Sys Rev Pharm 2020;11(9):431-438
A multifaceted review journal in the field of pharmacy

The Correlation of Google Trends as an Alternative Information Source in the Early Stages of COVID-19 Outbreak in Indonesia

Elly Usman[1], Ricvan Dana Nindrea[2]

[1]Department of Pharmacology, Faculty of Medicine, Universitas Andalas, Padang, Indonesia
[2]Department of Public Health and Community Medicine, Faculty of Medicine, Universitas Andalas, Padang, Indonesia

Corresponding Author: Elly Usman
Email: ellyusman@med.unand.ac.id

ABSTRACT

This study conducted correlation of google trends as an alternative information source in the early stages of COVID-19 outbreak in Indonesia. Time series approach was used in this study. We sought to compare the official COVID-19 report in Indonesia accessible on a daily basis as well as information searches in Google Trends regarding COVID-19. Data analysis was performed using moving average in Minitab version 18.0. Correlation were calculated using pearson correlation and Time lag. R value ≥ 0.7 ($p \leq 0.05$) was defined strong correlation. Moving average analysis showed a linear time series pattern between COVID-19 search trends and the official COVID-19 report. Pearson correlation analysis indicated strong correlation with R value ranging from 0.870-0.927 ($p \leq 0,05$) among the four keywords used in Google trends with the official report of COVID-19 in Indonesia. Time lag correlation inference COVID-19 search trends data could possibly be utilized for an early identification of public reaction against the increasing cases of COVID-19. In the early stages of COVID-19 outbreak found the correlations and similarity of linear time series patterns are shown between COVID-19 search trends and the official COVID-19 report. Public behavior in information seeking is useful in early identification of disease outbreaks in Indonesia.

Keywords: COVID-19; Digital Epidemiology; Google Trends; Indonesia; Information Seeking

Correspondence:
Elly Usman
Department of Pharmacology, Faculty of Medicine, Universitas Andalas, Padang, Indonesia, 25133
Telp: +6282169762531
Email: ellyusman@med.unand.ac.id

INTRODUCTION

In early 2020 the world was faced with an outbreak of novel coronavirus (COVID-19). This was a new type of corona virus that is transmitted via human to human [1]. This virus can affect anyone of any age including babies, children, adults, elderly, pregnant women, and breastfeeding mothers [2]. The phenomenon that illustrates the seriousness of this disease was seen through data showing that 216 countries were infected with COVID-19. Worldwide data counted there have been 13,070,095 infected cases and 572,539 died cases with a mortality rate of 4.4%. In addition to this worldwide situations, Indonesian statistics as of July 15, 2020 reported 80,094 reported as infected and 3,797 deaths with a higher mortality rate than global by 4.7% (https://infeksiemerging.kemenkes.go.id/) [3]. Thus, the COVID-19 outbreak is a serious infectious desease with a huge threat both to Indonesia and the world with high rates of transmission and casualty.

The COVID-19 can be transmitted through close contact and droplets initially and confirmed the possibility of airborne transmission recently. People who are at higher risk of getting infected are those who are closely contacted to COVID-19 patients and those who look after the COVID-19 patients [4]. Therefore, the world government as well as Indonesian government conduct mitigation measures as the crucial application in health and community services.

Knowledge about prevention and control measures is still limited about COVID-19 [3]. The impact on this condition almost all society seek for information of COVID-19 transmission, symptoms, and treatment by using various sources of media. However, since government's policy of physical distancing is implemented, access to acquire information was more commonly used through electronic media and internet [5]. The Google search engine is the most popular for searching online information [6].

In recent years the use of digital traces through the Google search engine has been used as a sources of information in knowing health-related conditions. The use of digital traces might contribute to the identification of COVID-19 epidemiology by exploring disease situations in the community and changes in health conditions by utilizing digital track records [7], [8].

Since the fact that the development of internet utilization has advanced. In consonance with the growth of mobile phone and the evolution of artificial intelligence. Therefore, in the current COVID-19 pandemic the use of digital epidemiology can assist to the identification of conditions in the society and COVID-19 surveillance [7], [8]. The utilize of Google search trends can possibly contribute the gap in general health surveillance in developing countries where the data often restrained from late reporting, incomplete data, or lack of infrastructure due to low budget [9], [10].

The utility of digital traces that continues to be investigated for epidemiology is reviewed in search engines [8]. It uses information search patterns through the use of search terms by determining the location and search period at a certain time. Google search data can be accesed through the website of Google Trends (https://trends.google.com/trends/). The Google Trends

data associated with conventional surveillance data [11-13]. The results of other studies described the advantages of utilizing Google Trends in the initial stages which can understand the initial conditions of health problems and low cost. But, the utilization of Google search data is still low because many obstacles from media and event the government compared to its epidemiological role [14], [15].

