An evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database

Jeff Friedlin, Clement J McDonald

Regenstrief Institute, Indianapolis, Indiana, USA

Correspondence to
Dr Jeff Friedlin, Regenstrief Institute, Suite 2000, 410 West 10th Street, Indianapolis, IN 46202, USA; jfriedlin@regenstrief.org

Received 29 September 2009
Accepted 3 March 2010

ABSTRACT

The logical observation identifiers names and codes (LOINC) database contains 55 000 terms consisting of more atomic components called parts. LOINC carries more than 18 000 distinct parts. It is necessary to have definitions/descriptions for each of these parts to assist users in mapping local laboratory codes to LOINC. It is believed that much of this information can be obtained from the internet; the first effort was with Wikipedia. This project focused on 1705 laboratory analytes (the first part in the LOINC laboratory name). Of the 1705 parts queried, 1314 matching articles were found in Wikipedia. Of these, 1299 (98.9%) were perfect matches that exactly described the LOINC part, 15 (1.14%) were partial matches (the description in Wikipedia was related to the LOINC part, but did not describe it fully), and 102 (7.76%) were mis-matches. The current release of RELMA and LOINC include Wikipedia descriptions of LOINC parts obtained as a direct result of this project.

Logical observation identifiers names and codes (LOINC) is a no-cost database of laboratory and clinical observations intended as a universal standard for identifying observations transmitted in HL7 and other standard messages.[1] The last release version 2.27 contains over 55 000 observation terms.[2] Each LOINC term consists of five to 10 'parts', which are the atomic components from which a full LOINC term is constructed. The database contains a total of 18 738 distinct LOINC parts. To make the database more useful and assist users in mapping local laboratory codes to LOINC codes, the Regenstrief Institute is seeking to create short (1–10 line) text definitions/descriptions for the part concepts in LOINC. The web is a potential source for the definitions we seek. The main question is, can these definitions be accurately extracted from such sources automatically by text matching techniques and if so how much can be found. Wikipedia is a web-based free-content multilingual encyclopedia project, and exists as a wiki (a website that allows any visitor to freely edit its content).[3] Started in January 2001, Wikipedia currently contains 2 416 460 English language entries (called 'articles') as of June 2008, and could provide many of the LOINC definitions we seek. The Wikipedia website presents its articles under the GNU free documentation license,[4] which permits the redistribution, creation of derivative works, and commercial use of content, provided that its authors are attributed and this content remains available under the GNU free documentation license. We developed a computer software tool to automate queries of the Wikipedia website with LOINC part names and quantified the coverage of Wikipedia for LOINC part content. Here we describe the software and our experience.

METHODS

We developed the Wikipedia to LOINC matcher (WLMA) software tool (written in JAVA) at the Regenstrief Research Institute in Indianapolis, Indiana.[5] WLMA queries Wikipedia website with a LOINC part, and attempts to find a matching Wikipedia article. WLMA automatically submits a list of words or phrases to Wikipedia and processes the results of the query using natural language processing (NLP) algorithmic rules to further narrow the search space and reduce the text available in Wikipedia via a series of steps shown in figure 1. We will now describe these steps in more detail.

Wikipedia contains several built-in search methods which help refine user queries and categorize search results, and WLMA leverages three of these in querying Wikipedia with LOINC parts. The first of these is disambiguation. Disambiguation in Wikipedia is the process of resolving conflicts in queries that occur when a single term can be associated with more than one topic. When such a query is made, Wikipedia displays a 'disambiguation page'—a non-article page that contains only links to other Wikipedia pages. In other words, disambiguations are paths leading to different articles which could, in principle, have the same title. For example, querying Wikipedia with the LOINC part 'Mercury' results in a disambiguation page because it can refer to several different things, including: an element, a planet, an automobile brand, a record label, a NASA manned-spaceflight project, a plant, and a Roman god. As only one Wikipedia page can have the generic name 'Mercury', unambiguous article titles must be used for each of these topics: mercury (element), Mercury (planet), Mercury (automobile), Mercury Records, Project Mercury, mercury (plant), Mercury (mythology). Therefore, searching Wikipedia for 'Mercury' causes the display of a disambiguation page with unambiguous article titles 'Mercury' might refer to. Wikipedia displays the most commonly referred to articles in a specific section at the top of the disambiguation page. The first article displayed for 'Mercury' is mercury (the element) which is the matching Wikipedia article for this LOINC part. When a LOINC part query results in a disambiguation page, WLMA extracts the top three article titles on this page. It then re-queries Wikipedia with each of these three titles until a match is found. We analyzed a sample of LOINC part queries that resulted in a disambiguation page

