LIST OF PEER REVIEWED PUBLISHED ARTICLES
2009

1. HOW ACCURATELY DOES A WEB-BASED DIAGNOSIS DECISION SUPPORT SYSTEM SUGGEST THE CORRECT DIAGNOSIS IN ADULT AND PEDIATRIC CASES?


2. WHAT IS THE IMPACT ON CHANGE OF DIAGNOSIS DECISION MAKING BEFORE AND AFTER USING A DIAGNOSIS DECISION SUPPORT SYSTEM?


3. HOW ACCURATELY DOES A WEB-BASED DIAGNOSIS DECISION SUPPORT SYSTEM SUGGEST THE CORRECT DIAGNOSIS IN 50 ADULT NEJM CPC CASES?


4. HOW DOES A WEB-BASED DIAGNOSIS DECISION SUPPORT SYSTEM IMPROVE DIAGNOSIS ACCURACY IN A CRITICAL CARE SETTING?

LIST OF PEER REVIEWED PUBLISHED ARTICLES
2009

5. WHAT IS THE MAGNITUDE OF DIAGNOSIS ERROR?

6. WHAT ARE THE PRIMARY & SECONDARY OUTCOMES OF DIAGNOSTIC PROCESS - THE LEADING PHASE OF WORK IN PATIENT CARE MANAGEMENT PROBLEMS IN EMERGENCY DEPARTMENT?

7. WHAT ARE THE CAUSES OF MISSED AND DELAYED DIAGNOSES IN THE AMBULATORY SETTING?

8. WHAT ARE THE FACTORS THAT CONTRIBUTE TO DIAGNOSIS ERROR?

9. DO PHYSICIANS KNOW WHEN THEIR DIAGNOSES ARE CORRECT?

-------------------------------------------------------------------------------------------------------------------------------
HOW ACCURATELY DOES A WEB-BASED DIAGNOSIS DECISION SUPPORT SYSTEM SUGGEST THE CORRECT DIAGNOSIS IN ADULT ER AND PEDIATRIC CASES?

Validation of a diagnostic reminder system in emergency medicine: a multi-centre study

Padmanabhan Ramnarayan, Natalie Cronje, Ruth Brown, Rupert Negus, Bill Coode, Philip Moss, Taj Hassan, Wayne Hamer, Joseph Britto

Background: Diagnostic error is a significant problem in emergency medicine, where initial clinical assessment and decision making is often based on incomplete clinical information. Traditional computerised diagnostic systems have been of limited use in the acute setting, mainly due to the need for lengthy system consultation. We evaluated a novel web-based reminder system, which provides rapid diagnostic advice to users based on free text search terms.

Methods: Clinical data collected from patients presenting to three emergency departments with acute medical problems were entered into the diagnostic system. The displayed results were assessed against the final discharge diagnoses for patients who were admitted to hospital (diagnostic accuracy) and against a set of “appropriate” diagnoses for each case provided by an expert panel (potential utility).

Results: Data were collected from 594 patients (53.4% of screened attendances). Mean age was 49.4 years (95% CI 47.7 to 51.1) and the majority had significant past illnesses. Most were assessed first by junior doctors (70%) and 266/594 (44.6%) were admitted to hospital. Overall, the diagnostic system displayed the final discharge diagnosis in 95% of inpatients and 90% of “must-not-miss” diagnoses suggested by the expert panel. The discharge diagnosis appeared within the first 10 suggestions in 78% of cases.

Conclusions: The Isabel diagnostic aid has been shown to be of potential use in reminding junior doctors of key diagnoses in the emergency department. The effects of its widespread use on decision making and diagnostic error can be clarified by evaluating its impact on routine clinical decision making.
its knowledge base (textbooks and journal articles), variations in natural language textual patterns in the medical sources used for the adult system may significantly influence its diagnostic suggestions.

METHODS
This preliminary validation study was designed purely to examine the clinical performance of the Isabel system and identify its potential utility in adult EDs, not to assess its impact on clinical practice or diagnostic errors. Therefore, clinicians were not allowed real-time access to the system during patient assessment in this study. Clinical data from ED patients presenting with a range of acute medical problems were used to validate the results of the DDSS. The study was approved by the London multi-regional ethics committee (04/MREC02/41) and relevant local research governance committees.

Study centres
A convenience sample of four UK EDs were selected for data collection during the study. Due to significant delay in completing research governance procedures at one centre, data were finally only collected from three participating sites. The characteristics of the study sites are summarised in table 1.

Study patient data
Data were collected from all consecutive patients over 16 years old presenting to the “majors area” or resuscitation rooms with an acute medical problem. Patients presenting to “minors” or a similar area, patients with surgical complaints (including trauma, orthopaedics, ENT, ophthalmology and gynaecology), post-operative surgical problems, psychiatric problems (including substance and alcohol abuse) and complaints directly related to pregnancy were excluded. Patients presenting for reassessment of the same clinical problem within a week of an earlier visit to the ED were also excluded.

Study procedure
This study was a prospective, multi-centre observational study utilising medical case note review. No interventions were performed on patients.

Screening for eligibility
The attendance records of all consecutive patients presenting to an ED “majors area” within a pre-designated 2-week period were screened for eligibility by the primary research assistant (RA). Screening was performed at study sites one after the other, ie, the 2-week period was different for each centre. The complete medical and nursing notes of patients with presenting complaints that fitted study criteria were selected for further review. Reasons for exclusion were recorded for the remainder using specific codes established a priori. Where the RA was unsure of study eligibility, patients were included and data were collected. Such notes were reviewed at regular intervals by the principal investigator, and a final decision regarding study eligibility was made. To assess inter-rater reliability during screening for eligibility, a second RA examined patient notes from two randomly chosen dates within the specified 2-week period at each study ED. The primary RA was blinded to these dates. Concordance between the two RAs was calculated using the kappa statistic (κ 0.65, 95% CI 0.62 to 0.68).

Data collection
From eligible patient notes, the primary RA extracted data regarding patient details; date and time of patient assessment; details of clinical presentation such as symptoms, past medical and family history; examination findings, tests performed and results available at the end of complete assessment by the first examining clinician; differential diagnosis and management plan of the first examining clinician; referral for specialist or senior opinion; and outcome of ED assessment. This was entered directly into an Access database (Microsoft, Reading, UK) by means of pre-designed electronic forms. During data entry, both positive and negative symptoms, signs and test results were collected as recorded in the patient notes. Following complete data collection at all centres, final discharge diagnoses for patients at the end of ED assessment, recorded either on discharge letters or on ED electronic systems, were ascertained. For patients admitted as inpatients, final primary diagnoses at hospital discharge were obtained from hospital electronic coding systems.

Data quality assurance was achieved by multiple means including: use of a training manual created before study commencement containing standardised case examples to practise screening, data collection and abstraction; the use of 25 medical records to practise data collection, with doubts being clarified by the study investigator at study outset; weekly meetings to discuss data abstraction issues and examine collected data; and a log of discussions for ready reference. Reliability of the data collection process was established by randomly assigning 50% of eligible patient notes from two randomly chosen dates to a second RA. Concordance for key variables was analysed using the kappa statistic (κ 0.58, 95% CI 0.52 to 0.64).

Expert panel
An expert panel was set up at each study site consisting of two consultants (attending physicians), which met regularly to provide gold standard diagnoses for a randomly selected subset of study patients. At each panel meeting, moderated by the primary RA, data collected during initial ED assessment for each patient were provided in the form of a pre-formatted clinical summary report generated from the Access database (table 2).

Presenting clinical symptoms were provided as recorded in the patient notes. Only relevant co-morbidities, family history, positive clinical signs and results of initial tests (if performed by the examining clinician) were provided. The panel were blinded...
to the ED where the patient was assessed, clinical decisions made and eventual patient outcome. The panel were instructed to provide, for each case following mutual consensus, a set of “must-not-miss” diagnoses, defined as key diagnoses that would influence patient management, ie, result in a specific action, either eliciting further history or physical examination or initiating new tests and/or treatments. To establish concordance between the three expert panels, a random selection of 5% of study patients was assigned to each panel in blinded fashion. For this subset, gold standard diagnoses were defined as those suggested by two or more panels.

**DDSS data entry**

Concurrent with the data collection process, an Isabel prototype was created from the original paediatric version. In order to generate a mature DDSS for use in adult patients, this prototype needed refinement. A total of 130 notes randomly drawn from patients not admitted to hospital were used for this final fine-tuning (development set). In an iterative process, the system’s results for each case were critically examined by clinicians in the development team. To generate a focussed list of diagnoses, each diagnosis in the Isabel database had to be tagged to particular age group(s), gender and regions in which the disease was commonly seen (paediatric tags could not be used for adult patients). Once a fully developed Isabel system was available, the remainder of the cases (464/594) were used to test its performance (validation set).

During the validation stage, the RA exported the clinical summary report (as presented to the panel) for each case from the Access database in the form of individual text files. Since the Isabel system accepted search terms only as text and negative findings could not be searched within its medical content, information from the patient summary report needed modification by the RA during data entry into Isabel. Patient characteristics (eg, age and gender) were input using a drop-down menu, numerical values from vital signs and test results were converted...
to text terms (e.g., temperature $38.8\,^\circ C$ into “fever”), and only positive findings and test results (when available) were entered. This procedure was standardised prior to data entry by establishing normal ranges for vital signs and test results. The patient summary was aggregated into a single block of text with each symptom, positive clinical finding, salient past illness and the result of each test placed on a separate line, and then pasted into the search box. Any amount of clinical information could be entered, although it was accepted that specificity of search results would improve with detailed data entry. Figure 1 illustrates this procedure. In fig 2, results as displayed in Isabel are shown.

### Outcome measures

Two separate outcome measures were assessed. Diagnostic accuracy was used to provide an indication of the system’s clinical performance and was defined as the proportion of cases admitted to hospital in which the final discharge diagnosis appeared among the DDSS suggestions. However, this did not provide an indication of the system’s utility in terms of guiding appropriate patient management in clinical practice. Therefore, the proportion of cases in which the DDSS included the entire set of key diagnoses that were deemed as “must-not-miss” by the expert panel indicated its utility in an ED setting.

