OPEN ACCESS Freely available online

Journal of Infectious Diseases & Preventive Medicine

ISSN: 2329-8731

Research Article

A Prediction Model for COVID-19 Prevalence Based on Demographic and Healthcare Parameters in Iran

Parimah Emadi Safavi[1], Karim Rahimian[1], Alireza Doustmohammadi[1], Mahla Safari Dastjerdei[2], Ahmadreza Rasouli[1], Javad Zahiri[2,3*]

[1]Department of Biophysics, Tarbiat Modares University, Tehran, Iran;[2]Department of Midwifery, Iran University of Medical Sciences, Tehran, Iran;[3]Department of Neuroscience, University of California, San Diego, USA

ABSTRACT

Background: Coronavirus Disease 2019 (COVID-19) pandemic has become the greatest threat to global health in only a matter of months. Iran struggling with COVID-19 coincidence with Nowruz vacations has led to horrendous consequences for both people and the public health workforce. Modeling approaches have been proved to be highly advantageous in taking appropriate actions in the early stages of the pandemic. To this date, no study has been conducted to model the disease to investigate the prevalence of COVID-19 cases, especially after travel restrictions in Iran. We investigated contributing factors of early-stage coronavirus spread *via* generating a model to predict daily confirmed cases in Iran.

Methods: In this study, we collected publicly available data of confirmed cases in 24 provinces from April 4, 2020, to May 2, 2020, with a list of explanatory factors. we exploited the opportunities that Artificial neural networks offer to investigate contributing factors of early-stage coronavirus spread *via* generating a model to predict daily confirmed cases in Iran. The factors were checked separately for any linear associations and to train and validate a multilayer perceptron network.

Results: The accuracy of the models was evaluated; the R2 scores were 0.842 for population distribution, 0.822 for health index, and 0.864 for the population in the provinces. No significant linear associations were seen between collected factors and COVID-19 incidence in provinces. Also, the relative distance from the disease epicenter had no relation to the disease incidence in provinces.

Conclusion: Our results show the significant impact of population, population density, and health infrastructure quality on the cross-province spread of COVID-19 in the time of travel restrictions when the vacation ended. Accordingly, this information can be implicated in assessing the risk of epidemics and future policy-making in this area.

Keywords: COVID-19; Healthcare; Geographical factors; Artificial neural network

INTRODUCTION

In late January 2020, Coronavirus disease 2019 (COVID-19) from China became to be an international concern [1]. The virus rapidly spread across the world with more than 60 million confirmed cases in 218 countries. Despite all countries efforts to mitigate the disease spread through imposing travel restrictions and lockdown, due to the high transmissibility of the virus, many countries including Iran became afflicted within the first months of the outbreak [2,3].

Iran is one of the first countries to confront the Coronavirus outbreak with a high number of confirmed cases in the world. The government decided to close public places and cancel all social events to mitigate the transmission. By early March all provinces were affected by the pandemic which was just a few weeks after the first cases were reported in Qom [4]. One great adversity faced during the pandemic was that the outbreak coincided with the Nowruz holidays which led to a sharp increase in the number of daily cases [5]. On March 27th in the middle of Nowruz vacations, Iranian authorities enacted travel restrictions to reduce the spreading of COVID-19 carriers to each province (Figure 1). The delay in decision-making has been debated to be one of many reasons the disease spread rapidly across all provinces [6].

Correspondence to: Dr Javad Zahiri, Department of Midwifery, Iran University of Medical Sciences, Tehran, Iran, E-mail: zahiri@modares.ac.ir

Received: October 05, 2021; **Accepted:** October 19, 2021; **Published:** October 26, 2021

Citation: Safavi PE, Rahimian K, Doustmohammadi A, Dastjerdei MS, Rasouli A, Zahiri J (2021) A Prediction Model for COVID-19 Prevalence Based on Demographic and Healthcare Parameters in Iran. J Infect Dis Preve Med. 9: 239.

