# Assessing the quality of health-related Wikipedia articles with generic and specific metrics

Luís Couto
Faculty of Engineering of the University of Porto
Porto, Portugal
mieic1204994@fe.up.pt

Carla Teixeira Lopes
INESC TEC
Faculty of Engineering of the University of Porto
Porto, Portugal
ctl@fe.up.pt

## ABSTRACT

Wikipedia is an online, free, multi-language, and collaborative encyclopedia, currently one of the most significant information sources on the web. The open nature of Wikipedia contributions raises concerns about the quality of its information. Previous studies have addressed this issue using manual evaluations and proposing generic measures for quality assessment. In this work, we focus on the quality of health-related content. For this purpose, we use general and health-specific features from Wikipedia articles to propose health-specific metrics. We evaluate these metrics using a set of Wikipedia articles previously assessed by WikiProject Medicine. We conclude that it is possible to combine generic and specific metrics to determine health-related content's information quality. These metrics are computed automatically and can be used by curators to identify quality issues. Along with the explored features, these metrics can also be used in approaches that automatically classify the quality of Wikipedia health-related articles.

## CCS CONCEPTS

• **Information systems → Wikis**; • **Applied computing → Consumer health**; Health informatics.

## KEYWORDS

Information Quality, Wikipedia, Health-related Content

**ACM Reference Format:**
Luís Couto and Carla Teixeira Lopes. 2021. Assessing the quality of health-related Wikipedia articles with generic and specific metrics. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3442442.3452355

## 1 INTRODUCTION

Wikipedia is a well-known encyclopedia that anyone can edit. That makes it both a powerful source of information because anyone can expand it with their knowledge, but simultaneously enables the possibility of adding wrong information, either deliberately or by mistake [21]. To mitigate the likelihood of users inserting inaccurate information, Wikipedia relies on curators that revise the published content and make sure it corresponds to the truth.


*WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia*

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452355

The quality of information is especially relevant in healthcare. The search for health-related information online often ends up taking people to Wikipedia [15]. In 2021, English Wikipedia has more than 40 thousand health-related articles [8]. Currently, the most visited medical articles on Wikipedia have more than two billion annual visits, and the most popular article, now "COVID-19 pandemic", has, on average, more than 40 thousand daily views [29]. Health-related information is empowering, helping users live better with their health issues, but can be simultaneously dangerous, as wrong or misleading information can lead to unwanted results [7]. It is, therefore, of uttermost relevance to ensure the quality of health-related information in Wikipedia.

Generic metrics have been proposed to evaluate the quality of Wikipedia information in different dimensions, such as its completeness, informativeness, and accuracy [26, 30]. These metrics are built on top of lower-level features such as the number of editors involved in an article. Existing metrics do not consider specific features that might be useful to improve evaluation in the health domain. In this work, we propose health-specific features such as medicine infoboxes to capture the quality of Wikipedia articles in this domain. After investigating the correlation between features and quality, we propose health-specific metrics and assess their effectiveness in capturing the quality of Wikipedia articles in the health domain.

Section 2 presents an overview of the work related to Wikipedia information quality. In Section 3, we describe our methodology, followed by the proposal of features and health-specific metrics, in Sections 4 and 5, respectively. Finally, Section 6 concludes the work.

## 2 WIKIPEDIA INFORMATION QUALITY

Quality has been a concern of Wikipedia that, since its start, has defined mechanisms to assure certain levels of quality. The English Wikipedia has, currently, more than 6 million articles [9]. Assessing the quality of so much information becomes a challenge and requires a certain degree of automation. Several authors have approached this issue, both generally and in the health domain.

### 2.1 Wikipedia internal quality mechanisms

Bearing in mind the questions raised about the quality of the information present in Wikipedia and its importance for its survival and growth as a source of information, the need to create internal mechanisms to guarantee acceptable quality levels arose from the beginning. Considering the advantage of editing content easily and immediately by anyone, Wikipedia considers the users as guarantors of quality while correcting errors detected when using the

information. Moreover, with quality in mind, Wikipedia defined a set of policies and guidelines used by teams of volunteers organized in specific departments to analyze the added material. Also, there are bots that automatically and continuously monitor the content, looking for errors. Also, there are bots that automatically and continuously monitor the content, looking for errors. Wikipedia also has a web service, ORES [12], that predicts edits and articles' quality through machine learning, helping humans in this task. This tool, however, supports only a limited set of Wikimedia wikis [1].

Wikipedia has a system[2] for classifying the quality of articles. Members of WikiProjects carry out quality assessments that make it possible to determine the quality of the information in specific areas and prioritize work according to expectations. In the health-related area, WikiProject Medicine[3] handles this task. This project started in 2004 to contribute to medical articles' improvement, which is part of the Wiki Project Med Foundation. It gathers expert curators to improve the healthcare-related quality of information. This project defined policies in addition to Wikipedia in general. Over time, a set of tools has been put together to help its members achieve the objectives. The main levels of the WikiProject Medicine Quality Scale[4] are described in Table 1.

