bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

COVID-19 research in Wikipedia

Giovanni Colavizza*[1]

[1] *University of Amsterdam, Netherlands*

Abstract

Wikipedia is one of the main sources of free knowledge on the Web. During the first few months of the pandemic, over 4,500 new Wikipedia pages on COVID-19 have been created and have accumulated close to 250M pageviews by early April 2020.[1] At the same time, an unprecedented amount of scientific articles on COVID-19 and the ongoing pandemic have been published online. Wikipedia's contents are based on reliable sources, primarily scientific literature. Given its public function, it is crucial for Wikipedia to rely on representative and reliable scientific results, especially so in a time of crisis. We assess the coverage of COVID-19-related research in Wikipedia via citations. We find that Wikipedia editors are integrating new research at an unprecedented fast pace. While doing so, they are able to provide a largely representative coverage of COVID-19-related research. We show that all the main topics discussed in this literature are proportionally represented from Wikipedia, after accounting for article-level effects. We further use regression analyses to model citations from Wikipedia and show that, despite the pressure to keep up with novel results, Wikipedia editors rely on literature which is highly cited, widely shared on social media, and has been peer-reviewed.

COVID-19, Coronavirus, CORD-19, Scientometrics, Bibliometrics, Wikipedia.

1 Introduction

Alongside the primary health crisis, the COVID-19 pandemic has been recognized as an information crisis, or an "infodemic" [62, 11, 21]. Widespread misinformation [53] and low levels of health literacy [40] are two of the main issues. In an effort to deal with them, the World Health Organization maintains a list of relevant research updated daily [64], as well as a portal to provide information to the public [2]; similarly does the European Commission [3], and many other countries and organizations. The need to convey accurate, reliable and understandable medical information online has never been so pressing.

*This is a working document which has not been peer reviewed yet. Please address any comment or remark to `g.colavizza@uva.nl`.

[1] `https://wikimediafoundation.org/covid19/data` [accessed 2020-05-10].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Wikipedia plays a fundamental role as a public source of information on the Web, striving to provide "neutral" and unbiased contents [34]. Wikipedia is particularly important as go-point to access trusted medical information [53, 51]. Fortunately, Wikipedia biomedical articles have been repeatedly found to be highly visible and of high quality [5, 31]. Wikipedia's verifiability policy mandates that readers can check the sources of information contained in Wikipedia, and that reliable sources should be secondary and published.[2] These guidelines are particularly strict with respect to biomedical contents, where the preferred sources are, in order: systematic reviews, reviews, books and other scientific literature.[3]

The COVID-19 pandemic has put Wikipedia under stress with a large amount of new, often non-peer-reviewed research being published in parallel to a surge in interest for information related to the pandemic [16]. The response of Wikipedia's editor community has been fast: since March 17 2020, all COVID-19-related Wikipedia pages have been put under indefinite sanctions entailing restricted edit access, to allow for a better vetting of their contents.[4] In parallel, a WikiProject COVID-19 has been established and a content creation campaign is ongoing [16, 22].[5] While this effort is commendable, it also raises questions on the capacity of editors to find, select and integrate scientific information on COVID-19 at such a rapid pace, while keeping quality high. In Figure 1 we show the time in number of months from publication to a first citation from Wikipedia for a large set of COVID-19-related articles (see Section 3). In 2020, this time has become negative on average: articles on COVID-19 are frequently cited in Wikipedia even before their official publication date, based on early access versions of articles.

In this work, we pose the following general question: *Is Wikipedia relying on a representative and reliable sample of COVID-19-related research?* We break this question down into the following two research questions:

1. RQ1: Is the literature cited from Wikipedia representative of the broader topics discussed in COVID-19-related research?

2. RQ2: Is Wikipedia citing COVID-19-related research during the pandemic following the same inclusion criteria adopted before and in general?

We approach the first question by clustering COVID-19-related publications using text and citation data, and comparing Wikipedia's coverage of different clusters before and during the pandemic. The second question is instead approached using regression analysis. In particular, we model whether an article is cited from Wikipedia or not, and how many citations it receives from Wikipedia. We then again compare results for articles cited before and during the pandemic, and with previous art.

[2] `https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources` [accessed 2020-05-10].

[3] `https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources_(medicine)` [accessed 2020-05-10].

[4] `https://en.wikipedia.org/wiki/Wikipedia:General_sanctions` [accessed 2020-05-10].

[5] `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_COVID-19` [accessed 2020-05-10].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

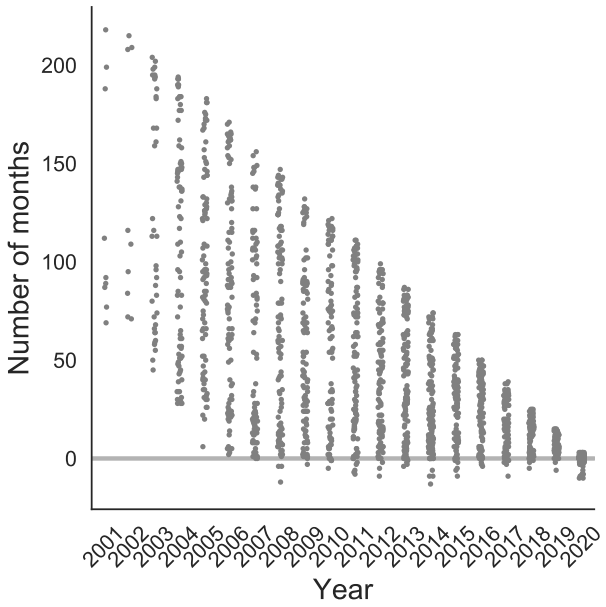


Figure 1: Number of months elapsed from publication to the first Wikipedia citation (scatterplot binned by year) of COVID-19-related research. In 2020, the average number of months from (official) publication to the first citation from Wikipedia has become negative, due to the effect of early releases by some journals. Also see Figure7.

Our main finding is that Wikipedia contents rely on representative and high-impact COVID-19-related research. (RQ1) During the past few months, Wikipedia editors have successfully integrated COVID-19 and coronavirus research, keeping apace with the rapid growth of related literature by including (a representative sample of) it. (RQ2) The inclusion criteria used by Wikipedia editors to integrate COVID-19-related research during the pandemic are consistent with those from before, and appear reasonable in terms of source reliability. Specifically, editors prefer articles from specialized journals over mega journals or pre-prints, and focus on highly cited and/or highly socially visible literature. Some altmetrics such as Twitter shares, mentions in news and blogs, Mendeley readers are complementing citation counts from the scientific literature as an indicator of impact positively correlated with citations from Wikipedia. After controlling for these article-level impact indicators, and for publication venue, time and size-effects, there is no indication that the topic of research matters with respect to receiving citations from Wikipedia, signaling that Wikipedia is currently not over nor under-relying on any specific COVID-19-related scientific topic.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2 Related work

Wikipedia articles are created, improved and maintained by the efforts of the community of volunteer editors [44, 10], and they are used in a variety of ways by a wide user base [50, 29, 42]. The information Wikipedia contains is generally considered to be of high-quality and up-to-date [44, 23, 17, 27, 43, 5, 51], notwithstanding margins for improvement and the need for constant knowledge maintenance [10, 30, 15].

Following Wikipedia's editorial guidelines, the community of editors creates contents often relying on scientific and scholarly literature [38, 18, 6], and therefore Wikipedia can be considered a mainstream gateway to scientific information [28, 19, 30, 48, 32, 42]. Unfortunately, few studies have considered the *representativeness and reliability* of Wikipedia's scientific sources. The evidence on what scientific and scholarly literature is cited in Wikipedia is slim. Early studies point to a relative low overall coverage, indicating that between 1% and 5% of all published journal articles are cited in Wikipedia [45, 49, 63]. Previous studies have shown that the subset of scientific literature cited from Wikipedia is more likely on average to be published on popular, high-impact-factor journals, and to be available in open access [37, 55, 6].

Wikipedia is particularly relevant as a means to access medical information online [28, 19, 51, 53]. Wikipedia medical contents are of very high quality on average [5] and are primarily written by a core group of medical professionals part of the nonprofit Wikipedia Medicine [48]. Articles part of the WikiProject Medicine "are longer, possess a greater density of external links, and are visited more often than other articles on Wikipedia" [31]. Perhaps not surprisingly, the fields of research that receive most citations from Wikipedia are "Medicine (32.58%)" and "Biochemistry, Genetics and Molecular Biology (31.5%)" [6]; Wikipedia medical pages also contain more citations to scientific literature than the average Wikipedia page [32]. Margins for improvement remain, as for example the readability of medical content in Wikipedia remains difficult for the non-expert [9]. Given Wikipedia's medical contents high quality and high visibility, our work is concerned with understanding whether the Wikipedia editor community has been able to maintain the same standards for COVID-19-related research.

3 Data and Methods

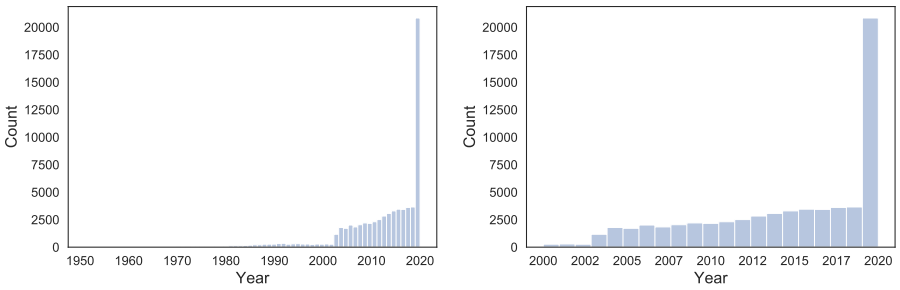
3.1 COVID-19-related research

COVID-19-related research is not trivial to delimit [13]. Our approach is to consider several public and regularly-updated lists of publications:

- The COVID-19 Open Research Dataset (CORD-19): a collection of COVID-19 and coronavirus related research, including publications from PubMed Central, bioRxiv and medRxiv [61].
- The World Health Organization Database [4].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- The Dimensions COVID-19 Publications list [1].



(a) Overall. (b) Since 2000 included.

Figure 2: COVID-19-related literature over time.

Publications from these three lists are merged, and duplicates removed using publications identifiers, including DOI, PMID, PMCID, Dimensions ID. Publications without at least one identifier among these are discarded. As of April 24 2020, the resulting list of publications contains 69,969 entries with a valid identifier, of which 20,841 have been released in 2020, as it can be seen from Figure 2. The research on corovaniruses, and therefore the accumulation of this corpus over time, has been clearly influenced by the SARS (2003+), MERS (2012+) and COVID-19 outbreaks. We use this list of publications to represent COVID-19 and coronavirus research in what follows. More details are given in the online repositories.

3.2 Auxiliary data sources

In order to study Wikipedia's coverage of this list of COVID-19-related publications, we use data from Altmetric [47, 39]. Altmetric provides Wikipedia citation data relying on known identifiers.[6] Despite this limitation, Altmetric data have been previously used to map Wikipedia's use of scientific articles [63, 58, 6], especially since citations from Wikipedia are considered a possible measure of impact [52, 25]. Publications from the full list above are queried using the Altmetric API by DOI or PMID. In this way, 43,561 publications could be retrieved. After merging for duplicates by summing Altmetric indicators, we have a final set of 40,866 distinct COVID-19-related publications with an Altmetric entry.

Furthermore, we use data from Dimensions [20, 33] in order to get citation counts for COVID-19-related publications. The Dimensions API is also queried by DOI and PMID, resulting in 64,040 matches. All auxiliary data sources have been queried on April 24 2020 too.

[6]The identifiers considered by Altmetric in order to establish a citation from Wikipedia to an article currently include: DOI, URI from a domain white list, PMID, PMCID, arXiv ID. `https://help.altmetric.com/support/solutions/articles/6000060980-how-does-altmetric-track-mentions-on-wikipedia` [accessed 2020-04-27].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3.3 Methods

We detail here the experimental choices made for a clustering analysis using publication text and citation data. Details on regression analyses are, instead, given in the corresponding section.

