Exploring Medical Expressions Used by Consumers and the Media: An Emerging View of Consumer Health Vocabularies

Tony Tse[a,b] and Dagobert Soergel[a]
[a]College of Information Studies, University of Maryland, College Park
[b]National Library of Medicine, Bethesda, MD

Healthcare consumers often have difficulty expressing and understanding medical concepts. The goal of this study is to identify and characterize medical expressions or "terms" (linguistic forms and associated concepts) used by consumers and health mediators. In particular, these terms were characterized according to the degree to which they mapped to professional medical vocabularies. Lay participants identified approximately 100,000 term tokens from online discussion forum postings and print media articles. Of the over 81,000 extracted term tokens reviewed, more than 75% were mapped as synonyms or quasi-synonyms to the Unified Medical Language System® (UMLS®) Metathesaurus®. While 80% conceptual overlap was found between closely mapped lay (consumer and mediator) and technical (professional) medical terms, about half of these overlapping concepts contained lay forms different from technical forms. This study raises questions about the nature of consumer health vocabularies that we believe have theoretical and practical implications for bridging the medical vocabulary gap between consumers and professionals.

INTRODUCTION

As health consumers seek an active role in their own care, as informed life-style choices and prevention are promoted to improve personal health, and as public health concerns require increased epidemiological surveillance, including public awareness and vigilance, communication of medical information across the lay-professional boundary increases in volume and importance:

- Health consumers need medical information and obtain it not only from their physicians but also through searching on the Web, through direct-to-consumer advertising, drug inserts, and patient education, mirroring a general trend towards end-user searching and end-user computing.
- Physicians need to understand patients' reports on their conditions (such as severity of pain, degree of discomfort), especially difficult in telemedicine or medicine otherwise mediated by technology.
- Policy makers and health administrators need to collect data from the public and to alert them about medical issues, including natural and artificial health threats (e.g., SARS, monkeypox virus, bioterrorism).

In all these cross-boundary communications, terminology and understanding of medical concepts are serious barriers. Non-specialists often do not understand technical terms and explanations or interpret them differently, based on their personal and cultural experiences, education levels, and cognitive and affective states of mind [1]. Conversely, professionals and medical information systems may have difficulty in correctly interpreting lay health expressions and associated conceptualizations. Thus, research on how consumers express medical concepts provides insights that help bridge the terminology gap in bi-directional health-related communication between lay persons and professionals.

Although the "consumer vocabulary problem" [2] has long been recognized, "personal health vocabularies have only recently been afforded importance in the literature…" [1:1485]. While terminologies for medical professionals continue to evolve, few consumer-level vocabularies have been explored ([3,4], for example). As Lewis et al. observed, "The development of a consumer vocabulary should be based on research that includes consumer information needs and consumers' ways of talking about and expressing those needs" [5:1530]. This suggestion parallels the trend towards user involvement in the development of better end-user systems.

A goal of this study is to identify and characterize *terms*—a *form* and its underlying *concept*—used by two groups of non-specialists: (1) consumers and (2) health information mediators (included because we hypothesized that they might represent a "natural bridge" between consumers and professionals). Problems in consumer-professional communication may occur at various levels, such as:

- shared forms/different concepts (e.g., *negative*: "unfavorable" vs. "no indication")
- different forms/shared concepts (e.g., *blood cancers* and *hematologic malignancies*)
- different forms/different concepts (e.g., *soul*; no equivalent professional term)

Understanding the extent of these differences may lead to ways to improve communication.

Several studies have analyzed terms extracted from electronic sources used by the public. McCray et al. [6] used queries put to the NLM homepage[1]. They removed lexical variations and mapped the terms to the UMLS. Unmapped terms included long descriptive phrases, misspellings, truncations of eponymic terms (*Crohn's* for *Crohn's Disease*), and abbreviations or word fragments (e.g., *cranio*).

Zeng et al. [7,8] used queries put to a Find-A-Doctor site[2] and MEDLINEplus[3]; they found problems at the lexical level (e.g., spelling, morphology, and word order) and the semantic level (e.g., synonymy).

Smith et al. [9] used email submitted by consumers to a cancer information service[4]. They extracted 504 unique terms representing "features and findings," and mapped them to the UMLS. The few (4%) unmapped terms consisted primarily of typographical errors, but included legitimate medical terms not in the UMLS and one abbreviation (*endo*).

The current study [10] contributes to this growing body of work and extends it.

METHODS

We generated the vocabularies in two steps and then analyzed the collected terms (see [11] for details).

Vocabulary generation. *Corpus Generation*. For the consumer corpus, we collected 1,936 archival postings from 12 Web-based health discussion forums; for the mediator corpus, we collected 208 documents: articles from popular magazines and newspapers, commercial ads, government publications, and patient pamphlets. Two controlled medical vocabularies were used as "surrogates" for a professional medical vocabulary (PMV): MeSH® (2002) and SNOMED International® (1988).

