Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 5, No. 4, November 2020, Pp. 299-308

Towards an effective tuberculosis surveillance in Indonesia through google trends

Dhomas Hatta Fudholi*[1], Khairul Fikri[2]
Department of Informatics, Universitas Islam Indonesia Yoyakarta, Indonesia[1]
Master Program in Informatics, Universitas Islam Indonesia Yogyakarta, Indonesia[2]

Article Info
Keywords:
Google Trends, Tuberculosis, Search Term, Surveillance

Article history:
Received 21 September 2020
Accepted 27 October 2020
Published 30 November 2020

Cite:
Fudholi, D., & Fikri, K. (2020). Towards an Effective Tuberculosis Surveillance in Indonesia through Google Trends. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 5(4). doi:https://doi.org/10.22219/kinetik.v5i4.1114

*Corresponding author.
Dhomas Hatta Fudholi
E-mail address:
hatta.fudholi@uii.ac.id

Abstract
Background. The search digital footprint, such as in Google Trend (GT), forms a large dataset that is suitable to be used as surveillance data and supports early warning systems. These advantages become great opportunities for disease surveillance agencies in Indonesia to get rapid early disease monitoring. Objective. Due to limited research in this area and the increasing level of internet penetration in Indonesia, a further study is needed in disease monitoring by utilizing Google Trends. In this research, we explore, analyze and create a set of the best search terms to be used in utilizing GT for disease surveillance in Indonesia, especially Tuberculosis. Method. We use correlation as the technique to define the relatedness between the real case data and GT results. We collect data from the Ministry of Health of Indonesia. From the data, we design a set of new search terms to take GT trend data. The collected data is analyzed using the Pearson correlation. Result. The analysis shows that the studied search terms give strong positive relationships between GT trend data and Tuberculosis cases number in Indonesia. From the correlation analysis, we get a set of proposed effective search terms with the highest score equals to 0.907. Conclusion. Finally, it is possible to monitor and make quick surveillance in tuberculosis in Indonesia through Google Trend and we have created a novel set of search terms that can be used as the basis in monitoring other diseases in Indonesia.

1. Introduction

Internet users left digital footprints when they used search engines. Google Trends (GT) collects these footprints and creates the information of search terms' trend. The use of the trend information led researchers to make studies to use it as the source of information in prediction and early warning systems. The research in this area is increasing [1]. Digital footprint as web-based data is used for monitoring and part of the Infodemiology topic. The studies in the topic itself has been started by [2][3] using Facebook, Twitter and Google sites as a source of determining information for policies and information for health domain.

The research that incorporates GT is not only stressed in the medical domain only but also other fields such as business, law, and education [4][5][6]. The use of GT could be formed differently: assistance in strategy modeling [7][8][9][10], optimization [11][12][13], tracking [4][14], public monitoring [15], verification [15][16][17][18][19], surveillance [4][5][6][14][15][20][21][22][23], checking [4][5][18][19][20][21][22][24][6][8][10][13][14][15][16][17], early warning [9][15], prediction [7][25], product promotion & marketing [11][13][26][27][28], and academic research [11][12][24].

While the GT trend data has been quite giving a prominent result as the aforementioned usages, some errors were found, especially in the health sector. One of them is the soaring 'bird flu' trend in GT, which in reality, there was no 'bird flu' incident found within the area in question [5]. This result is supported by the subsequent years of research, where there are known only under 50% (of the total cases) which correlates with the facts. The mass media coverage influences the results of the GT trend [21][22]. Moreover, the high amount of information search in the areas where there is no disease will appear as if the area is outbreaking and not the original infected area (since there might be very little information search through the internet). In this case, the GT dataset is concluded to be used as secondary data and cannot be considered as factual data [15].

The researcher continues to look for the concept that can closely match GT trend data with the value of factual data so that the GT dataset can become primary data. Of course, it requires transparency support from GT to increase public trust, the openness of the trend determination algorithm [5][10][15]. In short, standardization in how to look at GT trends is needed. Some factors can be the consideration to conclude whether GT trends can be a good use, such as the population in a city, the language used, internet penetration, educational background [15][22], and the selection of keywords/search terms which influences the trends generated by GT.

One of the factors which can drive a good use of GT is the search term. Consideration of determining the search term is needed so that the trend generated by GT positively correlates the factual information. One example, errors can occur while some users are looking for trends in apples. The user needs to use the search term 'fruit apple' and not just the word 'apple'. If only the word 'apple' used, there would be a bias that the following GT trends are about the technology brand Apple Inc. [29]. Also, an addition of certain keywords is possible to lead the trend data in the correct direction as the case of seeing General Electronic Company stock trends (www.ge.com) using the search term 'GE stock' [8]. GT is very sensitive to the usage of words and spelling errors. These two are the factors which may lead to the bad search term [30]. Although GT is not case sensitive, creating search terms must be even more thorough [30].

A careful search term selection is needed to get a good GT trend result (according to the purpose). In the case of the ability in using GT trends for disease monitoring, the disease prevention agency can determine the allocation of aid or prevention resources according to the target and avoid the accumulation of drugs in an area. Therefore, it is really necessary to create a search term template, especially in the health field, such as disease surveillance.