---

#### Table 3  Breakdown of patients screened and reasons for exclusion

<table>
<thead>
<tr>
<th></th>
<th>Centre A</th>
<th>Centre B</th>
<th>Centre C</th>
<th>Total</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient notes screened</td>
<td>344</td>
<td>403</td>
<td>366</td>
<td>1113</td>
<td></td>
</tr>
<tr>
<td>Eligible patient notes (%)</td>
<td>200 (58.1)</td>
<td>194 (48.1)</td>
<td>200 (54.6)</td>
<td>594</td>
<td>0.02</td>
</tr>
<tr>
<td>Exclusions (%)</td>
<td>144</td>
<td>209</td>
<td>166</td>
<td>519</td>
<td></td>
</tr>
<tr>
<td>Trauma</td>
<td>50 (34.7)</td>
<td>96 (45.9)</td>
<td>47 (28.3)</td>
<td>193</td>
<td>0.5</td>
</tr>
<tr>
<td>Surgical/post-surgical</td>
<td>15 (10.4)</td>
<td>10 (4.8)</td>
<td>20 (12.0)</td>
<td>45</td>
<td>0.03</td>
</tr>
<tr>
<td>Pregnancy-related</td>
<td>18 (12.5)</td>
<td>14 (6.7)</td>
<td>25 (15.0)</td>
<td>57</td>
<td>0.03</td>
</tr>
<tr>
<td>Psychiatric/psychosomatic</td>
<td>13 (9.0)</td>
<td>12 (5.7)</td>
<td>12 (7.2)</td>
<td>37</td>
<td>0.5</td>
</tr>
<tr>
<td>Substance abuse</td>
<td>15 (10.4)</td>
<td>23 (11.0)</td>
<td>21 (12.6)</td>
<td>59</td>
<td>0.8</td>
</tr>
<tr>
<td>Orthopaedic</td>
<td>2 (1.4)</td>
<td>19 (9.1)</td>
<td>5 (3.0)</td>
<td>26</td>
<td>0.001</td>
</tr>
<tr>
<td>Incomplete data</td>
<td>31 (21.5)</td>
<td>35 (16.7)</td>
<td>36 (21.7)</td>
<td>102</td>
<td>0.4</td>
</tr>
</tbody>
</table>

GP, general practitioner; NR, not recorded; OP, out patient.
Sample size and statistical analysis
Using an estimated diagnostic accuracy of 85% (from the paediatric system) and an acceptable error rate of 3.5%, data were required from 450 patients to ensure adequate power. Differences between study EDs were analysed using the $\chi^2$ test for proportions and ANOVA for continuous variables. Statistical significance was set at $p$ value <0.05.

RESULTS
During the study period, 1113 consecutive patient notes were screened for study eligibility at the three centres. A total of 489 patients were excluded by the RA and a further 30 patients were judged as ineligible on secondary review (overall 46.6%). Therefore, 594 medical notes were reviewed in detail. There were significant differences between centres with respect to the proportion of patient notes excluded and reasons for exclusion (table 3).

Patient characteristics are summarised in table 4. The mean age of patients included in the study was 49.4 years (95% CI 47.7 to 51.1) with an equal male to female ratio. A significant number of eligible patients were brought into EDs by ambulance (49.2%), and the majority were triaged into level 3 (55%). Most patients were seen by an SHO in the first instance (74.9%), and 40% were seen out of working hours (1800–0800). The majority of patients had past medical illnesses of note (91.4%). In addition, significant family history was present in a number of patients (12%). The primary examining clinician indicated having sought senior opinion in a number of patients (18%). The majority of patients had significant past medical illnesses of note (91.4%). In addition, significant family history was present in a number of patients (12%). The primary examining clinician indicated having sought senior opinion in a number of patients (18%).

Diagnostic accuracy was measured using 217 inpatient discharge diagnoses available from 266 admissions. Overall, the DDSS displayed 206/217 diagnoses, with an accuracy rate of 95% (CI 92% to 98%). Seventy eight per cent of the discharge diagnoses were displayed on the first page (first ten suggestions).

Panel members examined 129 cases to provide “must-not-miss” diagnoses. A total of 30 cases were assessed by all three expert panels and 99 others by a single panel. In the former set, 52 diagnoses were suggested (mean 1.7 per case). Isabel displayed 50/52 suggestions (96%) in the list of its diagnostic reminders, with the majority present on the first page (35/50, 70%). For the latter set, 100 “must-not-miss” diagnoses were provided by the panel (mean 1 per case). Isabel displayed 90/100 suggestions; 53/100 were present on the first page (first 10 reminders). Comprehensiveness improved significantly when all three pages were examined (fig 3). An example of one expert panel’s assessment of a case and relevant “must-not-miss” diagnoses is shown in table 5.

DISCUSSION
We have demonstrated in this large validation study that the Isabel diagnostic aid demonstrates significant accuracy for hospital discharge diagnosis among patients presenting to EDs with acute medical problems. The system also showed potential utility in the ED setting by including all key diagnoses among its diagnostic reminders.

The Isabel system has been previously evaluated in acute paediatrics, and shown to display the final diagnosis on the first page in >90% cases drawn from real life practice. The heterogeneous nature of acute presentations among adult patients and their lengthy past history may have resulted in more complex clinical data entry accounting for some decrease in accuracy in this study. Most adult patients had significant past medical illnesses and were on numerous medications. In addition, textual descriptions of diseases in the current Isabel knowledge base may be quite different between adult and paediatric sources resulting in a poorer match between the clinical data entered and the diagnostic results generated. Despite these findings, Isabel’s diagnostic performance is better than that of previously described systems.
DDSS. VanScoy et al showed in an ED setting using two diagnostic systems, QMR and ILIAD, that the final diagnosis was present in the top five choices in only 30% of cases. Data entry time was prolonged since detailed case descriptions needed to be matched to system-specific terminology and specific training was required in the use of the DDSS. In our study, clinical assessment was entered into Isabel in the examining clinicians’ own words in free text, leading to the conclusion that rapid and easy use of the system is possible without prolonged data entry. In this context, Graber et al have also recently shown that merely pasting the entire history and physical examination section from the Case Records of the Massachusetts General Hospital series as text without any modification into Isabel resulted in a diagnostic accuracy rate of 74%. Identifying the precise patient population in which DDSS might prove useful is a major challenge. Expert systems such as QMR were intended to be used in a diagnostic dilemma. However, there is sufficient evidence that diagnostic errors occur during routine practice, and that there is poor correlation between physicians’ diagnostic accuracy and their own perception of the need for diagnostic assistance. We chose to focus on validating Isabel against a pre-selected subset of ED patients (acute medical problems seen in the “majors area”). Nearly half of all patients screened qualified using our liberal study criteria, although it is improbable that in practice medical staff would have used Isabel in all these patients. Experience from our paediatric clinical study indicates that clinicians may seek diagnostic advice in 5–7% of acute medical presentations (three to five ED patients per day). Identifying the optimal parameter against which DDSS performance can be measured also remains controversial. We used hospital discharge diagnoses and an expert panel’s opinion of key diagnoses to provide a combined view of system accuracy and utility. During this initial evaluation, we deliberately denied clinicians access to Isabel. Yet, by extrapolation from these results, it seems likely that clinicians will use and benefit from its diagnostic advice in situations of uncertainty, especially since minimal data entry time was required. Integration of the DDSS into an electronic medical record may allow active diagnostic advice to be delivered to staff with minimal effort. Such an interface has been developed recently.

LIMITATIONS

The main limitation of this study was the fact that users did not interact with the system making it difficult to estimate its true utility in practice. The impact of a DDSS is best measured by its ability to improve clinicians’ diagnostic assessment; in addition, unexpected negative effects might be seen. We used electronic systems or discharge summaries to provide data on final diagnoses, but these sources have been shown to be unrepresentative and of variable quality. However, due to logistical reasons, we could not follow up all patients seen in this study. Also, as ED discharge diagnoses on patients were also missing in a number of patients, we used inpatient discharge diagnoses for our main outcome analysis.

CONCLUSIONS

Diagnostic assistance may be useful in a large proportion of patients seen in an emergency department. The Isabel diagnostic aid performs with an acceptable degree of clinical accuracy in this setting. Further studies to elucidate its effects on decision making and diagnostic error are essential in order to clarify its role in routine practice.

ACKNOWLEDGEMENTS

We gratefully acknowledge the advice during study design provided by Dr Paul Taylor.

Authors’ affiliations

Padminabhan Ramnarayan, Children’s Acute Transport Service, London, UK
Natalie Cronje, Isabel Healthcare, London, UK
Ruth Brown, Rupert Neighbour, St Mary’s Hospital, London, UK
Taj Hassam, Wayne Hamer, Leeds General Infirmary, Leeds, UK
Joseph Britto, Isabel Healthcare, Reston, VA, USA

Funding: This study was supported by a research grant from the National Health Service (NHS) Research & Development Unit, London. The sponsor did not influence the study design; the collection, analysis, and interpretation of data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Competing interests: The Isabel system is currently managed by Isabel Healthcare, and is available only to subscribed individual and institutional users. Dr Ramnarayan is a part-time research advisor for Isabel Healthcare, Ms Cronje was employed as a research assistant by Isabel Healthcare for this study, and Dr Britto is Clinical Director of Isabel Healthcare. All other authors declare that they have no competing interests.

REFERENCES

HOW ACCURATELY DOES A WEB-BASED DIAGNOSIS DECISION SUPPORT SYSTEM SUGGEST THE CORRECT DIAGNOSIS IN ADULT ER AND PEDIATRIC CASES?

ISABEL: a web-based differential diagnostic aid for paediatrics: results from an initial performance evaluation

P Ramnarayan, A Tomlinson, A Rao, M Coren, A Winrow, J Britto

Aims: To test the clinical accuracy of a web based differential diagnostic tool (ISABEL) for a set of case histories collected during a two stage evaluation.

Methods: Setting: acute paediatric units in two teaching and two district general hospitals in the southeast of England. Materials: sets of summary clinical features from both stages, and the diagnoses expected for these features from stage I (hypothetical cases provided by participating clinicians in August 2000) and final diagnoses for cases in stage II (children presenting to participating acute paediatric units between October and December 2000). Main outcome measure: presence of the expected or final diagnosis in the ISABEL output list.

Results: A total of 99 hypothetical cases from stage I and 100 real life cases from stage II were included in the study. Cases from stage II covered a range of paediatric specialities (n = 14) and final diagnoses (n = 55). ISABEL displayed the diagnosis expected by the clinician in 90/99 hypothetical cases (91%). In stage II evaluation, ISABEL displayed the final diagnosis in 83/87 real cases (95%).

