Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language

James M. Heilman, University of British Columbia
Andrew G. West

Available at: http://works.bepress.com/andrew_g_west/38/
Abstract

**Background:** Wikipedia is a collaboratively edited encyclopedia. One of the most popular websites on the Internet, it is known to be a frequently used source of healthcare information by both professionals and the lay public.

**Objective:** This document quantifies:

1. The amount of medical content on Wikipedia
   a. By number of articles
   b. By number of bytes
2. The citations supporting Wikipedia's medical content
3. The readership of medical content
   a. Comparison between Wikipedia and other healthcare websites
   b. Comparison between Wikipedia's natural language editions
   c. Correlation of Wikipedia article traffic and disease prevalence
4. The quantity/characteristics of Wikipedia's medical contributors
   a. Year-over-year analysis of editor numbers
   b. Contributor demographics and background via survey

**Methods:** Using a well-defined categorization infrastructure we identify medically pertinent English language Wikipedia articles and links to their foreign language equivalents (Objective 1). With these, Wikipedia's API can be queried to produce metadata and full texts for entire article histories (Objective 1-2). Wikipedia also makes available hourly reports that aggregate reader traffic at per-article granularity (Objective 3). An online survey was used to determine the background of contributors (Objective 4). Standard mining and visualization techniques (e.g., aggregation queries, cumulative distribution functions, and/or correlation metrics) are applied to each of these datasets. Analysis focuses on year-end 2013, but historical data permits some longitudinal analysis.
**Results:** Wikipedia’s medical content (at the end of 2013) is made up of more than 155,000 articles and 1 billion bytes of text across more than 255 languages. This content is supported by more than 950,000 references. Content was viewed more than 4.88 billion times in 2013. This makes it one of -- if not the most viewed -- medical resource(s) globally. The core editor community numbers less than 300 and has declined over the past 5 years. The members of this community are half health care providers and 85% have a university education.

**Conclusions:** While Wikipedia has a considerable volume of multi-lingual medical content that is extensively read and well-referenced, the core group of editors that contribute and maintain that content is small and shrinking in size.

**Keywords:** Wikipedia; wikis; Web 2.0; collaborative applications; user generated content; traffic analysis; multilingual content.

**Introduction**

Wikipedia is a multilingual, online, open-source encyclopedia which anyone with Internet access can edit. It is available in more than 275 languages and contains more than 32 million articles across a tremendously broad topic space [1]. While a considerable amount is known about the volume of content, readership, and editor population of Wikipedia as a whole, less is known about these aspects as they pertain to Wikipedia articles in the medical domain. Moreover, non-English language editions are dramatically understudied in comparison to the larger and more popular English version.

In January of 2014 Wikipedia was referred to as “the single leading source of medical information for patients and healthcare professionals” by the IMS Institute for Healthcare Informatics [2]. It is used as a source of healthcare information by 50% to 70% of physicians [3, 4] and has been reported as being the single most used resource by medical students (94%) [5]. A 2013 U.S. survey found people spend more than 52 hours a year searching for health information online, with 22% reporting using Wikipedia [6]. Wikipedia’s readership is also affected by current events, whether popular culture [7] or disease outbreaks [8, 9]. As Wikipedia’s health content is extensively read by the general public and in communities of practice, its authorship and reliability are important qualities. Additionally, quantifying topic popularity can help focus improvements towards greater impact.

With respect to measures of quality, the small amount of available research comes to differing conclusions [10]. In two small samples, Wikipedia’s accuracy was found to be similar to that of UpToDate, eMedicine and National Cancer Institute’s Physician Data Query (PDQ) comprehensive cancer database [10]. A narrow look at pharmacological articles assessed Wikipedia’s accuracy to be high based on significant overlap with textbook sources [11]. Other research found a selection of 50 English medical articles to be relatively well cited [12]. Since 2010 the number of health science academic articles using Wikipedia as a citation has increased.
substantially [13]. Differing research has found Wikipedia’s coverage to be incomplete or less than that of professional sources [10]. A paper examining gastroenterology articles from 2013 found insufficient discussion of the mechanisms of disease [14]. A comparison of pediatric otolaryngology articles between Wikipedia, MedlinePlus, and eMedicine found Wikipedia had a similar accuracy to MedlinePlus, but less than that of eMedicine [15].

In our subsequent analysis we will report on the amount of medical content on Wikipedia. This includes determining the number of references supporting this content and how this quantity has changed over the last 5 years. Readership for both English and non-English versions in 2013 will be analyzed, along with an attempt to determine how the popularity of Wikipedia’s medical content compares to that of other well-known Internet healthcare sites. We will determine if the most commonly viewed articles are those that cover major global health problems or more obscure ones. Finally, the size and makeup of the core editor community will be examined, including how this has changed since 2009.