Text-based clustering of publications was performed in two ways: topic modelling and k-means relying on SPECTER embeddings. Both methods made use of the titles and abstracts of available publications, by concatenating them into a single string. We detected 66,915 articles in English, out of 69,969 total articles (-3054 over total). Of these, 13,852 have no abstract, thus we only used their title. Before performing topic modelling, we applied a pre-processing pipeline using scispaCy's `en_core_sci_md` model [36] to convert each document into a bag-of-words representation, which includes the following steps: entity detection and inclusion in the bag-of-words for entities strictly longer than one token; lemmatisation; removal of (isolated) punctuation, stopwords and tokens composed of a single character; inclusion of frequent bigrams. SPECTER embeddings were instead retrieved from the API without any pre-processing.[7] We then trained and compared topic models using Latent Dirichlet Allocation (LDA) [8], Correlated Topic Models (CTM) [7], Hierarchical Dirichlet Process (HDP) [54] and a range of topics between 5 and 50. We found similar results in terms of topic contents and in terms of their Wikipedia coverage (see Section 4) across models and over multiple runs, and a reasonable value of the number of topics to be between 15 and 25 from a topic coherence analysis [35]. Therefore, in what follows we discuss an LDA model with 15 topics.[8] The top words for each topic of this model are given in the SI, while topic intensities over time are plotted as a heat map in Figure 9.

SPECTER is a novel method to generate document-level embeddings of scientific documents based on a transformer language model and the network of citations [12]. SPECTER does not require citation information at inference time, and performs well without any further training on a variety of tasks. We embed every paper and cluster them using k-means with $k = 20$. The number of clusters was established using the elbow and the silhouette methods; different values of k could well be chosen, we again decided to pick the smallest reasonable value of k.

We then turned our attention to citation network clustering. We constructed a bibliographic coupling citation network [24] based all publications provided by Dimensions and with references; these amount to 54,293. Edges were weighted using fractional counting [41], hence dividing the number of references in common between any two publications by the length of the union of their reference lists (thus, the max possible weight is 1.0). We only used the giant weakly connected component, which amounts to 53,131 nodes (-1162 over total) and 21,078,192 edges with a median weight of 0.0156. We clustered the citation net-

[7] `https://github.com/allenai/paper-embedding-public-apis` [accessed 2020-04-25].

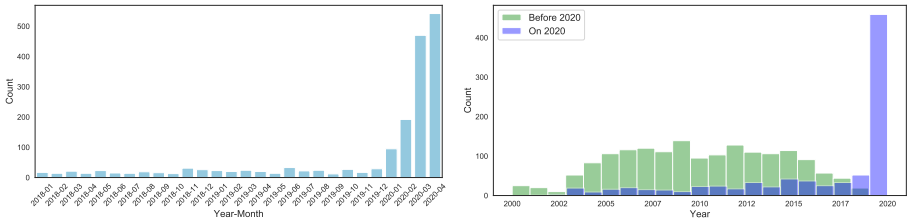
[8] We used `gensim`'s implementation for LDA [46] and `tomotopy` for CTM and HTM, `https://bab2min.github.io/tomotopy` [version 0.7.0]. The reader can find more results and the code to replicate all experiments in the accompanying repository.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

work using the Leiden algorithm [60] with a resolution parameter of 0.05 and the Constant Potts Model (CPM) quality function [59]. With this configuration, we found that the largest 13 clusters account for half the nodes in the network, and the largest cluster is composed of circa 7,000 nodes.

4 Results

An intense editorial work was carried out over the early weeks of 2020 in order to include scientific information on COVID-19 and coronaviruses into Wikipedia [22]. From Figure 3a, we can appreciate the surge in new citations added from Wikipedia to COVID-19 research. Importantly, these citations were not only added to cope with the growing amount of new literature, but also to fill gaps by including literature published before 2020, as shown in Figure 8b. The total fraction of COVID-19-related articles that are cited at least once from Wikipedia over the total is 3.1%. Yet, this number is uneven over languages and over time. Articles in English have a 3.2% chance of being cited from Wikipedia, while articles in other languages only a 0.036% chance. To be sure, the whole corpus is English dominated, as we discussed above. This might be an artefact of the coverage of the data sources, as well as the way the corpus was assembled. The coverage of articles over time is instead given in Figure 4, starting from 2003 when the first surge of publications happens due to SARS. We can appreciate that the coverage seems to be uneven, and less pronounced for the past few years (2017-2020), yet this needs to be considered in view of the high growth of publications in 2020. Hence, while 2020 is a relatively low-coverage year (2.2%), it is already the year with the most publications cited from Wikipedia in absolute number (Figure 8b).



(a) Number of citations from Wikipedia to COVID-19 literature, per month from January 2018 included.

(b) Publication year of the articles cited from Wikipedia, distinguishing between citations added before 2020 and in 2020.

Figure 3: Timing of new citations from Wikipedia, and publication years of the articles they refer to. See Figure 8 for the full timeline.

Citation distributions are skewed in Wikipedia as they are in science more generally. Some articles receive a high number of citations from Wikipedia and some Wikipedia articles make a high number of citations to COVID-19-related literature. Table 1 lists the top 20 Wikipedia articles by number of citations to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

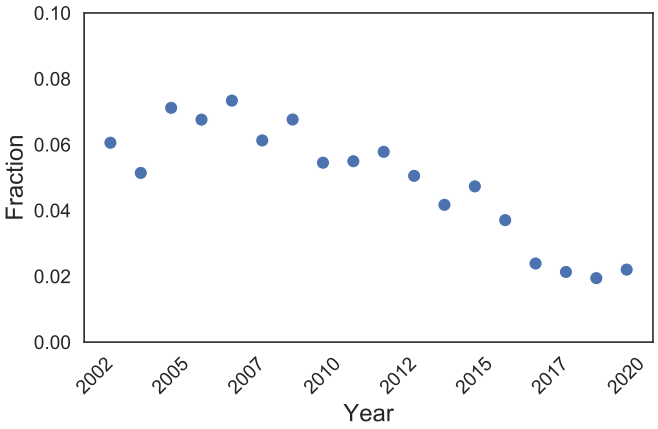


Figure 4: Fraction of COVID-19-related articles cited from Wikipedia per year, from 2003 included.

COVID-19-related research. These articles, largely in English, primarily focus on the recent pandemic and coronaviruses/viruses from a virology perspective, as already highlighted in a study by the Wikimedia Foundation [22]. Table 2 reports instead the top 20 journal articles cited from Wikipedia. These also follow a similar pattern: articles published before 2020 focus on virology and are made of a high proportion of review articles. The top cited article, for example, deals with virus taxonomy. Articles published in 2020, instead, have a focus on the ongoing pandemic, its origins, as well as its epidemiological and public health aspects. As we see next, this strongly aligns with the general trends of COVID-19-related research over time.

In order to discuss research trends in our CORD-19-related corpus at a higher level of granularity, we grouped the 15 topics from the LDA topic model into seven *macrotopics* and labelled them as follows:

- **Coronaviruses**: topics 2, 4; this macrotopic includes research explicitly on coronaviruses (COVID-19, SARS, MERS) from a variety of perspectives (virology, epidemiology, intensive care, historical unfolding of outbreaks).
- **Public health and epidemics**: topics 9, 12; research on global health issues, healthcare, epidemiology, including modelling the transmission and spread of pathogens.
- **Transmission**: topics 1, 7, 14; research on the origin and transmission of viruses from animals to humans and among humans.
- **Molecular biology**: topics 0, 5, 8; research on the genetics and biology of viruses.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- **Respiratory diseases**: topic 6; research on respiratory diseases (pneumonia, influenza), their detection and treatment.
- **Immunology**: topics 3, 10; research on vaccines, drugs, therapies.
- **Clinical medicine**: topics 11, 13; research on intensive care, hospitalization and clinical trials.

The grouping is informed by agglomerative clustering based on the Jensen-Shannon distance between topic-word distributions (Figure 12). To be sure, the labelling is a simplification of the actual publication contents. It is also worth considering that topics overlap substantially. The COVID-19 research corpus is dominated by literature on coronaviruses, public health and epidemics, largely due to 2020 publications. COVID-19-related research did not accumulate uniformly over time. We plot the relative (yearly mean, Figure 10a) and absolute (yearly sum, Figure 10b) macrotopic intensity. From these plots, we confirm the periodisation of COVID-19-related research as connected to known outbreaks. Outbreaks generate a shift in the attention of the research community, which is apparent when we consider the relative macrotopic intensity over time in Figure 10a. The 2003 SARS outbreak generated a shift associated with a raise of publications on coronaviruses and on the management of epidemic outbreaks (public health, epidemiology). Stable macrotopics instead include molecular biology, viral transmission, immunology and clinical medicine. A similar shift is again happening, at a much larger scale, during the current COVID-19 pandemic. When we consider the absolute macrotopic intensity, which can be interpreted as the number of articles on a given topic (Figure 10b), we can appreciate how scientists are mostly focusing on topics related to public health, epidemics and coronaviruses (COVID-19) during these first months of the current pandemic.

4.1 RQ1: Wikipedia coverage of COVID-19-related research

We address here our first research question: *Is the literature cited from Wikipedia representative of the broader topics discussed in COVID-19-related research?* We start by comparing the macrotopic coverage of articles cited from Wikipedia with those which are not. In Figure 5, three plots are provided: the macrotopic intensity of articles published before 2020 (Figure 5a), in 2020 (Figure 5b) and overall (Figure 5c). The macrotopic intensity is averaged and 95% confidence intervals are provided. From Figure 5c we can see that Wikipedia seems to cover COVID-19-related research well. The macrotopics on immunology, molecular biology and transmission seem slightly over represented, where clinical medicine, coronaviruses, public health and epidemics are slightly under represented. A comparison between publications from 2020 and from before highlights further trends. In particular, in 2020 Wikipedia editors have focused more on recent literature on coronaviruses, thus directly related to COVID-19 and the current pandemic, and proportionally less on literature on public health and epidemics,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

which is also dominating 2020 publications. The traditional slight over representation of immunology and viral transmission literature persists. Detailed Kruskal–Wallis H test statistics and Cohen's d effect sizes are provided in the SI (Figure 13 and Tables 3, 4, 5). While distributions are significantly different for most macrotopics and periodisations, the effect sizes are always small or very small. The coverage of COVID-19-related literature from Wikipedia appears therefore to be reasonably balanced from this first analysis, and to remain so in 2020. The topical differences we found, especially around coronaviruses and the current COVID-19 outbreak, might in part be explained by the criterion of notability which led to the creation or expansion of Wikipedia articles on the ongoing pandemic.[9]

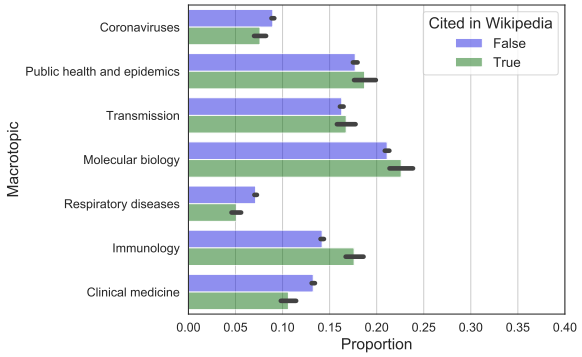
A complementary way to address the same research question is to investigate Wikipedia's coverage of publication clusters. We consider here both SPECTER k-means clusters and bibliographic network clusters. While we use all 20 SPECTER clusters, we limit ourselves to the top-n network clusters which are necessary in order to cover at least 50% of the nodes in the network. In this way, we consider 13 clusters for the citation network, all of size above 800. In Figure 6 we plot the % of articles cited from Wikipedia per cluster, and the clusters size in number of publications they contain. There is a general size effect, more pronounced for the SPECTRE clustering (Figure 6a), by which larger clusters are more represented than smaller clusters. When considering the citation network clustering solution, this applies to the two largest clusters, but not to the rest (Figure 6b).

When we characterise clusters using macrotopic intensities, some clear patterns emerge. Starting with SPECTER k-means clusters, the most cited clusters are number 6 (main macrotopics: molecular biology), 8 (main macrotopics: coronaviruses and public health, especially focusing on COVID-19 characteristics, detection and treatment) and 1 (main macrotopics: transmission, with an emphasis on genetics). The least cited clusters include number 18 (containing pre-prints from Research Square) and 5 (focused on the social sciences, and especially economics, e.g., from SSRN journals). Considering citation network clusters, the largest and most cited are number 0 (containing research on all coronaviruses from a variety of perspectives with a good balance among macrotopics) and 1 (with publications exclusively from 2020 on the current pandemic). The other clusters are smaller and hence more specialized. For example, at the two extremes we have cluster 11 (highly cited, mainly containing literature on viral infectious diseases) and cluster 10 (lowly cited, focused on animal to human transmission and immunology). The reader can explore all clusters using the accompanying repository.

