Journal of Intellectual Disability Research doi: 10.1111/jir.12730

VOLUME 64 PART 7 pp 475–481 JULY 2020

Brief Report

Coronaviruses and people with intellectual disability: an exploratory data analysis

J. Tummers,[1,4] C. Catal,[2] H. Tobi,[3] B. Tekinerdogan[1] & G. Leusink[4]

1 *Information Technology Group, Wageningen University & Research, Wageningen, The Netherlands*
2 *Department of Computer Engineering, Bahçeşehir University, Istanbul, Turkey*
3 *Biometris, Wageningen University & Research, Wageningen, The Netherlands*
4 *Department of Primary and Community Care, Radboud University Medical Center, Nijmegen, The Netherlands*

Abstract

Background Corona virus disease 2019 (COVID-19) has been announced as a new coronavirus disease by the World Health Organization. At the time of writing this article (April 2020), the world is drastically influenced by the COVID-19. Recently, the COVID-19 Open Research Dataset (CORD-19) was published. For researchers on ID such as ourselves, it is of key interest to learn whether this open research dataset may be used to investigate the virus and its consequences for people with an ID.
Methods From CORD-19, we identified full-text articles containing terms related to the ID care and applied a text mining technique, specifically the term frequency–inverse document frequency analysis in combination with *K*-means clustering.
Results Two hundred fifty-nine articles contained one or more of our specified terms related to ID. We were able to cluster these articles related to ID into five clusters on different topics, namely: mental health, viral diseases, diagnoses and treatments, maternal care and paediatrics, and genetics.
Conclusion The CORD-19 open research dataset consists of valuable information about not only COVID-19 disease but also ID and the relationship between them. We suggest researchers investigate literature-based discovery approaches on the CORD-19 and develop a new dataset that addresses the intersection of these two fields for further research.

Keywords coronavirus, COVID-19, intellectual disability, machine learning, text mining

Background

At the time of writing this article (April 2020), the world that we live in is drastically influenced by the corona virus disease 2019 (COVID-19), which is a new coronavirus. The first reports on the COVID-19 came in late December 2019 from the Wuhan province in China (World Health Organization 2020b). Since 11 March 2020, the COVID-19 is officially a pandemic (World Health Organization 2020d). At the time of writing this article, there are more than 900 000 confirmed cases and over 45 000 deaths, according to the World Health Organization (World Health Organization 2020a).

Correspondence: Mr Joep Tummers, Information Technology Group, Wageningen University & Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands (e-mail: joep.tummers@wur.nl).

Corona virus disease 2019 is spread by human-to-human transmission via small droplets or direct contact (Lai *et al.* 2020). People with an ID might have a higher risk of getting infected by the virus than others for two reasons: they often live close to one another in care facilities, and many rely on physical contact with caregivers for their daily life activities. People of all ages with an ID have a very high risk of early death due to respiratory infections (Glover *et al.* 2017), which makes research into coronaviruses that cause upper respiratory illnesses (e.g. Middle East respiratory syndrome and severe acute respiratory syndrome) (McIntosh & Perlman 2015) pivotal.

To help mitigate the current public health crisis, the COVID-19 Open Research Dataset (CORD-19) has been initiated (CORD-19 2020). In recent years, the large number of existing scientific publications makes it difficult for researchers to identify relevant information from this huge corpus of papers (Aljaber *et al.* 2010). This is even more of an issue nowadays, in times of the corona pandemic. The articles in CORD-19 are fragmented and scattered over multiple topics, and thousands of articles are added on a weekly basis. The clustering of articles into groups with similar topics allows to map commonalities and unexplored issues and helps to efficiently pave the way for further research. Cluster analysis, as used in text mining, helps researchers and other users of these bibliographic databases to obtain a quick overview of the topics (Amador Penichet *et al.* 2018). Because of the clusters, researchers can more effectively decide what topics are sufficiently covered to deserve a systematic review. Furthermore, these clusters may help policymakers and researchers to identify missing topics in research on ID and COVID-19. Currently, it is unknown whether the CORD-19 can be used to investigate the virus and its consequences for people with an ID. More information on the CORD-19 set can be found in the Methods section.

