UMLS-Based Analysis of Medical Terminology Coverage for Tags in Diabetes-Related Blogs

Zhe He[1*], Min Sook Park[1*], Zhiwei Chen[2]
[1]School of Information, Florida State University
[2]Department of Computer Science, Florida State University

Abstract
There is a well-known terminology disparity between laypeople and health professionals. Using the Unified Medical Language System (UMLS), this study explores an exploratory study on the terminology usages of laypeople, focusing on diabetes. We explain the analysis pipeline of extracting laypeople's medical terms and matching them to the existing medical controlled vocabulary system. The preliminary result shows the promise of using the UMLS and Tumblr data for such analysis.
Keywords: Health informatics; Controlled vocabulary; Terminology disparity; Natural language processing; Social tag
doi: 10.9776/16249

Contact: zhe.he@cci.fsu.edu[1], mp11j@my.fsu.edu[1], zc15d@my.fsu.edu[2]

1 Introduction

Health information users often use their own words and phrases to describe their experiences and knowledge regarding health issues on the World Wide Web. The advance of the social web fueled up ordinary health information users' active participation in health information consumption and creation. However, the words of laypeople are often different from those of medical professionals (Kim, 2013; Messai, Simonet, Bricon-Souf, & Mousseau, 2010; Poikonen & Vakkari, 2009; Smith & Wicks, 2008). The mismatch in vocabulary of laypeople and medical terminologies hinders health information seeking online for ordinary people (Gross & Taylor, 2005; Seedor et al., 2013; Smith & Wicks, 2008).

To bridge the terminology gap between laypeople and health professionals, recently studies have harnessed a massive amount of unstructured textual data available online such as social media data as a live source for health consumer vocabulary (Doing-Harris & Zeng-Treitler, 2011; Jiang & Yang, 2013). The rapidly growing user-generated resources in social media, such as blogs, reflect laypeople's knowledge, experiences, and opinions regarding their health issues in their own words (Oh, Zhang, & Min Sook, 2012). The privileges of the social web for health information users include not only generating contents but also attaching keywords (or tags) to their postings in order to represent their posting contents.

Incorporating laypeople's terms is an effective way to enrich controlled vocabularies that may better fill out users' needs. Moreover, new findings and new terms in the medical domain are constantly evolving (Messai et al., 2010). In this sense, there is a need to develop and update our in-depth understanding on the differences in the usage and the structure of terminology between medical professionals and laypeople. Understanding the differences between the established controlled vocabularies such as SNOMED CT, RxNORM, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and the terms used by ordinary health information users may narrow the terminology gap between experts and non-experts. Thus, the current study aims to understand the differences in vocabularies in laypeople-generated resources and associated tags, based on the most comprehensive biomedical terminological system, the Unified Medical Language System (UMLS) (Lindberg, Humphreys, & McCray, 1993).

The UMLS, created and maintained by the National Library of Medicine (NLM) of the U.S. National Institutes of Health, has integrated more than 9.1 million medical terms from over 120 English source vocabularies into 3.1 million medical concepts in its 2015AA version. Besides English, it has terms in 20 other languages. In this work, we focus on English source vocabularies. In the UMLS, the terms with the same meaning have been mapped to the same concept with a unique identifier. It has integrated all the major controlled vocabularies in biomedicine such as SNOMED CT, RxNORM, ICD-9-CM, LOINC, as well as open access and collaborative Consumer Health Vocabulary (CHV). Thus, the UMLS is an invaluable resource for terminology analysis and translation. Besides, each UMLS concept is assigned

[*] Min Sook Park and Zhe He contributed equally to this work.

one or more of the 127 semantic types, representing the broad semantics of the concept. In this work, we use the UMLS as the dictionary for the analysis.