The high internet penetration in Indonesian allows this country to use GT as a disease case surveillance [31][32]. So far, the only research carried out to find the possibility of using GT trend data as disease monitoring in Indonesia is [15], which takes the dengue hemorrhagic fever as the domain. This kind of research supports the disease monitoring process in Indonesia since the report data of the disease case is started to be collected at the district level which is then sent to the province and finally aggregated at the national level. The respective directorate that collects national level Tuberculosis case data is the Directorate of the Prevention and Control of Communicable Diseases, Ministry of Health of the Republic of Indonesia (Ditjen P2PML). The reporting process is done monthly at the district level. However, the final national level report is made when the full year of data is collected.

This research is motivated by the absence of standardization in search terms to see trend data in GT and the lack of web data-based disease surveillance. Problem: the problem that arises is the failure of information or received GT trends due to the wrong search term. The widespread disease and the delay in preventing the disease's spread have resulted in ineffective monitoring of existing diseases. Delays in submitting information on incidents of cases in the central (province) regions make it slow for prevention policies from central (provincial) institutions to emerge or be enacted. We answered two research questions: (i) what is the most appropriate search terms formulation to determine disease trends in Indonesia?; and (ii) how does the trend of GT correlate with disease case surveillance reports in Indonesia?. Hence, to strengthen the previous findings and accelerate country development, especially in the health sector, we specify two purposes of our research. First, investigating effective search terms to create a template of search terms to monitor disease outbreak through GT trend data. To do so, we used correlation analysis on related GT search term and keywords from well-known medical books. We picked the best term that gives a good correlation values to the actual disease case. Second, generalizing the previous finding by providing different domains to analyze. In this case, we choose Tuberculosis as the domain. Tuberculosis is one of the highest causes of death in Indonesia that reach more than 4% [33][34]. We use the Pearson correlation equation to determine the relatedness between GT trend data and factual data. Compared to the current research crowd in GT, this research falls within the 39.4% of correlation studies in GT [35], and within the 32.7% of modeling studies in GT [35]. Domain wise, this research falls within 27% of studies dealing with infectious diseases [10]. However, researchers believe that getting the best search term template and analysing them in a generalized domain should give high impact in the area of research.

This paper is organized as follows. Section 1 presents the background and motivation of our research. Moreover, it also highlights recent studies in the area. Section 2 elaborates on the methodology used within each research step. Section 3 shows the result of search term analysis and Tuberculosis trend data correlation analysis. Section 4 gives further discussion about whether our result leads to the good use of GT trend data. Section 5 concludes our paper.

2. Research Method

Researching effective search term templates for disease monitoring in Indonesia and using them in the Tuberculosis domain will prove its effectiveness is our research aim. This research adopts the steps of using Google Trend in the development of the health domain: (i) Measure online interest; (ii) Explore seasonality or variations; (iii) Find correlations; and (iv) Predicts, nowcast, and forecast [35]. We establish our research methodology in five steps: data collection and preparation, search term categorization and generation, GT trend data collection, correlation calculation and analysis, and evaluation.

2.1 Data collection and preparation

The dataset used in this research is the factual time series Tuberculosis case data in Indonesia. The data will be normalized in the range values of 0-100 to align the factual data to be in the same form as the GT trend data. The normalization follows Equation 1 [15], where vn_i is the normalized i-th data value, v_i is the actual i-th data value, and $max\ (v)$ is the maximum value within the factual data.

$$vn_i = \frac{v_i}{max\ (v)}\ 100 \quad (1)$$

2.2 Search term categorization and generation

Categories are created to group search terms. These categories are based on a medical reference book in Indonesia called Kapita Selekta Kedokteran (in English: Medical Capita Selecta) [36]. We define the categories from the topics used in the book while describing diseases. These topics include definition, risk factors, epidemiology, diagnosis, and clinical pathway.

The search term generation can be carried out in various ways. The study in [37] collects search terms from interviewing respondents who had been stricken with the disease and from consultation with a pathologist. A search term can also be generated from grouping several variables related to the object of research [29] and from related queries found in GT [15][38]. The term can be taken from the categories and sub-categories contained in the GT and can be self-modified [39]. A good search term can also be simply produced by a brainstorming focus group discussion [4] and a simple combination of the name of the disease with the term *penyakit* (in English: disease) [20]. Cho et al. [21], surveyed 100 hospital patients and took a general definition of the disease to create a search term. The use of common definitions of disease is also often used in creating search terms [15][22]. Our research tries to add a new method by using the text analytics technique by taking the word frequency from the Indonesian medical book reference *Kapita Selekta Kedokteran* (KSK) [36]. Hence the set of effective terms will be generated from the KSK combined with the related search term in GT.

When we create the set of the search term from KSK, we combine the main term and the combination term in Equation 2. The main term is derived from the general definition of the disease, while the combination term is derived from the frequency of words from a collection of disease discussion in KSK. In this case, we take the top 20 most frequent terms from each sub-chapter in KSK. The terms are then filtered with the rule that the term does not belong to a country name, disease name, letter of the alphabet, name of the day, number, number of words, unit dose of the drug, unit of the number, unit of time, unit of place or unit of weight.

$$search\ term\ =\ combination\ term\ +\ main\ term \quad (2)$$

2.3 GT trend data collection

The set of a created search term will be used to get the trend. Some configurations or settings are set in taking trends in GT. The searching areas is set to Indonesia, the category is set to 'All Categories', the type of search is set to 'Web search', and the period will be matched with the collected factual data.

