Research Paper ■

Consumer Health Concepts That Do Not Map to the UMLS: Where Do They Fit?

ALLA KESELMAN, PHD, MA, CATHERINE ARNOTT SMITH, PHD, GUY DIVITA, MS, HYEONEUI KIM, PHD, ALLEN C. BROWNE, MA, GONDY LEROY, PHD, QING ZENG-TREITLER, PHD

Abstract **Objective:** This study has two objectives: first, to identify and characterize consumer health terms not found in the Unified Medical Language System (UMLS) Metathesaurus (2007 AB); second, to describe the procedure for creating new concepts in the process of building a consumer health vocabulary. How do the unmapped consumer health concepts relate to the existing UMLS concepts? What is the place of these new concepts in professional medical discourse?

Design: The consumer health terms were extracted from two large corpora derived in the process of *Open Access Collaboratory Consumer Health Vocabulary* (*OAC CHV*) building. Terms that could not be mapped to existing UMLS concepts via machine and manual methods prompted creation of new concepts, which were then ascribed semantic types, related to existing UMLS concepts, and coded according to specified criteria.

Results: This approach identified 64 unmapped concepts, 17 of which were labeled as uniquely "lay" and not feasible for inclusion in professional health terminologies. The remaining terms constituted potential candidates for inclusion in professional vocabularies, or could be constructed by post-coordinating existing UMLS terms. The relationship between new and existing concepts differed depending on the corpora from which they were extracted.

Conclusion: Non-mapping concepts constitute a small proportion of consumer health terms, but a proportion that is likely to affect the process of consumer health vocabulary building. We have identified a novel approach for identifying such concepts.

■ **J Am Med Inform Assoc.** 2008;15:496–505. DOI 10.1197/jamia.M2599.

Introduction

Researchers increasingly speak to the need to reduce the discrepancy between the language of health consumers and health professionals; one means of bridging the gap is through the development of controlled consumer health vocabularies. Vocabulary development process typically involves identifying terms used by consumers and "translating" them into the language of health professionals by mapping consumer terms to their equivalents contained in professional controlled vocabularies (e.g., the Unified Medical Language System (UMLS) Metathesaurus).[1] The translation effort relies on an assumption that professional and consumer terms map to the same underlying concepts: for example, that a physician's *epistaxis* is a layperson's *nosebleed*. Accordingly, most research on consumer health vocabulary has focused primarily on consumer health terms, rather than the lay health concepts that underlie those terms, with the exception of Zeng and Tse.[2] However, the difference between the lay and professional knowledge base of health and disease is likely to extend beyond simple term labels, into the underlying concepts that are the basis for these terms.

Two general questions arise in considering the relationship between terms and concepts. First, to what extent are the health terms used by laypeople a reflection of a different set of concepts from those of professionals? Second, how should consumer health vocabulary developers handle the process of term mapping in the face of these potentially different conceptual models? This paper considers these issues in the context of the *Open Access and Collaborative Consumer Health Vocabulary (OAC CHV)* development project, presenting an analysis of OAC CHV terms that could not be mapped to UMLS concepts.

Affiliations of the authors: Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health (AK, GD, ACB), Bethesda, MD; Aquilent, Inc. (AK), Laurel, MD; School of Library and Information Studies, University of Wisconsin (CAS), Madison, WI; Lockheed Martin, Inc. (GD), Bethesda, MD; Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School (HK, QZ-T), Boston, MA; School of Information Systems and Technology, Claremont Graduate University (GF), Claremont, CA.

This project was supported the National Institutes of Health (NIH) grant R01 LM07222, the Intramural Research Program of the NIH, NLM and 2003 Donald A.B. Lindberg Research Fellowship sponsored by the Medical Library Association. The authors thank the National Library of Medicine (NLM) for sharing the MedlinePlus® query log data. The authors thank Sergey Goryachev for his technical help with the project.

Correspondence: Alla Keselman, PhD, MA, 7S713E, LHNCBC, National Library of Medicine, NIH, 8600 Rockville Pike, Bethesda, MD; e-mail: <keselmana@mail.nih.gov>.

Received for review: 08/20/07; accepted for publication: 02/08/08

Background

Why Do We Need Consumer Health Vocabularies?

Health consumers are increasingly expected to act as partners in their healthcare. This partnership requires that consumers communicate with health professionals about various treatment options; locate and comprehend information in various Internet and printed sources; and participate in making decisions. The gap between lay and professional health terminologies has been long identified as one of the significant barriers to empowerment of healthcare consumers. Studies suggest that lay people have difficulty understanding medical jargon,[3] and this affects their ability to search health-related websites,[4] comprehend printed materials, and communicate with their physicians.

The medical informatics approach to solving the vocabulary problem involves building structured vocabularies of consumer health terms and mapping them to professional medical vocabularies.[1,2] The process involves some challenges, the first and foremost being defining a consumer health vocabulary. Consumer health language lacks the stability of professional medical language. Consumer health vocabularies have greater variability and are more strongly affected by regional variations.[5] Consumer health language is affected by many factors, including the individual's geographic region, level of education, and personal experience with health and illness. Smith[5] cautions researchers that "the notion of a paradigmatic "consumer" who uses a particular vocabulary specific to her "consumer" status may be ill-founded." However, the complexity of the problem does not diminish the need to provide a bridge helping lay individuals communicate with health professionals and read health materials.

Zeng and Tse[2] define consumer health vocabularies as a collection of common health expressions, concepts, explanatory models, attitudes and beliefs "shared by most members of a consumer discourse group." While consumer health vocabularies can not perfectly reflect personal health constructs of every individual, they serve as an approximation of the world of consumer health language and understanding. They also provide a practical solution to the very real problem of a terminology gap between consumer and professional health discourse.

