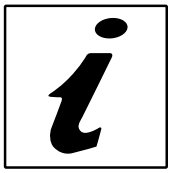
Information in practice



A taxonomy of generic clinical questions: classification study

John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, P Zoe Stavri

Abstract

Objective To develop a taxonomy of doctors' questions about patient care that could be used to help answer such questions.
Design Use of 295 questions asked by Oregon primary care doctors to modify previously developed taxonomy of 1101 clinical questions asked by Iowa family doctors.
Setting Primary care practices in Iowa and Oregon.
Participants Random samples of 103 Iowa family doctors and 49 Oregon primary care doctors.
Main outcome measures Consensus among seven investigators on a meaningful taxonomy of generic questions; interrater reliability among 11 individuals who used the taxonomy to classify a random sample of 100 questions: 50 from Iowa and 50 from Oregon.
Results The revised taxonomy, which comprised 64 generic question types, was used to classify 1396 clinical questions. The three commonest generic types were "What is the drug of choice for condition x?" (150 questions, 11%); "What is the cause of symptom x?" (115 questions, 8%); and "What test is indicated in situation x?" (112 questions, 8%). The mean interrater reliability among 11 coders was moderate ($\kappa = 0.53$, agreement 55%).
Conclusions Clinical questions in primary care can be categorised into a limited number of generic types. A moderate degree of interrater reliability was achieved with the taxonomy developed in this study. The taxonomy may enhance our understanding of doctors' information needs and improve our ability to meet those needs.

Introduction

Doctors often have questions about care as they see their patients: "How soon should I mobilise a patient with a deep vein thrombosis?" "How common is depression after infectious mononucleosis?" "Should a pregnant woman at full term with spontaneous rupture of membranes but not in labour come to the hospital now (3 am) or could she wait four hours?" Doctors answer only a minority of such questions authoritatively by consulting information resources.[1–3]

Answers might be more readily available if the authors of such resources knew what information needs arise in practice. In a previous study of doctors' questions, we developed a scheme to classify 1101 questions collected from 103 Iowa family doctors.[2] The purpose was to determine whether the essence of clinical questions could be captured by a limited number of generic question types. Questions with nearly identical structures (such as "How should I treat her Paget's disease?" and "How should I treat his epididymitis?") were placed into a single generic type ("How should I treat condition x?"). Through an iterative process of coding and revision, we developed a taxonomy of 69 generic types. This taxonomy may have limited applicability, however, because it was based on questions from a homogeneous group of doctors and because its interrater reliability was measured among a small group of investigators.

Therefore, in the current study, we modified this previously developed taxonomy to accommodate a different set of questions, and we measured interrater reliability in a more heterogeneous group of coders. Our goal was to produce a logical and concise classification scheme that could be applied reproducibly to the full range of questions that occur in primary care. We believe that such a scheme could increase the likelihood of finding answers to primary care questions. For example, the scheme could be used to identify frequently asked but problematic question types, enabling authors to develop better answers and more effective strategies for linking questions to their answers.

Participants and methods

Original taxonomy

In our previous study, 1101 questions about patient care were collected from 103 randomly selected Iowa family doctors.[2] The investigators visited doctors in their offices and recorded questions between patients' visits. Participants were asked to report everything from "clear cut questions (What's the dose of metformin?)" to the "vague, fleeting uncertainties" that they would normally keep to themselves. The purpose was to describe the questioning and answer seeking behaviour of family doctors and to develop a taxonomy of generic questions.

Additional questions

In a separate study Gorman and Helfand collected 295 questions from 49 Oregon primary care doctors (29 family doctors, 14 general internists, and 6 general

Department of Family Medicine, 01291-D PFP, University of Iowa College of Medicine, 200 Hawkins Drive, Iowa City, IA 52242-1097, USA
John W Ely *associate professor*

Praxis Press, 36 W 25th Street, 7th Floor, New York, NY 10010, USA
Jerome A Osheroff *director of informatics*

Division of Medical Informatics and Outcomes Research, Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portland, OR 97201, USA
Paul N Gorman *assistant professor*

Department of Family Practice, Michigan State University, B101 Clinical Center, East Lansing, MI 48824-1315, USA
Mark H Ebell *associate professor*

