Differential Diagnoses of the Heart Disease Program have better Sensitivity than Resident Physicians

Hamish S F Fraser MRCP, MSc[1, 2], William J Long PhD[1], Shapur Naimi MD[2]
[1]Clinical Decision Making Group, MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, MA 02139. Email: hamish@medg.lcs.mit.edu
[2]Cardiology and Clinical Decision Making, Tufts-New England Medical Center, Boston, MA

Abstract

We describe a prospective clinical evaluation of a computer program to assist with the diagnosis of heart disease. The Heart Disease Program (HDP) is a large diagnostic program covering most areas of heart disease and some related areas of general medicine. The program's output is a set of differential diagnoses with explanations and it can be deployed in a clinical setting using a web interface. A framework for assessing the complex diagnostic summaries generated by the HDP was developed and the program's diagnostic accuracy in a clinical setting was assessed. The diagnoses used for comparison came from the physician entering the case, a "gold standard" assigned by review of patient charts and investigations, and the opinions of expert cardiologists. The data collection, methods of comparison, example analyses and results on 114 cases are presented here. The HDP had a significantly higher sensitivity for both the gold standard (60%) and the cardiologist's diagnoses (58%) than the physicians did (39%, 34%). These findings were consistent in the 2 collection cohorts and for the more serious diagnoses alone. The significance of these findings and the many challenges in comparing these different diagnoses and minimizing bias are discussed.

INTRODUCTION

During 15 years of development, the Heart Disease Program (HDP) has grown to cover most areas of cardiology and related areas of general medicine. It was designed to assist physicians with the diagnosis of heart disease, particularly those conditions leading to hemodynamic dysfunction and heart failure [1,2]. The program is unusual in its ability to reason about the timing and severity of diseases. It also generates detailed explanations of differential diagnoses indicating the clinical data items which support each diagnosis. These characteristics provide many practical benefits in dealing with complex cases but also create challenges for evaluation. The multiple diagnoses from the program must be compared with some standard set of diagnoses either from expert opinion or patient follow-up. The diagnoses should ideally be based on definitive investigations, and should take into account the clinical importance of different diagnoses (based for example on case fatality, treatability or difficulty in making the diagnosis).

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.
2. Prospective, observational studies of the system in realistic clinical settings
3. Prospective, randomized intervention studies also in clinical settings.

Results here are from a prospective observational study of the Heart Disease Program (stage 2).

The Heart Disease Program (HDP) 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 diseases, 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 (figure 1). The details of the diagnostic algorithms and the construction of the Web interface are described elsewhere [1,2,3].

ISCHEMIC CARDIOMYOPATHY
(caused by CORONARY HEART
DISEASE), treated with NITROGLYCERIN ,
ACE INHIBITOR and DIGITALIS causing
VENTRICULAR ECTOPY (known diagnosis)
(also caused by CORONARY HEART
DISEASE), treated with ANTI-ARRHYTHMIC,
but ambulatory electrocardiography no PVCs)
LOW CARDIAC OUTPUT causing
RIGHT HEART FAILURE, as indicated by
mild pedal edema, treated with FUROSEMIDE
HIGH-LA-PRESSURE as indicated by
PULMONARY-CONGESTION

Figure 1: the HDP's main diagnostic hypothesis

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. Also 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 clarity of the HDP’s diagnostic summaries must be assessed.

Prior Evaluations

The Heart Disease Program has undergone two previous evaluations in the laboratory setting (stage 1 above). In the first evaluation [1], 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 of the HDP was divided into separate diagnoses, 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).

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. The physicians (mainly medical residents) selected cases that had at least one of: breathlessness, peripheral edema or ascites, abnormal heart sounds or murmurs or any reasonable suspicion of heart failure.