Some Recent Consumer Health Vocabulary Development Efforts

Several commercial groups work on building consumer health vocabularies with the potential to facilitate online health information seeking, indexing, and translation. Examples include Health Terminology (PHT) by Intelligent Medical Objects (Northbrook, IL; http://www2.e-imo.com), which maps the most common ICD-9 codes to consumer-friendly synonyms, and Apelon's terminology system of common consumer friendly terms (see Zielstorff[1] for review). In this work, we draw upon *Open Access and Collaborative Consumer Health Vocabulary (OAC CHV),* a consumer health terminology developed as a joint effort by several academic groups, with which the authors of this paper are affiliated.[6,7] At the present time this is the only non-commercial, open-access consumer health vocabulary in existence. It consists of actual terms commonly used by consumers. The goal of the OAC CHV initiative is to add a comprehensive source of consumer-used and consumer-preferred health-related words and phrases to the UMLS Metathesaurus.

Terms for the OAC vocabulary were identified on the basis of strings derived from two data sets, referred to as Set A and Set B. Set A consisted of 12 million queries, extracted from query logs of MedlinePlus® consumer health site (National Library of Medicine, Bethesda, MD; www.medlineplus.gov), reporting search strings submitted by users from October 2002 to September 2003. The queries were tokenized into 28,797,199 non-unique tokens (words) and 4,928,158 unique n-grams (1 to 7 tokens in length). Set B consisted of 23,657 unique health-related words and phrases manually extracted by experienced indexers from consumers' written utterances. The utterances were collected from consumer postings to more than 25 health-focused Web-based bulletin boards from October 2003 to November 2004. The boards were highly trafficked, frequently pointed to by health advice websites found through major search engines. For both data sets, the process of entering terms into the OAC Consumer Health Vocabulary consisted of similar steps (Figure 1).

In the first step, automated tools—HITEx[8] for Set A, MetaMap[9,10] for Set B—were used to map spelling-corrected terms or n-grams to UMLS Metathesaurus (2007 AB). In the second step, subsets of high-frequency unmapped 753 Set A n-grams and 293 Set B terms were manually mapped to Metathesaurus (2007AB) through collaborative review by the authors.* New concepts were created to represent the 44 terms from Set A and 20 terms from Set B that could not be manually mapped. The process of concept creation and the characteristics of these concepts are the focus of this paper.

Lay Mental Models of Health and Disease and Their Potential Impact on Consumer Health Concepts

Studies of lay understanding of health and disease demonstrate many instances of lay individuals' lack of knowledge and non-normative understanding of health issues. Unlike health professionals, lay individuals have minimal formal education in health matters, which translates into knowledge gaps and occasional misconceptions. For example, McGregor[11] interviewed patients with localized prostate cancer about their understanding of their disease. While participants differed greatly in their professional and educational backgrounds, most had little understanding of the function of the prostate gland and the side effects of the possible treatments.

Lay individuals often compensate for their lack of biomedical knowledge by drawing upon cultural, social, and experiential knowledge.[12] While the knowledge networks derived from these sources are often incompatible with the biomedical perspective, they make logical sense given the beliefs of their holders. However, most studies of lay health reasoning focused on mental models involved understanding of complex processes, such as childhood malnutrition and the mechanism of HIV infection.[13,14] For example, Keselman et al. con-

*In the case of Set A, the process required a preliminary step of extracting a larger pool of high-frequency n-grams and manually reviewing them for "termhood", since many high-frequency machine-extracted n-grams were not deemed true health terms (e.g., "treated with").

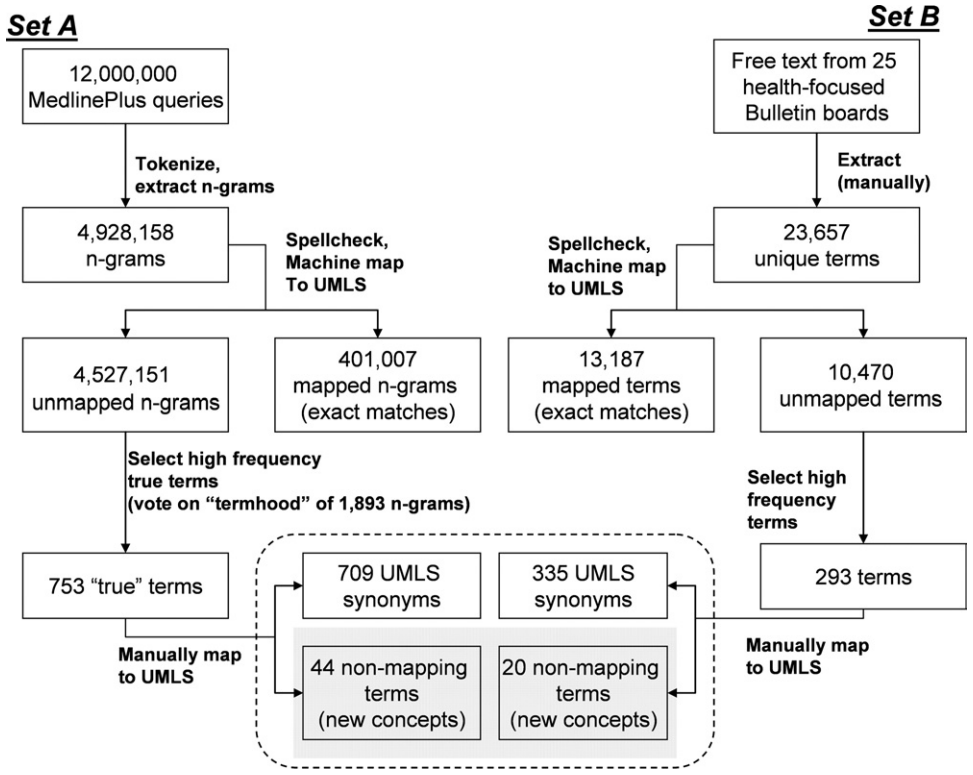

Figure 1. The process of OAC CHV development, leading to non-mapping terms identification.

ducted a study of adolescents' understanding of HIV and reasoning about HIV-related issues.[14] Many adolescents lacked understanding of the concept *virus*, and produced explanations that linked HIV infection to some "tangible" event, rather than to hidden biological processes (e.g., poor hygiene after sex). Non-normative lay mental models of health and disease suggest that the mismatch between lay and professional health language may partly reflect this mismatch in understanding. However, studies of conceptual understanding typically involve analysis at the level of mental models and mechanisms involving multiple concepts. Thus it is difficult to predict how and whether such misconceptions might affect consumer health vocabulary development.