2.4 Correlation calculation and analysis

Correlation calculations are done using the Pearson correlation equation [40] with the variable X is the factual data value of the disease and the variable Y is the search term trend dataset. This is intended to get a search term that can describe/represent the factual situation. This research has a strong desire to get a good search term template. To do so, the selected term will be filtered twice. The two filters are the significant coefficient r [41] and the threshold of 'very strong' correlation coefficient interval, which is 0.8-1.0 [41]. Selected search terms have to pass both filters by having a correlation value higher than r and fall within the 'very strong' category. To select, so we called, an effective term in the Tuberculosis domain, the correlation value between the GT trend data of the term and the actual case data should be higher than the significant r value and fall within the 'very strong' correlation category. This study sets a 95% confidence level in the results, so there is a 5% error or significance level. Every year has four data (quarterly). Since this study took five years, the number of data n equals 20. Referring to the table of significant r values in [41], this study has a significant coefficient of r ≥ 0.444. Also, to recall, the 'very strong' correlation category requires the correlation value should be 0.8 or higher. Since the 0.8 threshold is higher than 0.444, we take 0.8 as the correlation value threshold.

2.5 Evaluation

The sub-section has three parts. Firstly, we discuss the novel set of the effective search term in monitoring Tuberculosis through Google Trends. Secondly, we take the proposed effective set of search terms for disease monitoring which has been well experimented using Tuberculosis case data into other disease factual case data. Finally, we compare the proposed general search term with domain specific terms in Tuberculosis from previous study.

3. Results and Discussion

The five steps of the carried-out research have an important role in obtaining a set of effective search words and proving the use of GT for disease monitoring in Indonesia.

3.1 Data Collection and Preparation

The Tuberculosis case dataset was obtained from Ditjen P2PML, Ministry of Health, Republic of Indonesia. The dataset has the case record in quarterly reports from 2014 to 2018 (5 years) in .xls format. The number of cases used in the dataset is the number of cases of discovery. Figure 1 shows the normalization graph of the Tuberculosis cases in Indonesia.

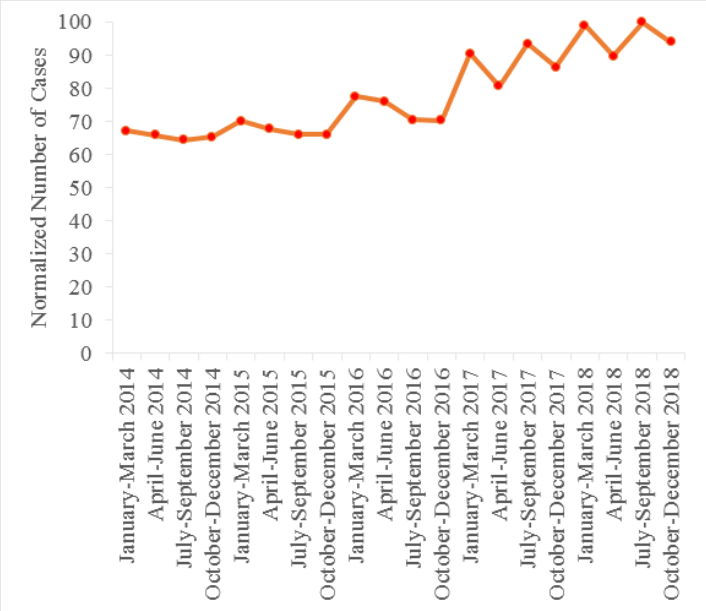

Figure 1. Normalized Tuberculosis Case in Indonesia from 2014-2018

3.2 Search Term Categorization and Generation

In terms of search term categorization, there are four specific categories in the health domain found in previous researches, namely Definition, Symptoms, Treatment, and Disease Vector [7][15][22]. A set of new categories was formed from the KSK book by looking at the generality of discussion of diseases. The sample of the diseases used is taken from the studies of the biggest causes of death in Indonesia [33][34].

The diseases used as the sample in this research are diarrhoea, hepatitis A, hepatitis B, hepatitis C, tuberculosis, bronchiolitis, pneumonia, stroke, heart failure, diabetes mellitus, chronic obstructive pulmonary disease, hypertension, and acute kidney disorders (kidney-hypertension) [41]. After researching for discussion topics within all diseases, we get the general discussion category for each disease. They are Definition, Diagnosis, and Procedures.

In this research, the set of search terms is formed from two sources: (i) the KSK book through text analytics, and (ii) the related query on GT. While we can also use the main term as the sole search term, we excluded them due to the same reason for the possibility of the ambiguity in searching ‘apple’ and not ‘Apple’ as explained in Section 1. The term generation process is explained as follows.

3.2.1 Search Term from KSK

The *search term* is made from the *main term* and *combination term* as in (2). The *main term* for the Tuberculosis domain is selected from the terms of how Indonesian people call this disease, which are 'tb', 'tbc' and 'tuberkulosis'. The *combination term* is obtained from analyzing the most frequent words in each discussion category. From the Definition category, there are twelve extracted terms: ‘infeksi’ ('infection'), ‘primer’ ('primary'), ‘akut’ ('acute'), ‘disertai’ ('accompanied'), ‘paru’ ('lung'), ‘sakit’ ('sick'), ‘afek’ ('affect'), ‘akibat’ ('effect'), ‘hati’ ('liver'), ‘kronis’ ('chronic'), ‘penyakit’ (‘disease’), and ‘virus’. Diagnosis category has fourteen terms: ‘pemeriksaan’ ('examination'), ‘dilakukan’ ('performed'), ‘infeksi’ ('infection'), ‘gejala’ ('symptoms'), ‘pasien’ ('patient'), ‘ditemukan’ ('found'), ‘paru’ ('lung'), ‘bta’ ('bta'), ‘darah’ ('blood'), ‘foto’ ('photos'), ‘normal’ ('normal'), ‘tanda’ ('sign'), ‘untuk’ ('for'), and ‘berat’ ('weight'). The Procedure category has six selected terms, which are ‘terapi’ ('therapy'), ‘untuk’ ('for'), ‘diberikan’ ('given'), ‘pasien’ ('patient'), ‘pemberian’ ('giving'), and ‘cairan’ ('fluid'). From these selected terms, there are 28 unique *combination terms* obtained and 84 new unique search terms from the combination of the *main terms* with the *combination terms.*