Term extraction, processing, and mapping. To reflect different consumer viewpoints, 14 laypersons identified medical expressions from the documents. Within the guidelines provided, extractors selected terms on the basis of their personal experience, knowledge, judgment, and context in the document. Each document was reviewed by two extractors.

The extracted terms were processed (including spelling correction, acronym and truncation expansion); normalized, using UMLS lexical tools Norm and LVG; and then mapped to concepts in the 2000-2001 UMLS Metathesaurus using MetaMap and the Knowledge Source Server. The many terms not mapped automatically were manually mapped to the UMLS by the first author, with assistance from a physician consultant. Mappings were categorized as *close* (identical or quasi-synonyms); *approximate* (other relationship, e.g., generic/ specific); and *none*. Due to time constraints, only 65% of CMV terms (selected at random) were processed.

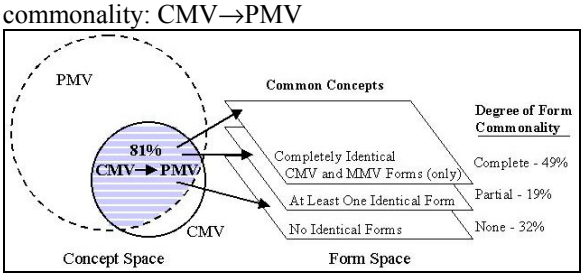
Analysis. We first analyzed the terms with respect to:

- form-based characteristics
- concept-based characteristics
- form-concept-based characteristics: *expressive variability* (number of forms per concept) and *consensus form* (preferred form for a concept)

Second, we analyzed relationships between vocabularies with respect to conceptual and form overlap. One-sided conceptual overlap between a source vocabulary (S) and a reference vocabulary (R) or S→R (Figure 1) is defined as follows:

S-concepts also in R / all S-concepts

Figure 1. Schematic of concept overlap and form commonality: CMV→PMV



For a given common concept, form commonality (Table 1) is defined as follows:

S-forms for the concept that are also R-forms for the concept / All S-forms for the concept

Table 1. Examples for different levels of form commonality

Form Commonality (Concept CUI)	CMV Form	PMV Form
Complete (C0003842)	artery	Arteries
Partial (C0042963)	vomit throw up	Vomiting Emesis
None (C0003449)	cough medicine suppressant	Antitussive Agent Antitussive Drug

Patterns found in this exploratory study may provide insights into the nature of consumer health vocabularies, methods for characterizing lay terms, and promising future research directions.

[1] http://www.nlm.nih.gov/
[2] http://www.brighamandwomens.org/mdSearch/
[3] http://medlineplus.gov/
[4] http://www.upci.upmc.edu/internet/about/contact.html

Table 2. Comparison of consumer term characteristics across studies

	Current Study [10]		Zeng et al.	Zeng et al.	Zeng et al.	McCray et al.
	CMV	MMV	[7]	[7]	[8]	[6]
General						
Sample population	consumer	mediator	consumer	consumer	consumer	consumer
Source	forums	print media	BWH Web	MLP Web	BWH Web	NLM Web
Source material	postings	articles	queries	queries	queries	queries
Form Characteristics						
Tokens	55,054	45,774	11,182	16,743		225,164
Types	24,952	21,282	3,246	10,342		
Average frequency (tokens/type)	2.2	2.2	3.4	2.1		
Normalized types	22,842	18,007				128,640
Forms appearing just once	74%	73%	64%	44%		
Mean words per form by token	1.6	1.6	1.5	2.0		
Mean words per normalized form by token	2.2	2.2			2.4	
UMLS Concept Characteristics						
Term mapping method	semi	semi	auto	manual	auto	auto
Total mapped (token)‡	99%	95%				
Total mapped (type)‡	59%	78%				
Closely mapped terms (token)	84%	75%	78%	88%		
Closely mapped terms (type)	36%	43%	49%		62%	41%
Subdomain Representation						
Disorders	34%	23%			23%	43%
Procedures	14%	11%			11%	9%
Chemicals and Drugs	10%	12%				20%
Concepts and Ideas	10%	12%				
Occupations	6%	6%			36%	

‡Includes both close and approximate mappings. Abbreviations: BWH (Brigham and Women's Hospital Find-a-Doctor site), MLP (MEDLINEplus), Semi (Semi-automatic mapping), Auto (Automatic mapping)

RESULTS

A total of 100,000 form tokens were extracted (Table 2); the pair-wise inter-extractor overlap was 55% complete, 22% partial, and 23% none. All forms were used for subsequent analysis. The first author reviewed the terms and modified approximately 6% to conform to the guidelines.

The overall results of the term-based analysis are juxtaposed with results from comparable studies, as shown in Table 2. Although these other studies have used different document sources and techniques for term analysis, the findings are comparable.

