Decision Support for Concurrent Utilization Review Using a HELP-Embedded Expert System

Brent D. Nelson and Reed M. Gardner

Department of Medical Informatics, LDS Hospital/University of Utah, Salt Lake City, UT

ABSTRACT

Utilization Review is the process of evaluating the efficiency of medical care, based on examination of the patient record. At LDS Hospital, the electronic patient record is in an advanced state. This paper describes the development and knowledge base verification of ASSURE (Automated Support System for Utilization Review), an application within the HELP hospital information system. ASSURE applies the Appropriateness Evaluation Protocol (AEP) Day of Care criteria to the electronic patient record, concurrent with the patient's stay. In blinded trials, an experienced Utilization Manager agreed with 92% of ASSURE's decisions on single AEP criteria for 560 acute care patients. Agreement was statistically significant, with kappa = 0.84, P < 0.0001.

INTRODUCTION

One purpose of Utilization Review (UR) is to detect inappropriate medical care, i.e., care provided in an unnecessarily expensive or otherwise resource intensive setting [1]. For example, a stable inpatient receiving only IV therapy twice a day would generally be considered inappropriate, because such care is effectively provided by home nursing services at a lesser cost.

Studies have shown that 24% of inpatient days and 20% of hospital admissions are inappropriate [2]. In a 1990 pilot study by LDS Hospital Utilization Managers and an author (BDN), 23 of 268 inpatient days reviewed were inappropriate (or 8.6%, ± 3.4%), and 1 of 103 admissions reviewed was inappropriate (or 1%, ±2%). Of 206 inpatients reviewed, 21 had at least one inappropriate day (or 10.2%, ± 4.1%).

Inappropriate care is a major concern for hospitals, since reimbursement may be fixed and/or may be contested, reduced, or denied by payers [3]. Inappropriate care unnecessarily exposes patients to iatrogenic conditions, inconvenience, and financial stress [4]. UR is part of the larger effort to allocate limited resources more efficiently [5].

When coupled with feedback to clinicians and institutions, UR reduces hospital expenditures by 8% to 12%, at a cost:benefit ratio of 1:8 [6,7,8]. In addition to intervening in individual cases, UR influences institutional policies through identification of systematic problems [2,5,7].

The Appropriateness Evaluation Protocol (AEP) [9] is a diagnosis-independent criteria set for appropriateness of admission and day of care. Satisfaction of any AEP criterion indicates appropriateness; satisfaction of none indicates presumptive inappropriateness. The AEP was designed as a review guide, to be overridden when the user's professional judgement indicates an exception is warranted. The AEP has been validated in multi-center clinical trials, and is widely accepted. A study of UR criteria sets concluded that the AEP has moderate validity and reliability, but that payment should never be denied on the basis of the AEP alone [10]. That is, professional judgement is still essential. The AEP is in the public domain, and has been accepted as the standard of appropriateness by the Utah Peer Review Organization (UPRO). The following are sample AEP Day of Care criteria:

A. Medical Services
1. Procedure in operating room that day...
7. Close medical monitoring by physician, tid...

B. Nursing/Life Support Services
1. Respiratory therapy (respirator use or inhalation therapy with chest PT), tid...
6. Major surgical wound or drainage care...

C. Patient Condition
2. Transfusion due to blood loss, within 24 hrs...
7. Acute hematologic disorders yielding signs or symptoms, within 24 hours...

Study Environment

LDS Hospital is a 520 bed tertiary care hospital in Salt Lake City, Utah. The HELP hospital information system was originally developed there [11], and active development has continued since.

Three major factors combined to make LDS

0195-4210/92/$5.00 © 1994 AMIA, Inc.

Hospital the ideal site for ASSURE's development. These include: 1) the HELP system, with its integrated data and applications environment, powerful processing capabilities, and extensive online patient data base, the majority of which is captured in real time, 2) a research-oriented Medical Informatics department, and 3) a progressive, supportive Quality Resources department. Taken as a whole, this was seen as an unique opportunity to automate some aspects of Utilization Review, particularly screening of acute care patients for appropriateness of inpatient day. At the time of project conception, it was not known to what extent the online data would support such screening, but preliminary research suggested the project was both feasible and a promising expert system application [12].

Basic features of the HELP Hospital Information System

The HELP system runs on a network of 12 fault-tolerant Tandem RISC computers. Data enter the system from user-operated terminals, various instrument interfaces, and from other systems.

