Are we there yet? Analyzing scientific research related to COVID-19 drug repurposing

Namu Park
Department of Digital Analytics, Yonsei University, Seoul, South Korea
Hyeyoung Ryu
The Information School, University of Washington, Seattle, WA, United States
Ying Ding
School of Information, University of Texas, Austin, TX, United States
Qi Yu
School of Management, Shanxi Medical University, Taiyuan, Shanxi, China
Yi Bu
Department of Information Management, Peking University, Beijing, China
Qi Wang
School of Management, Shanxi Medical University, Taiyuan, Shanxi, China
Jeremy J Yang
Department of Internal Medicine, School of Medicine, University of New Mexico, Albuquerque, NM, United States
Min Song (✉ min.song@yonsei.ac.kr)
Department of Library and Information Science, Yonsei University, Seoul, South Korea
https://orcid.org/0000-0003-3255-1600

Research Article

Keywords: Text mining, COVID-19, Bibliometrics, Drug repurposing, Biomedical informatics

Posted Date: September 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-80893/v1

Abstract

Drug repurposing may be a pivotal means of fulfilling urgent needs for treatment of the novel coronavirus disease 2019 (COVID-19), but current studies on drug repurposing for COVID-19 seem to show a lack of consensus in their drug candidate focus. Using bibliometric methods in a non-expert perspective, in a review of 34 published articles on the COVID-19 and drug-repurposing, we investigated obvious and less obvious points of consensus on drug candidates. To establish these two types of consensus, we first implemented document clustering. Within a set of five clustered papers, we established an obvious consensus, relying solely on the occurrence of entities by using term frequency and inverse document frequency and a comparison of mentioned drugs, finding that remdesivir and chloroquine were discussed with a certain degree of agreement. For the less obvious consensus, we created a drug entity co-occurrence network to establish low-high centrality combinations to probe the crucial drugs found in article clustering that are not plainly apparent through the mere counting of the occurrence of drug entities occurrences. Lopinavir emerged as having possibly potent effects in spite of underuse, while the mainstream of studies focus more on drugs such as chloroquine that enjoy explicit consent. Using an entitymetrics perspective, we expect that our research will support investigations of drug repurposing, expediting the process of establishing treatment for COVID-19.

Introduction

In the context of the current novel coronavirus disease 2019 (COVID-19) pandemic and the pressing need for adequate treatment of it, it can be instructive to revisit the history of treatment discoveries during past pandemics. Ebola virus disease (EVD) is a complex infectious disease with a much higher mortality rate than COVID-19, killing an overall average of 50% and up to 90% of those infected (World Health Organization 2020). The first outbreak of EVD was in Sudan and Congo in 1976, in a village near the Ebola River, and dozens of intermittent outbreaks followed throughout sub-Saharan Africa over the last 40 years. Between 2017 and 2018, the severe outbreaks of the disease in Congo led World Health Organization to declare EVD a world health emergency (Feldmann and Geisbert 2011). The first EVD vaccine, Ervebo, was only approved in the United States in December 2019 (FDA 2019), more than 40 years after the first outbreak, and still no antiviral drug has been approved by the FDA to treat it. More than 140 drugs have been repurposed for EVD, and there are reports of *in vitro* potency and *in vivo* effectiveness in animal models and clinical trials with EVD patients (Bai and Hsu 2019). Remdesivir, originally developed for treating hepatitis C, has been tested on EVD, but it failed in a recent clinical trial (Mulangu et al. 2019). However, remdesivir has been proven to effectively shorten recovery times for COVID-19 patients (Beigel et al. 2020). Recently, the FDA authorized the emergency use of remdesivir to treat severe COVID-19 (Dolin and Hirsch 2020). Unlike in earlier pandemics, an unprecedented worldwide scientific workforce has been brought together from universities, research labs, pharmaceutical companies, and organizations, exhibiting a laser focus on finding novel treatment for COVID-19 through drugs or vaccines.

Drug repurposing seeks new uses of existing drugs and is known to effectively shorten treatment development times (Pushpakom et al. 2019). Due to the urgent need to find a cure, drug repurposing has become a mainstream goal for COVID-19 research. It is thus crucial for us to understand the current status of research in this area as relates to COVID-19 and in particular to establish whether scientists have reached any consensus on a list of candidate drugs with potential for COVID-19.

We retrieved 34 papers on COVID-19 literature from Kaggle, which is the world's largest data science community, with the criterion that each used scientific method to propose a list of drug-repurposing candidates for COVID-19. By analyzing the resulting list of recommended candidate drugs for COVID-19 using bibliometric methods, we demonstrated a lack of consensus among drug repurposing literatures. Specifically, we observed two types of consensus: obvious and indistinct consensus.

Our research constitutes an early and objective investigation of scientific consensus without direct input from active researchers, relying solely on methods of data science for ease and speed of research. In short, this study outlined the key scientific outputs of drug repurposing for COVID-19 through the use of entitymetrics (Ding et al. 2013) that studies how drug and disease entities affect knowledge transfer and impact different fields or subjects.

Related Works

Drug Repositioning

Drug repositioning, also called drug repurposing or drug reprofiling, is a common term in the drug discovery context where new therapeutic opportunities for existing drugs are sought (Doan et al. 2011). This method is attracting more attention at present due to the advent of COVID-19, but it has been in widespread use since the beginning of the twenty-first century. In 2006, a study of drug repositioning was conducted that used a computational approach (Li et al. 2006), and 3 years later, *in silico* compound profiling was used to broaden the knowledge map of drug repurposing (Dubus et al. 2009). After the onset of the global pandemic in 2020, the drug-repositioning literature grew rapidly. The main reason for this is that since SARS-CoV-2, which causes COVID-19, is a type of a coronavirus, researchers expected that drug repurposing was a plausible method for finding treatment (Altay et al. 2020).