**Methods**

**Amount of Wikipedia Medical Content**

In order to quantify the number of medical articles and the amount of content within them, one must first determine the subset of Wikipedia which is medically relevant. Wikipedia has a category hierarchy that is built collaboratively, similar to how its core content is amassed and refined. These categories are the basis for identifying medical articles, drawn from the tagging work of “WikiProject Medicine” [16], which identified those English articles that fall within its project’s scope.

Examples of medical articles include: medical diseases and syndromes, medical procedures and diagnostic tests, medications and drugs, and articles related to the history of medicine. Some fitness, pathogenic, and microbiology topics are also categorized as medical; notable healthcare workers also often meet the threshold. However, articles for anatomy, individuals with specific conditions, pharmaceutical companies, and hospitals tend not to be categorized as “medical” being usually well covered by other projects [17].

To identify non-English language equivalents for English articles we rely on the “inter-language link” infrastructure. Also collaboratively built, these links build a graph of all articles – across all language editions – corresponding to a shared topic. Prior to 2013 these links were annotated in the articles themselves, in a distributed fashion. Throughout 2013 these links were migrated to a centralized location (“WikiData”) for ease of maintenance. When we measure the amount of content (in bytes) we account for this migration, else it would appear articles were losing content, when in fact duplicate content was just being more efficiently stored.
Determining the size of a language’s medical article membership is straightforward aggregation. We note our analysis only reports on “article” content, not the discussion or policy-based pages that surround it. Programmatic access to category and inter-language data is available via the Wikimedia API [18]. That same API permits us to obtain an article’s full content at any historical timestamp. We use snapshots from start-2013 and end-2013 in order to plot the byte-growth of medical content, measuring only textual content in this manner.

**Citations Supporting Wikipedia’s Medical Content**

One marker to estimate the quality of Wikipedia’s content is the number of references present in articles and the reputation of those sources being referenced. Leveraging the ability to obtain an article’s full content at any timestamp (as per the last section), we parse that content for standardized citation templates (i.e., the “<ref>” notation). Counting template usage is straightforward, and article snapshots at end-of-year 2009-2013 were used to analyze longitudinal trends. The citation templates also contain a “source” field. We use this to analyze the relative citation counts of leading medical journals, bearing in mind that non-standardized naming and abbreviation conventions (“New England Journal of Medicine, “NE Journal of Medicine.”, “NE J. Med.”) inhibit precise aggregation. In particular, we highlight citations to Cochrane reviews, as they are a highly regarded source. Parsers based on regular expressions were used in reference counting and source extraction.

**Readership of Wikipedia’s Medical Content**

Readership of specific articles and medical content in total is derived from the hourly “page view” aggregates [19] made available by the Wikimedia Foundation (WMF). These are large plaintext files where each line contains a language, article title, and view count -- with a single day’s volume (24 files) on the order of 10GB in size. We have authored scripts to obtain and process these files nightly, writing daily aggregates to a persistent database table indexed by language and article.

We emphasize that these files report only *desktop* views. However, mobile views are reported at project-scale [1, 20] (e.g., for all of English/French/Spanish Wikipedia), permitting some rough estimates if one assumes mobile traffic is uniformly proportional across all articles. An examination of the phenomena underlying this collection and broader readership trends can be found in [7].

Our database of daily views can be queried to produce aggregates by language, specific article, or the “topics” that span multiple language equivalents. To compare Wikipedia’s medical readership to that of other common healthcare websites we use SimilarWeb [21], a traffic measurement service. We multiply the “estimated visitors” and “page views per visit” metrics that service provides to produce a “page views” statistic comparable to the one reported by the WMF.
**Figure 1:** Relative amount of population/content by natural language group. “World by Language” is based on 2007-10 data per [23]; “Internet by Language” is derived per [24]; “Wikipedia by Language” is based on independent calculations over medical articles by language edition. The 10 European languages are German, French, Spanish, Polish, Italian, Portuguese, Russian, Dutch, Swedish, and Catalan.

To measure topic readership variance between languages we identify and rank (by traffic) a core set of equivalent articles that exist in all of Wikipedia’s 10 largest language editions. We first analyze these by topic, finding anomalous popularity patterns and outliers. For a more aggregate comparison we also calculate the Pearson correlation-coefficient between all language pairs.

We also sought to determine if diseases of greater global severity are more frequently viewed Wikipedia topics. To do so, we took the top 20 diseases by “disability adjusted life years” (DALYs) and the top 20 diseases by “years lived with disability” (YLDs) for 2012 as reported by the World Health Organization [22]; yielding 33 conditions in combination. We then found the 42 corresponding English Wikipedia articles for each disease (some, like “child behavioral disorders” can refer to both “ADHD” and “conduct disorder”). Traffic on these articles was compared against that on a broader set of Wikipedia articles corresponding to diseases, as identified by the presence of a standardized template (“infobox”) which concisely summarizes the identifiers of a disease.