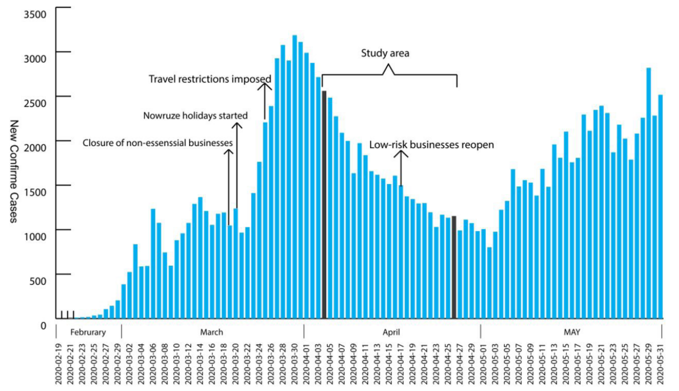

Figure 1: Demographics of COVID-19 confirmed cases during the early days of the pandemic in Iran.

Many factors have been known to be associated with the initial levels of the Coronavirus outbreak at the country level including geographical factors, demographical parameters, healthcare services, and economic status [7-10]. Meanwhile, some research projects have been carried out to investigate the early-stage dynamics of the COVID-19 epidemic in Iran. For instance, convey that there is a negative relationship between COVID-19 incidence and the relative distance of provinces from Qom as the epicenter of the disease [11]. Also, the study reported a relationship between the elderly populations with the outbreak. In another study, reported a direct relationship between population density and intra-provincial movement with the infection rate of each province [12]. This work also points out the weather impacts on the disease infection rate. Acknowledging such factors would be of great help in foreseeing inevitable conditions brought by epidemics in different climate zones.

Artificial neural network as a machine learning tool has been wildly used in studies to forecast the outbreak due to its proven application in epidemiologic risk assessments [13]. In a study, a limited time-series dataset of Coronavirus patients to obtain a quality neural network model of disease spread [14]. In another case in which artificial neural networks have become handy, authors collected a wide range of relevant factors and employed a neural network model to forecast COVID-19 incidence across the United States [15].

Interestingly most of the research studies in Iran were conducted during the period before travel restrictions in which it was permissible for provinces to receive travelers from other provinces. In this paper, we aimed to generate a spread model of COVID-19 and investigate the demographical and healthcare factors impact on disease incidence from April 4th to May 2nd in 24 provinces. To predict daily confirmed cases in the provinces, we collected potential risk variables and employed a multi-layer perceptron network trained by publicly available COVID-19 data from Iran authorities. Both adopting policies and assessing mitigation strategies are crucial in emergency cases in the future. Thus, this work would give a vision of the dynamic nature of this epidemic in the early stages.

MATERIALS AND METHODS

Data collection

The number of daily confirmed cases for the period from April 4, 2020, to May 2, 2020, was obtained from the Iranian Ministry of Health and Medical Education which shows the disease spread in provinces as time-series data. Due to the excess of missing values in reported cases, seven provinces were excluded from the study. Demographic data for each province were extracted from census data from the Statistical Centre of Iran (Table 1). To avoid any underlying association of variables, men's population divided by province populations was used. Since each province has medical infrastructures to cope with COVID-19, Province-level healthcare infrastructure indices were extracted from Shojai et al. to represent healthcare services access and quality in each province [16]. COVID-19 Incidence rate per 100,000 residents during the study period (4 April until 2 May) with all variables is reported for provinces (Table 2). The geographical distribution of the confirmed cases and healthcare quality indices are illustrated in Figure 2.

Table 1: Characteristics of provinces with reported COVID-19 cases used in the study.