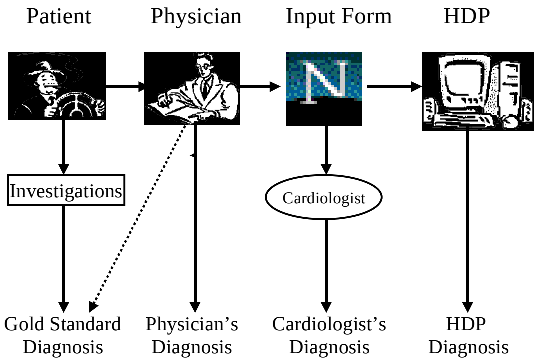

Figure 2: The origins of the 4 diagnostic groups

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 then 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.

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; At least one of cardiac catheter, echocardiogram, MUGA scan or thallium scan was performed in 99 of the 113 cases. Standard diagnostic criteria were used where applicable (e.g. WHO criteria for MI, valvular regurgitation had to be at least 2+ severity on echocardiogram or catheter). Finally the case summaries generated by the program were reviewed by one of several independent cardiologists (blinded to the diagnosis) who recorded their own differential diagnoses, to determine what was possible with the data generated by the interface. Figure 2 shows the origins and relationships between different diagnoses.

All 4 sets of diagnoses were entered into a database using the same diagnostic coding as the HDP. There are now over 200 codes for diagnoses and physiological states covering all main areas of cardiology. Diagnoses from the respiratory, renal and gastrointestinal systems are included if they may complicate cardiac problems or have a similar presentation. The process of matching diagnoses was complicated by the multiple terms used for similar physiological states of the cardiovascular system. For example "Left-Heart-Failure" is sometimes referred to as "Pulmonary-Edema" which is closely related to "High Left Atrial Pressure". A look-up table was therefore created with full matches for each term rated 1.0, and partially matching diagnoses, rated 0.5. Certain diagnoses are combinations of others such as “Cor Pulmonalae” which is “High Pulmonary Artery Pressure” and “Chronic Obstructive Pulmonary Disease” together. In these cases a rating of 0.5 was given for a match between the general term and the more specific one, and 1.0 for both parts. A severity score based on the clinical importance of each diagnosis (rated by 2 cardiologists) was also incorporated, allowing sub-analysis of diagnostic accuracy to determine performance on the more serious diagnoses.

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. 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. [5] *Specificity* is difficult to assess in this context 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 is a more useful measure in these circumstances though specificity could be calculated for any individual diagnoses which have sufficient data.

The diagnoses of the program, the physicians and the cardiologists were compared to the Gold standard. The HDP and physician were also compared to the cardiologist. The cardiologists diagnoses should help to separate the performance of the interface from that of the knowledge base and inference mechanisms (figure 2), and indicate items in the Gold Standard that are difficult to predict from the input data. In addition comparisons were made to the gold standard and the cardiologist's diagnoses using the HDP and physician's diagnoses combined (union of diagnostic items). After the diagnosis had been returned the physicians 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 for that case.

RESULTS

Clinical data on 127 patients were entered in 2 cohorts of 60 and 67. Full follow-up is available on 114, mean age 66 years (range 29 to 91), 40% female. The 13 patients without follow-up could not be traced from their identification number or did not have sufficient follow up data for analysis. 6 physicians and 4 cardiologists took part in the study. Figures 3&4 shows the results of comparing the three different diagnoses to the gold standard.

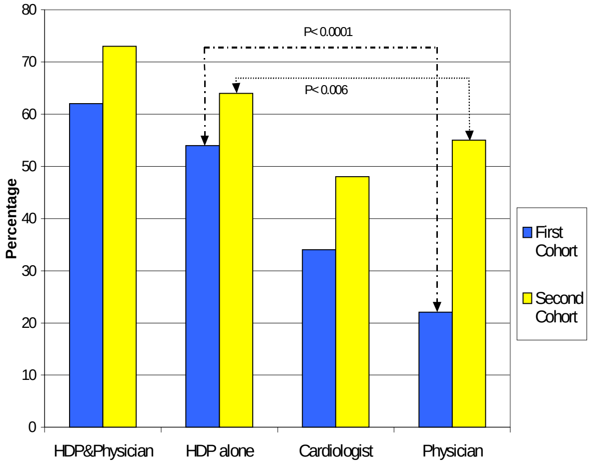

Figure 3: Sensitivity Compared to Gold Standard

Figure 4: Comparisons to the Gold Standard

	Sensitivity%	PPV%
HDP & Physician	68	32
HDP alone	60	30
Cardiologist	42	46
Physician	39	56