Moses Cone Family Medicine Residency, 1125 N Church Street, Greensboro, NC 27401, USA
M Lee Chambliss *assistant clinical professor*

continued over

BMJ 2000;321:429–32

bmj.com

A list of the taxonomy of generic clinical questions appears on the BMJ's website

Presbyterian Medical Center, 39th and Market Streets, Medical Arts Building, Suite 102, Philadelphia, PA 19104, USA
Eric A Pifer
assistant professor of medicine

School of Information Resources and Library Science, University of Arizona, 1515 East First Street, Tucson, AZ 85719, USA
P Zoe Stavri
assistant professor

Correspondence to:
J W Ely
john-ely@uiowa.edu

paediatricians).[3] The participants were asked to report questions about diagnosis or management. The purpose was to determine how doctors decide which questions to pursue and which to leave unanswered.

Generic question taxonomy

In our current study, seven investigators coded a random sample of 100 of the additional questions using the previously developed taxonomy.[2] Working independently, each investigator suggested changes to better accommodate the questions. We added new generic question types, changed the wording of existing types, and combined closely related types. A fourth option was to use an existing type to make a plausible, albeit imperfect, match. These decisions were based on consensus and guided by our goal of producing a concise, intuitive, reliable taxonomy.

The revised taxonomy was then distributed to all investigators, who independently coded a second random sample of 100 of the additional questions. Further changes were made and approved by all investigators. The 100 questions were categorised into three groups: all coders agreed on the generic type (n = 32), all but one coder agreed (n = 26), and more than one coder disagreed (n = 42). In the final round of taxonomy revisions, we focused on the last group, looking for problematic and ambiguous questions. For example, it was often difficult to determine whether a doctor was asking about diagnostic or therapeutic management: "Should silent ischaemia be pursued in an 80 year old woman with no symptoms and (known) bad coronary disease?" For this question, most coders assigned the code 3.1.1.1, which was the label for the generic question, "How should I manage condition x (not specifying diagnostic or therapeutic management)?" However, two coders assigned 1.3.1.1, which was the label for "What evaluation is indicated in situation x?" Analysis of such inconsistencies allowed us to address ambiguous elements in the taxonomy through changes in wording and a set of 37 coding guidelines. For example, one guideline states: "If diagnostic evaluation is the only reasonable consideration, use '1.3.1.1.' If treatment could be considered as part of the answer, use '3.1.1.1.'"

Taxonomy reliability

To measure the interrater reliability of the final taxonomy, the seven investigators coded a final random sample of 100 questions, 50 from Iowa and 50 from Oregon. In addition, four volunteers who were not familiar with the taxonomy coded the same 100 questions. The κ statistic[4] was used to estimate interrater reliability. This is a measure of agreement, which corrects for agreement that occurs by chance. It can be defined as $(P_o - P_e)/(1 - P_e)$, where P_o is the observed agreement and P_e is the agreement expected by chance.[4] From this formula, it can be seen that when the number of categories is large, as in this study, the agreement expected by chance (P_e) will be close to zero, and the κ will be close to the observed percentage agreement (P_o). We used a z test based on the κ values and their standard errors to compare reliability between groups of coders (investigators *v* volunteers) and between groups of questions (Iowa *v* Oregon). We chose a two tailed significance level of 0.05 and performed all analyses with Stata (Stata Corporation, College Station, TX).

Results

Demographic data

The mean age of the 103 Iowa doctors was 48 years, 23 (22%) were women, and 54 (52%) practised in a rural area.[2] The mean age of the 49 Oregon doctors was 45 years, 6 (12%) were women, and 24 (49%) practised in a rural area.[3] The investigators comprised three academic family doctors, three internists with interest and training in medical informatics, and a medical information scientist. The four volunteer coders comprised three family doctors at the University of Iowa and a medical librarian.

Generic questions

The generic questions were categorised using four hierarchical levels of specificity (see extra table on the *BMJ* website for details). The first level consisted of five broad areas: diagnosis, treatment, management, epidemiology, and non-clinical questions. Management questions asked what steps to take without distinguishing between diagnostic steps and treatment steps. A branching structure of secondary, tertiary, and quaternary levels further characterised the generic questions. Each quaternary category was exemplified by one or more closely related generic questions. For example, the question "Is there a way to continue lovastatin in patients with side effects of headache or indigestion (such as reduce the dose)?" would be coded as "treatment" (primary), "drug prescribing" (secondary), "adverse effects" (tertiary), and "administration in the face of adverse effects" (quaternary). The generic question corresponding to this quaternary category is "How can drug x be administered without causing adverse effect y?"