Province	Distance from Qom (km)	Population	Gender ratio (%)	Distribution (%)	Urban population (%)	Density (population per km^2)	Literacy rate (%)	Health infrastructure quality index	Over 65 years old (%)	Incidence per 100,000 inhabitants
Ardabīl	623	1270420	0.51	1.59	0.68	71	83.1	0.9271	6.4	231.1
Bushehr	897	1163400	0.53	1.46	0.72	51	89.2	0.9464	4.3	15.37
Chaharmahal	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
and Bakhtiari	384	947763	0.51	1.19	0.64	58	84.7	0.9254	5.8	27.68
East Azerbaijan	657	3909652	0.51	4.89	0.72	86	84.7	0.7342	7.2	151.49
Fars	760	4851274	0.51	6.07	0.7	40	88.8	0.4157	6.1	50.7
Hormozgān	1141	1776415	0.51	2.22	0.55	25	87.8	0.8945	3.9	37.7
Ilam	596	580158	0.51	0.73	0.68	29	84.9	0.9196	5.6	36.68
Isfahan	281	5120850	0.51	6.41	0.88	48	89.9	0.4977	7.1	66.46
Kerman	847	3164718	0.51	3.96	0.59	17	85.4	0.731	5.1	16.04
Kermanshah	425	1952434	0.51	2.44	0.75	78	84.5	0.8007	6.5	65.41
Khūzestān	666	4710509	0.51	5.89	0.75	74	86.3	0.6523	4.5	39.87
Kohgiluyeh and Boyer-Ahmad	611	713052	0.51	0.89	0.56	46	84.4	0.9879	4.6	38.91
Kurdistan	1154	1603011	0.51	2.01	0.71	55	81.5	0.9523	6.1	51.62

Lorestan	337	1760649	0.51	2.2	0.64	62	83	0.8723	5.8	89.07
Markazi	133	1429475	0.51	1.79	0.77	49	87	0.8642	7.6	35.85
Mazandaran	389	3283582	0.5	4.11	0.58	138	88.7	0.6364	7.6	81.64
North Khorasan	827	863092	0.5	1.08	0.56	30	83.3	0.9814	5.5	74.89
Qom	0	1292283	0.51	1.62	0.95	112	88.7	0.9465	4.8	137.22
Semnan	283	702360	0.51	0.88	0.8	7	91.5	0.8556	6.7	39.87
Sistan and Baluchestan	1344	2775014	0.51	3.47	0.48	15	76	0.7228	3.2	25.02
South Khorasan	1027	768898	0.51	0.96	0.59	5	86.7	0.948	6.9	30.66
Tehran	1027	13267370	0.5	16.6	0.94	969	92.9	0	6.8	63.68
West Azerbaijan	796	3265219	0.51	4.09	0.65	87	82	0.8106	5.6	61.35
Zanjan	369	1057461	0.51	1.32	0.67	49	84.8	0.9215	6.7	36.53

Table 2: The descriptive statistics of the variables.

Variable	Var	Min	Max	Mean
Distance from Qom	114919	0	1344	648.917
Population	3.00E+09	11240	183285	56675
Gender ratio (%)	7.00E+12	580158	1.00E+07	2592877
Distribution (%)	3.00E-05	0.50242	0.5336	0.50846
Over 65 years old (%)	8.7169	58.5	73	69.125
Urban population (%)	0.0138	0.48491	0.9518	0.69049
Density (population per area)	34453	5	969	91.7083
Literacy rate (%)	12.582	76	92.9	85.825
Health infrastructure quality	0.0491	0	0.9879	0.78934

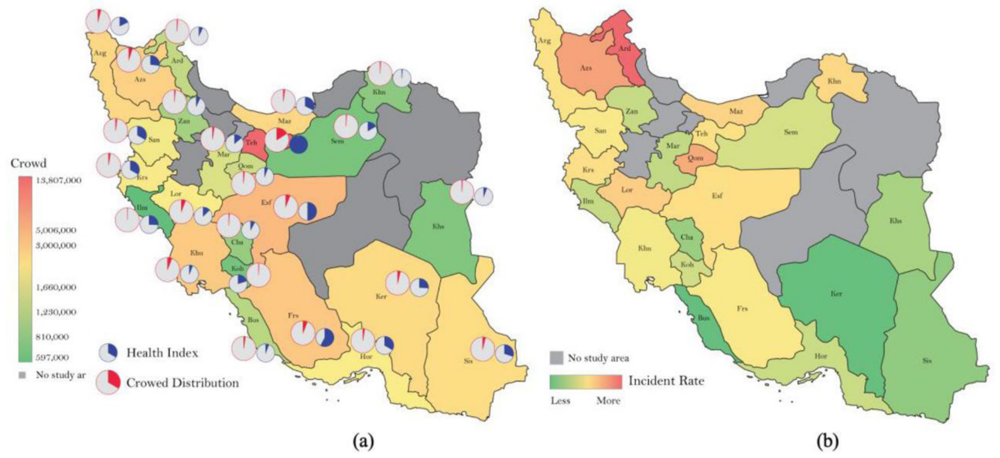

Figure 2: (a) Distribution of health infrastructure indices and population density in Iran, (b) COVID-19 incidence distribution across provinces in the study period.