3.2.2 Search Term from GT Related Search

While using the *main term* (‘tb’, ‘tbc’, and ‘tuberkulosis’) in GT, Google gives related search terms which then we took and analyzed. We took 49 unique related terms from ‘tb’ term, 50 unique related terms from ‘tbc’ term, and 20

unique related terms from ‘tuberkulosis’. The three collections of related terms are then combined and resulting in 118 new unique terms. We categorized these terms into the same general category Definition, Diagnosis, and Procedures.

3.3 GT Trend Data Collection

The created search terms are used to get GT data. The trends data is collected and subsequently tested for correlation. Some preferences are used in taking trends in GT. The preferences are: (i) the location of search is set to be within the country region of Indonesia; (ii) 'All Categories' is selected in category option; (iii) the type of search is 'Web search'; and (iv) the period is from January 1st, 2014 to December 31st, 2018. The GT trend data is generally stored in a weekly report format. Therefore, the weekly format is transformed to match the Tuberculosis disease case report, which is a quarterly format using the average. From the GT trend result, we found that 44 search terms are not able to retrieve any trend, therefore we put 0 as the correlation value between the GT trend through the respective search term and the real Tuberculosis disease case.

3.4 Correlation Calculation and Analysis

The correlation calculations use the Pearson equation. The correlation is calculated in country level data. Search terms that do not have or do not give any GT trend data are assumed as uncorrelated and each is given a value of correlation equals to zero (0).

Table 1 shows the filtered search terms that are derived from both KSK and related GT terms that comply with the effective criteria given (correlation value higher than 0.8). ST.ID in Table 1 is used as the identity of the search term that is considered effective. The search terms are categorized and have a correlation value. To show further the correlation between the GT trend data through certain search terms with the real case data, we take three terms and visualize both the real disease case and the GT trend data in a high-low line type chart. Figure 2, Figure 3, and Figure 4, show the high-low lines type chart of the search term ‘sakit tbc’, ‘gejala tbc’, and ‘obat tb’, where the top red line in the aforementioned figures are the normalized real Tuberculosis case data. The three visualizations show that the GT trend data form a similar shape, overall, with the factual Tuberculosis case data.

Creating a set of generic search term templates is one of our aims in this research. To do so, we do further analysis to carefully select generic terms that can be used as effective search terms to monitor another disease case through GT. Four search terms, which are st-2, st-4, st-18, and st-19, need to get further attention if we want to use it as a general term template of monitoring disease through GT. The reasons are: (i) the term ‘kelenjar’ (‘gland’) may be specific for internal organ related disease; (ii) the term 'mdr' is an abbreviation of Multi-drug Resistant which means that this is a condition of the patient who is resistant to the Tuberculosis drug and some diseases do not have this condition; (iii) the term 'tcm', can be interpreted not only as the abbreviation of Molecular Rapid Test, a method used in diagnosing Tuberculosis, but may also stands for Traditional Chinese Medicine; (iv) the term 'toss' which is likely the movement to search for disease drugs may be interpreted as applause, and there is no uniformity for all diseases.

Table 1. Filtered Search Term and Their Correlation Value

ST.ID	Category	Search Term	Correlation Value
st-1	Definition	*sakit tbc*	0.831721294
st-2		*kelenjar tb*	0.814140951
st-3		*tb adalah*	0.80776689
st-4		*tb mdr*	0.813800313
st-5		*tbc adalah*	0.878146589
st-6		*gejala tb*	0.808484277
st-7		*ciri ciri tbc*	0.858677014
st-8		*ciri tb*	0.832426032
st-9	Diagnosis	*icd 10 limfadenitis tb*	0.885142371
st-10		*icd 10 tb*	0.879452181
st-11		*icd 10 tb kelenjar*	0.822974118
st-12		*icd tb paru*	0.884785123
st-13		*kode diagnosa tb paru*	0.903883411
st-14		*kode icd 10 tb*	0.866965476
st-15		*kode icd tb paru*	0.841108996
st-16	Procedures	*obat tb*	0.906818458
st-17		*pro tb 4*	0.899244973
st-18		*tcm tb*	0.865422029
st-19		*toss tb*	0.944734838

