Jun 04, 2025

Assessing the quality of information about Parkinson's disease on Wikipedia

DOI
dx.doi.org/10.17504/protocols.io.14egn48xmv5d/v1

Simone Vieira da Silva[1], Alberto Leoncio Barreto Vasconcelos[2], Maria Elisa Pimentel Piemonte[1,3], João Alexandre Peschanski[1,4]

[1]Research, Innovation and Dissemination Center for Neuromathematics;
[2]Universidade Federal do Recôncavo da Bahia - UFRB;
[3]Faculdade de Medicina da Universidade de São Paulo – FMUSP; [4]Wikimidia Brasil

Simone Vieira da Silva
Centro de Pesquisa, Inovação e Difusão em NeuroMatematica - ...

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.14egn48xmv5d/v1

External link: https://cepid.fapesp.br/research-innovation-and-dissemination-center-for-neuromathematics

Protocol Citation: Simone Vieira da Silva, Alberto Leoncio Barreto Vasconcelos, Maria Elisa Pimentel Piemonte, João Alexandre Peschanski 2025. Assessing the quality of information about Parkinson's disease on Wikipedia . **protocols.io** **https://dx.doi.org/10.17504/protocols.io.14egn48xmv5d/v1**

Manuscript citation:

Protocol status: Working
We use this protocol and it's working

Created: March 28, 2025

Last Modified: June 04, 2025

Protocol Integer ID: 125603

Keywords: Parkinson's disease , Wikipedia , information quality , scientific dissemination, quality of wikipedia health information, wikipedia health information, quality of online health information, online health information, quality of health information, wikiproject medicine, health information, available health information, including wikipedia article, wikipedia article, quality of information, including parkinson, parkinson, wikipedia, health of user, health education tool, health, information, disease, health professional, clinical practice, evaluation guideline, medical student, available on wikipedia

Funders Acknowledgements:
Research Foundation's Research, Innovation and Dissemination Center for Neuromathematics
Grant ID: 2013/07699-0
State of São Paulo Research Support Foundation (FAPESP)
Grant ID: 2024/09360-4

Abstract

Wikipedia has been increasingly used as a health education tool, and is widely used by medical students and health professionals as a source of health information, which influences clinical practices and the health of users, highlighting the importance of the quality of available health information. Through the “WikiProject Medicine” Project, initiatives are developed to evaluate and improve the quality of Wikipedia health information, including Parkinson's disease (Matheson and Matheson-Monnet 2017; Joorabch Doherty and Dawson , 2020; Wikipedia, 2025; Herbert et al., 2015; Smith, 2020). However, there is still a lack of adequate methodology to evaluate the quality of health information available online, including Wikipedia articles (Dominguesand Lopes, 2019; Smith, 2020; da Silva Couto, 2021). Therefore, this protocol aims to present evaluation guidelines to compare the quality of online health information, specifically related to Parkinson's disease, available on Wikipedia in multiple languages, in order to ensure the quality of information about Parkinson's disease.

Guidelines

This protocol does not require ethical approval as it involves open access data made available on Wikipedia. This work was supported by the São Paulo Research Foundation's and Research, Innovation and Dissemination Center for Neuromathematics (FAPESP 2013/07699-0, 2024/09360-4).

Materials

Parkinson's disease.xlsx 8.3MB
Alzheimer's Disease.xlsx 9.5MB
Tourette's Syndrome.xlsx 3.3MB

Troubleshooting

1 A listing of articles in the category Parkinson's disease on the English Wikipedia was conducted using the PetScan tool (PetScan/en, 2025), on February 7, 2025. A query to PetScan generates a unique identifier, in this case: PSID 31781056. The process resulted in a total of 134 articles, covering those within the category Parkinson's disease on the English Wikipedia, as well as its subcategories (Category:Parkinson's disease, 2025).

2 Next, a data cleaning process was conducted to retain only relevant content. Entries from four subcategories (People with Parkinson’s disease, Dopamine agonists, Parkinson’s disease organizations and researchers) and one article from the main category (Contursi Terme, an Italian town) were excluded, leaving 121 articles for analysis.