Figure 5: Comparisons to the cardiologists

	Sensitivity%	PPV%
HDP & Physician	64	29
HDP alone	58	33
Physician	34	48

The HDP alone and also in combination with the physician was significantly more sensitive at detecting Gold Standard diagnoses than the physician or the cardiologist (P< 0.001, Wilcoxon Signed Rank). The HDP was also significantly more sensitive than the physician at matching the cardiologist's diagnoses (P< 0.001). These comparisons were significant in both cohorts separately (P<0.005 or better) as well as combined.

Figure 6: Analysis of the most serious diagnoses compared to the Gold Standard

	Sensitivity%	PPV%
HDP & Physician	56	25
HDP alone	47	24
Cardiologist	44	38
Physician	38	43

Sub-analyses looking at the more serious diagnoses are shown in figure 6. The cardiologists and the physicians maintained their performance but the HDP lost some sensitivity. However it was still significantly more sensitive than the physician even when tested alone.

Figure 7: Physician's Critique

Suggests Investigations	17	19%
Confirms the physicians opinion	26	29%
Organizes the data	8	9%
Possibilities (suggested by HDP)	23	25%
No help	8	9%
Total responses	91/113	81%

Figure 7 shows the results of the physicians' critique from the first cohort. The 7 "no help" cases include 5 in which the program did not initially run due to technical problems. This only occurred twice in the second cohort. During the collection of each cohort 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).

DISCUSSION

These results show that the Heart Disease Program had a sensitivity significantly greater than the physicians entering the cases did when compared with the gold standard. The sensitivity for the program compared to the cardiologist's diagnoses was also greater than that of the physicians. Combining the diagnoses of the HDP and the physicians gave a further benefit. This is likely to be a realistic way of using the program's output, but intervention studies are required to confirm this. It will be noted that sensitivity for all diagnoses improved somewhat on the second cohort (figure 3), this may reflect the fall in mean number of diagnoses in the gold standard between cohorts from 5.5 to 4.5 (for unknown reasons). The physicians showed the largest increase in sensitivity, their mean number of diagnoses rising from 2.2 to 3.9 between cohorts. This is likely to be due in part to requests to participating physicians to provide a broader differential diagnosis. Whether this represents more typical behavior for physicians or was an improvement in clinical practice induced by the study is not clear. It should be noted in assessing the performance of the HDP and doctors against the "Gold Standard" that the maximum data is available to the Gold Standard followed by the physician, the HDP and the Cardiologist both having the least data (figure 2).

The values for *sensitivity* and *PPV* may be compared to those obtained in a study of the medical diagnosis programs QMR, Dxplain, Iliad and Meditel, by Berner et al [6]. In that study they derived a measure equivalent to the *Sensitivity* used here (Comprehensiveness) which ranged from 25 - 38%. A measure similar to *PPV* used here (Relevance) ranged from 19 - 37%. The PPV for the HDP is in the same range but the sensitivity is much higher. The lower sensitivities in Berner's study are no doubt due in part to the wider range of medical diagnoses covered by those programs. However the HDP had to contend with direct entry of data by physicians and more direct standards for comparison. The relatively low PPV of these programs may be offset in the case of the HDP by the detailed explanations given, allowing the user to assess the significance of each diagnosis. This dialogue between the physician and program is a key assumption in ethical approval of clinical support systems [9].

As with many studies of this sort, case collection proved the most difficult part of the evaluation. Expecting physicians to enter substantial quantities of clinical data is likely to be a barrier to use. Survey of participating physicians indicates that these difficulties stem in part from lack of familiarity with computer systems (less of a problem with the residents), but particularly from the average 14.2 minutes required to enter the case. In addition 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. We plan to extract data directly from the hospital information system in future and also to explore collecting part of the clinical history directly from the patient. Refinement of the interface may also lead to more efficient case entry.