The revised taxonomy comprised 64 quaternary categories, down from 69 in the original version. For 42 categories (66%), the generic question examples consist of several closely related variations. For example, "What is the cause of symptom x?" "What is the differential diagnosis of symptom x?" "Could symptom x be condition y or be a result of condition y?" and "What is the likelihood that symptom x is coming from condition y?" were all considered variations of the same generic question. In each of these variations the doctor wants to know the cause of the patient's symptom. However, the answers to each question type would have a somewhat different focus. On average, there were 2.7 variations per generic question (range 1-11). If greater specificity is required in future applications of this taxonomy, a fifth level of specificity could separate these variations, enabling 175 question types to be hierarchically classified.

Question frequency

After combining the Iowa and Oregon questions (n = 1396), we found that the three most common generic types were "What is the drug of choice for condition x?" (150 questions, 11%), "What is the cause of symptom x?" (115 questions, 8%), and "What test is indicated in situation x?" (112 questions, 8%) (see table). Eight (0.6%) of the 1396 questions could not be classified beyond the primary level. To accommodate these questions, each primary level included a "not elsewhere classified" category.

Interrater reliability

The combined κ statistic for all 11 coders was 0.53 (55% agreement, indicating "moderate" reliability[5]). Agreement was slightly higher for the 50 Iowa questions than for the 50 Oregon questions (κ values 0.54 *v* 0.51, $P < 0.001$). When only the five broad areas in the primary level of the taxonomy were considered (diagnosis, treatment, management, epidemiology, non-clinical) agreement was "substantial"[5] ($\kappa = 0.70$), and agreement remained substantial when the primary and secondary levels (26 categories) were considered ($\kappa = 0.62$).

Agreement among the seven investigators was not significantly higher than among the four volunteers (κ values 0.55 *v* 0.54, $P = 0.33$). Agreement among the investigators with previous coding experience was higher than that among the other seven coders (κ values 0.68 *v* 0.47, $P < 0.001$).

Generic questions derived from questions by primary care doctors in Iowa and Oregon and their frequencies

	Frequency					
	Iowa questions (n=1101)		Oregon questions (n=295)		Total questions (n=1396)	
Question	Rank	No (%) of questions	Rank	No (%) of questions	Rank	No (%) of questions
What is the drug of choice for condition x?	1st	112 (10)	1st	38 (13)	1st	150 (11)
What is the cause of symptom x?	2nd	106 (10)	11th	9 (3)	2nd	115 (8)
What test is indicated in situation x?	3rd	84 (8)	3rd	28 (9)	3rd	112 (8)
What is the dose of drug x?	4th	84 (8)	10th	10 (3)	4th	94 (7)
How should I treat condition x (not limited to drug treatment)?	7th	53 (5)	2nd	29 (10)	5th	82 (6)
How should I manage condition x (not specifying diagnostic or therapeutic)?	5th	62 (6)	19th	5 (2)	6th	67 (5)
What is the cause of physical finding x?	6th	61 (6)	16th	6 (2)	7th	67 (5)
What is the cause of test finding x?	8th	53 (5)	7th	11 (4)	8th	64 (5)
Can drug x cause (adverse) finding y?	10th	42 (4)	4th	17 (6)	9th	59 (4)
Could this patient have condition x?	9th	49 (4)	26th	2 (1)	10th	51 (4)

Discussion

Main findings

In this study, we modified a taxonomy of generic clinical questions and measured how reproducibly 11 coders could assign questions to it. We found that a large number of questions could be categorised using a limited number of generic types. Coding reproducibility was moderate and was highest among the most experienced coders.

Comparison with other studies

Cimino matched clinicians' natural language inquiries with generic types for which computerised retrieval strategies had been previously developed.[6] The query types were based on semantic relations drawn from the National Library of Medicine's Unified Medical Language System.[7 8] The generic types were applied to questions that would be submitted to computerised retrieval systems rather than to the "on the spot" questions that we collected. These investigators did not describe a comprehensive taxonomy of generic queries.