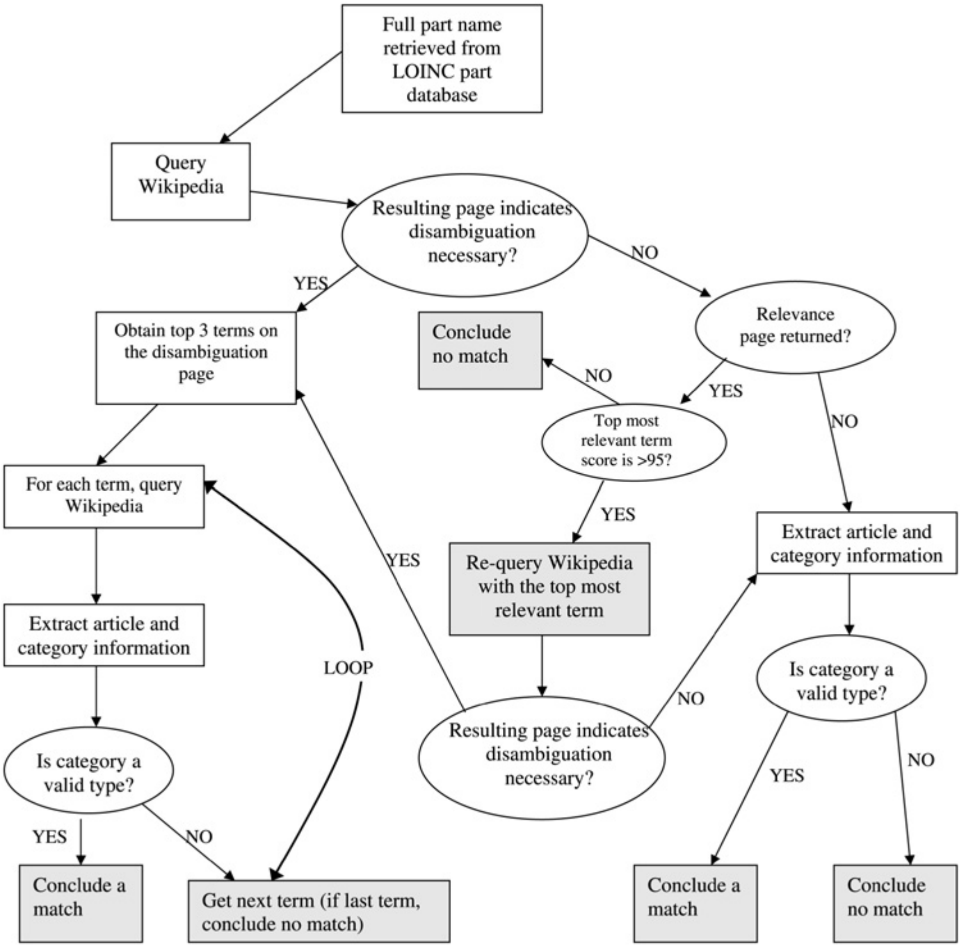
Full part name retrieved from LOINC part database
Query Wikipedia
Resulting page indicates disambiguation necessary?
YES
NO
Obtain top 3 terms on the disambiguation page
Conclude no match
Relevance page returned?
NO
YES
NO
Top most relevant term score is >95?
YES
For each term, query Wikipedia
YES
Re-query Wikipedia with the top most relevant term
Extract article and category information
Extract article and category information
LOOP
Resulting page indicates disambiguation necessary?
NO
Is category a valid type?
Is category a valid type?
YES
NO
YES
NO
Conclude a match
Get next term (if last term, conclude no match)
Conclude a match
Conclude no match

Figure 1 Matching algorithm used by the software to match logical observation identifiers names and codes (LOINC) part names to Wikipedia articles.

and we found that if Wikipedia contained a matching article, it was one of the top three articles approximately 98% of the time. When we changed this threshold to include the top six articles, we found we could capture the other 2% of the matching articles, but this caused an unacceptable decrease in specificity.

