COVID-19 research in Wikipedia

Giovanni Colavizza (University of Amsterdam)

Pre-print: https://www.biorxiv.org/content/10.1101/2020.05.10.087643v3
(forthcoming in Quantitative Science Studies)

Wikipedia response to COVID-19

424,894,924 Pageviews

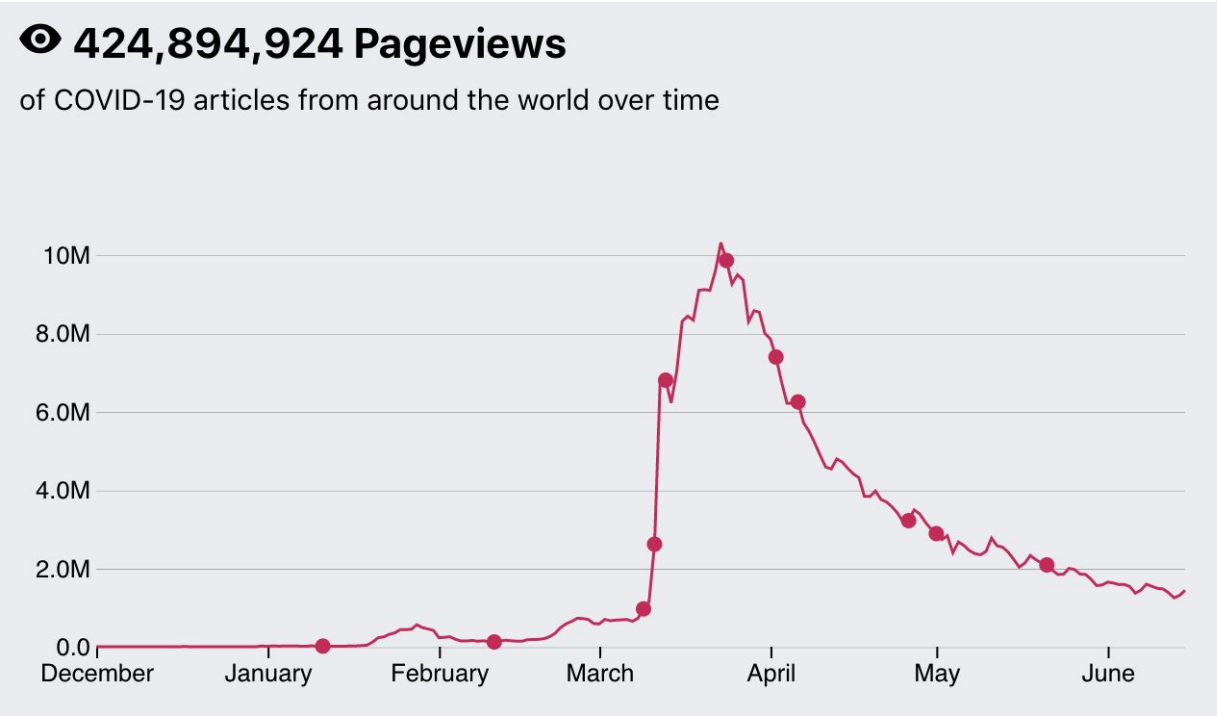
of COVID-19 articles from around the world over time

