Open Source and Collaborative Development of Consumer Health Vocabulary

Authors: Qing T. Zeng, Ph.D. [1]
         Tony Tse, Ph.D. [2]

Affiliations: [1] Department of Radiology
               Decision Systems Group
               Brigham and Women’s Hospital
               Harvard Medical School
               Boston, MA

               [2] Lister Hill National Center for Biomedical Communications
                   National Library of Medicine
                   National Institutes of Health
                   Department of Health and Human Services
                   Bethesda, MD

Correspondence Qing T. Zeng, Ph.D.
              Department of Radiology
              Decision Systems Group, Thorn 309
              Brigham and Women’s Hospital
              Harvard Medical School
              75 Francis Street
              Boston, MA 02115
              Phone: (617) 732-7694
              email: qzeng@dsg.bwh.harvard.edu

Running Head: Developing Consumer Health Vocabulary
Abstract

The development of consumer health vocabularies for facilitating information retrieval, comprehension of retrieved documents, and other consumer health information seeking tasks is essential for consumer health informatics. Mismatches in language between consumers and professionals hinder effective communication, health information seeking, and, ultimately, informed decision making. While research has been conducted recently to identify and characterize consumer health vocabularies (CHV) and to learn about the effects of the vocabulary gap on health information retrieval, there have been few systematic efforts to develop CHVs. This paper presents the point of view that, despite the substantial theoretical and practical challenges in defining and developing a CHV, there are practical methods for leveraging existing informatics resources. The authors explore issues surrounding CHVs and advocate a pragmatic approach to developing an open source, “first-generation” CHV based on the Unified Medical Language System® (UMLS®).

Introduction

Although it is long-recognized that lay persons (“consumers”) and health care professionals express and think about health-related concepts differently, such mismatches in language continue to hinder effective communication, health information seeking, and, ultimately, informed decision making. This “vocabulary gap” may be especially urgent today as more consumers are looking for health-related information on their own and assuming greater responsibility in personal health care. Further, technical terminology (or jargon) is a barrier to health literacy, defined as “the degree to which individuals have the capacity to obtain, process,
and understand basic health information and services needed to make appropriate health
decisions. [1]” And low health literacy has been associated with poor health (see [2] for literature
review).

In the general biomedical literature, research on the consumer understanding of medical words
and concepts has focused, primarily, on lists of discrete terms in various specialties: e.g., 17
oncology terms [3]; six word pairs among emergency department patients [4]; and 50 “common”
medical terms among surgical patients and family members [5]. While such studies consistently
confirm that consumers have difficulty comprehending medical jargon, they do not provide a
methodology for systematically obtaining alternate words or phrases (or collectively,
“expressions”) for general medical concepts that consumers are likely to recognize and
understand.

However, concerted informatics research on health-related expressions used by consumers, or
“consumer health vocabularies” (CHV), began relatively recently (e.g., [6, 7, 8, 9]). In 2001, a
distinguished panel on consumer vocabularies highlighted challenges in creating a standardized
consumer vocabulary at Medinfo and concluded that “The development of a consumer
vocabulary should be based on research that includes consumer information needs and
consumers’ ways of talking about expressing those needs. [10]” Analogous to professional
terminologies in the National Library of Medicine® Unified Medical Language System®
(UMLS®) [11], which links biomedical knowledge across disciplines and applications, CHV
might likewise link consumer expressions to terms used by professionals: “The healthcare
consumer then would not need to know the language of the healthcare provider in order to
effectively find relevant information resources, and he/she could search for resources in his/her
own language… [10]”
While exploratory research has been conducted recently to identify and characterize CHVs (e.g., [12, 13]) and to learn the effects of the vocabulary gap on health information retrieval (HIR) [14], there have been few systematic efforts to collect and develop CHV (see [15] for a comprehensive review). Although several companies have proprietary CHV products, the details of the development, maintenance, and evaluation of these CHVs have not been reported.