Another methodology in Wikipedia that helps refine query results is the concept of 'relevancy rank'. When Wikipedia finds no exact match for the search string, it re-directs the query to a special search page, which displays the names of existing articles ranked by a relevance score (range 1–100) based on the frequency of the queried term in the article itself. Through empirical analysis of 50 LOINC part Wikipedia queries that resulted in relevancy rank pages, we found that articles with relevance scores of 95 or greater had a high likelihood of being the matching Wikipedia article, and articles with scores less than 95 were much less likely to be a correct match. When a LOINC part query results in a relevancy rank page, WLMA pulls the article name with the highest relevance score (if the score is greater than or equal to 95) and re-queries Wikipedia. For example, querying Wikipedia with the LOINC part 'sesame seed (sesamum indicum)' results in the display of a relevancy rank page with the article entitled 'sesame' as the top ranking page with a score of 98. This article is the matching Wikipedia article for this LOINC part. If the top ranking relevance score is below 95, WLMA concludes there is no likely match in Wikipedia for this LOINC part.

Wikipedia helps refine query results with article 'categories'. Article categories are major topics that are likely to be useful to someone reading the article and are roughly a hierarchical representation of the query term. Categories allow articles to be placed in one or more groups, and allow those groups to be further categorized. They do not form a strict hierarchy since each article can appear in more than one category, and each category can appear in more than one parent category. This allows multiple categorization schemes to co-exist simultaneously. For example, the term 'penicillin' is categorized as 'β-lactam antibiotic' and the term 'colonoscopy' is categorized as 'diagnostic gastroenterology|medical tests'. WLMA uses this categorical information to verify that the matching article truly describes the LOINC part in question. To obtain categories of LOINC parts, we queried the Unified Medical Language System with each LOINC part to obtain their semantic types, in addition to manually reviewing the LOINC part database. We queried Wikipedia with a sample of LOINC parts from each category, and found 95 Wikipedia categories that represented LOINC part categories—examples include medication, virus, chemical, test, etc. We then programmed these Wikipedia categories into WLMA. If a matching article is not a member of one of these 95 categories, WLMA concludes that the match is incorrect. This helps prevent false-positive matches.

As shown in figure 1, when WLMA queries Wikipedia with a LOINC part, if neither a disambiguation page nor relevance rank page is encountered, and the article category is valid, a match is concluded. WLMA then extracts the entire introduction section of the article and trims this section to the first 1000 characters (expanded to the end of the current sentence). Our goal was to create definitions/descriptions of each LOINC part that provided enough information to be valuable, but were also concise enough to allow quick review by users and LOINC mappers. We discovered, through empirical analysis of matching Wikipedia articles, that descriptions of LOINC parts that satisfied these needs were generally found within the first 1000 characters of an article. (Users who wish to obtain the full Wikipedia article can follow the direct link to the article provided in the LOINC database.) Wikipedia uses its own special mark-up language and carries numerous strings that can not be represented in ordinary text displays without the Wikipedia text processor engine. WLMA removes all of this mark-up as well as HTML mark-up pronunciation graphics, links to other websites, references to tables or graphics, etc. This edited portion of the article is then placed into the LOINC 'description' field of the matching LOINC part and the 'description source' field is updated with 'Wikipedia' along with the URL to the full article.