As a preliminary analysis to observe laypeople's medical terminology usage practices, the current preliminary study explores the degree of resemblance between socially generated tags and the UMLS terms. A particular consideration was given to tags provided by laypeople in blog postings about diabetes in Tumblr.com (www.tumblr.com). Tags are short unstructured terms that laypeople use to represent information resources that they generated in the social media settings (Gruber, 2007; Mathes, 2004), and facilitates networking of related concepts (Abbas, 2010; Vander Wal, 2005; Weller, 2010). Thus, tags are often considered as a condensed form of terms that laypeople use to represent their main ideas in contents of resources they generated in social media (Shadbolt, Hall, & Berners-Lee, 2006; Weller, 2010; Yoon, 2010).

Among diverse medical conditions, the current study focuses on diabetes since it is one of the most prevalent chronic conditions that can lead to an array of serious health challenges. This study has the umbrella research question (RQ) "to what degree are the terms that ordinary health information users use to describe diabetes covered by Consumer Health Vocabulary (CHV)? What are they?" In order to answer the RQ, the following questions guided this study:

RQ 1. For those terms that are covered by the UMLS, what are they?

RQ 1-1. What terms are covered by CHV?

RQ 1-2. What terms are not covered by CHV? Are they covered by other controlled vocabularies in the UMLS (e.g., SNOMED CT, ICD-9-CM, RxNORM)?

RQ 2. For those tags that are not covered by the UMLS, what are they? Should they be considered as health-related terms?

These questions are to test the feasibility of enriching CHV with tags in Tumblr or existing terms in other source vocabularies in the UMLS. We hypothesized that besides CHV, laypeople also use UMLS terms from other source vocabularies and meaningful health-related terms that are not in the UMLS.

This study results could contribute to other research leveraging social media to learn laypeople's metadata practices in health. It could also contribute to developing computer-assisted tools for detecting consumer health representations.

2 Dataset & Methods

2.1 Data set & test bed

A total of 6,186 of tags associated with 709 blogs about diabetes were collected from Tumblr (www.tumblr.com) blogs written in English, using its application program interface (API). In particular, data collection focused on tags associated with text blogs that were posted between February and July 2015 and contained the term *diabetes* as a substring of one of their tags (e.g., 'diabetes', 'diabetes mellitus', 'Type 2 diabetes').

Tumblr was selected as the test bed for the current study because: a. Blogging is one of the two most outstanding online activities for those who live with chronic disease (Fox & Purcell, 2010; Larsen, 2015); b. Tumblr is one of the most popular blogging website with 29.3 million users as of July 2015 (Quantcast, 2015); and c. Tumbler allows its users to add tags with little limitation in terms of linguistic forms to index their blogs in a way that is more meaningful to them based on their subjective ideas.

2.2 Methods

The analytical pipeline of the study is illustrated in Figure 1. The overall analysis procedure of the study is comprised of two phases: in the first phase, tags that occur in blogs of the target disease (i.e., diabetes) were analyzed. The left box of the pipeline describes the analysis procedure. In the second phase (the right box of Figure 1), terms identified in textual blog contents, which are associated the tags, were analyzed. In this work-in-progress paper, we focus on the tag analysis (the left box of Figure 1).

We adapted an existing natural language processing (NLP) tool, OpenNLP (Baldridge, 2005; Sujit, 2015) to match tags in the blogs to UMLS terms and identify their corresponding source vocabularies. Two components of NLP were included in this current study: preprocessing and matching. In the preprocessing component, we processed both tags and UMLS terms (both are called terms hereafter) as follows:

Step 1: Apply tokenization to both tags in the blogs and UMLS terms because they may contain more than one word and each word needs to be processed in the following steps;

Step 2: Remove the punctuation and covert the case of a token to lowercase;

Step 3: Order the tokens of each term in an alphabetical order;

Step 4: Remove the stop words such as “a”, “the”, “is” from the tokens;

Step 5: Match the stemmed form of the tag with the stemmed form of UMLS terms. If there was an exact match between them, we considered the tag as a matching term. After term matching, every tag was matched to 0, one, or multiple UMLS terms irrespective of case, punctuation, order and stop words. The parsed results were stored in a MySQL database for analysis. Figure 2 illustrates the process.