Graesser and colleagues developed a taxonomy for categorising the full range of questions that adults ask.[9] They were able to categorise 1000 questions using 12 semantic categories, such as "verification" (Is x true or false?), "disjunctive" (Is x or y the case?), "causal antecedent" (What caused some event to occur?), and so forth. We did not apply this taxonomy to our clinical questions, which were relatively focused and utilitarian.

Implications of study

Our goal was to build a taxonomy that was valid, reliable, concise, intuitive, comprehensive, and useful. We have identified four areas of potential usefulness. Firstly, taxonomies such as ours, that are based on the generic type of information needed, could be used to organise large collections of clinical questions for efficient retrieval. Authors who want to produce clinically relevant material should address real questions that occur in practice. But there is an infinite number of such questions, and even frequently asked questions would be unmanageable without some way to organise them.

Secondly, the generic taxonomy could be used to develop tools for linking questions to answers. Different types of resources, such as textbooks, prescribing resources, and laboratory handbooks, have different types of information and are thus suited to different types of questions. Earlier research has suggested that resource selection can be problematic.[10 11] Since our taxonomy focuses on the types of information needed, it could potentially route questions to appropriate resources by means of automated computer interfaces. This function could be particularly useful when the resource is an electronic knowledge base.

Thirdly, the taxonomy could be used to characterise and classify areas where current sources of knowledge systematically fail to address specific types of questions. For example, current resources often explain how to treat a disease (heart failure) but not how to treat that disease when it is accompanied by a common comorbid condition (renal failure). By identifying and classifying such gaps, we could develop guidelines to help authors produce more useful resources.

Fourthly, the taxonomy could be used to set priorities for clinical research. Some types of questions are inadequately answered because definitive answers do not exist. For example, the family doctors in this study often asked how to distinguish viral upper respiratory infections from bacterial sinusitis on the basis of clinical findings. Questions of the form "What is the cause of clinical finding x?" are not often addressed by researchers.

Limitations of study

The kinds of questions collected in studies of doctors' information needs seem to depend on the methods used to collect them. For example, the Iowa doctors in this study were more likely than the Oregon doctors to ask about the cause of symptoms and physical findings. But the Oregon doctors were asked to report questions about "diagnosis or management," whereas Iowa doctors were asked to report everything from "clear cut questions" to "vague, fleeting uncertainties." Both datasets used office observations to collect questions, but other methods, such as the critical incident technique,[12] have been used and might influence the kinds of questions collected.

Most of the investigators in this study were primary care doctors (three internists and three family doctors). A taxonomy of questions for non-primary care developed by other kinds of investigators (such as librarians or neurosurgeons) might look different from ours.

What is already known on this topic

In a previous study, the essence of 1101 clinical questions asked by family doctors was captured in 69 generic types (such as, "What is the drug of choice for condition x?")

The applicability of this generic question taxonomy may be limited because of the homogeneous nature of the participants

What this study adds

After revision of the original taxonomy, questions asked by a different group of 49 primary care doctors could be classified with moderate reliability among 11 coders

The taxonomy has four potential uses: to organise large numbers of real questions, to route questions to appropriate knowledge resources by using automated interfaces, to characterise and help remedy areas where current resources fail to address specific question types, and to set priorities for research by identifying question types for which answers do not exist.

The information needs analysed in this study were "user based"—that is, questions were collected without regard to the most appropriate method for answering them. Other question sets are "system based," focusing, for example, on questions submitted to computerised information retrieval systems.[13] Some user based questions ("What is causing her abdominal pain?") would require a system based modification before they could be answered by a general information resource ("What is the differential diagnosis of right lower quadrant pain in adolescent females?").

We achieved only moderate interrater reliability. However, the coding reliability in this study compares favourably with other attempts to categorise medical topics. For example, highly trained Medline indexers achieve "consistency percentages" of only 43% when assigning medical subject headings and subheadings (MeSH terms) to journal articles.[14]

Conclusions

Doctors do not pursue answers to most of their questions, partly because they believe the answers are not readily available[3]—a belief that is often correct.[1 3 15] Doctors need rapid, accurate, and accessible answers to on the spot questions as they see their patients.[16] By learning about doctors' questions, we hope to influence the content of clinical information resources. By organising the full range of information needs that occur in practice, we can begin to address the most common types.