Internet and Google usage in Indonesia were 54.7% and 98.0%. This situation is an opportunity for the use of Google trends in supporting health problem solving [16]. This is the first study assessing correlation of google trends as an alternative information source in the early stages of COVID-19 outbreak in Indonesia.

MATERIALS AND METHODS

In this time series approach, we collected the official COVID-19 report from March 2, 2020 (COVID-19 first case found in Indonesia) to April 27, 2020. We used the official report of COVID-19 infection from the Ministry of Health, Republic of Indonesia (https://infeksiemerging.kemkes.go.id/) and data of search queries COVID-19 on Google Trends. The official report of COVID-19 was utilized as a standard of value to validate Google Trends web search information (https://trends.google.com/trends/).

COVID-19 positive results were confirmed from laboratory tests using a polymerase chain reaction (PCR) test in the form of a swab reported in the official report of COVID-19 infection from thirty four provinces in Indonesia of which accessible on a daily basis after its first confirmed case of COVID-19 hence Indonesia declared a national state of emergency. COVID-19 have been found in 34 provinces in Indonesia, which are Jakarta, Banten, Central Java, West Java, East Java, Yogyakarta, Bali, Aceh, North Sumatra, West Sumatra, Riau, Riau Islands, Jambi, Bengkulu, South Sumatra, Lampung, Bangka Belitung Islands, North Kalimantan, East Kalimantan, West Kalimantan, Central Kalimantan, South Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, West Sulawesi, West Nusa Tenggara, East Nusa Tenggara, Gorontalo, North Maluku, Maluku, Papua and West Papua. Data checking was conducted to ensure the sufficient of the data [3].

The official COVID-19 report data thereafter converted into the same interval based on the relative search volume (RSV) of Google search data, to contrast the official COVID-19 report and Google movements data in a linear time series. This method was also applied in previous research to transform COVID-19's official report into interval data and presented on a scale from 0 to 100 [6]. The interpretation of this interval elucidates the absence of COVID-19 cases are expressed by 0 while the highest prevalence of COVID-19 cases is expressed by 100 during March 2 - April 27, 2020.

We compared normalized COVID-19 cases with the number of searches on Google regarding to COVID-19 to be observed in the same period of time. The searches on Google was explained in several keywords used by the Indonesian netizen in online searching related to COVID-19 on Google. Data were downloaded in a comma-separated values (CSV) file from Google Trends website (https://trends.google.com/trends/) and are accessible on a daily basis. Data was attained by web searching through the use of search terms which include definitions of disease, symptoms, transmission and treatment (Table 1).

Keywords were gathered from Google Trends (search terms listing the most frequently used) and Google Correlate (search terms that have a similar pattern with the search term '*COVID-19*' or '*Coronavirus*').

Data collected from Google Trends was then changed from daily to weekly periods using averages. Subsequently, the data findings compared using a single graphical form. Graphs that have relatively similar linearity patterns can then be visualized using moving average analysis. Moving average analysis is used to measure the similarity of patterns between official COVID-19 reports and Google Trends data in more detailed ways. The similarity of graphs includes the pattern linearity, the similarity of leap, and the similarity in the number of COVID-19 cases.

Data analysis was performed using the Pearson Correlation test of search terms with the highest similarity pattern with the official COVID-19 report. R value ≥ 0.7 ($p \leq 0.05$) was defined strong correlation. Time lag correlation analysis with a p value ≤ 0.05 for the search term with the highest correlation. Time lag correlation is used to calculate the correlation between the time lag variables and the official COVID-19 cases. Statistical analysis was performed using Stata version 14.2 and Minitab version 18.0.

RESULTS

Data of positive confirmed COVID-19 cases by provinces in Indonesia (Figure 1).

In figure 1 known Indonesian data in the early stages of COVID-19 during March 2, 2020 to April 27, 2020 reported 9,096 confirmed as infected. The highest prevalence of COVID-19 cases was found in Jakarta (3,869 cases), West Java (951 cases), East Java (796 cases) and Central Java (666 cases).

Time series of COVID-19 cases in Indonesia (Figure 2).

Figure 2 known the time series of COVID-19 cases in Indonesia started from March 2, 2020 to April 27, 2020. There were four highest peaks of COVID-19 cases, with the highest peak on April 25 and April 26, 2020 which involved 436 new confirmed cases.

Moving average of COVID-19 cases and information search using Indonesian keywords '*COVID-19*', '*gejala*', '*penularan*' and '*pengobatan*' in Indonesia (Figure 3).

In figure 3 known the search keywords in bahasa Indonesia such as '*COVID-19*', '*gejala COVID-19*', '*penularan COVID-19*' and '*obat COVID-19*' have the same linearity of pattern with the official COVID-19 report. Information search using keyword '*COVID-19*' had increased at point 37 (84); 40 (92); 43 (90); 47 (90); and highest at point 48 and 56 (100).