Consumer Terms, Underlying Concepts and the Process of Vocabulary Development

The process of consumer health vocabulary development usually starts with mining consumer health resources and their logs for health terms that are commonly used by their lay users. This step typically yields lexical strings (terms), rather than concepts with definitions.[8] The next step involves reviewing each term extracted from a consumer source, with the goal of mapping it to a term-concept pair from a professional health terminology.

The challenge here is to reconstruct the meaning (concept) inherent in the lay usage of a term, and then to agree that consonance between lay and professional terms exists on the basis of this deeper meaning, rather than the lexical form. Thus we can envision that consumer term/concept pairs and professional term/concept pairs may have one of four possible relationships to each other.

Case 1 is an ***exact match*** between the pairs; this occurs when the term used by a lay person can be found in a professional terminology, and both terms correspond to the same concept (Figure 2). For example, the term "pain" used by a health consumer would map to a UMLS "pain" term, and both terms will be rooted in the same concept (UMLS Concept

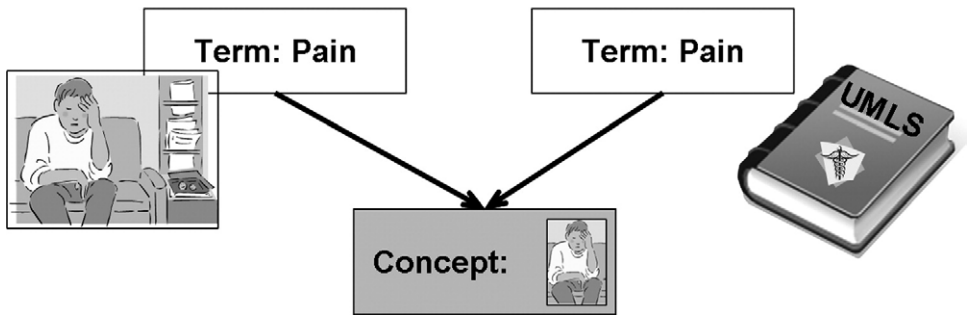

Figure 2. Exact match – Lay term is found in a professional terminology.

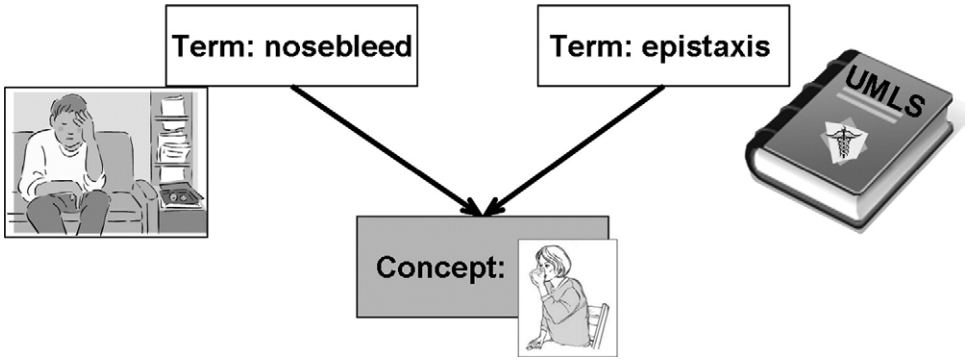

Figure 3. Lay synonym – Lay term corresponds to a professional term for the same concept.

Unique Identifier, or **CUI:** C0030193). While we call this kind of term-concept correspondence "exact", the reality of this mapping category entails many cases of what could be "near matches." Lay understanding of the majority of health concepts is less sophisticated than the understanding of the same concepts by healthcare professionals. For example, the lay concept for the term "heart" does not necessarily involve two atriums and two ventricles. However, for practical vocabulary building purposes, the concepts are close enough to warrant mapping. Many of the terms used by health consumers fall into the exact match category, as evidenced by the high proportion of terms collected by consumer health vocabulary initiatives that can be machine-mapped to professional vocabularies.[1]

Case 2 involves a ***lay synonym***. This occurs when the term used by a lay person does not exist in the controlled professional vocabulary, but corresponds to a professional term that denotes the same (or closely related) concept (Figure 3).[15] For example, "nosebleed" corresponds to "epistaxis" (UMLS **CUI:** C0014591). Mapping in this case involves finding such corresponding UMLS synonym.

Case 3 occurs when a term is used ***differently*** in lay communication and professional medical terminology (Figure 4). In this situation, while the lexical term string is the same, the concepts are different, and this difference inheres in more than simple conceptual unsophistication. An example is the term "leg". In everyday usage, this term string denotes the lower extremity that includes the hip. However, the UMLS Metathesaurus defines *leg* as "the inferior part of the lower extremity between the KNEE and the ANKLE" (UMLS **CUI:** C1140621). Lay term "leg" does not map onto the UMLS "leg", although the lexical strings are identical. It can, however, map to the UMLS concept (UMLS **CUI:** C1269079) "entire lower limb."