The HELP data base is patient centered, i.e., all patient data are keyed by patient number. The data base is central, and accessible to all HELP applications from all locations. Patient data are stored as time-stamped event "strings", with each string consisting of a variable number of coded data elements, known as PTXT (Pointer to Text) codes, which describe various aspects of the event described by the string. Some examples of event strings include drug orders, drug administration, nursing care and assessment, vital signs, laboratory test results, or surgery scheduling.

Most data in the HELP system are code/value pairs, where a numeric code identifying the specific data element, e.g., "CBC, White Cell Count", is paired with a value, e.g., "10.5". The value associated with the code is usually numeric, text, or time. Depending on the application, coded data may also be stored without an associated value, e.g., "Dyspneic, SOB" [13,14].

Other data, such as transcriptions of dictated reports, are stored in textual form. Some data are entered as freetext values attached to codes. In general, neither of these forms is useful for automated decision support systems.

HELP applications are well developed for Nurse Charting and Assessment, Pharmacy, Clinical & Blood Gas Laboratories, Blood Bank, Respiratory Therapy, Surgery, ADT, Medical Records, Radiology, and Microbiology [13,14].

A number of embedded decision support systems function routinely, taking advantage of the integrated database. Among these are adverse drug event surveillance [15], drug interactions [16], prophylactic antibiotic usage monitoring [17], hospital-acquired infection surveillance [17], antibiotic ordering support [18], clinician alerting to critical laboratory values [19], ARDS management protocols [20], and others [14].

No hospital information system as yet contains all clinical patient data. In a hospital with an information system, patient data is distributed among paper and electronic sources, posing a challenge to decision support applications. While some data are found only in the patient's paper chart, and are therefore accessible to clinicians only, others are found in both the electronic and paper charts. Since printed reports may omit or summarize data, some data are found in complete detail only in the electronic patient record.

THE ASSURE SYSTEM

The Automated Support System for Utilization Review, or ASSURE, is an expert system embedded within HELP that uses online patient data to detect inappropriate inpatient days, concurrent with the patient's stay, for adult acute care patients, using a current version of the AEP. Since the AEP actually detects appropriateness, inpatient days that cannot be documented as appropriate will be presumed inappropriate, and be referred to a Utilization Manager, who will perform a manual review and make any intervention required.

The ASSURE user interface allows Utilization Managers to review, accept, reject, or supplement the findings returned, to chart other UR data, manage a worklist, and generate reports.

The Knowledge Base

The ASSURE knowledge base is frame based. The AEP is implemented using 20 criterion frames and six auxiliary frames, which increase the modularity and efficiency of complex criterion frames, and/or carry out queries required by more than one criterion frame. All frames are Boolean. The inference engine uses a simple backward chaining algorithm, in which each AEP criterion is evaluated for the inpatient day being reviewed.

Each criterion frame finding data necessary for criterion satisfaction then searches for sufficient data to declare the criterion satisfied in fact. If sufficient data is found, an explanatory message is constructed, based on the retrieved patient data, indicating exactly how the patient satisfied the criterion (see Table 1). The explanatory message, labeled with the satisfied criterion, is placed in a data buffer for later retrieval by the inference engine, and the frame returns true. When all criterion frames have returned, the inference engine gathers up any explanations to pass to the user interface, and returns true if any of the criterion frames returned true.

Table 1. ASSURE explanations of how AEP appropriateness criteria are satisfied.

*** C4. Fever 38.3 C rectally (37.8 orally), if admitted for reason other than fever (24 hrs)**
EAR PROBE TEMP. 38.5 C; Time 03/28.13:10; ADMIT DIAGNOSIS: PNEUMONIA, ORGANISM UNSPECIFIED;

*** C6. Acute confusional state, not due to alcohol withdrawal (48 hrs)**
Not oriented to time; Not oriented to place; Short term memory not intact; Time 03/28.20:07; ADMIT DIAGNOSIS: PNEUMONIA, ORGANISM UNSPECIFIED;

*** C7. Acute hematologic disorders yielding signs or symptoms (24 hrs)**
Disorder: CBC, White Blood Count 16.5, Higher Than Normal 03/28.13:00;
Sign/Symptom: WBC high and increasing, up from CBC, White Blood Count 13.1 at 03/27/14:35;
Disorder: CBC, Platelet 545., Higher Than Normal 03/28.13:00;
Sign/Symptom: Tender, LUQ, 03/27.20:55;

Knowledge Engineering Process

Knowledge engineering for ASSURE was somewhat different for each AEP criterion implemented, but the basic steps were the following:

1. **Establishment of full criterion meaning** was based on the AEP criterion text and discussions with one or more Utilization Managers at LDSH and experts at UPRO. Current UPRO policies were taken into account to the extent possible.