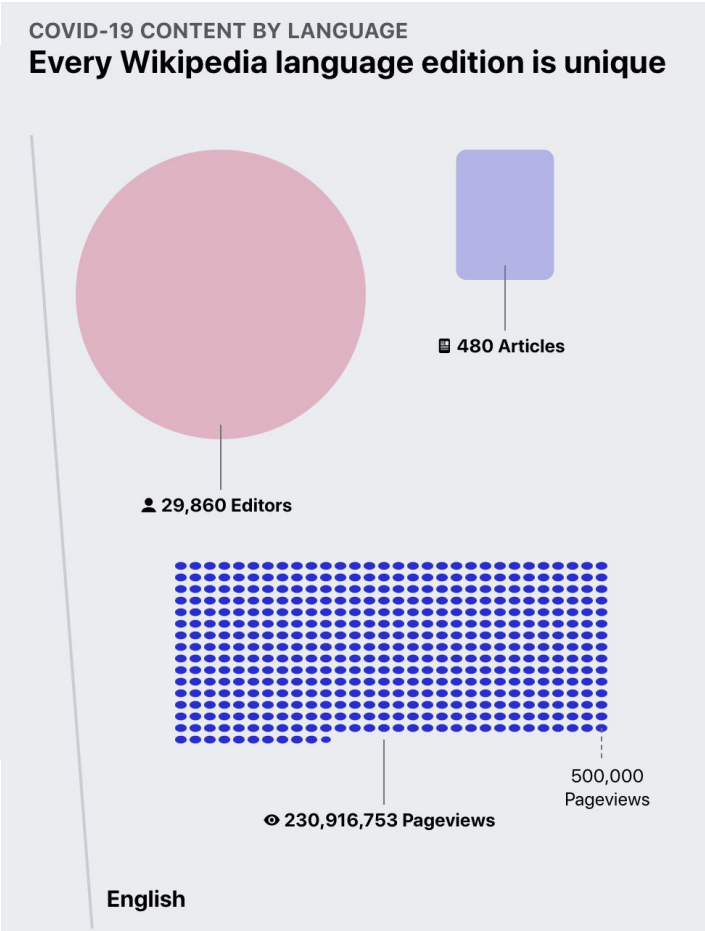
Views of Wikipedia articles about COVID-19 often reflect major developments in the timeline of the pandemic. For example, on March 12, 2020, the day after the World Health Organization (WHO) classified COVID-19 as a pandemic, the main English Wikipedia article about the pandemic had over 1.4 million views alone, an increase of 73 percent from the day before the WHO's announcement.[3]