For researchers on, ID, it is of key interest to know what is already in the literature on people with an ID and coronaviruses. With the help of an exploratory analysis using a text mining technique on the CORD-19 set, we want to investigate whether the CORD-19 may be useful for research on COVID-19 in people with an ID. We aimed to find clusters of articles that help researchers to identify interesting research topics on coronaviruses and people with an ID. The presented research and the methods shared not only yield clusters of articles but also can help other researchers when they consider pursuing a systematic literature review, and policymakers and researchers to identify research gaps.

The next sections of this paper are organised as follows: the Methods section presents the dataset and the tools that were used to analyse this dataset. The paper continues with the results of the exploratory analysis and ends with the discussion.

Methods

In this section, we present a brief description of the dataset CORD-19; subsequently, we explain how we identified articles about people with an ID and the text mining method that we used in this exploratory analysis.

Dataset

The CORD-19 set has over 44 000 scholarly articles, including over 29 000 with full text (CORD-19 2020). These articles are all about the COVID-19 and the family of coronaviruses. From the full-text articles, 260 contained the word 'COVID-19'. New articles are added on a weekly basis (White House Office of Science and Technology Policy 2020). All articles were retrieved with a query written in English from the following sources: PubMed (The United States National Library of Medicine at the National Institutes of Health 2020), a corpus maintained by the World Health Organization (World Health Organization 2020c), bioRxiv (Cold Spring Harbor Laboratory 2020a) and medRxiv (Cold Spring Harbor Laboratory 2020b). For this study, we used the dataset of 23 March 2020. The CORD-19 dataset is put together by the Allen Institute for AI, Chan Zuckerberg Initiative, Georgetown University's Center for Security and Emerging Technology, Microsoft, and the National Library of Medicine at the National Institutes of Health. Researchers were asked to apply text and data mining tools on this dataset to develop new insights into the COVID-19 via the Kaggle platform, which is a machine learning and data science community owned by Google Cloud (Kaggle 2020).

Identification of people with an intellectual disability

From the 29 000 articles with a full text, we aimed to identify a subset of articles that mention people with an ID. We decided to search for both synonyms of 'ID' and specific syndromes. We came up with a list of syndromes based on the website of the Dutch Expertise centre intellectual disabilities (Expertisepunt verstandelijke beperkingen 2020) and added syndromes based on our own domain knowledge. By combining synonyms of ID and syndromes, we searched for the following terms in the full text of each article (NB: not case sensitive):

> intellectual disab*, learning disab*, mental retardation, cognitive disab*, mental disab*, down syndrome, fragile x, prader willi, williams syndrome, fetal alcohol spectrum disorder, rett syndrome, velo-cardio-facial syndrome, angelman syndrome, tuberous sclerosis complex, cornelia de lange syndrome

The numbers of articles identified with this search are presented in Table 1.

Data analysis

Text mining is: '…. automatically extracting information from different written resources' (Hearst 2003, p. 1). In our case, we started with the term frequency–inverse document frequency (TF-IDF) algorithm to calculate the importance of each word for each article in the ID subset compared with the importance of that word in the rest of the CORD-19. Then, the outcomes of the TF-IDF were fed into the *K*-means clustering algorithm to cluster the articles. Articles from different clusters can have some important words in common; therefore, there will most probably be some overlap in the top words between the clusters.

For this data analysis, we have made a Kaggle notebook, which is available via the following link: https://www.kaggle.com/joeptummers/covid-2019-id-paper-version. Python version 3.6.6 (Python Software Foundation 2020) was used as the programming language in combination with the pandas package for data analysis and manipulation (pandas Community 2020) and the scikit-learn package for the (machine learning) algorithms (scikit-learn Community 2020).