Conclusion: ISABEL showed acceptable clinical accuracy in producing the final diagnosis for a variety of real as well as hypothetical case scenarios.

A large proportion of routine clinical decision making depends on the availability of good quality clinical information and instant access to up to date medical knowledge.1 There is increasing evidence that the use of computer aided clinical decision support to manage medical knowledge results in better healthcare processes and patient outcomes.2 However, the use of computer based knowledge management techniques in medical decision making has remained poor.3

Considerable advances have been made in making latest processed information from clinical trials available, exemplified by the Cochrane database and the Clinical Evidence series.4 The free accessibility of Medline on the internet (http://www.ncbi.nlm.nih.gov/PubMed) has enabled easy and universal search of the medical literature. However, these endeavours do not primarily serve the busy clinician at the bedside seeking bottom line answers to routine clinical questions. These questions involve diagnosis and immediate management for a patient in general practice or in the emergency department.5 Recent initiatives such as the ATTRACT project have attempted to provide quick, up to date answers to clinical queries for the bedside physician, but involve a substantial financial and resource commitment.6

In addition, the current model of healthcare delivery within the UK National Health Service (NHS) accentuates these problems by providing care in the form of an “inverted pyramid of knowledge”. Clinical wisdom and knowledge are concentrated at the top among senior staff, in many cases distant from the patient who is first seen by a junior doctor at the bottom of the pyramid. This may contribute to a significant proportion of medical error,7 resulting in extra bed days and a preventable financial burden.8

ISABEL (Isabel Medical Charity, UK) is a computerised differential diagnostic aid for paediatrics that is delivered via the world wide web. It was developed following a missed diagnosis on a 3 year old girl with necrotising fasciitis complicating chicken pox. It has currently over 9000 registered users (including doctors, nurses, and other healthcare professionals) and receives over 100 000 page requests every month. Powered by Autonomy, proprietary software that serves as an efficient information retrieval engine by matching patterns within unformatted text, the tool produces differential diagnoses for any set of clinical features by searching text from standard paediatric textbooks. Rather than provide a single diagnosis (a diagnostic tool), ISABEL is primarily intended to suggest only a differential diagnosis and serve as a “reminder” system to remind the clinician of potentially important diagnoses that might have been missed.

During the development of ISABEL, text from each textbook pertaining to each of 3500 different diagnostic labels (for example, measles, migraine) was added to the ISABEL database. These textbooks included Nelson’s textbook of pediatrics (16th edition, 2000, WB Saunders), Forfar and Arnell’s textbook of paediatrics (5th edition, 1998, Churchill Livingstone, UK), Jones and Dargan Churchill’s pocket book of toxicology (2001, Churchill Livingstone, UK), and Rennie and Roberton’s textbook of neonatology (3rd edition, 1999, Churchill Livingstone, UK). Each diagnostic label was allocated an age group classification (“newborn”, “infant”, “child”, or “adolescent”) to prevent inappropriate diagnoses being presented to the user for a specific patient age group.

In order to examine ISABEL’s utility, an evaluation programme was planned in a stepwise fashion: initial system performance to establish the safety of the tool, subsequent evaluation of impact in a simulated setting, and evaluation of impact in a real life clinical setting. This paper describes only the initial evaluation of ISABEL’s capability, and focuses on system performance. The tool was isolated from intended users (clinicians); its impact on clinical practice and decision making was not examined. This study was planned in two stages in two different settings, with the following considerations:

- The two stages to be conducted in series for ease of data collection.
- They were intended to provide a variety of hypothetical as well as real cases for the investigators to test the tool.
ISABEL is useful primarily as a reminder tool and not as an “oracle”. However, in reminding clinicians of diagnoses in a given clinical setting, it is imperative that the final diagnosis is also one of the “reminders” suggested. ISABEL would be considered “unsafe” for clinical use if many plausible diagnoses were suggested to the junior doctor, but the final diagnosis was not. For this reason, it is crucial that ISABEL showed acceptable clinical accuracy by displaying the final diagnosis (especially for inexperienced junior doctors who may not have considered the “correct” diagnosis).

METHODS
Differential diagnostic tool
ISABEL was delivered on the internet free of charge (www.isabel.org.uk). During the development phase, only the investigators had access to ISABEL, enabled by a secure log-in procedure. This would ensure that clinicians participating in the data collection would not be able to use ISABEL and impact on clinical management. On accessing the tool, the patient age group had to be chosen first (newborn, infant, child, or adolescent). Following this, the summary clinical features of the case were entered into a free text box. These features would normally be gathered from the history, physical examination, and results of initial investigations. Any additional findings could be subsequently entered into the tool to focus the differential diagnosis further.

To maximise the tool’s performance, clinical features had to be entered in appropriate medical terminology (rather than lay terms, as the database consisted of textbooks), the spelling had to be accurate (British or American), and laboratory results needed interpretation in words (“leucocytosis” or “increased white cell count” for a white cell count of $36.7 \times 10^6/\mu l$). These data were then submitted to the ISABEL database, searched by Autonomy, and a fresh web page displaying the results from ISABEL was produced. This list included around 10–15 unique diagnoses for consideration, classified on the basis of the systems from which they were drawn (respiratory, metabolic, etc). The diagnoses were not ranked in order of probability, thus reinforcing the function of the tool as a “reminder” system.

Study design and conduct
Stage I: We undertook this study in August 2000 within the Department of Paediatrics at St Mary’s Hospital, London, which has both general paediatric as well as specialist services in infectious diseases, neonatology, and intensive care. Clinicians with varying levels of experience (consultant, registrar, and senior house officer) were contacted to provide hypothetical case histories of acute paediatric presentations. For each case, they specified the age group category, summarised the clinical features, and listed the expected diagnosis(es).

Stage II: This study was undertaken in four acute paediatric units, two teaching hospitals (St Mary’s Hospital, London and Addenbrookes’ Hospital, Cambridge), and two large district general hospitals (Kingston Hospital, Surrey and Royal Alexandra Hospital, Brighton). This study was done from October to December 2000. Junior doctors working within these departments prospectively collected data on children presenting to the acute paediatric unit. Only data regarding the age group, a summary of clinical features at initial presentation, and the working diagnosis(es) were collected from the doctors. The final diagnosis for each patient, as decided by the clinical team at the end of the hospital stay (or at the end of the clinical assessment), was collected from the discharge summary.

Guidance was provided to the doctors to specify clinical features in medical terminology and to interpret results of initial investigations in words. This was done so that the investigators would not need to modify the content provided before entering the clinical features into ISABEL. One investigator (AT) collected these data in a structured form. In one sitting at the end of the data collection, she then entered the age group and the clinical features of each case into ISABEL as provided by the junior doctors, without modifying the content or spelling. This preserved the inherent user variability in summarising clinical features. Results generated by the tool for each case were collected for analysis.

Outcome measures
The main outcome measure used for both stages of the study was the presence of the expected or final diagnosis(es) within the results generated by ISABEL. This was a measure of the
Figure 2  Sample screen from ISABEL: input of clinical features.

Figure 3  Sample screen from ISABEL: output of diagnostic reminders classified by system.
The clinical accuracy of the tool, defined as reminding clinicians of the “correct” diagnosis.

RESULTS

Figure 1 summarises the outcome of all data that were collected for both stages of the evaluation. On the whole, 99 hypothetical cases (provided by 13 clinicians) were eligible for analysis in stage I, and 100 cases in stage II. Forms in which no diagnoses were entered and where the clinical features section included the final diagnosis in the wording were excluded from testing. Figures 2 and 3 show sample screens from ISABEL for a real case in stage II testing.

Tables 1 and 2 summarise the clinical characteristics of cases collected in stages I and II, providing an indication of the spectrum and frequency of final diagnoses as well as specialties covered during the evaluation.

Stage I: Presence of the expected/final diagnosis

This outcome measure is akin to the “sensitivity” of a test. In this respect, ISABEL showed a level of sensitivity between 83% and 91%. More importantly, ISABEL displayed the final diagnosis in stage II cases even when only the presenting clinical features were entered. Although the primary function of the tool is to remind clinicians to consider other reasonable diagnoses, in 13/100 cases, the final diagnosis was non-specific (such as “viral illness”). Since textbooks do not describe such non-specific diagnoses, and the ISABEL database was itself created from textbooks, these were not recognised as distinct final diagnoses. In the remaining 87 cases, the final diagnosis was present in the ISABEL output in 83 cases (95%). The four cases in which the final diagnosis was absent in the ISABEL output were: Stevens-Johnson syndrome, respiratory syncitial virus bronchiolitis, erythema multiforme, and staphylococcal cellulitis.

Recognising that non-specific final diagnoses are often made in routine clinical practice, ISABEL was separately tested against these diagnoses. In 10/13 such cases, ISABEL suggested diagnoses that were nearly synonymous. For a non-specific final diagnosis such as “viral illness”, ISABEL suggested alternatives such as roseola infantum, Epstein-Barr virus, and enteroviruses.

For each case, the ISABEL differential diagnostic tool suggested more than 10 diagnoses (mode 13, range 10–15).

DISCUSSION

Presence of the expected/final diagnosis

This outcome measure is akin to the “sensitivity” of a test. In this respect, ISABEL showed a level of sensitivity between 83% and 91%. More importantly, ISABEL displayed the final diagnosis in stage II cases even when only the presenting clinical features were entered. Although the primary function of the tool is to remind clinicians to consider other reasonable diagnoses, in 13/100 cases, the final diagnosis was non-specific (such as “viral illness”). Since textbooks do not describe such non-specific diagnoses, and the ISABEL database was itself created from textbooks, these were not recognised as distinct final diagnoses. In the remaining 87 cases, the final diagnosis was present in the ISABEL output in 83 cases (95%). The four cases in which the final diagnosis was absent in the ISABEL output were: Stevens-Johnson syndrome, respiratory syncitial virus bronchiolitis, erythema multiforme, and staphylococcal cellulitis.

Recognising that non-specific final diagnoses are often made in routine clinical practice, ISABEL was separately tested against these diagnoses. In 10/13 such cases, ISABEL suggested diagnoses that were nearly synonymous. For a non-specific final diagnosis such as “viral illness”, ISABEL suggested alternatives such as roseola infantum, Epstein-Barr virus, and enteroviruses.