While information search using the keyword '*gejala COVID-19*' had increased at point 42 (75); 43 (85); 44 (85); 47 (90); 48 (98); and 56 (100). Meanwhile, information seeking using the keyword '*penularan COVID-19*' had increased at point 42 (75); 43 (95); 48 (98); and 55 (96). Information search using the keyword '*obat COVID-19*' it increased at point 28 (69); 43 (96); 48 (98); and 56 (95). Based on official report of COVID-19 data provided by Indonesian Ministry of Health compared to Google Trends data utilizing information searches with these four keywords there was a similarity of increase in points 40, 43, 48, 55, and 56.

Results of Pearson correlation analysis to compare information seeking using Indonesian keywords '*COVID-19*', '*gejala*', '*penularan*' and '*obat COVID-19*' in Indonesia and the official report of COVID-19 cases (Table 2).

Table 2 known that there was a strong correlation with the value of R ≥ 0.7 (p ≤ 0.05) between the official report of COVID-19 in Indonesia and Google Trends data. The highest correlation of these four keywords from the highest to lowest were the symptoms of COVID-19 (R = 0.927), definition of COVID-19 (R = 0.914), transmission of COVID-19 (R = 0.894) and treatment of COVID-19 (R = 0.870).

The results of time lag analysis between information search using Indonesian search terms '*COVID-19*', '*gejala*', '*penularan*' dan '*obat COVID-19*' in Indonesia (Table 3).

Table 3 presented there was a high correlation with the value of R ≥ 0.7 (p ≤ 0.05) between the official report of COVID-19 in Indonesia with Google Trends data. This correlation occured one day before which has R value ranging from 0.781 to 0.900. Information search using the keyword '*gejala COVID-19*' on the previous day positively correlated with the official report of COVID-19 (R = 0.900; p ≤ 0.05).

DISCUSSION

Moving average analysis showed a linear time series pattern between Google trends data and the official COVID-19 report. Pearson correlation analysis indicated a clear correlation among the four keywords used in Google trends with the official report of COVID-19 in Indonesia. Time lag correlation inference Google trends data could potentially be used for an early identification of public respond against the increasing cases of COVID-19. This finding is relevant to previous related study revealed that Google Trends for certain questions using surveys about influenza correlated with national surveillance data in South Korea [6].

Prior study suggested that Google Trends traces in Korean languange can be utilized as supplementary data and validate influenza surveillance data [6]. Other studies also evaluated a significant correlation between information search through internet with conventional influenza surveillance [13]. Another study concluded that many people searching for information using internet when they are sick right before getting medical attention [17], [18]. In short, this implies that the trend of searching through internet can be an early sign of the disease outbreak compared to conventional surveillance systems.

In the context of global pandemic, many people are seeking for an up-to-date information regarding COVID-19 which includes the definition, transmission, prevention and treatment. During the last two months based on Google Trends data in Indonesia, information demand through web searching was significantly increased and there are 5 top provinces with high information searching through internet for the past two months namely Yogyakarta, East Java, Central Java, West Java and Jakarta. If we take a look at the very first case of the COVID-19 in Indonesia, the government confirmed it to public on March 2, 2020, whereas there has been a surge public searching for information related COVID-19 in those province in the past two months.

This finding is in line with the trend of information as the first COVID-19 case was identified in Jakarta. Furthermore, for less than a week the confirmation of positive cases of COVID-19 was reported in several other provinces, they were West Java, Central Java, Yogyakarta and East Java. Besides that, some province including Jakarta, West Java, Central Java and East Java are the top highest positive COVID-19 cases in Indonesia. Those provinces that confirmed the COVID-19 cases and those province with the highest prevalence of COVID-19 cases in Indonesia are the same province as the top information search through internet using Google search engine. Therefore, this illustrates that the increase of internet information search on Google Trends can be an early detection of an outbreak.

The utilizing of Google Trends data is considered to be an early identification system in countries with a weak surveillance system [6], [11], [12], [19], [20]. The finding offers a tantalising glimpse of the opportunity using Google Trends as a new instrument to observe and assess community reactions before the rising of COVID-19 cases and through the epidemic. Google trends is possibly utilized in meeting community needs for knowledge, information gaps and public concern that can be identified earlier, more easily and cost effective [6], [21-23].

However, there are some limitations in monitoring information on the internet. Firstly, it digitally can serve as a supporting source but not as a substitute for traditional surveillance systems [6], [11], [20], [24]. Secondly, there are some unreliable information develops on social media that cause public panicking as a result it affects the use of search terms. Consequently, it will change the algorithm of search engines that affect the outcome of monitoring [13], [24-27].

The key finding is that public behavior in information search is useful in early identification of disease outbreaks in Indonesia. For further studies, it is expected to validate the use of Google Trends data specifically in regions with the high prevalence of COVID-19 cases and compare them among regions with high and low internet penetration in Indonesia. Apart from that, involving other variables that influence the information search behavior by the community is needed to consider the relative search volume and then increase the correlation analysis between the observational variables.

