Connecting health research efforts and social attention: A dual analysis of local and international perspectives on Wikipedia and OpenAlex

Wenceslao Arroyo-Machado[*], Rodrigo Costas[**] and Adrián A. Díaz-Faes[***]

[*] *warroyom@asu.edu*
0000-0001-9437-8757
Center for Science, Technology and Environmental Policy Studies, School of Public Affairs, Arizona State University, Phoenix, AZ 85004, USA

[**] *rcostas@cwts.leidenuniv.nl*
0000-0002-7465-6462
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands
DSI-NRF SciSTIP, Stellenbosch University, South Africa

[***] *diazfaes@csic.es*
0000-0003-1928-4608
INGENIO (CSIC-UPV), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

Abstract

This research-in-progress paper examines the alignment between public interest, as evidenced by Wikipedia page views, and the distribution of academic resources across various health conditions. Utilising data from Wikipedia, Wikidata, and OpenAlex, the study reveals both relevant geographical correlations and notable gaps in how diseases are addressed in academic research compared to their visibility in social media. Moreover, discrepancies in content quality on Wikipedia pages indicate potential biases in the global research agenda. These findings underscore the importance of considering both social and academic metrics to address misalignments and advocate for a more equitable distribution of research resources in the biomedical sciences.

1. Introduction

Understanding the impact of health conditions and accelerating health improvements have often guided research priorities and the allocation of resources in biomedical research (Hanney et al., 2003; Zerhouni, 2003; Díaz-Faes et al., 2023). Within this context, the burden of disease, which measures the impact of health conditions on a population in terms of economic costs and loss of health and well-being, has emerged as a pivotal area of interest for quantitative science studies that seek to identify discrepancies in both attention and academic efforts (Evans et al., 2014; Yegros-Yegros et al., 2020). These studies aim to shed light on how the allocation of research resources does not consistently align with the actual health burdens faced by different populations (Howitt et al., 2012). This disparity suggests a potential bias in academic focus, driven by the needs of wealthy countries rather than purely by the public health needs (Kumar et al., 2024). Bibliometric methods are valuable proxies to monitor research efforts on disease (Yegros-Yegros, et al., 2020); however, careful attention is needed to avoid mechanistic approaches that may inadvertently steer research towards specific topics of interest or alter behavioural patterns (Baccini et al., 2019).

Analysis of Wikipedia offers an insightful perspective on public engagement and societal biases, complementing the findings of bibliometric analyses. As a universally accessible repository of knowledge, Wikipedia mirrors the varying levels of societal attention it receives, reflecting broader social interests (Arroyo-Machado et al., 2022). Studies into the activity patterns and page views of its articles have exposed substantial disparities in the representation and depth of coverage across different languages and regions, revealing preferences and neglect in the global discourse on health (Mittermeier et al., 2021). Within this framework, there has also been a focus on health-related topics, ranging from the quality of medical content (Smith, 2023) to the use of article traffic as a predictor of disease outbreaks (Santangelo et al., 2022). This connection between social attention and academic research enriches our understanding of how health-related concerns are prioritised and addressed across such as different communities.

However, literature lacks studies that more thoroughly address the social attention health issues receive on social media and connect this to the research attention they are given. While there are indeed preliminary proposals that have linked both publication data and Wikipedia(Arroyo-Machado et al., 2024)—there are no firm proposals that do so with a focus on such a sensitive topic. Our study aims to bridge this gap by investigating the alignment between social attention and research focus concerning health-related issues, using Wikipedia and OpenAlex. By doing so, we offer a first exploration on the potential discrepancies and biases that may exist between social interest and scientific research. Specifically, our study addresses the following research objectives:

1. To explore the extent to which social attention to health-issues parallels research efforts on disease.
2. To map health issues across different geographical regions, assessing whether the social attention varies significantly across areas.