2. **Identification of PTXT codes relevant to the criterion**, and currently in use, was done using data dictionary utilities, special purpose utilities, and consultations with applications programmers.

3. **Identification of PTXT code context** included matching clinical events with event strings and understanding the time-oriented aspects of the event strings. Identification of necessary codes, i.e., codes without which the criterion could not be met, and sufficient codes, i.e., codes that indicated that the criterion was met, was crucial.

4. **Criterion frame implementation** in PAL (PTXT Application Language) was a hypothesize-and-test cycle in which hypotheses regarding the data to be expected when the criterion was met were formed, programmed, tested against current patient records, and then iteratively refined.

5. **Training sets** were useful for difficult criteria, such as "Acute hematologic disorders yielding signs or symptoms, within 24 hours of the day reviewed." Criterion frame conclusions were compared with the judgement of an experienced RN Utilization Manager. When all available HELP data was being used optimally, the criterion frame was ready for formal verification.

PHASE I: KNOWLEDGE BASE VERIFICATION

Goals

The goals of ASSURE knowledge base verification were the following, in order of priority [21,22,23]:

1. To ensure that all available data were used to maximum advantage. In other words, to maximize the agreement achievable between the findings of the Utilization Manager, who made use of all available data sources (except ASSURE), and the findings of ASSURE, which used only coded HELP data. False positives, i.e., where the frame returned true when the patient in fact did not meet the criterion, were particularly to be avoided.

2. To detect and repair remaining systematic errors or programming bugs that survived into the test phase, so as to prevent their propagation into Phase II. However, this was not a comprehensive attempt to verify the correctness of all possible outputs. Therefore, the Phase I knowledge base was not absolutely frozen (see below).

3. To measure the agreement actually achieved between the Utilization Manager and ASSURE for each criterion. Measurement of agreement for the system as a whole will be done more rigorously in Phase II.

4. To identify sources of disagreement between the Utilization Manager and ASSURE.

Methods

Each of the twenty AEP criterion frames were separately verified using test patient sets, in which half of the patients met the criterion and half did not, according to the frame. Each day, a sampling program ran the criterion frame against current inpatients on the West 6, West 7, and West 8 acute care nursing divisions at LDS Hospital, examining the online data to see if each patient met the criterion on the day reviewed. The day reviewed was generally the day just prior to the day on which the sample was taken. Patients who had been in intensive care on the day reviewed were excluded. Positives (patients satisfying the criterion) and negatives (patients not satisfying the criterion) were sampled randomly in equal proportions.

For 17 of the 20 criterion frames, the planned sample size of 30 patients was obtained. Three frames had smaller sample sizes, due to a scarcity of positives and/or limited sampling time. Since the Utilization Manager had many other clinical duties, the test sample sets were acquired incrementally, generally at the rate of 12 or fewer patients per day. Frequently, multiple criterion frames were in verification phase simultaneously.

Data files from each sampling run were transferred to an IBM-compatible PC, where they were imported into a research database created using the Borland Paradox 3.5 relational database management system [24]. From the database, the day's test set patient list was generated, showing the AEP criterion being tested, the date reviewed, the names, numbers, and room numbers of the sampled patients, but no indication of the criterion frame results. The Utilization Manager took the list, examined each patient's chart, determined whether or not they met the indicated criterion on the indicated day, marked them "Y" or "N", and added any notes desired. All data sources but ASSURE were available to the Utilization Manager. For frames that made use of Nursing Notes and Assessments, the Utilization Manager was given printouts of these data covering the time period reviewed for each patient, instead of a patient list. These data were very rich in utilization-related information, but were so voluminous as to be time-consuming to review on the computer.

The Utilization Manager's findings were then entered into the Paradox database. If the Utilization Manager's findings disagreed with the criterion frame, a consultation was be held to make sure 1) we had the same understanding of the criterion, 2) ASSURE and the Utilization Manager had access to the same data sources, if possible, and 3), ASSURE and the Utilization Manager had both seen the same data items. These consultations were invaluable in ensuring that the criterion frames fit the real world.

