Covid-19 through the lens of the peer-reviewed literature

What can we learn about the coronavirus pandemic from text-mining the titles of journal papers? By **Samuel R. Neal** and **Cindy Zheng**

As medical students and budding clinical academics, we have been following the coronavirus pandemic with interest. But it has been very difficult to keep up with the so-called "infodemic" of information about Covid-19, particularly all the peer-reviewed journal articles.

No single person has time to read all the research that has been published so far. So, we decided to use text mining techniques to identify themes that have emerged from the literature and explore how these themes have changed through the course of the pandemic.

Text mining refers to the concept of drawing meaning from unstructured text and language. Since emerging in the late 1990s, text mining techniques have grown in popularity and use, coinciding with the explosion of data over the past two decades. We first discovered text mining through the book *Text Mining with R*.[1] We would highly recommend this book for those interested in exploring text mining in a concise and accessible way, and our own analysis is based on several examples from the book, implemented through the *tidytext* package in R.

We chose to focus on peer-reviewed publications published in PubMed up to 31 May 2020. We did not include preprints or articles from other methods of dissemination (such as social media) both for simplicity and because we were interested in exploring publications that had

successfully passed peer review. We searched for publications with the terms "covid" or "sars-cov-2" or "2019-ncov" or "novel coronavirus" in their titles. This search captured some articles related to prior novel coronaviruses, so the search was limited to articles published on or after 1 January 2020. Duplicates (defined as entries with identical titles) were identified and removed.

Of the 575,052 articles entered into PubMed between 1 January and 31 May this year, 15,779 unique publications were related to the current coronavirus pandemic according to our search terms. While this represents only 2.7% of the total, it implies an extraordinary daily output: up to the end of May, the median number of publications per day was 73 (interquartile range: 19–213), with a peak of 596 on 20 May 2020.

With this sizeable body of literature, we can learn a lot about the current pandemic through the lens of the vocabulary that these publications use. In an attempt to capture a snapshot of article topics, we have plotted the 100 commonest words from the titles of Covid-19-related publications as a word cloud in Figure 1. Note that we excluded synonyms of Covid-19 from the list as these words are unsurprisingly the commonest. The size of the word is proportional to its frequency, with the largest words (such as "pandemic", "patients" and "disease") occurring most often.

This word cloud gives us an overview of the whole outbreak thus far, but we can zoom in on each individual month to gain a more granular insight into the research topics that were important at particular times. The *term frequency-inverse document frequency* (tf-idf) is a widely-used text mining metric used to identify keywords in documents – that is, words that are of specific importance to a particular document within a larger collection.[2] It provides a measure of the frequency with which a term is used within an individual document, adjusted for how rarely it is used across all

documents – and the larger the tf-idf, the more useful the term is as a keyword.

In Figure 2 we apply this metric to the title words of Covid-19-related publications, treating each month's titles as our "documents" and all titles as our "collection", to identify the top 15 words that are characteristic of each month of the outbreak.

In January we can infer from words such as "begun", "tentative", "originating" and "just" that there was a lot of uncertainty following the first cases of Covid-19 and, already, comparisons were being drawn between this new virus and the

Figure 1: Word cloud showing the top 100 title words from Covid-19-related publications in PubMed from 1 January 2020 to 31 May 2020. Note that synonyms of Covid-19 are not included.

Samuel Neal is a medical student at the University of Aberdeen, currently studying for an MRes in Child Health at the UCL Great Ormond Street Institute of Child Health.