2. Methodology

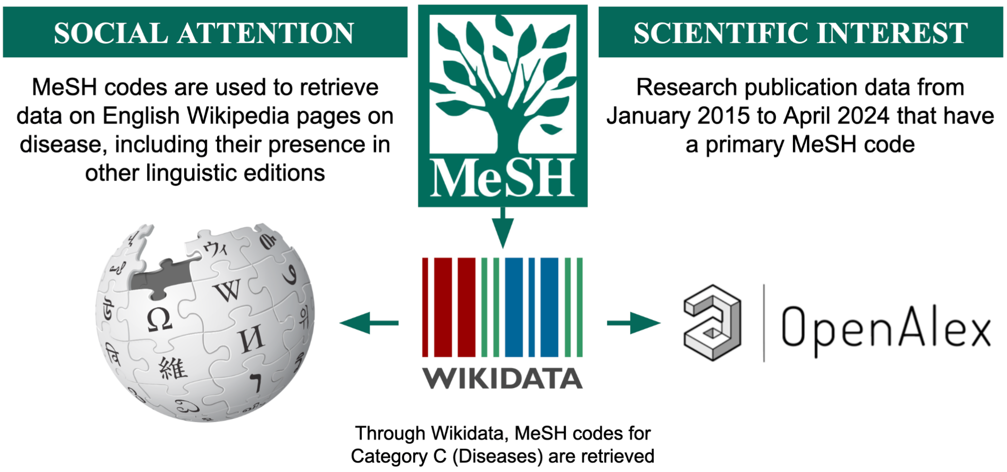
Data from three open sources have been utilised (Figure 1). First, Wikidata was used to locate all MeSH codes[1] descending from Category C (Disease) and the associated English Wikipedia article. This identified a total of 3,370 Wikipedia articles related to MeSH (the 67% of Category C of MeSH 2024), which served as a basis to explore the relationship between science and society concerning health issues. To gather data on social interest, the Wikipedia API was employed, identifying for each page its total number of editors, word count, bibliographic references (April 2024), and the total page views from when data has been available, from July 2015 to March 2024. Additionally, we retrieved the quality assessment for the articles linked to a WikiProject, which includes 97% of the articles. This assessment ranks articles according to established quality standards[2]. Regarding research performance, publication data were obtained from OpenAlex through Google Big Query[3], including the total number of publications and citations associated with each MeSH code, from January 2015 to April 2024.

[1] The National Library of Medicine's MeSH hierarchical classification system assigns categories to medical and scientific terms through a multi-levelled structure, where broader terms at the top of the hierarchy encompass more specific, related terms at subordinate levels, facilitating precise indexing and retrieval of biomedical information.

[2] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

[3] Public dataset courtesy of the InSySPo project in Campinas (Brazil).

Figure 1: Summary diagram of the data sources used.



MeSH codes are used to retrieve data on English Wikipedia pages on disease, including their presence in other linguistic editions

Research publication data from January 2015 to April 2024 that have a primary MeSH code

Through Wikidata, MeSH codes for Category C (Diseases) are retrieved

Second, Wikipedia articles were categorised based on their quality assessment (FA-Class, GA-Class, B-Class, C-Class, Start-Class, and Stub-Class). Similarly, the distributions of Wikipedia page views and number of publications in OpenAlex were directly compared as proxies for social attention and scientific interest, respectively. This was done to obtain a general overview of the potential relationship between these two dimensions. We also explored differences across linguistic editions to discern potential disimilarities in social interests across regions. Here we utilised page views from the four linguistic editions of Wikipedia with a significant presence of MeSH articles.