Case 4 comprises those concepts that ***cannot be mapped***, either through automated or manual methods, to the professional vocabulary. These can be legitimate health terms,

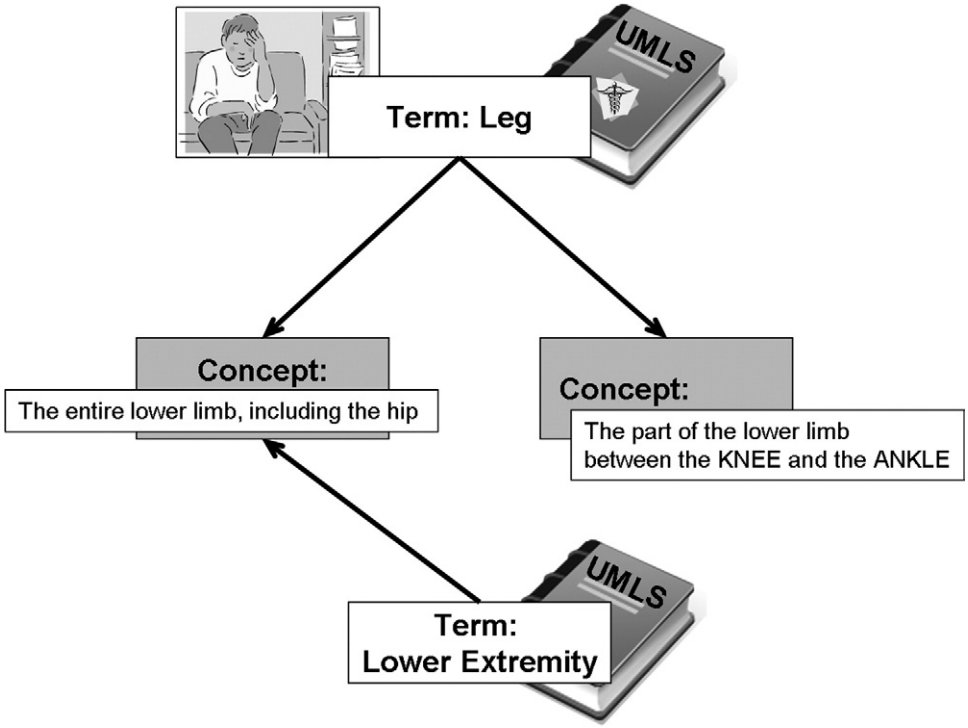

Figure 4. Lay usage – Lay and professional usage of a term map to different concepts.

the omission of which reflects real gaps in existing professional controlled vocabularies; or they can represent unique concepts reflecting lay models of health and disease. In either case, these need to be entered into the UMLS as new concepts. The challenge for vocabulary developers lies in assigning these new concepts semantic types and linking them to some existing concept(s) via semantic relations.

A number of studies and initiatives report statistics of their attempts to map consumer health terms to UMLS. For example, Apelon collected approximately 15,000 consumer health terms from a variety of sources, including consumer-oriented medical websites and NLM's MEDLINE searches by consumers.[1] Of these, 14,000 were algorithmically or manually mapped to concepts in the UMLS Metathesaurus, while 1,000 (10%) could not be mapped. In a different study, Smith and colleagues extracted 504 consumer health terms pertaining to features and findings from 139 e-mail patients' messages to a cancer information service.[5] These terms were mapped against the 2001 UMLS Metathesaurus. The authors found that that 185 (36%) of the terms were exact matches of the UMLS terms/concepts, 179 (35%) were partial string matches, and 119 (24%) were known synonyms for UMLS concepts. Brennan and Aronson[16] applied MetaMap[10] program to map health terms from 241 patients' e-mail to HeartCare intervention nurses to a set of UMLS vocabularies (Nursing Vocabularies, MeSH and SNOMED). The program identified 7,366 candidate concepts, only 5,078 (69%) were actually mapped to the formal vocabularies. The studies described above used different sources of consumer-produced health utterances and different extraction and mapping methods, so it is not surprising that their findings about non-mapping concepts vary significantly. These studies support the proposition that some consumer health concepts will not map to existing vocabularies. However, these studies do not explore *why* some concepts do not map, and what proportion of the unmapped terms represent gaps in knowledge representation on the part of professional terminologies, as opposed to truly novel and genuine lay concepts.

Specific Research Questions

The specific goals of this study were as follows:

1. Analysis of Set A and Set B terms that could not be manually mapped to the UMLS with respect to the nature and domain of their underlying concepts. The authors were particularly interested in making an important distinction between terms that represented legitimate medical concepts but which are not presently included in the UMLS, and those terms that reflected uniquely lay models of health and disease.
2. Characterizing of relationships between these new concepts and existing UMLS concepts, and understanding the implications of these relationships for consumer health vocabulary development;
3. Comparison of the characteristics of unmapped concepts in the two datasets.

Methods

A group of three or more raters conducted collaborative term reviews of the new concepts, with the goal of categorizing each term along the following five dimensions:

1. *Nearest related UMLS concept*. Introduction of new concepts into the UMLS Metathesaurus is only meaningful if these new entries can be semantically related to the existing ones. In this phase, once a new consumer concept had been identified, we sought existing UMLS concepts overlapping with the new one. The coding scheme distinguished between the following degrees of overlap:
   - ***Narrower-than*** an existing concept. This relationship was assigned if the new consumer concept could be represented as a child of an existing UMLS concept. For example, the new concept *Diet Pills* is narrower than the existing concept, *Weight-Loss Agents*, C0376606.
   - ***Broader-than*** an existing concept. This relationship was assigned if the new concept could be represented as a parent of an existing UMLS concept. For example, the new concept *Pelvic Area* is broader than the existing concept, *Pelvis*, C0030797. In the case where the new concept could be introduced at the intermediate hierarchical level, so that both parent and child concepts could be identified, it was related via this *Broader-than* relationship to the most specific node of the branch.
   - ***Vaguely related*** to two or more existing (or new) concepts. This relationship was assigned to new concepts which could potentially be used by consumers to refer to multiple related concepts. An example is the consumer concept of *Eye Genes*, which could be mapped to both the *EYCL1 gene* responsible for green/blue eye color expression (C1414494) and to the *EYCL3 gene* responsible for brown eye color expression (C1414495). Since an average lay user is unlikely to understand and imply the distinction, consumer *Eye Genes* was coded as vaguely related to C1414494 and C1414495. The case of *vaguely related concepts* should not be confused with the situation of *ambiguous terms*, when a term can be mapped to more than one concept, and the consumer can make the distinction between the synonyms.
   - ***Other*** relationship applies to a new consumer concept having a non-hierarchical relationship to an existing concept. For example, the new concept *Hairline* is non-hierarchically related to the existing concept *Hair*, C0018494.
2. *Semantic type*. New concepts having hierarchical relationships with existing UMLS concepts were assigned the semantic type of the parent or the child concepts. In the cases of concepts that could not be assigned UMLS parents or children (e.g., *Chakras*), the type was established via discussion and consensus, on the basis of existing UMLS semantic types.
3. *Domain*. If the new consumer concept unambiguously belonged to some medical specialty/domain, that domain was noted. For example, the concept *Bladder Cancer Treatment* was assigned to *Oncology*. Categorization by clinical domain, unlike semantic type, is not part of standard UMLS classification, and domain codes were developed specifically for this project, on the basis of this project's data sets. This feature was developed in order to facilitate identification of those health domains showing the greatest difference between professional and lay words.
4. *Lay nature*. New consumer concepts that did not have professional medical equivalents outside of UMLS were denoted as "lay". For coding purposes, lay concepts were