For each case, the ISABEL differential diagnostic tool suggested more than 10 diagnoses (mode 13, range 10–15).
diagnoses in their diagnostic plan and thus influence their management plan, it is important that the tool also generates accurate diagnoses in its output. The use of a final diagnosis as the gold standard is useful in this regard, to ensure that even inexperienced clinicians using the system remain “safe”. In this context, diagnostic accuracy rates of clinicians, in an unselected patient population, have been around 60%. These studies used necropsy findings or results of specific “diagnostic” tests as the gold standard. In a prospective evaluation of a prototype of the medical diagnostic aid Quick Medical Reference (QMR), the unaided diagnostic accuracy rate of physicians (an entire ward team considered as a single unit) was 60% for diagnostically challenging cases. In this highly selective patient population, it was possible to establish a final diagnosis only in 20 out of 31 cases, even after follow up for six months and extensive investigation.

In general, published reports of similar evaluations of other diagnostic systems are sparse. In one study, some commonly used systems for adult medicine (Dxplain, Iliad, QMR, and Meditel) had an accuracy of between 50% and 70% when tested against a set of challenging cases. Our figures compare favourably with results obtained from testing other diagnostic systems in clinical use. Most such systems are expert systems, and use a combination of rule based and Bayesian approaches to model diagnostic decision making. Most systems are also oriented towards adult medicine. Even the few existing paediatric diagnostic aids offer support related to very specific areas such as rheumatic diseases and abdominal

<table>
<thead>
<tr>
<th>System</th>
<th>Final diagnosis (n=55)</th>
<th>Primary diagnosis*</th>
<th>Supplementary diagnosis*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergy</td>
<td>Allergic reaction 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Egg allergy 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Urticarial reaction 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dermatology</td>
<td>Erythema multiforme 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Staphylococcal infection 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endocrine</td>
<td>Diabetes 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Diabetic ketoacidosis 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>Foreign body ingestion 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-specific gastritis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-specific abdominal pain 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pyloric stenosis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haematology</td>
<td>Sickle cell crisis 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infection</td>
<td>Acute gastroenteritis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cellitis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cervical lymphadenitis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Enteroviral infection 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Infected insect bite 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mycoplasma infection 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Osteomyelitis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Otitis media 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pneumococcal meningitis 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pneumonia 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Presacral abscess 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Preseptal cellulitis 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Roseola infantum 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rotavirus infection 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rubella 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scalded skin syndrome 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Septic arthritis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stevens-Johnson syndrome 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Streptococcal infection 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tonsillitis 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Toxic shock syndrome 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Upper respiratory tract infection 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Viral illness 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Viral meningitis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neonate</td>
<td>Opiate withdrawal 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nephrology</td>
<td>Nephrotic syndrome 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurology</td>
<td>Urinary tract infection 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respiratory</td>
<td>Acute severe asthma 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acute sinusitis 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asthma 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bronchiolitis 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Croup 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Laryngomalacia 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Respiratory syncytial virus bronchiolitis 12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Viral induced wheeze 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rheumatic</td>
<td>Acute exacerbation of systemic lupus erythematosus 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Henoch-Schönlein purpura 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skeletal</td>
<td>Disks of L1/L2 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surgical</td>
<td>Appendicitis 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Some patients had more than one final diagnosis (for example, one child had a primary diagnosis of urinary tract infection as well as a supplementary diagnosis of asthma).
pain, making them difficult to use in routine clinical practice. The evaluation of one such system for paediatric rheumatic diseases suggested 80% accuracy (92% when only diagnoses included in its knowledge base were used).

Limitations of the study
This study was designed to evaluate only the presence of the final diagnosis in the reminder list. It did not address the plausibility of the other diagnoses suggested by ISABEL. In this respect, it is possible that although the final diagnosis was present in the differential diagnosis list, the other suggestions made by ISABEL were misleading. However, in order to verify the safety of the tool, it was crucial to show that the final diagnosis formed part of the “reminder” list. Further evaluation is underway to test the relevance of the alternate diagnoses.

The hypothetical cases did not represent the complete spectrum of paediatrics and may not have tested the true limits of the tool. Similarly, the number of real cases used to test the tool may have been insufficient to ensure a complete evaluation of the system’s capability, especially in the specialties of neonatal paediatrics, oncology, and paediatric surgery. In addition, this study was conducted in secondary hospital care. Further testing is planned in a primary care setting to assess the safety of the tool.

Considering that this study was only a preliminary evaluation of the performance of the tool, an attempt to extrapolate these results to routine clinical use is not easily possible. This evaluation separates the tool from the intended user (the clinician). This tool, like other similar tools, will only be used as an adjunct to the existing system of clinical decision making. Despite this, it is possible that inexperienced doctors using the tool either in a primary care or hospital setting might cause an increase in referrals and investigations. The true measure of the tool’s benefits and risks will be clear only when tested by real clinicians involved in the development of the tool. These questions will be answered in subsequent clinical impact evaluations.

CONCLUSIONS
Our study shows that in a large proportion of cases, ISABEL has the potential to remind the clinician of the final diagnosis in a variety of hypothetical as well as real clinical situations.

The spectrum of cases tested is broad enough to ensure that the tool is not limited to a special subset of users. Delivered via the world wide web, ISABEL could impact on patient management on a global level. Because of a potential synergy between the doctor and ISABEL, the results of this study cannot be directly extrapolated to estimate the clinical impact of the tool. Further studies are underway to evaluate the effects of clinicians using ISABEL to aid clinical decision making in simulated as well as real settings. These studies are essential to explore the positive as well as negative impact ISABEL might produce on healthcare processes, health economics, as well as patient outcomes.

Table 3 Example from stage II evaluation data showing the clinical features, the ISABEL differential diagnosis list, and the final diagnosis

<table>
<thead>
<tr>
<th>Age group</th>
<th>Clinical features</th>
<th>Final diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infant</td>
<td>Cough</td>
<td>Respiratory syncytial virus bronchiolitis</td>
</tr>
<tr>
<td></td>
<td>Wheeze</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Increased work of breathing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subcostal recession</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wheeze and crackles on examination</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tachypnoea</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Poor feeding</td>
<td></td>
</tr>
</tbody>
</table>

REFERENCES
WHAT IS THE IMPACT ON CHANGE OF DIAGNOSIS DECISION MAKING BEFORE AND AFTER USING A DIAGNOSIS DECISION SUPPORT SYSTEM?


Diagnostic omission errors in acute paediatric practice: impact of a reminder system on decision-making.

BMC Med Inform Decis Mak. 2006 Nov 6;6(1):37
Diagnostic omission errors in acute paediatric practice: impact of a reminder system on decision-making

Padmanabhan Ramnarayan*1, Andrew Winrow2, Michael Coren3, Vasanta Nanduri4, Roger Buchdahl5, Benjamin Jacobs6, Helen Fisher7, Paul M Taylor8, Jeremy C Wyatt9 and Joseph Britto7

Address: 1Children’s Acute Transport Service (CATS), 44B Bedford Row, London, WC1H 4LL, UK, 2Department of Paediatrics, Kingston General Hospital, Galvowsby Road, Kingston-upon-Thames, KT2 7QB, UK, 3Department of Paediatrics, St Mary’s Hospital, Paddington, London, W2 1NY, UK, 4Department of Paediatrics, Watford General Hospital, Vicarage Road, Watford, WD18 0HB, UK, 5Department of Paediatrics, Hillingdon Hospital, Pield Heath Road, Middlesex, UB8 3NN, UK, 6Department of Paediatrics, Northwick Park Hospital, Watford Road, Harrow, Middlesex, HA1 3UJ, UK, 7Isabel Healthcare Ltd, Po Box 244, Haslemere, Surrey, GU27 1WU, UK, 8Centre for Health Informatics and Multiprofessional Education (CHIME), Archway Campus, Highgate Hill, London, N19 5LW, UK and 9Health Informatics Centre, The Mackenzie Building, University of Dundee, Dundee, DD2 4BF, UK

Email: Padmanabhan Ramnarayan* - ramnarayan@msn.com; Andrew Winrow - andrew.winrow@kingstonhospital.nhs.uk; Michael Coren - michael.coren@st-marys.nhs.uk; Vasanta Nanduri - v_nanduri@hotmail.com; Roger Buchdahl - rbuc@easynet.co.uk; Benjamin Jacobs - benjamin.jacobs@nwlh.nhs.uk; Helen Fisher - helen.fisher@st-marys.nhs.uk; Paul M Taylor - p.taylor@chime.ucl.ac.uk; Jeremy C Wyatt - j.wyatt@chs.dundee.ac.uk; Joseph Britto - joseph@isabelhealthcare.com

* Corresponding author

Abstract

Background: Diagnostic error is a significant problem in specialties characterised by diagnostic uncertainty such as primary care, emergency medicine and paediatrics. Despite wide-spread availability, computerised aids have not been shown to significantly improve diagnostic decision-making in a real world environment, mainly due to the need for prolonged system consultation. In this study performed in the clinical environment, we used a Web-based diagnostic reminder system that provided rapid advice with free text data entry to examine its impact on clinicians’ decisions in an acute paediatric setting during assessments characterised by diagnostic uncertainty.

Methods: Junior doctors working over a 5-month period at four paediatric ambulatory units consulted the Web-based diagnostic aid when they felt the need for diagnostic assistance. Subjects recorded their clinical decisions for patients (differential diagnosis, test-ordering and treatment) before and after system consultation. An expert panel of four paediatric consultants independently suggested clinically significant decisions indicating an appropriate and ‘safe’ assessment. The primary outcome measure was change in the proportion of ‘unsafe’ workups by subjects during patient assessment. A more sensitive evaluation of impact was performed using specific validated quality scores. Adverse effects of consultation on decision-making, as well as the additional time spent on system use were examined.

Results: Subjects attempted to access the diagnostic aid on 595 occasions during the study period (8.6% of all medical assessments); subjects examined diagnostic advice only in 177 episodes (30%). Senior House Officers at hospitals with greater number of available computer workstations in the clinical area were most likely to consult the system, especially out of working hours. Diagnostic workups construed as ‘unsafe’ occurred in 47/104 cases (45.2%); this reduced to 32.7% following system consultation (McNemar test, p < 0.001). Subjects’ mean ‘unsafe’ workups per case decreased from 0.49 to 0.32 (p < 0.001). System advice prompted the clinician to consider the ‘correct’ diagnosis (established at discharge) during initial assessment in 3/104 patients. Median usage time was 1 min 38 sec (IQR 50 sec – 3 min 21 sec). Despite a modest increase in the number of diagnostic possibilities entertained by the clinician, no adverse effects were demonstrable on patient management following system use. Numerical technical barriers prevented subjects from accessing the diagnostic aid in the majority of eligible patients in whom they sought diagnostic assistance.

