Mining Consumer Health Vocabulary from Community-Generated Text

V.G.Vinod Vydiswaran, PhD[1], Qiaozhu Mei, PhD[1,2], David A. Hanauer, MD, MS[3,1], Kai Zheng, PhD[4,1]
[1]School of Information; [2]Department of Electrical Engineering and Computer Science; [3]Department of Pediatrics; [4]School of Public Health Department of Health Management and Policy. University of Michigan, Ann Arbor, MI

Abstract

Community-generated text corpora can be a valuable resource to extract consumer health vocabulary (CHV) and link them to professional terminologies and alternative variants. In this research, we propose a pattern-based text-mining approach to identify pairs of CHV and professional terms from Wikipedia, a large text corpus created and maintained by the community. A novel measure, leveraging the ratio of frequency of occurrence, was used to differentiate consumer terms from professional terms. We empirically evaluated the applicability of this approach using a large data sample consisting of MedLine abstracts and all posts from an online health forum, MedHelp. The results show that the proposed approach is able to identify synonymous pairs and label the terms as either consumer or professional term with high accuracy. We conclude that the proposed approach provides great potential to produce a high quality CHV to improve the performance of computational applications in processing consumer-generated health text.

Introduction

Over the past decade, there has been a significant increase in the consumption of online health information by the general public. According to a 2013 Pew Research Center survey, 72% of U.S. adult Internet users have looked for health information online; and, among them, 59% have sought information about specific medical conditions.[1] However, it has been long recognized that laypersons and healthcare professionals think about and express health-related concepts very differently,[2] for example, "dry mouth" vs. "xerostomia" and "flu" vs. "influenza". This mismatch in the terminology and style of writing could diminish laypersons' ability to effectively find and comprehend online health information written by professionals or experienced patients.

On the other hand, community-generated data on the Internet have also been increasingly used as a source of information to support professional needs such as public health surveillance and scientific discovery.[3] For example, scientists at Google analyzed search queries submitted by millions of users worldwide to detect the outbreak and spread of influenza-like epidemics,[4] and pharmaceutical companies are routinely monitoring online social conversations for post-market drug research.[5-8] In order to properly extract relevant concepts from community-generated text corpora, a high-quality consumer health vocabulary (CHV) is often needed to specify how a particular health-related concept may be expressed differently in laypersons' terms vs. in a professional language. The accuracy and comprehensiveness of CHVs can be crucial to the performance of computational tools that make use of community-generated health text such as health-related tweets and patient posts in online health forums.

Many resources are available for describing and classifying medical concepts used in the professional settings, such as the Systemized Nomenclature of Medicine Clinical Trials® (SNOMED-CT®), the Logical Observation Identifiers Names and Codes (LOINC), and other biomedical vocabularies and ontologies included in the Unified Medical Library System (UMLS) Metathesaurus. However, vocabularies providing consumer-oriented health terms are relatively less mature. This fact diminishes the performance of named-entity recognition tools for processing community-generated text[9] as well as the potential for building applications that could "translate" professional language into layperson terms to improve readability and facilitate comprehension (e.g. to support the OpenNotes project that shares clinician notes with patients[10]).

Community-generated text corpora could serve as a valuable resource to extract laypersons' expressions of medical concepts (i.e., consumer terms) and their corresponding professional expressions. Wikipedia is one such rich resource that is frequently updated and popularly accessed by the general public. It is estimated that the medical entries on Wikipedia are accessed more than 180 million times a month, and about 1,000–2,000 edits are being made to them each day.[11] Further, it is also estimated that about half of the Wikipedia users who edit the medical entries are healthcare professionals; the remainder are patients, families, and the general public.[11] Thus, it is likely that the

Wikipedia medical entries contain both the professional terminology and laypersons' terms, linked by some semantic relationships (e.g., "influenza, commonly known as the flu"), which provides the basis for this research.

In this paper, we propose a pattern-based text-mining approach to identify pairs of professional terms and their consumer variants from Wikipedia, in addition to their alternative spellings and synonyms. We also describe a computational approach to validate the extractions and to label the terms in a pair as either "professional" or "consumer". This approach is based on the frequencies of a term appearing in MedLine,[12] which indexes scientific papers produced by the professional community, and in MedHelp,[13] a popular online health forum where the content is mainly generated by laypersons. A subsequent manual review of the extracted and labeled pairs of entities was conducted to validate the results generated by the computational approach. The results are very promising.