**Quantity/Characteristics of Wikipedia’s Medical Contributors**

Already leveraged for categories and language links, the Wikimedia API also permits one to crawl version histories towards gathering metadata about an article’s editors. Aggregating this across all medical articles (or just those of a particular language), we are able to plot participation at various thresholds. In particular, we identify 274 contributors who made more than 250 edits to medical articles in 2013. In May 2014 we utilized a Wikipedia messaging system to award 271 of these users a “barnstar”, a digital form of peer-to-peer recognition. Posting to user’s “talk” pages, the awards contained a request to complete a survey containing six questions:
Table 1: Wikipedia language editions ranked by number of medical articles (#MAs). Also provided is the amount of textual content in each language, in bytes.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
<th>#MAs</th>
<th>bytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>29,072</td>
<td>241MB</td>
</tr>
<tr>
<td>2</td>
<td>German</td>
<td>7,761</td>
<td>66MB</td>
</tr>
<tr>
<td>3</td>
<td>French</td>
<td>6,372</td>
<td>57MB</td>
</tr>
<tr>
<td>4</td>
<td>Spanish</td>
<td>6,367</td>
<td>51MB</td>
</tr>
<tr>
<td>5</td>
<td>Polish</td>
<td>5,999</td>
<td>28MB</td>
</tr>
<tr>
<td>6</td>
<td>Italian</td>
<td>5,677</td>
<td>44MB</td>
</tr>
<tr>
<td>7</td>
<td>Portuguese</td>
<td>5,269</td>
<td>25MB</td>
</tr>
<tr>
<td>8</td>
<td>Russian</td>
<td>4,832</td>
<td>48MB</td>
</tr>
<tr>
<td>9</td>
<td>Dutch</td>
<td>4,391</td>
<td>18MB</td>
</tr>
<tr>
<td>10</td>
<td>Japanese</td>
<td>4,303</td>
<td>36MB</td>
</tr>
<tr>
<td>11</td>
<td>Arabic</td>
<td>4,055</td>
<td>26MB</td>
</tr>
<tr>
<td>12</td>
<td>Swedish</td>
<td>3,661</td>
<td>12MB</td>
</tr>
</tbody>
</table>

1. What is your highest level of education?
2. Do you currently work in the healthcare field? Or have you previously?
3. Are you currently studying healthcare (a student)?
4. What language of Wikipedia do you mostly work on?
5. Did you receive a barnstar?
6. How do you identify your gender?

Question #5 was used to sanity-check respondents (as “barnstar” awards are public, uninvited participants could traverse the survey link). We also posed an open question: “Why do you edit Wikipedia’s medical content?”

Results

Amount of Wikipedia Medical Content

Number of Articles
Wikipedia had 155,805 medical articles across 255 natural languages at the end of 2013. A further 31 languages did not contain any medical articles per our methodology. Of the 155k articles, 29,072 are in English (18.6%). While a significant portion of Wikipedia’s content (both medical and otherwise) is in English, this imbalance is less than that observed across the broader Internet (Figure 1).

Table 1 presents the top languages by quantity of medical articles. Going beyond this list, the top 10 languages make up 51.4% of the total, while the top 25 languages account for 75%. Figure 2 plots the article quantity distribution, showing it to have a power-law distribution (i.e., few languages have many articles, and vice versa).

Bytes of Content
At the end of 2013 Wikipedia had 1,016MB of textual medical content, up 9.3% from one year earlier when the total was 922MB. English medical articles saw the most growth during this period, gaining some 19.7MB. Assuming the average word has 6 characters, this equates to 3.28M English words added in 2013. If the total
Figure 2: Distribution for the quantity of medical articles in a Wikipedia language edition, presented in rank order (note logscale on y axis).

Figure 3: Estimated volume of Wikipedia’s medical content, if printed.

(combined language) 1,016MB of content were printed in textbooks roughly the size of the Encyclopedia Britannica at 8 million characters per volume, it would consume 126.9 volumes, with Figure 3 visualizing this quantity. English language medical articles are responsible for 23.7% of all medical content (by bytes). The next largest languages per this metric are German, French, Spanish, Russian, Italian, Japanese, Polish, Arabic, and Portuguese (similar, but not identical to Table 1). Together the top ten languages account for 61.2% of all byte content.

Citations Supporting Wikipedia’s Medical Content
As a marker for Wikipedia’s reliability we count the number of references in year-end article versions between 2009 and 2013. This was done for medical portions of both English Wikipedia (Figure 4) and all languages (Figure 5). We found that English references more than doubled from 187,107 to 376,123, while across all languages the increase was more than 2.5× from 373,558 to 952,053. Note that this citation growth ratio significantly outpaces that observed for byte growth.