CONCLUSION

The similarity of linear time series patterns and correlations are shown between Google Trends data and the official COVID-19 report in the early stages of outbreak in Indonesia. This study confirmed there was a high correlation between the official report of COVID-19 in Indonesia with Google Trends data. Public behavior in information seeking is useful in early identification of disease outbreaks in Indonesia.

ACKNOWLEDGMENTS

The author would like to thanks to Ministry of Health, Republic of Indonesia for updating the official COVID-19 data in Indonesia.

SOURCE OF FUNDING

This study was a part of implementation in fundamental research project supported by Universitas Andalas Padang (17/UN.16.02/Fd/PT.01.03/2020).

CONFLICTS OF INTEREST

The author reports no conflicts of interest in this study.

ABBREVIATIONS

COVID-19: Coronavirus disease 2019
CSV: Comma-separated values
RSV: Relative search volume

REFERENCES

1. Lai CC, Shih TP, Ko WC, Tang HJ, Hsueh PR. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. Int J Antimicrob Agents. 2020; 55(3): 105924.
2. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet. 2020; 395(10223): 470-3.
3. Ministry of Health Republic of Indonesia. Emerging infections. 2020. [Last cited 2020 July 15]. Available from: https://infeksiemerging.kemkes.go.id/
4. World Health Organization. Coronavirus disease 2019 (COVID-19) situation report-66. 2020. [Last cited 2020 July 15]. Available from: https://www.who.int/
5. Gugus Tugas Percepatan Penanganan COVID-19. COVID-19 update in Indonesia. 2020. Last cited 2020 Apr 15]. Available from: https://www.covid19.go.id/
6. Cho S, Sohn CH, Jo MW, et al. Correlation between national influenza surveillance data and Google Trends in South Korea. PLoS One. 2013;8:e81422.
7. Salathé M. Digital epidemiology: what is it, and where is it going? J Life Sci Soc Policy. 2018;14:1–5.
8. Salathé M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. PLoS Comput Biol. 2012;8:1–5.
9. Runge-Ranzinger S, McCall PJ, Kroeger A, et al. Dengue disease surveillance: an updated systematic literature review. Trop Med Int Heal. 2014;19:1116–1160.
10. Das S, Sarfraz A, Jaiswal N, et al. Impediments of reporting dengue cases in India. J Infect Public Health. 2017;10:494–498.
11. Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. PLoS Negl Trop Dis. 2011;5:1–7.
12. Chan EH, Sahai V, Conrad C, et al. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl Trop Dis. 2011;5.
13. Seo D-W, Shin S-Y. Methods using social media and search queries to predict infectious disease outbreaks. Healthc Inform Res. 2017;23:343–348.
14. Alicino C, Bragazzi NL, Faccio V, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty. 2015;4:1-13
15. Cervellin G, Comelli I, Lippi G. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. J Epidemiol Glob Health. 2017;7:185–189.
16. StatCounter Global Stats. Search engine market share in Indonesia [Internet]. 2017 [Last cited 2020 Apr 15]. Available from: http://gs.statcounter.com/searchengine-market-share/all/indonesia
17. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232):1012–4.
18. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. J Med Internet Res. 2013; 15(7): e147.
19. Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012–2013 Influenza Epidemic. PLoS One. 2013; 8(12): e83672.
20. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012–2013 influenza season with daily geocoded twitter data from temporal and spatiotemporal perspectives. J Med Internet Res. 2014; 16(10): e236.
21. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. Theor Biol Med Model. 2014; 11(Suppl 1): S6
22. Kang M, Zhong H, He J, et al. Using Google Trends for influenza surveillance in South China. PLoS One. 2013;8:1–6.
23. Adawi M, Bragazzi NL, Watad A, et al. Discrepancies between classic and digital epidemiology in searching for the mayaro virus: preliminary qualitative and quantitative analysis of Google Trends. J Med Internet Res. 2017;3:1–11.
24. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science (New York, NY). 2014; 343(6176):1203–5.
25. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) google flu trends? Am J Prev Med. 2014; 47(3):341–7.
26. Nindrea RD, Sari NP, Lazuardi L, Aryandono T. Validation: the use of google trends as an alternative data source for COVID-19 surveillance in Indonesia. Asia Pac J Public Health. 2020; 32(6-7): 368-9.
27. Nindrea RD, Sari NP, Harahap WA, Haryono SJ, Kusnanto H, Dwiprahasto I, Lazuardi L, Aryandono T. Survey data of multidrug-resistant tuberculosis, tuberculosis patients characteristics and stress resilience during COVID-19 pandemic in West Sumatera Province, Indonesia. Data Brief. 2020; 32: 106293.