3 We identified corresponding articles in other language versions of Wikipedia using Wikidata, a (Vrandečić and Pintscher , Krötzsch, 2023). For example, the Wikidata item for Parkinson's disease is identified as Q11085, and as of 7 February 2025, it was linked to articles in 103 of 352 language editions of Wikipedia.

4 To retrieve corresponding articles, we implemented a script within the Django framework (Leoncio, 2023a) to identify the Wikidata item associated with each of the 121 articles collected from the Parkinson's Disease category on the English Wikipedia.

5 Wikidata items , we then run another scriptto extract the corresponding articles in all editions of Wikipedia (Leoncio, 2023b), finally identifying 919 entries in 105 languages.

6 For each of these 919 articles, we retrieved three types of information: textual characteristics, reference metrics, and editorial history of the articles. The decision to obtain these three types of
data refers to strategies documented in the literature on how to measure information quality on Wikipedia (Giles, 2005; Domingues and Lopes, 2019; Claes and Tramullas , 2021). A diverse set of tools was used to obtain information about articles across all Wikipedias . Duplicates between wikis were not removed, i.e., an editor who edited on more than one Wikipedia; references that were used in more than one article were also not removed as duplicates.

7 The database is organized into seven clusters: category, page, authorship, reference, analysis, parser, and query. The category, page, parser, and query clusters provide general information for collecting data from articles about Parkinson's disease across multiple Wikipedias. The page, analysis, and authorship clusters contain more detailed information for each page and are used to analyze the quality of the information.

8 For the cluster analysis in the database, the textual features of 919 articles were extracted using the MediaWiki API (API:Main page , 2025) and its routes, such as XML Parse Tree and Revisions (API:Parsing_wikitext, 2025; API:Revisions, 2025). Word count was used as a measure of article completeness, with computational code markers removed from the quantification process using the external data analysis library

BeautifulSoup (Thota and Elmasri, 2021). The number of links to other Wikipedia articles was considered a measure of interconnectivity and information access. The number of sections in an article was used to assess the structure of the information, considering only the main sections (Help:Section, 2025). Additionally, the number of images was used as an indicator of the quality of visual documentation.

9 Reference metrics were retrieved using the MediaWiki API for reference clustering. These metrics are the primary means of assessing reliability on Wikipedia. A total of 9,870 references were collected. DOI and PMID data were extracted from each reference using regular expressions via the MediaWiki API. In cases where a DOI was present but no PMID was found, an additional search was conducted using the PubMed API (PMC, 2025).

10 Data on the editorial history of Parkinson's Disease articles included in the analysis cluster were collected using the MediaWiki API and the WikiWho tool (WikiWho, 2024). Through the API, we retrieved the creation and last edit dates of all articles. Using WikiWho, we conducted a more detailed analysis of the histories of Parkinson's Disease articles to identify and aggregate information about editors. As a result, data were obtained only for the following Wikipedia editions: 'ar', 'de', ' en ', 'es', 'eu', ' fr ', ' hu ', 'id', 'it', ' ja ', ' nl ', ' pt ' and ' tr '. Anonymous editors were grouped as a single contributor. Collecting data on the editorial history of Wikipedia articles is recognized in the literature as a means of gaining insights into the reliability, accuracy, and evolution of content over time, as well as assessing the influence of different editors, detecting potential biases, and understanding how information quality develops (Martins and Carmo, 2019).

11 A listing of articles in the category Tourette syndrome on the English Wikipedia was conducted using the PetScan tool (PetScan/en, 2025), on February 7, 2025. A PetScan query generates a unique identifier, in this case: PSID 32867835 . The process resulted in a total of 126 articles, covering those within the category Tourette syndrome on the English Wikipedia as well as its subcategories (Category:Tourette syndrome, 2025).

12 Next, a data cleaning process was conducted to retain only relevant content. Entries from three subcategories (People with Tourette syndrome, Tourette syndrome organizations, Works on Tourette syndrome) were excluded, leaving 36 articles for analysis.