Consensus Check

There has been little treatment of the scientific consensus on COVID-19 treatment or any other subject because pursuing such research is time consuming and requires a certain level of domain knowledge. The most best-known research on scientific consensus, that of Cook et al. (2013), investigated the consensus on anthropogenic global warming . Its goal was to identify the evolution of this consensus, and it indicated that an increasing number of publications accepted the consensus. As a part of the citizen science project, volunteers with domain knowledge manually rated each publication's level of endorsement, adopting a criterion set by the authors. Nearly 25,000 abstracts were rated in this experiment. However, because all of the ratings were assigned manually, the work was time consuming,

and the different volunteers had different backgrounds and perspectives, which likely affected their ratings. This lack could lead to doubt regarding the results and may damage the reliability of the conclusions. Therefore, we suggest a streamlined method of observing consensus without resorting to potentially biased opinions from experts.

Proposed Approach

We selected papers from PubMed that proposed a drug candidate for repurposing using the following search query:

drug candidate or repurpos* or reposition* or re-purpos* or re-position**

56 publications were founded, among which we removed those that did not propose any repurposed drugs in their results parts. As a result, 34 papers were left. Our study had two main steps: document clustering and consensus analysis. Document clustering is the process of finding the set of candidate publications for review. Using the candidate documents established in the first stage, we conducted two analytical stages to check the consensus exhibited among them. Conspicuous consensus analysis involves methods that rely solely on term frequency and inverse document frequency (TF-IDF) and a comparison of the mention of certain items. It checks obvious consensus among candidate documents. Indistinct consensus analysis, by contrast, targets implicit agreement among documents, using a method that assesses the occurrence of entities and different types of network centrality. Figure 1 illustrates the overall analytical process.

Document Clustering

First, we presumed that the abstract of an article is the summary of the entire document, and all essential information is always included in it. We extracted the abstracts of 34 papers and then plotted them in vector space.
The word-embedding model used in this experiment was BioCovidBERT developed by Tonneau (2020), a fine-tuned version of BioBERT, created by Lee et al. (2020). BioCovidBERT was trained on a preprocessed COVID-19 dataset accessible in the COVID-19 Open Research Dataset Challenge (Kaggle 2020). It was created with the help of several leading research groups in response to the COVID-19 pandemic, and over 181,000 coronavirus-related publications are included in the data of the COVID-19 Open Research Dataset Challenge. We assumed that BioCovidBERT, on this rich resource, could effectively represent each document as a vector while preserving their semantics, especially for words related to COVID-19. After document embedding, each paper was represented as a 1,024-dimensional vector. To check the similarity of the documents, we used Kernel Principal Component Analysis (KernelPCA), which can effectively reduce the dimensions of non-linear high-dimensional data. With KernelPCA, we visualized each document vector in a 2- and 3-dimensional space, and K-means clustering was used to elaborate groups of vectors. Among the derived clusters, we selected the one with the shortest inter-class distance and the longest intra-class distance, meaning that its components are self-similar, and the differences between

points from other clusters are relatively high. The components of the selected cluster are determined as candidate documents and considered for further analysis in the next phase.

Conspicuous Consensus Analysis

With the candidate documents, two experiments were conducted for further investigation. To begin with, we used the TF-IDF for each candidate document. The following equation is the formula for TF-IDF, where refers to the number of documents in the corpus:

$$TFIDF_{word,doc} = tf_{word,doc} * \log\left(\frac{N}{df_{word}}\right)$$

TF-IDF is a sclar value and is used to determine how important a word is in a document. TF gives the number of appearances of the target word in each document, and IDF represents whether it is used commonly or rarely for each document. For example, the word “the” appears commonly in most document, so is relatively high, but inverse document frequency is low. Therefore, a high TF-IDF value indiates that the target word is rarely used in other documents but frequently appears in the target document.

For a more complete analysis, we extracted lists of all drugs mentioned in each paper and checked whether there were intersections between the drugs. This was done using PubTator Entity Tagger (U.S. National Library of Medicine and National Center for Biotechnology Information 2020) and with the help of the COVID-19 Drug and Gene Set Library developed by the Mount Sinai Health System (Icahn School of Medicine at Mount Sinai 2020), and we double-checked whether the extracted entities are included in the drug set. A total of 147 drugs were mentioned in 34 papers.

Indistinct Consensus Analysis

We formed drug entity co-occurrence networks for the clustered papers by setting the nodes as entity instances and the edges as the number of co-occurrences between the entity instances, making an entity co-occurrence network. In this network, we calculated the degree, betweenness, and closeness centrality values for each node (i.e., entity instance). After the logarithmic values for degree and closeness centrality values and the square roots of the betweenness centrality values were calculated, we cut the values into terciles, labeled as low, medium, and high. Using the values in the low tercile for one centrality and those in the high tercile for another, we created a low–high centrality combination for each entity. This combination indicates hidden gems within the entity network that cannot be observed from a direct observation of the centrality combinations.

In Table 1, we list the features of each centrality combination for the low–high centrality combinations (Zhang and Luo 2017). The instances in each entity co-occurrence network in the low–high centrality combination table could affect the network in a non-obvious but pivotal way. However, it should be noted that not all instances that occur in the low–high centrality combination table have low-profile importance, and that those that do appear in one of the following four low–high centrality combination groups: (a)

low degree–high betweenness, (b) low degree–high closeness, (c) low closeness–high degree, and (d) low closeness–high betweenness.