By parsing a standardized citation format, we are able to determine the journals that are most commonly used as references on Wikipedia are also some of the most respected, including: The Lancet, NEJM, Nature, BMJ, JAMA, Science, and the Cochrane Database of Systematic Reviews. While a lack of standardized naming/abbreviation conventions prevents precise aggregation, we were able to measure references to “Cochrane (reviews)” (Figure 6) across all languages have increased nearly three-fold from 2,717 in 2009 to 7,290 in 2013.
Readership of Wikipedia’s Medical Content

Comparison between Wikipedia and Other Healthcare Websites

Before embarking on traffic comparisons between Wikipedia and other healthcare sites, we first establish Wikipedia’s medical readership in isolation. In 2013, across all languages, Wikipedia’s medical content received 4.88 billion non-mobile views (estimates put the mobile-inclusive total close to 6.5 billion). Around 4.56 billion of these were in the top 12 languages (Table 2), with English accounting for 46.7%.

Medical content accounts for 0.6% of all articles on English Wikipedia, yet these receive 2.5% of all English Wikipedia page views. Similar patterns are observed across many language editions, with medical articles receiving far more than the mean expected traffic. As a portion of all content, among prominent languages, medical readership varies from 1% in Chinese to 4.4% in Spanish; the global percentage across all languages is 2.5% (the same as for English).
Recall that we use web monitoring service SimilarWeb [21] to estimate the traffic received at other healthcare websites. Despite having precise page view data for Wikipedia’s medical portions, in the interest of fairness, we also derive Wikipedia’s totals from SimilarWeb. That service’s sampling methodology likely introduces bias we would prefer to be uniform across all sites under evaluation. The healthcare sites we examine (NIH, WebMD, Mayo Clinic, NHS, WHO, UpToDate) host exclusively medical content. In contrast, the traffic statistics SimilarWeb reports for the “wikipedia.org” domain must be scaled down to its medical portion (2.5%).

Figure 7 presents the comparison after such adjustments for July 2014, with the light blue portion capturing that SimilarWeb slightly underreports traffic compared to the WMF data (recalling that neither reports mobile views). Regardless, Wikipedia appears to be the most utilized online healthcare information resource.

Comparison between Wikipedia’s Natural Language Editions

The popularity of individual topics across languages varies dramatically. Among the 100 most popular English articles, none of them were unanimously popular (in the top 100) across 9 other prominent languages in which a corresponding article exists. For example, “Down Syndrome” was 3rd most popular in German, 7th most accessed in Italian/Polish, and 17th in English. However, it was outside the top 1,000 in Russian, Japanese, French, Portuguese and Chinese. “Asperger Syndrome” was one of few articles close to being in the top 100 most viewed in all languages.

As enumerated in the caption to Table 5, these nine languages, plus English, comprise the 10 largest Wikipedia editions in 2013 by overall page views. This should not be confused with Tables 1 and 2, which report only on medical portions.
### Table 3: Medical topics with the most traffic summed across languages. View count is for 2013 and we list the number of languages (#LG) with a corresponding article.

<table>
<thead>
<tr>
<th>Topic/Article</th>
<th>Views</th>
<th>#LG</th>
<th>Topic/Article</th>
<th>Views</th>
<th>#LG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asperger syndrome</td>
<td>15.48M</td>
<td>56</td>
<td>Angelina Jolie ...²</td>
<td>9.51M</td>
<td>98</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>13.84M</td>
<td>74</td>
<td>Human</td>
<td>9.47M</td>
<td>165</td>
</tr>
<tr>
<td>Bipolar disorder</td>
<td>13.12M</td>
<td>55</td>
<td>Hemorrhoid</td>
<td>9.42M</td>
<td>73</td>
</tr>
<tr>
<td>Sexual intercourse</td>
<td>13.07M</td>
<td>96</td>
<td>Pneumonia</td>
<td>9.26M</td>
<td>90</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>11.86M</td>
<td>123</td>
<td>Blood type</td>
<td>9.20M</td>
<td>67</td>
</tr>
<tr>
<td>Diabetes mellitus</td>
<td>11.28M</td>
<td>113</td>
<td>HPV</td>
<td>8.97M</td>
<td>39</td>
</tr>
<tr>
<td>Autism</td>
<td>10.43M</td>
<td>80</td>
<td>Down syndrome</td>
<td>8.83M</td>
<td>66</td>
</tr>
</tbody>
</table>

but was nearly 1500th in Russian. Similarly, “Tuberculosis” fared well in all languages except French and Polish. Even a typical stronghold of Internet attention, “sexual intercourse” was only in the top 10 most popular articles for English, where it secured 3rd place. Table 3 presents the most popular topics overall, while Table 4 further highlights popularity variance.

While sometimes regional or cultural trends are observed (e.g., disease effected regions having high popularity for the corresponding article in the local language), a broader explanation of these patterns is a subject for future investigation.

Rather than looking at articles or topics in isolation we can calculate rank similarity between language pairs (Table 5). Working from the set of topics that have articles in all of the top ten language editions, we find Portuguese and Spanish visitors (PCC = 0.668) have the most similar browsing habits, while Russian and English visitors (PCC =0.207) are most dissimilar.