We narrowed the list of LOINC parts used to query Wikipedia from 18 738 to 1705 as described below. The LOINC part database includes parts that represent a base concept (often a disease or organism name) as well as more specific named entities derived from that base term. For example, it includes a part for Ebola virus (the base term) as well as many derivatives of that base term such as:
EBOLA VIRUS AB
EBOLA VIRUS AG
EBOLA VIRUS RNA

We identified derivatives by querying the LOINC part database for multiword part names that ended with terms known to indicate derivatives such as 'AB', 'AG', 'RNA' and 'IgM'. The list of derivative terms was obtained through inspection of the LOINC part database. We manually reviewed the results of the query to verify that all LOINC parts identified were truly derivatives and should be excluded. Because we did not expect that Wikipedia contained entries for the derivatives specifically, we excluded them and queried Wikipedia with only the base term (in this case Ebola virus). Of the total 18 738 LOINC parts, 14 787 contained such derivatives and were excluded. Furthermore, we excluded from our study some categories of LOINC parts unlikely to have corresponding Wikipedia articles such as test panel names, general terms (REFERENCE LAB NAME) and analyte test names (CELLS. CD3+CD4+CD45R+CD45). We also excluded LOINC parts with non-specific, cryptic names such as A1, B, AB, P2, etc. We identified these by querying the LOINC part database for part names that either contained special characters (such as '+') or were two characters or less in length. An additional 2246 LOINC parts were excluded based on these two criteria, resulting in a total of 1705 for inclusion in our study.

We tested WLMA's automatic matching by manually reviewing all description field entries in which a match occurred. We categorized each match into three groups: 1. a 'perfect' match occurred when the concept described in the Wikipedia article was identical to the LOINC part concept. 2. A partial match occurred when the Wikipedia article concept was related but not identical to the LOINC part. For example, the LOINC part 'dengue virus' matched to the Wikipedia article 'dengue fever' and the LOINC part Salmonella abortus matched to the Wikipedia article 'Salmonella'. 3. A mismatch occurred when the LOINC part and the matching Wikipedia article concept were unrelated (ie, the LOINC term 'turkey' (a part used in turkey (the bird) IGE antibody tests) matched to the Wikipedia article 'Turkey' (the country)). Some LOINC terms do contain country names (examples include Japanese encephalitis virus Ab and Venezuelan equine encephalitis virus Ab), so country is a valid LOINC part category.

OBSERVATIONS

WLMA queried the online encyclopedia Wikipedia website with 1705 parts in October 2007 using a broadband ethernet internet connection. The software completed all queries in approximately 45 min. Of the 1705 queries, 1416 returned a matching article, and 289 returned no match. Using criteria discussed previously, of the 1416 matches found by WLMA, 1299 (92%) were complete matches, 15 (1%) were partial matches, and 102 (7%) were non-matches. We manually queried Wikipedia with the 289 unmatched LOINC parts and found 35 parts (12%) matched either partly or completely to Wikipedia articles. Most of these false-negative errors (72%) were due to the stringent requirement for the relevance score (95) used by WLMA. The true matches had lower relevance scores than WLMA allowed. The remaining 18% of false-negative errors occurred when the Wikipedia article's category information was not contained in the database of 95 valid categories. The sensitivity of WLMA was 97.5%, its specificity 77.7%, and its positive predictive value 93.0%.

We found the likelihood of a false-positive match was much higher when WLMA matched an ambiguous Wikipedia article. We analyzed the 1416 part names in which WLMA declared a match (figure 2). Of the 1416 matches, 1243 were unambiguous according to Wikipedia (described earlier). Of these 1243 unambiguous matches, 1223 were true positive matches, and 20 were incorrect (false positives). Of the 173 matches needing disambiguation according to Wikipedia, only 91 were true matches and 82 were incorrect.