Form-Based Characterization. Average normalized form lengths by character and by word were similar for CMV (16.8 characters; 2.2 words) and MMV (18.2; 2.2). For comparison, the corresponding values for PMV were 23.5 and 2.4 See Table 2 for basic data. The frequency of normalized forms follows a Zipf distribution, with *doctor* occurring most frequently in both vocabularies.

We observed many non-regular forms in both vocabularies, with more in CMV than MMV:

- abbreviations/acronyms (*ANA*, *PSA*, *Dr.*)
- clippings/truncations (*med*, *oxy*, *doc*)
- idiom (*plumbing*, *going to the bathroom*)
- definitions/descriptions (*heart doctor*, *delay between heartbeats*, and *gallbladder removal*)
- misspellings/typos (*gaulbladder, lupis*)
- less frequent patterns (e.g., many modifiers, exemplars to represent classes, and neologisms)

Concept-Based Characterization. Close mappings to UMLS concepts (identical and similar meanings) were found for 36% CMV term types (representing 84% CMV tokens) and 43% MMV term types (75% MMV tokens), as shown in Table 2. Of the terms mapped to the UMLS, CMV and MMV show similar distributions of concept tokens by subdomains (Table 2); however, within the *Disorders* group, CMV showed a preponderance of semantic types related to symptoms, while MMV showed a preponderance of disease.

Form-Concept Relation. "Expressive variability," number of forms per UMLS concept, was ~1.3 for both vocabularies. The frequency distribution of expressive variability within each vocabulary follows a Zipf-type curve, with the majority of concepts represented by a single form. Preliminary analysis of the 30 concepts represented by the greatest numbers of forms in each vocabulary indicated that expressive variability tended to be greatest for concepts describing sensory experiences (Severe Pain: *severe pain*, *very painful*, *so much pain*, *terrible pain*) or qualitative observations (Increased: *increase*, *rise*, *off the charts*, *went up*).

"Consensus form" is defined as any form representing a concept preferred by members of a discourse group (similar to preferred terms). Among the concepts reviewed, only a few forms with greater than 50% representation by token were observed (e.g., *diagnosis*, *treatment*, *side effect*, and *health*).

Vocabulary Overlap Characterization. For CMV→PMV and MMV→PMV, one-sided conceptual overlap was nearly 81%, with the highest percentage in the subdomains *Anatomy* and *Chemicals and Drugs*. For CMV→MMV and MMV→CMV, concept overlap was nearly 50%, with the highest percentage in the subdomains *Concepts and Ideas* and *Physiology*. The higher overlap with PMV is likely due to its larger number of terms.

Of the CMV concepts present in PMV, nearly 70% had complete or partial form commonality (i.e., at least one form in common). Overall form commonality is approximately 75% for MMV→PMV and 82% for CMV→MMV and MMV→CMV. For all four pair-wise comparisons, the subdomains *Chemicals and Drugs* and *Anatomy* showed the highest number of concepts with forms shared between vocabularies.

We explored concept overlap between CMV and MMV for the 30 most frequent concepts in each vocabulary. Of these concepts, 14 were shared, for a one-sided overlap of 47% in either direction. Shared UMLS concepts[5] include "Physicians", "Pharmaceutical Preparations", and "Pain"; non-shared frequent concepts include "Problem, NOS" and "Test, Diagnostic" for CMV and "Human Females", "Risk", and "Hospitals" for MMV. The frequent concepts that showed high expressive variability in CMV were "Diagnosis" (19 forms; consensus form: *diagnosis*), "Therapeutic Procedure" (15 forms; consensus form: *treatment*), and "adverse effects" (12 forms; consensus form: *side effect*).

[5] UMLS concepts are represented by their preferred names.

DISCUSSION

These findings are consistent with recent publications comparing consumer and technical terms:

- overlap of consumer and technical terms, more so at the concept level than at the form level
- many form-level mismatches, such as:
  - word-formation problems such as spelling, truncation, and abbreviations/acronyms
  - general language expressions: definitions/ descriptions, colloquialisms, and slang
  - semantic relations, including specific for genetic (hypernymy), generic for specific (hyponymy), part for whole (meronymy), and specification by exemplar (e.g., Tylenol, representing over-the-counter analgesics)
- few concept-level mismatches, such as:
  - Notions outside the framework of allopathic medicine, such as concepts in complementary and alternative medicine
- form- and concept-level mismatches of legitimate medical terms not available in the UMLS

Hence, the evidence to date points to multiple categories of consumer expressions relative to technical terms: form, concept, and term in this study; lexical, semantic, and other in Zeng et al. [7]. Note that these results apply to a particular discourse group: people who had online access and used the Internet for personal health information seeking. These findings may well be different for other groups.