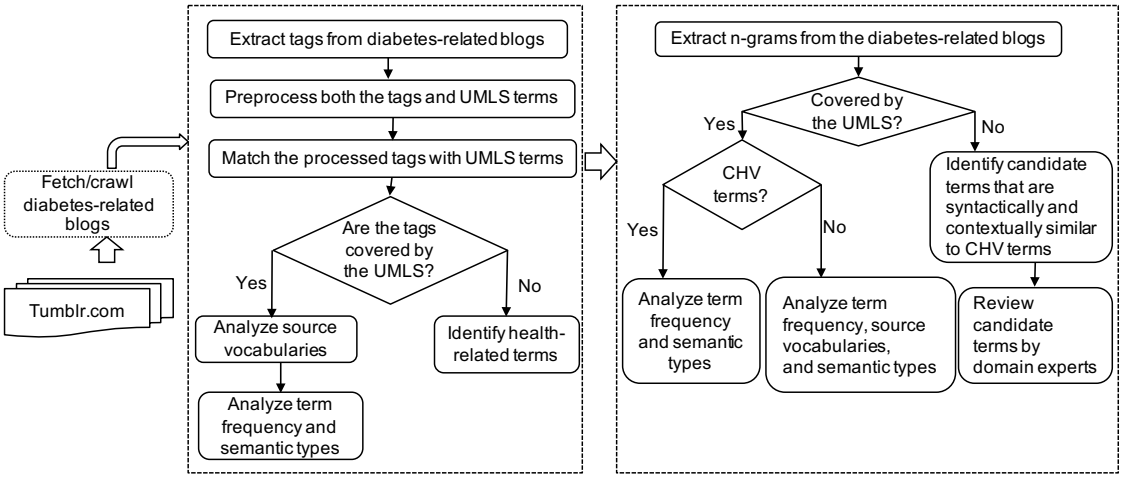

Figure 1. The analytical pipeline of the study.

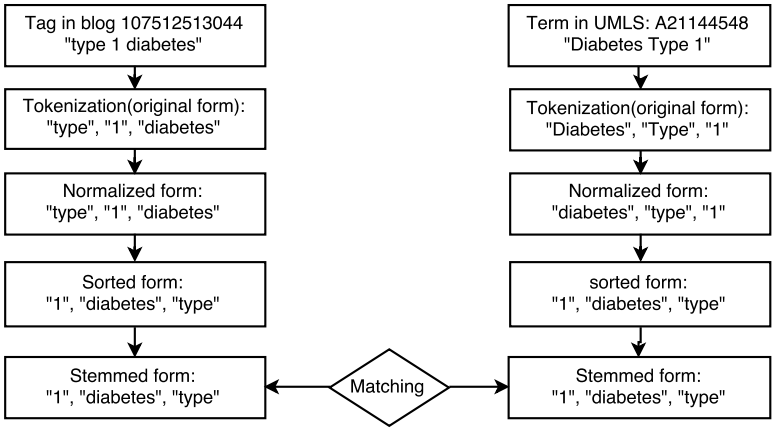

Figure 2: Illustration for the term preprocessing and matching.

For the terms covered by the UMLS, their terminology sources (e.g., SNOMED CT, ICD-9-CM, RxNORM) were identified. As for the tags that are not covered by the UMLS, two researchers of this study manually analyzed the tags that are noun phrases in English to decide if they can be considered as health-related terms. In this this process, the tags that are not in English terms or not noun phrases were manually excluded. Once the two researchers independently identified medical terms from tags that appeared more than twice, the rate of agreement was measured using Cohen’s κ (Cohen, 1960).

3 Results

Among 6,186 tags, 1,044 unique tags were identified. 574 out of 1044 tags (55%) were covered by at least one source vocabulary from the UMLS whereas 470 terms (45%) were not covered by any source vocabulary. For example, the tags such as ‘diabetes mellitus,’ ‘medicine,’ and ‘nursing’ are covered by the UMLS, whereas the tags such as ‘type 1 diabetic,’ ‘diabetes support,’ and ‘cardio’ were not.