defined as concepts related to terms without synonyms in professional medical discourse, and which could not be constructed by post-coordinating professional medical terms. Examples of lay terms include *Cure* and *Beauty Marks*.

5. *Postcoordination*. Concepts were denoted as postcoordinated, if they could be fully derived from existing concepts via conjunction or modification (e.g., Blood Pressure Medicine, Bone Cancer Treatment).

Results

Overview

For both sets, the proportion of terms that could not be manually mapped and which resulted in new concepts was small. Table 1 presents the summary of the 64 unmapped concepts' semantic relationships with the existing UMLS terms; their "lay" nature; and the possibility of constructing them by post-coordinating the existing terms.

Semantic Relationships with Existing Concepts

Set A

In Set A, the majority of new concepts (31 of 44) could be represented as children of some existing UMLS concepts via the narrower-than relationship. Of these 31, 25 could be constructed by post-coordinating existing UMLS concepts, usually via modification. Examples include *White Bumps* and *Red Bumps*, created as children of (or narrower-than) *C0577559, Mass of Body Structure;* and *Bladder Cancer Treatment* and *Bone Cancer Treatment*, created as children of *C0920425, Cancer Treatment*. Table 2 presents the distribution of semantic types, assigned to new narrower-than concepts derived from both sets. The two most common semantic types for narrower-than concepts in Set A included Therapeutic or Preventative Procedures, assigned to 10 concepts (e.g., *Bladder Cancer Treatment*); and Findings or Signs or Symptoms, assigned to 6 concepts (e.g., *White Bumps, Red Bumps, Cancer Symptoms, Sudden Weight Loss*).

The second most frequent relationship to nearest existing UMLS concepts in Set A was the "other" (11 concepts) (see Table 3). None of these could be produced via post-coordinating existing concepts. The distribution of semantic types of these concepts was much more varied than for *narrower-than* concepts.

Table 1 ■ Summary of Unmapped Concepts' Relationships with the Existing UMLS Concepts

	Set A (total = 44)	Set B (total = 20)	Total in OAC CHV (=64)
Semantic relationship*			
Narrower-than existing	31	5	36
Broader-than existing	1	2	3
Vaguely related to two or more	1	1	2
Other (non-hierarchical) relation	11	12	23
Lay terms†	6	11	17
Post-coordinated terms†	25	5	30

*The sum of the four semantic codes for each set equals the total number of concepts in that set (44 for Set A; 20 for Set B), since semantic relationship codes are non-overlapping and each concept was assigned a semantic relationship code.
†Lay terms and post-coordinated terms constitute a subset of total terms in Set A and Set B.

Table 2 ■ Semantic Types of "Narrower Than" Concepts

Semantic Type	Concept
Therapeutic or preventive procedures	Set A (N = 10 of 10*): dash diet, bladder cancer treatment, bone cancer treatment, cervical cancer treatment, cervical cancer treatment, colon cancer treatment, esophageal cancer treatment, gastric cancer treatment, skin cancer treatment, testicular cancer treatment, thyroid cancer treatment
Finding, signs or symptoms	Set A (N = 7 of 7): white bumps, red bumps, vaginal flora, sudden weight loss, red spots, white spot, cancer symptoms
Intellectual product	Set A (N = 7 of 7): growth chart, medical terminology, weight chart, research articles, nursing journals
Medical device	Set A (N = 3 of 3): growth chart, weight chart, coffin
Clinical drug or pharm. substance	Set A (N = 3 of 3): Blood pressure medicine, diet pills, leptoprin
Bodyloc/part, organ or organ component	Set B (N = 3 of 3): Left side of breast; foreskin stump; bangs
Disease or syndrome	Set A (N = 1 of 1): childhood obesity
Quantitative concept	Set A (N = 1 of 1) polycystic
Natural phenom. or process	Set A (N = 1 of 1): coffin birth
Neoplastic process	Set B (N = 1 of 1): Beauty marks
Daily or rec. activities	Set B (N = 1 of 1): Lateral LE's

*The second number is the total across both sets.

Table 3 ■ Semantic Types of "Other" Concepts

Semantic Type	Concepts
Bodyloc/part, organ or organ component	Set A (N = 1 of 5*): lap
	Set B (N = 4 of 5): privates; M-spot; G-spot; hairline
Intellectual product	Set A (N = 3 of 3): food pyramid, tutorials, illustration
Organization	Set A (N = 2 of 2): Mylan, Mayo Clinic
Organism function	Set B (n = 2 of 2): preejaculatory penile secretion; hormonal balance
Finding	Set A (N = 1 of 2): water in ear
	Set B (N = 1 of 2): brown eyes
Therapeutic or preventive procedures	Set A (N = 1 of 1): cure
Disease or syndrome	Set A (N = 1 of 1): leaky gut
Quantitative concept	Set A (N = 1 of 1): nursing shortage
Qualitative concept	Set A (N = 1 of 1): easy-to-read
Anatomical abnormality	Set B (N = 1 of 1 of 1): smegma pearl
Body substance	Set B (N = 1 of 1): preejaculatory fluid
Lab or test result	Set B (N = 1 of 1): endorphin levels
Temporal concept	Set B (N = 1 of 1): manhood
Unassigned	Set B (N = 1 of 1): chakras

*The second number is the total across both sets.

