with the peripheral contextual data accurately captures the accessibility to screening centers in each province. And the derived can be used to directly identify its impact on the spread or as a potent indicator of relevant socioeconomic factors of the province [9, 10].



(a) Causal model of *diagnosis_interval* on *state_deceased*. Confounder: *screening_center_accessibility*

(b) Causal model of *diagnosis_interval* on *state_deceased*. Confounder: *travel_frequency*

(c) Causal model of *policy A* on *infection_velocity_rate*

(d) Causal model of *policy B* on *infection_velocity_rate*

Figure 1

2.2 Method

The patient data in the DS4C-PPP dataset is generated from published raw data by various institutions of the Korean government. The Korean government generated very precise and accurate data by analyzing each patient's credit card history, public surveillance camera footage, and cellphone GPS locations. However, such valuable raw data was completely obsolete to data analysts due to three challenges:

- **Decentralized publication of data**: Over a hundred local governments independently collected and published results on their district websites. There was also no unified publication format throughout the websites and the publication format changed often as COVID-19 progressed and new types of data was collected.
- **Data embedded in natural language**: most patient data was embedded in natural language, and the natural language raw data had no heuristic or rule.
- **Short publication period**: Local governments permanently removed the patient data within a few days of publication.

In order to overcome the above challenges and effectively collect, engineer, and distribute data, 17 DS4C members manually scarped raw data and parsed natural language. Patient data's sources are in the supplementary material due to the large volume. Policy data is collected from [2, 25, 26] and provincial data is collected from [27].

3 Casual discovery

3.1 Feature Selection

3.1.1 Causal effect of diagnosis interval on state deceased

Main features which were used to identify the causal effect of diagnostic interval on state deceased were *diagnostic_ intervals* and *state_deceased*. The the different characteristics of patients' area of residence have a profound effect on both infection and mortality rates [8, 9, 10, 11, 12, 13]. We processed the feature *screening_center_accessibility* to reflect the provincial characteristics on our causal discovery. In addition, we also extracted *travel_frequency* which could represent the sensitivity of the COVID-19 infection and the lifestyle of the patients. Detailed explanations and the processed method of extracted features are shown below.

- **screening center accessibility:** Continuous variable which includes the accessibility for the COVID-19 screening centers per patient (higher is better). Used features for calculating *screening_center_accessibility* values were from *Provincial data*. We mapped *screening_center_accessibility* to each patient according to patients' home addresses found in the *province* column of *Patient data*. The metric for calculating the *screening_center_accessibility* values

$$A(p) = N(p)/(sS(p)\ hH(p)) \tag{1}$$

where, given province p, $A(p)$ is the *screening_center_accessibility*, $S(p)$ is the size in km^2, and $H(p)$ is the population of of the province. s and h are normalization constants for $S(p)$ and $H(p)$ respectively. Here we use $s = 1000000$ and $h = 10000$.

- **travel frequency:** Continuous variable that shows the number of locations that an infected but un-diagnosed patient visited before isolation.
- **diagnosis interval:** Continuous variable which shows the interval between the symptom onset date and the confirmed date of each COVID-19 patient. It was extracted from two columns in *Patient data*; *symptom onset date* and confirmed date. We got the time interval of two features.
- **state deceased:** Discrete variable that indicates whether the patient had been deceased or not(1 and 0 respectively).

The causal effect can be biased by *travel_frequency* and *screening_center_accessibility* variables. If we adjust on *travel_frequency* and *screening_center_accessibility*, we control for confounding bias, which gives us an unbiased estimate of the causal effect.

3.1.2 Causal effect of 2 policies on infection velocity rate

The South Korean government had enacted 61 notable policies to curb the spread of COVID-19 from January 2020 to Jun 2020. We have condensed the 61 policies into 2 by: 1) grouping policies with that are very similar in method (ex. different social distancing policies are grouped together), and 2) eliminating policies whose direct purpose is not to reduce the spread (ex. policies enacted to track movement of the population). Policies enacted before February 19th are removed to avoid the sparsity problem. We analyzed the effect of the two policies on the change of the number of newly confirmed cases per day. Detailed explanations of the variables are described below.