3. Results

3.1. General overview

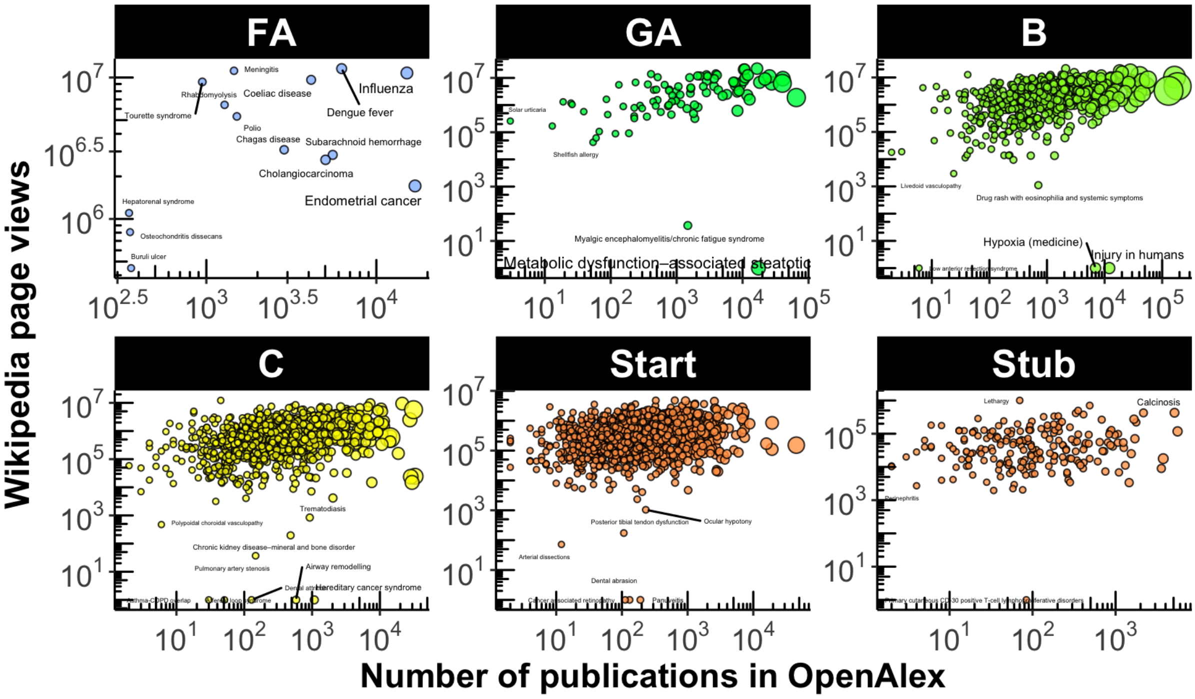
Table 1 provides descriptive values for Wikipedia and OpenAlex by assessment categories. 'Featured Articles' (FA-Class), with the highest quality rating, garner 5.48 million page views on average, pointing to high reader engagement and an average of over 4,433 works cited in OpenAlex. 'Good Articles' (GA-Class) maintain a high level of page views, slightly lower than FA-Class, and a greater average citation count, suggesting varied paths to achieving high article quality. With descending article assessments from B to 'Stub-Class', there is a clear trend of decreasing page views and research publication citations. This reveals a link between the depth of article development and public engagement, suggesting that a strong academic foundation attracts greater readership.

Table 1. Summary of Wikipedia and OpenAlex metrics by assessment category.

	Wikipedia					**OpenAlex**	
	Articles total	**Views** average	**Editors** average	**Words** average	**References** average	**Works** total	**Citations** average
FA-Class	14	5,482,618.43	990.50	5,332.79	101.43	4,433.57	71,673.36
GA-Class	92	4,536,433.71	1,008.41	4,729.83	108.42	5,382.59	93,107.14
B-Class	599	2,877,098.25	657.27	2,945.27	67.29	3,857.98	69,354.92
C-Class	1,112	1,095,431.48	237.87	1,559.43	32.02	1,140.17	19,569.52
Start-Class	1,234	506,138.97	109.10	742.13	14.50	594.22	8,260.34
Stub-Class	207	78,611.15	22.45	96.42	2.72	322.16	4,223.48
Non-Class	112	233,332.37	85.76	1,728.87	41.00	1,919.27	46,943.72

There is a noticeable, albeit not strict, trend where diseases with a high count of associated research publications in OpenAlex tend to have increased Wikipedia page views (Figure 2). This could imply that diseases with a substantial body of academic work gather greater interest, resulting in higher traffic on their Wikipedia pages. This trend is more pronounced for pages with moderate quality assessments (GA, B, and C classes), whereas Featured Articles (FA-Class), despite being fewer in number and presumably of higher content quality, show greater variability in page views. This diversity in viewership may reflect the context-dependency and varying interest of certain medical topics among the general public. Conversely, the Start and Stub categories, although numerous, attract fewer views, which could indicate that a lack of depth in content deter readers. 'Tuberculosis'[4] and 'Lyme disease'[5] articles are leading in Wikipedia pageviews, with 21,914,443 and 21,115,097 respectively. Tuberculosis is assessed as a Good Article, whereas Lyme disease is a B-Class article.

Figure 2: Distribution of the number of publications in OpenAlex for each MeSH term vs. Wikipedia articles page views. The size represents the average number of citations of the publications.