Background

The use of Wikipedia by laypersons and medical professionals has become a subject of active research in recent years. Several studies have shown that Wikipedia is one of the leading online destinations for health information seekers.[14-16] For example, recent surveys reported that 60% of European doctors use Wikipedia for professional purposes,[14] and nearly 50% of U.S. physicians who go online for information on specific medical conditions use Wikipedia.[15,16] These studies corroborate the findings published in peer-reviewed articles on the use of Wikipedia as a source of information for scientific and medical professionals,[17-19] as well as medical students.[20-25] Studies have also shown a growing use of online medical resources including Wikipedia by patients, caregivers, and healthcare consumers,[26-29] and that Wikipedia articles often appear in the top results provided by Web search engines.[30]

Identifying all medical entities from free text is an active area of research in natural language processing. Lexico-syntactic pattern-based approaches have been well established for over thirty years and have supported numerous information extraction tasks such as hyponym identification,[31] semantic classification,[32,33] meronym identification,[34] and large scale information extraction over the Web.[35] This paper builds on similar ideas to identify and leverage key textual patterns that are frequently used to present synonymous terms in Wikipedia. Automatic term recognition (ATR) techniques, sometimes called named entity recognition (NER) techniques, have also been proposed to identify valid candidate terms in biomedical text corpora,[36-40] using sequential models[41-43] and term scoring approaches.[44,45] In this study, we investigated if the rich formatting styles used by editors in Wikipedia and other wiki-based corpora give sufficient cues to extract consumer vocabulary terms with high accuracy.

Methods

In this section, we present the process used to identify and extract names of medical entities and their alternate synonym variants from Wikipedia. Figure 1 summarizes the system design, and the following paragraphs describe the process in detail.

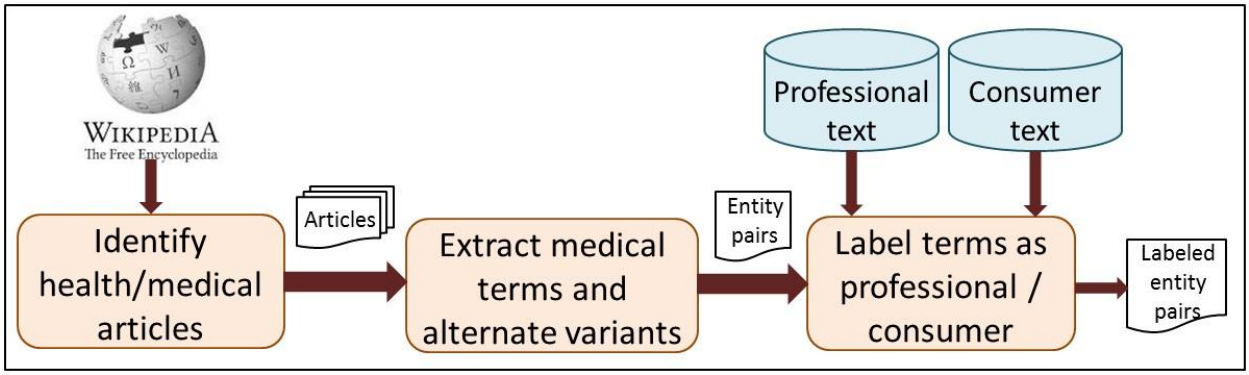

Figure 1. System work flow to extract and label professional and consumer health vocabulary.

Identifying relevant Wikipedia articles

Wikipedia releases periodic snapshots of all the articles published on the website, along with any associated metadata.[46] Each article contains the title, revision information, and the complete text. The text is typically unstructured and the formatting information, such as bold face or italicized fonts, citation information, and hyperlinks to other Wikipedia articles, is added with a marked up schema using special character sequences (such as two consecutive square brackets, three consecutive apostrophes, etc.). For this study, we considered the complete English language Wikipedia, which consists of over four million articles, and nine million additional pages for redirects, stubs, lists, and category pages.

Most Wikipedia articles also contain a list of tags that specify, for each article, a set of relevant categories from Wikipedia's hierarchical topic classification scheme.[47] At the top most level, this scheme consists of twenty five categories. Two of these are *Health* and *Medicine*. These top-level categories are further divided into additional sub-categories. For example, *Clinical medicine* is a sub-category of *Medicine*, and *Medical diagnosis* is a sub-category of *Clinical medicine*. The complete category hierarchy can be obtained by repeatedly traversing the sub-category links.

