An Approach to Evaluating the Completeness of a Medical Knowledge Base

Omar Bouhaddou, Eric Lepage, Homer Warner,
and Homer Warner, Jr.

Applied Informatics Inc. & Medical Informatics
Research Park & School of Medicine
University of Utah
295 Chipeta Way
Salt Lake City, Utah 84108

ABSTRACT

What is the list of Internal Medicine diseases? Which diseases in the list should be known by a student, a resident, an internist or a specialist? This papers reports on our approach to gathering the "complete" list of diseases in internal medicine and to categorizing in four groups the diseases to reflect the level of knowledge required of a student, a resident, an internist or a domain expert. This effort has provided us with a measure of how much an expert consultant program does "know" and what is the necessary coverage before the system is useful for each one of the four groups considered.

Introduction

What is the list of Internal Medicine diseases? Which diseases in the list should be known to a student, a resident, an internist or a specialist?

Medical diagnostic systems such as QMR (1, 2), DXplain (3, 4) and ILIAD (5) use knowledge frames to represent the relationship of each disease to all the disease manifestations. These systems can be used as an educational tool or as an expert diagnostic consultant. In the educational mode, the knowledge base of disease descriptions or disease frames offers multiple accesses and browsing capabilities to a well defined and structured database which include valuable quantitative information and selected literature references. In the consultation mode the database is used to generate a list of likely diseases given a patient case and hence, assists in differential diagnosis and cost conscious patient workup strategies.

The purpose of the QMR system is to emulate the performance of a practicing internist. The scope of its knowledge base is designed to cover what a good internist thinks a good internist ought to know. At present, QMR's knowledge base encompasses 572 diseases and 4,000 possible patient findings and the authors have identified 178 additional diseases yet to be covered in order to achieve "completeness" in internal medicine (2).

DXplain's purpose is to provide an extensive differential list to bring in light disease hypotheses that the user, principally an adult primary care provider, might not have thought of in the workup of his patient. The original set of diseases included in the DXplain's evolving database was developed based on the Current Medical Information and Terminology (CMIT) reference (5) and presently covers over 2000 diseases and 4000 disease manifestations (3, 4).

ILIAD is a medical educational tool that teaches differential diagnosis and cost conscious decision-making. ILIAD's knowledge base development aims to cover the list of disorders (diseases and disease presentations) included in medical schools clerkship syllabi.

As these systems, through the use of explicit models, are becoming exciting new pathways to the best of medical knowledge, it is natural to ask how much do they "know" or how much of internal medicine do they cover? How many diseases should they "know" before they are useful expert consultants and/or useful educational tools to a medical student, a resident, an internist or a domain specialist? Also, given an intended audience, which diseases are important to cover first?

0195-4210/89/0000/0110$01.00 © 1989 SCAMC, Inc.

In attempting to answer these later questions, it becomes necessary to investigate the two general questions stated at the opening of this paper.

This paper reports on our approach to gathering the "complete" list of diseases in internal medicine and to categorizing in four groups the diseases depending on the level of knowledge required of a student, a resident, an internist or a domain expert. This should provide some guidelines as to the coverage required of a consultant expert system for each one of these four groups.

Defining the list of diseases in internal medicine

The definition of the whole set of internal medicine diseases poses some fundamental problems. However, this work proposes a pragmatic approximation method based on multiple, complementary sources which include textbooks, manuals, the disease tree of the MESH dictionary and medical expert input.

Fundamental problems

The first fundamental problem was to identify a source that might provide an exhaustive listing of all the diseases in internal medicine. For this purpose, an investigation was conducted by contracting (mostly by telephone) local and distant experts, then continuing on with the local medical library, the National Library of Medicine, the American Association of Board Examination, the American Medical Association and, finally the American Association of Internal Medicine Specialities. It became apparent that no such a list existed and that it has to be built from combined sources.

Once a fairly reasonable set of potential literature resources has been identified, the task was to extract and enumerate the list of disease names. This posed the second fundamental problem related to the definition of what is a disease. In other words, what level of granularity a pathological entity is to be considered as a separate disease . For example, Genital/Oral/Encephalitic Herpes; Angina (disease or symptom of Coronary Artery Disease?); Acute Pulmonary Embolism and Pulmonary Thromboembolism.