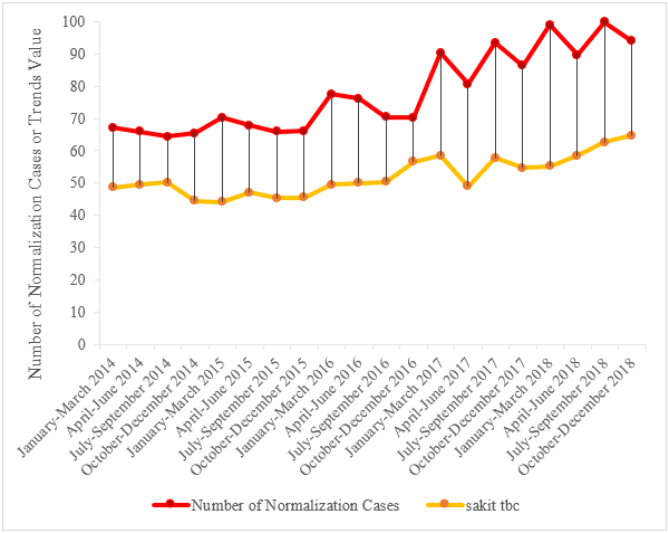

Figure 2. Comparison of Graphic Patterns between the real Tuberculosis Cases and the GT Trend Data of 'sakit tbc' Term (Country Level)

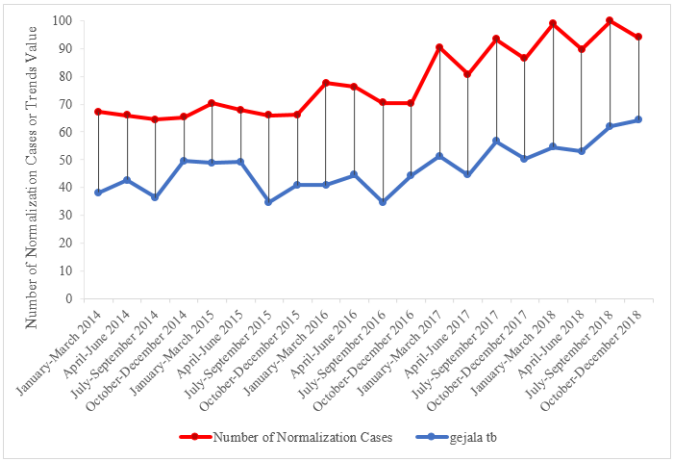

Figure 3. Comparison of Graphic Patterns between the real Tuberculosis Cases and the GT Trend Data of 'gejala tb' Term (Country Level)

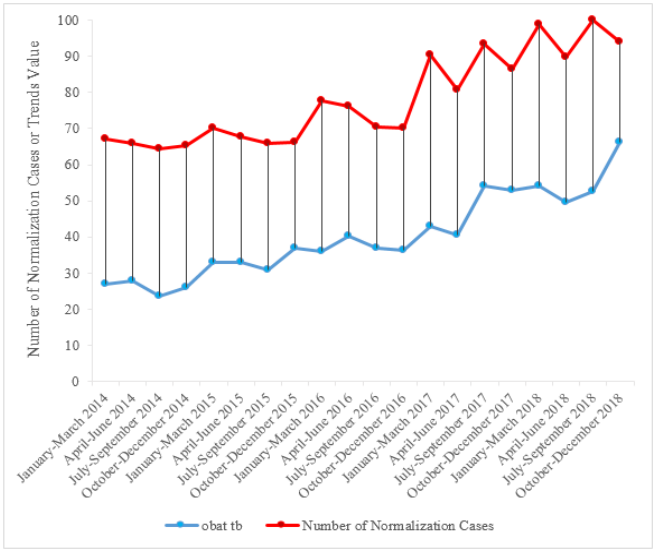

Figure 4. Comparison of Graphic Patterns between the real Tuberculosis Cases and the GT Trend Data of 'obat tb' Term (Country Level)

Terms other than st-2, st-4, st-18, and st-19, can be classified as general terms. The general term means that the term can be used for other diseases. The term 'obat' (‘medicine’) is possible to be attached to other diseases, for instance, ‘obat liver’ ('medicine for the liver') and ‘obat hipertensi’ ('medicine for hypertension'). Especially for the combination terms which has 'icd' ('icd 10 tb kelenjar, kode icd paru', 'kode icd 10 tb’, 'icd 10 tb', and 'icd 10 limfadenitis') and ‘ciri’ (‘characteristic’), we combine these terms into 'kode icd’ (‘icd code’) and ‘ciri’ (‘characteristic’) which is expected to be used for all diseases. Table 2 shows the seven selected general combination terms, which will be attached to the disease name as the main term when used in GT.

Table 2. Proposed Template of General Combination Terms

GCT.ID	Category	*Combination Term*	Usage
gct-1	Definition	*sakit*	*sakit + main term*
gct-2		*adalah*	*main term + adalah*
gct-3	Diagnosis	*gejala*	*gejala + main term*
gct-4		*ciri*	*ciri + main term*
gct-5		*kode icd*	*kode icd + main term*
gct-6		*kode diagnosa*	*kode diagnosa + main term*
gct-7	Procedures	*obat*	*obat + main term*

3.5 Discussion and Eveluation

The main terms used in our research for Tuberculosis are ‘tb’, ‘tbc’, and ‘tuberkulosis’. These terms are likely used by Indonesian people when referring to Tuberculosis. Interestingly, using the same combination term with different main terms results in different correlation values. Table 3 shows correlation values from the sample of different combinations of the main terms with the same combination term. From Table 3, we can see that in general, the abbreviation of the disease gives a higher correlation value. This is logical since people tend to call Tuberculosis with their abbreviations rather than using the full name. Another reason that might be logical is that the user prefers to make a quick typing for a quick search. Hence, we would recommend making a pilot study on how people call certain diseases.