We collected a list of all sub-categories for *Health* and *Medicine* down to a depth of three, using an external Wikipedia tool called CatScan.[48] CatScan recursively searches an article category to find all articles, sub-categories, images, etc. This resulted in a list of 2,331 candidate categories. Once the relevant categories were identified, all four million Wikipedia articles were programmatically checked to retain only those articles that were tagged with at least one of the candidate categories. This narrowed the number of potential medically relevant articles to about 46,000, which constitutes about 1.2% of all articles in English Wikipedia.

Extracting medical terms and their common alternate names

Wikipedia articles are often written and formatted to serve as an introduction to the topic of the article.[49] Typically, the first sentence is formatted such that the article title appears in bold face as the subject of the definition or introduction. If the title also has alternate forms, such as abbreviations, alternate spellings, or significant alternate titles, these also appear as bold face immediately following or adjacent to the first occurrence of the title. For example, the Wikipedia article on "xerostomia"[50] starts as:

> **Xerostomia** (also termed **dry mouth** or **dry mouth syndrome**) is the medical term for the subjective symptom of dryness in the mouth, …

informing readers that "dry mouth" and "dry mouth syndrome" are alternative names for "xerostomia".

This relatively consistent style lends itself to devising automated text-mining and information extraction techniques. Through an iterative process of reviewing medical articles on Wikipedia, we identified a list of phrases that generally connected a medical concept term to its alternate terminology. Table 1 lists twelve examples of common linking phrases.

Table 1. Common linking phrases.

also called	commonly called	sometimes called	also termed
also known as	commonly known as	sometimes known as	previously known as
also referred to as	commonly referred to as	sometimes referred to as	colloquially known as

In addition to the linking phrases, parentheses also serve as a textual clue to introduce alternate spellings, abbreviations, and synonyms. Multiple alternate terms are mentioned in a comma-separated list, or separated by conjunctions such as "or" and "and". Each alternate form is typeset in bold face or italicized text. Additionally, hyperlinks are used to link phrases to articles about other related concepts.

Wikipedia articles were parsed to identify the title, and the leading text paragraph. All bold face, italicized, and hyperlinked phrases in the lead paragraph were identified as candidates. The common linking patterns were applied to extract pairs of linked candidate entities. For instance, applying this technique to the example above would generate two pairs: ("xerostomia", "dry mouth") and ("xerostomia", "dry mouth syndrome").

Labeling terms as consumer or professional

Although Wikipedia articles mention the alternate variants for a term, they do not specify which terms are likely to be used by consumers and which ones by professionals. Hence, simply extracting these relationships directly from Wikipedia does not provide enough detail to build a CHV, or to map between consumer and professional terms because there is no label assigned to each term.

It can be difficult to categorize terms as belonging to professional or consumer vocabulary. Frequently, health terms used by professionals migrate or evolve into popular vernacular as they become well known.[51] The acceptance or preference of a term can be measured, however, based on how often the term is used by the community. Specifically, if a term is more prominently used in professional text that is generated by and primarily intended for medical professionals, it can be regarded as professional. Conversely, if a term is frequently used by laypersons but not as often by medical professionals, then the term is more likely to be in a consumer-preferred vocabulary.

To quantitatively measure the propensity of a term T to be a consumer-oriented term, we define the following measure:

CHV_propensity (T) = count(T occurs in a consumer text corpus) / (size of the consumer text corpus)

Similarly, we can measure the propensity of a term T to be a professional term, as

PROF_propensity (T) = count(T occurs in a professional text corpus) / (size of the professional text corpus)

We propose that a term is more likely to be a consumer-oriented term if and only if its CHV_propensity is higher than its PROF_propensity. We refer to this intuition as the *propensity argument*.