**Table 1: WikiProject Medicine article quality main levels**

| Level | Description |
|---|---|
| FA | The highest-rated article, exhaustively evaluated by independent reviewers. It is an article with good prose, comprehensive, with good underlying research, neutral in point of view, stable, follows the norms of style, has transferred multimedia elements, and has an appropriate extension to its content. |
| GA | Article analyzed by one or more impartial reviewers. It is well written, referenced, without unpublished research, it has comprehensive coverage of the topic, it is neutral from the point of view, stable, and illustrated when convenient. |
| B | Article analyzed by one or more impartial reviewers. It is well referenced, reasonably covers the topic, without omissions and apparent errors, has a defined structure, is reasonably well written, contains inappropriate multimedia elements, and its content is understandable. |
| C | Article where important content is missing or that contains much irrelevant content. It has references to credible sources, it is structured, following style norms, but it lacks some of the necessary criteria for level B. |
| Start | An incomplete article, which is still under development. It may not contain references from reliable sources, the prose may not be of high quality, but it must satisfy general Wikipedia policies. |
| Stub | Basic description of the topic. It may not be well written, has problems with the content itself, it is usually very short and runs the risk of ceasing to be considered an article. |
| List | Complies with the criteria of an autonomous list. An article comprising predominantly a list, typically consisting of links to articles in a particular subject field. |

[1]https://ores-support-checklist.toolforge.org
[2]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment
[3]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine
[4]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Assessment

## 2.2 Generic metrics for assessing quality

In 2010, Wu *et al.* [30] used 28 metrics divided into four groups: lingual - *e.g.* readability; structural - *e.g.* links; historical - *e.g.* article age and reputational - *e.g.* amount of editors. Li *et al.* [16] and De La Robertie *et al.* [4] proposed solutions based on the relationship between articles and editors. In 2019, Marrese-Taylor *et al.* [18] used metrics based on the articles' editions, but also consider the description that accompanies each edition.

Stvilia *et al.* [26] define seven metrics: authority, completeness, complexity, informativeness, consistency, currency, and volatility. These metrics use 19 features from Wikipedia articles and their history. Given the importance of these metrics for our work, we will describe them in the following paragraphs.

Authors define authority as "the degree of the reputation of an information object in a given community", computed as: ***Authority = 0,2 ∗ Num. Unique Editors + 0,2 ∗ Num. Edits + 0,1 ∗ Connectivity + 0,3 ∗ Num. Reverts + 0,2 ∗ Num. External Links + 0,1 ∗ Num. Registered User Edits + 0,2 ∗ Num. Anonymous User Edits***. The number of unique editors is the number of different authors involved in the article's editions, as extracted from its history. Connectivity corresponds to the number of articles linked to the article through joint editors. It is obtained by extracting each article's editors and the articles edited by them, using the history made available for the articles. This measure has the limitation that it can only be calculated based on the articles existing in the database, thus requiring a large dataset to be reliable. Reverts correspond to reversions made to previous editions of the article according to its editing history. External links refer to links present throughout the article that refer to content outside Wikipedia. Registered or anonymous editors can make editions.

Completeness is defined as "the granularity or precision of an information object's model or content values according to some general-purpose IS-A ontology such as WordNet". It is computed as: ***Completeness = 0,4 ∗ Num. Internal Broken Links + 0,4 ∗ Num. Internal Links + 0,2 ∗ Article Length***. Broken links are those that refer to pages that are currently unavailable. Internal links are links that refer to other pages of Wikipedia. The length of the article reflects the text size in characters.

The authors define complexity as "the degree of cognitive complexity of an information object relative to a particular activity". Its formula was defined as: ***Complexity = 0,5 ∗ "Flesch reading ease" - 0,5 ∗ "Kincaid grade level"***. Flesch Reading Ease [10] and Kincaid grade level [13] are instruments that assess readability using the number of phrases, words, and syllables in the text. Flesch Reading Ease is based on a ranking scale of 0-100, with low scores indicating text that is complicated to understand. Kincaid grade level assesses the American school grade necessary to understand the texts. They correlate inversely - a high score on the reading ease test corresponds to a lower grade level.

The definition of "Informativeness" is linked to the amount of information that an information object contains. It is computed as: ***Informativeness = 0,6 ∗ InfoNoise - 0,6 ∗ Diversity + 0,3 ∗ Num. Images***. InfoNoise is based on previous work [31] and refers to the ratio between the information present in an article and its total size, where the so-called noise exists. It refers to the ratio between the size of the information content, in words, after stemming and

stopping, and the object's total size. Diversity corresponds to the ratio between the number of unique edits and the number of total edits of an article. The number of images is obtained by counting them in each article among the different media objects present.

Consistency is defined as "the extent to which similar attributes or elements of an information object are consistently represented with the same structure, format and precision". It is calculated according to: ***Consistency = 0,6 ∗ Administrators Edit Share + 0,5 ∗ Age***. The ratio of edits by administrators corresponds to editions made by administrators out of the total editions. To obtain more reliable data, all language administrators should be considered, disregarding their activity status. The item's age is evaluated in days and corresponds to the difference between the collection date and the item's creation date.

Currency corresponds to "the age of an information object" in days, computed by the difference between the collection date and the date of the last edition made to the article.

Finally, volatility is defined as "the amount of time the information remains valid". It corresponds to the median number of hours the content was visible until a later edition reverted it and can be defined as: ***Volatility = Median Revert Time***.

## 2.3 Quality of health-specific contents

Given the importance of health-related subjects and the high use of Wikipedia in this field, the quality of Wikipedia's health information has been an analysis object. Works have been conducted in several medical specialties such as oncology [23], nephrology [28], neurosurgery [19], and anatomy [17, 27].

As a scientific area in constant evolution, the articles' age can reveal outdated information. It is one of the measures used in the works of Conti *et al.* [3] and Suwannakhan *et al.* [27]. Another measure that may reflect the article's update according to the science up-to-dateness is the number of editions, used by several authors [3, 23, 27].

