Comparing Complex Diagnoses: A Formative Evaluation of the Heart Disease Program

Hamish S F Fraser[1, 2], William J Long[1], Shapur Naimi[2]
[1]MIT Laboratory for Computer Science, Cambridge, MA,
[2]Tufts-New England Medical Center, Boston, MA

Abstract
To thoroughly evaluate the diagnostic accuracy of a decision support tool it is necessary to compare the diagnosis produced by that program on clinical cases with one or more standard sets of diagnoses. We are developing a framework for assessing the complex diagnostic summaries generated by the Heart Disease Program. The diagnoses used for comparison come from the physician entering the case, a "gold standard" assigned by patient chart review, and the opinions of expert cardiologists. The data collection, methods of comparison and example analyses are presented here. The many challenges in comparing these different diagnoses and minimizing bias are discussed.

INTRODUCTION

After more than three decades of inventive and high quality research in medical informatics, successful deployment of diagnostic programs is the exception rather than the rule. Reasons cited for this problem include technological barriers, user resistance to technology, variations in clinical practice and lack of common standards for exchanging data. However it is becoming increasingly evident that lack of a "user centered" approach to development, and inadequate on-site evaluation have contributed significantly to informatics implementation failures in many fields [1]. Good evaluation should provide important data on information needs of clinicians and the program design, as well as ensuring that the program performs as expected.

While user feedback should be available throughout the development and deployment process, the analogy of drug development suggests dividing the evaluation process into three stages:

1. Laboratory based measurements of accuracy, reliability and ease of use, using retrospectively collected data. (formative studies)
2. Prospective, observational studies of the system in realistic clinical settings (formative & summative studies)
3. Prospective, randomized intervention studies also in clinical settings. (summative studies)

This paper reports initial results from a prospective observational study of a cardiac diagnosis program (stage 2).

We have been developing the Heart Disease Program (HDP) for over a decade. It is designed to assist physicians with the diagnosis of all types of heart disease, particularly those leading to hemodynamic dysfunction and heart failure[2]. The program can be divided into 3 main components:

1. A user interface utilizing HTML forms created by PERL scripts.
2. The knowledge base and inference mechanisms (described below).
3. Mechanisms to summarize and explain diagnoses (which also generate HTML).

The diagnostic mechanism of the HDP is implemented as a network of nodes representing potential diseases and physiological states of the patient. The nodes are linked by probabilities, analogous to a Bayesian belief network. However nodes in the HDP can represent different severity levels of disease, and feedback loops are permitted. Mechanisms to reason about the time course of symptoms and diseases are incorporated. Clinical data sets are used to set up the prior probabilities in the network and instantiate the nodes. Diagnoses consisting of paths through the network are summarized to provide complete diagnostic hypotheses, ranked in order of probability. The HDP can provide a detailed analysis of complex cardiac cases including the underlying physiology. Though it is currently a "stand alone" system, incorporation into a hospital network is likely in the future. The details of the diagnostic algorithms and the construction of the Web interface are described at this meeting and elsewhere [2,5,6,7].

The output summary contains complete hypotheses about the patient including all the different possible diseases and physiological states. Each hypothesis includes explanations of how the clinical data items justify the various diagnoses. In addition there are a variable number of alternative hypotheses for each case. To assess the diagnostic accuracy of the program it is necessary to compare all the diagnostic items for each patient with some standard. Almost as important, the ability of physicians to use the program and successfully enter real cases has to be assessed. This includes training requirements, accuracy of data entry and speed of use. Finally the diagnostic summaries must clearly describe the case that was entered and the program's analysis.

PRIOR EVALUATIONS

The Heart Disease Program has undergone two previous evaluations in the laboratory setting (stage 1 above). In the first evaluation [3], cases collected by retrospective chart review were analyzed by the program. The case summaries and diagnoses were rated by cardiologists independent of the study. It was noted that cases where time course or severity of disease was important proved difficult for the program. The mechanisms for handling such cases were then added. In the second evaluation [4], a further set of cases were collected by chart review and analyzed by the revised program. The output diagnoses were divided into separate components, each of which was rated by two independent cardiologists using the scale "correct", "partially correct" and "wrong". Overall performance was rated as similar to the cardiologists in this controlled setting.

Having completed the initial (stage 1) studies, further testing was required to establish the performance of the program in realistic clinical settings (stage 2). All components of the program including interface, knowledge base, inference mechanisms and summarization were rigorously tested. Important issues regarding the input of data were uncovered, and are described in an accompanying paper. The main emphasis of this report is on development of tools to measure diagnostic accuracy on complex cardiac cases. independent cardiologists and final diagnosis/outcome from the hospital chart are then possible.