To label a pair of terms (A, B) as professional or consumer with respect to one another, we can extend the propensity argument to the following. A is said to be the professional term and B is said to be the consumer term, if

$$\frac{\text{PROF\_propensity (A)}}{\text{CHV\_propensity (A)}} > \frac{\text{PROF\_propensity (B)}}{\text{CHV\_propensity (B)}}$$

or alternately,

$$\frac{\text{PROF\_propensity (A)}}{\text{PROF\_propensity (B)}} > \frac{\text{CHV\_propensity (A)}}{\text{CHV\_propensity (B)}}$$

which is the same as saying

$$\frac{\text{count(A occurrences in professional text)}}{\text{count(B occurrences in professional text)}} > \frac{\text{count(A occurrences in consumer text)}}{\text{count(B occurrences in consumer text)}} \quad \text{(Eq. 1)}$$

Conversely, if the condition is not met, then A is said to be the consumer term and B is said to be the professional term. Note that when comparing two terms using Eq. 1, the relative sizes of the corpora do not matter. However, the statistics collected over large corpora are more robust. Even when text corpus sizes are large, some concepts might occur infrequently or not appear at all. Such cases are avoided by smoothing the counts using Laplace smoothing.[52]

We chose online health discussion forums as a representative of consumer language. We crawled all the questions and comments posted by members on community discussion forums on MedHelp.[13] MedHelp is one of the earliest and well-known online forums dedicated to supporting user-driven discussions on health or healthcare related topics. The dataset consists of approximately thirty million messages posted by about a million unique users, and contains approximately 450 million words. This dataset has been a subject of study in other research endeavors.[53] In the following analysis, we refer to this as the **consumer text corpus**.

As a representative of professional text, we chose the abstracts of articles published in scientific journals and included in the 2012 MEDLINE®/PubMed® Baseline distribution.[12] To create a comparable corpus (in terms of word counts) to the consumer text corpus described earlier, we processed two million citations from the Baseline distribution that corresponded to papers published between 2008 and 2012. Titles were excluded from the generated professional text corpus. In the following analysis, we refer to this corpus as the **professional text corpus**.

To evaluate the accuracy of labeling professional and consumer terms in the extracted pairs, a medical expert conducted a manual review. First, the extracted pairs were filtered such that both terms appeared at least five times in both professional and consumer text corpora. A sample of 100 pairs was then randomly selected from this filtered set and manually judged and coded by a medical expert as one of the following classes: (a) valid pairing with correct labeling, (b) valid pairing with incorrect labeling, (c) pairs of equivalent concepts that either have alternative spellings or are synonymous, (d) pairs of related items, but not in a professional-consumer setting, such as an "is-a" relationship, or (e) invalid pairings. The pairs coded as equivalent or related (classes (c) or (d) above) were further coded to check if they were spelling variants, in case of equivalent pairs, or had a hierarchical ("is-a") relationship, if they were initially coded as related. The analysis of the expert judgment is presented in the Results section.

Results

Extracting pairs of medical concepts and their alternate names

Applying these techniques over the filtered set of medical articles from Wikipedia, we obtained 2,721 pairs of concepts and their consumer-preferred alternate names. Table 2 lists the linking patterns used, along with number of concept pairs each pattern generated. We also list a few examples of pairs extracted using that pattern.

Table 2. Patterns used to find pairs of alternate names, along with the count of pairs extracted and a few examples.

Linking pattern	Count	Examples
also known as	1695	(hematocrit, packed cell volume); (hair removal, epilation); (leukopenia, leukocytopenia); (dentures, false teeth)
also called	604	(heat therapy, thermotherapy); (hypersalivation, ptyalism); (nephroptosis, floating kidney); (dark therapy, scototherapy)
commonly known as	157	(nitrous oxide, laughing gas); (calcium oxide, quicklime); (pleurothotonus, Pisa syndrome); (*nepeta cataria*, catnip)
also termed / referred to as	106	(vertebral osteomyelitis, spondylodiskitis); (periapical cyst, radicular cyst); (red blood cells, erythrocyte); (posterior ramus syndrome, Maigne syndrome)
commonly called / referred to as	61	(peripheral vascular disease, peripheral artery disease); (actaea, baneberry); (schizophasia, word salad); (unnecessary health care, overtreatment)
sometimes called / termed	45	(high blood pressure, arterial hypertension); (chalicosis, flint disease); (hemiballismus, ballism); (irritable male syndrome, Del syndrome)
sometimes known as / referred to as	33	(pentazonia, giant pill millipedes); (hypochondria, health phobia); (ocular dominance, eyedness); (sexual addiction, sex addiction)
previously known as / called / referred to as	14	(acute kidney injury, acute renal failure); (erythrovirus, parvovirus B19); (periodic limb movement disorder, nocturnal myoclonus); (ankylosing spondylitis, Bechterew's disease)
colloquially known as / called / referred to as	6	(halitosis, bad breadth); (coal workers pneumoconiosis, black lung disease); (asystole, flatline); (central facial palsy, central seven)
Total pairs	**2721**	

We observe that a majority of extracted pairs come from the pattern "also known as". Further, 90% of the extracted pairs come from the top three patterns. We also observe that for some patterns such as "commonly known as" or "colloquially known as", the consumer-preferred terminology usually occurs as the second part of the extracted pair. However, these patterns contribute only about 8% of the extracted pairs. In pairs extracted using other patterns, the consumer term could appear in either positions.