- **policy A:** Discrete variable. *policy A* represents a group of policies related to social distancing. From 29 Feb to 19 April and from 6 May, social distancing was in phase 2. From 20 April to 5 May the level of social distancing was in phase 1. We set 1 when the level of social distancing was in phase 1 and set 2 when the level of social distancing was in phase 2. Before any social distancing policy was implemented, we set the value to 0.
- **policy B:** Discrete variable. *policy B* represents a group of policies related to increasing mask supplies and distribution. As the demand for masks surpassed the supply, the South Korean government enacted a series of policies in order increase the supply of masks. It has been implemented since 9 March. We set 1 when it is implemented and 0 otherwise.
- **infection velocity rate:** Continuous variable. We defined the *infection_velocity_rate* as the difference between today's and yesterday's daily confirmed numbers divided by yesterday's confirmed numbers.

$$v(t) = \{x(t) - x(t-1)\}/x(t-1) \tag{2}$$

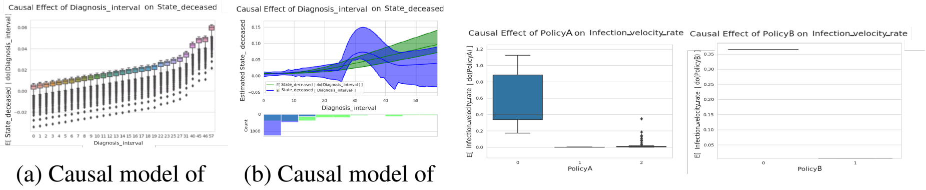
where $v(t)$ is the infection velocity and $x(t)$ is the number of daily confirmed patients on day t.

3.2 Ground Truth

3.2.1 Causal effect of diagnostic interval on state deceased

Recently, the diagnosis time delay turned out to be an important risk factor in the mortality rate for COVID-19 in the city of Rio de Janeiro, Brazil[10]. Accordingly, we set a causal structure to find out whether the diagnosis time delay works as a risk factor for the COVID-19 mortality in South Korea.

We utilized *screening_center_accessibility*, *travel_frequency*, *diagnostic_interval*, and *state_deceased* from DS4C-PPP dataset to construct the causal structure for the experiments. We found out that



(a) Causal model of *diagnosis_interval* on *state_deceased*. Confounder: *screening_center_accessibility*

(b) Causal model of *diagnosis_interval* on *state_deceased*. Confounder: *travel_frequency*

(c) Causal model of *policy A* on *infection_velocity_rate*

(d) Causal model of *policy B* on *infection_velocity_rate*

Figure 2

screening_center_accessibility and *travel_frequency* features are each related to *diagnostic_interval* through the PC algorithm. Therefore, we set two causal structures by setting *diagnostic_interval* and *state_deceased* as a treatment and an outcome, also regarding *screening_center_accessibility* and *travel_frequency* as confounders in each structure. The causal structure of each experiment is shown at (a) and (b) in Figure 1.

3.2.2 Causal effect of 2 policies on infection velocity rate

We utilized social distancing policy and mask distribution policy to construct causal structure. We found out that *policy A* and *policy B* are related to *infection_velocity_rate* by using PC algorithm. We set causal structures as the polices are the inputs and the *infection_velocity_rate* are the outputs. The causal structure is shown at (c) and (d) in Figure 1.

4 Experiments

4.1 Experimental setup

PC algorithm[28] and XGBoost[29] gradient boosting were used for causal analysis of the two experiment setups described above. First, PC algorithm was used to discover the causal structure from the selected features. Then, XGBoost was used to predict whether the experimental model based on the Ground Truth created fits the observational data well.

For the PC algorithm, Fisher's z transform was used to determine conditional independence amongst features where p-value was 0.05. XGBoost divided the data into 3-folds for cross validation. Superlearner[30] was used as the training method. Those algorithms were implemented in the Causal Fusion and Tetrad software package(version 6.5.4).

4.2 Results

In the first experiment, we identify a clear causal relation between *diagnosis_interval* and *state_deceased* as shown in Figure 2 (a) and (b). In both settings with different confounders, there exists a clear causation from *diagnosis_interval* to *state_deceased*. In the first 25 days, we observe higher accuracy when *travel_frequency* is set as the confounder, and in the latter 35 days when *screening_center_accessibility* is set as the confounder. This may hint that the causal relationship between *diagnosis_interval* and *state_deceased* changes as social distancing policies are implemented and medical supplies such as screening centers or test kits become more accessible.