For this exploratory data analysis, we started with the TF-IDF statistical measure, which is widely used in text mining and information retrieval. This statistical measure reflects how important a word is to a document relative to a larger set of documents (i.e. the corpus) (Rajaraman & Ullman 2011). The general formula for the TF-IDF is as follows:

$$w_{i,j} = tf_{i,j} * \log(N/df_i), \tag{1}$$

where $w_{i,j}$ is the TF-IDF characteristic for a term i in document j from the subset, that is, the articles identified by means of the search; $tf_{i,j}$ is the frequency of term i in document j divided by the total words in document j from the subset; df_i is the number of documents in the corpus containing term i; and N is the total number of documents in the corpus.

We calculated the TF-IDF characteristics for each term in each article that contained one of the ID-related terms. We used the complete set of documents (±29 000) as the corpus to calculate the inverse document frequency. We excluded English stop words (e.g. 'the' and 'and') and words that appeared in 95% of the articles from the analysis, to speed up calculations. Furthermore, we used the stem of each word to get rid of plurals, tenses and affixes (e.g. viruses → virus, infection → infect and infected → infect). More parameters can be accessed from our Kaggle notebook.

The TF-IDF values were used as input for the *K*-means algorithm in order to cluster the articles into different clusters. The objective of *K*-means is to divide n observations into k clusters, in such a way that each observation belongs to the cluster with the nearest mean. *K*-means does so by minimising the squared error between the mean of a cluster and the elements in the cluster. The goal of *K*-means is to minimise the sum of squared errors over all k clusters (Jain 2010).

$$O = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{I} \left|\left|\chi_{k,j,i} - c_k\right|\right|^2, \tag{2}$$

with O the objective function; K the number of clusters; $\mathcal{J}$ the number of articles; I the number of terms; $\chi_{k,j,i}$ the TF-IDF score for term i in article j in cluster k; and c_k the mean TF-IDF for cluster k.

Of the several available approaches to determine the number of clusters in *K*-means clustering (Kodinariya & Makwana 2013), we chose to use the widely accepted elbow method. In the elbow method, one starts with $K = 2$ and increases K by 1 in each step and calculates the sum of squared errors (Formula 2) for each step. While increasing K, the sum of squared errors will decrease dramatically in the beginning and will reach a plateau after increasing K further. The K, where this happens, is called the elbow, and that is where the value for K is set (Ng 2012).

The TF-IDF analysis yielded one TF-IDF value for each word for each paper (e.g. 259 values for the word 'quarantin'). This means that each article is described by about 8000 variables, and all these variables were used for the *K*-means clustering. For a visual representation of the 8000 variables for the articles, we needed to reduce the number of variables by means of a data reduction technique. We chose principle component analysis that combined all these variables in two linear combinations (so-called principal components) in a way that explains as much variance in the full data as possible (Wold *et al.* 1987). We chose two principal components to allow for a visual (two-dimensional) representation. In this visual representation, every paper is one dot, labelled by the cluster it belongs to.

Results

After loading and preprocessing the data, we ended up with 26 055 articles that contain the full-text article. In this section, we present the number of articles the terms related to ID appeared in and continue with the TF-IDF characteristics.

Papers containing intellectual disability care-related words

The number of articles in which the terms related to ID are present is shown in Table 1. Out of the 26 055 articles from the CORD-19 set, a subset of 259 mentioned one of the terms in their full text. Mental retardation was the most identified synonym for ID, and Down syndrome was the most identified syndrome. Fetal alcohol spectrum disorder and velo-cardio-facial syndrome were not found in the full texts. The subset of 259 articles was used for the TF-IDF analysis.