Labeling the extracted pairs

To evaluate the accuracy of labeling the terms in each extracted pair as "professional" or "consumer" terms, a medical expert reviewed a sample of 100 pairs, as described in the Methods section. Each pair was manually coded by the expert as one of the following classes: (a) valid pairing with correct labeling, (b) valid pairing with incorrect labeling, (c) pairs of equivalent concepts that either have alternative spellings or are synonymous, (d) pairs of related items, but not in a professional-consumer setting, such as an "is-a" relationship, or (e) invalid pairings. Table 3 summarizes the results.

Table 3. Classification of 100 pairs of extractions, with examples. In the examples provided, the professional term as defined by the expert reviewer is shown first and the consumer term is shown second.

Code	Class of instance	Counts	Example pairs
(a.0)	Valid pairing with correct labeling	54	(icterus, jaundice); (pyrosis, heartburn); (oral candidiasis, oral thrush); (somnambulism, sleepwalking); (tinea, ringworm)
(b.0)	Valid pairing with incorrect labeling	4	(chronic renal disease, chronic kidney disease); (ovum, eggs); (brucellosis, Mediterranean fever); (hyperplasia, proliferation)
(c.1)	Alternative spelling variants	8	(leukoplakia, leucoplakia); (fecal incontinence, faecal incontinence); (post-concussion syndrome, postconcussive syndrome)
(c.2)	Synonyms or equivalent	23	(orthostatic hypotension, postural hypotension); (viral load, viral titer); (Lugols iodine, Lugols solution); (mouthwash, mouth rinse); (fecal incontinence, anal incontinence);
(d.1)	Is-a mapping	2	(cannabinoid, endocannabinoid); (radiology, radiation oncology)
(d.2)	Related concepts	9	(hypersensitivity, intolerance); (medical test, diagnostic technique); (medical procedure, technique); (pain management, pain medicine); (keratosis, keratotic); (suffering, aversive); (nerve, innervation); (tracheotomy, tracheostomy); (coccidioidomycosis, cocci)

From Table 3, we first observe that 89% of the pairs are between synonymous or equivalent concepts (Table 3, rows (a.0), (b.0), (c.1), and (c.2)), while the remaining instances were mainly between related items that are not an exact synonym of one another. In the current study, since the focus is on identifying medical terms and their equivalent consumer terms, related concepts (Table 3, rows (d.1) and (d.2)) are less desirable than the valid mappings. None of the 100 extracted pairs were judged invalid or to be between unrelated medical concepts.

Among the 89 pairs identified as valid pairs, 58 pairs (65.2%) were judged to be valid mappings between a professional term and a consumer term. The remaining pairs were either synonyms or equivalent concepts, both of which are valid professional terms. It is instructive to note that even when two terms are profession terms, one is often more widely accepted and used in consumer-generated corpora. For example, the terms "orthostatic hypertension" and "postural hypotension" appeared with similar frequencies in the professional text corpus, but the latter variant was observed 7.5 times more frequently than the former in the consumer text corpus. Hence, our approach labeled "postural hypotension" as the consumer term in the above example pair.

The approach to label terms as consumer or professional was also fairly accurate. For four pairs, the expert labels mismatched those assigned automatically (using the propensity argument, Eq. 1). Comparing against the 54 instances that were judged and labeled the same, this leads to a labeling accuracy of 93.1%. All four instances have been listed as examples in the table (Table 3, row (b.0)).

Identifying new mappings between professional and consumer terms

Finally, we also compared the pairs obtained by the proposed method against the CHV files made available by the open access and collaborative (OAC) CHV initiative.[54] Our approach generated many new pairs that were not included in the existing databases of consumer health vocabulary. Table 4 lists a few such examples of pairs extracted from Wikipedia articles.

Table 4. Examples of new pairs of professional and consumer terms that are not included in available CHV datasets.