The level of granularity enforced was often the lowest level leading to entities that were distinguishable based on their presentation, their treatment or their illustration of a different pathophysiological mechanism. For instance, Acute Pulmonary Embolism and Pulmonary Thromboembolism were considered as the same disease and Angina as a complex symptom.

Another problem relates to the overlapping of diseases between several domains as, for example, with AIDS or Tuberculosis. In this case, only one category was taken into account (AIDS was counted as an infectious disease).

Finally, the third fundamental problem was the definition of the boundaries of Internal Medicine. The domains excluded were surgery, obstetrics, pediatrics, ENT, ophthalmology. Some diseases in psychiatry were included as they could be presented to a non specialist (e.g., Depression).

Pragmatic approximate approaches

The definition of the list of diseases in internal medicine would have been impossible if definitive answers had to be found for the above fundamental problems. Therefore, practical approaches were considered realizing that the final list obtained would only be an approximation, albeit objective, in the absence of a concensus in the domain.

The practical approaches consist of defining the list of diseases in internal medicine by reference to the current medical literature, hospital case mix and the knowledge of human experts.

- **Using current medical information resources**

The different medical information sources used to compile a list of diseases in internal medicine include textbooks, manuals, disease classifications and the disease tree of the MeSH dictionary as defined by the National Library of Medicine. Below is a brief statement on the perceived appropriateness of each considered source to effectively support our goal.

The Merck manual 5th edition 1987 (11): the main pathologies are listed in the table of contents. They are described in a synthetic way in the text which will also include reference to less important diseases. The number of disorders covered is relatively extensive.

Harrison's 11th edition 1987 (7): Diseases are not listed in any exhaustive way. The emphasis is on the pathophysiology of the disease and the mechanisms of symptoms. The textbook represents more of a reading and learning source than a list of internal medicine diseases. However, the diseases discussed are usually important and illustrative of major medical principles.

MeSH 1988 (8): The coverage in MeSH is biased by what the community publishes. The needs for classification are predominant and therefore the MeSH disease category (C category) doesn't include an exhaustive listing of the disease entities themselves. The classification is difficult to follow, includes many redundancies and important diagnoses are missing. It has been a difficult source to use for the purpose of listing disease names in internal medicine.

CMIT 1981 (6): The list of disorders in CMIT is the most extensive of all with a very fine splitting of disease states. However, the entities do not always reflect "medically" relevant differences. The listing also includes symptoms, signs and intermediate states.

ICD-9 1980 (9): ICD-9 represents probably the most complete disease classification system of all the sources considered. However, the nomenclature is often coarse making the identification of disease entities difficult. It includes symptoms, signs and procedures and is not limited to internal medicine diseases. It also includes the NOS (not otherwise specified) and NEC (not elsewhere classified) categories.

Medical expert systems: QMR (1988) and ILIAD (1989) lists of diseases (covered or yet to be covered) were used to build the final list of diseases.

- **Using hospital case mix**

The cumulative list of diagnoses seen on a medicine service during several years was considered for this purpose. Using the HELP system patient database (10), both the LDS and University hospital lists of discharge diagnoses were generated. However, there are several problems inherent to this source with regard to building a complete list of diagnoses: 1) the "vanishing patient", an increasing number of diseases are now treated in the physician's office and therefore are not part of a hospital inpatient or outpatient case mix, 2) undiagnosed problems are not reported and, 3) given the status of a hospital (e.g, primary, secondary or tertiary) the case mix is likely to be biased and lacking an accurate representation of the diseases in the real world; 4) the list of discharge diagnoses is based on the ICD9 classification which has already been consulted and discussed above. However, looking at the hospital case mix has proven useful in evaluating the nature or type of diseases a medical knowledge base covers as will be discussed below.

- **Using domain experts**

The list obtained using the previous sources was reviewed and upgraded by at least two experts in each of the domain of internal medicine to enhance its completeness.

In summary, the combination of sources used provided a relatively exhaustive and objective list of diseases. To evaluate the nature of the coverage of a medical expert system (e.g., educational) we need now to assess the importance of each disease covered. However, the descriptive disease information these sources provided was not readily useful for this purpose. Hence, the need to develop a method to define the importance of a disease.

Defining the importance of a disease

In order to maximize the utility of an expert consultant system at each step of the development effort, it seems useful to cover the important concepts first. However, given an intended purpose for the system (e.g., education, expert consultation), the importance would be defined differently both from the breadth and the depth point of views.