### Correlation of Wikipedia Article Traffic and Disease Prevalence

A 2014 IMS report [2] made the claim that “rarer diseases, which often have fewer available information sources and are less understood by patients and clinicians, show a higher frequency of [Wikipedia] visits than many more common diseases.” Given English is frequently the language used to search for information on Wikipedia regardless of a person’s country of origin, we used the English traffic data to gain perspective on this claim. We found that the articles associated with the 20 conditions having the greatest YLD and the 20 conditions with the greatest DALYs had an average view count of 1.68M in 2013. This compares to an average of

---

² The full article title is "Angelina Jolie Cancer Treatment", which was a newsworthy topic of 2013. However, that title redirects to a sub-section of the "Angelina Jolie" article, effectively making her (broadly popular) article and its (broadly popular) foreign language equivalents part of the medical category, despite the fact many visitors likely arrive there for non-medical reasons.
Table 4: Topics having most and least variable popularity rank across the top 10 languages. Relative variance (R-Var.) is the percentage of the maximum observed variance (i.e., it is not an absolute measure, but based on the variance calculated for the “Down syndrome” article). Popularity rank goes from 1 (most popular) to 1536 (least), as there are 1536 articles in all of the ten languages utilized herein; a constraint that helps to normalize these comparisons.

Table 5: Pearson’s correlation-coefficient (e.g., “rank similarity”) metric for medical topic popularity in ten prominent languages. More informally: “how similar is the popularity ordering for topics between two languages?” The measure is symmetric. Language codes are: EN=English; ES=Spanish, RU=Russian; JA=Japanese; DE=German; FR=French; PT=Portuguese; IT=Italian; PL=Polish; ZH=Chinese.

189,351 views for the 4,791 articles tagged with the disease “infobox”, and 78,000 views for the average English medical article. Clearly, globally prevalent and well-known medical conditions tend to receive considerable traffic.
Such macro-scale correlation is intuitive, but recent research [9] has also demonstrated the more nuanced capability to utilize traffic data for individual articles in near real time. That work found that the popularity of influenza articles not just correlated with the spread of the disease, but could also be temporally analyzed to create reasonably accurate infection forecasts. The extent to which this applies across the entire article base, and the ways the healthcare community can utilize such rapid signaling are topics for future work.

**Quantity/Characteristics of Wikipedia’s Medical Contributors**

**Year-over-year Analysis of Editor Numbers**

Given Wikipedia’s collaborative nature it is logical to investigate the editor community that has authored the content of such a frequently accessed resource. Most often, “editors” in this context are users with a persistent account name and login credentials. Although one may edit without an account, rarely do such users exhibit the consistent participation on which we focus. Of the 274 top contributors, just 4 edited without an account name.

We measure participation by looking at an editor’s quantity of contributions on medical articles in a given calendar year. The following are some participation thresholds measured across all languages: (5+ edits = 21,563 editors); (25+ edits = 5,573 editors); (100+ edits = 1,237 editors); (250+ edits = 274 editors); (1,000+ edits = 39 editors); (10,000+ edits = 1 editor; this paper’s lead author). There were 32 language versions that had at least one editor with 250+ edits in 2013.

Again highlighting the English edition, we can plot these same thresholds on a yearly basis from 2008 to 2013 (Table 8; Table 9). We found that at all participation thresholds the number of editors has decreased. Over this 5 year span the decrease...
Table 6: Survey responses from 121 top medical editors, across all language editions.

in editor numbers was around 40% for English Wikipedia, with 10-20% attrition typical for non-English languages.

Not included in the above totals is the work of non-human, automated, “bot” editors: computer programs that perform much repetitive maintenance. Bots and humans combined made 1,106,575 medical edits in 2013, with 406,003 (36.7%) of those in English. Bots accounted for 24.7% of the global total and 10.5% of the English one, numbers slightly inflated due to the bot-driven migration of inter-language links as described in our “Methodology” section.

Contributor Demographics/Background via Survey
In May of 2014 we sent out a survey to 271 of the 274 top medical editors in 2013. Three users were omitted as they had been blocked from contributing to Wikipedia due to various issues. Of these, 121 (45%) responded, with their answers summarized in Table 6.

We found more than half of editors -- 67 of 121 (55%) -- are either health care professionals or studying health care. Of the 54 outside health care, 17 used the open text area to describe their activities as primarily grammatical, formatting, language simplifications, and the removal of vandalism. Fifteen others reported more substantive editing despite lacking formal medical training. In some cases (two self-reported), contributors were arguably experts despite not being healthcare providers: One is a PhD biochemist and another is a SCUBA diver editing in related medical spaces.
Discussion

Principal Results
Wikipedia's medical content is made up of more than 155,000 articles and 1 billion bytes of text across 255 languages. This content is supported by more than 950,000 references and was viewed more than 4.88 billion times in 2013 (with mobile inclusive estimates at 6.5 billion). Third-party analytics suggests Wikipedia is the most viewed medical resource globally. The core editor community as of 2013 numbers less than 300 and has decreased over the past 5 years. The members of this community are half health care providers and 85% have a university education.