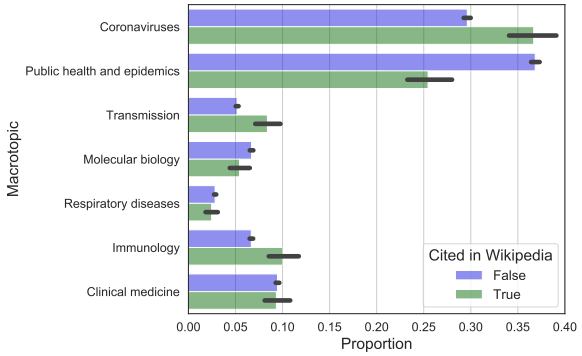
We have seen so far that Wikipedia relies on a reasonably representative sample of COVID-19-related literature, when assessed using topic models. During 2020, the main effort of editors has focused on catching-up with abundant new research (and some backlog) on the ongoing pandemic and, to a lower extent, on public health and epidemiology literature. When assessing coverage using

[9] `https://en.wikipedia.org/wiki/Wikipedia:Notability` [accessed 2020-05-10].

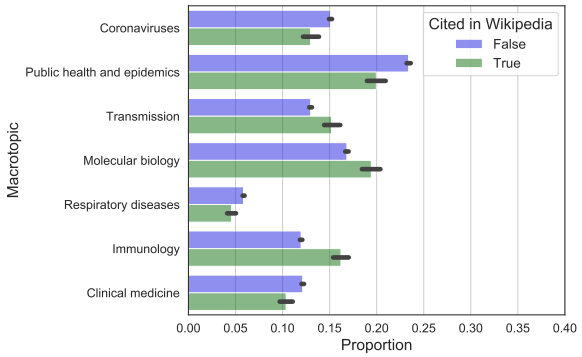
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) Published before 2020. Note: this plot also considers as cited from Wikipedia those publications published before 2020 and cited for the first time in 2020.



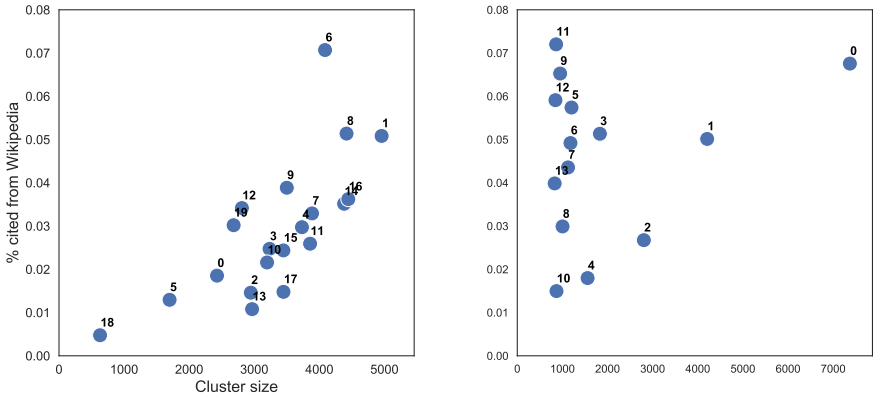
(b) Published in 2020.



(c) All publications.

Figure 5: Average macrotopic intensity of COVID-19-related publications cited in Wikipeda (green) or not (blue). 95% confidence intervals are given. See Figure 13 and Tables 3, 4, 5 for tests and effect sizes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) SPECTER k-means (all). (b) Bibliographic coupling (top-13).

Figure 6: Proportion of articles cited from Wikipedia (y axis) per cluster size (x axis).

different clustering methods, we find a size effect by which larger clusters are proportionally more cited from Wikipedia. Yet, we also find that, in particular with citation network clusters, smaller clusters can be either highly or lowly cited from Wikipedia on average. Lastly, we find an under representation of pre-print and social science research using text-based clustering, which we cannot find using citation network clustering. Despite this overall positive result, differences in coverage persist. In the next section, we further assess whether these differences can be explained away by considering article-level measures of impact.

4.2 RQ2: Predictors of citations from Wikipedia

In this section, we address our second research question: *Is Wikipedia citing COVID-19-related research during the pandemic following the same quality criteria adopted before and in general?* We use regression analysis in two forms: a logistic regression to model if a paper is cited from Wikipedia or not, and a linear regression to model the number of citations a paper receives from Wikipedia. While the former model captures the suitability of an article to provide encyclopedic evidence, the latter captures its relevance to multiple Wikipedia articles.

Dependent variables. *Wikipedia citation counts* for each article are taken from Altmetric. If this count is of 1 or more, an article is considered as cited from Wikipedia. We consider citation counts from Altmetric at the time of the data collection for this study. We focus on the articles with a match from Dimensions, and consider an article to have zero citations from Wikipedia if it is

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

not found in the Altmetric database. *Of 64,040 articles, 2175 (3.4%) are cited from Wikipedia.*

Independent variables. We focus our study on three groups of independent variables at the article level capturing impact, topic and timing respectively. Previous studies have shown how literature cited from Wikipedia tends to be published in prestigious journals and available in open access [37, 55, 6]. We are interested to assess some of these known patterns for COVID-19-related research, to complement them by considering citation counts and the topics discussed in the literature, and eventually to understand whether there has been any change in 2020.

Article-level variables include citation counts from Dimensions and a variety of altmetric indicators [47] which have been found to correlate with later citation impact of COVID-19 research [26]. Altmetrics include the number of: Mendeley readers, Twitter interactions (unique users), Facebook shares, mentions in news and blog posts (summed due to their high correlation), mentions in policy documents; the expert ratio in user engagement[10]. We also include the top-20 publication venues by number of articles in the corpus using dummy coding, taking as reference level a generic category 'other' which includes articles from all other venues. It is worth clarifying that article-level variables were also calculated at the time of the data collection for this study. This might seem counter-intuitive, especially for the classification task, as one might prefer to calculate variables at the time when an article was first cited from Wikipedia. We argue that this is not necessary, since Wikipedia can always be edited and citations removed as easily as added. As a consequence, a citation from Wikipedia (or its absence) is a continued rather than a discrete action, justifying calculating all counts at the same time for all articles in the corpus.

Topic-level variables capture the topics discussed in the articles, as well as their relative importance in terms of size (size-effects). They include the macrotopic intensities for each article, the size of the SPECTER cluster an article belongs to, and the size of its bibliographic coupling network cluster (for the 13 largest clusters with more than 800 articles each, setting it to zero for articles belonging to other clusters. In this way, the variable accounts for both size and thresholding effects). Cluster identities for both SPECTRE and citation network clusters were also tested but did not add contribute significantly to the models. Several other measures were considered, such as the semantic centrality of an article to its cluster centroid (SPECTER k-means) and network centralities, but since these all strongly correlate to size indicators they were discarded.

Lastly, we include the year of publication using dummy coding and 2020 as reference level. Several other variables were tested. The proposed selection removes highly correlated variables while preserving the information required

[10] Calculated using Altmetric data which distinguishes among the number of researchers (r), experts (e), practitioners (p) and members of the public (m) engaging with an article. The expert ratio is defined as $\frac{r+e+p}{r+e+p+m}$.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by the research question. The Pearson's correlations for the selected variables are shown in Figure 11. More details, along with a full profiling of variables, are provided in the accompanying repository.

Model. We consider two models: a Logistic model on being cited from Wikipedia (1) or not (0) and an Ordinary Least Squares (OLS) model on citation counts from Wikipedia. Both models use the same set of independent variables and the following transformations:

$$
\begin{aligned}
&is_cit_w|ln(n_cit_w+1) = C(publication_year) + ln(times_cited+1)+\\
&ln(counts_mendeley+1) + ln(counts_policy+1)+\\
&ln(counts_twitter_unique+1) + ln(counts_blogs_news+1)+\\
&ln(counts_facebook+1) + expert_ratio + C(top_journal)+\\
&tm_coronaviruses + tm_phe + tm_transmission+\\
&tm_molecular_biology + tm_respiratory_diseases + tm_immunology+\\
&tm_clinical_medicine + ln(spectre_cluster_size+1) + ln(network_cluster_size+1)
\end{aligned}
$$

All count variables are transformed by adding one and taking the natural logarithm, while the remaining variables are either indicators or range between 0 and 1 (such as macrotopic intensities, beginning with a *tm_* appendix; *tm_phe* is 'public health and epidemics'). OLS models including log transform and the addition of 1 for count variables such as citation counts, have been found to perform well in practice when compared to more involved alternatives [57, 56]. Furthermore, all missing values were set to zero, except for the publication year, venue (journal) and macrotopic intensities; removing those rows with missing values instead, yielded similar results.

Discussion. We discuss results for three models: two Logistic regression models one on articles published and first cited up to and including in 2020, and one on articles published and first cited up to an including 2019. The 2019 model only considers articles published in 2019 or earlier and cited for the first time from Wikipedia in 2019 or earlier, or articles never cited from Wikipedia, discarding articles published in 2020 or cited from Wikipedia in 2020 irrespective of their publication time. We also discuss an OLS model predicting (the log of) citation counts including all data up to and including 2020. We do not discuss a 2019 OLS model since it would require Wikipedia citation counts calculated at the end of 2019, which were not available to us. Regression tables for these three models are provided in the SI, Section 5, while Figure 14 shows the distribution of some variables distinguishing between articles cited from Wikipedia or not. Logistic regression tables provide marginal effects, while the OLS table provides the usual coefficients. The actual number of datapoints used to fit each model, after removing those which contained any null value, is given in the regression tables.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Considering the Logistic models first, we can show some significant effects.[11] First of all, the year of publication is always negatively correlated with being cited from Wikipedia, compared with the reference category 2020. This seems largely due to publication size-effects, since the fraction of 2020 articles cited from Wikipedia is quite low (see Figure 4. The 2019 model indeed shows positive correlations for all years when compared to the reference category 2019, and indeed 2019 is the year with lowest coverage since 2000. Secondly, some of the most popular venues are negatively correlated with citations from Wikipedia, when compared to an 'other' category (which includes all venues except the top 20). In the 2020 model, these less-cited-from-Wikipedia venues include pre-print servers (medRxiv in particular), mega-journals (PLoS One) and social sciences (SSRN). Positive correlations occur for few other specialized venues, such as Antiviral Research, The Lancet and Virology. When we consider indicators of impact, we see a significant positive effect for citation counts, Mendeley readers, Twitter, news and blogs mentions; we see instead no effect for policy document mentions and Facebook engagements. This is consistent in the 2019 model, except for smaller effects on citation counts and higher effects of Mendeley readers. This result, on the one hand, highlights the importance of academic indicators of impact such as citations, and on the other hand suggests the possible complementarity of altmetrics in this respect. Since certain altmetrics can accumulate more rapidly than citations [14], they could complement them effectively when needed [26]. Furthermore, the expert ratio in altmetrics engagement is negatively correlated with being cited from Wikipedia in 2020. This might be due to the high altmetrics engagement with COVID-19 research in 2020, but it could also hint at the possibility that social media impact need not be driven by experts in order to be correlated with scientific impact. We can further see how cluster size-effects are positively correlated with being cited from Wikipedia in 2020, and especially so for SPECTER clusters, but not in 2019. Lastly, we can see that *macrotopic intensities are never correlated with being cited from Wikipedia in either model*, underlining that Wikipedia appears to be proportionally representing all COVID-19-related research and that residual topical differences in coverage are due to article-level effects.

The OLS 2020 model largely confirms these results, except that mentions in policy documents and Facebook engagements become here positively correlated with the number of citations from Wikipedia. It is important to underline that, for all these results, there is no attempt to establish causality. For example, the

[11] Marginal effect coefficients should be interpreted as follows. For binary discrete variables (0/1), they represent the discrete rate of change in the probability of the outcome, everything else kept fix; therefore, a change from 0 to 1 with a significant coefficient of 0.01 entails an increase in the probability of the outcome of 1%. For categorical variables with more than two outcomes, they represent the difference in the predicted probabilities of any one category relative to the reference category. For continuous variables, they represent the instantaneous rate of change. It might be the case that this can also be interpreted linearly (e.g., a significant change of 1 in the variable entails a change proportional to the marginal effect coefficient in the probability of the outcome). Yet, this rests on the assumption that the relationship between independent and dependent variables is linear irrespective of the orders of magnitude under consideration. This might not be the case in practice.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

positive correlation between the number of Wikipedia articles citing a scientific article and the number of policy documents mentioning it, might be due to policy document editors using Wikipedia, Wikipedia editors using policy documents, both or neither. The fact is, more simply, that some articles are picked up by both.

