What We Know from Telemedicine Data in Indonesia? Study case using Alodokter, Dokter.id, and Honestdocs

F N Fuadina[1], **N W K Projo**[1], **S Mariyah**[1]

[1] Politeknik Statistika STIS, Jakarta, Indonesia

*Corresponding author's e-mail: 221709696@stis.ac.id

Abstract. The internet and technology development arise in various aspects of life in Indonesia, including in the health sector with e-health. Telemedicine utilization as a form of e-health was still rare among Indonesians because its existence is not as much as e-commerce that is more related to the economic sector. The COVID-19 pandemic has limited people's movement to get health care, but it made people use telemedicine in Indonesia. This research aims to analyze telemedicine utilization in Indonesia and see the health phenomena captured in the data. This research uses descriptive analysis and text mining to determine the utilization of telemedicine with the Named Entity Recognition (NER) and Latent Dirichlet Allocation (LDA) methods. In addition, a literature review is also used to identify the potential use of telemedicine data in collecting health statistics in Indonesia. The results show that telemedicine has been widely used in Indonesia. The clinical teleconsultation data and article titles on telemedicine produce various health topics. Therefore, telemedicine data can potentially be used as a source for collecting health statistics.

1. Introduction

Society needs technology to ease daily works, and the demand for internet facilities is increasingly high today. The Indonesian Association of internet service providers (APJII) conducted the Internet Penetration and Usage Behavior Survey in 2019. The results show that internet users increased by 23.5 million or 8.9% compared to 2018. This growth indicates the need for internet access in various aspects of life to support technological developments.

From an economic perspective, digital technology and internet development bring changes in business models. It drives an economic revolution towards a digital economy in various business lines such as online trading (e-commerce), technology-based startups (startups), and digital financial services or financial technology (fintech) [1]. Business actors could utilize the current number of internet users, especially to boost the economy through the e-commerce industry [2]. The impact of internet and technology development on the economic side might also happen on the social side, such as in the health sector.

According to the World Health Organization (WHO), e-health is the cost-effective and secure use of information and communications technologies to support health and health-related fields, including healthcare services, health surveillance, health literature, and health education, knowledge, and research [3]. E-health became one of the recommended information systems to be developed by the United Nations (UN) in line with the United Nations's request for member countries to implement Information and Communication Technology (ICT) in integrated development infrastructure [3]. In 2005, WHO issued WHA Resolution 28.28 (World Health Assembly Resolution on e-health, May

2005), which also discussed using ICT to support health problems. Along with the development of ICT in the health sector in Indonesia, various forms of e-health have emerged, both built by health agencies (under government) such as independent sites owned by public hospitals or public health clinics and private ones.

Based on the Indonesian Regulation of the Minister of Health Number 46 of 2017 (Permenkes RI Nomor 46 Tahun 2017) about the National E-Health Strategy, one of the e-health implementation initiatives focused on individual health service support is telemedicine [4]. On the other hand, according to the Indonesian Regulation of the Minister of Health Number 20 of 2019 (Permenkes RI Nomor 20 Tahun 2019), concerning the Implementation of Telemedicine Services between Health Service Facilities, Telemedicine is the provision of remote health services by health professionals using information and communication technology. It can take the form of clinical teleconsultation [5]. Clinical teleconsultation is defined as a remote clinical consultation service to help establish a diagnosis or provide management considerations/advice [5]. Some examples of privately owned telemedicine are Alodokter, Dokter.id, Halodocs, Honestdocs, and yesdok.

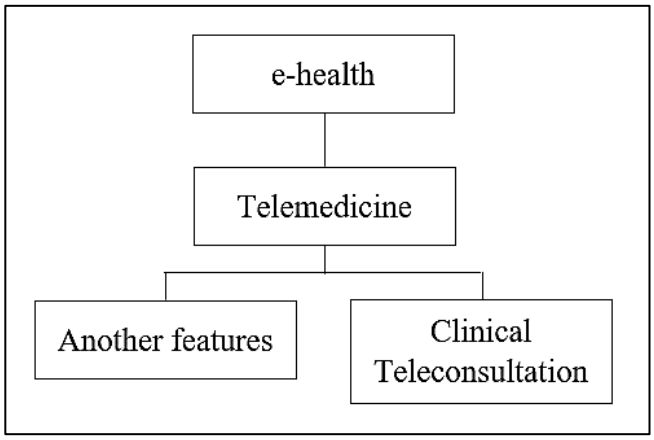

Figure 1. Visual depiction of the relationship between e-health, telemedicine, and clinical teleconsultation

Figure 1 shows a visual depiction of the relationship between e-health, telemedicine, and clinical teleconsultation. Clinical teleconsultation is part of telemedicine and telemedicine is part of e-health. As discussed earlier, telemedicine has several features and one of them is clinical teleconsultation. However, the features in telemedicine discussed in this study are clinical teleconsultation features, doctors, hospitals, health articles, drugs, and diseases.

The government has recognized the importance of e-health in Indonesia. As stated by Daryo Soemitro, the government's strategic work plan related to the implementation of e-health has been developed since 2015. One of the e-health that is currently emerging is telemedicine [6]. Therefore, many governments and private health agencies build e-health to support their services. Indonesia has great potential to implement e-health because electronic media can reach vast areas in Indonesia, although some areas still uncover by electricity and the internet [7]. However, the issue of e-health in Indonesia has not been widely discussed [8]. The opportunity for the development of telemedicine in Indonesia is tremendous [9]. Telemedicine programs are essential for people in disadvantaged and border areas [10].