Occasionally, the criterion frame was flawed, or the Utilization Manager had missed crucial data. Under certain circumstances, patients were dropped from test sets, or, after criterion frame revision, a test set became a training set and an independent test set was started. The following protocols reflect the verification goals, as well as the need to avoid both data dredging and dropping valuable outliers, and the need for efficient use of resources.

The protocol for dropping patients from test sets was the following:

1. If the patient had been discharged or the chart could not be located, the patient was dropped. For reasons of chart accessibility, only current inpatients were kept. A total of 38 patients were dropped for this reason, with a range of 1 to 16 patients per frame, over eight frames.

2. If a neatly dissectable, minor deficiency in the criterion frame was found, which could be fixed, the patient was dropped. After fixing the bug, correct output was verified for the entire test set. Where possible, other patients pertaining to the same subcase were sampled. A total of 10 patients were dropped for this reason, with a range of 1 to 4, over five frames.

Test sets were dropped or converted into training sets under the following circumstances:

1. If major errors in the working definition or implementation of the criterion were found.

2. If minor errors were widely diffused or errors affected frame output in ill-defined ways.

3. If significant, previously unused data sources were found that could be used by ASSURE.

4. If consultation with UPRO or the Utilization Manager indicated an erroneous understanding of the criterion.

5. If PTXT development in application areas rendered a previous verification set obsolete.

Analysis

Agreement between the Utilization Manager and ASSURE was measured using Cohen's kappa statistic, a classic measure of inter-rater reliability [25,26]. Kappa is an excellent measure of agreement, because it distinguishes between actual agreement and spurious agreement due to chance, and its significance is readily calculated. Kappa takes on continuous values between 0 and 1, where kappa=0 for agreement due to chance alone, and kappa=1 for perfect agreement. For each criterion frame, and for the Phase I study as a whole, kappa values were calculated using Stata software [27].

Results

Agreement. Results of the verification trials for individual criterion frames are shown in Table 2. Agreement was high and statistically significant for all criteria, though kappa values ranged from 1.00 down to 0.47. Over all frames in aggregate, a kappa of 0.84 was obtained, which is statistically significant, $P < 0.0001$, $Z = 19.8$. Overall agreement, where the Utilization Manager and a criterion frame made the same decision, was 92%.

Hit rates. For each criterion frame in Table 2, a hit rate is shown, which is the percentage of acute care patients meeting the criterion. Interestingly enough, all of the rare frames are highly reliable.

Sources of disagreement. The following sources of disagreement between the Utilization Manager and ASSURE frames were identified:

HELP data design or data capture deficiencies:

1. Crucial data were sometimes only available in human-readable forms. Occasionally, nurses entered freetext notes such as "V TACH" or "COMA", instead of selecting the coded equivalents from the charting menu. The "IV therapy tid" frame was highly reliable (kappa=0.82) on nursing floors with computerized drug administration data, and unreliable (kappa=0.15) on those without.
2. Frank data artifacts. For example, one patient had both "Oriented x 3" and "GLASGOW COMA SCORE 5" charted simultaneously.
3. Failure of personnel to carry out policies. Drug orders were not always discontinued and then reentered on patients transferred from ICU.
4. Ambiguous or unexpected uses of PTXT codes, such as the use of multiple codes for equivalent findings, were unusual but troublesome.

ASSURE knowledge engineering errors:

1. PTXT oversight. Set closure for relevant PTXT codes in current use was sometimes difficult.
2. Confusion over operative criteria definitions.
3. Interpreting ill-defined, subjective, or ambiguous patient states. E.g., the "Major wound care" frame was less reliable than the "Respiratory care" frame.

Utilization Manager error was nearly always data oversight, due to data overload. Nursing Notes and Assessments frequently amounted to 25 pages or more for a 48 hour period.

Table 2. Results of ASSURE criterion frame validation trials.

AEP criterion	n	kappa	$P \le$	Hit %
A. Medical Services				
OR procedure	31	1.00	0.0001	7.0
ER, OR next day	30	1.00	0.0001	0.18
Thoracentesis	30	0.93	0.0001	0.20
Experimental drug	3	1.00	0.0416	0.005
MD monitoring, tid	30	0.93	0.0001	6.3
Postop day	30	1.00	0.0001	10.
B. Nursing/Life Support				
Respiratory care	30	0.80	0.0001	4.3
IV therapy, tid	39	0.47	0.0017	29.
Continuous vitals	30	0.73	0.0001	16.
Injections tid	31	0.81	0.0001	9.5
Wound care	30	0.60	0.0005	17.
RN monitoring tid	30	0.93	0.0001	57.
C. Patient Condition				
GI/GU inability	30	0.93	0.0001	3.2
Transfusion	12	1.00	0.0003	0.58
V fib or ischemia	24	0.92	0.0001	0.53
Fever 37.8(O)	30	0.73	0.0001	11.
Coma 1 hour	30	0.93	0.0001	0.83
Acute confusion	30	0.80	0.0001	15.
Hematologic w/ SS	30	0.80	0.0001	36.
Acute neuro.	30	0.80	0.0001	3.0