5 Conclusion

The results of this study, while preliminary and given as the pandemic is still ongoing, provide some reassuring evidence. It appears that Wikipedia is well-able to keep track of COVID-19-related research. Of 64,040 articles in our corpus, 2175 (3.4%) are cited from Wikipedia: a similar share to what found in previous studies. Wikipedia editors are relying on scientific results representative of the several topics included in a large corpus of COVID-19-related research. They have been effectively able to cope with new, rapidly-growing literature. The minor discrepancies in coverage that persist, with slightly more Wikipedia-cited articles on topics such as molecular biology and immunology and slightly fewer on coronaviruses and public health, are fully explained away by article-level effects. Wikipedia editors rely on impactful and visible research, as evidenced by largely positive citation and altmetrics correlations. Importantly, Wikipedia editors also appear to be following the same inclusion standards in 2020 as before: in general, they rely on specialized and highly-cited results from reputed journals, avoiding e.g., pre-prints.

The main limitation of this study is that it is purely observational, and thus does not explain why some articles are cited from Wikipedia or not. While in order to assess the coverage of COVID-19-related research from Wikipedia this is of secondary importance, it remains relevant when attempting to predict and explain it. A second limitation is that this study is based on citations from Wikipedia to scientific publications, and no Wikipedia content analysis is performed. Citations to scientific literature, while informative, do not completely address the interrelated questions of Wikipedia's knowledge representativeness and reliability. Therefore, some directions for future work include comparing Wikipedia coverage with expert COVID-19 review articles, as well as studying Wikipedia edit and discussion history in order to assess editor motivations. Another interesting direction for future work is the assessment of all Wikipedia citations to any source from COVID-19 Wikipedia pages, since here we only focused on the fraction directed at COVID-19-related scientific articles. Lastly, future work can address the engagement of Wikipedia users with cited COVID-19-related sources.

Wikipedia is a fundamental source of free and unbiased knowledge, open to all. The capacity of its editor community to quickly respond to a crisis and provide high-quality contents is, therefore, critical. Our results here are encouraging in this respect.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data and code availability

All the analyses can be replicated using code and following the instructions given in the accompanying repository: `https://github.com/Giovanni1085/covid-19_wikipedia`. The preparation of the data follows the steps detailed in this repository instead: `https://github.com/CWTSLeiden/cwts_covid` [13]. Analyses based on Altmetric and Dimensions data require access to these services.

Acknowledgements

Digital Science kindly provided access to Altmetric and Dimensions data.

References

[1] Dimensions COVID-19 Publications, 2020. URL: `https://docs.google.com/spreadsheets/d/1-kTZJZ1GAhJ2m4GAIhw1ZdlgO46JpvX0ZQa232VWRmw/edit#gid=2034285255`.

[2] EPI-WIN: WHO Information Network for Epidemics, 2020. URL: `https://www.who.int/teams/risk-communication`.

[3] Fighting Disinformation - Official Sources on COVID-19 - Consilium, 2020. URL: `https://www.consilium.europa.eu/en/policies/covid-19-coronavirus-outbreak/fighting-disinformation`.

[4] WHO COVID-19 Database, 2020. URL: `https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov`.

[5] Clive E Adams, Alan A Montgomery, Tony Aburrow, Sophie Bloomfield, Paul M Briley, Ebun Carew, Suravi Chatterjee-Woolman, Ghalia Feddah, Johannes Friedel, Josh Gibbard, Euan Haynes, Mohsin Hussein, Mahesh Jayaram, Samuel Naylor, Luke Perry, Lena Schmidt, Umer Siddique, Ayla Serena Tabaksert, Douglas Taylor, Aarti Velani, Douglas White, and Jun Xia. Adding evidence of the effects of treatments into relevant Wikipedia pages: A randomised trial. *BMJ Open*, 10(2):e033655, February 2020. URL: `http://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2019-033655`, `doi:10.1136/bmjopen-2019-033655`.

[6] Wenceslao Arroyo-Machado, Daniel Torres-Salinas, Enrique Herrera-Viedma, and Esteban Romero-Frías. Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PLOS ONE*, 15(2):e0228713, February 2020. URL: `https://dx.plos.org/10.1371/journal.pone.0228713`, `doi:10.1371/journal.pone.0228713`.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[7] David M. Blei and John D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, 2007. URL: http://projecteuclid.org/euclid.aoas/1183143727, doi:10.1214/07-AOAS114.

[8] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL: http://dl.acm.org/citation.cfm?id=944937.

[9] Aleksandar Brezar and James Heilman. Readability of English Wikipedia's health information over time. *WikiJournal of Medicine*, 6(1):7, 2019. URL: https://en.wikiversity.org/wiki/WikiJournal_of_Medicine/Readability_of_English_Wikipedia's_health_information_over_time, doi:10.15347/wjm/2019.007.

[10] Chih-Chun Chen and Camille Roth. {{citation needed}}: the dynamics of referencing in Wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, Linz, Austria, 2012. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2462932.2462943, doi:10.1145/2462932.2462943.

[11] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *arXiv:2003.05004 [nlin, physics:physics]*, March 2020. arXiv: 2003.05004. URL: http://arxiv.org/abs/2003.05004.

[12] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level Representation Learning using Citation-informed Transformers. *arXiv:2004.07180 [cs]*, April 2020. arXiv: 2004.07180. URL: http://arxiv.org/abs/2004.07180.

[13] Giovanni Colavizza, Rodrigo Costas, Vincent A Traag, Nees Jan van Eck, Thed van Leeuwen, and Ludo Waltman. A scientometric overview of CORD-19. *bioRxiv*, April 2020. URL: https://www.biorxiv.org/content/10.1101/2020.04.20.046144v1, doi:10.1101/2020.04.20.046144.

[14] Zhichao Fang and Rodrigo Costas. Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics*, March 2020. URL: http://link.springer.com/10.1007/s11192-020-03405-9, doi:10.1007/s11192-020-03405-9.

[15] Andrea Forte, Nazanin Andalibi, Tim Gorichanaz, Meen Chul Kim, Thomas Park, and Aaron Halfaker. Information Fortification: An Online Citation Behavior. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork - GROUP '18*, pages 83–92, Sanibel Island, Florida, USA, 2018. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=3148330.3148347, doi:10.1145/3148330.3148347.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[16] Wikimedia Foundation. Responding to COVID-19. How we can help in this time of uncertainty, 2020. URL: https://wikimediafoundation.org/covid19.

[17] R. Stuart Geiger and Aaron Halfaker. When the levee breaks: without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration*, pages 1–6, Hong Kong, China, 2013. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2491055.2491061, doi:10.1145/2491055.2491061.

[18] Aaron Halfaker, Bahodir Mansurov, Miriam Redi, and Dario Taraborelli. Citations with identifiers in wikipedia, 2018. URL: https://figshare.com/articles/Citations_with_identifiers_in_Wikipedia/1299540/1, doi:10.6084/m9.figshare.1299540.

[19] James M Heilman, Eckhard Kemmann, Michael Bonert, Anwesh Chatterjee, Brent Ragar, Graham M Beards, David J Iberri, Matthew Harvey, Brendan Thomas, Wouter Stomp, Michael F Martone, Daniel J Lodge, Andrea Vondracek, Jacob F de Wolff, Casimir Liber, Samir C Grover, Tim J Vickers, Bertalan Meskó, and Michaël R Laurent. Wikipedia: A Key Tool for Global Public Health Promotion. *Journal of Medical Internet Research*, 13(1):e14, 2011. URL: http://www.jmir.org/2011/1/e14/, doi:10.2196/jmir.1589.

[20] Christian Herzog, Daniel Hook, and Stacy Konkiel. Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1):387–395, February 2020. URL: https://www.mitpressjournals.org/doi/abs/10.1162/qss_a_00020, doi:10.1162/qss_a_00020.

[21] John P.A. Ioannidis. Coronavirus disease 2019: The harms of exaggerated information and non-evidence-based measures. *European Journal of Clinical Investigation*, page e13222, March 2020. URL: http://doi.wiley.com/10.1111/eci.13222, doi:10.1111/eci.13222.

[22] Changwook Jung, Sun Geng, Meeyoung Cha, Inho Hong, and Diego Sáez-Trumper. Open data and COVID-19: Wikipedia as an informational resource during the pandemic, 2020. URL: https://medium.com/@diegosaeztrumper/open-data-and-covid-19-wikipedia-as-an-informational-resource-during-the-pandemic-dcca

[23] Brian Keegan, Darren Gergle, and Noshir Contractor. Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration - WikiSym '11*, Mountain View, California, 2011. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2038558.2038577, doi:10.1145/2038558.2038577.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[24] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, January 1963. URL: http://doi.wiley.com/10.1002/asi.5090140103, doi:10.1002/asi.5090140103.

[25] Kayvan Kousha and Mike Thelwall. Are wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, 68(3):762–779, 2017. URL: http://doi.wiley.com/10.1002/asi.23694, doi:10.1002/asi.23694.

[26] Kayvan Kousha and Mike Thelwall. COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts. 2020. URL: https://arxiv.org/abs/2004.10400.

[27] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, pages 591–602, Montréal, Québec, Canada, 2016. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2872427.2883085, doi:10.1145/2872427.2883085.

[28] M. R. Laurent and T. J. Vickers. Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, July 2009. URL: https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M3059, doi:10.1197/jamia.M3059.

[29] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM Press, 2019. URL: http://arxiv.org/abs/1812.00474, doi:10.1145/3289600.3291021.

[30] Włodzimierz Lewoniewski, Krzysztof Wecel, and Witold Abramowicz. Analysis of References Across Wikipedia Languages. In Robertas Damaševičius and Vilma Mikašytė, editors, *Information and Software Technologies*, volume 756, pages 561–573. Springer International Publishing, Cham, 2017. doi:10.1007/978-3-319-67642-5_47.

[31] Lauren A Maggio, Ryan M Steinberg, Tiziano Piccardi, and John M Willinsky. Reader engagement with medical content on Wikipedia. *eLife*, 9:e52426, March 2020. URL: https://elifesciences.org/articles/52426, doi:10.7554/eLife.52426.

[32] Lauren A Maggio, John M Willinsky, Ryan M Steinberg, Daniel Mietchen, Joseph L Wass, and Ting Dong. Wikipedia as a gateway to biomedical research: The relative distribution and use of citations in the English Wikipedia. *PLOS ONE*, 12(12):e0190046, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[33] Alberto Martín-Martín, Mike Thelwall, and Emilio Delgado López-Cózar. Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. 2020. URL: https://arxiv.org/abs/2004.14329.

[34] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245, 2015. URL: http://doi.wiley.com/10.1002/asi.23172, doi:10.1002/asi.23172.

[35] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, UK, 2011. ACM.

[36] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. 2019. arXiv:arXiv:1902.07669.

[37] Finn Årup Nielsen. Scientific Citations in Wikipedia. *First Monday*, 12, 2007.

[38] Finn Årup Nielsen, Daniel Mietchen, and Egon Willighagen. Scholia, Scientometrics and Wikidata. In Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Ławrynowicz, Fabio Ciravegna, and Olaf Hartig, editors, *The Semantic Web: ESWC 2017 Satellite Events*, volume 10577, pages 237–259. Springer International Publishing, Cham, 2017. URL: http://link.springer.com/10.1007/978-3-319-70407-4_36, doi:10.1007/978-3-319-70407-4_36.

[39] José Luis Ortega. Reliability and accuracy of altmetric providers: A comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics*, 116(3):2123–2138, September 2018. URL: http://link.springer.com/10.1007/s11192-018-2838-z, doi:10.1007/s11192-018-2838-z.

[40] Leena Paakkari and Orkan Okan. COVID-19: health literacy is an underestimated problem. *The Lancet Public Health*, 5(5):e249–e250, May 2020. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468266720300864, doi:10.1016/S2468-2667(20)30086-4.

[41] Antonio Perianes-Rodriguez, Ludo Waltman, and Nees Jan van Eck. Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10(4):1178–1195, November 2016. URL: http://linkinghub.elsevier.com/retrieve/pii/S1751157716302036, doi:10.1016/j.joi.2016.10.006.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[42] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. Quantifying Engagement with Citations on Wikipedia. In *Proceedings of The Web Conference 2020*, pages 2365–2376, Taipei Taiwan, April 2020. ACM. URL: https://dl.acm.org/doi/10.1145/3366423.3380300, doi:10.1145/3366423.3380300.