Data from December 1, 2019 - June 15, 2020

https://wikimediafoundation.org/covid19/data/

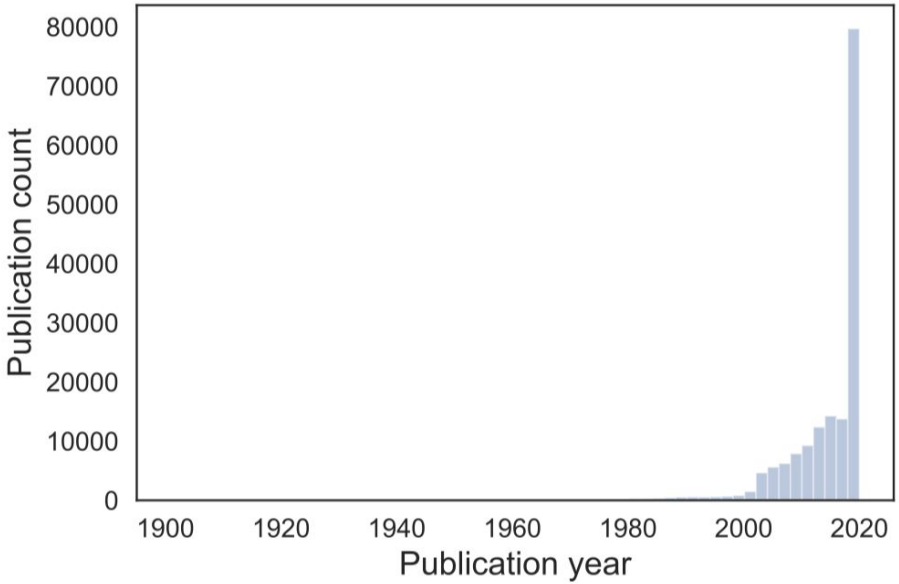
Wikipedia response to COVID-19

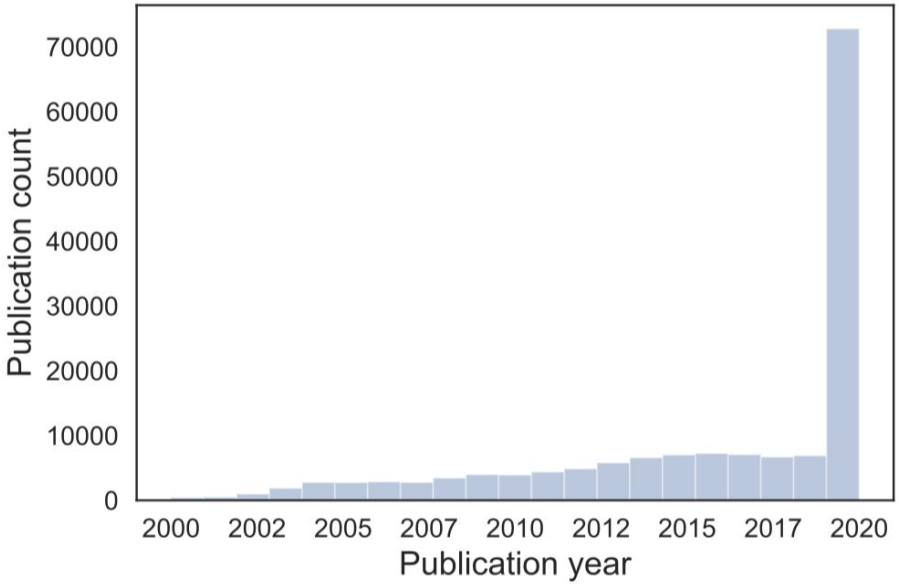




https://wikimediafoundation.org/covid19/data/

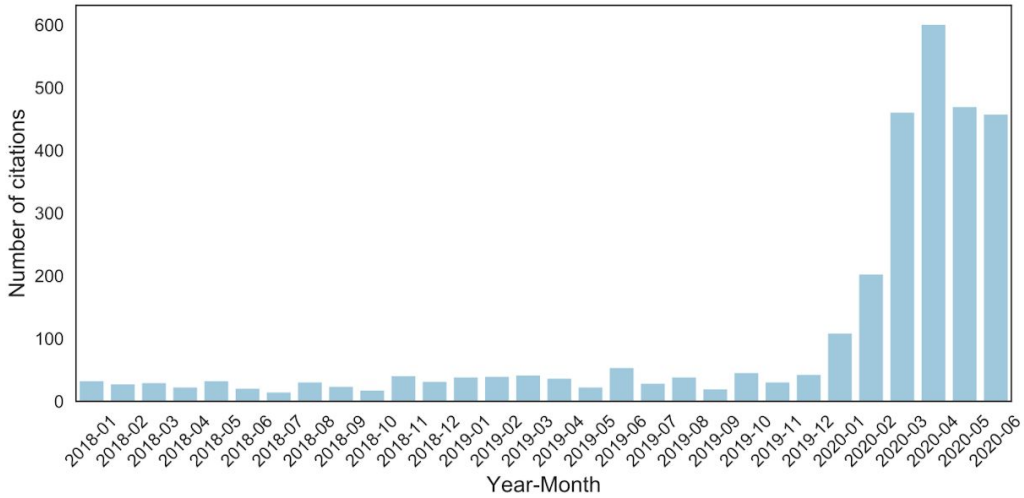
Research response to COVID-19





Data: July 1, 2020.

COVID-19 research in Wikipedia



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

Data: July 1, 2020.

COVID-19 research in Wikipedia?

Is Wikipedia relying on a representative and reliable sample of COVID-19-related research?

COVID-19 research in Wikipedia?

Is Wikipedia relying on a representative and reliable sample of COVID-19-related research?

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

COVID-19 research in Wikipedia?

Is Wikipedia relying on a representative and reliable sample of COVID-19-related research?