In the second experiment, we observe a causal relationship between the two policies and *infection_velocity_rate* as shown in Figure 2 (c) and (d). In the initial phase of social distancing level 2, *infection_velocity_rate* is higher than it is during social distancing level 1. This may seemingly contradict the effectiveness of social distancing policies. However, considering that such policies are reactive, not proactive, it is reasonable to see a short-term increase in *infection_velocity_rate*. Supporting this claim, we observe a significant decrease in *infection_velocity_rate* after days into the level 2 policy. Furthermore, we observe causality between mask distribution policies and *infection_velocity_rate*, further highlighting the effectiveness of wearing masks on curbing the spread of COVID-19.

Broader Impact

The published dataset was partially funded by the Korea Ministry of Science and Technology and MindsLab. The dataset was produced by the members of Data Science for COVID-19 (DS4C)[31]: Jihoo Kim, JoongKun Lee, SeoJin Jang, SeongHan Ryoo, YeonJun In, WonCheol Lee, DongHwan Jang, Jimi Kim, MuHwan Kim, BoYoung Song, KyeongWook Jang, MinSeok Jung, SangWook Park, TaeHyeong Park, WanSik Choi, YouNa Jung. Part of the raw route data was provided by the Corona Board team[32].

References

[1] Esteban Ortiz-Ospina Max Roser, Hannah Ritchie and Joe Hasell. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[2] Ministry of health and welfare.

[3] Yuanyuan Dong, Xi Mo, Yabin Hu, Xin Qi, Fan Jiang, Zhongyi Jiang, and Shilu Tong. Epidemiology of COVID-19 Among Children in China. *Pediatrics*, 145(6), 2020.

[4] Mirjam E Kretzschmar, Ganna Rozhnova, Martin C J Bootsma, Michiel van Boven, Janneke H H M van de Wijgert, and Marc J M Bonten. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. *The Lancet Public Health*, 5(8):e452 – e459, 2020.

[5] Xin Miao Rong, Liu Yang, Hui di Chu, and Meng Fan. Effect of delay in diagnosis on transmission of COVID-19. *Mathematical biosciences and engineering : MBE*, 17(3):2725–2740, mar 2020.

[6] Sofia Kyonhi Mettler, Jihoo Kim, and Marloes H. Maathuis. Diagnostic serial interval as a novel indicator for contact tracing effectiveness exemplified with the sars-cov-2/covid-19 outbreak in south korea. *medRxiv*, 2020.

[7] Giliberto Capano, Michael Howlett, Darryl S L Jarvis, M Ramesh, and Nihit Goyal. Mobilizing policy (in) capacity to fight COVID-19: Understanding variations in state responses. *Policy and Society*, 39(3):285–308, 2020.

[8] Madikay Senghore, Merveille K Savi, Bénédicte Gnangnon, William P Hanage, and Iruka N Okeke. Leveraging Africa's preparedness towards the next phase of the COVID-19 pandemic. *The Lancet Global Health*, 8(7):e884–e885, jul 2020.

[9] Benjamin Rader, Christina M Astley, Karla Therese L Sy, Kara Sewalk, Yulin Hswen, John S Brownstein, and Moritz U G Kraemer. Geographic access to United States SARS-CoV-2 testing sites highlights healthcare disparities and may bias transmission estimates. *Journal of Travel Medicine*, 27(7), 2020.

[10] Alexandre de FÃ¡tima Cobre, Beatriz BÃ, Mariana Millan Fachi, Raquel de Oliveira Vilhena, Eric Luiz Domingos, Fernanda Stumpf Tonin, and Roberto Pontarolo. Risk factors associated with delay in diagnosis and mortality in patients with COVID-19 in the city of Rio de Janeiro, Brazil. *CiÃ SaÃColetiva*, 25:4131 – 4140, 10 2020.

[11] Casey M Zipfel and Shweta Bansal. Health inequities in influenza transmission and surveillance. *medRxiv*, 2020.