3.1 Tags covered by the UMLS

Among the overall tags, CHV covered the largest portion of tags (44%, 464 out of 1044), followed by SNOMED CT (29.2%, 304 out of 1044). The other two important terminology sources, RxNORM, and ICD-9-CM covered only 2.2% and 1.82% respectively. For the tags covered by the UMLS, CHV appears to cover the largest portion of the matched tags (80%, 464 out of 574), followed by SNOMED CT (29.1%). Note that a tag can be covered by multiple source terminologies in the UMLS.

The UMLS integrates over 120 English source vocabularies. The size of its source vocabularies varies. For example, CHV is a relatively small vocabulary with about 58,000 concepts and 146,000 terms, whereas SNOMED CT, the most comprehensive clinical terminology, has about 315,000 active concepts and 793,000 terms. Many tags are covered by more than one source vocabulary. Out of 574 tags that are covered by the UMLS, 62 tags are covered by only one source vocabulary. Out of these 62 tags, 23 tags can be covered solely by CHV. Put it in other words, the 23 tags are purely lay terms that represent laypeople's language to describe diabetes issues. Some of these tags, such as 'blood glucose levels', 'lose weight', 'natural remedy', are highly relevant to diabetes. Only 35 tags can only be covered by both SNOMED CT and CHV. Examples are 'aliments,' 'appetite control,' and 'weight management.' Table 1 shows frequency and percentages of coverage by CHV and three other major controlled vocabularies.

Source Vocabulary	Number of tags	%* of covered tags	%** of all tags
CHV	464	80.8	44.4
SNOMED CT	304	53.0	29.1
RxNORM	23	4.0	2.2
ICD-9-CM	19	3.3	1.82
Unmatched	470	-	45.0%

Table 1. The number and percentage of terms covered by four major UMLS source vocabularies
* The percentage was calculated based on the total number of tags matched to the UMLS (n= 574).
** The percentage was calculated based on the total number of the unique tags (n= 1,044).

For the tags covered by the UMLS, the tag 'diabetes' is most frequently appeared, followed by 'depression,' 'exercise,' and 'alternative medicine'. One the other hand, the most frequently occurring tags covered by SNOMED CT is 'depression,' followed by 'exercise,' and 'acne.' Table 2 describes the top 10 most frequently occurring tags in diabetes-related blogs. Many frequent tags are covered by both CHV and SNOMED CT (e.g., 'depression,' 'exercise', and 'diabetes mellitus').

All Tags		**Tags Covered by the UMLS**					
		Tags covered by the UMLS		Tags covered by CHV		Tags covered by SNOMED CT	
Tags	n	Tags	n	Tags	n	Tags	n
diabetes	650	diabetes	650	diabetes	650	depression	521
depression	521	depression	521	depression	521	exercise	518
exercise	518	exercise	518	exercise	518	acne	515
alternative Medicine	518	alternative medicine	518	alternative medicine	518	diabetes mellitus	64
cardio	516	acne	515	acne	515	cancer	48
beauty	515	beauty	515	beauty	515	health	19
fitness Equipment	515	diabetes mellitus	64	diabetes mellitus	64	type 2 diabetes mellitus	17
acne	515	cancer	48	cancer	48	gastritis	17
diabetes mellitus	64	type 2 diabetes	21	type 2 diabetes	21	obesity	15
cancer	48	health	19	health	19	pancreas	15

Table 2. Top 10 most frequently occurring terms in diabetes-related blogs

3.2 Tags not covered by the UMLS

Tags that are not covered by the UMLS appear to have a wide range of linguistics variations and idiosyncratic terms were identified. Examples include linguistic variations of medical terms like 'type 1 diabetic,' idiosyncratic terms like 'T1IDDM,' and compounding terms like 'diabetes support.' Out of 470 tags that are not covered by the UMLS, 421 occurred only once. In order to identify health-related terms among these tags, two researchers of this study analyzed 36 tags occurring more than twice in the dataset after excluding 13 non-English or non-noun-phrases tags (e.g., adjectives). The researchers agreed on 17 tags (47.2%, 17 out of 36) to be meaningful medical terms. The agreement, measured by Cohen's κ between the two researchers reached 0.79, indicating a substantial level of agreement† (Landis & Koch, 1977). The terms 'cardio' and 'fitness equipment' were identified to be medical terms