[43] Alessandro Piscopo and Elena Simperl. What we talk about when we talk about Wikidata quality: a literature survey. In *Proceedings of the 15th International Symposium on Open Collaboration*, Skövde, Sweden, 2019. ACM Press. doi:10.1145/3306446.3340822.

[44] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Conference on supporting group work*, Sanibel Island, Florida, USA, 2007. ACM Press. URL: http://portal.acm.org/citation.cfm?doid=1316624.1316663, doi:10.1145/1316624.1316663.

[45] Jason Priem, Heather A. Piwowar, and Bradley M. Hemminger. Altmetrics in the Wild: Using Social Media to Explore Scholarly Impact, 2012. URL: https://arxiv.org/html/1203.4745.

[46] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[47] Nicolás Robinson-García, Daniel Torres-Salinas, Zohreh Zahedi, and Rodrigo Costas. New data, new possibilities: Exploring the insides of Altmetric.com. *El Profesional de la Informacion*, 23(4):359–366, May 2014. URL: https://recyt.fecyt.es/index.php/EPI/article/view/epi.2014.jul.03, doi:10.3145/epi.2014.jul.03.

[48] Thomas Shafee, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Häggström, and James Heilman. Evolution of Wikipedia's medical content: past, present and future. *Journal of Epidemiology and Community Health*, pages jech–2016–208601, August 2017. URL: http://jech.bmj.com/lookup/doi/10.1136/jech-2016-208601, doi:10.1136/jech-2016-208601.

[49] Xin Shuai, Zhuoren Jiang, Xiaozhong Liu, and Johan Bollen. A comparative study of academic and Wikipedia ranking. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*, Indianapolis, Indiana, USA, 2013. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2467696.2467746, doi:10.1145/2467696.2467746.

[50] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on*

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

World Wide Web, pages 1591–1600, Perth, Australia, 2017. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=3038912.3052716, doi:10.1145/3038912.3052716.

[51] Denise A. Smith. Situating Wikipedia as a health information resource in various contexts: A scoping review. *PLOS ONE*, 15(2):e0228786, February 2020. URL: https://dx.plos.org/10.1371/journal.pone.0228786, doi:10.1371/journal.pone.0228786.

[52] Cassidy R. Sugimoto, Sam Work, Vincent Larivière, and Stefanie Haustein. Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9):2037–2062, 2017. URL: http://doi.wiley.com/10.1002/asi.23833, doi:10.1002/asi.23833.

[53] Briony Swire-Thompson and David Lazer. Public Health and Online Misinformation: Challenges and Recommendations. *Annual Review of Public Health*, 41(1):433–451, April 2020. URL: https://www.annualreviews.org/doi/10.1146/annurev-publhealth-040119-094127, doi:10.1146/annurev-publhealth-040119-094127.

[54] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, December 2006. URL: http://www.tandfonline.com/doi/abs/10.1198/016214506000000302, doi:10.1198/016214506000000302.

[55] Misha Teplitskiy, Grace Lu, and Eamon Duede. Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, 68(9):2116–2127, 2017. URL: http://doi.wiley.com/10.1002/asi.23687, doi:10.1002/asi.23687.

[56] Mike Thelwall. The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, 10(2):336–346, 2016. doi:10.1016/j.joi.2015.12.007.

[57] Mike Thelwall and Paul Wilson. Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, 8(4):963–971, 2014. doi:10.1016/j.joi.2014.09.011.

[58] Daniel Torres-Salinas, Esteban Romero-Frías, and Wenceslao Arroyo-Machado. Mapping the backbone of the Humanities through the eyes of Wikipedia. *Journal of Informetrics*, 13(3):793–803, 2019. URL: https://linkinghub.elsevier.com/retrieve/pii/S1751157718302955, doi:10.1016/j.joi.2019.07.002.

[59] Vincent A. Traag, Paul Van Dooren, and Yurii Nesterov. Narrow scope for resolution-limit-free community detection. *Physical Review E*,

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

84(1):016114, 2011. URL: http://journals.aps.org/pre/abstract/10.1103/PhysRevE.84.016114.

[60] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, December 2019. URL: http://www.nature.com/articles/s41598-019-41695-z, doi:10.1038/s41598-019-41695-z.

[61] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The Covid-19 Open Research Dataset. *arXiv:2004.10706 [cs]*, April 2020. arXiv: 2004.10706. URL: http://arxiv.org/abs/2004.10706.

[62] Bo Xie, Daqing He, Tim Mercer, Youfa Wang, Dan Wu, Kenneth R. Fleischmann, Yan Zhang, Linda H. Yoder, Keri K. Stephens, Michael Mackert, and Min K. Lee. Global health crises are also information crises: A call to action. *Journal of the Association for Information Science and Technology*, March 2020. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24357, doi:10.1002/asi.24357.

[63] Zohreh Zahedi, Rodrigo Costas, and Paul Wouters. How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, 101(2):1491–1513, 2014. URL: http://link.springer.com/10.1007/s11192-014-1264-0, doi:10.1007/s11192-014-1264-0.

[64] John Zarocostas. How to fight an infodemic. *Lancet*, 395(10225), February 2020. URL: https://linkinghub.elsevier.com/retrieve/pii/S014067362030461X, doi:10.1016/S0140-6736(20)30461-X.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SI

Topics

Refer to Figures 9 and 10 for topic and macrotopic intensities over time. See Figure 12 for the topic clustering which informs their grouping into macrotopics. The macrotopic is given next to the topic number, for reference.

- **Topic #0, Molecular biology**: "protein", "domain", "membrane", "structure", "binding", "receptor", "site", "fusion", "bind", "protease", "activity", "interaction", "acid", "glycoprotein", "complex", "ace2", "residue", "form", "entry", "cleavage".
- **Topic #1, Transmission**: "sequence", "virus", "strain", "gene", "calf", "analysis", "isolate", "genome", "specie", "bat", "human", "genetic", "region", "identify", "ibv", "mutation", "host", "study", "variant", "different".
- **Topic #2, Coronaviruses**: "respiratory", "infection", "study", "year", "case", "child", "age", "mers-cov", "risk", "associate", "high", "associated with", "patient", "Middle", "rate", "factor", "illness", "95_ci".
- **Topic #3, Immunology**: "drug", "virus", "antiviral", "human", "activity", "target", "potential", "therapeutic", "new", "treatment", "viral", "compound", "review", "inhibitor", "study", "novel", "development", "host", "include", "infection".
- **Topic #4, Coronaviruses**: "covid-19", "patient", "COVID-19", "case", "coronavirus", "sars-cov-2", "2019", "2020", "China", "clinical", "novel", "Wuhan", "disease", "severe", "confirm", "report", "day", "pneumonia", "2019-ncov", "symptom".
- **Topic #5, Molecular biology**: "rna", "virus", "viral", "replication", "protein", "gene", "mrna", "expression", "synthesis", "cell", "genome", "transcription", "host", "translation", "hepatitis", "cellular", "viral_rna", "hcv", "subgenomic".
- **Topic #6, Respiratory diseases**: "virus", "influenza", "detection", "test", "viral", "sample", "detect", "assay", "respiratory", "method", "influenza_virus", "pcr", "result", "positive", "infection", "diagnostic", "human", "clinical", "specimen", "rsv".
- **Topic #7, Transmission**: "virus", "diarrhea", "animal", "pig", "rotavirus", "cat", "sample", "serum", "porcine", "infection", "day", "detect", "pedv", "bovine", "feline", "dog", "disease", "intestinal", "swine", "antibody".
- **Topic #8, Molecular biology**: "cell", "infection", "mouse", "response", "expression", "immune", "virus", "induce", "role", "type", "viral", "result", "increase", "cytokine", "level", "receptor", "study", "human", "activation", "disease".

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- **Topic #9, Public health and epidemics**: "model", "datum", "epidemic", "number", "case", "spread", "transmission", "disease", "time", "outbreak", "rate", "measure", "estimate", "population", "analysis", "result", "method", "different", "base", "control".
- **Topic #10, Immunology**: "vaccine", "antibody", "response", "antigen", "immune", "epitope", "mouse", "vaccination", "virus", "challenge", "neutralize", "monoclonal", "human", "development", "monoclonal_antibody", "mab", "immunity", "protection", "induce", "titer".
- **Topic #11, Clinical medicine**: "study", "group", "effect", "result", "high", "level", "method", "treatment", "increase", "compare", "significantly", "control", "low", "conclusion", "day", "concentration", "significant", "evaluate", "reduce", "difference".
- **Topic #12, Public health and epidemics**: "health", "disease", "public", "pandemic", "public_health", "outbreak", "care", "system", "Health", "risk", "country", "research", "covid-19", "global", "need", "response", "infectious", "provide", "public health", "information".
- **Topic #13, Clinical medicine**: "infection", "patient", "respiratory", "disease", "acute", "clinical", "cause", "lung", "associate", "viral", "pneumonia", "severe", "treatment", "associated with", "child", "tract", "bacterial", "common", "pulmonary", "chronic".
- **Topic #14, Transmission**: "protein", "coronavirus", "cell", "virus", "respiratory", "acute", "syndrome", "sars-cov", "severe", "SARS", "respiratory_syndrome", "severe_acute", "spike", "human", "mhv", "coronaviruse", "recombinant", "mouse", "culture", "express".

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Extra tables and figures

Table 1: Top-20 citing Wikipedia articles.

# citations	Wikipedia id	Wikipedia article title	Lang
127	63030231	Coronavirus disease 2019	en
54	62786585	Severe acute respiratory syndrome coronavirus 2	en
49	201983	Coronavirus	en
47	62750956	2019–20 Wuhan coronavirus outbreak	en
40	63676463	2019–20 coronavirus pandemic	en
31	211547	Severe acute respiratory syndrome-related coro...	en
27	63435931	COVID-19 drug development	en
27	39532251	Middle East respiratory syndrome	en
22	19572217	Influenza	en
21	4354646	Emergent virus	en
20	63430824	COVID-19 drug repurposing research	en
20	63319438	COVID-19 vaccine	en
18	22693252	Feline coronavirus	en
17	2717089	Angiotensin-converting enzyme 2	en
17	29802394	Social history of viruses	en
14	196741	Severe acute respiratory syndrome	en
14	477498	Human metapneumovirus	en
13	2925242	Severe acute respiratory syndrome coronavirus	en
13	22677497	Social distancing	en
12	854589	Coronavirus	sv

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2: Top-20 cited journal articles. The first column gives the number of distinct citing Wikipedia articles, while the last one gives the number of citations to these articles from the scientific literature (data from Dimensions).

# citations	DOI	Title	Publication year	Journal	Times cited
18	10.1007/s00705-012-1299-6	Ratification vote on taxonomic proposals...	2012	Arch Virol	207
15	10.3390/v2081803	Coronavirus Genomics and Bioinformatics Analysis	2010	Viruses	100
13	10.1007/978-1-4939-2438-7_1	Coronaviruses: An Overview of Their Replication...	2015	Methods in Molecular Biology	146
12	10.1016/s0140-6736(20)30183-5	Clinical features of patients infected...	2020	The Lancet	1706
10	10.1056/nejmoa2001191	First Case of 2019 Novel Coronavirus...	2020	New England Journal of Medicine	417
10	10.1083/jcb.148.5.931	Pex19 Binds Multiple Peroxisomal Membrane Proteins...	2000	Journal of Cell Biology	NaN
10	10.3390/v11020174	Global Epidemiology of Bat Coronaviruses	2019	Viruses	25
9	10.1128/jvi.06540-11	Discovery of Seven Novel Mammalian and Avian...	2012	Journal of Virology	384
9	10.1038/d41586-020-00548-w	Mystery deepens over animal source of coronavirus	2020	Nature	4
9	10.1056/nejmoa2001316	Early Transmission Dynamics in Wuhan, China...	2020	New England Journal of Medicine	935
9	10.1038/s41586-020-2012-7	A pneumonia outbreak associated with...	2020	Nature	651
9	10.1016/s0140-6736(20)30567-5	How will country-based mitigation measures...	2020	The Lancet	67
9	10.1038/s41422-020-0282-0	Remdesivir and chloroquine effectively...	2020	Cell Research	294
8	10.1016/j.pnpbp.2006.01.008	Human brain evolution and the “Neuroevolutiona...	2006	Progress in Neuro-Psychopharmacology...	61
8	10.1038/s41591-020-0820-9	The proximal origin of SARS-CoV-2	2020	Nature Medicine	68
8	10.1001/jama.2016.17324	Prevalence of Depression, Depressive Symptoms...	2016	JAMA	346
8	10.1086/511159	Infectious Diseases Society of America...	2007	Clin Infect Dis	3884
7	10.1056/nejmoa2001017	A Novel Coronavirus from Patients with Pneumonia...	2020	New England Journal of Medicine	1093
7	10.1016/s0196-6553(98)70046-x	Characterization of infectious aerosols...	1998	American Journal of Infection Control	119
7	10.1186/1743-422x-7-52	Origin of measles virus: divergence...	2010	Virol J	90