Table 3. Correlation Value of Different Main Term in GT Search Term

Search Term	Correlation Value
gejala tb	0,808484277
gejala tbc	0,474493561
gejala tuberkulosis	0,231676396
obat tb	0,906818458
obat tbc	0,793847879
obat tuberkulosis	-0,43939659
tb adalah	0,80776689
tbc adalah	0,878146589
tuberkulosis adalah	0,456973592
paru tb	0,732352909
paru tbc	0,735198943
paru tuberkulosis	-0,71430875
sakit tb	0,644330603
sakit tbc	0,831721294
sakit tuberkulosis	-0,02687441

The second evaluation is focused on the effectiveness of the proposed search term template. We use the dengue fever disease. The dengue fever case data in Indonesia is taken from [15]. In this experiment, we use three main terms: ‘dbd’, ‘demam berdarah’, and ‘demam berdarah dengue’. Table 4 shows the result of using the search term template in other disease domains is possible. Especially in the dengue fever domain, we can see that except gtc-5 and gtc-6, the terms from the proposed template give ‘very good’ correlation values. gtc-5 and gtc-6 may need to be further studied if they can be generalized to other disease case. In this domain, people seem to rarely use ‘dengue’ in their search. Our analysis is due to ‘dengue’ being a specific term that is not popular in representing ‘dengue fever’ in Indonesia.

In comparison to the previous study in [42], where 19 domain specific search terms are used, added with additional search terms from [36], the proposed template still holds better overall correlation against the Tuberculosis case data in Indonesia, and more suitable as a general term template. Table 5 shows the correlation of the search terms against Tuberculosis case data in Indonesia, where the correlation value is higher than 0.5. From Table 5, we could

see that the only terms that gives good correlation are not having disease identifier word (e.g. tuberculosis or tb) and might be not really correlate to Tuberculosis case, such as the terms 'sakit dada', 'nyeri dada', and 'panas dingin'.

Table 4. Implementation of the Proposed Search Term Template in Dengue Fever Domain

Search Term	Correlation Value
sakit demam berdarah dengue	0
sakit demam berdarah	0,865480198
sakit dbd	0,895286814
demam berdarah dengue adalah	-0,047942979
demam berdarah adalah	0,231742997
dbd adalah	0,602438585
ciri demam berdarah dengue	0
ciri demam berdarah	0,895916795
ciri dbd	0,869226212
gejala demam berdarah dengue	0,036540798
gejala demam berdarah	0,934665721
gejala dbd	0,904301374
kode icd demam berdarah dengue	0
kode icd demam berdarah	0
kode icd dbd	0,206385537
kode diagnosa demam berdarah dengue	0
kode diagnosa demam berdarah	0
kode diagnosa dbd	0
obat demam berdarah dengue	0,232927823
obat demam berdarah	0,95004515
obat dbd	0,901210981

Table 5. Correlation Values of Tuberculosis Case Data in Indonesia Using Term from [42] and [36]

[42] Types	[42] Terms	Terms (in Indonesia language)	[36] Related Terms	Coefficient Correlation
Symtom	chest pain	sakit dada	-	0.89729589
			nyeri dada	0.843159287
Symtom	chills	panas dingin	-	0.905928052
Symtom	cough up blood	batuk darah	-	0.763523351
Diagnosis	TB test	Tes TB	-	0.558742959

DOI Dataset : http://dx.doi.org/10.17632/zgh5j94hfc.1

GT has an ability to see the estimated regional distribution through geoMap of disease cases in each region of a country. Researchers examined the possibility of the Tuberculosis case data correlation in the regional level. From the case data, researchers rank ten provinces with the highest number of cases during 2014-2018. They are, from the highest order, a) West Java; b) East Java; c) Central Java; d) Special Capital Region of Jakarta; e) North Sumatra; f) Banten; g) South Sulawesi; h) South Sumatra; i) Lampung; and j) Papua. From the search term template that we proposed, three search terms that give the position of "West Java" in the first place. Even though this is consistent with the fact that in West Java the highest cases of tuberculosis through the given year, further analysis is needed to see the factors that affect this outcome.

This study has progress compared to previous studies regarding the search terms used to see disease case trends. The search terms obtained in this study have fewer errors than before. The most important finding is that monitoring the spread of tuberculosis can be done in Indonesia through GT. Previously [15], Dengue Hemorrhagic Fever was also known to be observed through GT. So that proves that GT can be a tool to monitor the spread of disease based on the latest web data today.

4. Conclusion

Rapid disease monitoring can be done by incorporating GT trend data. However, we need to further explore, analyze and create a set of the best search terms to be used in utilizing GT for disease surveillance in Indonesia. In this study, we use one disease, which is Tuberculosis to prove our concept on the possibility of doing rapid surveillance on diseases in Indonesia. We use the Pearson correlation to define the relatedness between the real case data and GT results. The Tuberculosis data is collected from the Directorate of Prevention and Control of Direct Communicable

Diseases, Ministry of Health of Indonesia. The analysis shows that the proposed template search terms give strong positive relationships between GT trend data and Tuberculosis cases number in Indonesia. We get seven proposed effective search terms with the highest score of correlation equals to 0.907. From the evaluation given, we can use the search term to monitor and make quick surveillance of tuberculosis in Indonesia through Google Trend. We also evaluate this template to be used in another domain. We can see that by using dengue fever case data, we still can get a really good correlation result. Hence, the search template with a prior study in terms that is widely used to represent certain diseases, can be utilized as the basis in monitoring other diseases in Indonesia.

