The Sensitivity of Medical Diagnostic Decision-Support Knowledge Bases in Delineating Appropriate Terms to Document in the Medical Record.

Mitchell J. Feldman, M.D., G. Octo Barnett, M.D. and Mary M. Morgan
Laboratory of Computer Science, Massachusetts General Hospital

Abstract

A pertinent, legible and complete medical record facilitates good patient care. The recording of the symptoms, signs and lab findings which are relevant to a patient's condition contributes importantly to the medical record. The consideration and documentation of other disease states known to be related to the patient's primary illness provide further enhancement. We propose that developing sets of disease-specific core elements which a physician may want to document in the medical record can have many benefits. We hypothesize that for a given disease, terms with high importance (TI) and frequency (TF) in the DXplain, QMR and Iliad knowledge bases (KBs) are terms which are used commonly in the medical record, and may be, in fact, terms which physicians would find useful to document. A study was undertaken to validate ten such sets of disease-specific core elements. For each of ten prevalent diseases, high TI and TF terms from the three KBs mentioned were pooled to derive the set of core elements. For each disease, all patient records (range 385 to 16,972) from a computerized ambulatory medical record database were searched to document the actual use by physicians of each of these core elements. A significant percentage (range 50 to 86%) of each set of core elements was confirmed as being used by the physicians. In addition, all medical concepts from a selection of full text records were identified, and an average of 65% of the concepts were found to be core elements. If this KB-directed method for obtaining core sets of problem-specific elements is appropriate for prevalent diseases (for which we have abundant medical records to perform the validation) then it may also prove worthwhile for collecting these terms for rarer diseases, for which patient records are scarce.

Introduction

Knowing what to record in a patient's chart is one of the medical student's most vexing concerns. In a chapter on the patient's record in her classic text on physical examination, Barbara Bates warns "... information can be buried in a mass of excessive detail, to be discovered by only the most persistent reader[1]." Clearly we can not record everything. Having a set of readily available, problem-oriented terms, however, could be quite useful. In addition to the potential educational benefits of imparting to students and residents a succinct list of problem-specific elements, these sets of terms could facilitate entry of a problem-based medical record in a physician workstation. Any increase in a standardized vocabulary which these sets might afford could help to facilitate communication among health care providers and among computer-based applications. As such, this work is closely related to the activity of the Unified Medical Language System (UMLS) project where biomedical concepts are identified as an aid to mapping between knowledge sources[2].

Previous Work

Three different examples of applications where controlled clinical vocabularies have been derived are illustrative:

The American Board of Pediatrics (ABP) has incorporated record review as part of its mandatory recertification process. To help pediatricians prepare for this activity, the ABP has started to produce several "Guides for Record Review" on selected clinical conditions. In the margins of these guidebooks, "important elements to be included in the record" appear. These problem-specific elements are thus intended to help a physician with quality assurance of his/her own personal patient records[3].

A second example is shown in a paper describing the Emergency Department Expert Charting System (EDECS). The authors note that "the American College of Emergency Physicians is attempting to create standards [of care] that are based on the patient's chief complaint." EDECS has been developed to present a physician with a series of screens which prompt an appropriate history and physical exam based on the chief complaint[4].

In a third application, Holbrook and Aghababian developed a list of critical pertinent positive and negative findings for each of five high-risk diagnoses. This work has become the basis for a more broad-based real-time risk management prompting medical record system which uses voice-recognition technology for input entry[5].

In each of the three applications just described, the authors have compiled problem-specific groups of elements. The examples demonstrate that such elements can be used in quality assurance for record review, facilitated entry into computer-based medical record systems and improving risk management.

In none of the applications just cited did the authors indicate how these sets of terms were derived. One must assume that the source is a combination of personal clinical experience (in an analogy to the origin of practice policies, Eddy would call this the 'global subjective judgement approach[6]') and the literature.