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

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

Data

References

1. ^ *a b c* Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. (February 2020). "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study". *Lancet*. **395** (10223): 507–513. doi:10.1016/S0140-6736(20)30211-7. PMC 7135076. PMID 32007143.

https://en.wikipedia.org/wiki/Coronavirus_disease_2019

Data

COVID-19 publications: ~160k articles (July 1, 2020) from CORD-19 and Dimensions; ***~141k with a DOI or PMID*** *matching in Dimensions.*

Wikipedia's citations to COVID-19: Altmetrics (July 1, 2020). ***3038 cited from Wikipedia (~1.9% overall; ~2.0% English; ~0.24% non-English).****

** For reference, across all Wikipedia compared to the Web of Science this average is instead of 3.5% (April 2020; https://arxiv.org/abs/2007.07022).*

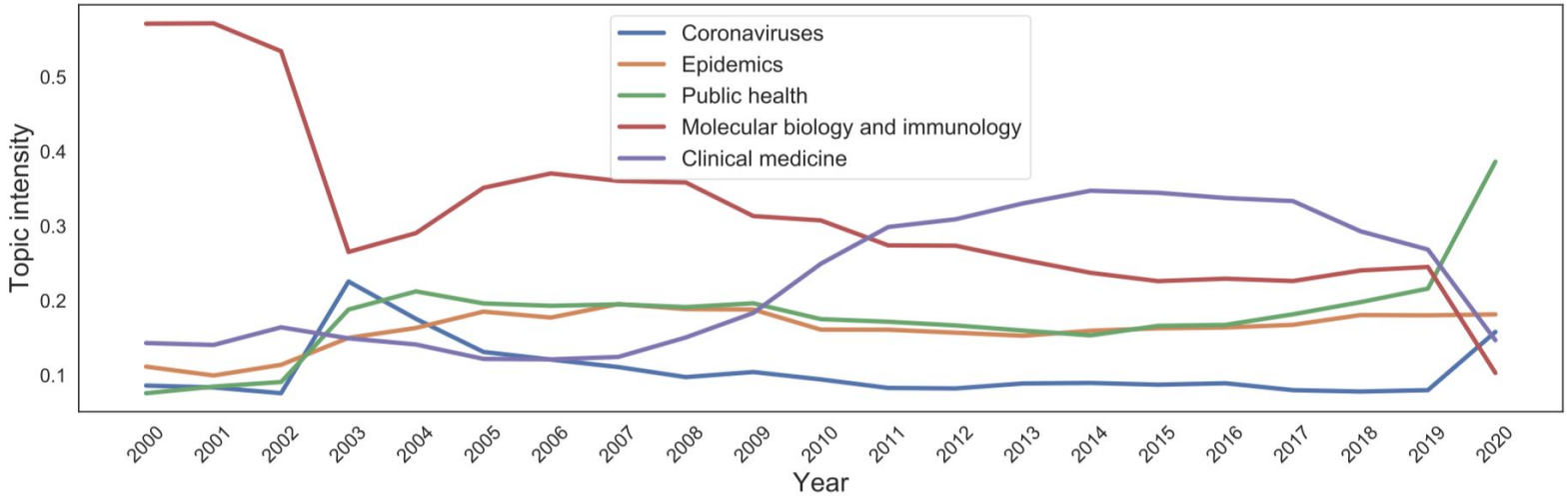
RQ1: Representativeness

Methods: apply different forms of clustering, what is the coverage of each cluster from Wikipedia?

RQ1: Representativeness

Methods: apply different forms of clustering, what is the coverage of each cluster from Wikipedia?