Reliability is another essential characteristic of health-related information, usually assessed through the number of references in the article [3, 23, 28]. Completeness is also expected from health-related information and is often evaluated through the article's size using, for example, the number of words [3, 27].

A large number of studies [3, 19, 23, 27, 28] consider readability as crucial for understanding. A very heterogeneous public seeks health-related information, from health professionals, in the various areas it covers, to lay public with more or less knowledge about health subjects. Naturally, a lay audience may have difficulty understanding the information that, in the health domain, is usually associated with lower levels of readability [1]. Besides the instruments mentioned in the above section to assess readability, others exist, such as the Simple Measure of Gobbledygook (SMOG) [14], Gunning Fog Index [11], Coleman-Liau Index [2], or Automated Readability Index (ARI) [24].

Conclusions vary in results, but generally, studies point to a good quality of information on Wikipedia, comparable to other scientifically recognized sources. However, the authors point out some flaws in quality, especially in terms of readability. Regarding methodology, we noticed that health-specific studies often involve manual analysis of content and generic studies use more automatic approaches. Lack of technological skills in health-related researchers may justify this difference. The size of the datasets used in both types of studies also varies due to the automation level involved.

# 3 METHODOLOGY

Our approach is based on five major tasks, as shown in Figure 1. Numbers identify the execution sequence, and arrows represent the information flow.

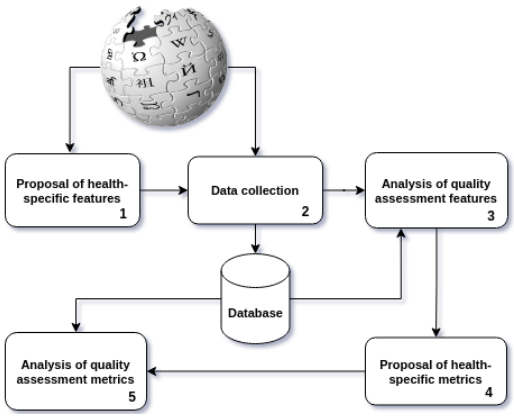


**Figure 1: Methodology**

We began by exploring health-related content features in Wikipedia articles as described in Subsection 4.1. Then, we collected the contents of these articles and the metadata needed to get the required features.We describe this task in Section 3.1. After that, we analyzed the quality of generic and the proposed health-specific features on assessing the articles as described in Section 4. From that, we proposed health-specific metrics that we detail in Subsection 5.2. In the end, we analyzed the quality of the proposed health-specific metrics compared to the generic metrics regarding their ability to assess the quality of Wikipedia health-related articles. These results are described in Subsection 5.3.

We address the following research questions:

(1) What specific features can be used to assess the quality of health-related Wikipedia articles?
(2) Which features are most important in capturing quality?
(3) What specific metrics can be used to assess the quality of health-related Wikipedia articles?
(4) Are specific metrics preferable to generic metrics in the health domain?

## 3.1 Data collection

To obtain a dataset with relevant and updated articles, we collected the top-1000 most viewed health-related pages[5]. The WikiProject Medicine maintain this list and data comes from the Wikimedia Pageview API. We expect to include current relevant articles, such as the COVID-19 pandemic, with this list. Articles from this list

[5]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages

are distributed by WikiProject Medicine quality levels shown in Table 1 as follows: FA (2.9%), GA (8.2%), B (42.8%), C (35%), Start (8.6%), Stub (0.4%), and List (1.4%). Existing works [25, 26] do not usually consider articles evaluated as *stub*, and we followed this same approach. We also discarded articles evaluated as *list* (1.4% of the dataset) for their nature and differences regarding the remaining articles.

Besides classifying articles by quality, Wikipedia also classifies them by importance. It assesses the article's priority for each WikiProject. In WikiProject Medicine, "the purpose of the importance rating is to direct the project's article improvement efforts towards the most important articles, and incidentally to provide a convenient shortlist of important topics for readers who are interested in medicine generally"[6]. In our dataset, articles importance is rated as: 7.2% *Top* importance; 25.4% *High* importance; 43.3% *Mid* importance; 23.4% *Low* importance; 0.7% of the articles are not rated for importance.

Following the approach of Domingues and Teixeira Lopes [6], we used the MediaWiki API to collect the current state of the article's contents and its metadata, revision history, language links, internal wiki links, and external links. Data not available through the API was obtained from the article's markup. Images are a good example of this because the API returns the entire set of images, including those not relevant for the article's content, such as the Wikimedia logo. From the article's markup, it was also possible to extract templates, infoboxes, and citations. Some measurements, such as readability scores, "InfoNoise" and the article's length, require plain text. To achieve so, we removed all the markup from the article's content.

## 4 QUALITY FEATURES

To answer the two first research questions, we propose and analyze health-specific features to assess health-related articles in Wikipedia. Given the relevance of the metrics and respective features proposed by Stvilia *et al.* [26] for the evaluation of Wikipedia's content, described in Section 2.2, we decided to use them in this work as generic features. At the end of this section, we compare generic and specific features in the health domain.

### 4.1 Wikipedia health-specific features

To identify specific characteristics, that can be used to assess the quality of health-related Wikipedia articles, we analyzed several articles from different health and medicine areas. We registered the specific elements common to these pages and tried to understand how they could contribute to quality evaluation.

*4.1.1 Num. health templates.* Templates are elements used to structure information on Wikipedia, allowing several pages to reuse the same element. Simultaneously, templates enable users to have quick, easy, and organized access to information. Templates can be included in any area of a Wikipedia article and are categorized according to their subject, in template categories and subcategories inside them. We only consider health-related templates. Figure 2 shows an example of a health-related template – the medical classifications template – for the Coronavirus disease 2019 page, with medical codifications.