Published: 06 November 2006


Received: 15 July 2006

Accepted: 06 November 2006

This article is available from: http://www.biomedcentral.com/1472-6947/6/37

© 2006 Ramnarayan et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
**Background**

Studies suggest that a significant proportion of adverse events in primary as well as in secondary care result from errors in medical diagnosis [1-3]; diagnostic errors also constitute the second leading cause for malpractice suits against hospitals [4]. Specialities such as primary care and emergency medicine have specifically been identified as high risk areas for diagnostic mishaps, where cognitive biases in decision making contribute to errors of omission, resulting in incomplete workup and 'missed diagnoses' [5-7]. Adverse events are also commoner in extremes of age, such as paediatric patients and the elderly [8,9]. Diagnostic decision support systems (DDSS), computerised tools that provide accurate and useful patient- and situation-specific advice have been proposed as a technological solution for the reduction of diagnostic errors in practice [10]. Although a number of 'expert diagnostic systems' exist currently, a recent systematic review showed that these systems were less effective in practice than systems that provided preventive care reminders and prescription advice [11]. To a large extent, this may be because most latter systems were integrated into an existing electronic medical record (EMR), enabling effortless and frequent use by clinicians; in contrast, expert DDSS such as Quick Medical Reference (QMR), ILLIAD and MEDITEL-PEDS were typically used in stand-alone fashion [12-14]. Due to a lengthy data input process, considerable clinician motivation and effort was required for their regular use, leading to infrequent usage [15]. As a result, busy clinical areas have been poorly served by existing DDSS.

Attempts to integrate diagnostic decision support into an EMR have been sporadic [16,17], mainly limited by the difficulties associated with converting a complex clinical narrative into structured clinical data in a standard EMR, especially for specialities such as paediatrics and emergency medicine. It appears likely that in the medium term, effortless provision of decision support for busy clinical areas at high-risk for diagnostic error seems possible only through alternative approaches. A Web-based paediatric DDSS that permits rapid use in a busy clinical environment using natural language free text data entry has been recently described [18,19]. Its underlying knowledge base consists of textual descriptions of diseases; using statistical natural language processing, the DDSS matches clinical features to disease descriptions in the database. This approach is similar to that adopted by the RECONSIDER program [20]. Diagnostic suggestions are displayed in sets of 10 up to a maximum of 30, and arranged by body system (e.g. cardiology) rather than by clinical probability. Between 2001 and 2003, >15,000 users registered for its use, 10% of whom used it on >10 separate occasions, resulting in >60,000 distinct user log-ins (personal communication). Thus, although poor usage has been a major confounding factor during evaluations of the clinical benefits of a number of DDSS [21], Isabel usage statistics led us to believe that a study evaluating its clinical impact would permit the assessment of its benefits and risks to be interpreted with confidence, and provide useful insights into the user-DDSS dynamic. Results from an independent email questionnaire survey also suggested that most regular users in the UK found it helpful during patient management [22].

In this study, we aimed to measure the clinical impact of the Isabel system on diagnostic decision making. We hypothesised that lessons learnt from our evaluation study could be generalised to the design, implementation and evaluation of other stand-alone DDSS, and clarify the risks associated with the use of such a system in real life. Diagnostic suggestions were provided to junior doctors during acute paediatric assessments in which they experienced diagnostic uncertainty.

**Methods**

The study was co-ordinated from St Mary's Hospital, Imperial College London, and was approved by the London multi-centre research ethics committee (MREC/02/2/70) and relevant local research ethics committees.

**Study centres**

From a list of non-London district general hospitals (DGH) at which >4 individual users were registered in the Isabel database, four paediatric departments (two university-affiliated DGHs and two DGHs without official academic links) were enrolled, based on logistical and clinical reasons - all sites were <100 miles driving distance from London; three were geographically clustered in the Eastern Region and two sites were separated by large distances from regional tertiary centres. The baseline characteristics of each of the participating centres are detailed in table 1. None of the study centres were involved in the development of the DDSS.

**Study participants**

All junior doctors (Senior House Officers [interns] and Registrars [residents]) in substantive posts at each of the
participating paediatric departments between December 2002 and April 2003 were enrolled after informed verbal consent. Consultants (attending physicians) and locum doctors were excluded.

**Study patients**
All children (age 0–16 years) presenting with an acute medical complaint, and assessed by a junior doctor in a designated Paediatric Assessment Area/Ambulatory Unit (PAU), were eligible for DDSS use. Outpatients, re-attendances for ward follow up, and day cases were ineligible. Based on subjects' feedback collected prior to the study start date, we made a pragmatic decision to allow junior doctors to selectively consult the DDSS only for patients in whom they experienced diagnostic uncertainty. This latter subset formed the actual study population.

**Study design and power**
Our study was a within-subject ‘before and after’ evaluation in which each study subject acted as their own control. Participants explicitly recorded their diagnostic workup and clinical plans (tests and treatment) for cases before seeking DDSS advice. Following real-time use, subjects either decided to act on system advice (by recording their revised diagnostic workup and clinical plans) or chose not to, thus ending the consultation. On the basis of a pilot study in an experimental setting, we calculated that the trial required data from 180 cases to detect a 33% reduction in clinically ‘unsafe’ diagnostic workups (80% power; type I error 5%). We defined diagnostic workups as being ‘unsafe’ if they deviated from a ‘minimum gold standard’ provided by an independent expert panel.

**Intervention**
Decision support system
A limited, secure (128-bit SSL encryption) and password-protected trial version was used. This differed from the publicly accessible DDSS – only the first 10 diagnostic suggestions were displayed, and reading material related to each diagnosis was not made available. The trial version was accessible only via a designated shortcut icon placed on each computer present at study commencement within the participating PAUs. This mechanism utilised cookies, and automatically facilitated identification of the centre at which the request originated, allowing subjects to access the system without an additional log-in procedure. On the trial DDSS, data were captured in two consecutive screens (figures 1 and 2). On screen 1, subjects recorded the patient's clinical details in their own words, and their own diagnostic workup and clinical plans (pre-DDSS). Based on the clinical details submitted in screen 1, diagnostic suggestions were displayed on screen 2. Subjects had the choice to revise their original diagnostic workup and clinical plans by adding or deleting decisions at this stage. It was not possible to go back from screen 2 to screen 1. The consultation ended when clinical decisions from screen 2 (unmodified or revised) were finally submitted, or after >2 hours of inactivity.

**Training**
Three separate group training sessions were organised by one investigator (PR) at each centre one month before the study start date, coinciding with weekly mandatory departmental teaching sessions. At each session, subjects used the trial DDSS with practice cases created for the study. Sessions were repeated twice during the study period to recruit and train new post-holders.

**Outcome measures**
The primary outcome measure was change in the proportion of ‘unsafe’ diagnostic workups following DDSS consultation. We defined ‘unsafe’ workups as instances in which subjects' diagnostic workup (pre- and post-DDSS consultation) deviated from a ‘minimum gold standard’ provided by an independent expert panel of clinicians.

---

**Table 1: Characteristics of participating paediatric departments**

<table>
<thead>
<tr>
<th>Centre</th>
<th>Nature</th>
<th>24 hours dedicated PAU</th>
<th>Annual PAU attendance</th>
<th>Number of junior doctors</th>
<th>Number of consultants (acute)</th>
<th>Computers in PAU</th>
<th>Metropolitan area</th>
<th>Dist to tertiary centre (miles)</th>
<th>Clinical activity (PAU attendances per hour PAU open)</th>
<th>Computer accessibility index (available computers per unit clinical activity)</th>
<th>Senior support (number of acute consultants per subject enrolled)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>university</td>
<td>no*</td>
<td>4194</td>
<td>31</td>
<td>10 (6)</td>
<td>2</td>
<td>yes</td>
<td>&lt;25</td>
<td>1.1</td>
<td>1.8</td>
<td>0.2</td>
</tr>
<tr>
<td>B</td>
<td>dgh</td>
<td>yes</td>
<td>4560</td>
<td>21</td>
<td>7 (4)</td>
<td>1</td>
<td>mixed</td>
<td>26</td>
<td>0.53</td>
<td>1.9</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>dgh</td>
<td>yes</td>
<td>4780</td>
<td>12</td>
<td>7 (4)</td>
<td>1</td>
<td>no</td>
<td>71</td>
<td>0.54</td>
<td>1.85</td>
<td>0.3</td>
</tr>
<tr>
<td>D</td>
<td>university</td>
<td>yes</td>
<td>4800</td>
<td>16</td>
<td>11 (7)</td>
<td>3</td>
<td>no</td>
<td>69</td>
<td>0.67</td>
<td>4.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Abbreviations: dgh = district general hospital, PAU = paediatric assessment unit

* PAU open from 0800 to 0000 only
Inclusion of the correct discharge diagnosis in the diagnostic workup (pre- and post-Isabel consultation), quality scores for diagnostic workup and clinical action plans; time taken by subjects to complete system usage; number of diagnoses included in the diagnostic assessment pre- and post-DDSS; inappropriate tests and treatments ordered by subjects following system advice and significant decisions deleted following consultation were examined as secondary outcome measures.