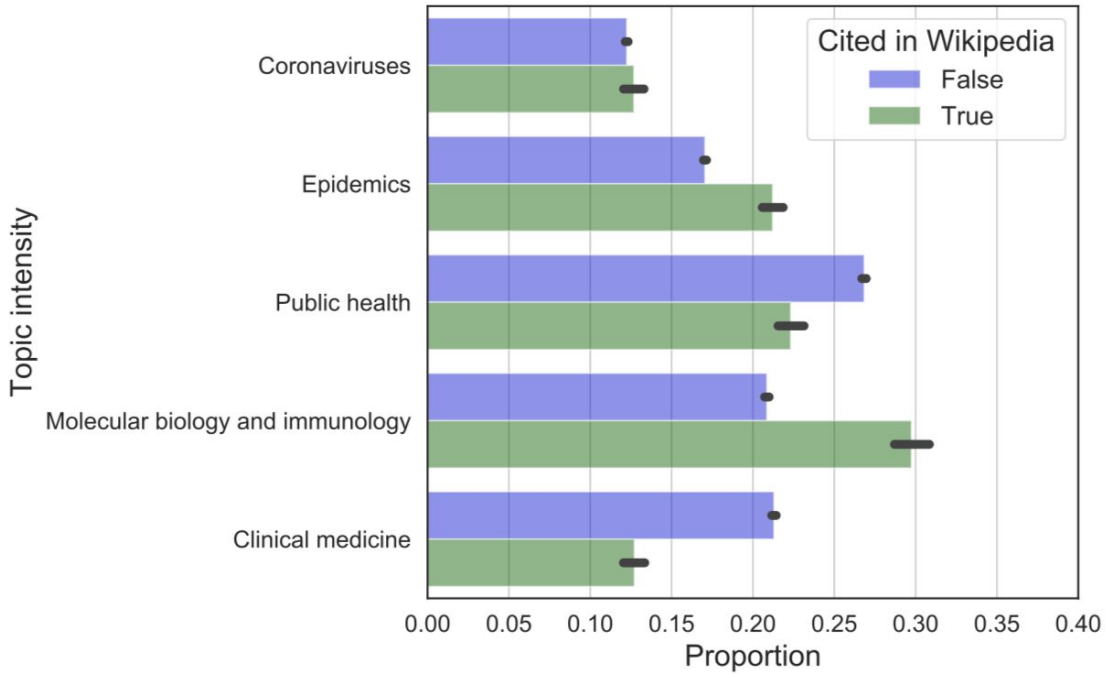
Results using **topic modelling** (on titles and abstracts)



RQ1: Representativeness

Topic intensities

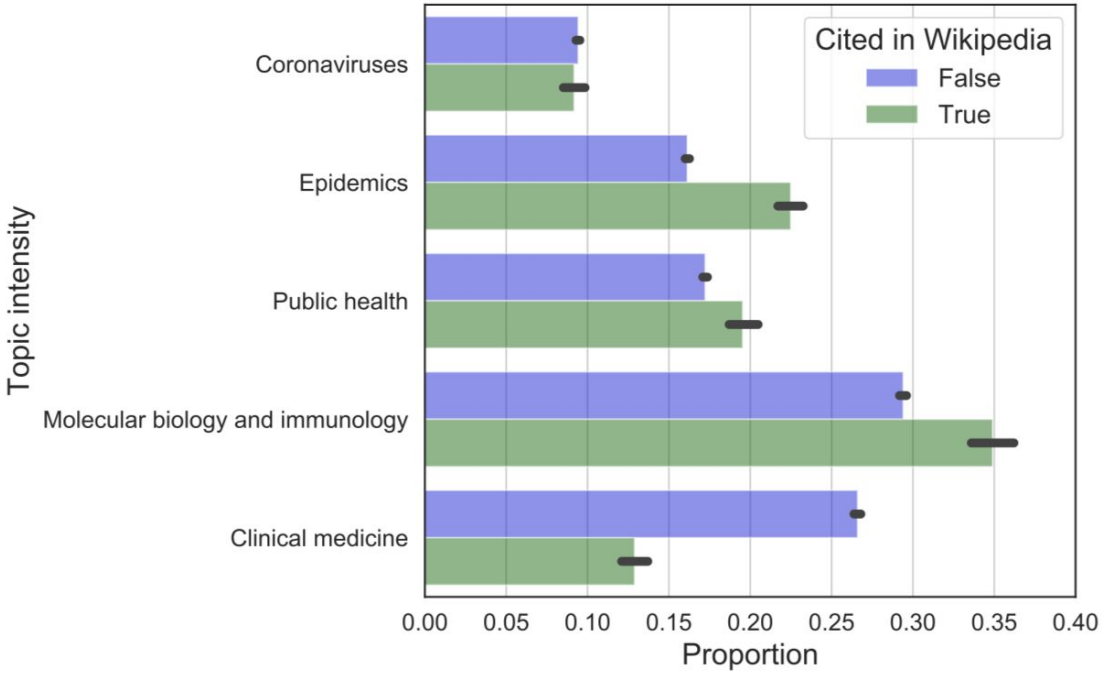
(overall)