Professional term	Equivalent consumer term
ambulatory surgery	outpatient surgery
arterial hypertension	high blood pressure
asphyxiation	lack of oxygen
Biofilms	Plaque
conjunctivitis	pink eye
dermatophytosis	Ringworm
Ethanol	drinking alcohol*
Ethanol	pure alcohol
Fatigue	Lethargy
fertilization	Conception

Professional term	Equivalent consumer term
immunoglobulins	Antibodies
nasopharyngitis	common cold
neuroleptics	Antipsychotics
nutrition	Nourishment
orthostatic hypotension	postural hypotension
periodontitis	gum disease
rofecoxib	Vioxx
social care	home care
stuttering	Stammering
uncompensated care	charity care

* The CHV mentions the pairing between ethanol and drinking alcohol, but lists it as an incorrect mapping instance.

Discussion

One of the benefits of using Wikipedia as a source to identify alternate names is that popular variant names are often nominated by the community and thus mentioned in the Wikipedia article about the primary concept. As language itself evolves, new variants might be introduced,[51] and existing variants might change in popularity. Regularly updating the CHV using the latest snapshot of Wikipedia articles provides one way to keep them current and relevant. Since errors in Wikipedia articles often get corrected by other editors, periodic update of extracted pairs can therefore help eliminate erroneous instances of extracted pairs. For instance, we observed that the latest version of Wikipedia (updated after our analyses were conducted) correctly removed the mention of "radiation oncology" as a variant form for "radiology" (see Table 3, row (d.1)). In the same vein, the professional and consumer text corpora could also be frequently updated to monitor the usage of such terms in the respective communities.

The approach of using free text to compute statistics is subject to the entity disambiguation problems. For example, the term "cocci" has multiple meanings – it could be used either as a shortened term for the disease "coccidioidomycosis", or as the plural form of "coccus". Failure to disambiguate between these two concepts could lead to mislabeling. Although this is a limitation of most statistical and shallow, frequency-based approaches, such occurrences are relatively infrequent.

We have not measured the recall of such pattern-based approaches, to extract *all* instances of consumer terms or alternate variants. Further study is also needed to understand if the words used in the patterns could improve the labeling of terms as professional or consumer. For example, if the pattern is "commonly known as" or "colloquially called", it is more likely that the succeeding concept is a consumer term.

Concerns have been raised about use of Wikipedia as a resource of information in the scientific literature. Bibliometric analysis has shown an increased rate of citing Wikipedia articles in peer-reviewed health science journal publications in recent years,[55,56] and studies have found several limitations with respect to depth of discussion and readability in Wikipedia articles.[57,58] Efforts are underway to encourage medical professionals to actively contribute to Wikipedia,[59] and, in collaboration with other medical and healthcare experts and medical journals, to improve the overall quality of medical articles in Wikipedia.[60,61] Although such concerns are important issues to address in the future, they are beyond the current scope of research presented in this paper.

Conclusion

In this study, we demonstrated the effectiveness of a novel approach that uses community-generated corpora such as Wikipedia to mine pairs of professional terms and their equivalent consumer terms. We measured the propensity of

a term to be a consumer term based on its relative frequencies appearing in the consumer or professional contexts, and demonstrated how this information could be used to properly label the terms. The empirical evaluation results are promising, suggesting that the proposed approach is able to identify and differentiate consumer and professional terms from the Wikipedia corpus with high accuracy. The methods proposed in this paper can therefore be used to augment, update, and maintain existing consumer health vocabularies to enhance the performance of computational applications designed to improve the readability, parsing, and understandability of community-generated health text.

Acknowledgements

The authors would like to acknowledge the contribution of Nirav Mehta, who helped collect data for this study. This study was supported in part by the University of Michigan MCubed Program, the National Center for Advancing Translational Sciences under Award Number UL1TR000433, the National Science Foundation under grant numbers IIS-1054199 and CCF-1048168, and the DARPA under award number W911NF-12-1-0037. The content is solely the responsibility of the authors and does not necessarily represent the official views of funding agencies.