Tabel 1: Keywords used by the Indonesian netizen in searching for online information related to COVID-19 on Google

Num	Category	Keywords	Description	Source
1.	Disease definition	'*korona virus*', '*corona*', '*coronavirus*', *covid-19*'	Disease definition of COVID-19 in bahasa Indonesia	Google Correlate Google Trends
2.	Symptom	'*corona gejala*' *gejala corona virus*', '*gejala covid*', '*gejala covid-19*', '*gejala korona*'	Symptom of COVID-19 in bahasa Indonesia	Google Correlate Google Trends
3.	Transmission	'*penularan virus corona*'	Transmission of COVID-19 in bahasa Indonesia	Google Correlate Google Trends
4.	Treatment	'*obat virus corona*', '*obat untuk corona*'	Treatment of COVID-19 in bahasa Indonesia	Google Correlate Google Trends

Table 2: Pearson correlation analysis

Keywords	COVID-19 cases
COVID-19	0.914*
'gejala COVID-19' (COVID-19 Symptom in bahasa)	0.927*
'penularan COVID-19' (COVID-19 transmission in bahasa)	0.894*
'pengobatan COVID-19' (COVID-19 treatment in bahasa)	0.870*

*significant in p≤0.05

Table 3: Time lag correlation analysis

Time lag	Keywords							
	'*COVID-19*' (COVID-19 definition)		'*gejala COVID-19*' (COVID-19 symptom)		'*penularan COVID-19*' (COVID-19 transmission)		'*obat COVID-19*' (treatment)	
	r	p value	R	p value	r	p value	r	p value
-2	0.809*	<0.001	0.815*	<0.001	0.737*	<0.001	0.689*	<0.001
-1	0.882*	<0.001	0.900*	<0.001	0.847*	<0.001	0.781*	<0.001
0	0.914	<0.001	0.927	<0.001	0.894	<0.001	0.870	<0.001