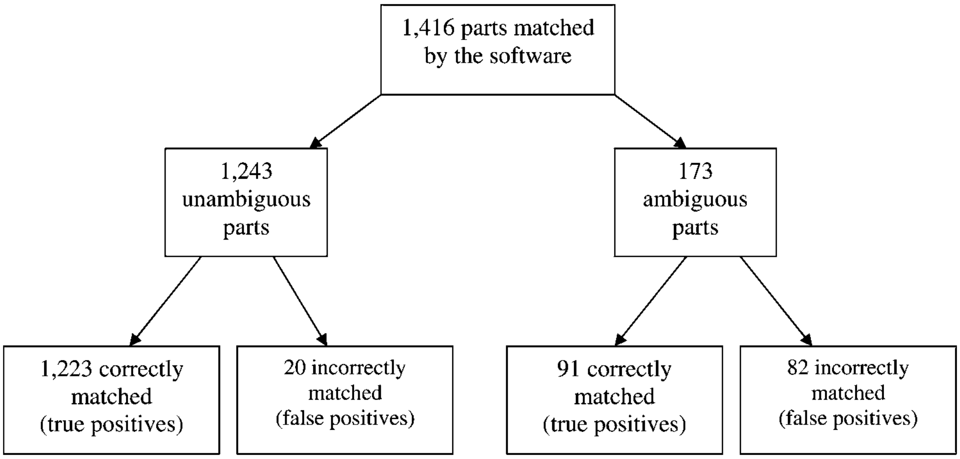
1,416 parts matched by the software
1,243 unambiguous parts
173 ambiguous parts
1,223 correctly matched (true positives)
20 incorrectly matched (false positives)
91 correctly matched (true positives)
82 incorrectly matched (false positives)

Figure 2 Numbers of true and false positives for all parts matched by the software.

Table 1 displays the percentage of Wikipedia matches for a sample of the 15 most clinically relevant LOINC part categories as determined by an experienced physician (JF). Five categories of LOINC parts had over 20% matching Wikipedia articles: gastrointestinal system, amino acids, body fluids, body parts, and medications. An experienced physician (JF) performed a review of a random sample of 100 complete matches for the purpose of assessing the accuracy of the information contained in the articles. The reviewer concluded that information in all 100 articles was accurate and that the information in the articles provided adequate definitions/descriptions of the LOINC parts.

DISCUSSION

The purpose of this study was twofold. First, we wished to evaluate the degree of medical knowledge contained in the online encyclopedia Wikipedia and the feasibility of using that knowledge as a means of adding description information to a laboratory and clinical observations database (LOINC). Second, we desired to test our software's ability to automatically extract relevant information from Wikipedia based on queries generated from part names taken from the LOINC part database.

We are surprised by the extensive amount of medical knowledge contained in the online encyclopedia Wikipedia. It contains large numbers of articles relating to science in general and medical topics in particular, and previous studies[6] show that these articles contain information comparable in accuracy to commercial online encyclopedias such as Encyclopedia Britannica.[7] This project demonstrates a unique medical informatics use of freely available online information. Other opportunities likely exist for similar uses using data from other online sources.

Due to space constraints in the LOINC database, we extracted only the introductory paragraphs of each matching Wikipedia article. Full Wikipedia articles contain much more information than we extracted. However, our analysis revealed that the data we extracted, although not complete, were accurate. To allow the LOINC user to easily obtain more information about a part, as well as to satisfy the copyright restrictions of Wikipedia, we include a direct link back to the specific article within the LOINC database.

Several new laboratory sources join our health information exchange on a yearly basis. Before a laboratory can become part of the health information exchange, all local laboratory codes must first be mapped to LOINC codes. The Regenstrief Institute employs four full-time employees to perform this mapping, but it is a time-consuming process. These descriptions have the potential to facilitate the mapping process. In the future, we plan to investigate the effect these descriptions have on the mapping process by comparing the mapping efficiency of mappers having access to the descriptions with those who do not. We also plan on surveying the mappers to evaluate the perceived usefulness of the descriptions and to elicit feedback on how they may be improved.

Table 1 Percentage of Wikipedia matches for various LOINC part categories

LOINC part category	No of LOINC parts	No of Wikipedia matches
Gastrointestinal system	13	11 (85%)
Amino acids	39	32 (82%)
Body fluids	13	10 (77%)
Body parts	92	70 (76%)
Medications	263	197 (75%)
Viruses	105	77 (73%)
Hormones	66	47 (71%)
Vascular system	118	84 (71%)
Musculoskeletal system	66	45 (68%)
Heavy metals	53	36 (68%)
Medical categories	105	70 (66%)
Bacteria	105	69 (66%)
Blood cells	53	34 (64%)
Animal	78	50 (64%)
Plant	131	84 (64%)

LOINC, logical observation identifiers names and codes.