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Regression tables

Model:	Logistic regression 2020
Dep. Variable:	in_wikipedia
Method:	dydx
No. Observations:	57,757
Pseudo R-squ.:	0.2613

Variable	dy/dx	std err	z	P> |z|	[0.025	0.975]
C(publication_year, Treatment(2020))[T.2000.0]	-0.0244	0.013	-1.930	0.054	-0.049	0.000
C(publication_year, Treatment(2020))[T.2001.0]	-0.0295	0.011	-2.751	0.006	-0.051	-0.008
C(publication_year, Treatment(2020))[T.2002.0]	-0.0269	0.011	-2.538	0.011	-0.048	-0.006
C(publication_year, Treatment(2020))[T.2003.0]	-0.0241	0.006	-3.723	0.000	-0.037	-0.011
C(publication_year, Treatment(2020))[T.2004.0]	-0.0263	0.005	-4.924	0.000	-0.037	-0.016
C(publication_year, Treatment(2020))[T.2005.0]	-0.0168	0.005	-3.264	0.001	-0.027	-0.007
C(publication_year, Treatment(2020))[T.2006.0]	-0.0190	0.005	-3.796	0.000	-0.029	-0.009
C(publication_year, Treatment(2020))[T.2007.0]	-0.0205	0.005	-4.005	0.000	-0.031	-0.010
C(publication_year, Treatment(2020))[T.2008.0]	-0.0192	0.005	-3.862	0.000	-0.029	-0.009
C(publication_year, Treatment(2020))[T.2009.0]	-0.0253	0.005	-5.049	0.000	-0.035	-0.015
C(publication_year, Treatment(2020))[T.2010.0]	-0.0279	0.005	-5.532	0.000	-0.038	-0.018
C(publication_year, Treatment(2020))[T.2011.0]	-0.0315	0.005	-6.392	0.000	-0.041	-0.022
C(publication_year, Treatment(2020))[T.2012.0]	-0.0327	0.005	-6.816	0.000	-0.042	-0.023
C(publication_year, Treatment(2020))[T.2013.0]	-0.0395	0.005	-8.416	0.000	-0.049	-0.030
C(publication_year, Treatment(2020))[T.2014.0]	-0.0402	0.005	-8.887	0.000	-0.049	-0.031
C(publication_year, Treatment(2020))[T.2015.0]	-0.0468	0.005	-10.130	0.000	-0.056	-0.038
C(publication_year, Treatment(2020))[T.2016.0]	-0.0459	0.005	-10.001	0.000	-0.055	-0.037
C(publication_year, Treatment(2020))[T.2017.0]	-0.0494	0.005	-10.306	0.000	-0.059	-0.040
C(publication_year, Treatment(2020))[T.2018.0]	-0.0424	0.005	-9.275	0.000	-0.051	-0.033
C(publication_year, Treatment(2020))[T.2019.0]	-0.0332	0.005	-6.978	0.000	-0.043	-0.024
C(top_j, Treatment('OTHER'))[T.Antiviral_Research]	0.0172	0.007	2.631	0.009	0.004	0.030
C(top_j, Treatment('OTHER'))[T.Arch_Virol]	0.0021	0.011	0.189	0.850	-0.019	0.023
C(top_j, Treatment('OTHER'))[T.Emerg_Infect_Dis]	-0.0048	0.005	-0.984	0.325	-0.014	0.005
C(top_j, Treatment('OTHER'))[T.Journal_of_Clinical_Virology]	-0.0008	0.012	-0.071	0.943	-0.024	0.022
C(top_j, Treatment('OTHER'))[T.Journal_of_Virological_Methods]	-0.0184	0.016	-1.176	0.240	-0.049	0.012
C(top_j, Treatment('OTHER'))[T.Journal_of_Virology]	-0.0027	0.003	-0.790	0.429	-0.009	0.004
C(top_j, Treatment('OTHER'))[T.PLoS_One]	-0.0081	0.005	-1.730	0.084	-0.017	0.001
C(top_j, Treatment('OTHER'))[T.PLoS_Pathog]	0.0047	0.005	0.859	0.390	-0.006	0.015
C(top_j, Treatment('OTHER'))[T.Research_Square]	-1.5120	4.06e+09	-3.72e-10	1.000	-7.96e+09	7.96e+09
C(top_j, Treatment('OTHER'))[T.SSRN_Electronic_Journal]	-0.0366	0.017	-2.113	0.035	-0.071	-0.003
C(top_j, Treatment('OTHER'))[T.Sci_Rep]	-0.0058	0.009	-0.662	0.508	-0.023	0.011
C(top_j, Treatment('OTHER'))[T.The_Lancet]	0.0112	0.005	2.378	0.017	0.002	0.020
C(top_j, Treatment('OTHER'))[T.The_Lancet_Infectious_Diseases]	0.0004	0.007	0.064	0.949	-0.013	0.014
C(top_j, Treatment('OTHER'))[T.Vaccine]	-0.0267	0.011	-2.339	0.019	-0.049	-0.004
C(top_j, Treatment('OTHER'))[T.Veterinary_Microbiology]	-0.0189	0.013	-1.402	0.161	-0.045	0.008
C(top_j, Treatment('OTHER'))[T.Virology]	-0.0033	0.006	-0.512	0.609	-0.016	0.009
C(top_j, Treatment('OTHER'))[T.Virus_Research]	0.0036	0.008	0.473	0.637	-0.011	0.019
C(top_j, Treatment('OTHER'))[T.Viruses]	0.0162	0.006	2.934	0.003	0.005	0.027
C(top_j, Treatment('OTHER'))[T.bioRxiv]	-0.0107	0.008	-1.415	0.157	-0.026	0.004
C(top_j, Treatment('OTHER'))[T.medRxiv]	-0.0800	0.015	-5.353	0.000	-0.109	-0.051
times_cited	0.0113	0.001	14.170	0.000	0.010	0.013
counts_mendeley	0.0122	0.001	18.476	0.000	0.011	0.014
counts_policy	-0.0006	0.002	-0.297	0.766	-0.005	0.003
counts_twitter_unique	0.0083	0.001	12.545	0.000	0.007	0.010
counts_blogs_news	0.0085	0.001	9.582	0.000	0.007	0.010
counts_facebook	-0.0007	0.002	-0.470	0.639	-0.004	0.002
expert_ratio	-0.0093	0.003	-2.733	0.006	-0.016	-0.003
tm_coronaviruses	0.0130	0.037	0.349	0.727	-0.060	0.086
tm_phe	0.0165	0.037	0.446	0.656	-0.056	0.089
tm_transmission	0.0335	0.037	0.906	0.365	-0.039	0.106
tm_molecular_biology	0.0205	0.037	0.556	0.578	-0.052	0.093
tm_respiratory_diseases	-0.0064	0.037	-0.173	0.863	-0.079	0.066
tm_immunology	0.0334	0.037	0.908	0.364	-0.039	0.106
tm_clinical_medicine	0.0144	0.037	0.392	0.695	-0.058	0.087
spectre_cluster_size	0.0141	0.004	3.506	0.000	0.006	0.022
network_cluster_size	0.0004	0.000	1.242	0.214	-0.000	0.001

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Model:	Logistic regression 2019
Dep. Variable:	in_wikipedia
Method:	dydx
No. Observations:	40,999
Pseudo R-squ.:	0.2631

Variable	dy/dx	std err	z	P> |z|	[0.025	0.975]
C(publication_year, Treatment(2019))[T.2000.0]	0.0488	0.014	3.468	0.001	0.021	0.076
C(publication_year, Treatment(2019))[T.2001.0]	0.0451	0.012	3.674	0.000	0.021	0.069
C(publication_year, Treatment(2019))[T.2002.0]	0.0380	0.013	2.827	0.005	0.012	0.064
C(publication_year, Treatment(2019))[T.2003.0]	0.0488	0.009	5.275	0.000	0.031	0.067
C(publication_year, Treatment(2019))[T.2004.0]	0.0464	0.008	5.544	0.000	0.030	0.063
C(publication_year, Treatment(2019))[T.2005.0]	0.0562	0.008	6.791	0.000	0.040	0.072
C(publication_year, Treatment(2019))[T.2006.0]	0.0507	0.008	6.210	0.000	0.035	0.067
C(publication_year, Treatment(2019))[T.2007.0]	0.0504	0.008	6.145	0.000	0.034	0.066
C(publication_year, Treatment(2019))[T.2008.0]	0.0491	0.008	6.045	0.000	0.033	0.065
C(publication_year, Treatment(2019))[T.2009.0]	0.0434	0.008	5.380	0.000	0.028	0.059
C(publication_year, Treatment(2019))[T.2010.0]	0.0359	0.008	4.388	0.000	0.020	0.052
C(publication_year, Treatment(2019))[T.2011.0]	0.0342	0.008	4.287	0.000	0.019	0.050
C(publication_year, Treatment(2019))[T.2012.0]	0.0359	0.008	4.600	0.000	0.021	0.051
C(publication_year, Treatment(2019))[T.2013.0]	0.0282	0.008	3.620	0.000	0.013	0.043
C(publication_year, Treatment(2019))[T.2014.0]	0.0306	0.008	3.999	0.000	0.016	0.046
C(publication_year, Treatment(2019))[T.2015.0]	0.0209	0.008	2.699	0.007	0.006	0.036
C(publication_year, Treatment(2019))[T.2016.0]	0.0200	0.008	2.582	0.010	0.005	0.035
C(publication_year, Treatment(2019))[T.2017.0]	0.0132	0.008	1.649	0.099	-0.002	0.029
C(publication_year, Treatment(2019))[T.2018.0]	0.0217	0.008	2.767	0.006	0.006	0.037
C(top_j, Treatment('OTHER'))[T.Antiviral_Research]	0.0061	0.008	0.729	0.466	-0.010	0.023
C(top_j, Treatment('OTHER'))[T.Arch_Virol]	-0.0066	0.013	-0.503	0.615	-0.032	0.019
C(top_j, Treatment('OTHER'))[T.Emerg_Infect_Dis]	1.411e-05	0.005	0.003	0.998	-0.010	0.010
C(top_j, Treatment('OTHER'))[T.Journal_of_Clinical_Virology]	0.0060	0.011	0.532	0.595	-0.016	0.028
C(top_j, Treatment('OTHER'))[T.Journal_of_Virological_Methods]	-0.0237	0.017	-1.422	0.155	-0.056	0.009
C(top_j, Treatment('OTHER'))[T.Journal_of_Virology]	-0.0022	0.004	-0.606	0.544	-0.009	0.005
C(top_j, Treatment('OTHER'))[T.PLoS_One]	-0.0133	0.005	-2.540	0.011	-0.024	-0.003
C(top_j, Treatment('OTHER'))[T.PLoS_Pathog]	0.0067	0.006	1.187	0.235	-0.004	0.018
C(top_j, Treatment('OTHER'))[T.Sci_Rep]	-0.0015	0.010	-0.160	0.873	-0.021	0.017
C(top_j, Treatment('OTHER'))[T.The_Lancet]	0.0050	0.006	0.817	0.414	-0.007	0.017
C(top_j, Treatment('OTHER'))[T.The_Lancet_Infectious_Diseases]	-0.0009	0.009	-0.101	0.920	-0.018	0.016
C(top_j, Treatment('OTHER'))[T.Vaccine]	-0.0268	0.013	-2.101	0.036	-0.052	-0.002
C(top_j, Treatment('OTHER'))[T.Veterinary_Microbiology]	-0.0132	0.013	-1.019	0.308	-0.039	0.012
C(top_j, Treatment('OTHER'))[T.Virology]	-0.0055	0.007	-0.784	0.433	-0.019	0.008
C(top_j, Treatment('OTHER'))[T.Virus_Research]	-0.0013	0.009	-0.144	0.885	-0.019	0.016
C(top_j, Treatment('OTHER'))[T.Viruses]	0.0121	0.007	1.836	0.066	-0.001	0.025
times_cited	0.0043	0.001	4.122	0.000	0.002	0.006
counts_mendeley	0.0212	0.001	19.705	0.000	0.019	0.023
counts_policy	0.0023	0.002	1.062	0.288	-0.002	0.007
counts_twitter_unique	0.0018	0.001	1.937	0.053	-2.11e-05	0.004
counts_blogs_news	0.0066	0.001	5.580	0.000	0.004	0.009
counts_facebook	0.0029	0.002	1.546	0.122	-0.001	0.007
expert_ratio	-0.0003	0.004	-0.077	0.938	-0.007	0.007
tm_coronaviruses	-0.0529	0.055	-0.955	0.340	-0.161	0.056
tm_phe	-0.0463	0.055	-0.842	0.400	-0.154	0.061
tm_transmission	-0.0195	0.055	-0.355	0.723	-0.127	0.088
tm_molecular_biology	-0.0318	0.055	-0.579	0.562	-0.139	0.076
tm_respiratory_diseases	-0.0512	0.055	-0.931	0.352	-0.159	0.057
tm_immunology	-0.0361	0.055	-0.659	0.510	-0.143	0.071
tm_clinical_medicine	-0.0374	0.055	-0.684	0.494	-0.145	0.070
spectre_cluster_size	0.0121	0.005	2.633	0.008	0.003	0.021
network_cluster_size	-0.0011	0.000	-2.548	0.011	-0.002	-0.000