Data collection
Clinical data were collected during the study period (December 2002-April 2003) by two complementary methods – automatically from trial DDSS logs, and manually by a single research assistant (HF). Data collected by the trial website during system use is shown in table 2. At the end of each clinical consultation, the user indicated how useful the diagnostic advice provided had been for a) patient management and b) as an educational exercise.
**Step 8:** tick the **additional** diagnoses you feel are relevant to your patient:

<table>
<thead>
<tr>
<th>INFECTIOUS DISEASES</th>
<th>RESPIRATORY SYSTEM DISORDERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ Bacterial Meningitis</td>
<td>☐ Pharyngitis/Tonsillitis</td>
</tr>
<tr>
<td>☐ Relapsing Fever</td>
<td>☐ Laryngitis</td>
</tr>
<tr>
<td>☑ Enterovirus Infections</td>
<td>☐ Parotitis</td>
</tr>
<tr>
<td>☐ Viral Meningoencephalitis</td>
<td>☐ Rhinovirus</td>
</tr>
<tr>
<td>☐ Parvoviral Infections</td>
<td>☐ Sinusitis</td>
</tr>
<tr>
<td>☐ Meningococcal Disease</td>
<td>☐ Sinusitis/Tonsillitis</td>
</tr>
<tr>
<td>☐ Herpes Simplex Encephalitis</td>
<td>☐ Strep Throat Swab</td>
</tr>
<tr>
<td>☐ Erythema Infectosum</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RD</td>
</tr>
<tr>
<td></td>
<td>RD</td>
</tr>
</tbody>
</table>

On the basis of the above differential diagnosis list,

**Step 9:** what does your differential diagnosis include now? (there is no need to add any that have been ticked above)

- [ ] bacterial meningitis encephalitis

**Step 10:** would you like to modify your investigations list?

- [ ] blood culture
- [ ] lumbar puncture
- [ ] viral throat swab

**Step 11:** would you like to modify your management list?

- [ ] antibiotics
- [ ] neuro-observations
- [ ] ivaciclovir

**Step 12:** Please rate how useful **Isabel** was for the clinical management of this patient

- [ ] not at all
- [ ] very

**Step 13:** Please rate how useful **Isabel** was as an educational tool on this occasion

- [ ] not at all
- [ ] very

**Step 14:** submit

---

**Figure 2**

This figure shows screen 2, which is displayed following submission of the information entered on screen 1. The subject has the opportunity to revise their decisions, including diagnoses and tests and treatments. A brief survey attempts to capture the user’s satisfaction with respect to educational value as well as clinical utility. Information submitted following this page completes the episode of data collection.
This was recorded on a Likert-scale from 0–5 (not useful to extremely useful). All subjects were sent two rounds of email and postal questionnaires each to collect feedback at trial completion.

The research assistant obtained study patients’ medical records that matched the patient identifiers collected automatically from the trial website during system use. It was not possible to use partial or invalid entries to match medical records. Copies of available medical records were made such that information was available only up to the point of DDSS use. Copies were anonymised by masking patient and centre details. Diagnostic workup and clinical plans recorded by subjects on the trial website were verified for each case against entries in the medical records and in hospital laboratory systems. Discharge diagnoses were collected from routinely collected coding data for all study patients, and additionally from discharge summaries where available. Discharge diagnoses were validated by a consultant involved in study conduct at each centre. In addition, limited demographic and clinical details of all eligible patients at each study centre were collected from hospital administrative data.

A panel of four consultant paediatricians independently examined study medical records, in which subjects’ clinical decisions were masked to ensure blinding. In the first instance, each panel member provided a list of ‘clinically significant’ diagnoses, tests and treatments (the latter two slightly more than half the records were assessed by a single panel member). Clinically significant decisions (diagnoses, tests and treatments) were collated as a ‘minimum gold standard’ set for each case. For cases assessed by multiple panel members, significant decisions provided by a majority of assessors were used to form the gold standard set. Concordance between panel members for clinical decisions was moderate to good, as assessed by the intra-class correlation co-efficient for decisions examined by all four members (0.70 for diagnoses, 0.47 for tests and 0.57 for treatments).

### Analysis

We analysed study data from two main perspectives: operational and clinical. For operational purposes, we defined each attempt by a subject to log into the DDSS by clicking on the icon as a ‘DDSS attempt’. Each successful display of screen 1 was defined as a ‘successful log in’; a unique study identifier was automatically generated by the trial website for each successful log in. Following log in, DDSS usage data was either ‘complete’ (data were available from screens 1 and 2) or ‘incomplete’ (data were available from screen 1 only, although screen 2 may have been displayed to the subject). Time spent by the user processing system advice was calculated as the difference between the time screen 2 was displayed and the end of the consultation (or session time out).

In the first instance, we used McNemar’s test for paired proportions to analyse the change in proportion of ‘unsafe’ diagnostic workups. In order to account for the clustering effects resulting from the same subject assessing a number of cases, we also calculated a mean number of ‘unsafe’ diagnostic workups per case attempted for each subject. Change in this variable following DDSS consultation was analysed using two-way mixed-model analysis of variance (subject grade being between-subjects factor and occasion being within-subjects factor). In order to exclude re-thinking effect as an explanation for change in the primary outcome variable, all episodes in which there was a case that would ensure a safe clinical assessment. The absence of ‘clinically significant’ items in a subject’s workup was explicitly defined during panel review to represent inappropriate clinical care. For this reason, the panel members did not include all plausible diagnoses for each case as part of their assessment, and instead focused on the minimum gold standard. Using this list as a template, the appropriateness of each decision suggested by subjects for each case was subsequently scored by the panel in blinded fashion using a previously validated scoring system [23]. This score rewarded decision plans for being comprehensive (sensitive) as well as focussed (specific). 25% of medical records were randomly assigned to all four panel members for review; a further 20% was assigned to one of the six possible pairs (i.e. slightly more than half the records were assessed by a single panel member). Clinically significant decisions (diagnoses, tests and treatments) were collated as a ‘minimum gold standard’ set for each case. For cases assessed by multiple panel members, significant decisions provided by a majority of assessors were used to form the gold standard set. Concordance between panel members for clinical decisions was moderate to good, as assessed by the intra-class correlation co-efficient for decisions examined by all four members (0.70 for diagnoses, 0.47 for tests and 0.57 for treatments).

### Table 2: Study data automatically collected by the DDSS logs

<table>
<thead>
<tr>
<th>Patient details</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Surname</td>
<td></td>
</tr>
<tr>
<td>Date of birth</td>
<td></td>
</tr>
<tr>
<td>Age group (neonate, infant, child or adolescent)</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>User details</td>
<td></td>
</tr>
<tr>
<td>Centre code (based on identity of icon clicked)</td>
<td></td>
</tr>
<tr>
<td>Subject identity (including an option for anonymous)</td>
<td></td>
</tr>
<tr>
<td>Subject grade</td>
<td></td>
</tr>
<tr>
<td>Operational details</td>
<td></td>
</tr>
<tr>
<td>Date and time of usage (log in, submission of each page of data)</td>
<td>Unique study ID assigned at log in</td>
</tr>
<tr>
<td>Clinical details</td>
<td></td>
</tr>
<tr>
<td>Patient clinical features at assessment</td>
<td></td>
</tr>
<tr>
<td>Doctor’s differential diagnosis (pre-ISABEL)</td>
<td></td>
</tr>
<tr>
<td>Doctor’s investigation plan (pre-ISABEL)</td>
<td></td>
</tr>
<tr>
<td>Doctor’s management plan (pre-ISABEL)</td>
<td></td>
</tr>
<tr>
<td>Isabel list of differential diagnoses</td>
<td></td>
</tr>
<tr>
<td>Diagnoses selected from Isabel list by user as being relevant</td>
<td></td>
</tr>
<tr>
<td>Doctor’s differential diagnosis (post-ISABEL)</td>
<td></td>
</tr>
<tr>
<td>Doctor’s investigation plan (post-ISABEL)</td>
<td></td>
</tr>
<tr>
<td>Doctor’s management plan (post-ISABEL)</td>
<td></td>
</tr>
<tr>
<td>Survey details</td>
<td></td>
</tr>
<tr>
<td>Satisfaction score for patient management</td>
<td></td>
</tr>
<tr>
<td>Satisfaction score for educational use</td>
<td></td>
</tr>
</tbody>
</table>

BMC Medical Informatics and Decision Making 2006, 6:37

http://www.biomedcentral.com/1472-6947/6/37
difference between the workup pre- and post-DDSS consultation were examined. If diagnoses that were changed by subjects were present in the Isabel suggestion list, it could be inferred that the DDSS was responsible for the change. A more objective marker of clinical benefit was assessed by examining whether the post-Isabel diagnostic workup (but not the pre-Isabel workup) included the discharge diagnosis. We analysed changes in pre- and post-DDSS diagnostic quality scores, as well as clinical action plan scores, using subjects as the unit of analysis. We tested for statistical significance using one way analysis of variance (grade was the between-subjects factor) to provide a sensitive measure of changes in diagnostic workup, and tests and treatments. The median test was used to examine differences between grades in system usage time. Diversity of suggestions displayed by the DDSS during the study, and therefore its dynamic nature, was assessed by calculating the number of unique diagnoses suggested by the system across all episodes of completed usage (i.e. if the diagnostic suggestions remained constant irrespective of case characteristics, this number would be 10). Statistical significance was set for all tests at p value <0.05.

Subjects were expected to use the DDSS in only a subset of eligible patients. In order to fully understand the characteristics of patients in whom junior doctors experienced diagnostic difficulty and consulted the DDSS, we examined this group in more detail. We analysed available data on patient factors (age, discharge diagnosis, outcome of assessment and length of inpatient stay if admitted), user factors (grade of subject), and other details. These included the time of system usage (daytime: 0800–1800; out-of-hours: 1800–0800), centre of use and its nature (DGH vs university-affiliated hospital), an index of PAU activity (number of acute assessments per 60 min period the PAU was functional), computer accessibility index (number of available computers per unit PAU activity) and level of senior support (number of acute consultants per subject). We subsequently aimed to identify factors that favoured completion of DDSS usage using a multiple logistic regression analysis. Significant variables were identified by univariate analysis and entered in forward step-wise fashion into the regression model. Characteristics of patients where subjects derived clinical benefit with DDSS usage were also analysed in similar fashion. We correlated subjects' own perception of system benefit (Likert style response from user survey) with actual benefit (improvement in diagnostic quality score) using the Pearson test. Qualitative analysis of feedback from subjects provided at the end of the study period was performed to provide insights into system design and user interface.

Results
During the 5-month study period, 8995 children were assessed in the 4 PAUs; 76.7% (6903/8995 children) presented with medical problems and were eligible for inclusion in the study. Subjects attempted to seek diagnostic advice on 595 separate occasions across all centres (8.6%). Subjects successfully logged in only in 226 episodes. Data were available for analysis in 177 cases (complete data in 125 cases; screen 1 data only in an additional 52 cases). The summary of flow of patients and data through the study is illustrated in figure 3. Centre-wise distribution of attrition in DDSS usage is demonstrated in table 3. Medical records were available for 104 patients, and were examined by the panel according the allocation scheme shown in figure 4.