There are several methods by which one might try to identify systematically these sets of terms. One alternative would be to search the medical literature for the appropriate information. Another possibility would entail reviewing large numbers of medical records. Both of these alternatives are inordinately time consuming. We propose that computer-based medical decision support KBs, by virtue of their containing a wealth of data on disease-term relationships, provide an excellent source of information from which to derive problem-specific sets of terms with minimal effort. By using three individual KBs, we can pool the expertise of all the knowledge sources used in the development of these systems as well as decrease the regional or geographic influence of any one. Additionally, deficiencies in any one KB may be balanced by the other two.

Cimino and Barnett previously described the validation of a prototype vocabulary of cardiac exam terms. These terms, derived from MESH, SNOMED, textbooks and other sources, were validated by recreating the cardiac exam portion of 75 actual medical records using this controlled vocabulary[7]. In the present study, we will also use actual patient records to validate the problem-specific terms. We are not assuming that either the medical records or the KBs are a "gold standard." Rather, we propose that the KBs provide an easily accessible, fast way to identify problem-specific terms, and the medical records provide a fertile "testing-ground" to determine the usage of these terms by clinicians.

In arriving at a set of core terms, it is *not* our goal to specify prescriptively which terms should always be recorded for a particular disease as assessed by validity, appropriateness or relevance. Eddy might advocate that such standards be derived only after outcomes-based research, possibly considering patient preferences[8], a method which "can cost tens of thousands of dollars and require several months[7]"– and this only for one disease or problem! Rather we hope to show that the quantitative, easily accessible information contained in computer-based KBs can be used to suggest a list of elements which a physician may want to document in the medical record.

Methods and Procedures

Background

Three KBs were used to derive the core term lists.

QMR has been under development since 1985 by Miller et al at the University of Pittsburgh. This KB contains some 600 disease profiles and 4300 disease manifestations or findings. For each finding in a disease profile, the term frequency (TF, ranging from 1-5) describes how frequently the finding is found in a patient with the given disease. The evoking strength (ES ranging from 0 to 5) indicates how strongly a disease should be considered in a patient with the finding. Each term also has associated a disease-independent term importance (TI, range 1-5) which describes how significant the term is (if the TI=1, the term could be disregarded, if 5, the term should be explained by a disease in the differential)[9]. For each disease profile, 50 to 100 articles from the medical literature are reviewed and consultation from clinical experts obtained[10].

The DXplain project has been evolving since 1985 under the direction of Barnett at the Massachusetts General Hospital. This KB contains over 2000 diseases and 4500 findings. DXplain also uses the concepts of TF, ES (range 1 to 9) and TI (range 1 to 5). The knowledge sources for DXplain include CMIT[11], medical textbooks and articles from the medical literature[12].

The development of the Iliad project began in 1987 by Warner et al at the University of Utah. The KB contains over 1000 diseases and 5600 disease findings. A disease is represented as a series of Bayesean and Boolean frames containing findings and associated probability information. In Iliad, the counterpart to the TF of DXplain and QMR is the sensitivity or P(finding|disease) (range 0 to 1)[13]. The Iliad KB does not use the concept of TI. The KB is derived from clinical experts in knowledge engineering sessions, from the HELP system patient database and from literature searches[14].

KBs represent a source of quantitative information about disease-term relationships. They derive from a diverse set of resources including expert clinicians, medical texts, journal articles and patient record databases.

Derivation of core element lists

Ten common diseases were chosen from a list of prevalent diseases in a current COSTAR patient database in use at Massachusetts General Hospital[15]. The first ten diseases on this list which were profiled in DXplain, QMR and Iliad were chosen. For each disease, a core-element list (hereafter called the KB list) was constructed which included the union of all terms in the three KBs where TI $\geq$ 3 (for QMR and DXplain), TF $\geq$ 3 or 7 (QMR or DXplain) or sensitivity $\geq$ 0.5 (Iliad). TF $\geq$ 3 or 7 for QMR or DXplain, respectively, corresponds to terms with a frequency of greater than 50%.