[12] Jason Corburn, David Vlahov, Blessing Mberu, Lee Riley, Waleska Teixeira Caiaffa, Sabina Faiz Rashid, Albert Ko, Sheela Patel, Smurti Jukur, Eliana Martínez-Herrera, Saroj Jayasinghe, Siddharth Agarwal, Blaise Nguendo-Yongsi, Jane Weru, Smith Ouma, Katia Edmundo, Tolu Oni, and Hany Ayad. Slum Health: Arresting COVID-19 and Improving Well-Being in Urban Informal Settlements. *Journal of urban health : bulletin of the New York Academy of Medicine*, 97(3):348–357, jun 2020.

[13] Victor A Alegana, Joseph Maina, Paul O Ouma, Peter M Macharia, Jim Wright, Peter M Atkinson, Emelda A Okiro, Robert W Snow, and Andrew J Tatem. National and sub-national variation in patterns of febrile case management in sub-Saharan Africa. *Nature Communications*, 9(1):4994, 2018.

[14] Yejin Kim and Xiaoqian Jiang. Evolving transmission network dynamics of covid-19 cluster infections in south korea: a descriptive study. *medRxiv*, 2020.

[15] Sukru Yagiz Olmez, Jameson Mori, Erik Miehling, Tamer Basar, Rebecca Lee Smith, Matthew West, and Prashant Mehta. A data-informed approach for analysis, validation, and identification of covid-19 models. *medRxiv*, 2020.

[16] Yoshiro Suzuki and Ayaka Suzuki. Machine learning model estimating number of covid-19 infection cases over coming 24 days in every province of south korea (xgboost and multioutputregressor). *medRxiv*, 2020.

[17] Xinhua Yu, Jiasong Duan, Yu Jiang, and Hongmei Zhang. Distinctive trajectories of covid-19 epidemic by age and gender: a retrospective modeling of the epidemic in south korea. *medRxiv*, 2020.

[18] Byungwon Kim, Seonghong Kim, Woncheol Jang, Sungkyu Jung, and Johan Lim. Estimation of the case fatality rate based on stratification for the covid-19 outbreak. *medRxiv*, 2020.

[19] Nicholas G Davies, Petra Klepac, Yang Liu, Kiesha Prem, Mark Jit, , and Rosalind M Eggo. Age-dependent effects in the transmission and control of covid-19 epidemics. *medRxiv*, 2020.

[20] Xianding Deng, Wei Gu, Scot Federman, Louis Du Plessis, Oliver Pybus, Nuno Faria, Candace Wang, Guixia Yu, Chao-Yang Pan, Hugo Guevara, Alicia Sotomayor-Gonzalez, Kelsey Zorn, Allan Gopez, Venice Servellita, Elaine Hsu, Steve Miller, Trevor Bedford, Alexander Greninger, Pavitra Roychoudhury, Michael Famulare, Helen Y Chu, Jay Shendure, Lea Starita, Catie Anderson, Karthik Gangavarapu, Mark Zeller, Emily Spencer, Kristian Andersen, Duncan MacCannell, Suxiang Tong, Gregory Armstrong, Clinton Paden, Yan Li, Ying Zhang, Scott Morrow, Matthew Willis, Bela Matyas, Sundari Mase, Olivia Kasirye, Maggie Park, Curtis Chan, Alexander Yu, Shua Chai, Elsa Villarino, Brandon Bonin, Debra Wadford, and Charles Y Chiu. A genomic survey of sars-cov-2 reveals multiple introductions into northern california without a predominant lineage. *medRxiv*, 2020.

[21] Atina Husnayain, Eunha Shim, Anis Fuad, and Emily Chia-Yu Su. Assessing the community risk perception toward covid-19 outbreak in south korea: evidence from google and naver relative search volume. *medRxiv*, 2020.

[22] Xinhua Yu. Risk interactions of coronavirus infection across age groups after the peak of covid-19 epidemic. *medRxiv*, 2020.

[23] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(22):613–636, 2007.

[24] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[25] Ministry of science and ict.

[26] Ministry of economy and finance.

[27] Statistics korea.

[28] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[29] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[30] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[31] Data science for covid-19.

[32] Coronaboard.