This paper presents the point of view that, despite the substantial theoretical and practical challenges in defining and developing a “standardized consumer vocabulary” as envisioned by Lewis and colleagues [10], there are practical methods for leveraging existing resources to begin developing an open-source CHV, an important if not essential tool for consumer health informatics. In this paper, we contend that:

- Consumers use words and phrases (expressions) to describe health-related concepts and relationships among such concepts that frequently differ from those of professionals;
- Such consumer-designated health-related forms and concepts are sufficiently consistent—at least among discourse groups for particular tasks—to be identified and characterized as distinct CHVs; and
- Research and development of CHV is an important goal for consumer and public health informatics.

We then outline a pragmatic approach to developing a first-generation CHV based on the UMLS while conducting further research on CHV. Our intention is to increase awareness of the need for such a vocabulary, stimulate dialog and action on the development of an open-source CHV, and ultimately, help to meet the challenges of bridging consumers with relevant and appropriate medical information.
General Language and the Health Domain

Consumers use words and phrases (expressions) to describe health-related concepts and relationships among such concepts which frequently differ from those of professionals. While health care specialists share foundational domain knowledge based on formal education and professional experience, laypersons have some socially- and culturally-derived notions of health and illness, basic knowledge from formal and informal education (e.g., media exposure), and unique personal experiences. Thus, whereas terminology facilitates discourse among professionals in their everyday routines, jargon is likely to be incomprehensible or confusing for consumers, who typically have a fragmented knowledge of health terminology—both the form (written or spoken), underlying concepts they refer to, and relationships to other concepts.

During face-to-face communication between consumers and professionals, the interactive nature of such encounters allows for the possibility of negotiating ideas at a common level of discourse. For example, either person may ask for clarification or repeat important points when confusion is detected. However, in actuality, such idealized discourse patterns often do not occur between provider and consumer. Interactions between humans and computers create even greater challenges for adapting discourse to facilitate understanding. Finally, “static” and public-oriented media such as brochures, pamphlets, newspapers, and television, where many people obtain their health information, rely on a “one-size-fits-all” or “lowest common denominator” approach and the knowledge and judgment of individual authors. During any of these interactions with a provider, computer, or document, consumers may “fill in” any vocabulary gaps (correctly or incorrectly), using their own knowledge, experience, and preferences. In the best case, consumers will infer meaning from a term correctly and even adopt the expression as their own in health communication (e.g., “attention deficit disorder”). Among the worst case scenarios,
consumers misinterpret or (mis)associate a term with context or connotations not intended by the health care professional, content provider, or author. Rubin [16] recounted a physician-patient encounter where the term “local treatment” was interpreted by a patient in a geographical, rather than anatomical, sense. Conversely, consumers may recognize or use a technical form, but associate it with a different or incomplete concept in common usage (e.g., “depression” for “sadness” rather than “depressive disorder”). Consumers who do not or cannot “fill in” the gaps at all (e.g., due to low health literacy) would then simply be confused, the information having been “lost in translation.”

On the other hand, consumers may use general language expressions with personal or cultural meanings or connotations not commonly used in professional discourse. Further, expressions such as “hair loss” or “heart failure” have more precise meanings for health professionals than for consumers. The range of such expressions may vary from general and descriptive (e.g., “device to look inside my ear” for “otoscope”) to specific but colloquial (e.g., “sugar” for “diabetes”). Wood [17] describes social, cultural, and economic influences on associations that consumers may make with medical terms. Sontag [18] has written about popular and literary conceptions of illness and common metaphors for various diseases including cancer, tuberculosis, and AIDS. Thus, consumer discourse on the health-related topics may include a combination of technical terminology and general language expressions (Figure 1), with many possible interpretations based on individual, contextual, societal, and cultural associations. Whether the study of consumer expressions is better suited for a specialized lexicology of the health domain, a “popularized” health terminology, or some combination is debatable. What is clear is that consumers have distinct ways of communicating (general language in a health domain) than professionals (specialized health domain-specific language).
Informatics and Consumer Health Vocabularies

Research and development of CHVs is an important goal for consumer and public health informatics. Eysenbach [19] has observed that “consumer health informatics is designed to empower consumers by putting health information into their hands, including information on their own health, such as diagnoses, lab results, personal risk factors, and prescribed drugs (p. 1714).” Given the current public interest in health information and advances in consumer health informatics, we believe that a concerted effort at developing a first-generation open-source CHV at the present time would improve consumer access to health information considerably, especially through computerized health information applications and systems.