Table 1 Papers containing intellectual disability care-related words

Term	*N*
Intellectual disability	30
Learning disability	26
Mental retardation	123
Cognitive disability	5
Mental disability	24
Down syndrome	50
Fragile x	33
Prader Willi	1
Williams syndrome	1
Fetal alcohol spectrum disorder	0
Rett syndrome	2
Velo-cardio-facial syndrome	0
Angelman syndrome	2
Tuberous sclerosis complex	9
Cornelia de Lange syndrome	1
Total	259

Some articles mentioned multiple terms; therefore, the total is lower than the sum of the individual terms.

Term frequency–inverse document frequency and *K*-means

The elbow of the plot of the sum of squared errors versus K appeared to be at $K = 5$ (Fig. 1). Therefore, we decided to use $K = 5$ and let the algorithm group the articles in five clusters. These clusters are depicted in the principal component plot in Fig. 2. An overview of the subset of 259 articles and the clusters they belong to can be found at the end of the Kaggle notebook.

Table 2 shows the five clusters in the subset, with their top 10 TF-IDF terms. These top 10 words are stemmed; therefore, they might look a bit unusual. We also added a column with 'topic' that contains our own interpretation of the top 10 words. The biggest cluster was the 'Genetics' cluster, with 105 articles, and the smallest cluster was the 'Mental health' cluster, with 26 articles.

Discussion

From the CORD-19 set, we identified that 259 out of the 26 055 articles had one of the terms related to the ID care in their full text. We are aware that this list of used terms is not exhaustive, but considered the list sufficient for exploring whether CORD-19 has any

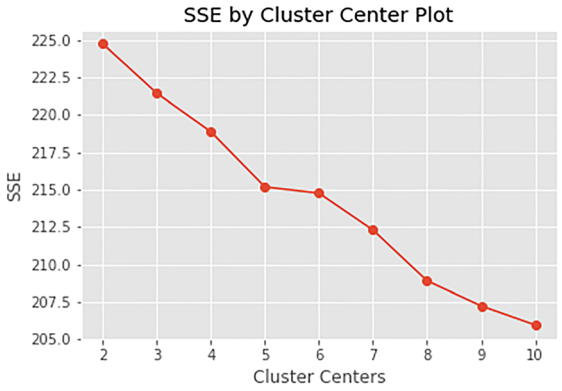

Figure 1. The sum of squared errors (SSEs) versus the number of clusters (see Formula 2). The 'elbow' of the graph appears to be at $K = 5$. [Colour figure can be viewed at wileyonlinelibrary.com]

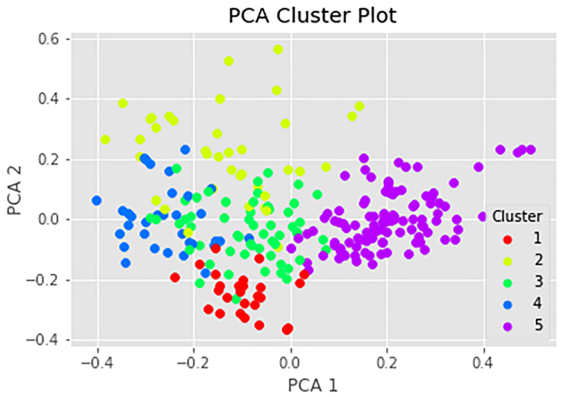

Figure 2. Principle component analysis (PCA) plot with two principal components. Contains 259 data points (articles), divided over five clusters. [Colour figure can be viewed at wileyonlinelibrary.com]

Table 2 Five clusters, their top 10 stemmed words from the TF-IDF analysis on the subset, their topics and number of articles in the cluster

	Top 10 words based on TF-IDF score	Topic	*N*
1	disord, social, diseas, studi, public, mental, patient, psychiatr, ptsd, health	Mental health	26
2	viral, may, caus, patient, diseas, immun, cell, vaccin, virus, infect	Viral diseases	38
3	cell, treatment, disord, studi, infect, diseas, zikv, may, children, patient	Diagnoses and treatments	57
4	breastfeed, hospit, patient, pneumonia, vaccin, infect, bronchiol, rsv, infant, children	Maternal care and paediatrics	33
5	virus, interact, dna, bind, express, activ, rna, gene, cell, protein	Genetics	105