In addition to containing manifestations of the disease at hand, the KB list can be enhanced by including as elements, other ('trigger') disease states which should be considered when evaluating a patient with the ('given') particular disease. For example, the KB list for the disease "angina pectoris" would include not only the manifestations "exertional chest pain" and "crushing substernal chest pain" but also the 'trigger' diseases "aortic valve stenosis" and "acute MI." The KBs provide the structure to obtain these additional diseases. In QMR, these 'trigger' diseases can be obtained using the 'LINK' structure. LINKed diseases with TF or ES $\geq$ 3 correspond, respectively, to diseases which occur in a significant number of patients with the given disease, or diseases which one should think of when the given disease is known to be present. Additionally, 'trigger diseases' may be obtained by searching the KBs for disease states for which the given disease, if it exists as a term in the KB, has an ES $\geq$ 3 or 7 (QMR or DXplain), or diseases which have a probability of > 50% when the given disease, if it exists as a finding in the KB is EXPLAINed (Iliad).

The first object of the experiment was to determine whether these KB-derived elements or concepts are used by clinicians. We reasoned that if these concepts were identified as present in the medical record, and if a significant percentage of the concepts in the medical record were found on the KB list, then the concepts derived from the KBs had been demonstrated to be a valid source of problem-specific descriptors.

In order to manage the searching process through medical records, a single dimensional cross tabulation frequency count was performed through all patient records in the COSTAR system for each disease considered. Using MQL (Medical Query Language, a database retrieval language which can be used with COSTAR[16]), one word, two word and three word phrase word frequency lists (WF lists) were generated. Each list contains, in descending order, the most frequent one, two or three word phrases encountered in the medical records for a given disease. Two other questions which this study addresses pertain to the WF lists: (1)Since arbitrary cutoffs were used in generating the WF lists (see footnote Table II), to what degree do the full text medical records contain concepts which cannot be found on these WF lists, and (2)To what degree do the WF lists contain concepts which cannot be found on the KB lists.

Concept Matching

The first issue addressed was whether the KB derived concepts could be documented in the WF lists. Each concept on the KB list was evaluated in turn. The concept was considered to be documented if any of the following criteria were met when searching the three WF lists:

(1)Exact word match.

(2)Synonym match e.g. the concept "Distal Interphalangeal Nodule" (DXplain term) was matched to "Heberdon's Node" (a medical record term), and "Heartburn" was matched to "Indigestion".

(3)Different stem but equivalent concept match e.g. "Proteinuria" is a KB term not present on the WF lists generated from 1187 records of UTI encounters. When the search of the WF lists was constrained to the morpheme "protein," the two word phrase "1+ protein" was found, which in the context of UTI medical records, is a concept match to "proteinuria"

(4)Hierarchically subsumed concept- a match to a 'parent' or 'child' concept, e.g. the term "Total lung capacity increased" is a KB term for the disease Asthma. Although this term is not on the WF list, the phrase "Check PFT's" is used 31 times in the 2236 asthma records examined. Because the information contained in "PFT's" will address the issue of "Total lung capacity," this type of concept matching was permitted. Only items at different levels of the hierarchy were allowed to match. For example, the KB term "Pyuria" would not match to the medical record term "Hematuria" even though these concepts are related, e.g. both would be obtained in a UA.

The next issue was whether there were other concepts on the word frequency lists not present on the KB lists. To ascertain this, the one, two and three word WF lists were reviewed manually. No therapy-related terms were considered since none of the three KBs contains therapy information. For each word phrase, two criteria needed to be satisfied:

(1)Is the word phrase clinically important?

(2)Is the word phrase relevant to the disease under consideration? e.g. "Renal insufficiency," while clinically important is not primarily germane to the disease 'Osteoarthritis.'

If the two criteria were satisfied and the word phrase was not on or conceptually related to any term on the KB list, a tally was incremented.