† Cohen's κ, the value between 0.81 and 1.00 indicates "almost perfect" in the degree of concordance, followed by "Substantial": 0.61–0.80, "Moderate:" 0.41–0.60, and "Fair:" 0.21–0.40 (Landis & Koch's, 1977).

with the highest frequency. A range of term variations also observed such as misspelled terms, compounding terms, and ellipses. Some terms such as 'artritis' and 'conjuntivitis' were misspelled health terms of 'arthritis' and 'conjunctivitis' respectively. Examples of compounding terms are 'diabetes treatment,' 'diabetes cause,' and 'diabetes treatment.' Table 3 shows the top 10 frequent tags that both researchers identified as health terms, along with the attribute of the tags, and the frequency of the tags in the 709 diabetes-related blogs.

Medical terms uncovered by the UMLS					
Term	Attribute	n	Term	Attribute	n
cardio	-	516	healthinnovations	Compounding words/ grammar violation	9
fitness equipment	-	515	actually diabetic	Compounding words	5
artritis	Typo of arthritis	12	diabetes treatment	Compounding words	4
conjuntivitis	Typo of conjunctivitis	12	type 1 diabetic	Word variation	4
Hodking	Ellipses of Hodgkin lymphoma	12	diabetes diet	Compounding words	3

Table 3. Top 10 most frequently occurring health terms uncovered by the UMLS

4 Conclusion & Plan for the Future Study

The UMLS covered slightly more than half (55%) of the tags in diabetes-related blogs. Although overall coverage by CHV was salient across the tags covered by the UMLS, only a small number of tags (n=23) were covered solely by CHV. This result of this preliminary study supports the assertions in previous studies that that consumer-authored blogs have the potential to enrich laypeople's vocabulary in CHV.

Among the tags uncovered by the UMLS, health-related terms were identified. Yet, they are varying in their forms, misspelled, or with grammar violations. Considering that controlled vocabulary requires not only conceptual coverage (Cimino, 1998) but also consistency in quality (Peters, 2009; Svenonius, 1989), the identified tags must be further processed before incorporating into a controlled vocabulary such as CHV. Nevertheless, the inclusion decisions of the potential new terms should be made by the curators of the controlled vocabulary. By keeping incorporating laypeople's terms, CHV would help index health resources in a more user-centered approach. This project also lays a necessary foundation for building a consumer-oriented health information search engine.

The current study has a few limitations. The current study concentrated on term coverage of the tags. A tag may be matched to multiple terms with different semantics (i.e., multiple UMLS semantic types). In order to capture the semantics of the identified terms, we will perform word sense disambiguation (WSD) on the terms that are mapped to multiple terms in the UMLS. For WSD, we will enhance existing methods by combining knowledge based similarity metric (path based, adapted Lesk, information content based method (Jimeno-Yepes & Aronson, 2010) and statistical based metric (e.g., location-based TF-IDF (term frequency- inverse documents frequency)). Then the researchers will give each metric a weight, which can be obtained by leveraging NLM labeled corpus (Jimeno-Yepes, McInnes, & Aronson, 2011; U.S. National Library of Medicine, n.d.) as the training data in a linear regression model. We will add new similarity metric like POS (Part-of-Speech) and semantic type similarity.

5 References

Abbas, J. (2010). *Structures for organizing knowledge: Exploring taxonomies, ontologies, and other schemas.* New York, NY: Neal-Schuman Publishers.

Baldridge, J. (2005). The opennlp project. Retrieved from http://opennlp. apache. org/index

Cimino, J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, *37*(4-5), 394.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement. *Educational and Psychological Measurement*, *20*, 37–46.