**Figure 2: Medical resources template for Coronavirus disease 2019 Wikipedia page**

*4.1.2 Num. health infobox values.* Infoboxes are a specific type of template and one of the most recurring types of templates. They are commonly used in health-related articles. Inboxes are a fixed-format table usually available in the top right-hand corner of the pages. These contain facts and statistics relevant to the articles related to this and improve navigation between them. Infoboxes can also include metadata. They are a way to summarize important aspects in an easy and quick format to read for the user. Figure 3 represents an infobox, also extracted from the Coronavirus disease 2019 page. We only collected health-related infoboxes. Infoboxes contain key/value pairs, the keys being previously defined for each infobox, and each one's values can vary. There is also the possibility of including or not each of the key/value pairs, so counting the number of pairs may indicate the degree of development and, therefore, the article's quality. In health-related articles, these pairs often include symptoms, complications, treatment, and medication, for diseases, or pharmacokinetic data, for medication, as examples.

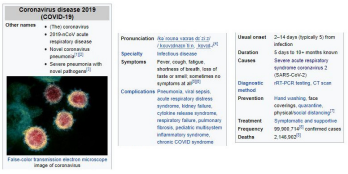


**Figure 3: Infobox present in Coronavirus disease 2019 Wikipedia page**

*4.1.3 Num. health infobox images.* Images can also be included in infoboxes, as shown in Figure 3. As a multimedia element, images enrich the content made available to users, assuming particular relevance in some themes such as health, where they look, for example, for signs of diseases, which are commonly visual. In this case, their number in infoboxes, categorized only in the health-related topic, is counted.

*4.1.4 WPM edit share.* In WikiProject Medicine, there is no training requirement for its members. Still, most of them are doctors, medical students, nurses, scientists, and laypeople with a specific interest in certain medical topics. So, WikiProject Medicine administrators' editions may be an indicator of quality in health-related articles, potentially more reliable than the fact that they are just general Wikipedia administrators. We collected the list of active and inactive administrators. Subsequently, we intersected this list with the list of users responsible for the editions to determine the share of edits made by WikiProject Medicine Admins.

[6]https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Assessment

*4.1.5 TF translated.* The Healthcare Translation Task Force[7] was created as a joint venture between WikiProject Medicine, Wiki Project Med Foundation, and Translators Without Borders. It is a project based on volunteering, counting since 2019 with the help of a translation tool. At the beginning of 2021, it already has over 1,900 articles translated into more than 90 languages. The selection of articles to be translated can be an indicator of a higher quality of these articles. We collected the list of articles already translated by Task Force Translation and intersected it with our dataset.

*4.1.6 Num. medical codes.* A particular feature of health-related articles is the link to medical classifications. Medical classifications aim to code medical diagnoses or procedures. An example of it is the International Statistical Classification of Diseases and Related Health Problems - ICD [20]. The codes, which may be present in the different templates, such as the example in Figure 2, were collected and counted for each article. Wikipedia gathers a list of codes that may be included in templates[8].

*4.1.7 Num. reputed Links.* A commonly used metric for assessing the quality of information on Wikipedia is external links, guaranteeing the information's reliability. However, the number of these links is not, in itself, a guarantee of this reliability since authority is not guaranteed. Thus, we propose to estimate the reliability of these links using the information sources' reputation. To analyze this reputation, we scrapped those suggested by the National Institute of Health[9], a part of the U.S. Department of Health and Human Services. We later matched it with the list of external links from each article.

*4.1.8 Num. recommended sections.* Article length is often a measure used by authors to assess the quality of an article [3, 26, 27]. However, quantity is not a synonym for quality. To evaluate the text's semantics' quality, a manual validation is usually done, or, using tools, which always imply a manual evaluation, as is the case with the DISCERN tool. As a way of automatically evaluating the semantics, albeit slightly, and at the same time the correct structuring of the text, we propose to assess the different sections of the articles, checking which of these sections are in a list of recommended sections in the WikiProject Medicine guidelines[10].

## 4.2 Analysis of generic features

To analyze which generic features are most important in capturing quality, we computed the features used by Stvilia *et al.* [26] in our dataset. As the dataset does not follow a normal distribution, we computed each feature's median in the dataset. We then analyzed the correlation of each feature with the quality levels. For this purpose, we used Spearman's rank correlation since the data is ordinal. We converted each quality level to a numerical value, from 1 - Start to 5 - FA. Table 2 shows each feature's median value, its correlation with quality, the $p$-value obtained in a standard test of the null hypothesis that the correlation is zero. We apply the Bonferroni correction to the $p$-values to account for multiple hypothesis tests and indicate statistical significance. We present features in decreasing order of correlation with quality.