THE CLINICAL TRIAL

Case entry was by hospital physicians at the New England Medical Center. Each participating physician signed a consent form and received approximately 45 minutes training. They then selected cases according to the following clinical entry criteria:

- Breathlessness
- Peripheral edema or ascites
- Abnormal heart sounds or murmurs
- Any reasonable suspicion of heart failure

Case selection and entry were carried out entirely independently of the researchers. After each case was submitted to the diagnostic system, the physician completed another form detailing his/her diagnoses. Once that form was submitted and the program's analysis was returned (keeping the diagnoses from the program and the physician separate). Finally the participating physician filled in a critique form on the HDP's diagnoses. This entire process took between 15 and 25 minutes depending on the complexity of the case and the physician's experience with the program. To provide a comparison for the program's diagnosis each case was assigned a "Gold Standard" diagnosis from detailed review of the relevant discharge summary and/or patient chart. Particular attention was paid to definitive cardiac investigations such as echocardiography or cardiac catheterisation. Standard diagnostic criteria were used where applicable (e.g. WHO criteria for MI). Finally the case summaries generated by the program (without the diagnoses) were reviewed by one of several independent cardiologists who recorded their own diagnoses, to determine what diagnoses were possible with the data generated by the interface. All 4 sets of diagnoses were entered into a database using the same diagnostic coding as the HDP (table 1).

Table 1: The 4 Sets of Diagnoses and Relevant Input Summary Data Illustrating Analysis Methods

HDP Diagnosis	**Cardiologist Diagnosis**	**Physician's Diagnosis**	**Gold Standard Diagnosis**
Congestive-Failure Dilated-Cardiomyopathy Pulmonary-Embolism Hypertensive-Heart-Disease ***Anemia Score 0.5***	Congestive-Failure Dilated-Cardiomyopathy Pulmonary-Embolism Endocarditis ***Anemia Score 0.5*** Renal-Failure Anti-Arrhythmic-Toxic	Congestive-Failure Cor-Pulmonale	Runs-of-VT Dilated-Cardiomyopathy High-Degree-Block Pneumonia ***Hypertension Score 0*** Tricuspid-Regurgitation High-PA-Press
Total items = 4.5	**Total items = 6.5**	**Total items = 2.0**	**Total items = 6.0**
Items in input summary are down-graded as shown by the Score: Known Diagnosis = 0, Investigation result =0.5			

There are now over 200 codes for diagnoses and physiological states covering all main areas of cardiology. Diagnoses from the respiratory, renal and gastro-intestinal systems are included if they may complicate cardiac problems or have a similar presentation.The four sets of diagnoses were entered into the database with items in the clinical summary produced by the input interface checked to determine whether the final diagnoses would be biased. For example if the entering physician lists "Dilated Cardiomyopathy" as a "known diagnosis" then clearly that diagnosis cannot be used to analyze the program's performance. Diagnoses that were present in investigations such as mitral stenosis on echocardiogram, or anemia on a lab test (as in table 1), were also noted and down-rated in the analysis (currently they receive a 0.5 rating compared to the other diagnoses). Certain diagnoses are combinations of others such as "Cor Pulmonalae" which is "High PA Pressure" and "Chronic Obstructive Pulmonary Disease" together. In these cases a rating of 0.5 was given (rather than 1) for a match between the general term and the more specific one (table 1). A comparison was then made between the different sets of diagnoses. The *sensitivity* was calculated by counting how many diagnoses in the second list are present in the first list (tables 1&2). For example, if the comparison is between the Heart Disease Program (HDP) and the cardiologist, and the cardiologist has 6 diagnoses, 3 of which are present in the HDP diagnosis, and then sensitivity is 3/6 i.e. 0.5. If the HDP had 10 diagnoses for this case the *positive predictive value* (PPV) was 3/10 i.e. 0.3.

Table 2: Calculation of Sensitivity and PPV for the Case Shown In Table 1

Assessment of HDP			Assessment of Standard Dx		
HDP->Card	Sens	3.5/6.5	Card-> Gold	Sens	1/6
HDP->Card	PPV	3.5/4.5	Card-> Gold	PPV	1/6.5
HDP->Gold	Sens	1/6	Phys.->Gold	Sens	1/6
HDP->Gold	PPV	1/4.5	Phys.->Gold	PPV	1/2
HDP->Phys	Sens	1/2	Phys.->Card	Sens	1/6.5
HDP->Phys	PPV	1/4.5	Phys.->Card	PPV	1/2

Specificity is difficult to assess due to the large number of possible diagnoses in each case. Typically there are more than 185 true negatives and less than 10 false positives giving a *Specificity* of 95% or higher. PPV appears to be a more useful measure.