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Model:	OLS regression 2020
Dep. Variable:	counts_wikipedia
Method:	OLS
No. Observations:	57,757
R-squ.:	0.138

Variable	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	-0.0238	0.033	-0.717	0.473	-0.089	0.041
C(publication_year, Treatment(2020))[T.2000.0]	-0.0333	0.013	-2.566	0.010	-0.059	-0.008
C(publication_year, Treatment(2020))[T.2001.0]	-0.0380	0.011	-3.589	0.000	-0.059	-0.017
C(publication_year, Treatment(2020))[T.2002.0]	-0.0385	0.011	-3.562	0.000	-0.060	-0.017
C(publication_year, Treatment(2020))[T.2003.0]	-0.0213	0.006	-3.658	0.000	-0.033	-0.010
C(publication_year, Treatment(2020))[T.2004.0]	-0.0263	0.005	-5.511	0.000	-0.036	-0.017
C(publication_year, Treatment(2020))[T.2005.0]	-0.0144	0.005	-3.013	0.003	-0.024	-0.005
C(publication_year, Treatment(2020))[T.2006.0]	-0.0167	0.004	-3.719	0.000	-0.026	-0.008
C(publication_year, Treatment(2020))[T.2007.0]	-0.0182	0.005	-3.934	0.000	-0.027	-0.009
C(publication_year, Treatment(2020))[T.2008.0]	-0.0167	0.004	-3.723	0.000	-0.025	-0.008
C(publication_year, Treatment(2020))[T.2009.0]	-0.0234	0.004	-5.366	0.000	-0.032	-0.015
C(publication_year, Treatment(2020))[T.2010.0]	-0.0301	0.004	-6.834	0.000	-0.039	-0.021
C(publication_year, Treatment(2020))[T.2011.0]	-0.0295	0.004	-6.912	0.000	-0.038	-0.021
C(publication_year, Treatment(2020))[T.2012.0]	-0.0283	0.004	-6.908	0.000	-0.036	-0.020
C(publication_year, Treatment(2020))[T.2013.0]	-0.0407	0.004	-10.355	0.000	-0.048	-0.033
C(publication_year, Treatment(2020))[T.2014.0]	-0.0407	0.004	-10.774	0.000	-0.048	-0.033
C(publication_year, Treatment(2020))[T.2015.0]	-0.0474	0.004	-13.031	0.000	-0.055	-0.040
C(publication_year, Treatment(2020))[T.2016.0]	-0.0443	0.004	-12.553	0.000	-0.051	-0.037
C(publication_year, Treatment(2020))[T.2017.0]	-0.0451	0.004	-12.866	0.000	-0.052	-0.038
C(publication_year, Treatment(2020))[T.2018.0]	-0.0342	0.003	-10.075	0.000	-0.041	-0.028
C(publication_year, Treatment(2020))[T.2019.0]	-0.0180	0.003	-5.409	0.000	-0.024	-0.011
C(top_j, Treatment('OTHER'))[T.Antiviral_Research]	0.0261	0.009	2.991	0.003	0.009	0.043
C(top_j, Treatment('OTHER'))[T.Arch_Virol]	0.0042	0.009	0.453	0.650	-0.014	0.023
C(top_j, Treatment('OTHER'))[T.Emerg_Infect_Dis]	-0.0264	0.006	-4.284	0.000	-0.039	-0.014
C(top_j, Treatment('OTHER'))[T.Journal_of_Clinical_Virology]	0.0080	0.008	1.060	0.289	-0.007	0.023
C(top_j, Treatment('OTHER'))[T.Journal_of_Virological_Methods]	-0.0083	0.009	-0.932	0.351	-0.026	0.009
C(top_j, Treatment('OTHER'))[T.Journal_of_Virology]	-0.0068	0.004	-1.635	0.102	-0.015	0.001
C(top_j, Treatment('OTHER'))[T.PLoS_One]	-0.0212	0.004	-5.024	0.000	-0.029	-0.013
C(top_j, Treatment('OTHER'))[T.PLoS_Pathog]	0.0092	0.009	1.059	0.289	-0.008	0.026
C(top_j, Treatment('OTHER'))[T.Research_Square]	0.0077	0.009	0.844	0.399	-0.010	0.025
C(top_j, Treatment('OTHER'))[T.SSRN_Electronic_Journal]	0.0022	0.005	0.423	0.672	-0.008	0.013
C(top_j, Treatment('OTHER'))[T.Sci_Rep]	-0.0138	0.007	-1.851	0.064	-0.028	0.001
C(top_j, Treatment('OTHER'))[T.The_Lancet]	0.0363	0.007	5.163	0.000	0.023	0.050
C(top_j, Treatment('OTHER'))[T.The_Lancet_Infectious_Diseases]	0.0059	0.009	0.695	0.487	-0.011	0.023
C(top_j, Treatment('OTHER'))[T.Vaccine]	-0.0358	0.008	-4.282	0.000	-0.052	-0.019
C(top_j, Treatment('OTHER'))[T.Veterinary_Microbiology]	-0.0249	0.009	-2.648	0.008	-0.043	-0.006
C(top_j, Treatment('OTHER'))[T.Virology]	-0.0123	0.007	-1.813	0.070	-0.026	0.001
C(top_j, Treatment('OTHER'))[T.Virus_Research]	-0.0028	0.008	-0.355	0.723	-0.019	0.013
C(top_j, Treatment('OTHER'))[T.Viruses]	0.0268	0.007	3.949	0.000	0.013	0.040
C(top_j, Treatment('OTHER'))[T.bioRxiv]	-0.0140	0.008	-1.764	0.078	-0.030	0.002
C(top_j, Treatment('OTHER'))[T.medRxiv]	-0.0327	0.004	-7.327	0.000	-0.041	-0.024
times_cited	0.0154	0.001	21.657	0.000	0.014	0.017
counts_mendeley	0.0094	0.001	18.200	0.000	0.008	0.010
counts_policy	0.0611	0.003	18.033	0.000	0.054	0.068
counts_twitter_unique	0.0046	0.001	6.341	0.000	0.003	0.006
counts_blogs_news	0.0427	0.001	32.416	0.000	0.040	0.045
counts_facebook	0.0409	0.002	16.842	0.000	0.036	0.046
expert_ratio	-0.0244	0.003	-7.550	0.000	-0.031	-0.018
tm_coronaviruses	-0.0258	0.027	-0.956	0.339	-0.079	0.027
tm_phe	-0.0194	0.027	-0.730	0.466	-0.072	0.033
tm_transmission	0.0074	0.027	0.281	0.779	-0.045	0.059
tm_molecular_biology	-0.0193	0.026	-0.728	0.467	-0.071	0.033
tm_respiratory_diseases	-0.0420	0.027	-1.566	0.117	-0.095	0.011
tm_immunology	0.0019	0.027	0.071	0.943	-0.050	0.054
tm_clinical_medicine	-0.0177	0.027	-0.669	0.503	-0.070	0.034
spectre_cluster_size	0.0048	0.003	1.806	0.071	-0.000	0.010
network_cluster_size	-0.0005	0.000	-1.936	0.053	-0.001	6.3e-06

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3: Test statistics for macrotopic intensities of articles cited in Wikipedia or not, limited to articles *published before 2020*. In W: cited in Wikipedia; Not in W: not cited in Wikipedia; KWH: Kruskal–Wallis H test.

Macrotopic	**In W**		**Not in W**		**Test**	**p-value**	**Effect size**
	Mean	SD	Mean	SD	KWH	KWH	Cohen's d
Coronaviruses	0.076	0.143	0.090	0.163	7.637	**0.006**	0.084
Public health and epidemics	0.187	0.256	0.177	0.265	26.911	**0.000**	0.038
Transmission	0.168	0.206	0.163	0.210	3.297	**0.069**	0.022
Molecular biology	0.226	0.268	0.211	0.265	6.397	**0.011**	0.056
Respiratory diseases	0.051	0.114	0.071	0.145	28.787	**0.000**	0.144
Immunology	0.176	0.205	0.142	0.190	67.383	**0.000**	0.178
Clinical medicine	0.106	0.169	0.133	0.190	39.684	**0.000**	0.140

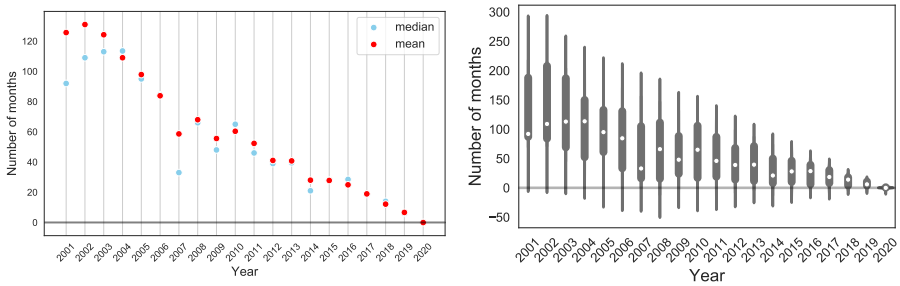
Table 4: Test statistics for macrotopic intensities of articles cited in Wikipedia or not, limited to articles *published in 2020*. In W: cited in Wikipedia; Not in W: not cited in Wikipedia; KWH: Kruskal–Wallis H test.

Macrotopic	**In W**		**Not in W**		**Test**	**p-value**	**Effect size**
	Mean	SD	Mean	SD	KWH	KWH	Cohen's d
Coronaviruses	0.367	0.252	0.296	0.258	36.880	**0.000**	0.274
Public health and epidemics	0.255	0.254	0.368	0.312	44.686	**0.000**	0.366
Transmission	0.084	0.138	0.052	0.110	25.407	**0.000**	0.291
Molecular biology	0.054	0.104	0.067	0.137	1.154	0.283	0.094
Respiratory diseases	0.025	0.068	0.028	0.084	0.041	0.840	0.044
Immunology	0.100	0.166	0.067	0.138	11.717	**0.001**	0.241
Clinical medicine	0.093	0.144	0.095	0.145	0.427	0.513	0.008

Table 5: Test statistics for macrotopic intensities of articles cited in Wikipedia or not; *all publications*. In W: cited in Wikipedia; Not in W: not cited in Wikipedia; KWH: Kruskal–Wallis H test.

Macrotopic	**In W**		**Not in W**		**Test**	**p-value**	**Effect size**
	Mean	SD	Mean	SD	KWH	KWH	Cohen's d
Coronaviruses	0.130	0.203	0.151	0.217	26.649	**0.000**	0.098
Public health and epidemics	0.200	0.257	0.234	0.293	2.674	0.102	0.118
Transmission	0.152	0.198	0.13	0.193	50.273	**0.000**	0.115
Molecular biology	0.194	0.255	0.168	0.244	24.151	**0.000**	0.106
Respiratory diseases	0.046	0.108	0.059	0.131	10.660	**0.001**	0.097
Immunology	0.162	0.201	0.120	0.179	135.070	**0.000**	0.235
Clinical medicine	0.104	0.164	0.121	0.179	26.859	**0.000**	0.098

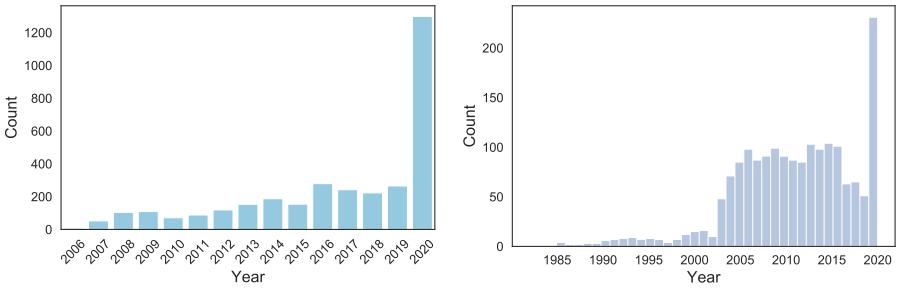
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) Mean and median.