This work is concerned with the breadth of a system and, therefore, four increments were defined. They are defined as the breadth of knowledge of a student, a resident, an internist and a domain specialist. The disease categorization should reflect the disease prevalence, the urgent attention it requires or the educational value it holds as illustrative of major pathophysiological principles.

With these guidelines, sublists of diseases were extracted and distributed to specialists in each subspecialty. At least, two experts were consulted to validate the results. A copy of the recommendations provided to each domain expert is shown below:

Thank you for contributing to the evaluation of ILIAD's knowledge base coverage.
The purpose of this evaluation is to help us answer the following questions:
-how much does ILIAD know today?
-what's left to be known in internal medicine?
-what priorities to set on the diseases not yet covered?

The attached list of disease names has been developed by 4th year medical students and myself using medical textbooks (e.g., HARRISON, THE MERCK MANUAL), the Current Medical Information and Terminology book (CMIT), the International Classification of Diseases (ICD9) and the MEdical Subject Headings (MeSH) as used in Medline by the National Library of Medicine.

The importance scores we're asking you to attach to each disease are defined as follow:
1 = the disease should be known by a medical student
2 = the disease is part of an intern or a resident's knowledge
3 = the disease is part of an internist's knowledge
4 = the disease is part of a domain expert knowledge

"known" is stated here as the ability to diagnose the disease. This importance score should reflect the high prevalence of the disease, the urgent attention it requires or the educational value it illustrates.

Also, the following considerations should be kept in mind while reviewing the list. Please
- mention synonyms to avoid double counting
- use the most common name for a disease
- eliminate names in the list that refer to a topic (class of diseases), and intermediate state (syndrome) or an observation (finding)
- eliminate diseases that are more appropriately part of a different domain
- consider deeper levels of granularity or more breakdowns of a disease or different stages of a disease only if medically, pedagogically meaningful (use your best judgement)
- add any missing important disease
- conclude with a qualifier describing the completeness of the list (e.g., complete, too many missing diseases in category 1,2, 3 and/or 4)

If you have any further question, please call me, Omar Bouhaddou, at Medical Informatics 581-4080 or at Applied Informatics 584-3062.
Thank you again.

It should be noted that this scoring scheme leads to inclusive sets, in the sense that, a disease known at one level is known at the higher levels of knowledge. Therefore, a score 1 is equivalent to a score 1 to 4, a score 2 to a score 2 to 4 and a score 3 to a score 3 to 4. For instance, if a disease is known to a student (category 1) then it is known to the resident, the internist and the domain specialist (category 2-4).

Results of the evaluation

The results are 1) a list of all internal medicine diseases as a union set representing the medical literature resources and experts' input, and 2) a categorization of the diseases in four groups reflecting a level of importance as discussed in the previous section.

DOMAIN	Imp 1	Imp 2	Imp 3	Imp 4	Total
Cardiovascular	63	33	25	15	136
Dermatology	11	38	15	30	94
Endocrinology	16	32	18	10	76
Environmental/poisoning	6	25	17	7	55
Gastrointestinal/Hepatic	50	42	19	56	167
Genitourinary	20	24	4	23	71
Hematology	60	24	36	10	130
Immunology/Allergy	5	2	8	4	19
Infectious Diseases	33	45	24	54	156
Metabolic/Nutrition	30	28	31	13	102
Musculoskeletal/Rheum	30	21	27	16	94
Neurology	54	29	19	2	104
Psychiatry	17	13	11	3	44
Pulmonary	28	22	37	28	115
Renal	14	31	33	41	119
Totals	437	409	324	312	1482

Table 1: Diseases count according to priority category and internal medicine domain.

Table 1 presents the total count of diseases in each internal medicine domain considered and each of the four importance categories. As, at least two experts looked at the list, it is necessary to define an "average" score. For our purpose, we are interested in identifying, in each domain, those diseases that both experts judged important and, hence, a maximum of the two scores was considered. This allows us to implement the critical mass first.

However, other choices are possible. But, in the context of this experiment, the main purpose of having a second expert was to verify the non-random nature of the scoring. Hence, the difference between the two scores was calculated and is presented in Table 2 to illustrate the concordance observed between the two experts.