The next task was to assess the percentage of concepts in the full text medical records which were contained in the KB lists. 13 to 16 randomly selected full text records for each disease were reviewed. For each record, all pertinent concepts were identified. Again, therapy-related elements were not considered, and concepts deemed unrelated to the disease were also disregarded e.g. in a record about UTI, the concepts "concerned that mother died" or "cardiac 2/6 systolic ejection murmur" were judged not related. Concept matching was performed to determine the proportion of full text record concepts which could be found on the KB list.

Results

Table I shows the results on the ten diseases studied. Displayed in the first column is the number of full text records from which the WF lists were generated. The denominator in the second column is the number of KB terms (core elements) obtained for each disease from the three KBs, while the numerator indicates the number of terms actually documented in the WF lists. The extent to which the proportion of KB terms documented in the medical record is less than one is the extent to which extra terms exist on the KB list. An average of 62% (range 49-86%) of the KB-derived concepts were found to be present in the WF lists.

TABLE I

Proportion of total KB terms documented in the WF lists

	# rec'ds	Terms documented in WF lists/ Total KB terms
HTN	16,972	20/40
Diabetes mell.	4,393	24/45
Osteoarthritis	2,806	24/28
UTI	1,187	24/29
Asthma	2,236	24/39
Duodenal ulcer	385	17/28
Hypothyroidism	1,254	25/51
Angina pectoris	1,568	27/38
CHF	1,083	32/64
Sinusitis	444	13/24

Table II demonstrates the degree to which terms on the WF list were not contained on the KB list. Table II lists the number of one, two and three-word phrases which were used frequently in the medical records and deemed relevant medical concepts for the problem under consideration, but are not on the KB list.

TABLE II

Number of concepts on the WF lists NOT on KB list

	1 word	2 word	3 word
HTN	5	12	3
Diabetes mellitus	0	3	2
Osteoarthritis	2	9	13
UTI	3	14	5
Asthma	2	13	3
Duodenal ulcer	5	9	5
Hypothyroidism	0	6	2
Angina pectoris	0	8	6
CHF	1	19	4
Sinusitis	1	9	4

note: For the '1-word' word frequency list,terms were reviewed to a frequency of 0.2%. For the 2 and 3-word phrase word frequency lists, the 500 most frequent phrases were reviewed (0.04% to 0.17%)

Table III's results are perhaps the most interesting. Here we see the proportion of concepts from full text medical records which can be found on the KB list, on the WF list but not on the KB list, and on neither list. An average of 65% of the concepts identified in the full-text records sampled could be found on the

KB lists, and an additional 21% on the WF lists. One might wonder how 15% of the concepts identified in the full-text records could be missing from the WF lists, since the WF lists originated from the full-text records. The answer is that some of the WF lists cut off at a frequency of e.g. 4 (see note, Table II). Thus these WF lists would not reflect those concepts found in the medical records with a frequency of less than 4. In addition, not all concepts can be expressed as one, two or three word phrases, e.g. from the full text "Can walk up 1 flight of stairs with difficulty," in the context of Asthma medical records, the concept of "Dyspnea on exertion" can be surmised. However this concept would not appear on the WF lists which contain only one, two or three word contiguous phrases found in the text of the patient records.

(3)Findings with a high TI are likely to be of greater clinical significance than findings with a low TI (this is part of the definition of TI); as such they are 'higher priority' terms to include in a problem-specific set.

A possible source for error can occur when the authors assumed a KB list element was equivalent to a term actually used in the medical record when in fact the terms were not equivalent. For example, the KB-derived term "No Pheochromocytoma" in the Hypertension profile was matched to the one word concept "VMA" from the WF list which was used 31 times in 16,972 hypertension records. It was assumed that the use of the term "VMA" was describing vanillylmandelic acid, which is often measured in patients suspected of having pheochromocytoma, but this was not a straightforward concept match.