(b) Violin plot.

Figure 7: Number of months elapsed from publication to the first Wikipedia citation. Alternative views on Figure 1.



(a) Number of citations from Wikipedia to COVID-19 literature, per year, overall.

(b) Publication year of the articles cited from Wikipedia, overall.

Figure 8: Timing of new citations from Wikipedia, and publication years of the articles they refer to.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

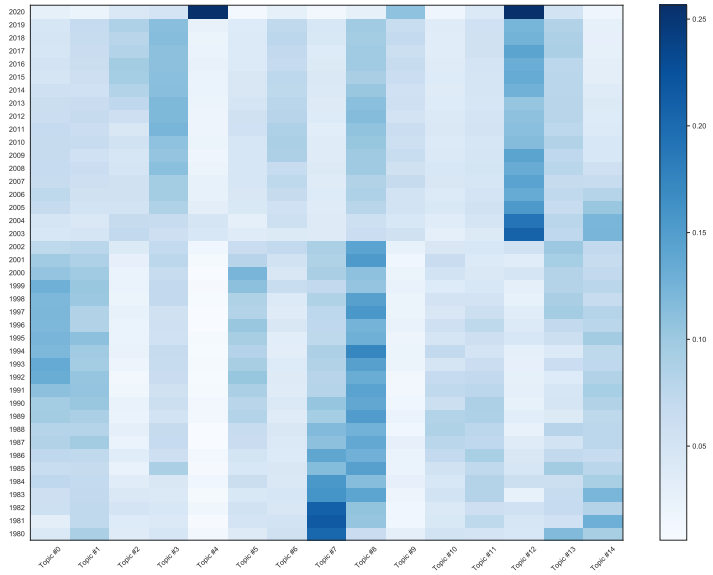
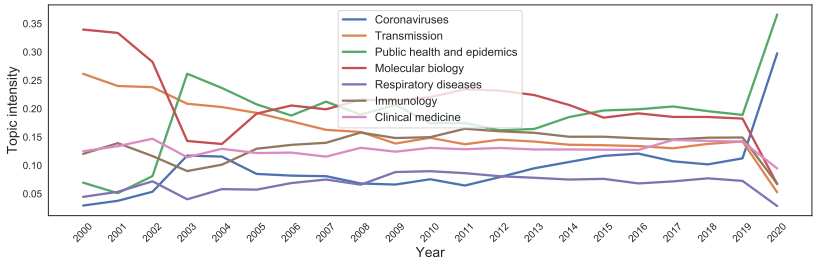
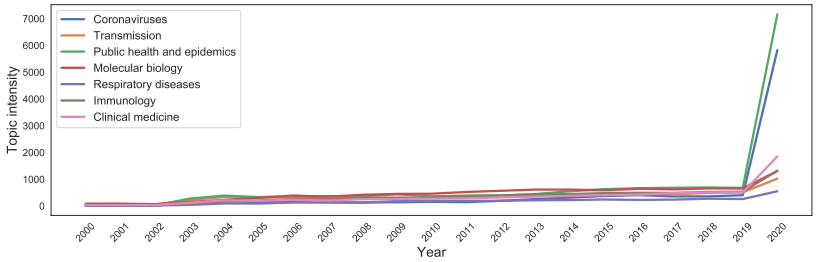


Figure 9: Heatmap of topic intensities over time.



(a) Average aggregate; this can be interpreted as the average topic intensity.



(b) Cumulative aggregate; this can be interpreted as the number of papers per topic.

Figure 10: Macrotopic intensities over time.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

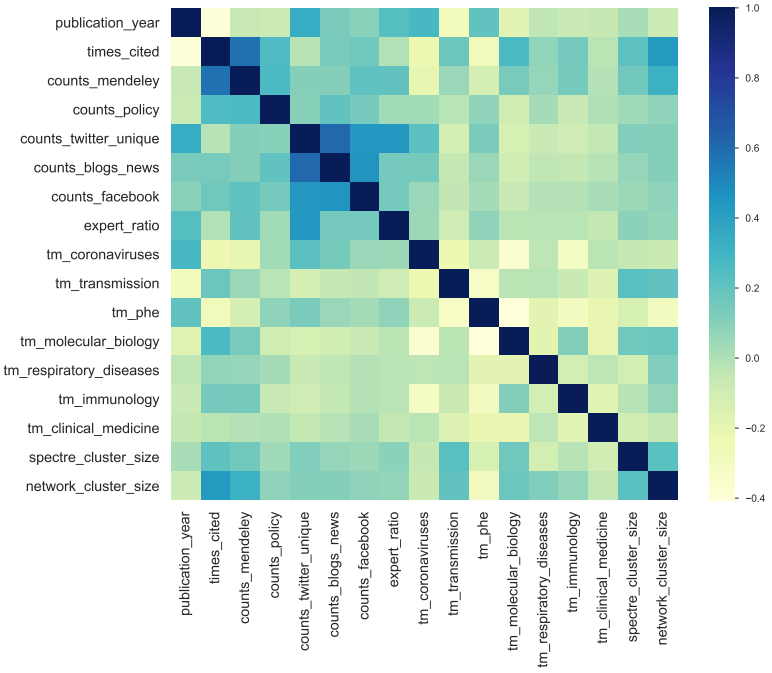


Figure 11: Heatmap of regression variables correlations (Pearson's).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

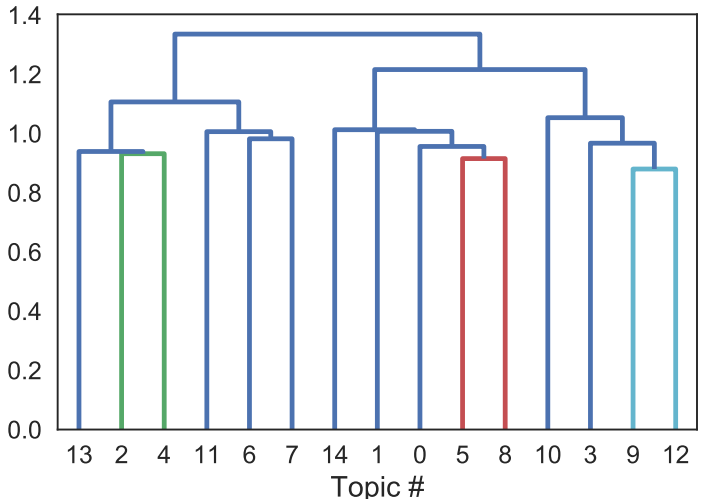


Figure 12: Agglomerative clustering dendrogram over topics, based on Jensen-Shannon distances. Considering a cut at 1.1, the left-most cluster (topics 2,4,13) focuses on coronaviruses and related clinical medicine; next is a cluster with topics related to other respiratory diseases (pneumonia, influenza), their transmission and treatment (topics 6,7,11); next is a cluster on molecular biology studies on viruses and their transmission, in particular from animals to humans (topics 0,1,5,8,14); lastly, on the right, is a cluster on public health, epidemics and immunology (topics 3,9,10,12).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

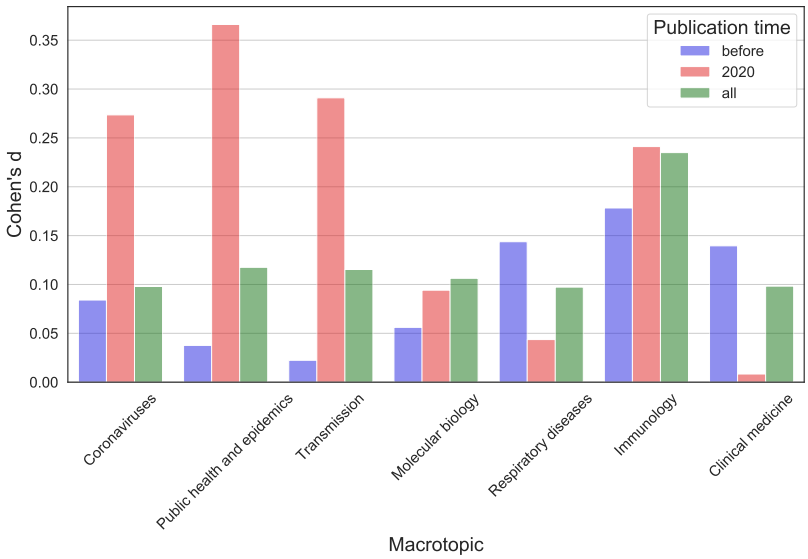
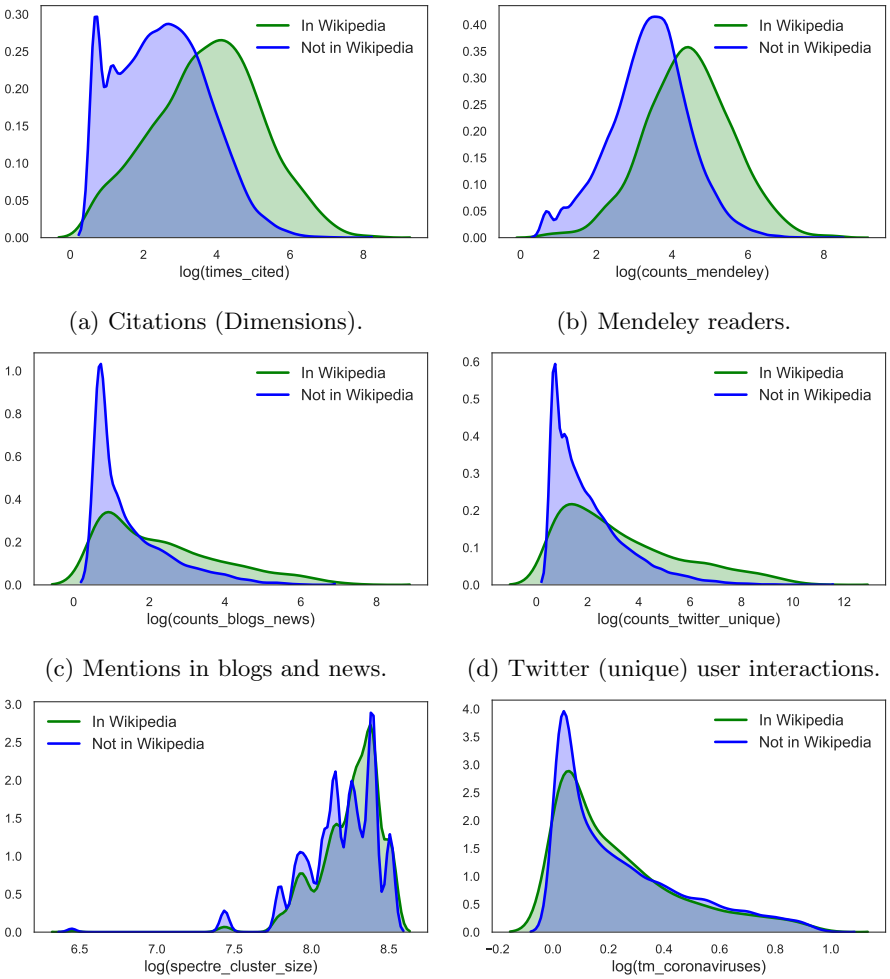


Figure 13: Cohen's d effect statistic for macrotopic intensity differences between articles cited in Wikipedia and not. Publications published before 2020, in 2020, and overall are considered. See Table 3, 4 and 5. Effect sizes are considered very small when below 0.2, small when below 0.5 and medium when below 0.8.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.10.087643; this version posted May 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) Citations (Dimensions).

(b) Mendeley readers.

(c) Mentions in blogs and news.

(d) Twitter (unique) user interactions.

(e) SPECTER cluster size.

(f) Macrotopic coronaviruses.

Figure 14: Some variables used for regression analyses. The plots distinguish variable values for articles cited from Wikipedia (green) or not (blue).