Table 4 ■ Content Domains of New Concepts

Domain	Concepts
Gynecology/sexual health	Set A (N = 2 of 11*): vaginal flora, vaginal bacteria
	Set B (N = 9 of 11): privates, vaginal area, foreskin stump, M-spot, preejaculatory fluid, smegma pearl, preejaculatory penile secretion, G-spot, manhood
Oncology	Set A (N = 10 of 10): cancer symptoms, bladder cancer treatment, bone cancer treatment, cervical cancer treatment, colon cancer treatment, esophageal cancer treatment, gastric cancer treatment, skin cancer treatment, testicular cancer treatment, thyroid cancer treatment
Wellness; fitness; nutrition	Set A (N = 5 of 6): xenadrin, leptoprin, childhood obesity, diet pills, food pyramid
	Set B (N = 1 of 6): lateral LE's
Beauty	Set B (N = 3 of 3): beauty marks, bangs, hairline
Hypertension	Set A (N = 2 of 2): blood pressure medicine, dash diet
Alternative medicine	Set A (N = 1 of 2): leaky gut
	Set A (N = 1 of 2): chakras
Pathology	Set A (N = 1 of 1): coffin birth
Genetics	Set B (N = 1 of 1): eye genes
Meta	Set A (N = 5 of 5): Medical terminology, research articles, nursing journals, tutorials, illustrations

*The second number is the total across both sets.

Only one concept in this set was broader than its closest UMLS relative (*Xenadrine*, broader than C1572218, *XENADRINE EFX EPHEDRINE FREE FAT LOSS CAP/TAB*). Similarly, only one concept was vaguely related to two others (*Vaginal Bacteria*, vaguely realated to C0085166, *Bacterial Vaginosis* and *Vaginal Flora, itself a new concept*).

Set B

In this dataset, the most frequent relationship between new and existing UMLS concepts was non-hierarchical *Other* −12 out of 20− (see Table 3 for semantic types). This set also contained five new concepts related to existing concepts via narrower-than relationship (Table 2); three concepts related via broader-than relationship; and one vaguely related to more than one concept.

Content Domains of New Concepts

Set A

In this set, 21 out of 44 concepts could be assigned to a specific health domain, and five additional concepts could be labeled as meta-concepts, characterizing desired information rather than specifying its content (e.g., *Research Articles*) (see Table 4). The most commonly assigned content domains in this set were oncology (10 concepts) and wellness and nutrition (5 concepts). All oncology concepts could be creating by post-coordinating existing UMLS concepts. Nine referred to treatments for a specific type of cancer (e.g., *Colon Cancer Treatment*, *Skin Cancer Treatment*) and did not contain lay undertones; while one, *Cancer Symptoms*, did not have a professional equivalent.

Set B:

In this data set, 15 out of 20 concepts could be assigned to some specific content domain, with the most common ones being sexual health (8) and beauty (4).

Lay Concepts

Most new concepts created from both data sets were concepts that could be legitimately used by health professionals. Across both sets, 17 concepts were classified as primarily lay (11 in Set B and 6 in the Set A). Since the number of lay concepts is small, we chose not to separate them into Sets A and B for this analysis. Table 5 presents the distribution of these concepts within different semantic types across both sets. The table also lists the UMLS "relatives" of these new concepts and the type of relationship by which they are connected.

In many cases, the definition of lay concepts is obvious from the terms that denote them (e.g., *Cure*). Other cases require some explanation.

- *M-Spot* is claimed by some to be an area which is especially sensitive to sexual stimulation (similar to the more commonly known *G-Spot*). As this area is believed to be located on the skin surface in the waist area, we related the concept to *Sexuality C0036915*, rather than to a specific organ or structure.
- *Vaginal Bacteria* encompasses *Bacterial Vaginosis C0085166* and another new concept, *Vaginal Flora.* Overall, the context of consumer usage of health terms suggests frequent lack of clear distinction between an illness and a microorganism that causes it.
- *Manhood* is an ambiguous term mapping to two existing UMLS concepts, *Masculinity C0042757* and *Penis, C0030851*, as well as denoting the new concept, which can be defined as the period of sexual potency in a male's life.
- *Coffin Birth* is a postmortem delivery of a fetus from the decomposing uterus of the mother, due to buildup of gases.

Six of the 17 lay concepts identified in our data sets could be assigned to the domain of sexual health/gynecology: *vaginal bacteria*, *privates*, *M-spot*, *G-spot* and *manhood*. The next most represented domain was beauty (*hairline*, *bangs* and *beauty marks*). The following domains were represented by one concept each: wellness (*diet pills*), oncology (*cancer symptoms*), pathology (*coffin birth*), and genetics (*eye genes*). Finally, the following four concepts could not be assigned to ant specific domain: *cure*, *lap*, *pelvic* area and *brown eyes*.