Table 1 Low–high centrality combination features

	High Degree	High Betweenness Centrality	High Closeness Centrality
Low Degree	-	An individual's few ties are crucial for network flow.	An individual has ties to some active/important actors
Low Betweenness Centrality	An individual's connections are redundant and communication bypasses him/her	-	The network may hold many paths where an individual is near many actors, but so are many others.
Low Closeness Centrality	An individual is embedded in a cluster that is away from the rest of the network.	An individual can monopolize the ties of a few people to many others.	-

Results

Document Clustering

Figure 2 represents document embedding in a 2-dimensional space. At first glance, it is not easy to observe the existence of meaningful clusters or find a consensus among the 34 drug repurposing papers. To develop our investigation, we used the K-means clustering algorithm, an unsupervised machine learning apporach that is widely used for cluster detection. Our goal was to check potential clusters within this 2-dimensional space, where horizontal and vertical axes represent the principal components. In other words, we assumed that if there are to be similar points among some of the papers, they would form a cluster in vector space, in which the components may share elements. Before applying the K-means method, it was necessary to develop a process to derive the optimal number of clusters. Using the elbow method and silhouette scoring, we discovered that three is an optimal number for our clusters. The results of our clustering are given in Figure 3-a, where the red points form a relatively meaningful cluster, as the inter-class distance is longer there than in the other clusters, and intra-class distance is shorter. Therefore, we anticipated that documents corresponding to the red points (6, 7, 11, 18, 30, 32, and 33) have the highest probability of addressing issues common to all 34 papers. We also checked KernelPCA in 3-dimensional space to improve the validity of the candidate documents (Figure 3-b). We found that documents 7 and 11 had low similarity in 3-dimensional space, although they seemed were located close together in 2-dimensional space. Using these means, five documents (6, 18, 30, 32, and 33) were shortlisted for possible consensus publications.

Table 2 Part of the TF-IDF matrix for the candidate documents

	Paper 6	Paper 18	Paper 30	Paper 32	Paper 33
Receptor	0.00000	2.78645	0.00000	0.34831	0.34831
Hydroxychloroquine	0.00000	7.34265	1.22378	0.00000	40.38459
Plasma	13.46153	0.00000	0.00000	0.00000	0.00000
ACE2	0.00000	3.27324	0.00000	0.00000	2.45493
Chloroquine	0.00000	45.38988	2.32769	1.16384	29.67800
Remdesivir	0.00000	0.43532	0.87064	3.48254	0.87064
Protease	0.00000	0.30748	0.00000	5.22724	0.61497
Pneumonia	0.00000	0.00000	1.17260	0.39087	0.00000
Azithromycin	0.00000	0.00000	0.00000	0.00000	21.08615
Lopinavir	0.53063	0.00000	1.06126	1.59188	3.18377

After removing stop words, we identified the 1,374 most frequently used words from 34 papers to generate a TF-IDF matrix. Table 2 presents part of this matrix for the five candidate documents, and certain several characteristics of each paper can be observed. In Paper 6 (Da Silva 2020), entitled “Convalescent plasma: A possible treatment of COVID-19 in India,” convalescent plasma, which can be collected from an infected individual, is proposed as a treatment to COVID-19. The TF-IDF for plasma in Paper 6 is high, but it is not mentioned in other four documents. In Paper 18 (Devaux et al. 2020), it is reported that chloroquine interferes with ACE2, a protein on the surface of the cell that is known to be a pathway for coronavirus penetration. The TF-IDF values for chloroquine and ACE2 are relatively higher than its TF-IDF values from other documents. Thus, our TF-IDF matrix reflects each document’s main ideas quite well. Following this assumption, we calculated the cosine similarity for each TF-IDF vector from each publication (Fig. 4).

The average cosine similarity of TF-IDF vectors was 0.1556 when calculated with all 34 papers, whereas the similarity values for the five candidate papers were lower, with an average of 0.1133. The fact that all five candidate documents, which formed the most meaningful cluster in the clustering section, showed a lower level for TF-IDF vector similarity implies two things.

A first is that there is information loss when the dimensions of document embeddings are reduced. On this view, some common issues may appear to be shared in a 2- or 3-dimensional space, but there is no consensus in reality. A second is that, if there was no information loss, the cross sections among these

publications may be too trivial. To be specific, the words that contribute to the clustering may not be about the 1,374 core words we used to generate TF-IDF vectors. If some critical information were shared, shared vocabulary used to present that discovery would result in a high similarity in TF-IDF vectors. Overall, only eight drugs were mentioned more than twice: lopinavir, hydroxychloroquine, Arbidol, chloroquine, ritonavir, remdesivir, favipiravir, and ribavirin. Remdesivir and chloroquine were each mentioned in four of the five candidate documents, followed by lopinavir, ribavirin, and ritonavir, which appeared three times each. Thus, remdesivir and chloroquine appear to enjoy a certain consensus, but there is still no broad agreement, as these documents discuss too wide of a range of drugs. Figure 5, presenting the network visualization of Table 3, which reports the list of drugs mentioned in the five publications, buttresses our point. The self-loops in Figure 5 highlight drugs that appear only in a single document. For example, potassium, amiodarone, and azithromycin are only mentioned in 33. The multiple self-loops observed in Figure 5 and the multitude of edges labeled remdesivir or chloroquine reinforces our claim that no strong consensus exists, except for these two drugs.