Limitations

Amount of Wikipedia Medical Content
Our analysis depends heavily on the Wikipedia editor community to establish: (1) what constitutes a medically related article, and (2) the inter-language links between corresponding articles. Whether or not something is related to medicine, or related “enough” to justify a tagging is a subjective distinction. Inter-language links are often less ambiguous, but still require a bi-lingual speaker who is familiar with Wikipedia syntax.

Though subjectivity might shift these bounds slightly, more articles have likely never been considered in these contexts, either because they are undiscovered entirely, or they are too emergent, tangential, or unpopular to draw the attention of the editors who typically make category and inter-language annotations. Although usually quickly restored [25], “vandals” also sometimes destroy tags or links with malicious intent.

Following the very nature of collaborative work, it is our subjective experience that “major” topics are more likely to be correctly tagged and linked than more obscure ones. Thus, tagging and linking inaccuracies likely have a greater impact on article quantity measurements than readership totals. Categorization omissions, in particular, could be estimated by searching English Wikipedia using a database of terms such as the ICD10 and verifying that corresponding articles have been appropriately tagged. We leave this as a topic for future research.

Lastly, our analysis uses tagged English articles as the starting point for inter-language link discovery. A medical topic that does not have a corresponding English article version would not be included in our analysis.

Citations Supporting Wikipedia’s Medical Content
Wikipedia strives for verifiable content, rather than the less agreeable notion of absolute “truth.” As such, information drawn from reputable sources upholds the notability and verifiability requirements that Wikipedia promotes.
In this work we quantify the number of references (and highlight some particularly well reputed sources) as a proxy for reliability. We recognize that the number of references is just one mark of quality. Content may be inaccurate despite having a citation and vice versa. Our data does not look at whether or not the text of Wikipedia accurately reflects the sources in question or if the sources are outdated. Both would be interesting questions to investigate further.

**Readership of Wikipedia’s Medical Content**

Language-scale aggregates regarding Wikipedia readership are influenced by the number of member articles. Thus, previously discussed limitations surrounding category tagging and inter-language links also cascade into this analysis.

It is important to emphasize that none of our traffic data (Wikipedia or third-party) includes readership from mobile devices. These shortcomings in the WMF's collection infrastructure were remedied during our writing in October 2014; mobile readership will be analyzed in future work. While making for fair comparison, this also means we underreport the scale at which online healthcare resources operate. Across all of English Wikipedia (not just medical portions), mobile views are 30%+ of the total traffic, and growing [20]. Thus, readership as we present it may underrepresent the browsing habits of certain economies, languages, and regions (e.g., where mobile networks are the only means of connectivity, and/or cellular devices are the only affordable means of access) or certain demographics (e.g., youth demonstrating a preference for mobile browsing).

Moreover, when comparing Wikipedia’s medical readership to other healthcare websites, one must be mindful of the varying coverage and scope. While it would be interesting to compare per-topic page views, alternative sites (some proprietary) have not made publicly available such granular traffic data.

In our broad comparison of readership on healthcare websites we relied on the third-party SimilarWeb [21]. That service’s measurement methodology and accuracy is not known. However, it is reassuring that SimilarWeb’s page view estimates for the entirety of “en.wikipedia.org” differ only by about 3% from the more authoritative data published by the Wikimedia Foundation.

To some extent, all information sources find themselves mirrored across the Internet and combined into other sources. This; however, occurs more frequently with Wikipedia and government sources as they are freely licensed or in the public domain which encourages reuse. Such transitive/downstream consumption (both online and off) is difficult to quantify. For example, low-cost “alternative textbook” provider Boundless.com amasses such open-source content when compiling its texts [26], with some becoming popular in practice [27]. Further, the NIH and Wikipedia often see their content integrated directly into Google search results, and regardless, these sources often have high search-engine ranking [28].
Quantity/Characteristics of Wikipedia’s Medical Contributors
Our survey to medical editors had a response rate around 50%. This raises the concern that those with the time and willingness to complete the questionnaire are somehow non-representative. While roughly half of recipients primarily edit a non-English Wikipedia, our survey was available only in English, potentially limiting and biasing the response pool. Our validation question (“Q5: Did you receive a barnstar?”) also takes respondents at their word, in addition to trusting the feedback received for all other questions.

We identified 4 IP accounts that made more than 250 edits, assuming that those IP addresses are statically assigned to a single contributor. Dynamic IP assignment is common in residential and wireless networks (i.e., DHCP) and could have impacts such that multiple human users inhabit a single IP over time (causing an overestimation on our part), or that a single user’s contributors are unknowingly spread across IP space (an underestimation).