Using a more extensive set of stopwords might have filtered 'may' and 'caus' out. 'zikv' is the often used abbreviation for Zika virus.

potential for research on people with an ID. We did not include the 22q11.2 deletion syndrome and autism spectrum disorders on purpose because these syndromes do not always co-occur with an ID. We do not expect a big rise in the number of articles identified if we had added more terms. The current number of 259 articles seems to be sufficient for a systematic literature review, particularly one aiming at genetics aspects. Moreover, thousands of articles are added to the CORD-19 dataset every week. Please note that our method and code are publicly available and can be reused by other researchers within just a few minutes.

Because of the application of the *K*-means algorithm on the outcomes of the TF-IDF calculations, we were able to cluster the articles that

contained the ID terms in five clusters on different topics. Researchers on ID can use these clusters to identify research areas on the topic of coronaviruses and people with an ID, and possible relevant articles already published on it. However, we have to be aware that setting the number of clusters at five was, to some extent, a design choice. When fixing the number of clusters at four, we saw more overlap of the clusters in the principle component analysis plot, probably because the articles in our cluster number 3 (from Table 2) were dispersed over the other clusters. When we set the number of clusters at six, the extra cluster was a small cluster, very similar to our cluster five (Genetics). Repeating the analyses on an extended edition of CORD-19 may result in a different number of clusters with other topics.

Currently, there are search engines available for the CORD-19 (Allen Institute for AI 2020) that can be useful to identify papers related to a certain topic. However, these search engines do not map the commonalities between different articles nor do they help to identify what has not been investigated yet, and this is where the methods described in the present paper have their added value.

Our exploratory analysis has shown that the CORD-19 set can be useful for researchers on ID and may be approached with text mining techniques. If we aim to create new hypotheses on the relations between coronaviruses and people with an ID based on literature, we need to utilise other text mining-based techniques. One approach that seems to have great potential is the literature-based discovery. This approach uses various computational methods to discover previously unknown links between two pieces of existing knowledge by analysing their relevant pieces of literature (Swanson 2008; Sebastian *et al*. 2017). In our case, one piece of knowledge would be the CORD-19 and the other a set of articles on people with an ID. Especially, the cluster similarity technique (Fujita 2012) and bibliographical coupling technique (Kostoff 2014) seem to be suitable to construct new hypotheses on possible relations between corona viruses and people with an ID.

Concluding, the CORD-19 has shown to be interesting for research on ID and coronaviruses. At first sight, there seemed to be relatively few articles (259 out of 26 055) in the CORD-19 related to ID. Nonetheless, the full-text articles were approachable by means of text mining techniques. With TF-IDF and *K*-means clustering, we were able to identify five clusters of articles on different topics. Researchers on ID can use the presented five clusters to identify research areas onID and coronaviruses and to decide on pursuing a systematic literature review (or not). Also, they may apply our approach again and again on the weekly updates of CORD-19, as this will take little time. For creating new hypotheses on the relations between coronaviruses and people with an ID, based on CORD-19, we suggested other text mining approaches.

Acknowledgement

We would like to thank the White House Office of Science and Technology Policy and their partners for publishing the CORD-19.

Source of Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors. It is part of a larger set of projects that have been funded by the Dutch Ministry of Health, Welfare and Sport.

Conflict of Interest

There are no conflicts of interest.

References

Aljaber B., Stokes N., Bailey J. & Pei J. (2010) Document clustering of scientific texts using citation contexts. *Information Retrieval* Springer **13**, 101–31.

Allen Institute for AI (2020) CORD-19 Explorer. Available at: https://cord-19.apps.allenai.org (retrieved 1 April 2020).

Amador Penichet L., Magdaleno Guevara D. & García Lorenzo M. M. (2018) New similarity function for scientific articles clustering based on the bibliographic references. *Computación y Sistemas*. Centro de Investigación en Computación, IPN **22**, 93–102.