Patient and public involvement

Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Statistical analysis

All the variables including population-based, health indices, literacy indices, altitude, and distance from the epicenter, were normalized before the analysis. Any linear association between incidence rate per 100,000 residents and the variables was measured *via* Pearson's correlation coefficient (r). All analyses were performed using python version 3.8.

Artificial neural network

Multi-layer perceptron (MLP) is a common network model for supervised learning regression studies. It consists of input, hidden, and output layers. In this study, the capability of each variable as predictors is evaluated through the network model to predict new COVID-19 cases of Iran provinces.

The variables were imported as neurons into the network *via* the input layer. These values then were affected by weights of the connections and imported to the next hidden layers which ended to produce the output variable in the output layer. The flexible ability of hidden layers to recognize patterns in the data makes this machine-learning method superior to the standard regression model.

To restructure the time-series dataset of confirmed cases to a supervised learning problem, we used the sliding window method. An approximated window iteratively segments the time-series data until optimal window size (m) with the least possible error is chosen. A matrix of historical daily case data in all provinces is then retrieved to train the neural network. The output variable, in this case, is province-level forecasted daily COVID-19 cases.

K-fold cross-validation technique carried out for model internal validity assessment. This technique randomly partitioned the dataset into k subsets of equal size (in this case, k=10), which was repeatedly used in training and testing to the point every subset served once to test the model. The entire process was repeated 10 times. The result is then presented as the average of achieved scores, with standard deviation noted.

A base model which only contains daily cases as input variables were defined to assess the performances of neural network each time variables were added as an input to the model (Figure 3). Root mean square error (RMSE) and the absolute fraction of variances (R2) metrics were utilized to evaluate the performance of the model and to investigate the capabilities of variables to predict the output variables. All the procedure was performed in the python environment using the scikit-learn library on a PC with a six-core 2.6 GHz processor and 16 GBs RAM.

RESULTS

Iran has a population of over 83,000,000 with a territory of 1,648,195 km2. All analyses were carried out in the period of 4 April until 2 May for 24 provinces. Most cases were reported in Tehran as a highly-populated and developed province.

The linear relationships between COVID-19 incidence and variables were investigated Table 3. The results revealed that COVID-19 incidence in provinces has no significant linear associations with either demographic variables or healthcare indices. Also, no statistically significant relation was found between disease incidence in provinces with their Distance from Qom as the epicenter of the disease (r=-0.26, p>.05).

A multilayer perceptron was used in the study. The optimal number of hidden layer neurons was obtained by a trial-and-error procedure and selection was made based on the lowest RMSE and highest R2 metrics. In the MLP model, 18 combinations of inputs were examined to evaluate the performance of variables to act as predictors, each time they were added as an input to the base model. Most accurate models were obtained when variables population distribution (R2=84%), health index (R2=82%), and population (R2=86%) were added as an input to the base model (Figure 4).

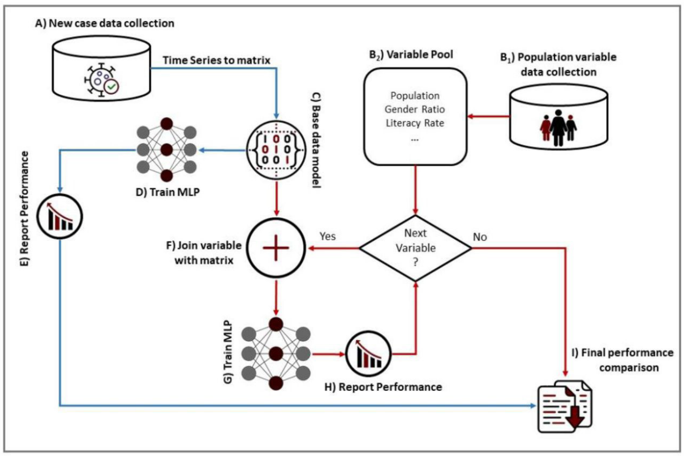

Figure 3: Flow diagram of the modeling process. Confirmed case data is collected (A) and converted to a base matrix, the matrix is then used to train the MLP (D) with a performance report (E). Each demographic variable (B1) is added to the base matrix in a loop-like manner (F) to train the network(G) with a performance report(H), the performances are compared to investigate the effects.