Acknowledgement

The research is fully funded by the Department of Informatics, Universitas Islam Indonesia, Yogyakata, Indonesia, and is supported by the Directorate of the Prevention and Control of Communicable Diseases, Ministry of Health of the Republic of Indonesia.

References

[1] S. P. Jun, H. S. Yoo, and S. Choi, "Ten years of research change using Google Trends: From the perspective of big data utilizations and applications," *Technol. Forecast. Soc. Change*, vol. 130, no. December, pp. 69–87, 2018. https://doi.org/10.1016/j.techfore.2017.11.009

[2] G. Eysenbach, "Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet.," *J. Med. Internet Res.*, vol. 11, no. 1, 2009. https://doi.org/10.2196/jmir.1157

[3] G. Eysenbach, "Infodemiology : The Epidemiology of (Mis) information," vol. 113, pp. 763–765, 2002. https://doi.org/10.1016/s0002-9343(02)01473-0

[4] C. Pelat, C. Turbelin, A. Bar-Hen, A. Flahault, and A.-J. Valleron, "More Diseases Tracked by Using Google Trends," *Clin. Infect. Dis.*, vol. 15, no. 8, pp. 1327–1328, 2009. https://dx.doi.org/10.3201%2Feid1508.090299

[5] H. A. Carneiro and E. Mylonakis, "Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks," *Clin. Infect. Dis.*, vol. 49, no. 10, pp. 1557–1564, 2009. https://doi.org/10.1086/630200

[6] D. R. Olson *et al.*, "Searching for better flu surveillance? A brief communication arising from Ginsberg et al. Nature 457, 1012-1014 (2009)," *Nat. Preced.*, vol. 1014, no. August, pp. 1012–1014, 2009. https://doi.org/10.1038/npre.2009.3493.1

[7] H. Choi and H. Varian, "Predicting the Present with Google Trends," *Econ. Rec.*, vol. 88, no. SUPPL.1, pp. 2–9, 2012. https://doi.org/10.1111/j.1475-4932.2012.00809.x

[8] L. Kristoufek, "Can google trends search queries contribute to risk diversification?," *Sci. Rep.*, vol. 3, pp. 1–5, 2013. https://doi.org/10.1038/srep02713

[9] T. Preis, H. S. Moat, and H. Eugene Stanley, "Quantifying trading behavior in financial markets using google trends," *Sci. Rep.*, vol. 3, pp. 1–6, 2013. https://doi.org/10.1038/srep01684

[10] S. V. Nuti *et al.*, "The use of google trends in health care research: A systematic review," *PLoS One*, vol. 9, no. 10, 2014. https://doi.org/10.1371/journal.pone.0109583

[11] M. Siswanto and A. Fahriannur, "Google Trend untuk Analisa Pasar Bisnis Online & Pemilihan Keywords pada E-Commerce Web," *Semin. Has. Penelit. dan Pengabdi. pada Masy. Dana BOPTN*, pp. 272–277, 2016.

[12] A. F. Satibi, Suharyono, and Y. Abdillah, "Analisis Pemanfaatan Search Engine Optimization dalam Meningkatkan Penjualan Produk UMKM di Pasar Internasional," *J. Adm. Bisnis*, vol. 50, no. 6, pp. 96–105, 2017.

[13] M. Nuruddin and A. R. A. Udin, "Penerapan Internet Sehat Dan Produktif (Insap) Bagi Kelompok Remaja Di Lingkungan Sumber Ketangi Kelurahan Wirolegi Kecamatan Sumbersari Kabupaten Jember," *Semin. Nas. Has. Pengabdi. kepa Masy.*, pp. 247–250, 2017.

[14] V. Nijman, "CITES-listings, EU eel trade bans and the increase of export of tropical eels out of Indonesia," *Mar. Policy*, vol. 58, pp. 36–41, 2015. https://doi.org/10.1016/j.marpol.2015.04.006

[15] A. Husnayain, A. Fuad, and L. Lazuardi, "Correlation between Google Trends on dengue fever and national surveillance report in Indonesia," *Glob. Health Action*, vol. 12, no. 1, 2019. https://doi.org/10.1080/16549716.2018.1552652

[16] M. P. T. Sulistyanto and D. A. Nugraha, "Implementasi IoT (Internet of Things) dalam pembelajaran di Universitas Kanjuruhan Malang," *SMARTICS J.*, vol. 1, no. 1, pp. 20–23, 2015.

[17] I. Suharyadi, "Peran Penting Asia Africa Smart City Summit (AASCS) 2015 Terhadap Perkembangan Paradiplomasi Kota Bandung," *GLOBAL*, vol. 18, no. 1, pp. 95–107, 2016. https://doi.org/10.7454/global.v18i1.37

[18] J. P. Abiyyu, R. Andreswari, and M. A. Hasibuan, "Implementasi Aplikasi Mobile Modul Penyelenggara Dan Konsumen Kegiatan Di Kota Bandung Menggunakan Metode Iterative Incremental Untuk Meningkatkan Minat Terhadap Kegiatan Di Kota Bandung," *e-Proceeding Eng.*, vol. 5, no. 1, pp. 1381–1391, 2018.

[19] C. O. N. Susanto, "Internet Sehat," 2016.

