In Their Own Words?
A Terminological Analysis of E-mail to a Cancer Information Service

Catherine Arnott Smith¹, P. Zoë Stavri², and Wendy Webber Chapman¹

¹Center for Biomedical Informatics, University of Pittsburgh
²Division of Medical Informatics & Outcomes Research, Oregon Health & Science University

Abstract

Objective: To better understand the terms used by consumers to describe their health information needs and determine if this “consumer terminology” differs from those used by health care professionals.

Methods: Features and findings identified in 139 e-mail messages to the University of Pittsburgh Cancer Institute’s Cancer Information and Referral Service were coded and matched against the 2001 Unified Medical Language System Metathesaurus. Results: 504 unique terms were identified. 185 (36%) were exact matches to concepts in the 2001 UMLS Metathesaurus (MTH). 179 (35%) were partial string matches; 119 (24%) were known synonyms for MTH concepts; and 2 (<1%) were lexical variants. Only 19, or 4% of the total terms, were not found to be present in the 2001 MTH. Conclusion: 96% of the clinical findings and features mentioned in e-mail by correspondents who did not self-identify as healthcare professionals were described using terms from controlled healthcare terminologies. The notion of a paradigmatic “consumer” who uses a particular vocabulary specific to her “consumer” status may be ill-founded.

Introduction

Patrick et al.¹ have called the “consumer vocabulary problem” a fundamental issue in health information provision. This is the problem of mismatch between terms used by healthcare professionals and those used by the consumers who receive their services.

Compatibility, or the lack of it, between consumer and clinical terminology has been investigated in domains including website usability²,³,⁴, information prescriptions⁵, and HMO report cards⁶. Researchers have argued that modifying or mapping existing controlled medical terminologies to those used by consumers can remove barriers to patient empowerment and facilitate communication for the benefit of all involved. The issue is seen as so important that one packaged consumer vocabulary has been advertised as the “Rosetta Stone for the consumerization of healthcare.”⁷

One guiding assumption in this research domain is that consumers have their own “language” susceptible to analysis. However, as Lewis et al. have cautioned: “To presume that [consumer language] contains a knowable, stable vocabulary and grammar similar in structure to that of the formal languages of health care imposes a professional structure on a very personal experience.”⁸ Any study of consumer “language” will beg the question of what terms consumers actually use. Two significant challenges to consumer terminologists are the definition and the capture of terms that accurately reflect consumer reality.

One strategy to enhance usability of information targeted at consumers is to involve patients themselves in the development of their own educational material; for example, clinical practice guideline composition⁹. Another approach mines terms from existing text. Patrick et al.¹⁰ not only extracted words and phrases from print publications explicitly authored by consumers, but raided the Dictionary of American Regional English for their folk equivalents.

Content analysis of query log files and e-mail messages can also provide a new perspective on consumer vocabulary. For example, McCray et al.¹¹ examined 3 months’ worth of queries submitted to the National Library of Medicine’s home page. The principal focus of this study was the content and scope of the information requests themselves, and not the language in which those requests were expressed. However, in the process of log analysis, these authors were able to identify common terminological problems in query formulation, ranging from translation errors (“psicology”) to transcribed verbal slips (“prostrate cancer”). This work points to email analysis as one fruitful source of consumer terms.

E-mail service remains the predominant popular reason for Internet usage; in March 2002, one
telephone survey found that 92% of all Internet users had sent or received email\textsuperscript{14}. The strengths and weaknesses of this communications medium have been described by D’Alessandro and colleagues, who examined e-mail as a window into user needs for a digital health sciences library. These authors concluded that because e-mail was direct and asynchronous at the same time, benefits included

the convenience of replying when one chooses…easy clarification of meaning or instructions…retention of the interaction, especially information conveyed, which would need to be referenced in the future.\textsuperscript{[13]}

Duff and Johnson looked at e-mail expressions of information need submitted to North American archivists. Their reason was that “E-mail questions [afford] the opportunity to examine users’ requests in their own words” and because questions asked via this medium differed from telephone, or face-to-face, inquiries: The sender of e-mail “knows she needs to include as much information as possible because of the lack of immediate feedback”\textsuperscript{[14]}. For all these reasons, e-mail messages would seem an ideal source of consumer terminology for analysis.