Even though CHVs have been researched and developed in other consumer health non-informatics domains (e.g., health communication), free-text lists and dictionaries of lay expressions are not sufficient for consumer informatics applications any more than medical dictionaries could replace controlled medical vocabularies in clinical informatics systems. Health communication research has provided much insight into the “how” and “why” issues regarding the CHV “problem”. For instance, Scott and Weiner [20] developed a typology to describe technical terms that appear in their “Patientspeak” Dictionary, including:

**Figure 1. Consumer health vocabularies as “termino-lexicology”**
• Difficult general language words having the same meanings as technical terms (e.g., alleviate, apprehensive, and intermittent)
• Technical terms (e.g., angina, bronchus, and diastolic)
• General language words with different technical meanings (e.g., depression – clinical depressions vs. feeling sad – and negative)

Informatics research as an applied field should and could build on such insight to provide computing solutions to the vocabulary gap.

From an application systems standpoint, computers need to recognize consumer expressions and their intended meaning. For instance, a natural language processing component of an HIR system would need to recognize that “water on the knee” is a consumer expression synonymous with the medical term “knee joint effusion.” From a consumer perspective, laypersons need to be able to understanding the information that is retrieved. For example, consumers are more likely to understand health information when “gastroesophageal reflux” and “GERD” are supplemented by the common term “heartburn,” thereby providing readers with both technical and lay terms. Thus, a CHV could provide support for both HIR and lay comprehension of health information.

Thus far, professional medical vocabulary support has been employed by virtually every type of biomedical application from microarray data mining to disease outbreak surveillance. Many specialties in the biomedical domain (e.g., pathology or nursing) have also created their own vocabularies. As consumer health applications increase in number and sophistication, we predict that generations of CHV will play the same important role as professional vocabularies currently do for clinical and biomedical systems.

**Defining Consumer Health Vocabularies**
At the intersection of lexicology and terminology, general language for medical concepts involves complex interactions between forms and concepts, represented by everyday expressions that refer to technical concepts. As Lewis and colleagues [10] point out, “to presume that [the language of consumers] contains a knowable stable vocabulary and grammar similar in structure to that of the formal languages imposes a professional structure on a very personal experience. …a consumer health vocabulary will need to consist of ‘normal standard ways of expressing things’ (in everyday life)…” Yet, tailoring language dynamically or “just-in-time” for each patient and her/his circumstance is not likely to be realized in the near future.

Although what exactly constitutes a CHV requires further study, a practical first step should be to identify expressions that a group of consumers commonly use for medical concepts. For example, research by us and others suggests that some general language health expressions used by consumers—forms referring to health-related and medical concepts—are sufficiently consistent, at least among particular discourse groups, to constitute a CHV [13, 14]. For a narrowly defined population (e.g., online health information seekers at a particular resource) and task (e.g., HIR), there exists a “common level of discourse” in which a majority of the people expresses concepts using particular “consensus” forms. While acknowledging the “expressive variability” (e.g., synonymy and lexical variations) inherent in general language, it may be possible to define an effective “common” vocabulary that will be accessible to most members of a discourse group, and be easily learnable by others. For instance, analysis of queries submitted to NLM MedlinePlus® <http://www.medlineplus.gov>, a consumer health information system, for a 12-month period indicates that the expression “kidney stones” appears over 95% of the time for the concept “kidney calculi.” Such consensus forms (and common synonyms), for example,
might be used to label medical concepts, thereby improving consumer access to relevant health information.