*significant in p≤0.05

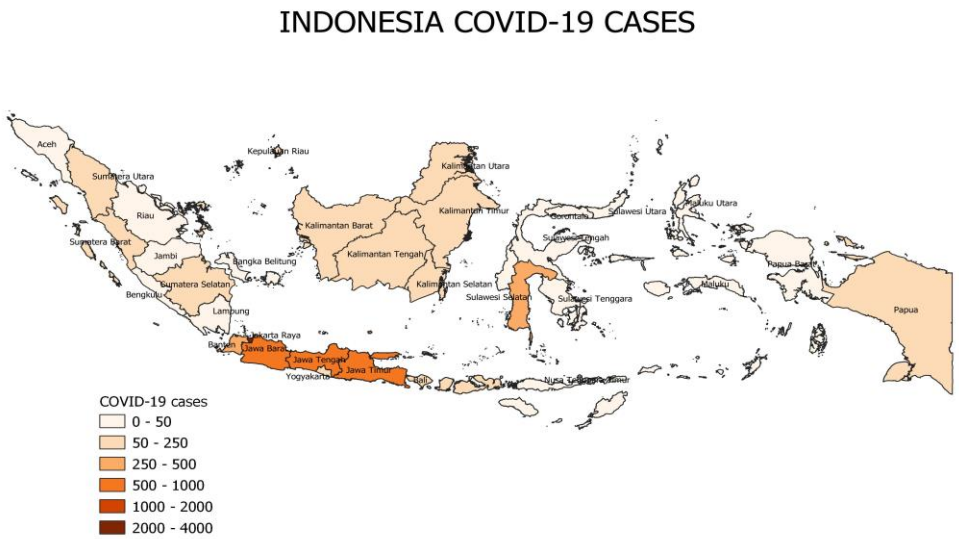


Figure 1: Data of positive confirmed COVID-19 cases by provinces in Indonesia