Studies of e-mail content in the healthcare context are few in number. An early literature review by Moyer, Stern, Katz and Fendrick\textsuperscript{15} identified 12 e-mail studies, of which the majority focused on patient-provider communication, e.g., Fridsma, Ford, and Altman\textsuperscript{16}. Some early studies were concerned with demographics; for example, Gillespy and Richardson\textsuperscript{17} classified 98 unsolicited messages according to the sender’s job responsibility, gender, and physical location.

However, several authors have looked at e-mail content. It was suggested by Till\textsuperscript{18} as early as 1995 that messages exchanged in an online cancer support group were a form of “journaling … eminently suitable for qualitative data analysis” (p. 379). Klemm and colleagues\textsuperscript{19} classified similar support group messages by theme, and found that information-seeking was the principal topic. D’Alessandro and colleagues\textsuperscript{20,21} analyzed unsolicited email sent to the maintainers of a digital health sciences library (751 messages; 1998) and the author of a Web-based textbook (315 messages, 2000) to determine the nature and scope of senders’ information needs.

These studies reveal that the needs expressed are great, particularly among those who self-identified as consumers or as patients. Gillespy and Richardson, for example, found that 54 of their 98 unique correspondents were “nonclinical or nondental”\textsuperscript{22} and that 50% of these were writing to request either clinical advice or a referral. Patients and laypeople also made up the largest single category of senders in the two studies done by D’Alessandro and colleagues: 36.8% in 1998 and 62% in 2000.

Cancer patients are a group for which the need for health information “about diagnosis, prognosis, treatment and the purpose and results of tests is particularly great”\textsuperscript{23}. These patients and their friends and family members understand early that obtaining information regarding their disease and its treatment is critical to their care. In fact, hematology and oncology were among the most frequently expressed information topics sought by consumers in D’Alessandro’s first study.\textsuperscript{24} Quade, Far and Püschel\textsuperscript{25} surveyed users of a German implementation of the CancerNet website (http://www.meb.uni-bonn.de/). These authors found that nearly 70% of visits by “workers, housewives, and businessmen” to the site took place because of an individual’s cancer diagnosis.

The present study examines e-mail messages submitted to a cancer information service at a large academic medical center. The purpose is to better understand the terms used by consumers to describe their health information needs. The hypothesis is that these terms will be different from those used by healthcare professionals. To assess the complexity with which consumers expressed themselves, these e-mail messages were analyzed for their reading levels and reading ease according to established instruments. In particular, we are interested in the degree to which this “consumer terminology” agrees with health care terminologies in use by health care professionals and incorporated as source vocabularies in the Unified Medical Language System. Correspondence between the two distinct terminologies will obviously affect the ease of mapping from one world to the other.

**Setting/Methods**

The University of Pittsburgh Cancer Institute comprises 12 diagnosis-specific centers. It is one of 41 NCI Comprehensive Cancer Centers in the country, and the only such Center in Western Pennsylvania. The UPCI Cancer Information and Referral Service (CIRS) was inaugurated in 1988 and is staffed entirely by oncology nurses and social workers. Cancer information seekers can call an 800 number or, since 1997, ask questions via e-mail.
Since March, 2001, questions have been submitted via the “Ask A Question” form at the CIRS web page; form content is routed as an e-mail message to staffers.

The first author was provided with deidentified digital copies of every such e-mail received by the CIRS during an 8-month period between November 15, 2000, and July 1, 2001, a total of 179 messages. Thirty-one messages were discarded because they were: (a) spam, or automated junk mail (10%; 6%); (b) a copy of another message already present in the file (18%; 10%); or (c) empty of content (3%; 2%). Finally, nine messages (5%) in which the sender self-identified as a health care professional (“I am an RN”) were removed from the set. This was not done because health care professionals’ information needs are unimportant, but rather to establish that the terms obtained for the study would be as close to those of “consumer,” in the sense of “lay,” language as possible. This left a total of 139 messages for further analysis.