Table 3 Mentioned drugs list in five candidate papers

Paper 6	Paper 18	Paper 30	Paper 32	Paper 33
Arbidol	Chloroquine	Chloroquine	Adenosine	Amiodarone
Darunavir	Chlorpromazine	Lopinavir	Arbidol	Azithromycin
Favipiravir	Doxycycline	Remdesivir	Chloroquine	Chloroquine
Lopinavir	Hydroxychloroquine	Ribavirin	Disulfiram	Hydroxychloroquine
Methylprednisolone	Quinine	Ritonavir	Favipiravir	Potassium
Ritonavir	Remdesivir	Teicoplanin	Galidesivir	Remdesivir
-	Ribavirin	-	Lopinavir	-
-	Tyrosine	-	Nitazoxanide	-
-	-	-	Oseltamivir	-
-	-	-	Remdesivir	-
-	-	-	Ribavirin	-
-	-	-	Ritonavir	-
-	-	-	Tenofovir	-

Indistinct Consensus Analysis

In the drug entity co-occurrence network for the five clustered papers, chloroquine and teicoplanin were found to have high degree centrality scores, but they also had low betweenness and closeness centrality scores, which meant that their connections were redundant: important communication simply bypassed

them, and they were embedded in a cluster that was distant from the rest of the network (Table 4). Lopinavir had few ties that were pivotal to network flow but had ties to other important drug instances in the network. This implies that lopinavir may be a hidden gem for the cluster formation of drug repositioning research in a high entropy situation.

Table 4 Drug entity co-occurrence network in low–high centrality combination for the five clustered papers

Drug	High Degree	High Betweenness Centrality	High Closeness Centrality
Low Degree	-	Lopinavir	Lopinavir
Low Betweenness Centrality	Chloroquine, Teicoplanin	-	-
Low Closeness Centrality	Chloroquine, Teicoplanin	-	-

Discussion

The data used to find conspicuous and indistinct consensus in treatments for COVID-19 were abstracts of COVID-19 drug repositioning papers published before April 15, 2020. To determine whether our predicted focusable drug was actually a target of COVID-19 drug repositioning research, we examined research trends before and after April 15, 2020, for the three drug instances found in low–high centrality combinations in the drug entity co-occurrence network. We investigated the number of research papers on each drug before and after April 15, 2020, on PubMed and calculated the rate of increase in the research by dividing the number of papers published after April 15, by the number of papers published before that date.

In Table 5, it is shown that all three drugs increased in the number of studies targeting them, but the higher research increase rate for chloroquine and teicoplanin should be noted. Although these two were not in the main part of the co-occurrence network, and important communication bypassed them, the number of studies conducted on these two were higher than those on lopinavir, which has few ties but ones that are crucial the network flow. The difference in research increase rate between lopinavir and chloroquine may not be significantly different, but a comparison of the absolute number of studies conducted shows that chloroquine was researched approximately 4.517 times more often than lopinavir, which indicates a notably low research focus on the latter. Thus, since the inconspicuous consensus analysis did not receive sufficient attention, probing this drug may help develop innovatory drug repositioning results.

Table 5 Lopinavir, chloroquine, and teicoplanin research trends before and after April 15, 2020

	Lopinavir	Chloroquine	Teicoplanin
Before April 15, 2020	11	47	1
After April 15, 2020	60	271	7
Research Increase Rate (# of Papers After April 15, 2020/# of Papers Before April 15, 2020)	5.455	5.766	7.0

Conclusion

In this research, we examined the consensus among publications on COVID-19 drug repurposing using two data science approaches. First, we derived several conspicuous features using document clustering, a TF-IDF matrix, and co-occurrence of drug entities. Among five papers that formed a cluster, we found that the cosine similarity for each publication's TF-IDF was low, and that only eight drugs were mentioned more than twice in this group. Remdesivir and chloroquine appeared in four out of the five clustered papers, which implied the conclusion that these drugs are the subjects of a certain agreement. Second, using co-occurrence network analysis, we conducted an additional experiment to determine an indistinct consensus. Lopinavir was the only drug to show high betweenness centrality, high closeness centrality, and low degree, which implies that it may have potential for repurposing. However, the rate of increase in research on chloroquine is still higher than that of lopinavir, which indicates that most publications are still concentrating on chloroquine.

In future work, we hope to extend our search for drug repurposing publications on the rapidly growing area of COVID-19. Applying these methods, we can continue to monitor what overlaps will emerge from the proposed repurposed drug candidates for COVID-19 from novel studies. These overlaps may eventually lead to a research consensus on an established list of repurposed drug candidates for COVID-19.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.
Availability of data and material The data that support the findings of this study are available from the corresponding author upon reasonable request
Code availability All code for data cleaning and analysis associated with the current submission is available at https://github.com/namupark/COVID_19_Consensus_Analysis

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1A2C2002577). Qi Yu gratefully acknowledges financial supports from National Natural Science Foundation of China (Grant Number: 71573162) and the Shanxi Scholarship Council of China (Grant Number: HGKY2019057).

References

Altay, O., Mohammadi, E., Lam, S., Turkez, H., Boren, J., Nielsen, J., et al. (2020). Current status of COVID-19 therapies and drug repositioning applications. Iscience, 101303.

Bai, J. P. F., & Hsu, C. W. (2019). Drug repurposing for Ebola virus disease: Principles of consideration and the animal rule. *Journal of Pharmaceutical Sciences*, *108*(2), 798-806.

Beigel, J., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., et al. (2020). Remdesivir for the treatment of COVID-19: Preliminary report. *New England Journal of Medicine*, https://www.doi.org/10.1056/NEJMoa2007764

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., et al. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, *8*(2), 024024.

Da Silva, J. A. T. (2020). Convalescent plasma: A possible treatment of COVID-19 in India. Medical journal, Armed Forces India, 76(2), 236–237. Advance online publication. https://doi.org/10.1016/j.mjafi.2020.04.006

Devaux, C. A., Rolain, J. M., Colson, P., & Raoult, D. (2020). New insights on the antiviral effects of chloroquine against coronavirus: what to expect for COVID-19?. International journal of antimicrobial agents, 105938.