2 X 2 Table for all criterion frames:

		Utilization Manager		
		Y	N	
ASSURE	Y	258	21	279
	N	24	257	281
		282	278	560

Overall results:
N=560
kappa=0.84
$P < 0.0001$
Agreement = (258+257)/560 = 92.0%

Conclusions

1. The LDSH HELP system database supports concurrent Utilization Review.
2. ASSURE produces a high level of agreement with a human expert under clinical conditions. The agreement is statistically significant.

PHASE II: VALIDATION OF ASSURE

Phase II, the test of the ASSURE system's performance in detection of inappropriate inpatient days, began in April 1993, and will be completed in July 1993. The sample will consist of 168 randomly sampled acute care patients from the West 6, West 7, and West 8 nursing divisions of LDS Hospital. Patients transferred from an ICU on the day reviewed will be excluded.

Purposes

1) To measure ASSURE performance in detection of inappropriate inpatient days. The kappa statistic will be used to measure agreement on detection of inappropriate inpatient days between ASSURE, a Utilization Manager using ASSURE output, and a second Utilization Manager using only other data sources. The Utilization Managers will be fully crossed with these two review methods. Patients will be their own control, being reviewed for the same inpatient day by all three methods. The knowledge base will be "frozen", i.e., unchanged, during Phase II.

A decrease in agreement due to the relatively higher hit rates for some of the less reliable criteria found in Phase I (see Table 2) may be offset by an increase in agreement due to patients meeting multiple criteria. However, introduction of a second Utilization Manager may produce more variability, as will criteria overrides. Overall, a small but significant decrease in agreement is anticipated. The agreement will be clinically useful if AEP criteria satisfaction is strongly associated with appropriateness of inpatient day.

2) To compare ASSURE and administratively initiated reviews in terms of sensitivity, specificity, and positive and negative predictive values. Administrative review initiation mechanisms include both insurance company requests and internal review policies. The comparison will be possible only for patients that are both sampled for Phase II and have an administratively initiated review. It is expected that ASSURE will be more sensitive and specific than administrative mechanisms, because of its power to search the clinical database for indicators of appropriateness.

3) To examine the effects of age, patient type, and payer type on the incidence of inappropriate inpatient days. Logistic regression will be used to measure predictive value for these variables. The patient set will be stratified for age, surgical versus medical patient type, and payer type.

DISCUSSION

The characteristics of the ASSURE project are well-suited to expert system implementation [12,28]. First, the standard of "truth", the AEP criteria set, has already been extensively studied and validated. The AEP, while not ideal, is better than competing criteria sets [10]. Second, the expected users are very similar to the expert involved in the knowledge engineering process. Thus, the users can reasonably be expected to exercise professional judgement when using the system, a prerequisite to beneficial outcomes [28]. Third, the moderate number of frames in the knowledge base and the two possible outcomes make the system maintainable. Fourth, treatment decisions will not be affected without review by a human expert. Fifth, ASSURE will be able to contribute significantly to utilization review and management without a level of cognitive performance equal to that of a human expert, since it is meant ultimately as an automated screening tool. Based on the Phase I results, which indicate that LDSH Utilization Managers are very likely to agree with ASSURE's findings with respect to individual AEP criteria, it is expected that a good level of agreement will be found for ASSURE's determination of inpatient day appropriateness.

In the long term, we believe that ASSURE will prove useful as a UR screening tool, by which patients found appropriate by ASSURE can reliably be considered appropriate without further review, and patients found inappropriate by ASSURE will be very likely to be found, upon manual review, to be inappropriate. The ultimate usefulness of ASSURE will be it's ability to review all acute care patients each day, and alert Utilization Managers to patients at high risk for being inappropriate.

[The authors are deeply indebted to Gail Hedrick, RN, who, tirelessly and expertly, both reviewed charts and consulted on the AEP criteria.]

References

[1] Payne SMC. Identifying and Managing Inappropriate Hospital Utilization: A Policy Synthesis. Health Serv Res, Dec 1987;22(5):709-769.