We analyzed the messages for their reading level using the popular Flesch-Kinkaid method. To obtain a list of “consumer terms”, three medical librarians were instructed to code the messages using the methodology previously tested by Lowe et al. These authors focused specifically on the “features” and “findings” present in radiology reports to evaluate results of an automatic text UMLS-coding system. Each librarian coded all 139 messages.

The unique words and phrases representing “features” and “findings” identified by any of the three indexers were then submitted for batch file processing by the Unified Medical Language System’s Knowledge Source Server (http://umlsks.nlm.nih.gov) and NLM’s MetaMap to determine matches to the 2001 edition of the UMLS Metathesaurus (MTH). Terms were classified as either (1) exact matches to the MTH; (2) present as synonyms for a concept present in the MTH; (3) partial string matches to concepts in the MTH, as determined by MetaMap; or (4) not present in the MTH.

Results
Message characteristics. Relatives of cancer patients, not including spouses, comprised the largest category of e-mail senders (45%; 32%). Cancer patients writing on their own behalf were the second largest group (25%; 18%). Spouses (11%; 8%), friends or colleagues (7%; 5%), school research or employment-related writers (8%; 6%) and persons who stated that they were not patients, but acting on a patient’s behalf (8%; 6%) made up the rest of the population. In 35 cases (25%) it was not possible to determine the role of the e-mail writer in requesting information.

The writer’s gender was not discernible in 113 (81%) of the messages. The 26 messages in which the sender self-identified as male or female fell almost evenly between males (54%) and females (46%).

Reading level. The Flesch-Kinkaid reading level score for each message was obtained using Microsoft Word’s grammar-checking utility. The subject line and body of each message was evaluated; e-mail header information was ignored. The average Flesch-Kinkaid score of the 139 messages was a reading level of 9.5 grade.

Term analysis. The 139 e-mail messages analyzed in this study yielded 504 unique terms that the librarian indexers considered either Features or Findings. 185 (36%) were exact matches to concepts in the 2001 UMLS Metathesaurus (MTH). 179 (35%) were found by MetaMap to be partial string matches to MTH concepts. Another 119 (24%) were found to be known synonyms to MTH concepts, and 2 (<1%) were lexical variants. Examples of exact matches [M], synonyms [S], and partial string matches [PSM] appear in Table I.

The remaining 19, or 4% of the total terms, were not found to be present in the 2001 MTH. The complete list of terms present as features or findings in the CIRS e-mails, but not matched in the MTH, appears in Table II.

Medical misspelling. McCray and colleagues in their 1999 NLM query log analysis found that misspellings accounted for some query problems. The results of their study informed development of a terminology server, which today facilitates queries submitted at www.clinicaltrials.gov by suggesting alternate spellings for medical words not recognized. To investigate misspelling as a possible contributor to UMLS mismatch, the author and a fourth librarian reviewed the 19 nonmatching terms and found that 14 were obvious spelling mistakes or typographical errors. These have been italicized in Table II. Four of the remaining 5 terms appear to be legitimate but not included in the UMLS (i.e., "cribriform"). Only "endo", common slang for "endometriosis", conforms to the popular notion of consumer terminology.
**Table I**

Selected “Consumer Terms”
Found in UMLS 2001 Metathesaurus

- Burkitt lymphoma [M]
- diffuse small cleaved cell lymphoma [M]
- glioblastoma multiforme [M]
- Hepatosplenomegaly [M]
- lymph node involvement [M]
- pseudocholinesterase deficiency [M]
- side effects [S]
- Paget's disease [S]
- Lymphnode [S]
- Linfonodos [S]
- FUO [S]
- CBD [S]
- Radiation Necrosis [PSM]
- jugulare glomus skullbase tumor [PSM]
- hurting in privates [PSM]
- figo stage II adenocarcinoma of the endometrium [PSM]