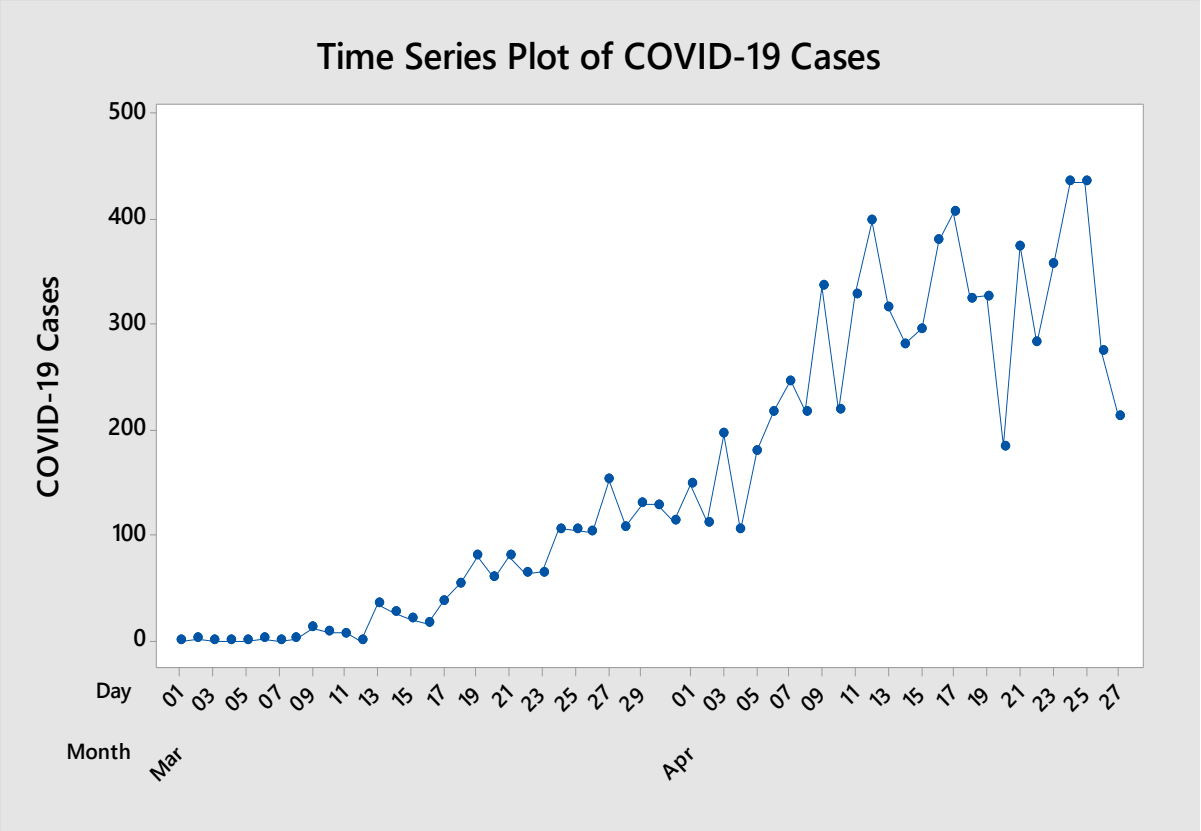
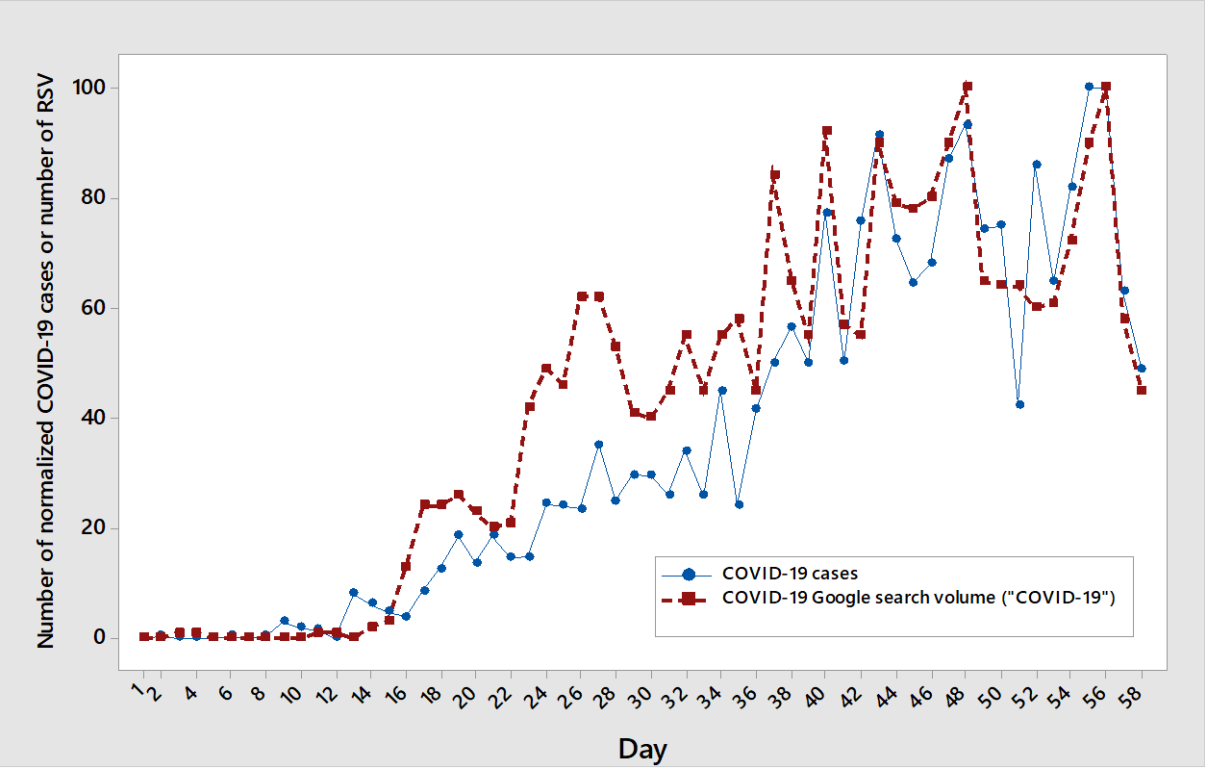
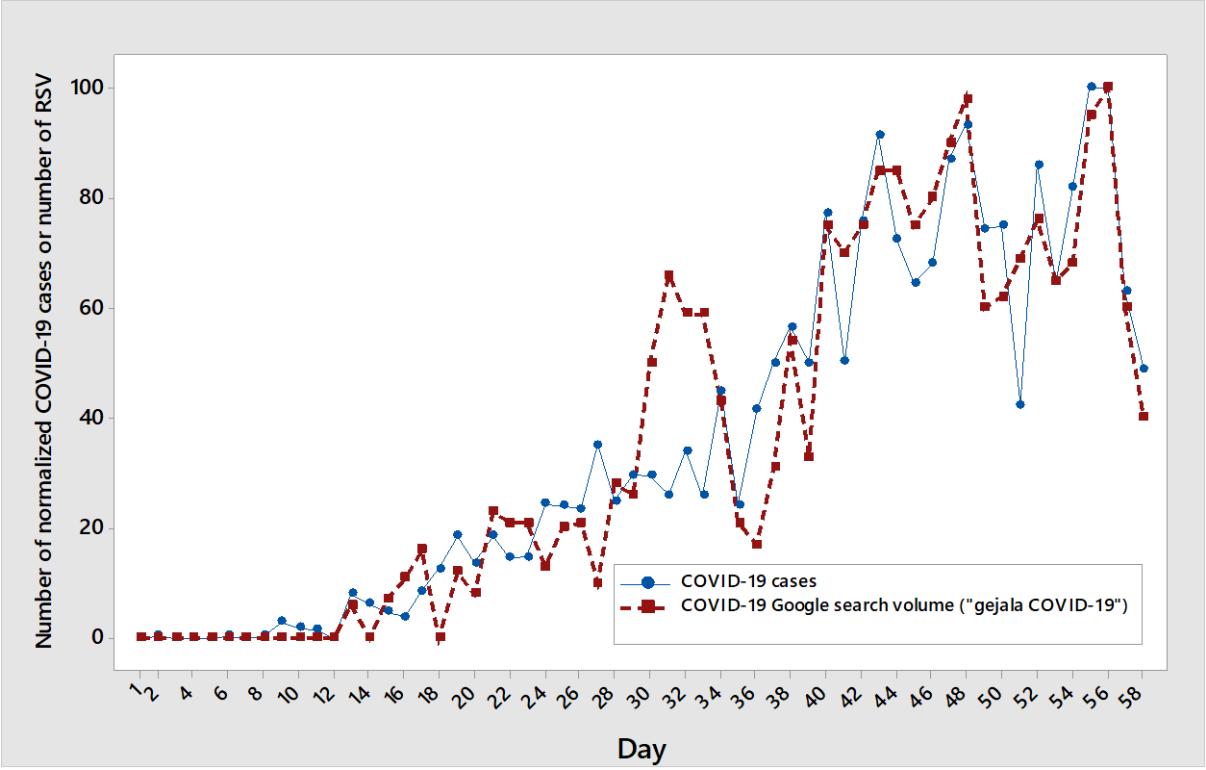