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., et al. (2013). Entitymetrics: Measuring the impact of entities. *PLoS ONE*, *8*(8), 1–14.

Doan, T. L., Pollastri, M., Walters, M. A., & Georg, G. I. (2011). The future of drug repositioning: Old drugs, new opportunities. In *Annual reports in medicinal chemistry* (Vol. 46, pp. 385-401). Academic Press. https://doi.org/10.1016/B978-0-12-386009-5.00004-7

Dolin, R., & Hirsch, M.S. (2020). Remdesivir - An important first step. New England Journal of Medicine, https://www.doi.org/10.1056/NEJMe2018715

Dubus, E., Ijjaali, I., Barberan, O., & Petitet, F. (2009). Drug repositioning using in silico compound profiling. *Future Medicinal Chemistry*, *1*(9), 1723-1736.

FDA. (2019). First FDA-approved vaccine for the prevention of Ebola virus disease, marking a critical milestone in public health preparedness and response. https://www.fda.gov/news-events/press-announcements/first-fda-approved-vaccine-prevention-ebola-virus-disease-marking-critical-milestone-public-health. Accessed 2 June 2020.

Feldmann, H., & Geisbert, T. W. (2011). Ebola haemorrhagic fever. *Lancet*, *377*(9768), 849-862.

Icahn School of Medicine at Mount Sinai. (2020). The COVID-19 Drug and Gene Set Library. https://amp.pharm.mssm.edu/covid19. (Accessed 3 June 2020.

Kaggle. (2020). COVID-19 Open Research Dataset Challenge (CORD-19) [Data file]. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed 2 June 2020.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Li, Y. Y., An, J., & Jones, S. J. (2006). A large-scale computational approach to drug repositioning. *Genome Informatics*, *17*(2), 239-247.

Mulangu, S., Dodd, L.E., Davey, Jr. R. T., Mbaya, O. T., Proschan, M., Mukadi, D., Manzo, M. L., et al. (2019). A randomized, controlled trial of Ebola virus disease therapeutics. *New England Journal of Medicine*, *381*, 2293-2303.

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery*, *18*(1), 41-58.

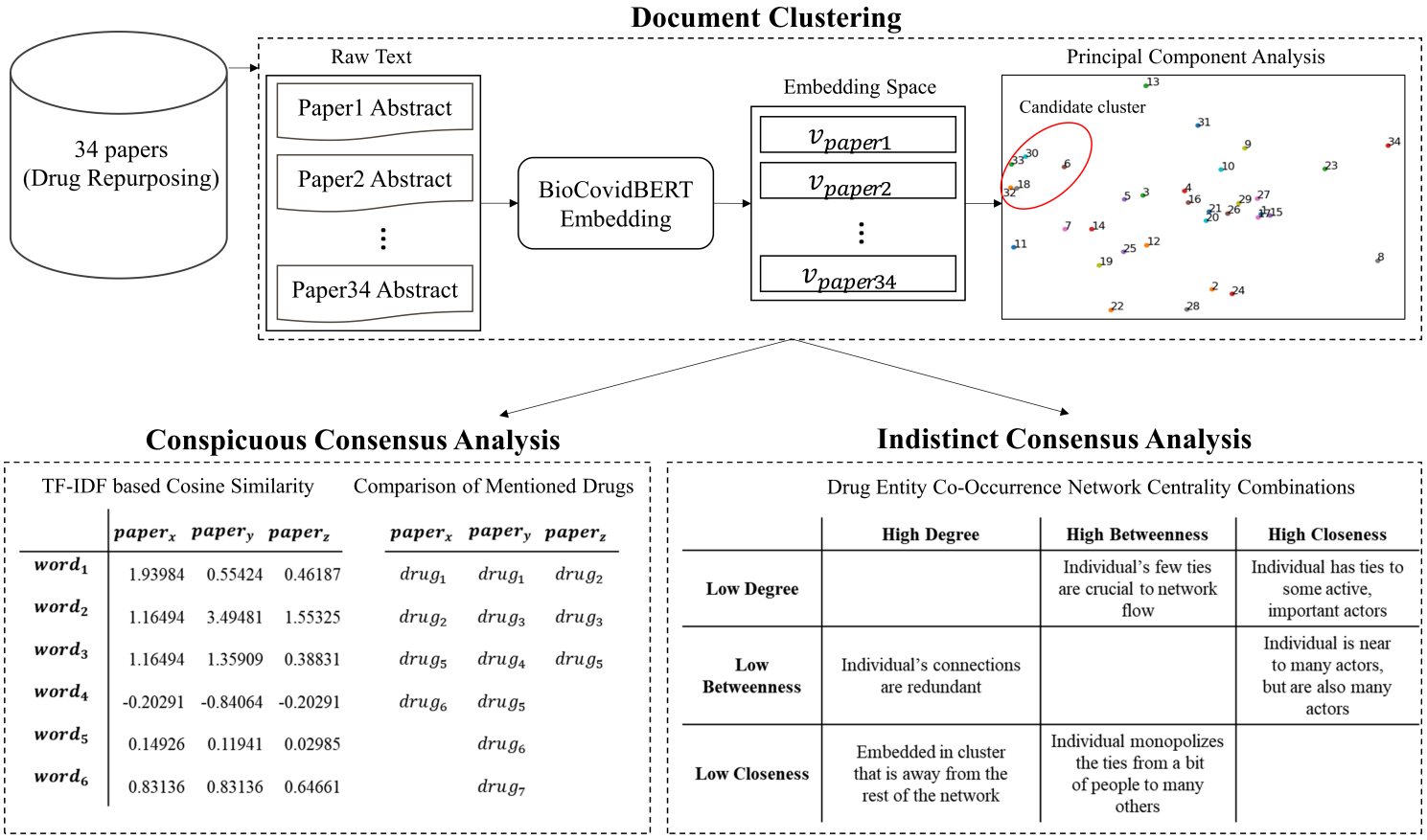
Tonneau, M. (2020). Covid-BERTs. https://github.com/manueltonneau/covid-berts. Accessed 2 June 2020.