RQ1: Representativeness

Topic intensities

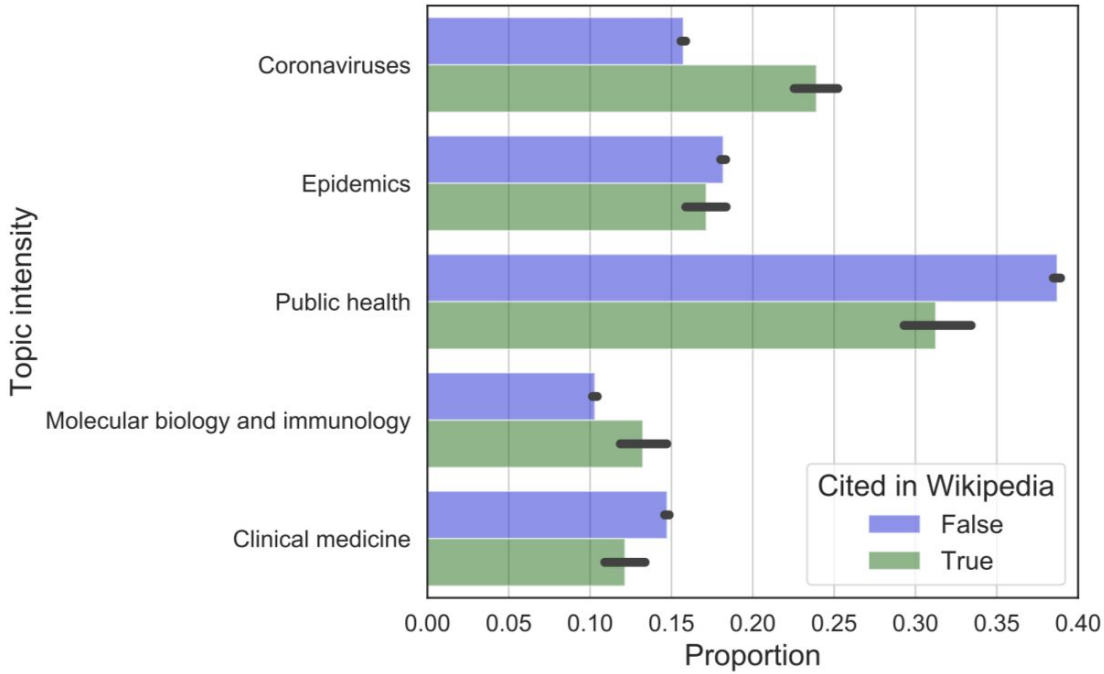
(pre-2020)



RQ1: Representativeness

Topic intensities

(2020)