Comparison with Prior Work
Our “Introduction” section enumerates some of the prior research that qualitatively relates to this work. A purely quantitative point of reference comes from the parallel work of Nusa Faric [29], which also surveyed English Wikipedia’s most active medical editors. That research found 50% of those surveyed had a medical background, 70% were over 30 years old, most were male, and 75% had a college degree. All data points are quite similar to our findings, which additionally considered non-English editors.

Conclusions
Amount of Wikipedia Medical Content
While Wikipedia has a tremendous amount of medical content, it is primarily concentrated in English and a few major European languages. As a user-generated website this reflects the populations that are willing and able to contribute. Wikipedia’s distribution of content by language, however, better matches global language popularity than the Internet does as a whole. Additionally there are ongoing efforts to improve Wikipedia’s medical coverage in non-English languages via a partnership with the not for profit, Translators Without Borders.

Citations Supporting Wikipedia’s Medical Content
Wikipedia is relatively well referenced and by this marker is becoming increasingly reliable over time. Encouragingly, references to high quality sources such as the Cochrane collaboration are rising at a greater rate than references on the whole.
Readership of Wikipedia’s Medical Content

A previous IMS report [2] claimed that Wikipedia is the single most used medical resource on the Internet. Our statistical work herein appears to confirm this assertion, with conservative analysis putting Wikipedia’s readership on par with NIH and surpassing that of WebMD (two sites traditionally atop the “health” category). With the Internet likely to be the most consulted information medium, Wikipedia may well be the most used medical resource overall.

Our study unexpectedly found more than a 4× variation in the proportional popularity of health content across different languages. The catalyst for this variation is unclear. Is it the case that Spanish speakers care more about their health than Chinese ones? Or do Chinese populations prefer a different source?

We also found that popular topics/articles differed wildly amongst languages. This has interesting ramifications as emergent language editions try to expand their medical content (either organically or through translation). Simply assuming content that is well-read in one language will draw audiences in another is insufficient, and more careful cultural consideration may be prudent.

Quantity/Characteristics of Wikipedia’s Medical Contributors

While Wikipedia’s medical content has tremendous readership, the number of significantly active contributors is few. It is concerning that these editor numbers, at all thresholds, have decreased over the past 5 years. This trend is one exhibited not just by medical contributors, but the overall Wikipedia community. A number of explanations have been proposed for this poor retention and recruitment: (1) deterrents such as stricter reference requirements and more policy, (2) growing competition for participant attention in the open-source and user-generated content communities, (3) xenophobia and a community unwelcoming of new users [30], and (4) the perception that in some languages there remains little “low hanging fruit” to be authored. Understanding and reversing this trend is an area of active research for Wikipedia and its sub-communities.

The community of medical editors, like Wikipedia overall, is male dominated [31]. The reasons are not entirely clear, but some possibilities include: technical barriers, lack of self-confidence, minimal social activity, and the adversarial nature of some discussions [32]. Efforts to make Wikipedia more female friendly are ongoing.

Our survey of Wikipedia’s medical contributors found many are healthcare professionals and most are university educated. While just 29% of the U.S. population has a Bachelor’s degree [33], 85% of Wikipedia’s core medical editors have attained one (with more than 50% going beyond that level). Educational levels attained were similar between editors for English and non-English versions. These educational and professional benchmarks put into doubt the claims by some that Wikipedia is “anti-expert” [34].
Acknowledgements
Nikita Belavate, Soumyadeep Bhaumik, Abhijith Jayanthi, Manu Mathew, and Jake Orlowitz are thanked for their help on this project. Data analysis was done on machines hosted at the University of Pennsylvania.

Conflicts of Interest
The authors are active contributors to Wikipedia. They have no financial conflicts of interest related to the subject matter of the article.

Appendix
Adjacent (or appended) to the publication of this document should be an appendix, where the authors link to raw data and extend the tables found in this writing. If this supplemental file is not available, please contact the authors directly.

References


17. Wikipedia Community. WikiProject Medicine assessments. 2014: 

Archived at: http://www.webcitation.org/6TXxXXrGN

http://dumps.wikimedia.org/other/pagecounts-rat/. Archived at: http://www.webcitation.org/6TXcLcpq6m