U.S. National Library of Medicine, National Center for Biotechnology Information. (2020). Pubtator Central. https://www.ncbi.nlm.nih.gov/research/pubtator/. Accessed 3 June 2020.

World Health Organization (2020). Ebola virus disease: Key facts. https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease. Accessed 2 June 2020.

Zhang, J., & Luo, Y. (2017). Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017)*. Atlantis Press.

Figures



TF-IDF based Cosine Similarity

	$paper_x$	$paper_y$	$paper_z$
$word_1$	1.93984	0.55424	0.46187
$word_2$	1.16494	3.49481	1.55325
$word_3$	1.16494	1.35909	0.38831
$word_4$	-0.20291	-0.84064	-0.20291
$word_5$	0.14926	0.11941	0.02985
$word_6$	0.83136	0.83136	0.64661

Comparison of Mentioned Drugs

$paper_x$	$paper_y$	$paper_z$
$drug_1$	$drug_1$	$drug_2$
$drug_2$	$drug_3$	$drug_3$
$drug_5$	$drug_4$	$drug_5$
$drug_6$	$drug_5$	
	$drug_6$	
	$drug_7$	

Drug Entity Co-Occurrence Network Centrality Combinations

	High Degree	High Betweenness	High Closeness
Low Degree		Individual's few ties are crucial to network flow	Individual has ties to some active, important actors
Low Betweenness	Individual's connections are redundant		Individual is near to many actors, but are also many actors
Low Closeness	Embedded in cluster that is away from the rest of the network	Individual monopolizes the ties from a bit of people to many others	