According to the Indonesian Ministry of Health (MoH), Coronavirus Disease 2019 (COVID-19) is a disease caused by the Sars-CoV-2 virus. The rapid and widespread spread of COVID-19 led the WHO to declare COVID-19 a pandemic. The existence of this pandemic has led to various new regulations that limit social life, one of which is Government Regulation Number 21 of 2020 (PP No. 21 Tahun 2020), about the Large-Scale Social Restrictions (PP PSBB) to accelerate COVID-19 reduction. The regulation states the restrictions on the movement of people and goods to prevent the spread of the COVID-19 virus. All activities are recommended at home, including worship, study, and work [11]. However, health care is an essential need in daily life. If there are health problems, the community needs health facilities and services to provide better treatment. The community has to

strive to fulfill these needs amidst the enactment of movement restrictions. The utilization of e-health and telemedicine is a possible solution to these problems.

Health statistics are essential to help the government or health agencies set policies. The collection of health statistics is carried out by collecting basic statistics by BPS Statistics Indonesia (BPS) and sectoral statistics obtained by MoH. Health data collection carried out by BPS is Indonesian National Socio-Economic Survey (SUSENAS) and the Indonesian Demographic and Health Survey (SDKI). At the same time, the MoH organizes Basic Health Research (RISKESDAS), National Health Indicator Survey (SIRKESNAS), Health Facility Research (RIFASKES), and Manpower Research in the Health Sector (RISNAKES). The data collected by BPS aims to determine health conditions in Indonesia. The majority of the data collected by MoH are regarding facilities, services, and health personnel. Both data collection by BPS and MoH have a long implementation period. The collection of health statistical data is considered less up-to-date and incomplete because of the limited health phenomena covered in those surveys.

Big Data includes data and information that provides volume, velocity, variety, veracity, and values. Humans generate Big Data produced in cyberspace through technological mechanisms; computers, artificial intelligence, and mobile internet [12]. The existence of e-health in the form of an electronic system also produces Big Data that stores data and information. A process of discovering new knowledge through data mining is called the Knowledge Discovery Process (KDP). It can be interpreted as an analysis stage to identify new, helpful, and understandable patterns from large and complex data sets. Big Data is large and complex data so that KDP can be applied in analyzing it.

The Latent Dirichlet Allocation (LDA) produces a more coherent and consistent topic among other topic modeling methods of LDA, Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF) [13]. The LDA method can model health topics from social media data, news, and health articles. For example, research on health topics with Twitter data [13] [14], research on women's health with health news [15], research using data on patient and disease characteristics [16], and research using the information on health paper titles in journals [17]. However, no previous research uses the clinical teleconsultation data on telemedicine. In addition, no studies have identified the potential of such data to complement the current health statistics by BPS and MoH.

Based on the introduction above, this research has the following objectives:

- To describe the use of telemedicine in Indonesia.
- To model health topics from clinical teleconsultation data and article titles.
- To identify the potential use of telemedicine data to complement health statistics in Indonesia.

2. Methods

2.1. Scope of Research

The research was carried out from October 2020 to June 2021. The study uses three telemedicine websites, namely Alodokter, Dokter.id, and Honestdocs. The selection of these three telemedicine is based on the feature similarities and the ease of collecting data. We use data from telemedicine from May 2019 to October 2020. This research uses metadata for collecting health data from BPS and MoH to identify the potential use of telemedicine data.

2.2. Data collection methods

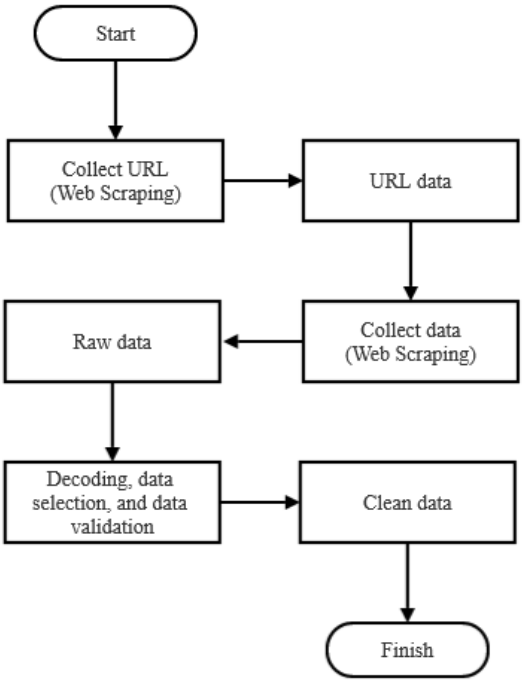

Figure 2. Flowchart of telemedicine data collection

Telemedicine data is collected through Web Scraping using the Python 3.9 programming language with request and selenium modules. Meanwhile, metadata is collected by downloading related documents from the BPS and MoH websites. Figure 2 shows the flow chart of the telemedicine data collection process. The first thing to do in the process is to collect a Uniform Resource Locator (URL). The URL will take the user to the telemedicine page that displays some of the information needed for research. The data to be collected is contained in several different web pages. It is necessary to collect all URLs to program the data collection process efficiently and effectively. After the URL is compiled, then the data can be collected. Before entering the analysis process, the scraped data will be decoded to change the code format into the Indonesian language writing format. Then data selection and data validation are carried out to eliminate data that does not match the required data criteria. The final result of this data collection is "Clean data", and it is ready to enter the data analysis stage.

2.3. Data analysis methods

Telemedicine data will be analyzed to fulfill the first and second objectives of this research. Meanwhile, the metadata for collecting health statistics is analyzed using the literature review method to complete the third research objective. All data will be analyzed, followed by the process of interpretation and visualization of data. In the final approach, conclusions are drawn from the research results.