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 or conflicting data. This lead to 5 cases failing to run without correction in the first cohort. The problem was virtually solved by interface upgrading in the second cohort. Alternatively too many "known diagnoses" can leave the clinical situation so well specified that meaningful measurement of diagnostic accuracy is not possible. This would be less of a problem if advice was sought by the physician for a difficult case, or in an intervention study. Some of the initial data collection problems related to the need to provide default values for data that is not fully described by the physicians. A more interactive interface such as used with QMR could ensure higher quality input data, however the time penalty may be large, QMR may require 1 to 4 hours for the entry of one case [7].

CONCLUSIONS

As this was a formative evaluation, experience from the first phase of the study guided the improvements in the knowledge base and algorithms of the program which included providing better analysis of coronary artery disease risk factors. There was a significant improvement in sensitivity with the second version of the program and also a smaller rise in sensitivity from 54 to 57 percent when the new program was run on the data from the first cohort (the second version of HDP also used some additional data items). The assessment mechanisms detailed here allow comparison between the performance of different versions of the program, guiding the development of a better knowledge base and algorithms. The assessment of clinical severity used here might be improved by the use of more objective measures such as case fatality or treatability.

In collecting the next cohort, cases should ideally be sampled either at random, or consecutively to minimize bias. However, the difficulty in recruiting cases makes these more complex and sophisticated techniques a significant barrier. A definitive study of the performance of the program will require a randomized intervention trial measuring changes in physicians' diagnoses and management decisions, and ultimately patient outcomes. Randomization of physicians or better still clinics is required to reduce biases [8]. Given the safety and cost implications of such studies it is clearly important that the performance of a decision support program is well-characterized in typical clinical settings first. This study provides much of the evidence required to demonstrate the suitability of the program for an intervention trial.

Acknowledgments This study was approved by the New England Medical Center HIRB. The work was supported by the National Heart Lung and Blood Institute grant R01-HL33041. Thanks to the cardiologists, medical residents and GMA staff at New England Medical Center for assistance with case collection, and John Wong, Peter Szolovits, James Stahl and Laura Smeaton for technical assistance and advice.

References

1. Long WJ, Naimi S, Criscitiello MG. Development of a Knowledge Base for Diagnostic Reasoning in Cardiology. Comp. Biomed. Res. 1992; 25: 292-311.
2. Long WJ, Fraser HSF, Naimi S. Reasoning Requirements for Diagnosis of Heart Disease. Artif Intell Med. 1997; 10(1): 5-24.
3. Long WJ, Fraser H S F, Naimi S, A Web Interface for the Heart Disease Program, Proceedings of the 1996 AMIA Fall symposium, October 26 to 30, 1996; 762-766
4. Long WJ, Naimi S, Criscitiello MG. Evaluation of a New Method for Cardiovascular Reasoning. J Am Med Inform Assoc. 1994; 1(2): 127-141.
5. Fraser HSF, Long WJ, Naimi S, Comparing Complex Diagnoses: A Formative Evaluation of the Heart Disease Program. Proc AMIA Annu Fall Symp. 1997; : 853
6. Berner, ES, Webster, GD, Shugerman, AA, Jackson, JR, et al. Performance of Four Computer-Based Diagnostic Systems, NEJM 1994; 330(25): 1792-1796.
7. Bankowitz, RA, NcNeil, MA, Challinor, SM, Parker, RC, et al. A Computer-Assisted Medical Diagnostic Consultation Service. Annals of Internal Medicine, 1989; 110:824-832.
8. Friedman CP, Wyatt JC, Evaluation Methods in Medical Informatics, Springer-Verlag, 1996.
9. Forsythe DE, Buchanan BG, Broadening our Approach to Evaluating Medical Information Systems, Proc 16th Annu Symp Comput App Med Care,1992: 8 - 12.