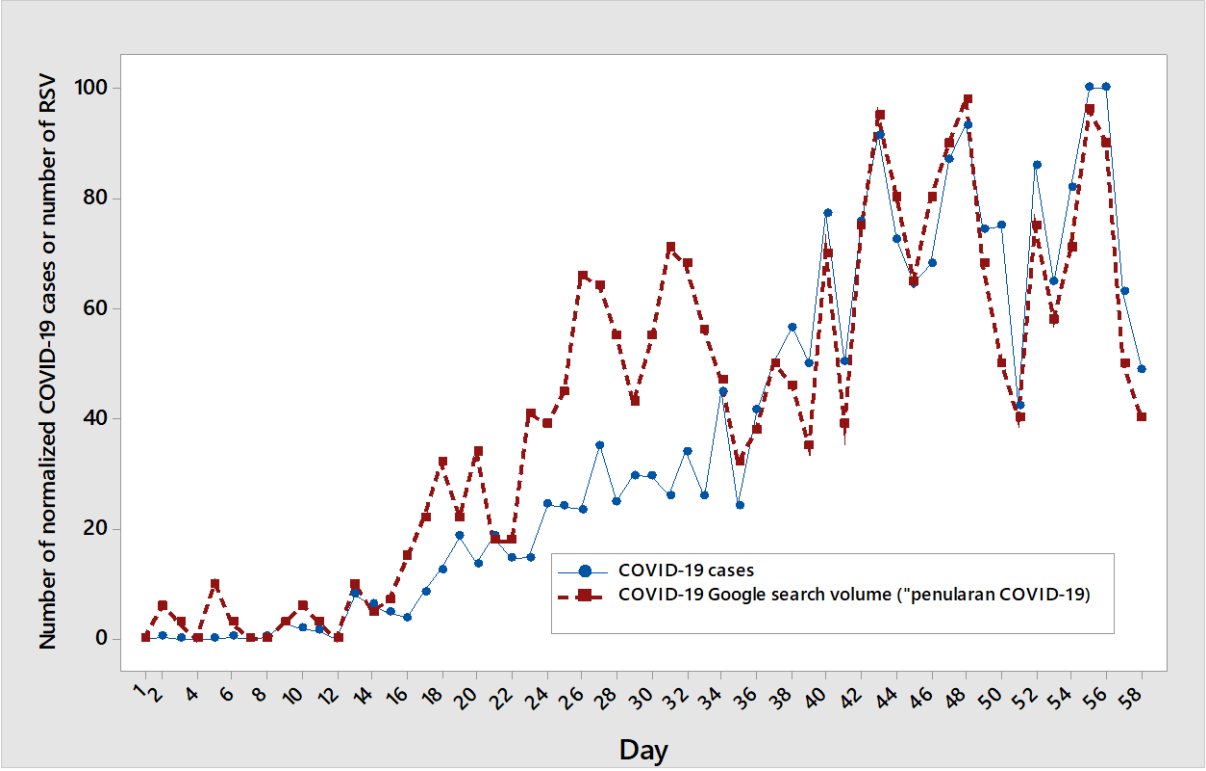
Figure 2: Time series of COVID-19 cases



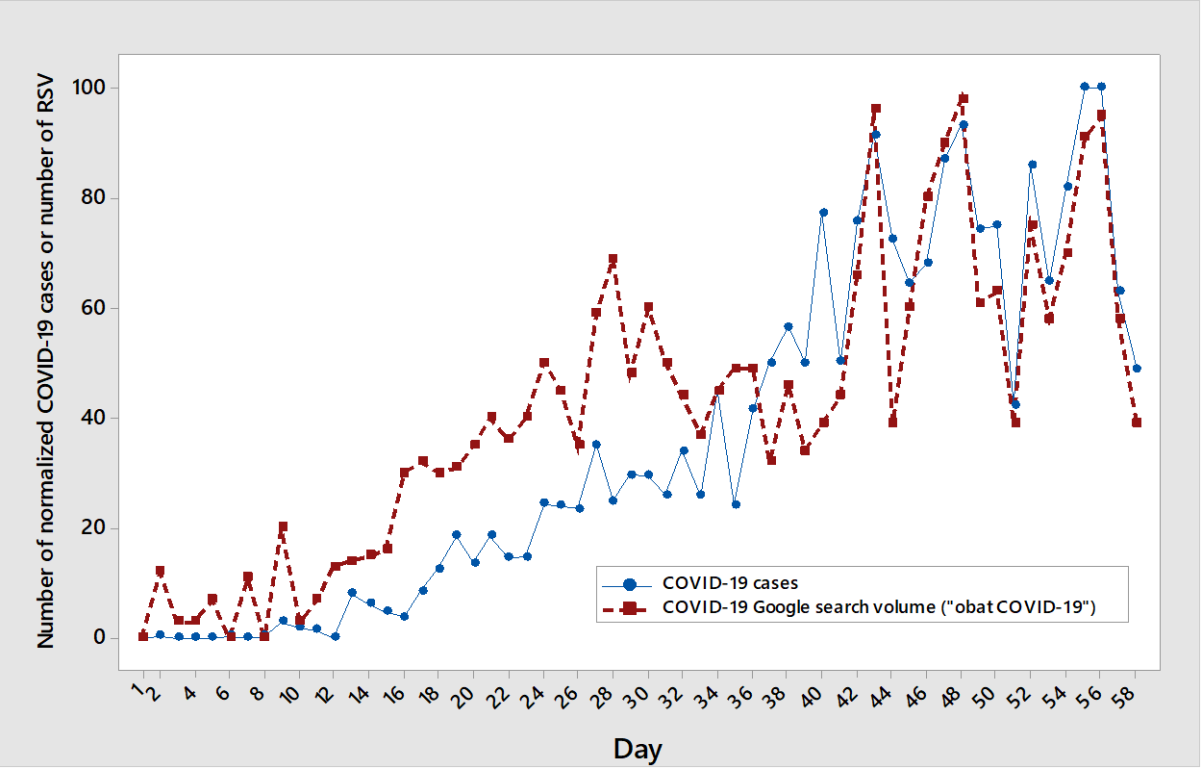
a) Moving average of the COVID-19 cases and information search using Indonesian keywords '*COVID-19*'



b) Moving average of the COVID-19 cases and information search using Indonesian keywords '*gejala COVID-19*'



c) Moving average of the COVID-19 cases and information search using Indonesian keywords '*penularan COVID-19*'



d) Moving average of the COVID-19 cases and information search using Indonesian keywords '*obat COVID-19*'

Figure 3: Moving averages of COVID-19 cases and information search using keywords a) '*COVID-19*'; b) '*gejala*'; c) '*penularan*'; d) '*pengobatan*' in Indonesia