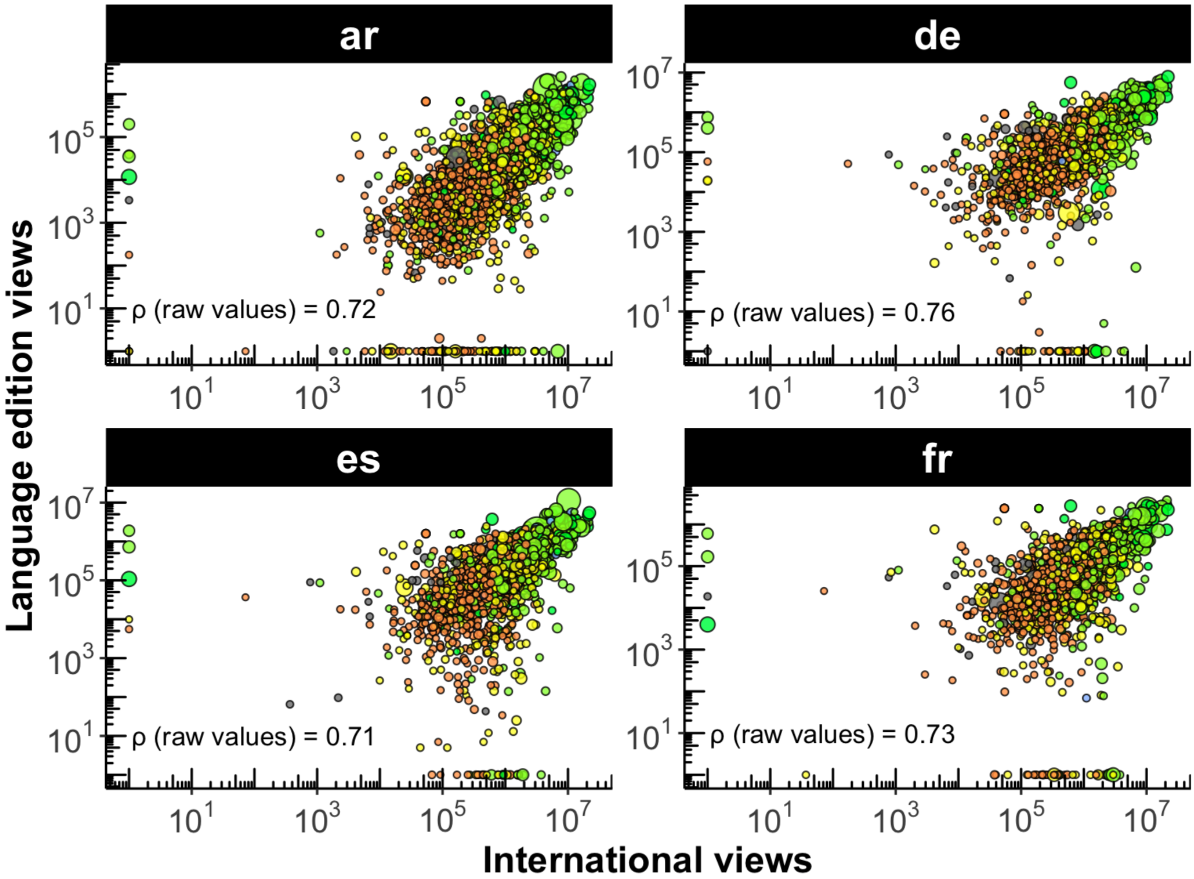
3.2. International attention vs local attention

Figure 3 compares the distributions of page views across Wikipedia editions where MeSH articles are most prevalent, specifically Arabic (ar), German (de), Spanish (es), and French (fr), against the benchmark of international views (English). The scatterplots show a strong correlation between the language edition views and international views, reflecting a consistent cross-linguistic interest in medical topics.Pearson correlation coefficients (ρ) are 0.72, 0.76, 0.71, and 0.73 for the Arabic, German, Spanish, and French editions respectively. This pattern underscores the universal relevance of medical information transcending linguistic barriers and suggesting a common thread of interest in health-related topics across these diverse cultures.

[4] https://en.wikipedia.org/wiki/Tuberculosis

[5] https://en.wikipedia.org/wiki/Lyme_disease

Figure 3: The association between international views and other language edition views of Wikipedia pages associated with MeSH codes. The colours correspond to the English Wikipedia assessment, with those lacking a class being in black.



4. Conclusions and further research

This study underscores a transition in altmetrics from simple numerical tallies to a more interactive and process-oriented approach. Our conceptualisation of social attention through Wikipedia metrics, alongside academic attention gauged by publication counts, has unveiled a robust correlation between these domains. Nevertheless, significant discrepancies are evident, particularly in how socially pertinent diseases are represented in academic research compared to their presence on Wikipedia. These discrepancies highlight the potential misalignment between public concerns and scientific interest, especially in terms of health needs.