Doing-Harris, K., & Zeng-Treitler, Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. Journal of medical Internet research. *Journal of Medical Internet Research*, *13*. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3221384/

Fox, S., & Purcell, K. (2010). *Chronic disease and the Internet*. Pew Research Center. Retrieved from http://www.pewinternet.org/2010/03/24/chronic-disease-and-the-internet/

Gross, T., & Taylor, A. (n.d.). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries*, *66*(3), 212–230.

Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems*, *3*(1), 1–11.

Jiang, L., & Yang, C. (2013). Using co-occurrence analysis to expand consumer health vocabularies from social media data (pp. 74–81). Philadelphia, PA.

Jimeno-Yepes, A., & Aronson, A. (2010). Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics*, *11*(1), 569.

Jimeno-Yepes, A., McInnes, B., & Aronson, A. (2011). Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics*, *12*(1), 223.

Kim, S. (2013). An exploratory study of user-centered indexing of publishing biomedical images. *Journal of Medical Library Association*, *101*(1), 73–76.

Landis, L., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Larsen, P. (2015). *Lubkin's Chronic Illness: Impact and Intervention* (9th ed.). SudburyUnited States: Jones and Bartlett Publishers, Inc.

Lindberg, D., Humphreys, B., & McCray, A. (1993). The Unified Medical Language System. *Methods of Information in Medicine*, *32*(4), 281–291.

Mathes, A. (n.d.). Folksonomies - cooperative classification and communication through shared metadata. *Computer Mediated Communication*, *47*(10). Retrieved from https://scholar.google.com/scholar?q=Folksonomies+-+cooperative+classification+and+communication+through+shared+metadata.&btnG=&hl=en&as_sdt=0%2C10

Messai, R., Simonet, M., Bricon-Souf, N., & Mousseau, M. (2010). Characterizing consumer health terminology in the breast cancer field. *Studies in Health Technology and Informatics*, *160*(1), 991–994.

Oh, S., Zhang, Y., & Min Sook, P. (2012). Health information needs on diseases: A coding schema development for analyzing health questions in social Q&A. In *Proceedings of the 75th Annual Conference of the American Society for Information Science & Technology (ASIST' 12)*. Baltimore, MD.

Peters, I. (2009). *Folksonomies: Indexing and retrieval in Web 2.0*. Berlin, German: Deutsche Nationalbiblothek.

Poikonen, T., & Vakkari, P. (2009). Lay persons' and professionals' nutrition-related vocabularies and their matching to a general and a specific thesaurus. *Journal of Information Science*, *35*(2), 232–243.

Quantcast. (2015). Tumblr.com. Retrieved from https://www.quantcast.com/tumblr.com

Seedor, M., Peterson, K., Nelesen, C., McCormick, J., Chute, C., & Pathak, J. (2013). Incorporating expert terminology and disease risk factors into consumer health vocabularies. In *Pacific Symposium on Biocomputing* (pp. 421–432).

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web revisited. *Intelligent Systems, IEEE*, *21*(3), 96–101. http://doi.org/10.1109/MIS.2006.62

Smith, C., & Wicks, P. (2008). PatientsLikeMe: Consumer health vocabulary as a folksonomy. In AMIA annual symposium proceedings (p. 682).

Sujit, P. (2015), Dictionary based annotation at scale with spark SolrTextTagger and OpenNLP, Retrieved from https://data.mendeley.com/datasets/4xdkh7xdtt/1. doi: 10.17632/4xdkh7xdtt.1

Svenonius, E. (1989). Design of controlled vocabularies. In *Encyclopedia of Library and Information Science* (pp. 82–109).

U.S. National Library of Medicine. (n.d.). NLM WSD Test Collection. Retrieved from http://wsd.nlm.nih.gov

Vander Wal, T. (2005). Tagging for fun and finding. Retrieved from http://okcancel.com/archives/article/2005/07/tagging-for-fun-and-finding.html

Weller, K. (2010). *Knowledge representation in the Social Semantic Web*. Berlin, German: Walter de Gruyter GmbH & Co.

Yoon, J. (2010). Indexing. In M. Norton (Ed.), *Introductory concepts in information science* (2nd ed., pp. 67–86). Medford, NJ: American Society for Information Science and Technology.