13 We identified articles corresponding to Tourette syndrome in other language versions of Wikipedia using Wikidata (Vrandečić , Pintscher , Krötzsch , 2023). For example, the Wikidata item for Tourette syndrome is identified as Q11085, and as of 7 February 2025, it was linked to by 81 articles in 103 of 352 language editions of Wikipedia.

14 To retrieve matching articles, we implement a script within the Django framework (Leoncio, 2023 a) to identify the Wikidata item associated with each of the 36 articles collected from the Tourette Syndrome category on the English Wikipedia.

15 Using the obtained Wikidata items, we then run another script to extract the corresponding articles in all editions of Wikipedia (LEONCIO, 2023b), ultimately identifying 370 entries in 81 languages.

16 For each of these 370 articles, we retrieved three types of information: textual characteristics, reference metrics, and editorial history of the articles. The decision to obtain these three types of data refers to strategies documented in the literature on how to measure information quality on Wikipedia (Giles, 2005; Domingues and Lopes, 2019; Claes and Tramullas, 2021). A diverse set of tools was used to obtain information about articles across all Wikipedias . Duplicates between wikis were not removed, i.e., an editor who edited on more than one Wikipedia; references that were used in more than one article were also not removed as duplicates.

17 The database is organized into seven clusters: category, page, authorship, reference, analysis, parser, and query. The category, page, parser, and query clusters provide general information for collecting data from Tourette Syndrome articles across multiple Wikipedias. The page, analysis, and authorship clusters contain more detailed information for each page and are used to analyze the quality of the information.

18 For the cluster analysis in the database, textual features of 370 articles were extracted using the MediaWiki API (API:Main page , 2025) and its routes, such as XML Parse Tree and Revisions (API:Parsing_wikitext , 2025; API:Revisions, 2025). Word count was used as a measure of article completeness, with computational code markers removed from the quantification process using the external data analysis library BeautifulSoup (Thota and Elmasri, 2021). The number of links to other Wikipedia articles was considered a measure of interconnectivity and information access. The number of sections in an article was used to assess the structure of the information, considering only the main sections (Help:Section , 2025). Additionally, the number of images was used as an indicator of the quality of visual documentation.

19 Reference metrics were retrieved using the MediaWiki API for reference clustering. These metrics are the primary means of assessing reliability on Wikipedia. A total of 3,861 references were collected. DOI and PMID data were extracted from each reference using regular expressions via the MediaWiki API. In cases where a DOI was present but no PMID was found, an additional search was conducted using the PubMed API (PMC, 2025).

20 Data on the editorial history of Tourette Syndrome articles included in the analysis cluster were collected using the MediaWiki API and the WikiWho tool (WikiWho , 2024). Through the API, we retrieved the creation and last edit dates of all articles. Using WikiWho, we conducted a more detailed analysis of the histories of Tourette Syndrome articles to identify and aggregate information about editors. As a result, data were obtained only for the following Wikipedia editions: 'ar', 'de', 'en' , 'es', 'eu', ' fr ', ' hu ' , 'id', 'it', ' ja ', ' nl ', ' pt ' and ' tr '. Anonymous editors were grouped as a single contributor. Collecting data on the editorial history of Wikipedia articles is recognized in the literature as a means of

gaining insights into the reliability, accuracy, and evolution of content over time, as well as assessing the influence of different editors, detecting possible biases, and understanding how information quality develops (Martins and Carmo, 2019).

21 A listing of articles in the category Alzheimer's disease on the English Wikipedia was conducted using the PetScan tool (PetScan/en, 2025), on February 7, 2025. A PetScan query generates a unique identifier, in this case: PSID 32867927. The process resulted in a total of 557 articles, covering those within the category Alzheimer's disease on the English Wikipedia as well as its subcategories (Category:Alzheimer's disease, 2025).

22 Next, a data cleaning process was conducted to retain only relevant content. Entries from five subcategories (People with Alzheimer's disease, Alzheimer's disease papers, Alzheimer's and dementia organizations, Alzheimer's disease activists, Alzheimer's disease researchers) were excluded, leaving 138 articles for analysis.