The largest subset of patients in whom the DDSS was consulted belonged to the 1–6 year age group (61/177, 34.5%). The mean age of patients was 5.1 years (median 3.3 years). The DDSS was most frequently used by SHOs (126/177, 71.2%). Although 25% of all eligible patients were seen on PAUs outside the hours of 08:00 to 18:00, more than a third of episodes of system usage fell outside the hours of 0800–1800 (64/177, 36.2%). Subjects at Centre D used the DDSS most frequently (59/177, 33.3%). In general, usage was greater at university-affiliated hospitals than at DGHs (106/177, 59.9%). Discharge diagnosis was available in 77 patients. The commonest diagnosis was non-specific viral infection; however, a wide spread of diagnoses was seen in the patient population. 58 patients on whom the DDSS was consulted were admitted to hospital wards, the rest were discharged home following initial assessment. Inpatients stayed in hospital for an average of 3.7 days (range: 0.7–26 days). Clinical characteristics of patients on whom Isabel was consulted are summarised in table 4. A list of discharge diagnoses in table 5 indicates the diverse case mix represented in the study population.

80 subjects enrolled during the study. 63/80 used the system to record some patient data (mean range: 1–12); 56/80 provided complete patient information and their revised decisions post-DDSS consultation (mean range: 1–6). Due to limitations in the trial website design, it was unclear how many of the subjects who did not use the system during the study period (17/80) had attempted to access the DDSS and failed. It was evident that a small number of subjects used the DDSS on multiple (>9) occasions but did not provide their revised clinical decisions, leading to incomplete system use. System usage data with respect to subjects is shown in figure 5.

'Unsafe' diagnostic workups
104 cases in which medical records were available were analysed. Before DDSS consultation, ‘unsafe’ diagnostic workups occurred in 47/104 cases (45.2%); they constituted episodes in which all clinically significant diagnoses were not considered by subjects during initial decision
8995 patients seen in PAU in study period

6903 patients with medical problems

Number of patients in whom diagnostic uncertainty expressed - unknown

595 attempts to access DDSS

Log in unsuccessful in 369 episodes

226 completed log in episodes

177 episodes in which initial decisions submitted

125 episodes in which complete data available

104 medical records available and used for study

2092 patients excluded with surgical problems

52 incomplete episodes

Valid medical records unavailable in 21 cases

Figure 3
Flow diagram of patients and data through the study.
making. Overall, the proportion of ‘unsafe’ workups reduced to 32.7% (34/104 cases) following DDSS consultation, an absolute reduction of 12.5% (McNemar test p value <0.001, table 6). In a further 5 cases, appropriate diagnoses that were missing in subjects’ workup did form part of Isabel suggestions, but were ignored by subjects during their review of DDSS advice. In 11/13 cases in which ‘unsafe’ diagnostic workups were eliminated, the DDSS was used by SHOs. Mean number of ‘unsafe’ workups per case reduced from 0.49 to 0.32 post-DDSS consultation among subjects (p < 0.001); a significant interaction was demonstrated with grade (p < 0.01). Examination of the Isabel suggestion list for cases in which there was a difference between pre- and post-DDSS workup showed that in all cases, the additional diagnoses recorded by subjects formed part of the system’s advice. Similar results were obtained for clinical plans, but smaller reductions were observed. In 3/104 cases, the final

Table 3: Centre-wise attrition of DDSS usage and study data

<table>
<thead>
<tr>
<th>Centre</th>
<th>Centre B</th>
<th>Centre C</th>
<th>Centre D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients seen in PAU</td>
<td>2679</td>
<td>1905</td>
<td>1974</td>
<td>2437</td>
</tr>
<tr>
<td>Medical patients seen in PAU</td>
<td>2201</td>
<td>1383</td>
<td>1405</td>
<td>1914</td>
</tr>
<tr>
<td>Number eligible for diagnostic decision support</td>
<td>unknown</td>
<td>unknown</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>DDSS attempts</td>
<td>338</td>
<td>118</td>
<td>52</td>
<td>87</td>
</tr>
<tr>
<td>DDSS successful log in</td>
<td>unknown</td>
<td>unknown</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>Step 1 completed†</td>
<td>47</td>
<td>26</td>
<td>45</td>
<td>59</td>
</tr>
<tr>
<td>Steps 1&amp;2 completed</td>
<td>30</td>
<td>25</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>Medical records available</td>
<td>24</td>
<td>24</td>
<td>16</td>
<td>40</td>
</tr>
</tbody>
</table>

* Each successful log in was automatically provided a unique study identifier which was not centre-specific. The number of successful log-ins was thus calculated as the total number of study identifiers issued by the trial website.
† Step 1 completed indicates that following successful log-in, the subject entered data on the first screen, i.e. patient details.

Figure 4
Allocation of medical records for expert panel assessment consisting of four raters. All panel members rated 25% of cases and two raters (six possible pairs) rated an additional 20% of cases.
diagnosis for the patient was present in the post-DDSS list but not in the pre-DDSS list, indicating that diagnostic errors of omission were averted following Isabel advice.

**Diagnostic quality scores**
104 cases assessed by 51 subjects in whom medical records were available were analysed. Mean diagnostic quality score across all subjects increased by 6.86 (95% CI 4.0–9.7) after DDSS consultation. The analysis of variance model indicated that there was no significant effect of grade on this improvement (p = 0.15). Similar changes in clinical plan scores were smaller in magnitude (table 7).

**Usage time**
Reliable time data were available in 122 episodes. Median time spent on system advice was 1 min 38 sec (IQR 50 sec – 3 min 21 sec). There was no significant difference between grades with respect to time spent on screen 2 (median test, p = 0.9). This included the time taken to process DDSS diagnostic suggestions, record changes to original diagnostic workup and clinical plans, and to complete the user satisfaction survey.

**Impact on clinical decision making**
Pre-DDSS, a mean of 2.2 diagnoses were included in subjects' workup; this rose to 3.2 post-DDSS. Similarly, the number of tests ordered also rose from 2.7 to 2.9; there was no change in the number of treatment steps. Despite these increases, no deleterious tests or treatment steps were added by subjects to their plans following DDSS consultation. In addition, no clinically significant diagnoses were deleted from their original workup after Isabel advice.

### Table 4: Characteristics of patients in whom Isabel was consulted

<table>
<thead>
<tr>
<th>Factor</th>
<th>Number of DDSS consultation episodes (completed episodes)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PATIENT FACTORS</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Age (n = 177)</strong></td>
<td></td>
</tr>
<tr>
<td>Neonate (19)</td>
<td></td>
</tr>
<tr>
<td>Infant (33)</td>
<td></td>
</tr>
<tr>
<td>Young child (1–6 yrs) (61)</td>
<td></td>
</tr>
<tr>
<td>Older child (6–12 yrs) (38)</td>
<td></td>
</tr>
<tr>
<td>Adolescent (26)</td>
<td></td>
</tr>
<tr>
<td><strong>Primary diagnostic group (n = 77)</strong></td>
<td></td>
</tr>
<tr>
<td>Respiratory (9)</td>
<td></td>
</tr>
<tr>
<td>Cardiac (0)</td>
<td></td>
</tr>
<tr>
<td>Neurological (6)</td>
<td></td>
</tr>
<tr>
<td>Surgical (3)</td>
<td></td>
</tr>
<tr>
<td>Rheumatology (5)</td>
<td></td>
</tr>
<tr>
<td>Infections (31)</td>
<td></td>
</tr>
<tr>
<td>Haematology (3)</td>
<td></td>
</tr>
<tr>
<td>Other (20)</td>
<td></td>
</tr>
<tr>
<td><strong>Outcome (n = 104)</strong></td>
<td></td>
</tr>
<tr>
<td>IP admission (58)</td>
<td></td>
</tr>
<tr>
<td>Discharge (46)</td>
<td></td>
</tr>
<tr>
<td><strong>USER FACTORS</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Grade (n = 177)</strong></td>
<td></td>
</tr>
<tr>
<td>SHO (126 (79))</td>
<td></td>
</tr>
<tr>
<td>Registrar (51 (46))</td>
<td></td>
</tr>
<tr>
<td><strong>OPERATIONAL FACTORS (n = 177)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Time of use</strong></td>
<td></td>
</tr>
<tr>
<td>In hours (0800–1800) (113 (84))</td>
<td></td>
</tr>
<tr>
<td>Out of hours (1800-0800) (64 (41))</td>
<td></td>
</tr>
<tr>
<td><strong>Centre</strong></td>
<td></td>
</tr>
<tr>
<td>A (47 (30))</td>
<td></td>
</tr>
<tr>
<td>B (26 (25))</td>
<td></td>
</tr>
<tr>
<td>C (45 (20))</td>
<td></td>
</tr>
<tr>
<td>D (59 (50))</td>
<td></td>
</tr>
</tbody>
</table>
Using forward step-wise regression analysis, grade of subject (registrar), time of system usage (in-hours), centre identity, senior support and computer accessibility index were identified as independent factors predicting completion of DDSS usage. Patients in whom actual benefit was demonstrated on diagnostic decision making were more likely to stay longer in hospital.

469 unique diagnostic suggestions were generated by the DDSS during its use on 125 cases. This represented a high degree of diversity of responses appropriate for the diverse case mix seen in this study – a static list would have consisted of the same 10 diagnoses, and a unique set of suggestions for each single episode of use would have generated 1250 distinct suggestions.