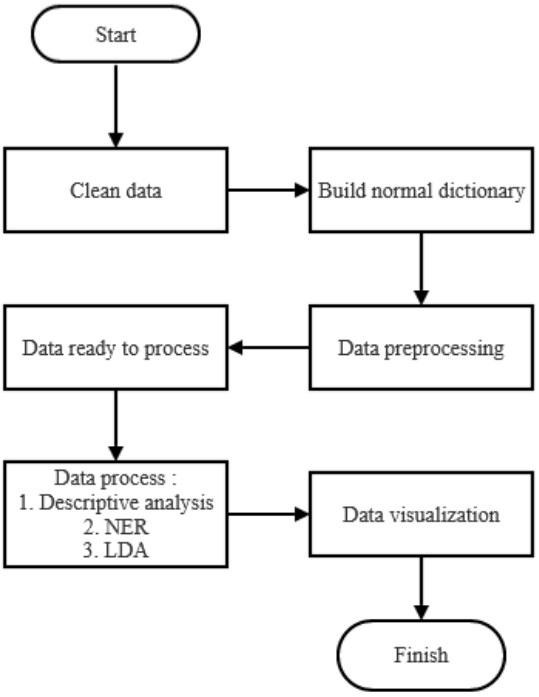

Figure 3. Flowchart of telemedicine data collection

The flow of data analysis carried out on telemedicine data is shown in Figure 3. "Clean data" generated from the data collection process (Figure 2) requires preprocessing to form data that is ready to be analyzed. A "normal dictionary" is first created from the unique word tokens of the data used in preprocessing. The normal dictionary includes words that have gone through the normalization and stemming stages to become basic words. The normal dictionary is then applied to the data, followed by a new stop word process that becomes "Data Ready to Process." This data can then be carried out through descriptive analysis and text analysis (text mining). Text Mining is extracting data to solve the problem of information needs by applying data mining techniques, machine learning, natural language processing, information retrieval, and knowledge management [18].

This research will apply two text mining methods: Latent Dirichlet Allocation (LDA) and Named Entity Recognition (NER). The LDA is a three-level hierarchical Bayesian model in which each collection item is modeled as a finite mix of a set of topic sets. Each topic is modeled as an infinite mixture through an underlying set of topic probabilities. In the context of text modeling, topic probabilities provide an explicit representation of a document [19]. This study uses this method to model health topics from clinical teleconsultation data and health article titles. The NER is a set of statistical, algorithm, and dictionary-based techniques used to search for unstructured text and classify words into predefined categories [20]. This method is intended to identify terms for medicine, diseases, and COVID-19. The NER method used in this research is a dictionary-based NER method by utilizing the FuzzyWuzzy algorithm.

As a form of data validation, researchers compared telemedicine data with official statistics. Doctor data from telemedicine is matched with doctor data from the Indonesian Doctors Association (IDI). Meanwhile, hospital data were matched with data from accredited hospitals belonging to MoH. The data matching process is in the form of measuring the similarity of terms by utilizing the FuzzyWuzzy algorithm.

3. Result

3.1. Analysis of telemedicine utilization based on features

The analysis compares data from three telemedicine and studies each feature using the combined data from the telemedicine pages. The features analyzed are the "Tanya Dokter" feature (clinical teleconsultation), hospital directory, doctor directory, articles, medicine, and diseases.

3.1.1. Tanya dokter (clinical teleconsultation). The "tanya dokter" feature, the embodiment of clinical teleconsultation, is a feature in the form of questions and answers between the user and other doctors/medical personnel in telemedicine. Table 1 shows the comparison and combined data of the three telemedicine that can deliver the condition of the features and actual evidence of using these features.

Table 1. Results of analysis of clinical teleconsultation features

	Telemedicine			Combined data
	Alodokter	Dokter.id	Honestdoc	
Number of questions	163,115	1,188	22,954	187,257
Average answer time	14.19 hours	11.13 hours	1.07 hours	3.90 hours
The impact of COVID-19	not exist	exist	not exist	not exist
Trend	up	down	down	down

The total questions from May 2019 - October 2020 are 187,257 from the three telemedicine pages. Most of the questions came from Alodokter. On average, Honestdocs has the fastest response speed. With the three telemedicine combined, the trend of users of this feature is down, and only Alodokter is showing an uptrend (over time, there are more and more users). Only Doctor.id demonstrates the influence of COVID-19, with a significant increase in the number of users during the entry of COVID-19 in Indonesia, in March 2020. However, identifying the COVID-19 discussion with the NER method shows a 2 % discussion in this feature. The highest discussion in March 2020 coincides with the entry of COVID-19 into Indonesia. During this period, discussing COVID-19 in this feature shows an upward trend (increasing usage). The existence of this trend proves that telemedicine is quite up-to-date to describe a health phenomenon that is widely discussed by the public.

3.1.2. Hospital. The Hospital is a feature that contains complete hospital directory/list information along with general information (address, contact). Table 2 shows the comparative and combined data of these features. There are 2,122 registered hospitals from the three pages, most of which are hospital data from Alodokter. Alodokter has the highest completeness of attributes and the coverage of provinces with hospitals registered in telemedicine in terms of completeness of attributes. From combined data of three telemedicine, the regional coverage is 97.06% or 33 of 34 provinces. The five provinces with the most hospitals are DKI Jakarta, West Java, East Java, Central Java, Banten.

Table 2. The hospital features analysis results

	Telemedicine			Combined data
	Alodokter	Dokter.id	Honestdoc	
Number of hospitals	1,183	692	247	2,122
Attribute completeness	90.11%	7.08%	89.88%	-
Coverage of provinces	97.06%	91.18%	61.76%	97.06%

3.1.3. Doctor. Like hospitals, the Doctors feature displays a directory/list of doctors accompanied by their personal information (specialist, address, contact, place & practice schedule). This feature is only owned by Alodokter and Dokter.id. The results of this feature analysis are shown in table 3. In total, there are 12,859 registered doctors. Alodokter has more doctors and greater completeness of attributes

than Dokter.id. For regional coverage, the presence of doctors has the same value, namely 91.18%, but when the data is combined, the coverage of provinces reaches 97.06%. The five provinces with the most doctors are DKI Jakarta, West Java, Banten, East Java, and Central Java. For both hospital data and doctor data, the five provinces with the most hospitals/doctors are on the island of Java. The fact shows that 56.10% of Indonesia's population live on Java Island [21]. When viewed from the specialty, the five most specialties of doctors are obstetricians, dentists, pediatricians, general practitioners, and internal medicine doctors.

Table 3. Doctor features analysis results

	Telemedicine		Combined data
	Alodokter	Dokter.id	
Number of doctors	12,712	147	12,859
Attribute completeness	45.08%	6.12%	-
Coverage of provinces	91.18%	91.18%	97.06%

3.1.4. Article. The article feature is a page that displays articles related to health. From three telemedicine, a total of 9,096 articles are released during the research data period, with the most significant contributor from the Honestdocs page. Table 4 displays the results of the article feature analysis. Judging from the completeness of the attributes, only Alodokter has incomplete attributes. With regards to COVID-19, only Honestdocs does not show any influence in the published articles. However, when we look at the number of discussions about COVID-19, only 2.21% discuss COVID-19, with the most discussion in April 2020 or a month after COVID-19 entered Indonesia. For the trend of the number of articles, only Dokter.id shows an uptrend (increasing over time).

Table 4. Articles features analysis results

	Telemedicine			Combined data
	Alodokter	Dokter.id	Honestdoc	
Number of articles	2,351	1,120	5,625	9,096
Attribute completeness	95.36%	100%	100%	-
The impact of COVID-19	Exist	Exist	Not Exist	Not Exist
Trend	Down	Up	Down	Down

3.1.5. Medicine and Disease. The Medicine and Diseases feature presents a list and explanation related to medicine/diseases. In this study, we do not analyze the medicine and disease data but rather identify a general discussion of health when a person consults a doctor. People generally ask about the disease they are suffering from, the medicine that must be consumed, or the explanation about medicine. The NER method based on a medicine and disease dictionary will show how medicine and disease are discussed in clinical teleconsultation data from telemedicine.

The NER results show that the number of medicine terms mentioned is 129,043 times and 2,230 words with unique words. Meanwhile, the disease terms are mentioned 224,934 times and 6,634 words with the unique disease terms. The discussed terms between medicine and disease have a ratio of 1: 1.74. A similar thing at each discussion time is that the disease is more talked about than medicine.

Regarding COVID-19, the term "covid" became the top 25 most talked-about diseases during the study data period. At the time of the emergence of COVID-19 in Indonesia (March 2020), "covid" was included in the top 10 most discussed diseases during March - April 2020. For the term medicine related to COVID-19, the public's widely discussed medicine, namely vitamin, supplements, and vaccines., Those are included in the top 25 discussions. During March-April 2020, the three medicine were also included in the top 10 discussions.

3.2. Health topic modeling

The second result is the text analysis by grouping existing text data based on topics hidden from the data. The analysis is carried out using the Latent Dirichlet Allocation (LDA) method. The data are from the "ask a doctor" feature (clinical teleconsultation) and article title data. Before being used for analysis, the two data pass the preprocessing stage to become "ready to process data." This discussion will explain the health topics hidden behind the existing data and the trend of the number of talks on each subject for each time.

3.2.1. Modeling of clinical teleconsultation data topics. In building a topic model, we first do parameter tuning to determine the best topics to get the best model. The tuned parameter is the number of topics, while the value used as a marker of whether the model is good or bad is the coherence score. In the parameter tuning process, the model that produces the highest coherence score is the best model. Figure 4 shows a line chart of the results of tuning 1 to 20 topics and the coherence score for each topic. The highest coherence score is 0.5425, with 12 topics. This value is slightly different from the 15 topics model, which has a coherence score of 0.5408. Several possible discussions can be taken on each topic by utilizing the ten words with the highest probability of constructing the twelve topics, as shown in table 5.

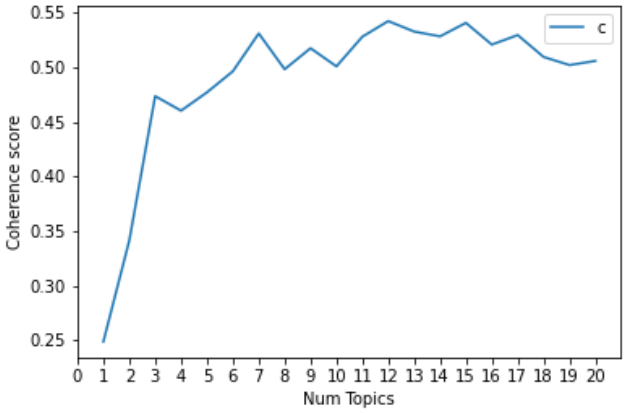

Figure 4. Clinical teleconsultation data LDA parameter tuning results

Table 5. Modeling results of clinical teleconsultation data topic

	Discussion
Topics 1	Digestion in infants includes appetite, drinking, defecation, milk.
Topics 2	Pregnant program; Big/small size of the breast or the penis and its dangers; The presence of a lump in the breast/penis is large/small, and its dangers.
Topics 3	It covers some common symptoms of colds that attack the body, may occur at night/wake up, or last continuously.
Topics 4	The feeling of shortness of breath occurs in older people, or that is felt when the result is being afraid or causes a person to be scared.
Topics 5	Complaints of pain/pain in the right/left side of the head/bones/stomach is occasional.
Topics 6	Itching/spots on the skin that are red/white, there may also be red/white fluid.
Topics 7	Complaints of pain on the right/top/bottom of cavities; Discuss body weight (kg).
Topics 8	Side effects of drinking/consuming something, causing stomach acid problems; Discussing the palms that may indicate stomach acid problems.
Topics 9	Eyes/eyelids/ears/legs/swollen feet; Chicken pox.; Eye/eyelid/ear/foot/foot surgery.

	Discussion
Topics 10	Menstrual problems due to being late or about bleeding; Intimate relations out of blood; Discussing pregnancy in general, or indications of pregnancy due to late menstruation or because of intimate relationships; Bleeding during sex.
Topics 11	KB and its products (pills, injections); Acne on the face.
Topics 12	Scar/surgery; Nausea/vomiting due to surgery or other things; sores on the fingers/toes; Surgery/injury to the heart.

Topic number 10 is the most discussed topic, namely menstruation, bloody intercourse, and pregnancy. The discussion of topic ten is generally related to reproduction which is a bit sensitive and taboo when discussed directly (face to face) when consulting a doctor. This result indicates that when people want to consult about taboo topics, they prefer to use telemedicine rather than consult directly. On the other hand, the least discussed topics are topic number 12 regarding scars/surgeries and nauseous vomit; There are sores on the fingers/toes, operation, and wounds.

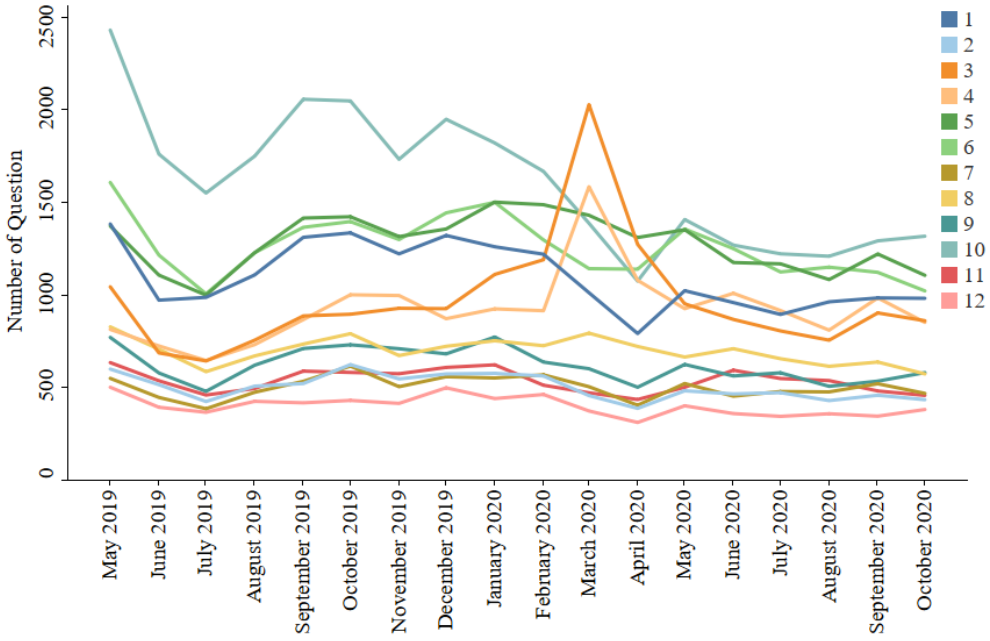

Figure 5. The topic modeling from clinical teleconsulting data

The health trends by topic for each month can be seen as a line chart in Figure 5. It can be seen that apart from topics 3, 4, and 10, the emergence of topics tends to be stable (not showing an extreme change in movement that is different from other topics). Topic 10 appears the most at the beginning of the period and looks very different from the other topics, but over time it decreases and becomes stable like any other topic. Topics 3 and 4 have evident changes at the same time, namely in March 2020. Topic 3 discusses shortness of breath, while topic 4 discusses complaints of pain. These two topics are slightly related to the symptoms of COVID-19, which have appeared in Indonesia since March 2020. Based on a circular issued by MoH and Healthy Living Community Movement (GERMAS), the clinical symptoms of COVID-19 are fever, cough, runny nose, shortness of breath, sore throat, fatigue, and lethargy [22]. The significant changes in topics 3 and 4 indicate that telemedicine shows Indonesia's actual health trends.

3.2.2. Modeling of article title data topics. The next data to be analyzed is the article title data. Before modeling the topic, it is necessary to tune the parameters first. Like the previous data, the tuned parameter is the number of topics, while the marker of whether the model is good or bad is the

coherence score. The highest coherence score is the best model. Figure 6 shows a line chart of tuning of 1 to 20 topics and the coherence score for each topic. The highest score is 0.4485, which is obtained from the 14 topics model. This value is slightly different from the 19 topics model, which has a coherence score of 0.4482. Several possible discussions can be taken on each topic by utilizing the ten words with the highest probability that build on the fourteen topics, as shown in table 6.

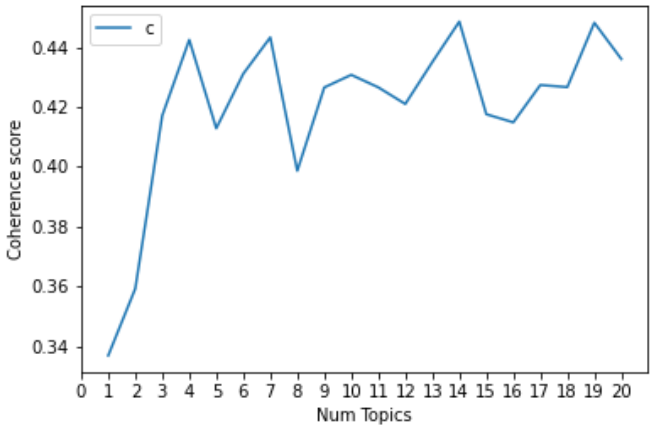

Figure 6. Article title data LDA parameter tuning results

Table 6. Modeling results of articles title data topic

	Discussion
Topics 1	Method/characteristic/cause/medicine for anemia or venereal disease.
Topics 2	Prevent/impact of smoking; How to quit smoking; How to maintain child resistance.
Topics 3	Side effects of fractures/smoking/e-cigarettes on the body or women in particular
Topics 4	Benefits/tips/interesting facts of sport (general) or specifically for children; Benefits/tips/interesting facts of contraception/(for) sex
Topics 5	Causes or ways to get pregnant / urinate/overcome itching in the vagina or the teeth.
Topics 6	Causes/how to deal with bleeding during sexual intercourse; How to keep children healthy.
Topics 7	Causes of headache/dry skin; Natural ingredients good for lungs/vagina/dry skin.
Topics 8	Facts/myths/harmful effects of fungal infections/bowel cancer in general or in particular for the elderly.
Topics 9	Identify the causes/symptoms/treatment of menstrual pain or bone pain in general or specifically for children; Vaping treatment.
Topics 10	Sexually transmitted diseases; strokes; Diseases and Child Health. Bacterial/vaginal/burn infections; Effective way to deal with odors/infections/burns.
Topics 11	Bacterial/vaginal/burn infections; Effective way to deal with odors/infections/burns.
Topics 12	Intimate relationships and how to avoid them; KB, taking birth control pills, how to avoid them; Food; Down body.
Topics 13	Signs of heart disease/coronavirus in general or at a young age; Vitamin.
Topics 14	Dangers/tips/healthy ways to drink or use drugs in general or specifically for children.

Topic 14 is the most frequently appeared during the research data period, namely about the dangers/tips/healthy ways to drink or use the medicine. The topic that came up the least is topic 12 on Intimacy, Family Planning, and Food.

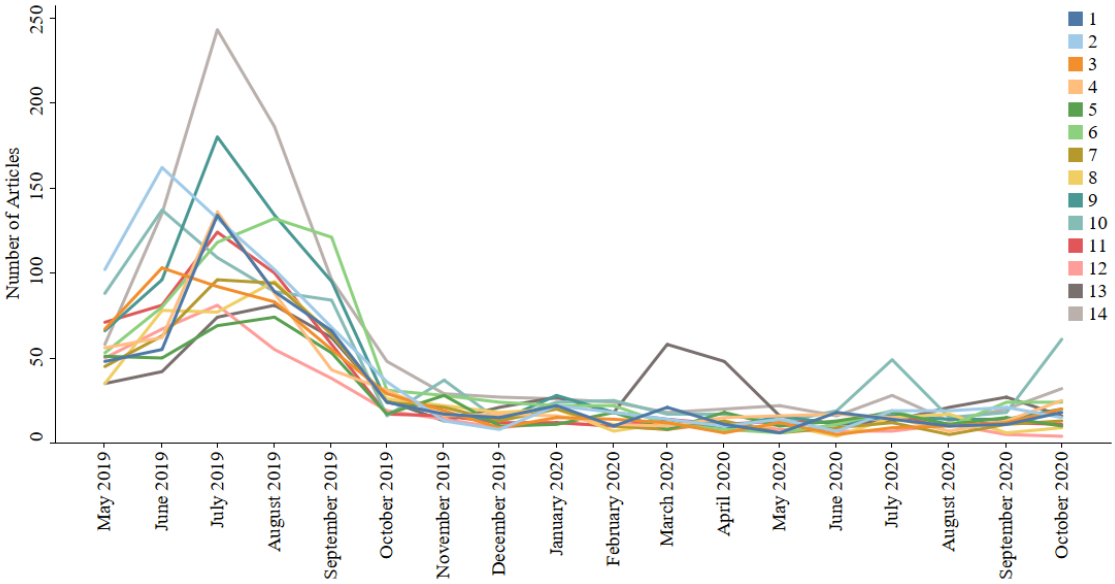

Figure 7. The topic modelling from the articles title data

The topic trend for each month can be seen in Figure 7. It shows that the lines are lined up, showing that, every time, all topics tend to appear in almost the same number. However, one line has a different movement in the middle of the period, namely topic line 13 about signs of heart disease/coronavirus in general and discussing vitamins. A sudden increase appeared in March - April 2020. This topic is related to the health phenomenon in Indonesia at that time, namely the beginning of the emergence of COVID-19.

The researchers explore the data topic modelling results to explain why there is a high value for some article topics at the beginning of the research data period. The number of articles released on the Honestdocs page in those months is very high, reaching more than 500 articles per month. It is different from other months, which are only in the range of under 150 articles. Some topics are highly discussed at the beginning of the data period, coinciding with the high number of articles released in those months. We do not explore other extreme lines in the diagram because the focus of the research is to explain COVID-19 as a form of health phenomenon that can be validated for its occurrence.

3.3. Potential Use of Telemedicine Data in Health Data Collection in Indonesia

3.3.1. Potential use of telemedicine data to complement official statistics. Official Statistics are produced based on high-quality standards nationally and internationally and made/provided by official statistical agencies or other government agencies or international organizations. In Indonesia, one of the laws that discuss official statistics is law No. 17 of 1997 (UU No.17 Tahun 1997) concerning Statistics. According to that law, there are three types of statistics: basic, sectoral, and special statistics [23]. Health statistics in Indonesia are provided from basic statistical data by BPS and sectoral statistics by MoH. There are several surveys to obtain health statistics. In this study, the researchers only discussed several surveys on health data collection related to data obtained from telemedicine.

From BPS surveys, two surveys are closely related to health data, namely The Indonesian Demographic and Health Survey (SDKI) and The National Socio-Economic Survey (SUSENAS). The SDKI is a survey whose main objective is to provide updated basic demographic and health indicators

[24]. The IDHS is held for over five years (last 2017). Regarding topic modeling results, several topics are related to the discussion in the SDKI, namely family planning, reproduction, contraception, and child health. The SUSENAS survey aims to prove data availability on household welfare, including education, health, and purchasing power [25]. The SUSENAS is held every year, in March and September. Regarding the results of topic modeling, the discussion in the SUSENAS also includes pregnancy, family planning, immunization, health care, and healthy food.

MoH conducts four surveys related to research data, namely *Basic Health Research (RISKESDAS) and National Health Indicator Survey (SIRKESNAS)* relate to the community's health conditions. At the same time, *Health Facility Research (RIFASKES) and Manpower Research in the Health Sector (RISNAKES) relate to* health facilities and resources. The RISKESDAS is a survey to provide information on the health status that has been achieved during the last five years and information on the magnitude of the problem of risk factors related to the measured health status, as a consideration in formulating health development policies in Indonesia [26]. The SIRKESNAS is a survey that aims to determine the achievements of the 2015-2019 Ministry of Health Strategic Plan indicators that are not included in the central report, recording, and routine reporting. In addition, it is also to find out the achievements of the National Mid-Term Development Plan (Rencana Pembangunan Jangka Menengah Nasional, RPJMN) indicators in the health sector for 2015-2019, and both are at a national scope [27]. The RIFASKES is facility-based national health research that aims to obtain recommendations for improving the implementation of the National Health Insurance [28]. The RISNAKES is a research conducted by the MoH which seeks to get an overview of personnel in the health sector in health care facilities [29].

The surveys by MoH are more detailed and specific about health. In contrast to the BPS surveys that usually combine health issues with demography or the economy. The RISKESDAS and SIRKESNAS have identical aims with BPS surveys. However, RIFASKES and RISNAKES collect more detailed data on health service resources in Indonesia.

The surveys conducted by BPS and MoH are equally utilized by the government or special agencies who wish to make policies. It supports managing population and health programs, including planning, monitoring, and evaluating government policies. The potential application of telemedicine in supporting the existing surveys are:

- Describing current and past health conditions.
- Planning the development of health services
- Identifying the pattern of health over some time.
- Determining policies on the management of health care programs, such as the availability of services and medicine or other health support tools/products.
- Updating the latest health trends in a shorter period
- Knowing health phenomena broadly, inside and outside the scope of existing surveys/research.
- Identifying rare events or new events that have not been included in existing surveys/research.
- Becoming a trusted publication and directory.

The potential uses apply to telemedicine owned by health companies and government health institutions/agencies. Our results suggest that the main feature that has great potential to complement the current health statistics is clinical teleconsultation. Therefore, the improvement of these features is needed to support the realization of telemedicine data.

Apart from the clinical teleconsultation feature, health service resource data can also be utilized. For state-owned facilities, there is no need to doubt the completeness of the data. However, there is a possibility that the data has not been recorded by the government for private-owned facilities because it is new or not registered or is illegal. If the facility is registered in telemedicine, the authorities can follow up on this incident to not cause problems in the future.

3.3.2. Telemedicine data validation. To validate the telemedicine data, we compare the telemedicine data with official health statistics. In this study, the proof is carried out by matching doctor data from

telemedicine with doctor data at the Indonesian Doctors Association (IDI). While for hospitals, the matching process is between the hospital data from telemedicine with data belonging to MoH

First, matching is done for doctor data by utilizing the IDI "Member Directory" feature to search for doctors on the official IDI web page (http://www.idionline.org/about/direktori-member/). From the Alodokter and Dokter.id pages, 12,859 doctor data are collected. From this data, only 96.85% or as many as 12,454 doctors have a valid address and match IDI membership data. The results show that 53.43% or as many as 6,654 telemedicine doctors are registered (validated) in IDI membership. This result is quite a lot considering more than half of the total doctors registered with telemedicine. Alodokter has more registered (validated) doctors than Dokter.id when broken down for each page.

The second validation process is validating hospital data on telemedicine with accredited hospital data belonging to MoH (http://sirs.kemkes.go.id/fo/home/accreditation). However, the data from MoH only contains hospital data. It is different from telemedicine data, where it also includes several types of health facilities registered as hospitals, for example, clinics, maternity hospitals, and other facilities. Of the 2,122 hospitals data, only 66.21% or 1,405 data are hospitals. The data is then selected based on the validity of the hospital address. The results show that the number of hospitals with valid addresses is 1,390 hospitals, or 98.93%. However, after comparing to MoH data, only 38.49% or as many as 535 hospitals from telemedicine data are registered (validated). The order of telemedicine with more validated hospitals are Dokter.id, Alodokter, then Honestdocs.

4. Conclusions

Based on the research, the researcher draws several conclusions as follows:

- Based on the analysis conducted on its features, the public has used telemedicine a lot. Telemedicine data have already captured the COVID-19 phenomenon through their features. For example, clinical teleconsultation, articles, and medicine and diseases.
- From this research, health topics can be modeled into 12 topics for clinical teleconsultation data and 14 topics for article data, displaying trends from these health topics.
- Telemedicine data has the potential to be used to support official health statistics in Indonesia.

Further studies may enlarge the scope of research, such as using other telemedicine, or it can also be additional features. In addition, it may use different analysis tools such as article content analysis, sentiment analysis of question data, or the use of artificial intelligence for disease detection.

Acknowledgments

The authors would like to thank Politeknik Statistika STIS, BPS Statistics Indonesia for their full support regarding this research.

References

[1] Setiawan A B 2018 Revolusi bisnis berbasis platform sebagai penggerak Jurnal Masyarakat Telematika dan Informasi 9 pp 61-76

[2] Palinggi S and Limbongan E C 2020 Pengaruh internet terhadap industri e-commerce dan regulasi perlindungan data pribadi pelanggan di Indonesia *Seminar Nasional Riset dan Teknologi (SEMNAS RISTEK) 2020*

[3] United Nations 2017 *Global Information Society Watch APC and iTem*

[4] Minister of Health of the Republic of Indonesia (Kemenkes) 2017 Minister of Health of the Republic of Indonesia Regulation Number 46 of 2017 (Peraturan Menteri Kesehatan Republik Indonesia Nomor 46 Tahun 2017)

[5] Minister of Health of the Republic of Indonesia (Kemenkes) 2019 Minister of Health of the Republic of Indonesia Regulation Number 20 of 2019 (Peraturan Menteri Kesehatan Republik Indonesia Nomor 20 Tahun 2019)

[6] Soemitro D 2016 Tantangan e-kesehatan di Indonesia *Buletin Jendela Data dan Informasi Kesehatan* pp 1-21

[7] Kristianto E 2013 Telemedicine Di Indonesia *Jurnal Teknik dan Ilmu Komputer Universitas Kristen Krida Wacana – Jakarta* **02** No.06 pp 167-171
[8] Widiyastuti I 2008 Analisa aplikasi *telemedicine* berbasis *website* di instansi kesehatan pemerintah dan swasta serta potensi implementasinya di Indonesia *Jurnal Penelitian Ilmu Pengetahuan dan Teknologi Komunikasi* **10** No.2
[9] Santoso et al 2015 Perkembangan dan masa depan telemedika di Indonesia *CITEE 2015*
[10] Ariyanti S and Kautsarina 2017 Kajian tekno-ekonomi pada *telehealth* di Indonesia *Buletin Pos dan Telekomunikasi* **15** No.1 pp 43
[11] Government Regulation Number 21 of 2020 (Peraturan Pemerintah Nomor 21 Tahun 2020)
[12] Salahudin S 2019 Mengenal *Big Data* (BD) dan *Big Data Analysis* (BDA)
[13] Islam T 2007 Yoga-veganism: correlation mining of twitter health Data *Conference: 8th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM) 2019* (Alaska, USA)
[14] Paul M J and Dredza M 2014 Discovering health topics in social media using topic models *PLOS ONE e103408 Issue 8* **9**
[15] Cho K W, Kim S Y, and Woo Y W 2019 Analysis of women's health online news article using topic modeling *Osong Public health and Research Perspectives* **10**(3) 158-169
[16] Karami A, Gangopadhyay A, Zhou B, and Kharrazi H 2017 Fuzzy approach topic modelling for health corpora *International Journal of Fuzzy Systems*
[17] Sahria Y and Fudholi D H 2020 Analisis topik penelitian kesehatan di Indonesia menggunakan metode topic modeling LDA (Latent Dirichlet Allocation) *Jurnal RESTI (Rekayasa Sistem dan teknologi Informasi)* **4** (2) 336-344
[18] Feldman R and Sanger J 2007 *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (New York: Cambridge University Press)
[19] Blei D M 2003 Latent dirichlet allocation *Journal of Machine Learning Research 3* 993-1022
[20] Verhoeven E 2019 Fuzzy matching entities in a custom entity dictionary
[21] Yanuar Riezqi Yovanda. Tribun News [newspaper on the Internet]. 2021 Jan 22 [cited 21 Jun 20]; Umum, Nasional. Available from: https://www.tribunnews.com/nasional/2021/01/22/5610-persen-penduduk-indonesia-tinggal-di-pulau-jawa
[22] Minister of Health of the Republic of Indonesia (Kemenkes) [article on the Internet]. 2020 Jan 25 [cited 21 Mar 15]; Available from: https://promkes.kemkes.go.id/informasi-tentang-virus-corona-novel-coronavirus
[23] Laws of The Republic Indonesia Number 16 1997 (Undang-Undang Republik Indonesia Nomor 16 Tahun 1997)
[24] Badan Pusat Statistik 2018 *Laporan SDKI 2017 I* (Jakarta : BPS)
[25] Badan Pusat Statistik 2019 *Ringkasan Metadata Statistik Dasar* (Jakarta : BPS)
[26] Minister of Health of the Republic of Indonesia (Kemenkes) 2019 *Laporan Nasional Riskesdas 2018* (Jakarta : Badan Penelitian dan Pengembangan Kesehatan Kemenkes)
[27] Minister of Health of the Republic of Indonesia (Kemenkes) 2016 *Laporan Survei Indikator Kesehatan Nasional (Sirkesnas) 2016* (Jakarta : Badan Penelitian dan Pengembangan Kesehatan Kemenkes)
[28] Minister of Health of the Republic of Indonesia (Kemenkes) 2019 *Laporan Riset Fasilitas Kesehatan 2019* (Jakarta : Badan Penelitian dan Pengembangan Kesehatan Kemenkes)
[29] Minister of Health of the Republic of Indonesia (Kemenkes) 2017 *Laporan Risnakes 2017* (Jakarta : Badan Penelitian dan Pengembangan Kesehatan Kemenkes)