At this stage, the preliminary findings of our work reveal a clear connection between research efforts and social attention, although variations across different language editions of Wikipedia were minimal. However, the differences by geographical area and content quality of Wikipedia pages deserve further scrutiny. Moving forward, we expect to expand our methodology and provide a more granular account of the factors affecting this relationship. This includes investigating the aforementioned discrepancies to better identify socially relevant diseases that may lack corresponding scientific research. By refining the methodologies employed and broadening the scope of use cases, we aim to enrich previous findings, such as those by Evans et al. (2014) and Yegros-Yegros et al. (2020), that have pointed to a misalignment between health needs and research efforts. Enhancing our understanding of the dynamics at play between social attention and scholarly engagement will undoubtedly provide valuable insights into aligning societal needs with academic research efforts more effectively.

5. References

Arroyo-Machado, W., Díaz-Faes, A. A., Herrera-Viedma, E., & Costas, R. (2024). From academic to media capital: To what extent does the scientific reputation of universities translate into Wikipedia attention? *Journal of the Association for Information Science and Technology*, *75*(4), 423-437. https://doi.org/10.1002/asi.24856

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes. *Quantitative Science Studies*, 1-22. https://doi.org/10.1162/qss_a_00226

Baccini, A., De Nicolao, G., & Petrovich, E. (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLOS ONE*, *14*(9), e0221212. https://doi.org/10.1371/journal.pone.0221212

Díaz-Faes, A. A., Llopis, O., D'Este, P., & Molas-Gallart, J. (2023). Assessing the variety of collaborative practices in translational research: An analysis of scientists' ego-networks. *Research Evaluation*, 32(2), 426-440. https://doi.org/10.1093/reseval/rvad003

Evans, J. A., Shim, J.-M., & Ioannidis, J. P. A. (2014). Attention to Local Health Burden and the Global Disparity of Health Research. *PLoS ONE*, *9*(4), e90147. https://doi.org/10.1371/journal.pone.0090147

Hanney, S. R., Gonzalez-Block, M. A., Buxton, M. J., & Kogan, M. (2003). The utilisation of health research in policy-making: Concepts, examples and methods of assessment. *Health Research Policy and Systems*, *1*(1), 2. https://doi.org/10.1186/1478-4505-1-2

Howitt, P., Darzi, A., Yang, G.-Z., Ashrafian, H., Atun, R., Barlow, J., Blakemore, A., Bull, A. M., Car, J., Conteh, L., Cooke, G. S., Ford, N., Gregson, S. A., Kerr, K., King, D., Kulendran, M., Malkin, R. A., Majeed, A., Matlin, S., … Wilson, E. (2012). Technologies for global health. *The Lancet*, 380(9840), 507-535. https://doi.org/10.1016/S0140-6736(12)61127-1

Kumar, A., Koley, M., Yegros, A., & Rafols, I. (2024). Priorities of health research in India: Evidence of misalignment between research outputs and disease burden. *Scientometrics*. https://doi.org/10.1007/s11192-024-04980-x

Mittermeier, J. C., Correia, R., Grenyer, R., Toivonen, T., & Roll, U. (2021). Using Wikipedia to measure public interest in biodiversity and conservation. *Conservation Biology*, *35*(2), 412-423. https://doi.org/10.1111/cobi.13702

Santangelo, O. E., Gianfredi, V., & Provenzano, S. (2022). Wikipedia searches and the epidemiology of infectious diseases: A systematic review. *Data & Knowledge Engineering*, *142*, 102093. https://doi.org/10.1016/j.datak.2022.102093

Smith, D. A. (2023). It's Time to Recognize Wikipedia as a Health Information Resource. *Journal of Consumer Health on the Internet*, *27*(2), 210-220. https://doi.org/10.1080/15398285.2023.2211498

Yegros-Yegros, A., Van De Klippe, W., Abad-Garcia, M. F., & Rafols, I. (2020). Exploring why global health needs are unmet by research efforts: The potential influences of geography, industry and publication incentives. *Health Research Policy and Systems*, *18*(1), 47. https://doi.org/10.1186/s12961-020-00560-6

Zerhouni, E. (2003). The nih roadmap. *Science*, 302(5642), 63-72. https://doi.org/10.1126/science.1091867

Open science practices
For further insights into our InSySPo infrastructure and associated projects, access the GitHub repositories at https://github.com/insyspo and https://github.com/alyssonmazoni.

Author contributions
WAM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing—original draft.
RC: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing—review & editing.
AADF: Conceptualization, Project administration, Resources, Supervision, Validation, Writing—review & editing.

Competing interests
Authors declare that they have no competing interests.