Figure 1

Overview of the adopted approach

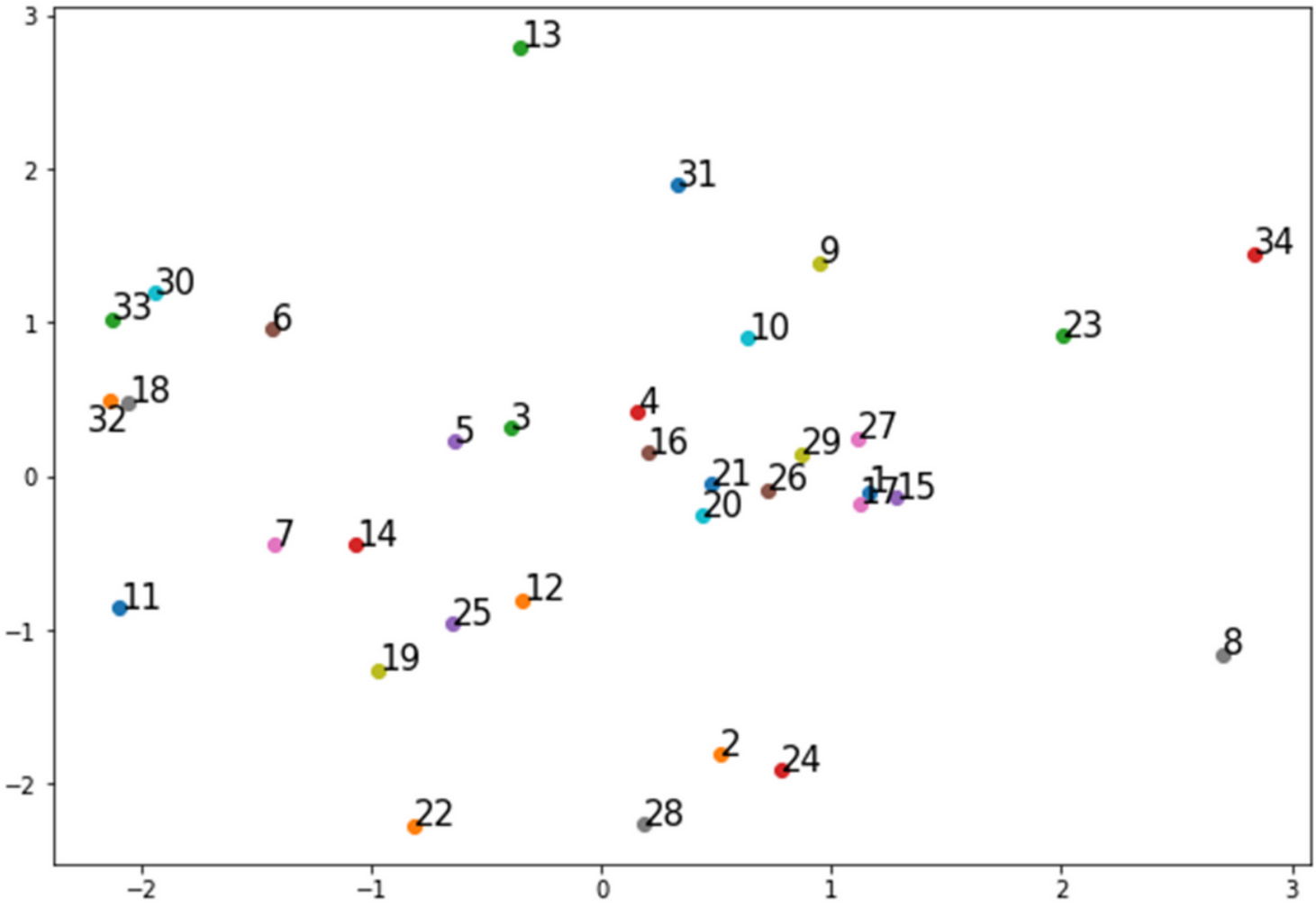


Figure 2

BioCovidBERT embedding of 34 papers projected in a 2-D space using Kernel PCA

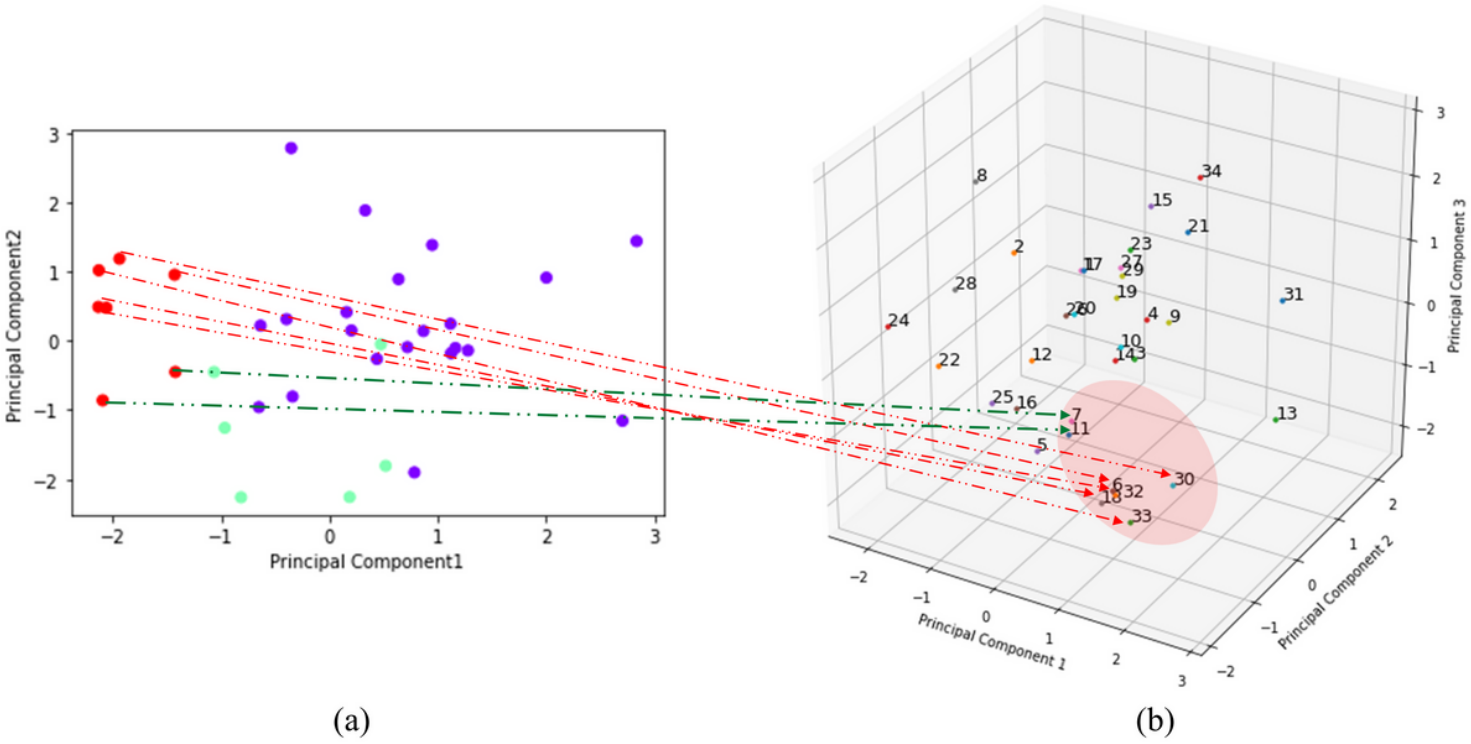
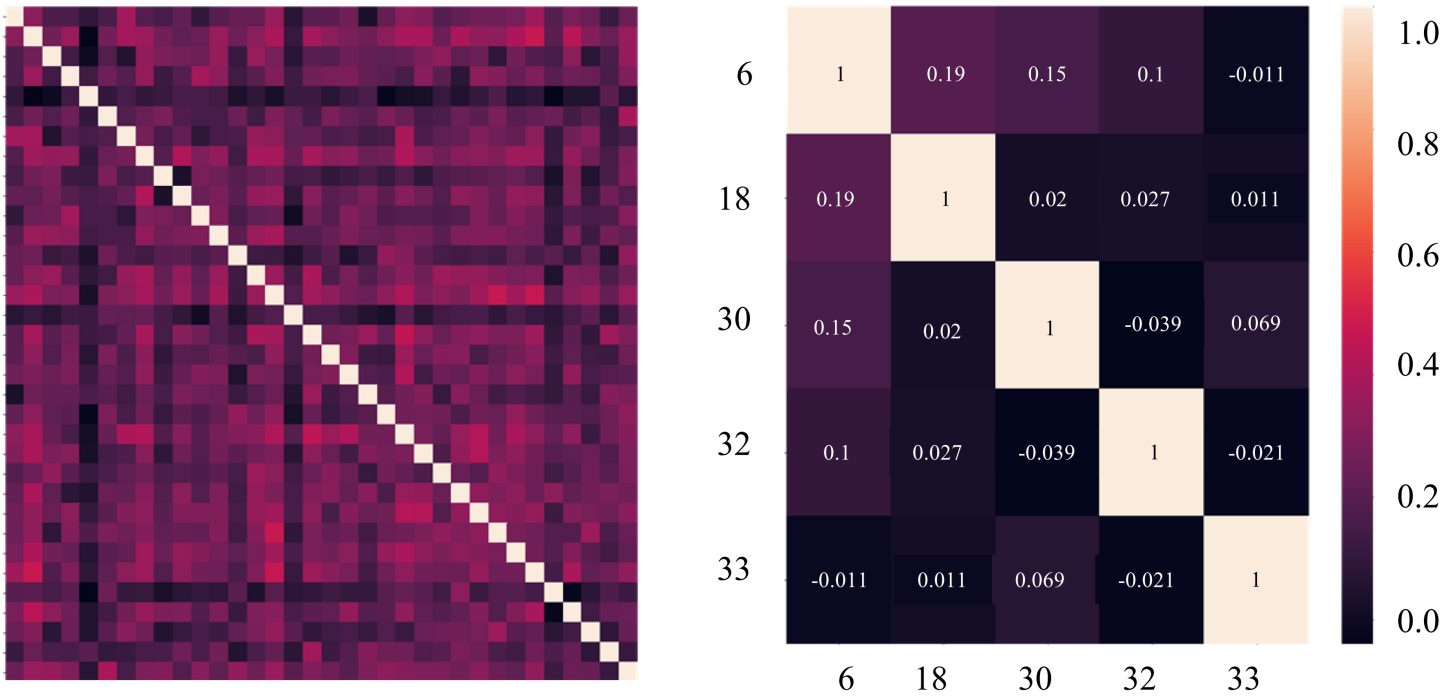