Table 3: Correlation analysis between Correlation between the incidence of COVID-19 and the variables.

Variable	Pearson		Spearman	
	Correlation	p-value	Correlation	p-value
Distance from Qom (km)	-0.26	0.21	-0.31	0.14
Population	0.03	0.87	0.32	0.13
Over 65 years old population (%)	0.26	0.22	0.34	0.1
Urban population ratio (%)	0.26	0.21	0.31	0.14
Literacy rate (%)	-0.06	0.79	-0.1	0.64
Population distribution (%)	0.04	0.85	0.32	0.13
Health infrastructure quality	0.02	0.92	-0.11	0.62
Gender ratio (%)	-0.08	0.69	-0.15	0.47
Population density (population per area)	0.09	0.66	0.61	0.001

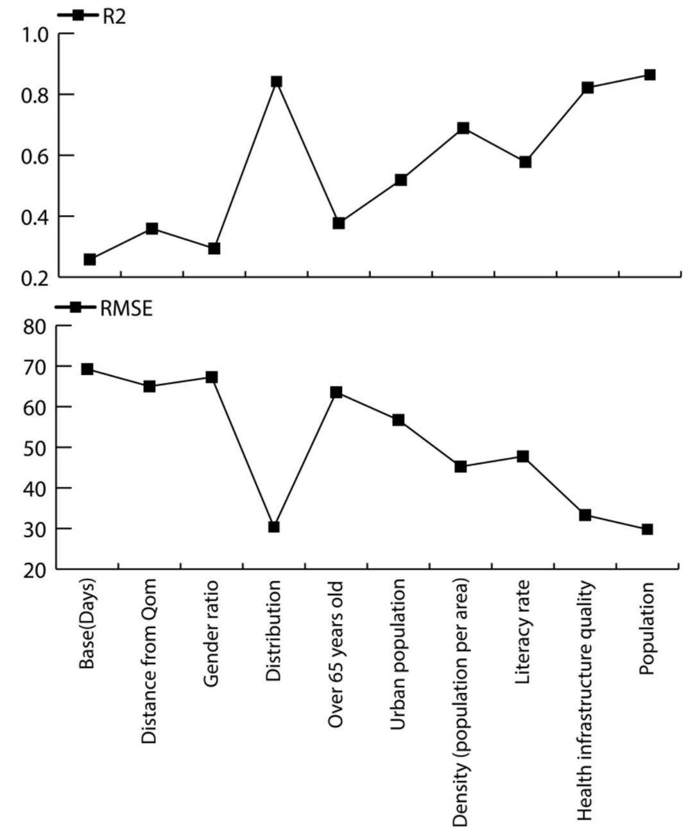

Figure 4: Summary comparative performance reports of the models.

DISCUSSION

Coronavirus's early-stage spread was investigated in this study with the help of an artificial neural network due to its proven potentials in epidemic forecasting. New COVID-19 cases were predicted with the use of daily diagnosed cases from 24 provinces of Iran from 2 April until 4 May, along with demographic and healthcare factors to discover new possible influencers of the disease spread in the country.

The most notable variable is the distance from Qom which shows how provinces are vulnerable to spread from the epicenter of the disease. It was shown in a study 11 in the earlier days of the outbreak; the distance was significantly associated with COVID-19 spread. Our study showed that no significant relationships existed in the observation period of the study. One worth-mentioning issue is that the Nowruz holidays had a great impact on the spread dynamics of the pandemic [5]. Our study was carried out in the period after the holidays in which many travelers had contributed to rapid growth in daily COVID-19 cases and disease spread across the country. This can be in tune with the postulation that with the amount of time after all provinces are affected by the virus, especially after holidays, the distance would not be a criterion for predicting the epidemic.