[2] Restuccia JD, Payne SMC, Lenhart G, Constantine HP, Fulton JP. Assessing the Appropriateness of Hospital Utilization to Improve Efficiency and Competitive Position. Health Care Man Rev, 1987;12(3):17-27.

[3] Friedman E. Hospital Uncompensated Care:

Crisis?. JAMA, Dec 1, 1989;262(21):2975-2977.

[4] Franks P, Clancy CM, Nutting PA. Gatekeeping Revisited - Protecting Patients from Overtreatment. N Engl J Med, Aug 6, 1992;327(6):424-429.

[5] Rosenstein AH. Utilization Review: Health Economics and Cost-Effective Resource Management. Qual Assur Util Rev, Fall 1991;6(3):85-90.

[6] Wickizer TM, Wheeler JRC, Feldstein PJ. Does Utilization Review Reduce Unnecessary Hospital Care and Contain Costs? Med Care, June 1989;27(6):632-647.

[7] Wickizer TM, Feldstein PJ, Wheeler JRC, McDonald MC. Reducing Hospital Use and Expenditures Through Utilization Review. Qual Assur Util Rev, Aug 1990;5(3):80-85.

[8] Feldstein PJ, Wickizer TM, Wheeler JRC. Private cost containment: The Effects of Utilization Review Programs on Health Care Use and Expenditures. N Engl J Med, May 19, 1988;318(20):1310-1314.

[9] Gertman PM, Restuccia JD. The Appropriateness Evaluation Protocol: A Technique for Assessing Unnecessary Days of Hospital Care. Med Care, Aug 1981;19(8):855-871.

[10] Strumwasser I, Paranjpe NV, Ronis DL, Share D, Sell LJ. Reliability and Validity of Utilization Review Criteria. Med Care, Feb 1990;28(2):95-109.

[11] Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP System. J Med Sys, 1983;7:87-102.

[12] Beckman TJ. Selecting Expert-Systems Applications. AI Expert, Feb 1991;:42-48.

[13] Pryor TA. The HELP Medical Record System. MD Comp, 1988;5(5):22-33.

[14] Kuperman GJ, Gardner RM, Pryor TA. HELP: A Dynamic Hospital Information System. New York: Springer-Verlag Inc., 1991.

[15] Evans RS, Pestotnik SL, Classen DC, Bass SB, Burke JP. Prevention of Adverse Drug Events through Computerized Surveillance. SCAMC, 1992; 16:437-441.

[16] Gardner RM, Hulse RK, Larsen KG. Assessing the effectiveness of a computerized pharmacy system. SCAMC, 1990;14:668-672.

[17] Evans RS. The HELP system: A Review of Clinical Applications in Infectious Diseases and Antibiotic Use. MD Comp, 1991;8(5):282-288.

[18] Evans RS, Pestotnik SL, Classen DC, Burke JP. Development of an automated antibiotic consultant. MD Comp, 1993;10(1):17-22.

[19] Bradshaw KE, Gardner RM, Pryor TA. Development of a Computerized Laboratory Alerting system. Comp Biomed Res, 1989;22:575-587.

[20] Sittig DF, Pace NL, Gardner RM, et al. Implementation of a Computerized Patient Advice System Using the HELP Computerized Hospital Information System. Comp Biomed Res, 1989;22:474-487.

[21] Berry DC, Hart AE. Evaluating Expert Systems. Expert Systems, Nov 1990;7(4):199-207.

[22] Miller PL. The Evaluation of Artificial Intelligence Systems in Medicine. Comp Meth Prog Biomed, 1986;22:5-11.

[23] Miller PL, Sittig DF. The Evaluation of Clinical Decision Support Systems: What is Necessary Versus What is Interesting. Med Infor, 1990; 15(3):185-190.

[24] Paradox Relational Database, Version 3.5, 1990. Borland International, 1800 Green Hills Road, P.O. Box 660001, Scotts Valley, CA 95067.

[25] Siegel S, Castellan NJ. Non-Parametric Statistics for the Behavioral Sciences. 2nd Ed., New York, McGraw-Hill, 1988.

[26] Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.

[27] Stata Statistics/Data Analysis, Version 3.0, 1992. Computing Resource Center, 1640 Fifth St., Santa Monica, CA 90401

[28] Schoolman, HM. Obligations of the Expert System Builder: Meeting the Needs of the User. MD Comp, 1991;8(5):316-321.