**Table 2: Median value of generic features and their correlation with quality**

| | Median | Correl. | $p$-value |
|---|---|---|---|
| **Num. Reg. Edits** | 1115.0 | 0.53 | < 2.2e-16** |
| **Num. Edits** | 1729.0 | 0.52 | < 2.2e-16** |
| **Connectivity** | 131.5 | 0.50 | < 2.2e-16** |
| **Num. Unique Editors** | 802.0 | 0.49 | < 2.2e-16** |
| **Num. Ext. Links** | 141.0 | 0.49 | < 2.2e-16** |
| **Num. Anon. Edits** | 550.5 | 0.47 | < 2.2e-16** |
| **Num. Reverts** | 148.5 | 0.47 | < 2.2e-16** |
| **Article Length** | 24291.5 | 0.43 | < 2.2e-16** |
| **Age** | 6726.5 | 0.38 | < 2.2e-16** |
| **Num. Images** | 13.0 | 0.37 | < 2.2e-16** |
| **Diversity** | 0.5 | -0.32 | < 2.2e-16** |
| **Admin Share** | 0.2 | 0.31 | < 2.2e-16** |
| **Num. Inner Links** | 388.0 | 0.29 | < 2.2e-16** |
| **Median Rev. Time** | 11.0 | -0.28 | < 2.2e-16** |
| **Kincaid** | 17.5 | -0.25 | 8.473e-15** |
| **Flesch** | 27.0 | 0.20 | 7.241e-10** |
| **Num. Broken Links** | 1.0 | 0.13 | 2.471e-5** |
| **InfoNoise** | 0.88 | 0.13 | 5.989e-5* |
| **Currency** | 6.0 | -0.04 | 0.21 |

* significance level $p$<2.6e-3, ** significance level $p$<5.3e-5. (Bonferroni corrected from $p$=0.05 and $p$=0.001, 19 tests)

From the values presented in Table 2, we conclude that correlation values vary from a negligible correlation of -0.04 for the currency feature to a moderate correlation of the number of registered users edits (0.53). To describe strength of the correlation values, we adopt the scale and terminology used by Prion and Haerling [22]. All features except currency have a correlation value significantly different from 0. Among these, all but InfoNoise are significant at $p$=0.001.

As expected, Kincaid grade level and diversity have a negative correlation, as they are subtractive pairs. Currency and median revert time also have negative correlations, as lower values should correspond to higher quality. Results associated with the currency feature might be related to the fact that we are working with the topmost viewed articles, with a median of 6 days. Note that in the Stvilia *et al.* [25] median for articles with grade-level FA was three days, while the median for a set of random articles was 46 days.

Analyzing the medians values, we can note some particularities, such as the high rate of registered editions among the total editions. Another particularity worth mention is the number of unique editors. Stvilia *et al.* [25] computed a median of 108 unique editors for the FA grade-level set and only five for the random set; in our dataset, we computed a median of 802 unique editors.

## 4.3 Analysis of health-specific features

To investigate the importance of the specific features proposed in Section 4.1 in capturing quality, we conducted an analysis similar to the one described in Section 4.2, computing the median of each feature and its Spearman correlation with quality. Results are shown in Table 3, decreasingly ordered by correlation. Moreover,

[7] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Translation_task_force
[8] https://en.wikipedia.org/wiki/Template:Medical_resources
[9] https://www.nlm.nih.gov
[10] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Medicine-related_articles

we conducted an analysis of the distribution of each feature by evaluation level. These distributions are shown in the form of boxplots, in Figure 4, where an "X" represents the mean.

**Table 3: Median values of specific features and their correlation with quality**

| | Median | Correl. | *p*-value |
|---|---|---|---|
| **Num. reputed links** | 46 | 0.52 | < 2.2e-16** |
| **Num. rec. sections** | 5 | 0.40 | < 2.2e-16** |
| **TF translated** | 0 | 0.32 | < 2.2e-16** |
| **WPM edit share** | 0.05 | 0.25 | 1.9e-15** |
| **Num. health templates** | 3 | 0.23 | 4.2e-13** |
| **Num. medical codes** | 0 | 0.23 | 1.1e-12** |
| **Num. health inf. values** | 8 | 0.21 | 8.2e-11** |
| **Num. health inf. images** | 1 | 0.19 | 9.6e-10** |

* significance level $p$<6.3e-3, ** significance level $p$<1.3e-4. (Bonferroni corrected from $p$=0.05 and $p$=0.001, 8 tests)

The correlation analysis shows that values are relatively homogeneous, varying from the minimum of 0.19 for the number of infobox images, to the maximum of 0.52 for number of reputed links, meaning a moderate correlation. Number of reputed links has a correlation value equivalent to that of the second most correlated feature (num. Edits) of the generic features, as described in Table 2. Note that translated and medical codes have a median value of 0, as most of the articles are not in the translated considered list and have no medical codes. All features have a correlation value significantly different from 0 at $p$=0.001.

The boxplots allow a closer look at the differences by quality level. In general, the mean values shown in the boxplots tend to decrease as quality increases. Exceptions to this sometimes occur, as is the WPM Edit Share case, between the first two levels – FA and GA. This may indicate that these two types of articles are very similar to each other, so the distinction is difficult to make. On the other hand, we can notice the fact that the last level of quality - Start - is the one that generally distinguishes itself more from the rest. In the analysis of the boxplots, two of them differ from the remaining, related to the *Num. health infobox images* and the *TF Translated*. The first case is due to the little variation of the values - from zero to the maximum of two images and the second is caused by the binary nature of the variable - 1 if translated, 0 if not translated. In these cases, the mean, represented in the graph, provides a clearer image of the tendency. We can see that there is a big difference in the two last levels C, and Start - regarding these aspects, influencing these results.

# 5 PROPOSAL OF HEALTH-SPECIFIC METRICS

We proposed specific metrics that can be used to assess the quality of health-related Wikipedia articles and later, we evaluated them and compared them to generic ones, proposed by Stvilia *et al.* [26], in the health domain.

## 5.1 Features importance for generic metrics

Metrics proposed by Stvilia *et al.* [26] were calculated according to the formulas set out in the Section 2.2. We computed the Pearson correlation between each metric and its features to determine each feature's overall contribution to the final value of each metric. Table 4 shows features organized by metric, with correlation values and the $p$-value obtained in a standard test of the null hypothesis that the correlation is zero.