Contributors: JWE collected the Iowa questions, coordinated the development of the taxonomy, and wrote the first draft of the paper. JAO had the original idea of organising clinical questions by generic type. He helped with constructing the taxonomy and with writing the paper. PNG collected the Oregon questions, guided the early development of the study and helped design the taxonomy. MHE, MLC, EAP, and PZS helped plan the study, and each coded 250 Oregon questions and 50 Iowa questions. They modified the first draft of the taxonomy to better accommodate these questions and approved the final version. All authors contributed to editing the paper. Dedra Diehl, Susan Langbehn, Robert Garrett, and Mark Graber helped test the taxonomy, and Jeffrey Dawson provided statistical support. JWE and JAO are guarantors for the study.

Funding: This study was supported by a grant (G9518) from the American Academy of Family Physicians Foundation.

Competing interests: None declared.

1 Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? *Ann Intern Med* 1985;103:596-99.
2 Ely JW, Osheroff JA, Ebell MH, Bergus GR, Levy BT, Chambliss ML, et al. Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999;319:358-61.
3 Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. *Med Decis Making* 1995;15:113-9.
4 Fleiss JL. *Statistical methods for rates and proportions*. 2nd ed. New York: Wiley, 1981.
5 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
6 Cimino J. Generic queries for meeting clinical information needs. *Bull Med Libr Assoc* 1993;81:195-206.
7 Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The unified medical language system: an informatics research collaboration. *J Am Med Inform Assoc* 1998;5:1-11.
8 McCray AT, Miller RA. Making the conceptual connections: the unified medical language system (UMLS) after a decade of research and development. *J Am Med Inform Assoc* 1998;5:129-30.
9 Graesser AC, Lang K, Horgan D. A taxonomy for question generation. *Questioning Exchange* 1988;2:3-15.
10 Osheroff JA, Bankowitz RA. Physicians' use of computer software in answering clinical questions. *Bull Med Libr Assoc* 1993;81:11-9.
11 Curley SP, Connelly DP, Rich EC. Physicians' use of medical knowledge resources: preliminary theoretical framework and findings. *Med Decis Making* 1990;10:231-41.
12 Northup DE, Moore-West M, Skipper B, Teaf SR. Characteristics of clinical information-searching: investigation using critical incident technique. *J Med Educ* 1983;58:873-81.
13 Lindberg DA, Siegel ER, Rapp BA, Wallingford KT, Wilson SR. Use of MEDLINE by physicians for clinical problem solving. *JAMA* 1993;269:3124-9.
14 Funk ME, Reid CA. Indexing consistency in MEDLINE. *Bull Med Libr Assoc* 1983;71:176-83.
15 Chambliss ML, Conley J. Answering clinical questions. *J Fam Pract* 1996;43:140-4.
16 Huth EJ. "In the balance": weighing the evidence. *Ann Intern Med* 1994;120:889.

(Accepted 22 May 2000)

One hundred years ago

The Temperature of Underground Railway Tunnels

Not very long ago it was gravely suggested that certain stations on the Underground Metropolitan Railway, where the air is most sulphurous and the general conditions simply Stygian, were useful as sanatoria for convalescent railway porters and for those suffering from phthisis and "delicate chests." It seems now that the latest accession to our underground railway system has hygienic properties of a different kind. It has been stated that a person who had suffered from anorexia for 18 months suddenly developed ravenous appetite after a single journey by the new underground electric railway, popularly known as the "twopenny tube." From the repetition of a similar therapeutic journey every two or three days this satisfactory state of affairs has been so well maintained that in his case an actual pecuniary profit has been achieved from the discontinuance of tonic remedies which have become needless. The two facts—the journey and the appetite—certainly seem conceivably connected when it is possible that the passenger went from an above ground atmosphere with a temperature of perhaps 80° F. to a tubal air of, say, 50°, thus adopting a convenient and expeditious method of counteracting any tendency to heat prostration that may have been present. But the more general opinion seems to be rather in the other direction, that the abrupt change in temperature together with the gale of wind, maintained as is probable by a system of fans, acts injuriously on persons who descend the lifts in a condition of profuse perspiration. (*BMJ* 1900;ii:595.)