TABLE III

Percentage of Full text record concepts on KB list, Not on KB list but on Word Frequency list (+WF/-KB), or on Neither List (-KB, -WF)*. (Percentages may not add to 100% due to rounding)

	KB	+WF/-KB	-KB/-WF	Avg # Concepts per record
HTN	41%	39%	20%	4.4
Diabetes mellitus	68%	9%	23%	5
Osteoarthritis	50%	30%	20%	3.7
UTI	62%	28%	11%	4.3
Asthma	82%	11%	7%	3.7
Duodenal ulcer	73%	22%	5%	4.2
Hypothyroidism	81%	14%	6%	2.8
Angina pectoris	71%	13%	16%	3.5
CHF	63%	22%	15%	7.3
Sinusitis	56%	21%	22%	5.9
AVERAGES	65%	21%	15%	4.5

note: 13 to 16 full text records were reviewed for each disease (one record was reviewed for each of the 16 physicians that use the COSTAR system). A 'record' is defined as one encounter recorded by a physician on one patient for the specific disease under consideration. For osteoarthritis, as an example, a record averaged 4 lines of text.

*Some concepts from the medical record listed in the "neither list" column would have been on the Word frequency lists, if lower cutoffs for generating the WF lists had been used.

Discussion

Arriving at a set of problem-specific elements for use in the medical record may not seem a difficult task. The simplest strategy would be to read an article about the disease, and write down a list of symptoms, signs, lab findings and related conditions mentioned. Unfortunately this strategy may be suboptimal since the list will be biased from the perspective of one author compared to the large number of journal references, texts and clinical expertise present in three KBs. It may be difficult to distill from the article read even semi-quantitative information about the frequency of occurrence or clinical significance of the terms, whereas the KBs contain this data in a useful quantitative form. This allows the construction of a list of clinical elements tailor-made to specific criteria in a way not possible otherwise. We chose the criteria of high TI and TF when constructing our lists for the following reasons:

(1)To narrow down the list of elements to a manageable size for each disease.

(2)Findings which occur frequently in a disease are very likely the terms which physicians will need to use frequently in recording a patient's condition.

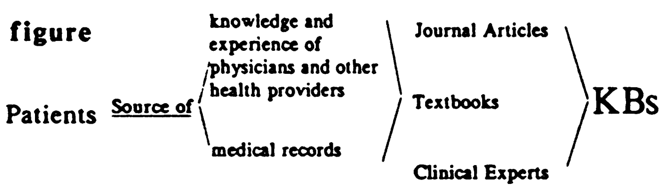
Table III shows that an average of 65% of the concepts identified in the full text medical records were found to be KB terms. For each disease, a single encounter was reviewed for each of the 13-16 physicians whose patients were in the COSTAR database. It is to be expected that only a few of the KB terms might be used in any single patient encounter; a physician would not need to use all the KB terms in each and every encounter. It is therefore possible that a higher percentage of the concepts in the full text records may have been found to be KB terms had we used multiple encounters for the same patient, rather than a single encounter.

Many of the terms found on the WF lists which were not on the KB lists are in fact present in the KBs, but with a lower frequency or TI than the arbitrary cutoffs necessary for inclusion on the KB list. For example the term "Pleural effusion" is present in the WF list for CHF, but is not on the KB list. The term is in both the DXplain and QMR KBs with TFs of 6 (20-50%) and 2 (6-35%), respectively. This example illustrates nicely the tradeoff in achieving a concise KB element list when choosing higher cutoffs, versus compiling an exhaustive list with lower cutoffs.

We are encouraged by our results which show that, by selecting the elements from the 3 KBs as outlined, a list of terms can be composed in which an average of 65% of the concepts from the medical record can be found. Previously, authors have described problem-specific lists of terms with application to the medical record [3,4,5]. While these authors did not describe their methods for arriving at these lists, we have described a straightforward method which we believe provides a good foundation for establishing them.

It would be enormously time consuming manually, and technically very difficult (if not impossible), to take large numbers of common text-based medical records and express the medical concepts within them using the controlled vocabulary. Conversely, if this problem-specific vocabulary structure were imposed at input, a physician might willingly substitute a KB derived term for a conceptually equivalent one, when entering the record. This is particularly true if the mechanism for doing this is easy and if incentives are great (e.g. useful patient statistics feedback, ease of retrieval of patient data for research purposes).