Cold Spring Harbor Laboratory (2020a) bioRxiv. Available at: bioRxiv.org (retrieved 1 April 2020).

Cold Spring Harbor Laboratory (2020b) medRxiv. Available at: medRxiv.org (retrieved 1 April 2020).

CORD-19 (2020) COVID-19 Open Research Dataset (CORD-19), version *2020-03-23*. Available at: 10.5281/zenodo.3715506.

Fujita K. (2012) Finding linkage between sustainability science and technologies based on citation network analysis. In *2012 Fifth IEEE International Conference on Service-Oriented Computing and Applications (SOCA)*, pp. 1–6.

Glover G., Williams R., Heslop P., Oyinlola J. & Grey J. (2017) Mortality in people with intellectual disabilities in England. *Journal of Intellectual Disability Research* Wiley Online Library **61**, 62–74.

Hearst M. (2003) What is text mining, SIMS, UC Berkeley, 5.

Jain A. K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* Elsevier **31**, 651–66.

Kaggle (2020) Kaggle homepage. Available at: https://www.kaggle.com (retrieved 23 March 2020).

Kodinariya T. M. & Makwana P. R. (2013) Review on determining number of cluster in *k*-means clustering. *International Journal* **1**, 90–5.

Kostoff R. N. (2014) Literature-related discovery: common factors for Parkinson's disease and Crohn's disease. *Scientometrics*. Springer **100**, 623–57.

Lai C.-C., Shih T. P., Ko W. C., Tang H. J. & Hsueh P. R. (2020) Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *International Journal of Antimicrobial Agents* Elsevier **55**, 105924.

McIntosh K. & Perlman S. (2015) Coronaviruses, including severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). In: *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases, Updated Edition*, 8th edn. Elsevier Saunders, Philadelphia, PA.

Ng A. (2012) Clustering with the *k*-means algorithm, Machine Learning.

pandas Community (2020) Pandas version 1.0.3. Available at: https://pandas.pydata.org (retrieved 23 March 2020).

Python Software Foundation (2020) Python Language Reference, version 3.6.6. Available at: https://www.python.org/psf/ (retrieved 23 March 2020).

Rajaraman A. & Ullman J. D. (2011) *Mining of Massive Datasets*. Cambridge University Press, Cambridge.

scikit-learn Community (2020). scikit-learn version 0.22.2. Available at: scikit-learn.org.

Sebastian Y., Siew E.-G. & Orimaye S. O. (2017) Emerging approaches in literature-based discovery: techniques and performance review. *The Knowledge Engineering Review*. Cambridge University Press. Available at: https://www.cambridge.org/core/journals/knowledge-engineering-review/article/emerging-approaches-in-literaturebased-discovery-techniques-and-performance-review/09D7E9623190AA85C14350AEA64AD3E4.

Swanson D. R. (2008) Literature-based discovery? The very idea. In: *Literature-based Discovery*, pp. 3–11. Springer, Berlin, Heidelberg.

The United States National Library of Medicine at the National Institutes of Health (2020) PubMed. Available at: https://ncbi.nl.nih.gov/pubmed/ (retrieved 1 April 2020).

White House Office of Science and Technology Policy (2020) Call to action to the tech community on new machine readable COVID-19 dataset. Available at: https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/.

Wold S., Esbensen K. & Geladi P. (1987) Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*. Elsevier **2**, 37–52.

World Health Organization (2020a) Coronavirus disease (COVID-19) pandemic. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (retrieved 23 March 2020).

World Health Organization (2020b) Novel coronavirus – China update. Available at: https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/ (retrieved 23 March 2020).

World Health Organization (2020c) WHO COVID-19 database. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov (retrieved 1 April 2020).

World Health Organization (2020d) WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020. Available at: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (retrieved 23 March 2020).

Accepted 7 April 2020