23 We identified articles corresponding to Alzheimer's disease in other language versions of Wikipedia using Wikidata (Vrandečić , Pintscher , Krötzsch , 2023). For example, the Wikidata item for Alzheimer's disease is identified as Q11085, and as of 7 February 2025, it was linked to by 127 articles in 103 of language editions of Wikipedia.

24 To retrieve corresponding articles, we implemented a script within the Django framework (Leoncio, 2023a) to identify the Wikidata item associated with each of the 138 articles collected from the Alzheimer's disease category on the English Wikipedia.

25 Using the obtained Wikidata items , we then run another script to extract the corresponding articles in all editions of Wikipedia (Leoncio, 2023a), finally identifying 918 in 127 languages.

26 For each of these 918 articles, we retrieved three types of information: textual characteristics, reference metrics, and editorial history of the articles. The decision to obtain these three types of data refers to strategies documented in the literature on how to measure information quality on Wikipedia (Giles, 2005; Domingues and Lopes, 2019; Claes and Tramullas, 2021). A diverse set of tools was used to obtain information about articles across all Wikipedias . Duplicates between wikis were not removed, i.e., an editor who edited on more than one Wikipedia; references that were used in more than one article were also not removed as duplicates.

27 The database is organized into seven clusters: category, page, authorship, reference, analysis, parser, and query. The category, page, parser, and query clusters provide general information for collecting data from articles about Alzheimer's disease across multiple Wikipedias. The page, analysis, and authorship clusters contain more detailed information for each page and are used to analyze the quality of the information.

28 For the cluster analysis in the database, textual features of 918 articles were extracted using the MediaWiki API (API:Main page, 2025) and its routes, such as XML Parse Tree and Revisions (API:Parsing_wikitext, 2025; API:Revisions, 2025). Word count was used as a measure of article completeness, with computational code markers removed from

the quantification process using the external data analysis library BeautifulSoup (Thota & eElmasri, 2021). The number of links to other Wikipedia articles was considered a measure of interconnectivity and information access. The number of sections in an article was used to assess the structure of the information, considering only the main sections (Help:Section, 2025). Additionally, the number of images was used as an indicator of the quality of visual documentation.

29 Reference metrics were retrieved using the MediaWiki API for reference clustering. These metrics are the primary means of assessing reliability on Wikipedia. A total of 15,886 references were collected. DOI and PMID data were extracted from each reference using regular expressions via the MediaWiki API . In cases where a DOI was present but no PMID was found, an additional search was conducted using the PubMed API (PMC, 2025).

30 Data on the editorial history of the articles on Alzheimer's Disease, included in the analysis cluster, were collected using the MediaWiki API and the WikiWho tool (WikiWho , 2024). Through the API, we retrieved the creation and last edit dates of all articles. Using WikiWho , we conducted a more detailed analysis of the histories of the articles on Alzheimer's Disease to identify and aggregate information about editors. As a result, data were obtained only for the following Wikipedia editions: ‘ar’, ‘de’, ‘ en ’, ‘es’, ‘eu’, ‘ fr ’, ‘ hu ’, ‘id’, ‘it’, ‘ ja ’, ‘ nl ’, ‘ pt ’ and ‘ tr ’. Anonymous editors were grouped as a single contributor. Collecting data on the editorial history of Wikipedia articles is recognized in the literature as a means of gaining insights into the reliability, accuracy and evolution of content over time, as well as to assess the influence of different editors, detect potential biases and understand how information quality develops (Martins and Carmo, 2019).

31 With the collected data, a database will be structured with information on the three diseases: Parkinson's Disease, Tourette Syndrome and Alzheimer's Disease for a comparative analysis.

Protocol references

API:Main page. MediaWiki. (2025, February 23). Retrieved from https://www.mediawiki.org/w/index.php?title=API:Main_page&oldid=6703085.

API:Parsing_wikitext. MediaWiki. (2025, February 23). Retrieved from https://www.mediawiki.org/wiki/API:Parsing_wikitext.