**Table II**

“Consumer Terms”
Unmatched in UMLS 2001 Metathesaurus

- Abdomine
- Adenocarinoma
- Atropy
- Barier
- Bloodwork
- carcinods of the liver
- Cribiform (used in 2 unique messages)
- Disease
- Diaporrea
- Endo
- fibroadenomnas
- Kydney
- light-headness
- Melenoma
- metastic adencarcinoma
- OVERIES
- Perineum
- Quadrant
- Sulci

**Discussion**

The typical e-mail message submitted to the University of Pittsburgh Cancer Institute between November, 2000, and July, 2001, was 134 words long and was written by a relative of a cancer patient at a 9th-grade reading level. “Low literacy” is defined as reading, writing, and understanding information written between the levels of 5th and 8th grade.3 Thus the persons composing these e-mails could be considered to be of average literacy. Interestingly, researchers who evaluated patient education materials on commercial, academic, government and association websites in 1999 found the sampled material to average a 10th grade Flesch-Kinkaid reading level, which they considered too high for most patients to understand. This suggests that the writers of e-mails examined in this study might constitute a reasonable target population for typical Web-based patient education materials.

However, the terms these e-mail writers used to express their health information needs overlapped in 96% of the cases with terms from the 92 healthcare terminologies that comprise the 2001 UMLS. This dissonance between low reading level and high vocabulary level needs to be explained. One possible explanation is that although the words used in these e-mails are long and clinical in nature, the sentences in which they are expressed are short, thus providing a distorted sense of reading level; an identical result was noted by Chapman et al.3 in their comparison of speech-generated and traditional radiology reports, both of which showed Flesch-Kinkaid reading levels halfway through 7th grade. The radiology reports averaged 50 words per sentence; the CIRS e-mail messages, 35.

Second, without the opportunity to observe the writers of these messages, we can never know if an e-mail is more like a letter, or a transcription—of pathology reports or physician letters detailing the cancer patient’s condition. Finally, another explanation, and a limitation of this study, rests in the anonymous nature of e-mail itself. While it is possible to eliminate those correspondents who self-identify as MDs or RNs, it is impossible to know the occupational status, let alone infer the degree of education, of those who do not so self-identify. This means that “consumer vocabulary” may ultimately be as nebulous a concept as is “consumer” itself.
Conclusion

Mi father have melanome, and i wanna know some more information about it PLEASE send to me more information about o(6)-benzyguanine!!!!!! PLEASE!!!

from an e-mail message to the CIRS

Analysis of 8 months’ worth of e-mails to a cancer information service reveals that almost all of the clinical findings and features mentioned by correspondents who did not self-identify as healthcare professionals were described using terms from controlled healthcare terminologies. As the extract above illustrates, the notion of a paradigmatic “consumer” using a vocabulary specific to her “consumer” status may be ill-founded. In future work we will examine the types of questions posed by these seekers of cancer information in hopes of more fully developing a picture of such a consumer.

Acknowledgments

Ms. Arnott Smith is supported by a U.S. National Library of Medicine (NLM) Medical Informatics Trainee Grant # 5-T15 LM07059. The authors express their thanks to Alan Aronson, PhD, Lister Hill National Center for Biomedical Communications, National Library of Medicine, for his assistance with this project. Carolyn Nunnelly, RN, University of Pittsburgh Cancer Institute, graciously provided access to the emails analyzed in this study.

References

8Lewis D, Brennan PF, McCray AT, Tuttle M, Bachman J. If we build it they will come: Standardized consumer vocabularies. Medinfo 2001, 10(1), 1530.
10Patrick et al., 2001.
11McCray et al., 1999.
21D’Alessandro et al., 2000.
24D’Alessandro et al., 1998
26University of Pittsburgh Cancer Institute CIRS website: http://www.upci.upmc.edu/internet/about/contact.html
30McCray et al., 1999.
31McCray et al., 2000.
33Graber et al., 1999.