The study aimed to develop a COVID-19 spread forecasting model in search of variables' contribution to disease prevalence. Models with high accuracy were obtained by taking population, population density, and health indices to account as predictor variables in the neural network. Thus, no direct relationships were found between population densities or population with the incidence rate of COVID-19, both factors was effective in the increase of accuracy of our predictive model. Also, Population density had been proved to be an effective parameter in the outbreak as mentioned in earlier days of the pandemic which could be a sign of dynamic behavior of the pandemic [12].

CONCLUSION

In the course of our research, we faced some limitations mostly due to the lack of data in the case of both variables and daily cases. Many factors may be involved in the virus spread across the country which was not accessible at the province level. Also, we were not able to verify our predicted data to the real data since the data was not published for provinces.

Characterizing the main factors of the early-stage outbreak of such a contagious virus provides opportunities to comprehend the early dynamics of the outbreak. Also, it could be of great importance in confronting future epidemics, in terms of taking preventive action or implementing control measures by policymakers based on prior knowledge.

ACKNOWLEDGMENTS

Authorship

All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Funding

This research received no specific grant from any funding agency in the public or commercial sectors.

Contributor ship statement

JZ designed and supervised the study. MS and AR collected the data, KR, PE, and AD performed data analysis, PE and KR wrote the manuscript, PE edited the manuscript. All authors read and approved the final manuscript.

Disclosures

The authors declare no conflicts of interest or competition.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Patient consent for publication

Not required.

REFERENCES

1. World Health Organization. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV).
2. Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerg Infect Dis. 2020;26(7):1470.
3. Hamzelou J. World in lockdown. New Sci. (2020);245(3275):7.
4. Venkatesan P. COVID-19 in Iran: round 2. Lancet Infect Dis. 2020;20(7):784.
5. Heidari M, Sayfouri N. Did Persian Nowruz aggravate the 2019 coronavirus disease crisis in Iran? Disaster Med Public Health Prep. 2020;14(4):e5-6.
6. Abdi M. Coronavirus disease 2019 (COVID-19) outbreak in Iran: Actions and problems. Infect Control Hosp Epidemiol. 2020;41(6):754-755.
7. Sun Z, Zhang H, Yang Y, Wan H, Wang Y. Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China. Sci Total Environ. 2020;746:141347.
8. Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. Proc Natl Acad Sci. 2020;117(18):9696-9698.
9. Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, et al. Fair allocation of scarce medical resources in the time of Covid-19. N Engl J Med. 2020;382:2049-2055.
10. Lai CC, Wang CY, Wang YH, Hsueh SC, Ko WC, Hsueh PR. Global epidemiology of coronavirus disease 2019 (COVID-19): disease incidence, daily cumulative index, mortality, and their association with country healthcare resources and economic status. Int J Antimicrob Agents. 2020;55(4):105946.
11. Dadar M, Fakhri Y, Bjørklund G, Shahali Y. The association between the incidence of COVID-19 and the distance from the virus epicenter in Iran. Arch Virol. 2020;165(11):2555-2560.
12. Ahmadi M, Sharifi A, Dorosti S, Ghoushchi SJ, Ghanbari N. Investigation of effective climatology parameters on COVID-19 outbreak in Iran. Sci Total Environ. 2020;729:138705.
13. Philemon MD, Ismail Z, Dare J. A review of epidemic forecasting using artificial neural networks. Int J Epidemiol Res. 2019;6(3):132-143.
14. Car Z, Baressi Šegota S, Anđelić N, Lorencin I, Mrzljak V. Modeling the spread of COVID-19 infection using a multilayer perceptron. Comput Math Methods Med. 2020.
15. Mollalo A, Rivera KM, Vahedi B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. Int J Environ Res Public Health. 2020;17(12):4204.
16. Shojaei P, Ghanbarzadegan A, Najibi M, Bastani P. Ranking of Iranian provinces based on healthcare infrastructures: before and after implementation of Health Transformation Plan. Cost Eff Resour Alloc. 2020;18(1):1-3.