**Table 4: Correlation of metrics with their features**

| | Correlation | *p*-value | Metric |
|---|---|---|---|
| **Num. Edits** | 0.99 | < 2.2e-16** | Authority |
| **Num. Un. Editors** | 0.95 | < 2.2e-16** | |
| **Num. Reg. Edits** | 0.93 | < 2.2e-16** | |
| **Num. Anon. Edits** | 0.92 | < 2.2e-16** | |
| **Num. Reverts** | 0.89 | < 2.2e-16** | |
| **Num. Ext. Links** | 0.60 | < 2.2e-16** | |
| **Connectivity** | 0.28 | < 2.2e-16** | |
| **Article Length** | 1.00 | < 2.2e-16** | Completeness |
| **Num. Inner Links** | 0.31 | < 2.2e-16** | |
| **Num. Broken Links** | 0.02 | 0.49 | |
| **Flesch** | 1.00 | < 2.2e-16** | Complexity |
| **Kincaid** | -0.95 | < 2.2e-16** | |
| **Num. Images** | 1.00 | < 2.2e-16** | Informativeness |
| **Diversity** | -0.27 | < 2.2e-16** | |
| **InfoNoise** | 0.08 | 1.1e-2* | |
| **Age** | 1.00 | < 2.2e-16** | Consistency |
| **Admin Share** | 0.17 | 7.31e-8** | |
| **Currency** | 1.00 | < 2.2e-16** | Currency |
| **Median Rev. Time** | 1.00 | 5.3e-2 | Volatility |

* respective significance levels: $p$< [7.1e-3, 1.6e-2, 2.5e-2, 1.6e-2, 2.5e-2, 5e-2, 5e-2], ** respective significance level $p$< [1.4e-4, 3.3e-4, 5e-4, 3.3e-4, 5e-4, 1e-3, 1e-3]. (Bonferroni corrected from $p$=0.05 and $p$=0.001, [7, 3, 2, 3, 2, 1, 1] respective tests)

We can see significant heterogeneity in correlation values. There are features with a very strong correlation, including values of 1, showing a high contribution to the metric final value. From these strong correlations, only the correlation of the median revert time with volatility is not significantly different from 0.

On the other hand, there are negligible correlations, such as the number of broken links (0.02) and InfoNoise (0.08). The first is not significantly different from 0 but InfoNoise is significantly different from 0 at a $p$-value of 0.05.

Consistency represents a notable case, as the two constituting features achieved different results; age feature had a very strong correlation (1), while the administrators share had a negligible value of 0.17. For currency and volatility, features had a value of 1, probably influenced by having only one feature. Yet, the correlation of the median revert time with volatility is not significantly different from 0.

## 5.2 Health-specific metrics

To propose specific metrics that can be used to assess the quality of health-related Wikipedia articles, we adapted the generic metrics of Stvilia [26], adding or replacing features.

To obtain each feature's weight, we considered its importance and specificity to evaluate the quality of health information. We also considered the weight of the features included in each generic metric and assigned each proposed feature weight, according to

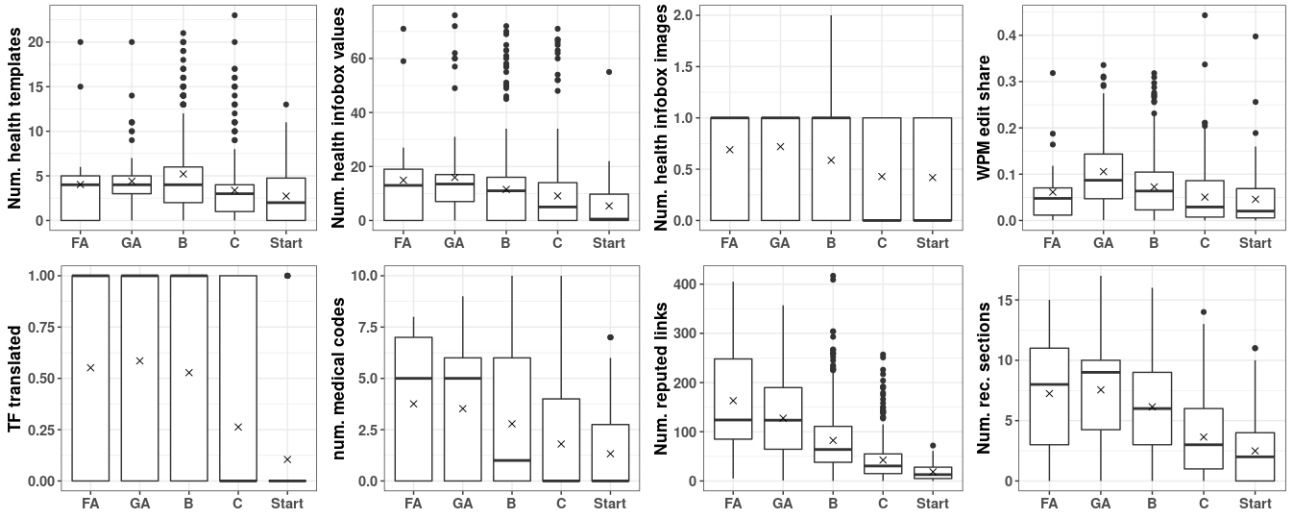


**Figure 4: Distributions of health-specific features by quality level**

the computed median value, so that the final result for that metric matches the same range of values as the existing ones. For example, the administrator's edit share had a median value of 0.2 and a weight of 0.6; replacing WPM edit share had a lower median - 0.05, so the weight raised to 1.9. These values also benefited from a health professional's opinion - one of the authors of this work - a nurse.