Therefore, we operationally define a first-generation CHV as a collection of forms that represent health-related concepts (and their relations) used in health-oriented communication for a particular task (e.g., information retrieval) by a significant percentage of people from a particular discourse group (e.g., users of a consumer health information system). One implication is that CHV has significant overlap with controlled professional vocabularies, such as those in the UMLS. As Zielstorff [15] asserts, “[developing a CHV] would likely be a multidisciplinary effort, including not only nurse informaticians and nurse clinicians, but linguists, medical librarians, and other health disciplines and, of course, patients themselves (p. 332).” CHV and various professional vocabularies can have both degrees of significant overlap as well as substantial differences.

Another implication in taking a pragmatic approach is that the primary criterion for considering a consumer expression is not its clarity, correctness, or biomedical basis, but rather usage and usability in health communication among consumers. For example, even though “myocardial infarction” is a more accurate and scientific name than “heart attack,” it is less familiar to laypersons. In a CHV, “heart attack” should thus be designated as the preferred “consumer form” of the concept, based on evidence that most health information seekers use this “consensus form” to express the same concept, myocardial infarction.

**Research and Development: A Practical Approach**

Given the complexities of developing a “complete” CHV that would take into account various factors that influence how consumers express and think about health-related topics, we have provided a practical definition of a first-generation CHV. In our view, development of such a
CHV using the existing informatics technologies should begin immediately, while research on broader issues of consumer lexico-terminology and various cultural and social influences should be conducted in parallel. Overall, we advocate an open source, collaborative, and text-corpus-based approach.

Typically, vocabularies contain three components (Figure 2): (1) forms; (2) concepts; and (3) relations (form-form, form-concept, or concept-concept). Indeed, preliminary studies suggest that consumer health language does differ from that used by professionals at all three levels [14].

Figure 2. Relationships between components of consumer and professional vocabularies

*Forms:* Differences at the form level are well known, such as “heart attack” for “myocardial infarction.” In our view, a CHV should not only simply list all consumer forms for a concept, but also provide a measure of comprehensibility and other attributes of the concept names. For instance, one concept may have multiple consumer or professional forms and the readability of each form by a lay audience would be represented by a probability gradient. In other cases, extremely common misspellings or non-standard abbreviations may be included in CHV for parsing consumer generated text, but not for display.
Concepts. Whether or not lay medical concepts exist or not is questionable. Some researchers hold the view that consumers do not have health concepts independent of professional medical concepts, just different forms or inaccurate perceptions. Others (among whom we count ourselves) hold the opinion that there are indeed consumer concepts. As an example, “overweight” is a term that both consumers and physicians commonly use. In the UMLS and its source vocabularies, “overweight” may be mapped to two concepts: (1) Obesity and (2) Overweight (BMI < 30).

Both concepts have strictly defined medical semantics: a person’s condition is obese when his/her body mass index (BMI) >= 30 and overweight when the BMI falls between 25 and 30. In consumer usage, the meaning of “overweight” is better captured by the American Heritage® Dictionary of the English Language, Fourth Edition [21]: “more weight than is normal, necessary, or allowed.” We would thus consider such a general language notion of “overweight,” which is not represented in the UMLS, to be a “consumer concept.” It would then be necessary to represent “consumer concepts” in a CHV for a computer application to successfully map consumer entered forms to the appropriate concepts. In this case, the application would have the common English concept of “overweight” in its vocabulary, which would be mapped to both concepts: obesity and overweight (BMI < 30).

Relations. Differences in relations (form-form, form-concept, and concept-concept) between consumer and professional vocabularies also exist. This partly results from differences in form and concept usage, but also reflects considerable differences in professional and lay mental models [22]. When we compared terms used by professionals and consumers to look up physicians in an online directory, we found that organs and body parts were more frequently used by consumers to search for medical specialties [14]. In UMLS, “wrist” and “carpal tunnel
syndrome” are not directly linked. In a set of consumer health queries we examined, “wrist” and “carpal tunnel syndrome” co-occurred quite often, suggesting a short semantic distance in consumers’ mental models (e.g., “carpal tunnel syndrome” occurs_in “wrist”).