Figure 3

Projection of document embedding in 2D, 3D space



(a) Cosine similarity of TF-IDF vectors from all 34 papers

(b) Cosine similarity of TF-IDF vectors from 5 candidate papers

Figure 4

Heatmap of cosine similarity based on the TF-IDF vectors

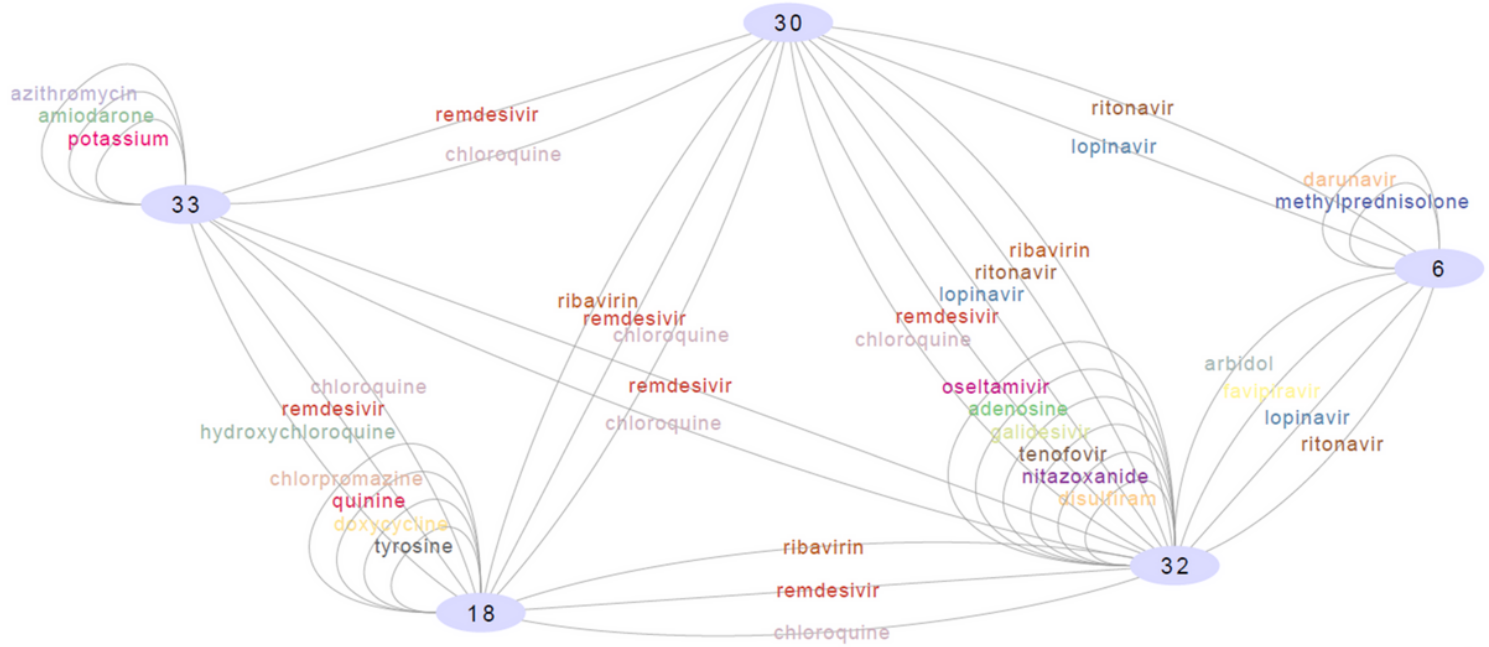


Figure 5

Multigraph network illustration of the mentioned drugs in the five candidate papers

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx
- All34andSelectedPapersEntityNetworkLowMediumHighCentralityTableUpdated0709.py
- litcovid34papersabstractupdated.ipynb
- docdruglist.xlsx
- clustereddrugnetwork.ipynb
- drugnetwork.ipynb
- rawtextpreprocess.ipynb
- fiveclusteredpapersdrugnetwork.cys
- cosinesimtfidfupdated.ipynb