The program's diagnosis was compared to the cardiologist, the entering physician and the Gold standard. In addition the cardiologist and the entering physician were compared to the Gold standard. The cardiologists diagnoses help to separate the performance of the interface from that of the knowledge base and inference mechanisms, and indicate items in the Gold Standard that are difficult to predict from the input data.

When the physicians critiqued the program after the diagnosis had been returned they were asked to comment on which diagnostic elements were most relevant, which were not relevant, and what was missing. They were also asked how useful the program had been in that particular case. These results allow further assessment of the accuracy and usability of the program.

INITIAL RESULTS

Initial case collection is complete on 60 cases entered by physicians. During this phase of the study the program's knowledge base and inference mechanisms were fixed. The interface was also stabilized (except for minor adjustments to provide better explanation of inputs). Complete follow-up on 27 cases is currently available; follow-up of the whole cohort should be complete by fall of 1997. Table 3 shows the initial results of comparing the four different diagnoses.

Early in the evaluation the program occasionally failed to return a diagnosis. Usually conflicting data had been entered. These cases were re-run after correction. In addition certain cases gave inadequate diagnoses due to mistakes in patient data entry and/or interface problems. For example entering an echocardiogram but failing to indicate any abnormalities is taken by the program to indicate that the test was entirely normal, preventing certain diagnoses from appearing in the summary. Occasional problems of this sort were also dealt with by re-running the case. After the initial 60 cases the program was modified to prevent most of these problems. These are preliminary results and should not be seen as fully representative of the overall evaluation. It must also be noted that in assessing the performance of the HDP and doctors against the "Gold Standard" that the data on which the program is carrying out the analysis is much less than is available at the time of discharge.

Table 3: Comparison of the Four Diagnosis Lists for 27 Cases

Comparison	HDP -> Cardiologist	HDP -> Gold Standard	HDP -> Physician	Cardiologist-> Gold Standard	Physician-> Gold Standard
Sensitivity	0.43	0.46	0.62	0.34	0.19
PPV	0.33	0.37	0.24	0.41	0.61

Table 4: Repeat diagnosis comparison for 27 cases (10 cases rerun without "known diagnoses")

Comparison	HDP -> Cardiologist	HDP -> Gold Standard	HDP -> Physician	Cardiologist -> Gold Standard	Physician -> Gold Standard
Sensitivity	0.51	0.39	0.59	0.26	0.15
PPV	0.32	0.31	0.26	0.26	0.39

The values for *sensitivity* and *PPV* are therefore relatively low for all comparisons against the Gold Standard, but are comparable to those obtained in a recent study of the Medical diagnosis programs QMR, Dxplain, Iliad and Meditel, by Berner et al [8]. In that study they derived a measure equivalent to the *Sensitivity* used here (Comprehensiveness) which ranged from 0.25 – 0.38. A measure similar to *PPV* used here (Relevance in Berner's study) ranged from 0.19 – 0.37. This compares with a figure of 0.5 for the sensitivity and 0.31 for the PPV of the HDP taking the mean of the first 3 columns in table 3.

These initial cases included several where the entering physician had provided too much data and certain correctly diagnosed items could not be rated. Table 4 shows the effect of rerunning the program on these 10 cases without the known diagnoses and recent cardiac investigations. This provides a more realistic assessment of the program's diagnostic capabilities. The table also shows a comparison of the Entering Physician and the Cardiologist. As would be expected the closest agreement is between the HDP and the cardiologists. The sensitivity of only 51% is partly due to the tendency of cardiologists to focus on the most important problems then provide detailed differential diagnoses for those areas. The physician entering the case will typically give a smaller number of diagnoses, which may be incorrect due to the early stage in the patient's assessment. This leads to poorer sensitivity but better PPV for these physicians.

DISCUSSION

As with many studies of this sort, case collection proved the most difficult part of the entire evaluation. While we were reasonably successful in recruiting physicians interested in the study, the number of cases entered was low in the early stages. Getting physicians to enter clinical data is always difficult, particularly when much detail is required as in this study. Discussions with participating physicians indicate that these difficulties stem in part from lack of familiarity with computer systems, but particularly from the 15 to 20 minutes required to enter the case. In addition, we have specifically targeted hospital physicians, to simplify case follow-up and to ensure that the program is less likely to mislead the physician. However, hospital physicians have easy access to further investigations and expert colleagues and may see less need for decision support systems.

The requirement for physicians to enter their own cases meant that the input interface had to be as simple and as quick to use as possible. Ideally decision support tools should be used when there is partial data about the case and further information may be sought by the physician (but before definitive investigations). Giving physicians flexibility to enter cases in their own fashion is a powerful test of programs such as this. However it can lead to cases being entered with inadequate data. Alternatively too many "known diagnoses" can leave the clinical situation so over-specified that these input diagnoses have to be removed from the analysis as described earlier. Some of the initial data collection problems relate to the need to provide default values for data that is not fully described by the physician. A more interactive interface such as to one used by QMR could ensure higher quality input data however the time penalty may be large, QMR may require 1 to 4 hours for a case [9].