Implications. The ultimate purpose of this work is to support the design of systems that mitigate the language barrier between the health consumer and professional medical domains. Knowing the forms used by laypersons and how such forms map to medical concepts supports assistance to health consumers (1) in query formulation and (2) in understanding medical documents retrieved. Knowledge of how consumers express themselves about health-related topics will also help professionals and information systems interpret patient and lay utterances, as, for example, during patient interviews and interpreting lay responses on health surveys.

There are many ways in which the results of studies on health consumer language can be used, such as:

- using "consensus forms" as category names for browse hierarchies and as suggested words for health text authoring systems for lay audiences
- creating a consumer-oriented entry vocabulary for professional medical vocabularies to map or expand query terms

- identifying and linking professional or lay medical terms in text to definitions or authoritative resources for consumers automatically, either through pre-computed or on-the-fly mechanisms

However, before such systems could be developed, obstacles need to be overcome, including:

- rapidly changing uses and variability in general language expressions, culturally and temporally
- variable length of lay health expressions: short and often cryptic or nonstandard expressions and long descriptive phrases
- reliance on local and personal context for meaning, contrary to terminology where, ideally, terms are unambiguous in meaning within a domain
- imperfect and nebulous lay understanding of medical concepts (e.g., requires field studies and ethnographic research [1])

These observations suggest that lay health expressions, situated "midway" between the lexicology–terminology spectrum as postulated by García de Quesada's unified theory [12], will require additional research in several areas, such as:

- maintaining the currency and accuracy of consumer expressions and their associated concepts
- sorting, parsing, and understanding variations in long definitional phrases, using lexical and natural language processing techniques
- disambiguating homonymous forms, either through form context or direct interaction with users
- detecting lay users' conceptualizations of medical terms and relationships that may benefit from "just in time" explanations or other educational interventions

CONCLUSIONS

While we believe that useful improvements in consumer health-oriented systems may be made now, many challenges continue to limit the role of computational systems as mediators of lay health expressions and professional medical terminology. Thus, research to understand lay use of language in communicating health concepts remains to be done and its results will likely help to improve consumer health information systems. In particular it would be helpful to establish a collaborative framework in which much of the knowledge on the consumer medical vocabulary gained in multiple studies could be pooled and integrated to form a rich knowledge base for extensive user support.

Until a framework for consumer health vocabulary research is developed, we propose a three-pronged approach:

- implementing systems that use existing resources such as consumer "synonyms" (e.g., in the UMLS);
- researching cognitive models and information-seeking behavior of consumers; and
- addressing questions about consumer health expressions, such as those raised in this paper.

Building truly consumer-oriented systems is a huge challenge and a new frontier for medical informatics. We hope that in this paper we have made a small contribution towards meeting this challenge.

REFERENCES

1 Stavri, PZ. Personal health information-seeking: a qualitative review of the literature. *MEDINFO*. 2001;10:1484-8.

2 Patrick TB, Monga HK, Sievert MC, Hall JH, Longo DR. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *J Med Internet Res*. 2001;3(3):e24.

3 Marshall PD. Bridging the terminology gap between health care professionals and patients with the Consumer Health Terminology (CHT). *Proc AMIA Symp*. 2000;1082.

4 Nath R. *Consumer Health Vocabularies*. Meeting of the Workgroup on National Health Information Infrastructure. Chicago: National Committee on Vital and Health Statistics (NCVHS), 2002 July 24.

5 Lewis D, Brennan PF, McCray AT, Tuttle M, Bachman J. If we build it, they will come: standardized consumer vocabularies. *MEDINFO*. 2001;10:1530.

6 McCray AT, Loane RF, Browne AC, Bangalore AK. Terminology issues in user access to Web-based medical information. *Proc AMIA Symp*. 1999;107-11.

7 Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. *Meth Inf Med*. 2002;41(4):289-98.

8 Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: how different are they? *MEDINFO*. 2001;10:399-403.

9 Smith CA, Stavri PZ, Chapman WW. In their own words? A terminological analysis of e-mail to a cancer information service. *Proc AMIA Symp*. 2002;697-701.

10 Tse AY. *Identifying and characterizing a "Consumer Medical Vocabulary."* 2003. Doctoral Dissertation. College Park (MD): College of Information Studies, University of Maryland. 261 pp.

11 Tse T, Soergel D. Procedures for Mapping Vocabularies from Non-Professional Discourse. A Case Study: "Consumer Medical Vocabulary". *Proc. ASIST Annual Meeting*. 2003;*in press*.

12 García de Quesada M. *Estructura definicional terminográfica en al subdominio de la oncología clínica*. Estudios de Lingüística. 2001. Doctoral Dissertation. Spain: University of Granada. Available from: URL: http://elies.rediris.es/elies14/. [Machine translated by SPANAM® from the Pan-American Health Organization (PAHO).]