Discussion

Building consumer health vocabularies by mapping to standardized medical vocabularies requires an approach for dealing with consumer terms that refer to concepts not yet represented in those vocabularies. The findings of this study suggest that the overlap between the conceptual universes underlying lay health language and professional terminologies is large. Of the 1,046 terms extracted by the OAC CHV development team, only 64 could not be mapped to existing UMLS concepts. Moreover, 47 of these terms denoted concepts that could be present in professional medical discourse. Some of these legitimate concepts could reasonably be expected to appear in future versions of the UMLS (for

Table 5 ■ Lay Concepts — Semantic Types and Relationships to Existing UMLS Concepts

Semantic Type	New Concept	Related Concepts	Relation
Body location or region	M-spot	Sexuality C0036915	O
	G-spot	Vagina C0042232	O
	Vaginal area	Vagina C0042232	B
	Pelvic area	Pelvis C0030797	B
	Lap	NONE	O
	Hairline	Hair C0018494	O
Finding	Brown eyes	Eye color C0015396	O
	Vaginal bacteria	Bacterial vaginosis C0085166; Vaginal flora - NEW	VB
Body part, organ or organ component	Privates	Genitalia C0017420; Breast C0006141; Buttocks C0006497	O
	Bangs	Hair C0018494	N
Gene or genome	Eye genes	EYCL1 gene C1414494; EYCL3 gene C1414495	VB
Temporal concept	Manhood	Male Gender C0024554	O
Neoplastic process	Beauty marks	Nevus C0027960	N
Sign or symptom	Cancer symptoms	Symptoms C1457887	N
Therapeutic or preventative procedure	Cure	NONE	O
Clinical Drug/pharmacological substance	Diet pill	Weight-loss agents C0376606	N
Natural phenomenon or process	Coffin birth	Birth C0005615	N

O = Other; N = Narrow then; B = Broader than; VB = Vague and Broader than.

example, novel drugs and procedures); and most of these legitimate concepts could be constructed by post-coordinating existing UMLS concepts. Only 17 terms referred to concepts that would make a health professional frown or shrug at in puzzlement.

The findings also point to some interesting differences between the concepts derived from the two datasets used in this study. While most concepts derived from the query-based set were narrower (more specific) than their closest UMLS relatives, and oncology was the single most represented domain, most concepts derived from the free text set had non-hierarchical relationships with their UMLS relatives. The most widely represented domain was sexual health, which was not surprising, given the high proportion of sexual health-oriented messages on the bulletin boards.

Implications for Vocabulary Building

These findings suggest that most of the labor in building consumer health vocabularies indeed lies in bridging consumer-preferred terms and physician-preferred terms referring to the same concept—for example, equating *shakes* and *tremors*, *sugar* and *glucose*, *cancer* and *malignant neoplasm*. This should be a cause for optimism, as translation between languages that describe the same realities is a manageable, albeit labor-intensive, task. In addition, almost all new concepts identified in the course of this study were found to be closely related to existing UMLS concepts. Finding such relationships makes the new concepts potentially useful for information retrieval.

The study also provides some pointers to conceptual differences in lay and professional thinking about health and disease, which requires further investigation by vocabulary builders. The prevalence of terms that can be derived by post-coordination supports the findings of other researchers that patients' organize their health knowledge in a way that is different from professionals.[17] From the professional perspective, cancer therapies may be organized according to their mechanism of action (e.g., chemotherapy, immunotherapy), irrespective of the cancer type. From the perspective of the patient, however, information needs are directly connected to a specific diagnosis and its effect on their life course; thus, cancer therapies are more likely to be organized according to the bodily systems affected by the disease (e.g., *Colon Cancer Treatment*, *Cervical Cancer Treatment*). Despite the common-sense vocabulary builders' notion that lay concepts are "fuzzier" than the professional ones, this study reveals that lay concepts are more likely to be "narrower-than" than "broader-than" their closest UMLS relatives. This finding also appears to support the notion that individuals' thinking of health issues is very specific to the details of the individual situation. Understanding patients' information organization is essential in building information portals and supporting information retrieval.

The findings also suggest that in some domains and settings, the number and breadth of semantic coverage of non-mapping concepts may be greater than in the others. One of the goals of this study was comparison of the non-mapping concepts in two sets, one extracted from MedlinePlus® queries and the other extracted from consumers' free text exchanges. The query-based set produced many more concepts that were narrower than their closest relatives and could be constructed by post-coordinating existing concepts. In contrast, the free text set produced many concepts having non-hierarchical relationships with existing UMLS concepts. This suggests that the degree of the lay-professional language overlap in query analysis may be deceptive. When lay individuals communicate in what they perceive as the professional setting, they may adjust their language to that of health professionals.[18] However, when talking with people they perceive as peers, they may use a somewhat different language, the one with which they are more familiar and comfortable. They may also be more likely to operate with concepts that are not part of standard medical worlds. Furthermore, these conceptual differences may be more prominent in some domains than in others. The free text set included many terms from the domains of sexual health and wellness/beauty/physical fitness. These domains generated some concepts that were truly lay, such as the mysterious *M-Spot*. It is desirable for future studies to focus on identifying lay health concepts in the domains where deviations from traditional professional views are

likely to abound (e.g., sexual health, alternative medicine). This study also suggests that uncovering truly lay health concepts is a slow process; innovative methodologies for streamlining the task are desirable.

Implications for Interpreting Lay Models of Health and Disease

Existence of lay concepts that do not overlap with professional medical concepts suggests that patients and consumers may have unique models of health and disease, which differ from those used by professionals. Does the scarcity of such concepts identified in this study suggest that the differences between lay and professional health models are negligible? Studies that investigate lay understanding of specific diseases point to the contrary.[17,19] In the background section of this paper, we proposed four possible relationships between lay and professional term/concept pairs. This study investigated the (relatively uncommon) situation, when lay individuals use terms that ***cannot be mapped*** to the professional vocabulary via automated or manual methods, and require the creation of new concepts. We did not, however, consider the case of ***lay usage*** of professional terms, when a lay individuals use existing professional terms, but ascribe to them meaning that differs from their professional definition; for example, *depression*. This case may be as common as it is difficult to investigate.

One can argue, however, that the usage of almost any health term used by a non-health professional will involve some vagueness or alteration of meaning. For example, as mentioned earlier, when consumers use the term "heart", they are likely to know that it is an organ that pumps the blood through the body, but may not think of it as a "four-chambered organ that receives the blood from the veins and contracts to send it through the arteries." In addition to containing fewer details and having some vagueness, concepts in lay models are likely to differ from the professional ones in their organization and relationship to one another. Understanding these relationships is important for connecting concepts in consumer health vocabularies.