API:Revisions, 2025. MediaWiki. (2025, February 23). Retrieved from https://www.mediawiki.org/wiki/API:Revisions.

Category: Alzheimer's Disease. (2025, February 14). Wikipedia/en. Retrieved 14:41, February 14, 2025 from https://en.wikipedia.org/wiki/Category:Alzheimer%27s_disease.

Category:Parkinson's disease. (2025, February 14). Wikipedia/en. Retrieved 14:41, February 14, 2025 from https://en.wikipedia.org/w/index.php?title=Category:Parkinson%27s_disease&oldid=894981590.

Category: Tourette's Syndrome. (2025, February 14). Wikipedia/en. Retrieved 14:41, February 14, 2025 from https://en.wikipedia.org/wiki/Category:Tourette_syndrome.

Claes, F. and Tramullas, J. (2021). Estudios sobre la credibilidad de Wikipedia: una revisión. *Área abierta*, ART-2021-127192.

da Silva Couto, L. P. (2021). Avaliação da qualidade da Wikipédia enquanto fonte de informação em saúde. *Repositório Aberto da Universidade do Porto*.

Domingues, G, and Lopes, C. T. (2019, May). Characterizing and comparing Portuguese and English Wikipedia medicine-related articles. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 1203-1207).

Giles, J. (2005). Special report: Internet encyclopaedias go head to head. *Nature, 438*(15), 900-901. Available from: https://www.mediawiki.org/w/index.php?title=Help:Section&oldid=6967655

Help:Section. MediaWiki. 23 Feb 2025, <https://www.mediawiki.org/w/index.php?title=Help:Section&oldid=6967655>.

Joorabchi, A., Doherty, C. and Dawson, J. (2020). 'WP2Cochrane', a tool linking Wikipedia to the Cochrane Library: Results of a bibliometric analysis evaluating article quality and importance. *Health Informatics Journal, 26*(3), 1881-1897.

Leoncio A. *load_articles* [Internet]. GitHub; 2023 [cited 2025 Mar 5]. Available from: https://github.com/albertoleoncio/parkinson/blob/main/data/management/commands/load_articles.

Leoncio A. *load_wikidata.py* [Internet]. GitHub; 2023 [cited 2025 Mar 5]. Available from: https://github.com/albertoleoncio/parkinson/blob/main/data/management/commands/load_wikidata.py.

Martins, D. L and do Carmo, D. (2019). Dinâmica de participação social na construção coletiva de informação no campo museal: Estudo de caso dos museus na Wikipédia no âmbito do Instituto Brasileiro de Museus. *Liinc em Revista, 15*(1).

Matheson, D., and Matheson-Monnet, C. (2017). Wikipedia as informal self-education for clinical decision-making in medical practice. *Open Medicine Journal, 4*(1).

PetScan/en. *Meta* [Internet]. 2025 Feb 11 [cited 2025 May 27]. Available from: https://meta.wikimedia.org/w/index.php?title=PetScan/en&oldid=28249805.

PMC. For developers. *PubMed Central* [Internet]. 2025 Feb 17 [cited 2025 May 27]. Available from: **https://pmc.ncbi.nlm.nih.gov/tools/developers/**.

Smith, D. A. (2020). Situating Wikipedia as a health information resource in various contexts: A scoping review. *PloS One, 15*(2), e0228786.

Thota, P. and Elmasri, R. (2021). Web scraping of COVID-19 news stories to create datasets for sentiment and emotion analysis. *The 14th PErvasive Technologies Related to Assistive Environments Conference*, 306–314. https://doi.org/10.1145/3453892.3461333.

Vrandečić, D., Pintscher, L and Krötzsch, M. (2023). Wikidata: The making of. *Companion Proceedings of the ACM Web Conference 2023*.

Wikipedia–challenges and new horizons in enhancing medical education. *BMC Med Educ*. 2015;15:1–6.
WikiWho. *MediaWiki* [Internet]. 2024 Aug 4 [cited 2025 May 27]. Available from: https://www.mediawiki.org/w/index.php?title=WikiWho&oldid=6686647.