### Table 5: Discharge diagnoses in children in whom the diagnostic aid was consulted

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viral infection</td>
<td>8</td>
</tr>
<tr>
<td>Acute lymphadenitis</td>
<td>3</td>
</tr>
<tr>
<td>Viral gastroenteritis</td>
<td>3</td>
</tr>
<tr>
<td>Pelvic region and thigh infection</td>
<td>3</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>3</td>
</tr>
<tr>
<td>Acute lower respiratory infection</td>
<td>3</td>
</tr>
<tr>
<td>Allergic purpura</td>
<td>2</td>
</tr>
<tr>
<td>Acute inflammation of orbit</td>
<td>2</td>
</tr>
<tr>
<td>Chickenpox with cellulitis</td>
<td>2</td>
</tr>
<tr>
<td>Gastroenteritis</td>
<td>2</td>
</tr>
<tr>
<td>Rotaviral enteritis</td>
<td>2</td>
</tr>
<tr>
<td>Feeding problem of newborn</td>
<td>2</td>
</tr>
<tr>
<td>Syncope and collapse</td>
<td>2</td>
</tr>
<tr>
<td>Lobar pneumonia</td>
<td>2</td>
</tr>
<tr>
<td>Kawasaki disease</td>
<td>2</td>
</tr>
<tr>
<td>Abdominal pain</td>
<td>2</td>
</tr>
<tr>
<td>Angioneurotic oedema</td>
<td>1</td>
</tr>
<tr>
<td>Erythema multiforme</td>
<td>1</td>
</tr>
<tr>
<td>Constipation</td>
<td>1</td>
</tr>
<tr>
<td>Irritable bladder and bowel syndrome</td>
<td>1</td>
</tr>
<tr>
<td>Coagulation defect</td>
<td>1</td>
</tr>
<tr>
<td>G6PD deficiency</td>
<td>1</td>
</tr>
<tr>
<td>Sickle cell dactylitis</td>
<td>1</td>
</tr>
<tr>
<td>Cellulitis</td>
<td>1</td>
</tr>
<tr>
<td>Clavicle osteomyelitis</td>
<td>1</td>
</tr>
<tr>
<td>Kerion</td>
<td>1</td>
</tr>
<tr>
<td>Labynrithitis</td>
<td>1</td>
</tr>
<tr>
<td>Meningitis</td>
<td>1</td>
</tr>
<tr>
<td>Myositis</td>
<td>1</td>
</tr>
<tr>
<td>Purpura</td>
<td>1</td>
</tr>
<tr>
<td>Scarlet fever</td>
<td>1</td>
</tr>
<tr>
<td>Staphylococcal scalded skin syndrome</td>
<td>1</td>
</tr>
<tr>
<td>Mitochondrial complex I deficiency</td>
<td>1</td>
</tr>
<tr>
<td>Adverse drug effect</td>
<td>1</td>
</tr>
<tr>
<td>Eye disorder</td>
<td>1</td>
</tr>
<tr>
<td>Musculoskeletal back pain</td>
<td>1</td>
</tr>
<tr>
<td>Trauma to eye</td>
<td>1</td>
</tr>
<tr>
<td>Disorders of bilirubin metabolism</td>
<td>1</td>
</tr>
<tr>
<td>Foetal alcohol syndrome</td>
<td>1</td>
</tr>
<tr>
<td>Neonatal erythema toxicum</td>
<td>1</td>
</tr>
<tr>
<td>Physiological jaundice</td>
<td>1</td>
</tr>
<tr>
<td>Stroke</td>
<td>1</td>
</tr>
<tr>
<td>Acute bronchiolitis</td>
<td>1</td>
</tr>
<tr>
<td>Acute upper respiratory infection</td>
<td>1</td>
</tr>
<tr>
<td>Asthma</td>
<td>1</td>
</tr>
<tr>
<td>Hyperventilation</td>
<td>1</td>
</tr>
<tr>
<td>Juvenile arthritis with systemic onset</td>
<td>1</td>
</tr>
<tr>
<td>Polyanarthitis</td>
<td>1</td>
</tr>
<tr>
<td>Reactive arthropathy</td>
<td>1</td>
</tr>
<tr>
<td>Anorectal anomaly</td>
<td>1</td>
</tr>
</tbody>
</table>

Using forward step-wise regression analysis, grade of subject (registrar), time of system usage (in-hours), centre identity, senior support and computer accessibility index were identified as independent factors predicting completion of DDSS usage. Patients in whom actual benefit was demonstrated on diagnostic decision making were more likely to stay longer in hospital.

469 unique diagnostic suggestions were generated by the DDSS during its use on 125 cases. This represented a high degree of diversity of responses appropriate for the diverse case mix seen in this study – a static list would have consisted of the same 10 diagnoses, and a unique set of suggestions for each single episode of use would have generated 1250 distinct suggestions.

### User perception

Data were available in 125 cases in which subjects completed DDSS usage. Mean satisfaction score for patient management was 1.6 (95% CI 1.4–1.96); for Isabel use as an educational adjunct, this was higher (2.4, 95% CI 1.98–2.82). There was moderate correlation between subjects' perception of DDSS usefulness in patient management and actual increment in diagnostic quality score (r value 0.28, p value 0.0038, figure 6). Feedback from questionnaires indicated that many subjects found the trial website cumbersome to use in real time since it forced them to record all their decisions prior to advice, thus taking up time during patient assessment. This was especially problematic since many subjects had used Isabel in its original form. A number of subjects were dissatisfied with computer access during the trial; these related to unavailability of passwords to access the Internet, slow computer connections, unavailability of adequate workstations at the point of clinical use and lack of infrastructure support. Another theme that emerged from user feedback involved the lack of access to reading material on diagnoses during the trial period – most users felt this was an important part of the system and the advice provided.

### Discussion

This study demonstrates that diagnostic uncertainty occurs frequently in clinical practice, and that it is feasible for a DDSS, unintegrated into an EMR, to improve the process of diagnostic assessment when used by clinicians in real life practice. We have also shown that this improvement prevented a small but significant number of diagnostic errors of omission. A number of barriers to computer and Internet access in the clinical setting prevented system use in a significant proportion of eligible patients in whom subjects sought diagnostic assistance.

The DDSS studied provided advice in the field of diagnosis, an area in which computerised systems have rarely been shown to be effective. In an early clinical study, Wexler et al showed that consultation of MEDITEL-PEDS, a DDSS for paediatric practice, resulted in a decrease in the number of incorrect diagnoses made by residents [24]. However, subjects did not interact with the DDSS themselves; advice generated by the system was provided to cli-
nicians, and diagnostic decisions were amended by subjects on the basis of the information provided. The impact of QMR was studied in similar consultative fashion: a beneficial impact was demonstrated on diagnostic decisions as well as test ordering [25]. In a subsequent laboratory study examining the impact of two different systems (QMR and ILIAD) on simulated cases, a correct diagnosis was added by subjects to their diagnostic workup in 6.5% episodes [26]. Diagnostically challenging cases were deliberately used: it was not clear that junior clinicians would seek diagnostic advice on similar cases in routine practice. Since the user-DDSS dynamic plays a key role in whether these systems are used and the extent of benefit derived from them [27,28] the above-mentioned studies provide limited information on how clinicians would interact with computerised DDSS to derive clinical benefits in practice, especially in a busy environment.

Our study was notable for utilising a naturalistic design, in which subjects used the Isabel system without extensive training or monitoring, allowing results to be generalised to the clinical setting. This design allowed us to explore the complex interplay between user-DDSS interaction, user decisions in the face of diagnostic advice, and barriers to usage. The DDSS selected was already being used frequently in practice; a number of previous system evaluations have been confounded by inadequate usage. The clinical performance of the DDSS studied has also been previously validated [29]. A preliminary assessment of Isabel impact on subjects' diagnostic decisions has already been made in a simulated environment, results of which closely mirror our current findings [30]. Although the nature and frequency of clinicians' information needs have been previously described, we were able to estimate the need for diagnostic decision support, and characterise the subgroup of patients in whom junior clinicians sought diagnostic advice. Since diagnostic uncertainty only occurs in a subset of acutely ill patients, similar interventions in the future will need to be targeted, rather than being universally applied. However, this has to be balanced against our finding that there was poor correlation between subjects' own perception of system utility and actual clinical benefit, which suggests that a universal approach to usage may be more beneficial. This phenomenon has been previously described [31]. We have also identified that junior doctors, such as SHOs, are more

Table 7: Changes in mean quality scores for diagnostic workup and clinical action plans

<table>
<thead>
<tr>
<th></th>
<th>SHO</th>
<th>Registrar</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic quality score change (SD)</td>
<td>8.3 (11.6)</td>
<td>3.8 (6.1)</td>
<td>6.9 (10.3)</td>
</tr>
<tr>
<td>Clinical action plan score change (SD)</td>
<td>1.4 (6.3)</td>
<td>1.7 (7.4)</td>
<td>1.5 (6.7)</td>
</tr>
</tbody>
</table>

Figure 5
DDSS usage data shown as distribution of number of subjects by episodes of system use.
likely to use and benefit from DDSS, including in an educational role. Cognitive biases, of which 'premature closure' and faulty context generation are key examples, contribute significantly to diagnostic errors of omission [32], and it is likely that in combination with cognitive forcing strategies adopted during decision making, DDSS may act as 'safety nets' for junior clinicians in practice [33].

Fundamental deviation in function and interface design from other expert systems may have contributed to the observed DDSS impact on decision-making in this study. The provision of reminders has proved highly effective in improving the process of care in other settings [34]. Rapid access to relevant and valid advice is crucial in ensuring usability in busy settings prone to errors of omission – average DDSS consultation time during this study was <2 minutes. It also appears that system adoption is possible during clinical assessment in real time with current computer infrastructure, providing an opportunity for reduction in diagnostic error. EMR integration would allow further control on the quality of the clinical input data as well as provision of active decision support with minimum extra effort; such an interface has currently been developed for Isabel and tested with four commercial EMRs [35]. Such integration facilitates iterative use of the

<table>
<thead>
<tr>
<th>Unsafe diagnostic workup</th>
<th>Pre-DDSS consultation</th>
<th>Post-DDSS consultation</th>
<th>Relative Reduction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHO</td>
<td>28</td>
<td>17</td>
<td>39.3</td>
</tr>
<tr>
<td>Registrar</td>
<td>19</td>
<td>17</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Figure 6
Correlation between user perception of system utility and change in diagnostic quality score.
system during the evolution of a patient's condition, leading to increasingly specific diagnostic advice. A number of other observations are worthy of note: despite an increase in the number of diagnoses considered, no inappropriate tests were triggered by the advice provided; the quality of data input differed widely between users; the system dynamically generated a diverse set of suggestions based on case characteristics; the interpretation of DDSS advice itself was user-dependent, leading to variable individual benefit; and finally, on some occasions even useful advice was rejected by users. User variability in data input cannot be solely attributed to the natural language data entry process; considerable user variation in data entry has been demonstrated even in DDSS that employ controlled vocabularies for input [36]. Further potential benefit from system usage was compromised in this study due to many reasons: unavailability of computers, poor Internet access, and slow network connections frequently prevented subjects from accessing the DDSS. Paradoxically, the need to enter detailed information including subjects' own clinical decisions into the trial website (not required during real life usage) may itself have compromised system usage during the study, limiting the extent to which usage data from the study can be extrapolated to real life.

This study had a number of limitations. