A bibliometric analysis of Pubmed literature on coronavirus: All time period

Siti Setyawati Mulyono Putri[1*], *Anis* Fuad[1], and *Ahmad* Watsiq Maula[2]

[1]Department of Biostatistics, Epidemiology and Population Health, Faculty of Medicine, Public Health and Nursing, Universitas Gadjah Mada, Indonesia
[2]Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, United States America

Abstract. In late December 2019, there are several reported pneumonia-liked cases with the new strain coronavirus in China. The World Health Organization then assigned this new disease with COVID-19. Coronavirus has been declared as the most responsible agent for a recent public health emergency (PHEIC) in early 2020. The need for further research regarding coronavirus is essential, considering its high threat of public health without any available specific antiviral or vaccine yet. The growth and development of coronavirus related research and thematic trends are still unknown. This study aimed to depict the bibliographic trend of coronavirus all time and pictured the coronavirus research patterns and dynamics throughout the years. Therefore, the objective of this study was aimed to generate a comprehensive bibliometric analysis of coronavirus infection, research topic dynamic and the development of Medical subject heading (MeSH). The study retrieved data from PubMed for the source. Pubmed is chosen because it is the biggest freely available health and medicine electronic database. The R software and Microsoft Excel used for the data analysis. For data visualization, it extracted from VOS viewer. The graph from VOS viewer used as a source for social network analysis.

1 Introduction

Coronavirus has been declared as the most responsible agent for the recent public health emergency (PHEIC) in early 2020. This virus mostly detected in the respiratory tract and gastrointestinal is a family of the virus responsible for outbreaks that occurred in 2002 and 2012. They were not considered to be highly pathogenic to humans until the outbreak of severe acute respiratory syndrome (SARS) as the coronaviruses that circulated before that time in humans mostly caused mild infections in immunocompetent people [1, 2].

In 2002, there was the first case reported from Guangdong, China caused by species of coronavirus. It caused atypical respiratory disease known as Severe Acute Respiratory Syndrome- coronavirus (SARS-CoV). SARS-CoV was an animal virus that adapted to human-human transmission in the recent past. The presence of this animal reservoir implies that it is possible for this virus to again cross into humans and initiate disease outbreaks in the future [3].

Another similar pneumonia case reported in Saudi Arabia in 2012. It also caused by coronavirus strain and spread rapidly in the Middle-East region, so then it called Middle East Respiratory Syndrome- coronavirus (MERS-CoV). MERS-CoV sequences have been found in bats and in many dromedary camels. In humans, MERS attacks lower respiratory tract (LRT) involving fever, cough, breathing difficulties and pneumonia that may progress to acute respiratory distress syndrome, multiorgan failure, and death in 20 % to 40 % of those infected [4]. In late December 2019, there are several reported pneumonia-liked cases with the new strain coronavirus in China. The World Health Organization (WHO) then assigned this new disease with corona virus disease 2019 (COVID-19). It is linked with the seafood and livestock market in the city of Wuhan, China. The new disease spread more rapidly compared to SARS-CoV and MERS-CoV, within weeks COVID-19 already infected thousands of people in mainland China. As of July 30th 2020, 17 540 901 confirmed cases and 677 924 deaths globally [5]. The cases found outside China are confirmed linked to travel history from Wuhan then spreading all over the world through local infection.

On March 2020, the WHO had declared the outbreak of COVID-19 as a global pandemic. The cases found outside China are proof that human-human transmission is possible and can be a threat to global health, especially greater risk to the countries with the weaker health systems. The diseases caused by the corona type virus from time to time resulted to outbreak and become a public health threat. The need for further research regarding coronavirus is essential, considering its high threat of public health without any available specific antiviral or vaccine yet. The growth and development of coronavirus related research and thematic trends are still unknown. This study aimed to depict the bibliographic trend of coronavirus all time and pictured the coronavirus research patterns and dynamics throughout the years. Therefore, the objective of this study was

* Corresponding author: ssmputri@yahoo.co.id

aimed to generate a comprehensive bibliometric analysis of coronavirus infection, research topic dynamic and the development of Medical subject heading (MeSH).

2 Methods

2.1 Data source

Data is collected from Pubmed. Pubmed is chosen because it is the biggest freely available health and medicine electronic database. It is devoted to biomedical sciences and is affiliated with several other National Library of Medicine (NLM) tools that can help optimize the analysis of biomedical subjects. It also provides Medical Subject Heading (MeSH), a professional indexing tool, whereupon adding a new article to Pubmed database, the article will be searched by experts for the main topics it discusses, and a list of MeSH will be assigned for each article [6]. These data extracted from Pubmed at July 20th 2020. The entered query is "coronavirus"[mesh] AND "Middle East Respiratory Syndrome"[tiab] OR "Severe Acute Respiratory Syndrome"[tiab] OR "COVID-19"[tiab] OR "coronavirus"[tiab].

2.2 Data analysis

Medline format from PubMed used as the main data in this study. The R software used in the data cleaning process to retrieve the affiliation country of all authors, and determine the four publication periods (pre-SARS, SARS, MERS-CoV, and COVID-19). In the descriptive analysis step, Microsoft Excel used to explore the number of publications by years and journals. For data visualization, it extracted from VOS viewer. The graph from VOS viewer used as a source for social network analysis (SNA) to show network and flow within entities. Subanalysis is conducted by dividing dataset into 4 time periods : pra-SARS (1949-2002), SARS (2003-2012), MERS (2013-2019) and COVID-19 (2020).

2.2.1 Cleaning country method

The bibliography data from PubMed has one variable affiliation which was consists of affiliation information from each author separated by semicolons (;). In the first step, we split this variable into multiple affiliation variables for each affiliation/author in MS Excel. Each variable had detailed information of affiliation, for example, "*State Key Laboratory of Respiratory Diseases National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, Department of Respiratory Medicine, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou 510120, China.*".

Then we conducted a text mining method in R to scan each variable. We used an R package "countrycode" in this step (the R code listed below).

```
library(cou
ntrycode)
library(read
xl)
medline_afi
ls <-
read_excel("medline_afils.xls",gues
s_max = 20000)
for(n in names(medline_afils)){
medline_afils[paste(n,"adc",
sep="")]<-
countrycode(medline_afils[[n]],
'country.name', 'country.name')}
medline_afils$country <-
paste(toString(medline_afils[,109:2
12],sep = ";",na.rm=TRUE))
```

The result from the process above was the country name from each affiliation variable as a new variable. And the final step, we merge all new variables into one variable "country" separated by semicolons (;).

2.2.2 Social Network Analysis

SNA map was created using Vos Viewer software analysing bibliographic data extracted from Pubmed. Co-occurance analysis was performed for this study. The unit of analysis for pre-SARS, SARS and MERS were using 'MeSH keywords', meanwhile for COVID-19 period was using 'All keywords' which are contains of author keywords and MeSH keywords. Then we choosing minimum numbers of occurance to appear, it set the minimum bar of keywords to become one dot in the SNA map. The next step is verifying selected keywords. We eliminate irrelevant and general keywords that we exclude to analyze such as female,male, child, old, adult etc.

No IRB is required for bibliometric analysis due to no human subject were involved in this study.

3 Results

3.1 General bibliometric indices

The first publication about coronavirus indexed Pubmed appeared in 1949. This publication discuss about the coronavirus infection in animal with title 'Demonstration of an interfence phenomenon associated with infectious bronchitis virus of chickens'. In total until July 20th 2020, we have collected 47 045 publications. The growth rate of coronavirus and coronavirus infection publications increased over time and has great surge in 2020. Most of the publication were journal articles (26 738; 57 %) followed by review (4 839; 10 %), letter (3 561; 8 %), comment (3 206; 7 %), editorial (2 245; 5 %), case report (1 966; 4 %), comparative study (1 490; 3 %) and remaining.

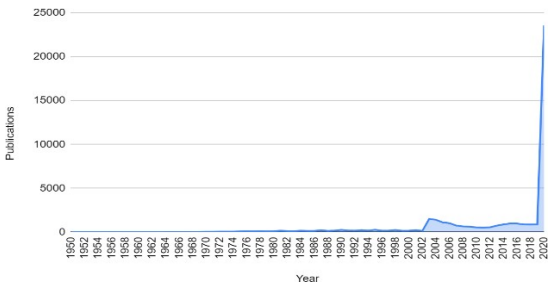

Fig. 1. Publication growth rate all year

3.2 Bibliometric indices by period

We divided the analysis into 4 groups Pre-SARS, SARS, MERS and COVID-19. Pre-SARS data was collected within the range of 1949-2002. SARS period of data is extracted from 2003 where SARS was outbreak and became international concern, to 2012 right before another outbreak of MERS coronavirus infection. MERS data collected in the range of 2013 to 2019. And the last group, COVID-19 was collected from publications in 2020 to the present. The largest amount of publications is within the COVID-19 period with 28 752 publications (61 %), consecutively followed by SARS period with 8 066 publications (17 %), MERS with 5 568 publications (12 %) and pre-SARS with 4 659 publications (10 %).

3.2.1 Phase 1: pre-SARS (1949–2002)

The growth of coronavirus-related publications showed plateau trend since the first study until 1960's. The publications keep steadily increasing after that. In total there were 4 659 publications within 55 yr or average of 84 publications anually. The study of coronavirus increased in the 1990s, approximately around 200 publications each year. First reported atypical pneumonia, later known as SARS was in December 2002 in Guangdong, China [7]. The publication trend started to raise from there, in the year 2002 there were 134 published articles about coronavirus. It was in March 2003 when WHO issued the global alert of a severe form of pneumonia related to coronavirus infection.

3.2.2 Phase 2: SARS (2003-2012)

The total number in SARS period is 8 066 publications (17 %) within 10 years. Approximately there were 806 published article anually and 10 times more productive, a significant increasing number compared to the last period (pre-SARS). The publication trends is increasing until it reach the peak in 2003 by 1 471 publications, right after the SARS outbreak but the keep decreasing years by years until the end of SARS period.

3.2.3 Phase 3 : MERS (2013 – 2019)

In 2013 the publication growth started to increase again with total of 623 publications. Total publications in the MERS period are 5.568 (12 %) publications. Annually, 795 articles published in this period. It showed that in the period after coronavirus outbreak always followed by raising number of publications, even for some years after. The average publications in both SARS and MERS are consistently above 700 publications/year, a very significant number compared to the period where coronavirus were not a threat to human.

3.2.4 Phase 4: COVID-19 (January 2020 – July 2020)

Meanwhile the COVID-19 group was analysed per month since its number is growing rapidly. Total of 28 752 publications within 7 mon. The fast growth of educational sharing is the impact of the great technology that now we have. There are lot of pre-print journals available make it easier for scientist to share their study. It is also the advantage in this outbreak situation when knowledge needs to be shared as fast as it can to help other scientists, medical practitioner and policy maker.

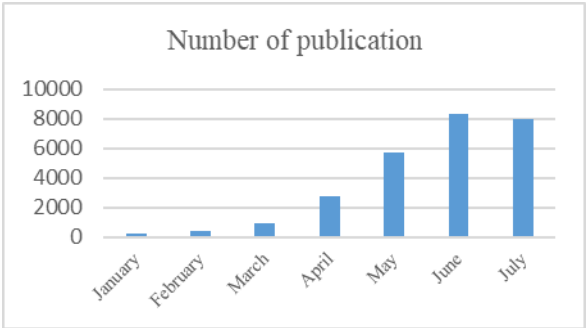

Fig. 2. Publication growth COVID-19 period

3.3 Distribution of publications by countries

The country contribution on the publications was extracted from the country origin of the first author. If we see from the country contribution, (Table 1) the most publication were done by the United States with total of 13 104 articles published, contributed for 27.85 % of all the study. Second most productive country is United Kingdom and followed by China. The United Kingdom published 6 983 articles (14.84 %) meanwhile China, the country where SARS and COVID-19 firstly emerged published 3937 articles or 16.3 % of all the publications. Pre-SARS period publications extracted from 1949 to 2002. From the database, we collected a total of 4 659 publications and 48 countries' contribution. It was the period before the emerging case of SARS and MERS, coronavirus not yet become an international threat since it only resulted in diseases in animal or mild diseases in human. China and Saudi Arabia are not in the top ten countries with the most publications. China only published 4 publications, meanwhile, Saudi Arabia only contributed to one study. China started to become

productive in the research after the outbreak of SARS. The publication was growing significantly, increased for 305 times from 4 publications to 1 221 publications between 2003–2012. A total of 76 countries contributed to 8 066 publications. Within the MERS period, the dataset collected from the year 2013 until 2019. A total of 5 568 publications were published by 86 countries. The domination of the United States as the most productive country for publishing related articles was replaced by China with 16.8 % of total publications. The first reported cases of MERS were from Saudi Arabia. Saudi Arabia manage to published 224 publications or 4.29 % triggered by the MERS outbreak in its region. In the COVID-19 period, by July 2020 there are already 28.752 publications related to the coronavirus, bigger number compared to other periods combined. The growing number of publications is increasing rapidly, approximately 232–233 publications each day up to July 20th 2020.

3.4 Distribution of publication by journals

Meanwhile, all-time most productive journal publisher (Table 2) is Journal Virology. Journal Virology is part of American Society for Microbiology (ASM) journals. This journal covers the updated research on the nature of viruses, its scope including structure and assembly, genome replication and regulation of viral gene expression, genetic diversity and evolution, virus- cell interactions, cellular response to infection, transformation and oncogenesis, gene delivery, vaccines and antiviral Agents and pathogenesis and immunity [8]. The United States is on the top list both in most productive country based on author origin and the most productive country based on journal publisher origin

3.5 Evolution of research topics by Medical Subject Heading

Coronavirus and coronavirus infection both indexed in the Pubmed since 1994. According to the MeSH, coronavirus is a member of coronaviridae which causes respiratory or gastrointestinal disease in a variety of vertebrates [7]. And coronavirus infection is virus diseases caused by the coronavirus genus. Some specifics include transmissible enteritis of turkeys (enteritis, transmissible, of turkeys); feline infectious peritonitis; and transmissible gastroenteritis of swine (gastroenteritis, transmissible, of swine) [8] Social network analysis in (Figure 3) displayed co-occurrence of MeSH in journal articles from 1949 to 2002. Four clusters of co-occurrence MeSH were identified. Red cluster is the biggest one shows coronaviridae, antibodies viral, antigen viral, swine, coronaviridae infection and remaining. Second biggest, the blue cluster showed the most keyword use are murine hepatitis virus, coronavirus infection, virus replication and the remaining. The green cluster is dominated by molecular sequence data. Meanwhile, the smallest yellow cluster informed that 'viral envelope protein' is the most discussed topic. Overall the publications in the Pre-SARS period dominated with biomolecular research. The closer distance between nodes and or thicker edge connecting nodes indicate the higher intensity of co-occurrence.

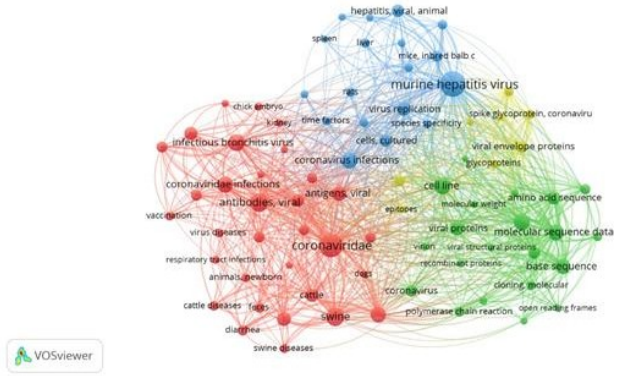

Fig. 3. Social Network Analysis for coronavirus research topic in pre-SARS period.

From the SARS era (Figure 4), the SNA showed an obvious co-occurrence MeSH as shown in the biggest green cluster. It is lead with severe acute respiratory syndrome and followed with disease outbreaks, Hongkong, China, global health etc. The yellow cluster, as the second informed that the main MeSH is sars virus. Another cluster, the red cluster is dominated by coronavirus infection, molecular sequence data, and coronavirus. The smallest blue cluster mostly consists of molecular aspects such as antibodies, spike glycoprotein, membrane glycoproteins, and viral envelope proteins. In this period, the publication starts to change with additional terms emerges such as Hongkong, SARS and global health. The pattern not only focused on biomolecular study but starting to concern on epidemiology and global health

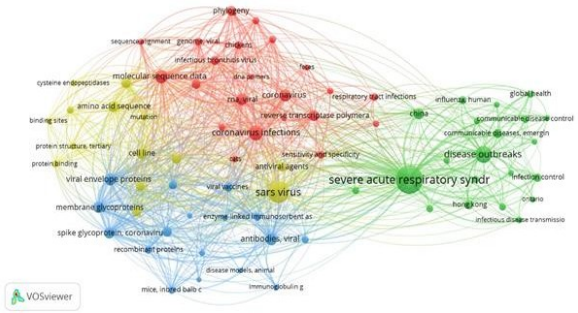

Fig. 4. Social Network Analysis for coronavirus research topic in SARS period.

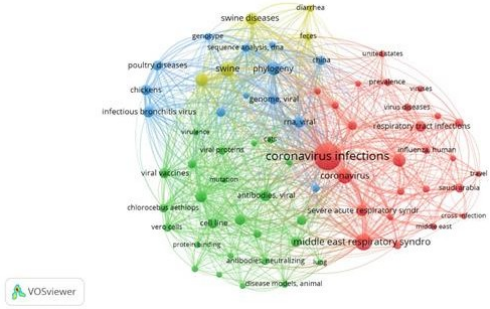

Fig. 5. Social Network Analysis for coronavirus research topic in MERS period

In the SNA from 2013 to 2019 (Figure 5), shows the co- occurrence of the MeSH that divide into four big clusters. This period is when the MERS outbreak in 2013, supposedly the MeSH had changed from SARS to MERS. The biggest red cluster has a prominent of coronavirus infection followed by middle east respiratory syndrome, respiratory tract infection, human, severe acute respiratory syndrome and remaining. The SARS dot is still visible but smaller than MERS dot. It means that the publications are mentioned MERS more frequently. The green cluster consists of chlorocebus aethiops, cell line, antibodies, viral vaccines, etc. The blue cluster consists of infectious bronchitis virus, poultry diseases, phylogeny, genome and several remaining. And the smallest yellow cluster consists of swine, swine diseases, diarrhea, and feces.

Table 1. Most productive country

No	Pre-SARS			SARS			MERS			COVID-19		
	Countries	Σ	%	Countries	Σ	%	Countries	Σ	%	Countries	Σ	%
1	United States	652	14.07%	United States	1388	17.75%	China	878	16.80%	United States	10287	35.78%
2	Canada	228	4.92%	China	1221	15.61%	United States	777	14.87%	United Kingdom	6944	13.26%
3	Japan	196	4.23%	Hong kong	627	8.02%	Saudi Arabia	224	4.29%	China	1838	6.39%
4	Netherlands	148	3.19%	Canada	371	4.74%	Korea	215	4.11%	Italy	1271	4.42%
5	Germany	119	2.57%	Taiwan	348	4.45%	Germany	147	2.81%	Switzerland	854	2.97%
6	France	111	2.40%	Singapore	251	3.21%	Japan	144	2.75%	Brazil	763	2.65%
7	Spain	81	1.75%	Japan	240	3.07%	Netherlands	125	2.39%	India	690	2.40%
8	United Kingdom	39	0.84%	Netherlands	193	2.47%	Hong Kong	120	2.30%	Germany	508	1.77%
9	Belgium	34	0.73%	Germany	190	2.43%	Canada	107	2.05%	Canada	445	1.55%
10	Switzerland	30	0.65%	Italy	146	1.87%	Taiwan	107	2.05%	Netherlands	445	1.55%
	Total countries: 48 4659			Total countries: 75 8066			Total countries: 86 5568			Total countries: 58 28752		

Table 2. Most active journal

No	pre-SARS				SARS		MERS			COVID-19		
	Journal Title	Σ	%	Journal Title	Σ	%	Journal Title	Σ	%	Journal Title	Σ	%
1	Advances in experimental medicine and biology	551	11.82%	Journal of virology	507	6.48%	Journal of virology	245	4.69%	Journal of medical virology	449	1.56%
2	Journal of virology	365	7.83%	Emerging infectious diseases	207	2.65%	PloS one	168	3.21%	The New England journal of medicine	326	1.13%
3	Virology	254	5.45%	Virology	146	1.87%	Emerging infectious diseases	140	2.68%	Lancet (London, England)	324	1.13%
4	Avian diseases	218	4.68%	Lancet (London, England)	109	1.39%	Viruses	119	2.28%	Nature	290	1.01%
5	The Journal of general virology	196	4.20%	Advances in experimental medicine and biology	104	1.33%	Virus research	106	2.03%	BMJ (Clinical research ed.)	268	0.93%
6	American journal of veterinary research	177	3.80%	Journal of virological methods	90	1.15%	Veterinary microbiology	105	2.01%	International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases	231	0.80%
7	Archives of virology	174	3.73%	Biochemical and biophysical research communications	80	1.02%	Virology	99	1.89%	International journal of environmental research and public health	214	0.74%
8	The Veterinary record	119	2.55%	The Journal of general virology	75	0.96%	Archives of virology	97	1.86%	Science (New York, N.Y.)	207	0.72%
9	Laboratory animal science	71	1.52%	Virus research	74	0.95%	Scientific reports	75	1.43%	The Journal of infection	199	0.69%
10	Veterinary microbiology	68	1.46%	Hong Kong medical journal = Xianggang yi xue za zhi	73	0.93%	The Journal of general virology	71	1.36%	Head & neck	189	0.66%
	Total journals 520			Total journals 1424			Total journals 985			Total journals 3083		

The social network analysis (Figure 6) shows the pattern of article keywords in 2020. Instead of using MeSH, we used keyword to visualized data considering the recent article might not yet indexed into MeSH. The biggest cluster portrayed the most used keyword among the published articles, which is 'coronavirus infection'.

The most mentioned MeSH (Figure 7) in all time publications of coronavirus consecutively are immunology, genetics, virology, epidemiology, metabolism, animals, humans, veterinary, isolation purification and chemistry. It can indirectly indicate the theme of publications. Total MeSH term indexed in all time publication is 7 494 terms and mentioned 352 508 times. There is no new MeSH term indexed after the outbreak of COVID-19. However, there were 4.236 new MeSH terms appearing in the SARS period. 'Bibliometrics' MeSH terms appeared 5 times in the SARS period and 4 times in the MERS period, which means that bibliometrics study for coronavirus is still limited.

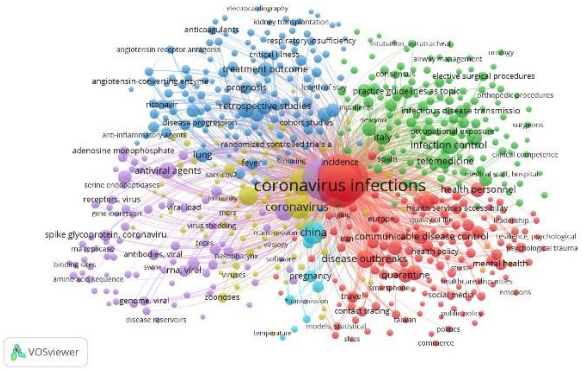

Fig. 6. Social Network Analysis for coronavirus research topic in COVID-19 period

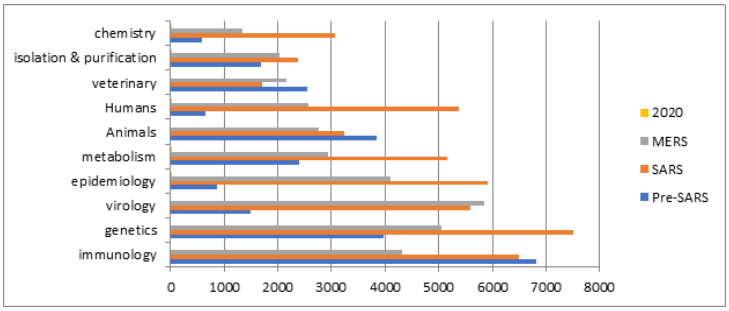

Fig. 7. Most mentioned MeSH

4 Discussions

All-time analysis of coronavirus publication trends is increasing globally. Especially from the first year of publication to 2002. At this period coronavirus was not a deathly disease, mostly found in the animals and if it was found in the humans it only caused mild symptoms. The first outbreak of coronavirus was SARS in 2003, the second outbreak was MERS in 2012 and recently COVID-19 outbreak. It is known that every outbreak was a new type of coronavirus that had not been identified before. Furthermore, all of those outbreaks are coming from animals to humans infection then continue spreading as a humans-humans transmission. After the outbreak period, the publication growth rate is increasing significantly. Most publications regarding coronavirus and coronavirus infection are dominated by the USA and followed by China. Generally, all publications from 1995 to 2016 extracted from PubMed also showed that the USA and China are two leading countries for health science publications. Research publications from the United States (USA) showed a steady rise and a doubling of publications in the 20-year review period[11]. Aside of the domination from those two countris, in each period, other countries where the first case emerged and most affected tend to have higher number of publication. From the MeSH term analysis, keywords related to coronavirus infection appeared mostly after the outbreak in 2002. Before the outbreak, MeSH term keywords that mostly appeared were related to biomolecular and virology of the coronavirus. After the first outbreak, the keyword trend is shifting to diseases-related. The top ten most mentioned MeSH keywords that are close related to vaccine are immunology, virology and isolation&purification. Term 'vaccine' itself is not in top ten list. But it only reflect the bibliometric until before COVID-19 since it may not already indexed in MeSH.

5 Conclusion

The growth of coronavirus publications worldwide is improving significantly, especially in most infected countries such as United States, China and Italy. Evolution of research topic are changing overtime. After the first outbreak, the keyword trend is shifting to diseases-related. Specific terms for 'vaccine' is not appeared frequently, but another terms related to vaccine appeared quite often. This term appearance in COVID-19 period is not seen yet due to limitation on MeSH analysis, since many recent publications in 2020 have not been indexed in MeSH. We assume in the coming years, publications related vaccine will appear in bibliometric study. More studies about coronavirus vaccines and virus evolution are highly suggested.

Study found that using PubMed database can prove the growth of research related to coronavirus. Massive growth in 2020 can be seen from total of 61 % of all coronavirus publications were conduct only in this year. The fast growth of educational sharing is the impact of the great technology that now we have. There are lot of pre-print journals available make it easier for scientist to share their study. However, this can raise questions about the quality of the article due to the brevity of the research and review processes.

5.1 Limitation

This study using only retrieved bibliography data from Pubmed database. Currently Pubmed has new feature called LitCovid, a curated literature hub for tracking up-to-date sciencetific information about COVID-19 which we did not analyse in this study.

References

1. J. Cui, F. Li, Z. Shi, Nature Reviews Microbiology, **17**,3,181–192 (2018).
2. A.J. Rodriguez-Morales, Infez. Med., **28**, 3–5 (2020).
3. J. Nicholls, X. Dong, G. Jiang, M. Peiris, Respirology, **8**,s1, S6–S8 (2003).
4. I. Mackay, K. Arden, MERS coronavirus: diagnostics, epidemiology and transmission. Virology Journal, **12**,1, (2015).
5. Worldometer. [online] Available at: https://www.worldometers.info# (2020). [Accessed 30 Jul. 2020].
6. Pubmed. Coronavirus. (2020). NCBI NLM. [online] Available at: https://www.ncbi.nlm.nih.gov/mesh/68017934# [Accessed 15 Feb. 2020)

7. Cdc.gov. CDC SARS Response Timeline, CDC. (2020). [online] Available at: https://www.cdc.gov/about/history/sars/timeline.htm# [Accessed 9 Feb. 2020].
8. American Society for Microbiology. About JVI. ASM (2020). [online] Available at: https://jvi.asm.org/content/about-jvi [Accessed 24 Feb. 2020]
9. Pubmed. Coronavirus Infections. NCBI NLM. (2020). [online] Available at: https://www.ncbi.nlm.nih.gov/mesh/68018352# [Acccessed 15 Feb. 2020]
10. A.R. AlRyalat, L.W. Malkawi, S.M. Momani. J. Vis. Exp, **152**,1–12 (2019).
11. P. Fontelo, F. Liu, Systematic reviews, **7**,1, 147 (2018). https://doi.org/10.1186/s13643-018-0819-1