References

1. Fox S, Duggan M. Health online 2013. Pew Research Center's Internet & American Life Project. Published 2013 Jan 15.
2. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc 2006;13:24–9.
3. Okun S, McGraw D, Stang P, et al. Making the case for continuous learning from routinely collected data. Institute of Medicine Discussion Paper, National Academy of Sciences. 2013.
4. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012–4.
5. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. J Med Internet Res 2011;13(1):e6.
6. Pearson JF, Brownstein CA, Brownstein JS. Potential for electronic health records and online social networking to redefine medical research. Clin Chem 2011;57(2):196–204.
7. Bian J, Topaloglu U, Yu F. Towards large-scale Twitter mining for drug-related adverse events. Proc Workshop on Smart Health and Wellbeing 2012;25–32.
8. Riding the information technology wave in life sciences: priorities, pitfalls and promise. Retrieved 2014 Mar 12 http://www.imshealth.com/portal/site/imshealth/menuitem.762a961826aad98f53c753c71ad8c22a/?vgnextoid=743a7a4c18394410VgnVCM10000076192ca2RCRD.
9. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc 2013;0:1–8.
10. Delbanco T, Walker J, Darer JD, et al. Open notes: doctors and patients signing on. Ann Intern Med 2010;15(2):121–5.
11. Wi /. kipedia is a massively popular (yet untested) doctor. Retrieved 2014 Feb 22 from http://m.nextgov.com/health/2014/02/wikipedia-massively-popular-yet-untested-doctor/79154/.
12. 2012 MEDLINE®/PubMed® Baseline Database Distribution. Retrieved 2013 Nov 23 from http://www.nlm.nih.gov/bsd/licensee/2012_stats/baseline_med_filecount.html.
13. MedHelp. Retrieved 2013 Mar 20 from http://www.medhelp.org/.
14. Eade D. Dr Wikipedia will see you now… Insight Research Group. 2011 Jun 7. Retrieved 2014 Mar 8 from http://www.pmlive.com/pharma_news/dr_wikipedia_will_see_you_now..._280528.
15. Rosen D. Engaging patients through social media. Report by the IMS Institute for Healthcare Informatics. Published 2014 Jan.
16. Comer B. Docs look to Wikipedia for condition info: Manhattan research. Medical Marketing & Media. 2009 Apr 21. Retrieved 2014 Mar 11 from http://www.mmm-online.com/docs-look-to-wikipedia-for-condition-info-manhattan-research/article/131038/.
17. Hughes B, Joshi I, Lemonde H, Wareham J. Junior physician's use of Web 2.0 for information seeking and medical education: a qualitative study. Int J Med Inform 2009;78:645–55.
18. Brokowski I, Sheehan AH. Evaluation of pharmacist use and perception of Wikipedia as a drug information resource. Ann Pharmacother 2009;43:1912–3.