[20] A. Seifter, A. Schwarzwalder, K. Geis, and J. Aucott, "The utility of 'Google Trends' for epidemiological research: Lyme disease as an example," *Geospat. Health*, vol. 4, no. 2, pp. 135–137, 2010. https://doi.org/10.4081/gh.2010.195

[21] S. Cho *et al.*, "Correlation between national influenza surveillance data and Google Trends in South Korea," *PLoS One*, vol. 8, no. 12, 2013. https://doi.org/10.1371/journal.pone.0081422

[22] M. Kang, H. Zhong, J. He, S. Rutherford, and F. Yang, "Using Google Trends for Influenza Surveillance in South China," *PLoS One*, vol. 8, no. 1, pp. 2009–2014, 2013. https://doi.org/10.1371/journal.pone.0055205

[23] R. A. Strauss, J. S. Castro, R. Reintjes, and J. R. Torres, "Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case," *Int. J. Med. Inform.*, vol. 104, no. November 2016, pp. 26–30, 2017. https://doi.org/10.1016/j.ijmedinf.2017.05.003

[24] L. R. Swari and R. Lakoro, "Perancangan Video Promosi 'Jelajah Pantai Tulungagung' untuk Menunjang Potensi Wisata Pantai di Kabupaten Tulungagung," *J. Sains dan Seni ITS*, vol. 5, no. 2, 2016. https://dx.doi.org/10.12962/j23373520.v5i2.19934

[25] S. Vosen and T. Schmidt, "Forecasting private consumption: Survey-based indicators vs. Google trends," *J. Forecast.*, vol. 30, no. 6, pp. 565–578, 2011. https://doi.org/10.1002/for.1213

[26] D. Rahmanto, Wiyadi, and M. Isa, "Analisis Permintaan Pasar Online Produk Batik di Indonesia," Universitas Muhammadiyah Surakarta, 2011.

[27] K. Falgenti, "Transformasi UKM ke Bisnis Online dengan Internet Marketing Tools," *Ilm. Fakt. Exacta*, vol. 4, no. 1, pp. 62–73, 2011. http://dx.doi.org/10.30998/faktorexacta.v4i1.39

[28] A. D. Riyanto, "Pemanfaatan Google Trends Dalam Penentuan Kata Kunci Sebuah Produk Untuk Meningkatkan Daya Saing Pelaku Bisnis Di Dunia Internet," *Semin. Nas. Inform.*, pp. 52–59, 2014.

[29] M. Dilmaghani, "Workopolis or The Pirate Bay: what does Google Trends say about the unemployment rate?," *J. Econ. Stud.*, vol. 46, no. 2, pp. 422–445, 2019. https://doi.org/10.1108/JES-11-2017-0346
[30] A. Mavragani and G. Ochoa, "Google trends in infodemiology and infoveillance: Methodology framework," *J. Med. Internet Res.*, vol. 21, no. 5, 2019. https://doi.org/10.2196/13439
[31] B. Orenzi, "Statistik Pengguna Digital Dan Internet Indonesia 2019," *BOC Indonesia*, 2019.
[32] APJII, "Mengawali Integritas Era Digital 2019.," 2019.
[33] WHO, "Indonesia: WHO statistical profile Basic," 2015.
[34] DEPKES, "Tekan Angka Kematian Melalui Program Indonesia Sehat dengan Pendekatan Keluarga," *Departemen Kementrian Kesehatan RI*, 2017.
[35] A. Mavragani, G. Ochoa, and K. P. Tsagarakis, "Assessing the methods, tools, and statistical approaches in Google trends research: Systematic review," *J. Med. Internet Res.*, vol. 20, no. 11, pp. 1–20, 2018. https://doi.org/10.2196/jmir.9366
[36] C. Tanto, F. Liwang, S. Hanifati, and E. A. Pradipta, *Kapita Selekta Kedokteran*, IV. Jakarta: Media Aesculapius, 2014.
[37] S. SeyyedHosseini, A. Asemi, A. Shabania, and M. CheshmehSohrabi, "An infodemiology study on breast cancer in Iran: Health information supply versus health information demand in PubMed and Google Trends," *Electron. Libr.*, vol. 36, no. 2, pp. 258–269, 2018. https://doi.org/10.1108/EL-03-2017-0062
[38] E. D'Avanzo, G. Pilato, and M. D. Lytras, "Using Twitter sentiment and emotions analysis of Google Trends for decisions making," *Progr. Electron. Libr. Inf. Syst.*, vol. 51, no. 3, pp. 322–350, 2017. https://doi.org/10.1108/PROG-02-2016-0015
[39] M. A. Dietzel, "Sentiment-based predictions of housing market turning points with Google trends," *Int. J. Hous. Mark. Anal.*, vol. 9, no. 1, pp. 108–136, 2016. https://doi.org/10.1108/IJHMA-12-2014-0058
[40] W. Teng, L. P. Cheng, and K. J. Zhao, "Application of kernel principal component and Pearson correlation coefficient in prediction of mine pressure failure," *Proc. - 2017 Chinese Autom. Congr. CAC 2017*, vol. 2017-Janua, pp. 5704–5708, 2017. https://doi.org/10.1109/CAC.2017.8243801
[41] Sugiyono, *Statistika untuk Penelitian*. Bandung: Alfabeta, 2007.
[42] X. Zhou, J. Ye, and Y. Feng, "Tuberculosis surveillance by analyzing google trends," IEEE Trans. Biomed. Eng., vol. 58, no. 8, pp. 2247–2254, 2011. https://doi.org/10.1109/tbme.2011.2132132