Complexity, currency, and volatility remained unchanged since none of the proposed features falls within these metrics. The remaining are proposed as follows:

***HealthAuthority*** *= 0,2 ∗ Num. Unique Editors + 0,2 ∗ Num. Edits + 0,1 ∗ Connectivity + 0,3 ∗ Num. Reverts + 0,6 ∗ Num. Reputed Links + 0,1 ∗ Num. Registered User Edits + 0,2 ∗ Num. Anonymous User Edits + 200 ∗ TF Translated*

***HealthCompleteness*** *= 0,4 ∗ Num. Internal Broken Links + 0,4 ∗ Num. Internal Links + 0,2 ∗ Article Length + 970 ∗ Num. Recommended Sections + 4850 ∗ Num. Medical Codes*

***HealthInformativeness*** *= 0,6 ∗ InfoNoise - 0,6 ∗ Diversity + 0,3 ∗ Num. Images + Num. Health Infobox Values + Num. Health Infobox Images + 0,08 ∗ Num. Health Templates*

***HealthConsistency*** *= 1,9 ∗ WPM Edit Share + 0,5 ∗ Age*

We added translated TF in HealthAuthority, as the selection for translation by the Task Force might reinforce the authority of that article. Reputed links have replaced external links to filter the external links by their reputation in the health area. In HealthCompleteness, we added the number of recommended sections, and the number of medical codes as both features may indicate a satisfactory extent of the information. In HealthInformativeness, we added the number of images and values in health-related infoboxes and the number of medical templates because it addresses the amount of information in an article. For HealthConsistency, Administrators Edit Share has been replaced by WPM Edit Share, representing health-related administrators.

## 5.3 Evaluation

To evaluate specific metrics and compare them to generic metrics in the health domain, we computed the correlation of both types of metrics with Wikipedia's quality levels in each dimension of analysis: authority, completeness, informativeness, and consistency. We have also conducted statistical tests to compare the two correlations for each dimension [5]. Correlation values and the *p*-values associated with comparisons are presented in Table 5.

**Table 5: Correlation of generic and specific metrics with quality**

| | Generic | Specific | *p*-value |
|---|---|---|---|
| **(Health)Authority** | 0.43 | 0.46 | 0** |
| **(Health)Completeness** | 0.34 | 0.36 | 0.58 |
| **(Health)Informativeness** | 0.13 | 0.23 | 9.0e-4** |
| **(Health)Consistency** | 0.30 | 0.30 | 1.0 |

* significance level $p<0.05$, ** significance level $p<0.001$.

From the table's analysis, we concluded that we improved all metrics, although with heterogeneous differences. The metric that showed the most marginal improvements (rounded to zero) was the HealthConsistency. This result is most likely associated with the fact that the weight of the changed feature - WPM edit share - is too insignificant in the formula's total, as already pointed out in Section 4.2. At the other extreme, HealthInformativeness, represents a very significant improvement, revealing the likely importance of templates and their characteristics in assessing the quality of Wikipedia articles. Results suggest that specific metrics can be preferable to generic metrics in the health domain.

## 6 CONCLUSION

We describe an on-going work on quality metrics in the health domain.

To answer our first research question, we explored what specific features can be used to assess health-related Wikipedia articles' quality. In this regard, we proposed eight features: number of health templates, number of health infobox values, number of health infobox images, WikiProject Medicine edits share, Task Force translated articles, number of medical codes, number of reputed links, and number of recommended sections.

After this proposal, we have analyzed which features are most important in capturing quality, our second research question. We conclude that the number of editions made by registered users, connectivity, and total editions are the top three generic features to assess the quality of articles. These positions are occupied by the number of reputed links, the number of recommended sections, and articles translated by the Task Force, for specific features.

Based on this analysis, we explored what specific metrics can be used to assess the quality of health-related wikipedia articles. Working on top of generic metrics, we proposed four metrics: HealthAuthority, HealthCompleteness, HealthInformativeness, and HealthConsistency.

In the end, we assessed the proposed metrics and compared them with generic metrics. We concluded that it is possible to improve the quality assessment of medical articles on Wikipedia using specific metrics. HealthInformativeness and HealthAuthority represent two statistically significant improvements.

A more in-depth analysis could lead to adjustments in the proposed metrics, including new features or assigning different weights to features. In future work, we plan to use the health-specific features and metrics here proposed to automatically classify the quality of Wikipedia articles in the health domain.

## REFERENCES


[1] H. Antunes and C. T. Lopes. 2019. Readability of web content. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*. 1–4. https://doi.org/10.23919/CISTI.2019.8760889

[2] Meri Coleman and T. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology* 60 (04 1975), 283–284. https://doi.org/10.1037/h0076540

[3] Riccardo Conti, Emanuel Marzini, Angelo Spognardi, Ilaria Matteucci, Paolo Mori, and Marinella Petrocchi. 2014. Maturity assessment of Wikipedia medical articles. *Proceedings - IEEE Symposium on Computer-Based Medical Systems* (2014), 281–286. https://doi.org/10.1109/CBMS.2014.69

[4] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. 2015. Measuring article quality in Wikipedia using the collaboration network. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015* (2015), 464–471. https://doi.org/10.1145/2808797.2808895

[5] Birk Diedenhofen and Jochen Musch. 2015. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE* 10 (04 2015), e0121945. https://doi.org/10.1371/journal.pone.0121945