Limitations of the Study and Directions for the Future Research

The main limitation of this study is the lack of the context in which communication was taking place; a context which would help us interpret the full meaning ascribed by individuals to the health terms they used. Set A, the query data set, provided us with isolated search engine queries; Set B, the free-text data set, provided us with more context, but did not allow us to probe the message writers about the terms they used and what they meant. An additional problem with the query data set is the potential tendency of web portal users to imitate what they perceive as the professional medical language, which may have limited the opportunity for discovering unique lay concepts.[18] This limitation is the negative but necessary aspect of our methodological approach, which allowed the extraction of large numbers of consumer health terms from the corpus and establishing use frequency statistic for different terms. Other researchers have conducted patient surveys, analyzed transcripts of doctor-patient interactions, and recorded physicians' recall of patients' words they found difficult to understand.[3,20,21] While these methods provide more context in which to understand the usage of individual words, they are less systematic than those employed in this study, and are additionally more likely to yield regionalisms and extremely rare concepts. However, when examined with caution, the findings of such studies can be used to supplement the list of non-mapping concepts in our consumer health vocabulary.

The analysis of the free text corpus suggests that this may be a fruitful venue for lay concept discovery. As mentioned previously, our analysis suggests that not all content domains may be equally abundant with lay health concepts. Additional studies should identify and explore promising domains. Future work should also focus on other categories of professional/lay concept mismatch, including cases where lay individuals ascribe unique meaning to professionally sounding terms. Finally, further studies are needed to characterize the difference between consumer knowledge of health of terms and their understanding of the underlying concepts (e.g., Keselman et al.[22]). The field of consumer health vocabulary development is relatively new. As it matures and vocabularies grow, the aspect of creating new, well defined uniquely lay concepts will become more prominent, and the quality of procedures for defining such concepts and relating them to the existing ones will affect the quality and usefulness of the vocabularies.

Conclusions

Non-mapping terms represent a small, but non-negligible, proportion of health terms used by lay people. Consumer health vocabulary development requires a standardized approach for creating concepts that denote these terms, assigning their semantic types and relating them to existing concepts in professional medical vocabularies. Most non-mapping terms refer to concepts that are present in professional medical discourse, and are either relatively novel (not yet included in the professional vocabularies) or can be represented via post-coordinating professional vocabulary concepts. The task of introducing these concepts into consumer health vocabularies is relatively straightforward. Other consumer concepts, however, are uniquely lay and reflect the difference between lay and professional understanding of health and disease. While introducing these concepts into consumer health vocabularies represents a greater challenge, these concepts are also a valuable resource for understanding the structure of lay conceptual knowledge in health. Future studies should focus on extracting lay health concepts from different kinds of consumer health discourse and on understanding the relationship of these concepts to one another and to existing professional concepts.

References ■

1. Zielstorff RD. Controlled vocabularies for consumer health. J Biomed Inform. 2003 Aug-Oct; 36(4–5):326–33.
2. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc. 2006 Jan-Feb; 13(1):24–9.
3. Chapman K, Abraham C, Jenkins V, Fallowfield L. Lay understanding of terms used in cancer consultations. Psychooncology. 2003 Sep; 12(6):557–66.
4. Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: how different are they? Medinfo. 2001;10(Pt 1):399–403.
5. 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.

6. Open Source, Collaborative Consumer Health Vocabulary Initiative. Open Source, C; Available at: http://www.consumerhealthvocab.org/. Accessed May 22, 2008.
7. Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC, editors. Exploring lexical forms: first-generation consumer health vocabularies. AMIA Annu Symp 2006.
8. Zeng Q, Tse T, Divita G, Keselman A, Crowell J, Browne A, et al. Term Identification Methods for Consumer Health Vocabulary Development. Journal of Medical Internet Research. 2007;9(1):e4.
9. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001. 17–21.
10. MetaMap program. Available at: http://MMTx.nlm.nih.gov. Accessed May 22, 2008.
11. McGregor S. What information patients with localised prostate cancer hear and understand. Patient Educ Couns. 2003 Mar; 49(3):273–8.
12. Patel VL, Kaufman DR, Arocha JF. Conceptual change in the biomedical and health sciences domain. In: Glaser R, editor. Advances in Instructional Psychology. Hillsdale, NJ: Lawrence Erlbaum Associates; 2000. p. 329–92.
13. Sivaramakrishnan M, Patel VL. Reasoning about childhood nutritional deficiencies by mothers in rural India: A cognitive analysis. Social Science & Medicine. 1993;37(7):937–52.
14. Keselman A, Kaufman DR, Patel VL. "You can exercise your way out of HIV" and other stories: The role of biological knowledge in adolescents' evaluation of myths. Science Education. 2004;88(4):548–73.
15. Ogden J, Branson R, Bryett A, Campbell A, Febles A, Ferguson I, et al. What's in a name? An experimental study of patients' views of the impact and function of a diagnosis. Fam Pract. 2003 June 1; 20(3):248–53.
16. Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. J Biomed Inform 2003;36(4–5):224–41.
17. Patel VL, Arocha JF, Kushniruk AW. Patients' and physicians' understanding of health and biomedical concepts: relationship to the design of EMR systems. J Biomed Inform 2002 Feb; 35(1):8–16.
18. Jucks R, Bromme R. Choice of words in doctor-patient communication: an analysis of health-related internet sites. Health Communication 2007;21(3):267–77.
19. Keselman A, Massengale L, Ngo L, Browne A, Zeng Q, editor. The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge about Gender-Specific Illnesses. ISBMDA; 2006.
20. Sugarman J, Butters RR. Understanding the patient: Medical words the doctor may not know. North Carolina Medical Journal 1985;46(7):415–7.
21. Davidson B, Blum D, Cella D, Hamilton H, Nail L, Waltzman R. Communicating about chemotherapy-induced anemia. J Supp Oncol 2007;5(1):36–40, 6.
22. Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing consumer health vocabulary familiarity: an exploratory study. J Med Internet Res 2007;9(1):e5.