To develop a first-generation CHV, we believe the focus should be on forms and especially the more frequently used consensus forms. For such development, a large corpus is required to account for variability in consumer health language and to calculate the frequency of usage. Past research has relied on interviewing consumers for their understanding of different terminology. While this type of research provides in-depth information about what consumers actually comprehend, qualitative methodologies such as interviews and surveys do not scale up sufficiently for a CHV. Also, such approaches are difficult to automate for use with informatics tools. Thus, a first-generation CHV will identify common consumer expressions in a machine-readable format rather than a comprehensive view of consumer forms, concepts, and relations. Further research will be required to fine-tune the CHV.

Materials for the corpus are to be obtained from various machine-readable sources of consumer-authored text. With the proliferation of health applications for the public, such as health information Web sites and newsgroups, we are able to sample large, representative quantities of consumer-authored health-related text and search queries. A CHV based on forms, concepts, and relations most often used on the Web will approximate common usage by online health information seekers. The familiarity of a word or phrase may be inferred using statistical analyses, such as frequency of occurrence and co-occurrence of n-grams. A practical strategy is to map the most frequently occurring forms to the UMLS, providing an entry vocabulary for consumers into the professional vocabulary.
The resulting CHV may be evaluated on several levels. Completeness/coverage of the CHV may be evaluated in the same manner as professional vocabularies: validating automatically encoded text through human review, such as the large scale vocabulary test of UMLS [23]. Its utility may be evaluated in conjunction with natural language processing and information retrieval applications.

One unique feature of the said CHV is an indicator of a form’s comprehensibility in the vocabulary, for instance, marking one or more terms as “consumer preferred forms”. Such designation may be evaluated by using a modified Cloze procedure where lay participants are asked to read materials with missing words. The participants must select the most appropriate word that fits in the blank from among closely related technical terms and consumer expressions. Lay participants, controlled for domain knowledge, should be able to select preferred consumer terms with greater speed and accuracy than the corresponding technical terms.

We and colleagues in our respective institutions have started laying the groundwork to develop such a CHV. In particular, we have started collecting a text-corpus, mapping it to the UMLS, using automated tools, and creating Web-based systems for human review of the mapped and unmapped forms. Given the enormity of the task, and its potential for contributing to consumer health information seeking, we invite members of the community to join in this and other efforts to develop open source CHVs.

**Conclusion**

Patel and Kaufman [24] have suggested that medical informatics is a “local science of design.” That is, the field is “local” to specific aspects or parts of the broader domain (contrast with “global”) and “design” because it looks toward pragmatic outcomes: “in practice, design is strongly bound by domain-specific constrains and grounded in contexts in which an artifact is to
be used.” We believe that our view of developing a first-generation CHV for a limited group of consumers for specific tasks exemplifies their thesis. Development of a CHV should not be hampered by the complexities of what is currently unknown about this multi-dimensional construct, but embrace the challenge of making medical information more accessible to the public. “One of the goals of a local science of design is to discover what works and then determine why some things work and others don’t. A working system is an outcome not merely of technology but of the social and cognitive processes of integrating such a system into the daily workflow. [24] (p. 490)”

Preliminary studies suggest that the “vocabulary gap” between consumers and professionals is a substantial barrier to health information access for laypersons. A first-generation CHV is likely to support incremental improvements in consumer health information seeking while providing new insights into the overall problem of bridging lay forms, concepts, and relations with the medical domain. Simultaneously, research on the contextual, socio-cultural, and other factors that affect how laypersons express and think about health topics will illuminate the other dimensions of CHV. By tackling small pieces of the larger problem, it is hoped that a convergence in understanding of the overall phenomenon will result from a concerted, interdisciplinary effort by the consumer health informatics research community.
References