Cindy Zheng is a medical student at the University of Aberdeen, currently studying for a Master of Public Health.
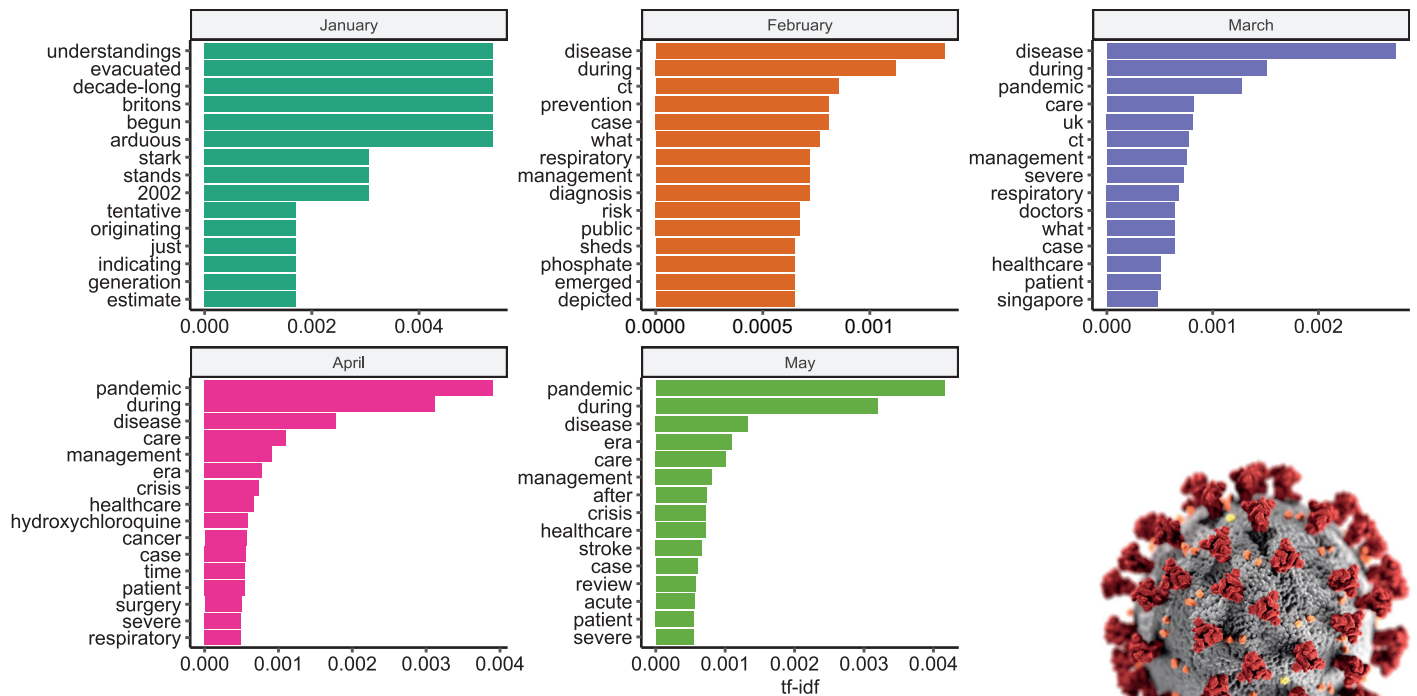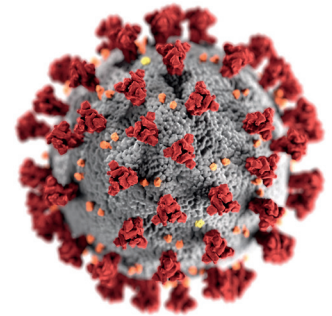
Figure 2: Comparison of the 15 title words with the highest term frequency-inverse document frequency for each month of the outbreak. Synonyms of Covid-19 are not included as, although they occur frequently and not in all months, they are relatively uninformative.

SARS epidemic that started in "2002". There was a realisation that "arduous" times might lie ahead.

In February key themes show the expansion of knowledge about this virus, with "case" series reporting on its presentation as a "respiratory" disease and the diagnostic utility of "CT" scans in distinguishing between Covid-19 and other respiratory infections. We can also see an emphasis on this outbreak being a "public" health problem, with an essential need for disease "prevention" and "management" of affected patients. Interestingly, the term "what" appears

disproportionately frequently in February and March, highlighting how many questions were generated as the world grappled with an emerging disease.

The word "pandemic" is more important from March to May than in January and February, and this coincides with the World Health Organization declaring Covid-19 a pandemic on 11 March 2020. There also appears to be an increasing focus on "severe" cases in the literature as management strategies were developed and shared for this group of patients. Furthermore, in March, there were more publications on the Covid-19 situation in "Singapore" as cases in the country surged.

In April, there was increasing interest in "hydroxychloroquine" as a possible treatment for Covid-19 and a focus on how Covid-19 would impact the "healthcare" of "cancer" patients and patients who require "surgery". Similarly, in May, the care of "stroke" patients received particular attention, with a suggestion that Covid-19 can cause neurological symptoms and complications, including stroke. And we also started to see researchers consider what might occur "after" the Covid-19 "era".

This, of course, is only a snapshot of the 15,000 peer-reviewed publications that were published in PubMed up to the end of May. The number of papers continues to increase, and the pandemic continues to evolve. Many questions have been answered, yet much remains unknown. ∎

No single person has time to read all the research that has been published so far. So, we used text mining to identify themes that have emerged from the literature

Note
Code for this article can be found on GitHub (bit.ly/2Zj5t9Z). This is an edited version of an article originally published on the *Significance* website. See significancemagazine.com/covid19 for the full version of this and other articles about coronavirus.

References
1. Silge, J. and Robinson, D. (2017) *Text Mining with R: A Tidy Approach*. Sebastopol, CA: O'Reilly Media.
2. Havrlant, L. and Kreinovich, V. (2017) A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, **46**(1), 27–36.