19. Masters K. For what purpose and reasons do doctors use the Internet: a systematic review. Int J Med Inform 2008;77:4–16.
20. Burgos C, Bot A, Ring D. Evaluating the effectiveness of a wiki Internet site for medical topics. J Hand Microsurg 2012;4:21–4.
21. Varga-Atkins T, Dangerfield P, Brigden D. Developing professionalism through the use of wikis: a study with first-year undergraduate medical students. Med Teach 2010;32:824–9.
22. Haigh CA. Wikipedia as an evidence source for nursing and healthcare students. Nurse Educ Today 2011;31:135–9.
23. Jalali A, Mioduszewski M, Gauthier M, Varpio L. Wiki use and challenges in undergraduate medical education. Med Educ 2009;43:1117.
24. Snodgrass S. Wiki activities in blended learning for health professional students: enhancing critical thinking and clinical reasoning skills. Aus J Educ Technol 2011;27:563–80.
25. Weiner SA, Stephens G, Nour AY. Information-seeking behaviors of first-semester veterinary students: a preliminary report. J Vet Med Educ 2011;38:21–32.
26. Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the World Wide Web: a systematic review. JAMA 2002;287:2691–700.
27. Deshpande A, Jadad AR. Web 2.0: Could it help move the health system into the 21st century? J Men Health Gender 2006;3:332–6.
28. Thomas GR, Eng L, de Wolff JF, Grover SC. An evaluation of Wikipedia as a resource for patient education in nephrology. Semin Dial 2013;26:159–63.
29. Kinnane NA, Milne DJ. The role of the Internet in supporting and informing carers of people with cancer: a literature review. Support Care Cancer 2010;18:1123–36.
30. Laurent MR, Vickers TJ. Seeking health information online: does Wikipedia matter? J Am Med Inform Assoc. 2009;16:471–9.
31. Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proc Conf Comput Linguist (COLING) Assoc Comput Linguist 1992;539–45.
32. Snow R, Jurafsky D, Ng AY. Learning syntactic patterns for automatic hypernym discovery. Adv Neural Inf Process Syst 2004;17:1297–304.
33. Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the web: an experimental study. Artif Intell 2005;165(1):91–134.
34. Girju R, Badulescu A, Moldovan D. Automatic discovery of part-whole relations. Comput Linguist Assoc Comput Linguist 2006;32(1):83–135.
35. Yates A, Etzioni O. Unsupervised Methods for Determining Object and Relation Synonyms on the Web. J Artif Intell Res 2009;34:255–96.
36. Aronson AR, Lang FM. An overview of MetaMap: historic perspectives and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.
37. Torii M, Hu Z, Wu CH, Liu H. BioTagger-GM: a gene/protein name recognition system. J Am Med Inform Assoc 2009;16(2):247–55.
38. Harkema H, Gaizauskas R, Hepple M, et al. A large scale terminology resource for biomedical text processing. Proc Workshop Linking Biological Literature Ontologies and Databases Assoc Comput Linguist 2004:53–60.
39. Kageura K, Umino B. Methods of automatic term recognition: a review. Terminology 1996;3(2):259–89.
40. Krauthammer M, Nenadic G. Term identification in the biomedical literatue. J Biomed Inform 2004;37(6):512–26.
41. Collier N, Nobata C, Tsujii J. Extraction of the names of genes and gene products with a hidden Markov model. Proc Conf Comput Linguist (COLING) Assoc Comput Linguist 2000:201–7.
42. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine learned solutions for three stages of clinical information extraction: the state-of-the-art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557–62.
43. Jindal P., Roth D. Using soft constraints in joint inference for clinical concept recognition. Proc Conf Empirical Methods in Natural Language Processing (EMNLP) Assoc Comput Linguist 2013:1808–14.
44. Frantzi KT, Ananiadou S, Mima H. Automatic term recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries 2003;3(2):115–30.
45. Zeng QT, Tse T, Divita G, et al. Term identification methods for consumer health vocabulary development. J Med Internet Res 2007;9(1):e4.
46. Wikipedia dump of English language articles. Retrieved 2013 Sep 15 from http://dumps.wikimedia.org/enwiki/.
47. Wikipedia: main topic classifications. Retrieved 2013 Sep 15 from http://en.wikipedia.org/wiki/Category:Main_topic_classifications.

48. CatScan. Retrieved 2013 Oct 12 from http://tools.wmflabs.org/catscan2/catscan2.php.
49. Wikipedia: manual of style / lead section. Retrieved 2014 Mar 9 from http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section.
50. Wikipedia article on Xerostomia. Retrieved 2014 Mar 11 from http://en.wikipedia.org/wiki/Xerostomia.
51. Doing-Harris KM, Zeng-Treitler, Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J Med Internet Res 2011;13(2):e3.
52. Manning CD, Raghavan P, Schütze M. Introduction to Information Retrieval. Cambridge University Press 2008:240.
53. Vydiswaran VGV, Liu Y, Mei Q, Zheng K, Hanauer D. User-created groups in health forums: What makes them special? Proc Conf Weblogs and Social Media (ICWSM) Assoc Adv Artif Intell 2014:515–24.
54. Open Access and Collabolative Consumer Health Vocabulary Initiative. Retrieved 2013 Nov 10 from http://consumerhealthvocab.org/.
55. Bould MD, Hladkowicz ES, Pigford AA, et al. References that anyone can edit: review of Wikipedia citations in peer reviewed health science literature. BMJ 2014;348:g1585.
56. Noruzi A. Editorial: Wikipedia popularity from a citation analysis point of view. Webology 2009 Jun;6(2):e20.
57. Giles J. Internet encyclopaedias go head to head. Nature 2005;438:900–1.
58. Azer SA. Evaluation of gastroenterology and hepatology articles on Wikipedia: are they suitable as learning resources for medical students? Eur J Gastroenterol Hepatol 2014 Feb;26(2):155–63.
59. Metcalfe D, Powell J. Should doctors spurn Wikipedia? J R Soc Med 2011;104:488–9.
60. Heilman JM, Kemmann E, Bonert M, et al. Wikipedia: a key tool for global public health promotion. J Med Internet Res 2011;13:e14.
61. Mathew M, Joseph A, Heilman J, Tharyan P. Cochrane and Wikipedia: the collaborative potential for a quantum leap in the dissemination and uptake of trusted evidence. Cochrane Database Syst Rev 2013 Oct 22;10:ED000069.