[6] Gil Domingues and Carla Teixeira Lopes. 2019. Characterizing and comparing Portuguese and English Wikipedia medicine-related articles. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019* (2019), 1203–1207. https://doi.org/10.1145/3308560.3316758

[7] En.Wikipedia.org. [n.d.]. Wikipedia:Researching with Wikipedia. Retrieved jan 26, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Researching_with_Wikipedia

[8] En.Wikipedia.org. [n.d.]. Wikipedia:Size of Wikipedia. Retrieved jan 26, 2021 from https://wp1.openzim.org/#/project/Medicine/articles

[9] En.Wikipedia.org. [n.d.]. Wikipedia:Size of Wikipedia. Retrieved jan 26, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[10] R FLESCH. 1948. A new readability yardstick. *The Journal of applied psychology* 32, 3 (June 1948), 221—233. https://doi.org/10.1037/h0057532

[11] R. Gunning. 2021. The techniques of Clear Writing. (02 2021).

[12] Aaron Halfaker and R.Stuart Geiger. 2019. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. (09 2019).

[13] J. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

[14] G. LAUGHLIN. 1969. SMOG Grading —- a New Readability Formula. *Journal of Reading* 12 (01 1969).

[15] Michaël R. Laurent and Tim J. Vickers. 2009. Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association* 16, 4 (2009), 471–479. https://doi.org/10.1197/jamia.M3059

[16] Xinyi Li, Jintao Tang, Ting Wang, Zhunchen Luo, and Maarten de Rijke. 2015. Automatically assessing wikipedia article quality by exploiting article–editor networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9022 (2015), 574–580. https://doi.org/10.1007/978-3-319-16354-3_64

[17] Daniel A. London, Steven M. Andelman, Anthony V. Christiano, Joung Heon Kim, Michael R. Hausman, and Jaehon M. Kim. 2019. Is Wikipedia a complete and accurate source for musculoskeletal anatomy? *Surgical and Radiologic Anatomy* 41, 10 (2019), 1187–1192. https://doi.org/10.1007/s00276-019-02280-1

[18] Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2019. An Edit-centric Approach for Wikipedia Article Quality Assessment. (2019), 381–386. https://doi.org/10.18653/v1/d19-5550 arXiv:1909.08880

[19] Omeed Modiri, Daipayan Guha, Naif M. Alotaibi, George M. Ibrahim, Nir Lipsman, and Aria Fallah. 2018. Readability and quality of wikipedia pages on neurosurgical topics. *Clinical Neurology and Neurosurgery* 166, January (2018), 66–70. https://doi.org/10.1016/j.clineuro.2018.01.021

[20] World Health Organization. [n.d.]. International Statistical Classification of Diseases and Related Health Problems (ICD). Retrieved jan 26, 2021 from https://www.who.int/standards/classifications/classification-of-diseases

[21] Sud ouest. [n.d.]. Léon-Robert de L'Astran, celui qui n'a jamais existé. Retrieved jan 26, 2021 from https://www.sudouest.fr/2010/06/07/leon-robert-de-l-astran-celui-qui-n-a-jamais-existe-110539-7.php

[22] Susan Prion and Katie Haerling. 2014. Making Sense of Methods and Measurement: Spearman-Rho Ranked-Order Correlation Coefficient. *Clinical Simulation in Nursing* 10 (10 2014), 535–536. https://doi.org/10.1016/j.ecns.2014.07.005

[23] Malolan S. Rajagopalan, Vineet K. Khanna, Yaacov Leiter, Meghan Stott, Timothy N. Showalter, Adam P. Dicker, and Yaacov R. Lawrence. 2011. Patient-Oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database. *Journal of Oncology Practice* 7, 5 (2011), 319–323. https://doi.org/10.1200/jop.2010.000209

[24] E Smith and R Senter. 1967. Automated Readability Index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)* iii (06 1967), 1–14.

[25] Besiki Stvilia, Michael Twidale, Linda Smith, and Les Gasser. 2005. Assessing Information Quality of a Community-Based Encyclopedia. *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005* (01 2005).

[26] B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. 2005. Information quality in a community-based encyclopedia. *Knowledge Management: Nurturing Culture, Innovation, and Technology-Proceedings of the 2005 International Conference on Knowledge Management* (2005), 101–113.

[27] Athikhun Suwannakhan, Daniel Casanova-Martínez, Laphatrada Yurasakpong, Punchalee Montriwat, Krai Meemon, and Taweetham Limpanuparb. 2019. The Quality and Readability of English Wikipedia Anatomy Articles. *Anatomical Sciences Education* 13 (2019), 1–13. https://doi.org/10.1002/ase.1910

[28] Garry R. Thomas, Lawson Eng, Jacob F. de Wolff, and Samir C. Grover. 2013. An Evaluation of Wikipedia as a Resource for Patient Education in Nephrology. *Seminars in Dialysis* 26, 2 (2013), 159–163. https://doi.org/10.1111/sdi.12059

[29] Wikipedia. 2020. Wikipedia:WikiProject Medicine/Popular pages. Disponível em https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages.

[30] Kewen Wu, Qinghua Zhu, Yuxiang Zhao, and Hua Zheng. 2010. Mining the factors affecting the quality of Wikipedia articles. *Proceedings - 2010 International Conference of Information Science and Management Engineering, ISME 2010* 1, 1 (2010), 343–346. https://doi.org/10.1109/ISME.2010.114

[31] Xiaolan Zhu and Susan Gauch. 2000. Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 288–295. https://doi.org/10.1145/345508.345602