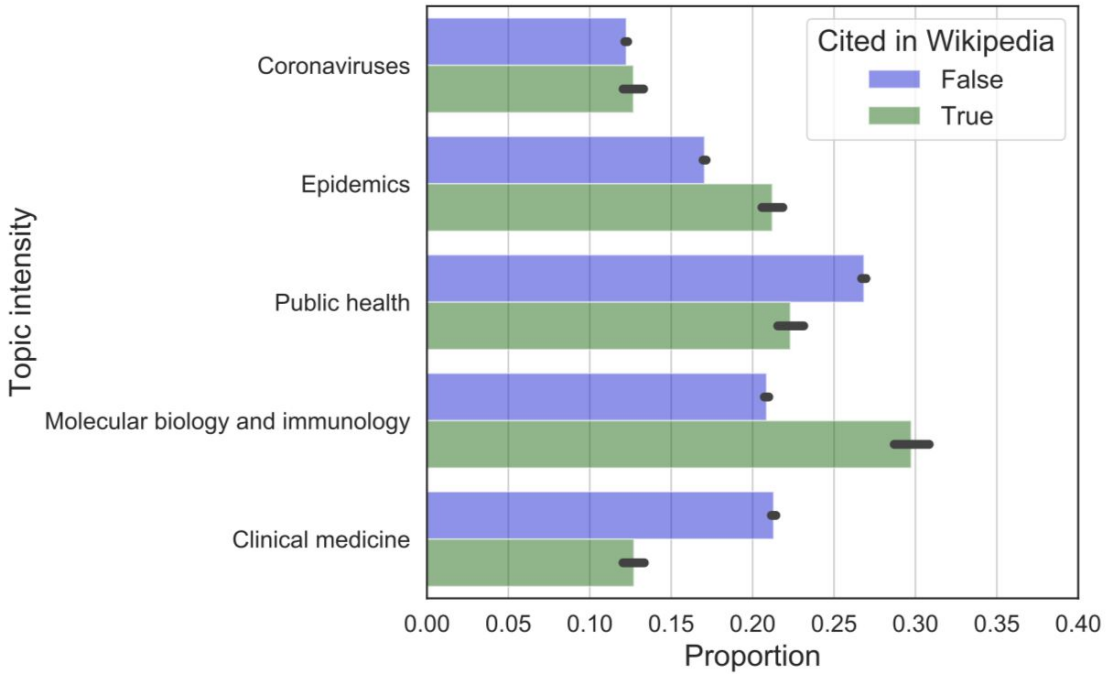


RQ1: Representativeness

Topic intensities

(overall):

- Epidemics, molecular biology and immunology remain slightly overrepresented.
- Public health and clinical medicine remain slightly underrepresented.
- Coronavirus-specific research is balanced.



RQ2: Reliability

Methods: we want to understand inclusion criteria, that is to say what correlates with being cited from Wikipedia? We use regression analysis and compare citations given before 2020 to those from 2020.

RQ2: Reliability

Results (2020 but also applies to 2019):

- **+ Reputed/specialized journals** (Nature, BMJ, Lancet)
- **+ Citations and altmetrics** (number of received citations, number of mentions on Twitter, Mendeley, blogs and news, etc.)
- **- Pre-prints** (medarXiv, bioarXiv)
- **Topics do not matter** (i.e., article-level effects explain away the residual variations in topic intensity we saw before).

RQ2: Reliability

Results (2020 but also applies to 2019):

- **+ Reputed/specialized journals** (Nature, BMJ, Lancet)
- **+ Citations and altmetrics** (number of received citations, number of mentions on Twitter, Mendeley, blogs and news, etc.)
- **- Pre-prints** (medarXiv, bioarXiv)
- **Topics do not matter** (i.e., article-level effects explain away the residual variations in topic intensity we saw before).

These are good news as they align with previous findings on literature cited from Wikipedia, and point to consistent inclusion criteria being used.

COVID-19 research in Wikipedia

Summary:

- Very rapid and large-scale effort to integrate COVID-19 knowledge in Wikipedia.
- Coverage of COVID-19 research topics mostly well-balanced, residual differences explained away by article-level effects.
- COVID-19 research cited from Wikipedia is highly impactful/visible.

-> Well done editors!

Pre-print: https://www.biorxiv.org/content/10.1101/2020.05.10.087643v3
(forthcoming in Quantitative Science Studies)