Scores Difference	%
0	58
1	27
2	12
3	03

Table 2: Distribution of the difference obtained by subtraction of the two expert scores (1 to 4). This illustrates the non-random nature of the scores obtained.

Discussion

The final list obtained reflects our disease counting strategy, hence the total count listed in Table 1. It is difficult to compare this count to different counts derived from other lists (e.g., DXplain), as the level of granularity of a disease might be arguable. However, the list used by ILIAD represents a more exhaustive approach to the problem as it accounts for all disorders referenced in the combined set of medical literature sources, medical information systems and human expert knowledge.

The validity of the methodology developed to evaluate the importance of each disease is supported by different observations which became apparent as we began to analyse the scope of ILIAD's knowledge base.

As mentioned above, ILIAD is a teaching tool that aims at start to provide diagnostic assistance with those disorders included in medical schools syllabi. This pattern is confirmed when observing from Table 3 that most (67%) of ILIAD's diagnoses are in category "1".

DOMAIN	% of ILIAD Dx in "1"
Cardiovascular	0.96
Endocrinology	0.38
Environmental/poisoning	1.00
Gastrointestinal	0.63
Hematology	0.77
Immunology/Allergy	
Infectious Diseases	0.53
Metabolic/Nutrition	
Musculoskeletal/Rheum	0.87
Neurology	
Pulmonary	0.60
Renal	0.50
Totals	0.67

Table 3: Most (67%) of ILIAD diagnoses are in the category "1" defined by the experts (i.e., medical student knowledge). For example, of all cardiovascular diseases included in ILIAD's knowledge base, 96% are in category "1".

Also, ILIAD is more likely to cover the high prevalence diagnoses seen in a hospital. In other words, the higher the prevalence of a disease in a medical ward, the more likely it is included in the current ILIAD knowledge base (Table 4).

Diagnosis Frequency Count	Total	ILIAD coverage	%
over 85	3	3	100
30 to 60	6	5	83
20 to 30	14	8	57
10 to 20	40	19	48
6 to 10	35	15	43
3 to 6	98	44	45
Totals	196	94	48

Table 4: The more frequent a disease in a hospital setting, the more likely it is included in ILIAD knowledge base.

Hence, the analysis of the current ILIAD knowledge seems to support that the disease categorization defined by the domain experts reflects a high prevalence and an educational value associated with each disease.

Conclusion

Yes indeed, we recognize the approximations adopted in the construction of the list of diagnoses in internal medicine, however, this list attempts to objectively represent multiple sources including literature sources and experts input. As a result, the list is now available to estimate the breadth of medical expert systems. The pragmatic categorization of the diseases in four levels of knowledge seems to provide a reliable answer as to what is the critical mass to cover before a system is useful for a given audience. Given limited time and resources, this work has given us a measuring device against which we can monitor ILIAD's development and optimize its intended utility at any point in time.

Acknowledgements

This work wouldn't have been possible without the contribution of the 4th year medical students (Mark Ott, Bruce Newton and Rob Price) who helped build the list of internal medicine diseases and all the domain experts from the University of Utah School of Medicine who very generously agreed to review and categorize the list. The authors also acknowledge the useful feedback received from the experts contacted by phone.

References

1. Miller RA et al. Quick Medical Reference (QMR) for diagnostic assistance. M.D. Computing; May 1986; 3(5); 34-48

2. Miller RA et al. The Internist-I/Quick Medical Reference project - status report. The western journal of medicine December 1986; 145; 816-822

2. Barnett GO et al. DXplain: an evolving diagnostic decision-support system. JAMA 1987; 258(1); 67-74

3. Packer MS et al. Updating the DXplain database.
SCAMC 1988; 96-100

4. Warner HR et al. ILIAD as an expert consultant to teach differential diagnosis. SCAMC 1988; 371-376

5. Gordon B (ed). Current Medical Information and Terminology. 4th edition. American Medical Association, Chicago 1971

6. Finkel G. (ed). Current Medical Information and Terminology. 5th edition. American Medical Association, 1981

7. Braunwald E et al (eds). Harrison's principle of internal medicine.
MacGraw-Hill Book company;11th edition 1987

8. Medical Subject Headings - Tree structure, 1988
Library operations, National Library of Medicine, Bethesda, Maryland July 1987

9. The international classification of diseases 9th revision clinical modification (ICD-9-CM). Second edition; September 1980; DHHS publication