It is not surprising that medical diagnostic decision support KBs should identify terms which are useful to document in a patient's record. Ultimately, the chain of events leading to the development of a KB begins with a patient (see figure). KBs are thus, in a sense, 'higher level' or 'refined' aggregates of medical records, coupled with clinical experience and organized in a highly structured, quantitative fashion. The ability for these same KBs, then, to provide information which can serve as an outline or template for medical record data could be predicted.



Conclusion

Medical decision-support KBs contain detailed quantitative information about terms and disease-term relationships. We have described a relatively quick method to identify pertinent problem-specific elements for use in medical record documentation. For a specific disease, those KB terms with high TI and TF may be a good starting point for a core set of elements to document in the medical record. For the ten prevalent diseases examined, a substantial percentage of these KB terms were validated by documenting their presence in word frequency lists obtained from actual patient records. Moreover, a significant proportion of the concepts noted in the full-text medical records were KB terms. Any set of core elements will have to be adapted to the actual clinical site where used. The results obtained so far are encouraging. We suggest that this method may hold even greater value in the setting of rare diseases, since physicians may be less familiar with these conditions. KBs are filled with detailed information on rare diseases and physicians may find suggestions about appropriate terms to document especially useful in this context.

Acknowledgements

The authors would like to thank the following individuals: Marvin Packer, Joel Kahn, Diane Oliver and Ed Hoffer served as Internal Medicine consultants.
Chris Cimino provided helpful advice about vocabulary issues. This work was supported, in part, by an educational grant from the Hewlett Packard Corporation, and by UMLS contract [N01-LM-8-3513]. Dr. Feldman is supported by NLM training grant [2-T15-LM07037-04].

1 Bates, B. *A Guide to Physical examination*, 3rd ed. Philadelphia: J.B. Lippincott, 1983.

2 Lindberg DAB and Humphries BL. The UMLS knowledge sources: Tools for building better user interfaces. *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*. New York: IEEE Computer Society Press; 1990;121-25.

3 Nazarian LF. Reviewing the pediatrician's personal medical records. *Pediatr Rev.* 1990;12(5):133.

4 Schriger DL, Baraff LJ, Wilton R and McCoy JM. EDECS-The emergency department expert charting system. *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care.* New York: IEEE Computer Society Press; 1989;611-15.

5 Holbrook J and Aghababian R. A computerized audit of 15,009 Emergency Department Records. *Ann Emerg Med.* 1990;19:139-44.

6 Eddy DM. Practice policies: where do they come from? *JAMA.* 1990;263(9):1265-75.

7 Cimino JJ and Barnett GO. The physician's workstation: recording a physical examination using a controlled vocabulary. *Proceedings of the Eleventh Annual Symposium on Computer Applications in Medical Care.* New York: IEEE Computer Society Press; 1987;287-91.

8 Eddy DM. Practice policies: guidelines for methods. *JAMA.* 1990;263(13):1839-41.

9 Miller RA, Masarie FE and Myers JD. Quick Medical Reference (QMR) for diagnostic assistance. *MD Computing* 1986; 3(5):34-48.

10 Masarie FE and Miller RA. QMR A diagnostic decision-support program for internal medicine. *Group Practice Journal.* 1990; May/June:16-25.

11 Gordon B (ed). Current medical information and terminology, ed. 4. Chicaco, American Medical Association, 1971.

12 Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain- An evolving diagnostic decision-support system. *JAMA* 1987;258:67-74.

13 Iliad User Manual. Applied Informatics, Salt Lake City, 1990.

14 Yu H, Haug PJ and Turner CW. Embedding clinical statistics on a data dictionary for use in an expert system. *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care.* New York: IEEE Computer Society Press; 1989;235-41.

15 Barnett, GO. The application of computer-based medical-record systems in ambulatory practice. *N Engl J Med.* 1984;310:1643-50.

16 Medical Query Language User's Manual. The General Hospital Corp., Laboratory of Computer Science, Boston, 1986.