Variations in the amount and nature of data entered on a particular patient lead to other problems. These differences seem to be due in part to the nature of the problem and the setting where the patient is being seen. Out patients from the General Medicine clinic tend to present differently from hospital inpatients and often had fewer investigations. In addition the well-recognized variations in how doctors obtain and record clinical data cause many interesting problems. Physical examination is a particularly difficult issue as the program puts considerable weight on this data. The reporting of cardiac murmurs varies from the over-simple "systolic murmur" without qualification, to detailed and accurate descriptions of character, location, radiation, presence of thrill, and effect of maneuvers such as deep inspiration. This evaluation also highlighted the importance of time and duration of diseases in determining an accurate diagnosis in Cardiology. In several cases the precise timing of previous surgery and investigations caused conflicts that prevented accurate diagnoses being obtained. For example a patient with previous mitral and aortic valve replacement was also recorded as having an echocardiogram that indicated abnormalities of these valves. If the time is not stated the HDP now assumes that the test occurred before the valve surgery.

CONCLUSIONS

As this is a formative evaluation, experience from the first phase of the study has guided the improvements in the knowledge base and algorithms of the program which include providing better analysis of coronary artery disease risk factors. Collection of a second cohort of 70 patients is nearly complete. The assessment mechanisms detailed here will allow comparison between the old and new versions of the program, as well as providing overall measures of the program's effectiveness. We are in the process of automating the mechanisms for comparison of diagnoses.

An important extension to the method of diagnostic assessment described here is rating the importance or seriousness of a diagnosis. For example myocardial infarction is clearly a dangerous disease and missing it is much more serious than missing more benign or chronic problems. Measures of clinical importance have been developed to provide the appropriate ranking for diagnoses in the output summary. These factors are being incorporated into the comparison evaluation on the important diagnoses.

Ideally, cases should be chosen either at random, or consecutively to minimize bias. However, the difficulty in recruiting cases makes these more complex and sophisticated techniques somewhat premature. The challenge of allowing physicians to carry out the entire case selection and entry in typical clinical settings provides an unusually tough and realistic test compared to previous evaluations of this and many other programs [8,9]. It may also indicate the type of case where advice is sought. Once the program has a high degree of robustness and accuracy in this setting, it will be appropriate to carry out the third stage of evaluation; a full randomized controlled trial.

REFERENCES

1)Landauer T K, The Trouble with Computers, Usefulness, Usability and Productivity, MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02142. ISBN 0-262-12186-7
2) Long W. Medical Diagnosis Using a Probabilistic Causal Network. Applied Artificial Intelligence. 1989;3:367-83.
3) Long W J, Naimi S Criscitiello M G, Development of a Knowledge Base for Diagnostic Reasoning in Cardiology, Computers in Biomedical Research, 25: 292-311, 1992.
4) Long WJ, Naimi S, Criscitiello MG. Evaluation of a New Method for Cardiovascular Reasoning. Journal of the American Medical Informatics Association. 1994;1:127-141.
5) Long WJ, Naimi S, Criscitiello MG. Summarization of Complex Causal Diagnostic Hypotheses. In: Ozbolt, JG, ed. Symposium on Computer Applications in Medical Care. Washington,DC: Hanley & Belfus, 1994:970.
6) Fraser H S F, Long W J, Naimi S A, Testing a Heart Disease Diagnosis Program in a Practical Clinical Setting. AAAI symposium on Artificial Intelligence Applications in Health Care, Stanford, March 1996.
7) Long W J, Fraser H S F, Naimi S, A Web Interface for the Heart Disease Program, Proceedings of the 1996 AMIA annual fall symposium, October 26 to 30, 1996. Editor James J. Cimino MD.
8) Bankowitz, R A, NcNeil, M A, Challinor, S M, Parker, R C, et al. A Computer-Assisted Medical Diagnostic Consultation Service. Annals of Internal Medicine, 1989; 110:824-832.
9) Berner, E S, Webster, G D, Shugerman, A A, Jackson, J R, et al. Performance of Four Computer-Based Diagnostic Systems, NEJM, 330, No. 25, 1792-1796

Acknowledgments this study was approved by the New England Medical Center HIRB. We would like to thank Marvin Konstam and James Udelson for permission to use patient data, members of the medical resident and GMA staff for case entry, and John Wong, Peter Szolovits and James Stahl for their advice. The work was supported by the National Heart Lung and Blood Institute (R01-HL33041).