20. Zachte, E. Page views for Wikimedia, all projects, all platforms, normalized. 
Archived at: http://www.webcitation.org/6SFYGxqW9

for Wikipedia, archived at: http://www.webcitation.org/6SG2o8hD ;
for NIH, archived at: http://www.webcitation.org/6SG2tSPG ;
for WebMD, archived at: http://www.webcitation.org/6SG2wYzwE ;
for Mayo Clinic, archived at: http://www.webcitation.org/6SG306pg9 ;
for NHS, archived at: http://www.webcitation.org/6SG32tqhC ;
for WHO, archived at: http://www.webcitation.org/6SG35Vc3 ;
for UpToDate, archived at: http://www.webcitation.org/6SG37Cq2t
   Archived at: http://www.webcitation.org/6SFYiWhnZ
25. West AG. Damage Detection and Mitigation in Open Collaboration Applications. 
27. Weddell C. Feel the love: Educators tell us why they love Boundless. 2014.  
   http://blog.boundless.com/2014/02/educators-on-boundless-teaching-platform/#more-991. Archived at: http://www.webcitation.org/6RIP7bNAQ
   16(11):e260. doi:10.2196/jmir.3569
   10.1371/journal.pone.0065782
   Archived at: http://www.webcitation.org/6SG4EhBSI
Wikipedia and Medicine: Quantifying Readership, Editors, and the Significance of Natural Language (Appendix)

James M. Heilman -- Andrew G. West

Appendix

This appendix is intended to complement the above named paper. It does so primarily by: (1) extending the tables presented in that document, (2) providing raw data for figures presented in that document, and (3) providing additional measurements that were not presented due to space constraints.

Given the scale of these measurements, and the goal of data re-use, we choose to provide these datasets in plaintext format. Here, we provide a brief description of the datasets and link to those files. Unless otherwise indicated, datasets capture calendar year 2013 data regarding Wikipedia’s medical portions:

- **Traffic to individual articles (year sums):** A large CSV file that lists all medical articles in English. Tab intended beneath each English entry are its foreign language equivalents. For all article rows, three columns of data are provided: (1) total 2013 page views, (2) that article’s views as a percentage of the medical view total in the specified language, and (3) that's article's views as a percentage of the project-scale views for the specified language. [http://www.andrew-g-west.com/data/human_readable_yearly.csv.zip](http://www.andrew-g-west.com/data/human_readable_yearly.csv.zip)

- **Traffic to individual articles (monthly breakout):** Identical in format to the above, but after the column showing total page views in 2013 there are now 12 additional columns that provide monthly breakdowns. [http://www.andrew-g-west.com/data/human_readable_monthly.csv.zip](http://www.andrew-g-west.com/data/human_readable_monthly.csv.zip)

- **Number of medical articles and page views by language:** Also calculates the average number of views per article. [http://pastebin.com/D0Z8Xjwg](http://pastebin.com/D0Z8Xjwg)

- **Top n most popular articles, by language:**
  - Top 50 for “popular languages”: [http://pastebin.com/uQgCHAs0](http://pastebin.com/uQgCHAs0)
  - Top 25 for all languages: [http://pastebin.com/EGTn4Ey9](http://pastebin.com/EGTn4Ey9)

- **Most popular topics (across languages):** List containing the 2000 most popular topics (page views summed across corresponding articles, for all languages where an article exists). There are three columns: (1) the
corresponding English article, (2) the sum of all views, and (3) the number of languages that had a corresponding article.  http://pastebin.com/7pPAQ29q

- **Top traffic variance (temporally) for English medical articles:** The 500 English medical articles whose daily page views demonstrated the most variance in 2013. Individual daily totals were provided to the variance calculation. Articles with high variance are often linked to breaking news events, Google Doodles, Reddit threads, etc. Moderate variance is often observed with seasonal diseases and conditions.  
  http://pastebin.com/erLELAqf

- **Top traffic variance, measured across languages:** The 500 topics that have the most traffic variance across the 10 largest Wikipedia editions. The 10 inputs to the variance function are not view counts, but normalized view rankings (i.e., the most popular article in a language receives rank “1”). This measurement captures articles that are very popular in one/some languages, and dramatically less so in others.  
  http://pastebin.com/SJm1p9cg. As a complement to this, we also provide the 500 topics with the least variance.  
  http://pastebin.com/H3eAHfd5

- **Medical content size and growth, 2012-2013:** Per language breakdown of textual byte size of medical articles at year end 2012 and 2013. The impact of Wikidata link migration is also included, so that year-over-year growth can be accurately calculated while excluding that influence.  
  http://pastebin.com/8ia4U6ab

- **Quantity of edits by language/editor thresholds:** The number of edits per language in calendar year 2013. Additional columns enumerate the number of users in that language having {1+, 5+, 10+, 25+, 50+, 100+, 250+, 500+, and 1000+} edits.  
  http://pastebin.com/3j4CWbpS

- **Editors having 250+ med. edits in 2013:** This includes bots, whose names by rule include the “bot” substring  
  http://pastebin.com/GJ3GU75T

- **“Cochrane” references by language/year:**  
  http://pastebin.com/3AJK1L0X

- **Most commonly cited journals:**  
  http://pastebin.com/euvaTiNT

- **Yearly citation growth:**  
  http://pastebin.com/H1tA1mxX

- **Miscellaneous queries:** Some quick queries and sums that include results such as: (1) topics having a corresponding article in most language editions, and (2) project-scale view counts for 2013 (not just medical portions).  
  http://pastebin.com/WsSnC4E2