We are pleased by the overall specificity, sensitivity, and precision of our software's matching algorithm. As one of the goals of this study was to determine the degree of medical knowledge contained in Wikipedia, we set the criteria for a true match at a relatively low level, to attempt to find all possible Wikipedia matches and minimize false negatives. WLMA's specificity could be improved by setting the criteria for a true match to a stricter level thereby decreasing the number of false-positive matches. However, this would likely result in a decrease of the software's sensitivity.

Several changes have been made to Wikipedia since we performed this study. First, the relevance rank page is displayed in a different format. When Wikipedia finds no exact match for a query, a page with results from that query of a search engine external to Wikipedia is displayed. The user can select which external search engine Wikipedia uses and includes Google, Yahoo, Wikiwix, and Microsoft Live. Wikiwix displays results with a relevancy score very similar to that described above and found in previous versions of Wikipedia.

Wikipedia has also recently made other changes which could be of interest to the medical informatics community. Wikipedia has long included classification information for certain articles, such as the chemical abstracts service registry number for chemicals and the scientific classification for animals. Recently, it has also begun to include classification information for medically related articles, such as the International Classification of Diseases (ICD) version 10, ICD-9[8] and medical subject headings (MESH) codes.[9] For example, a Wikipedia query for pneumonia results in an article containing not only text describing the condition, but also the corresponding ICD-10, ICD-9 and MESH codes. Additional codes from other coding systems such as the Online Mendelian Inheritance in Man,[10] the Diseases Database,[11] MedlinePlus,[12] and eMedicine[13] (from WebMD[14]) are included in some medically related Wikipedia articles. Other standardized coding systems, such as LOINC codes and Systematized Nomenclature of Medicine—clinical terms (SNOMED-CT)[15] do not appear in Wikipedia articles at this time.

We conclude that Wikipedia contains a surprisingly large amount of scientific and medical data and could effectively be used as an initial knowledge base for specific medical informatics and research projects. The software we developed to automate the matching of LOINC part names to Wikipedia articles performed satisfactorily with high sensitivity and moderate specificity. The current release of RELMA and LOINC include descriptions of LOINC parts obtained from Wikipedia as a direct result of this project.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **McDonald CJ,** Huff SM, Suico JG, *et al*. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem* 2003; **49**:624—33.
2. **LOINC.** Logical Observation Identifiers Names and Codes (LOINC), 2009. http://www.loinc.org/ (accessed Aug 2009).

3. **Wikipedia.** 2009. http://www.en.wikipedia.org/wiki/Main_Page (accessed Aug 2009).
4. **GNU.** GNU Operating System, 2008. http://www.gnu.org/ (accessed Aug 2009).
5. **Regenstrief Institute Inc.** 2009. http://www.regenstrief.org/ (accessed Aug 2009).
6. **Giles J.** Internet encyclopaedias go head to head. *Nature* 2005;**438**:900—1.
7. **Encyclopedia Brittanica.** Encyclopedia Brittanica Online Encyclopedia, 2009. http://www.britannica.com/ (accessed Aug 2009).
8. **World Health Organization.** *Manual of the international statistical classification of diseases, injuries and causes of death—ICD-9*. 9th edn. Geneva: World Health Organization, 1977.
9. **MESH.** Medical Subject Headings, 2009. http://www.nlm.nih.gov/mesh/ (accessed Aug 2009).
10. **Online Mendelian Inheritance in Man (OMIM).** 2008. http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim (accessed Aug 2009).
11. **Diseases Database.** Medical lists and links. V.1.8. 2009. http://www.diseasesdatabase.com (accessed Aug 2009).
12. **MedlinePlus.** MedlinePlus Health Information from the National Library of Medicine, 2009. http://www.nlm.nih.gov/medlineplus (accessed Aug 2009).
13. **eMedicine.** http://www.emedicine.medscape.com (accessed Aug 2009).
14. **WebMD.** 2009. http://www.webmd.com (accessed Aug 2009).
15. **SNOMED Clinical Terms (SNOMED CT).** 2008. http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html (accessed Aug 2009).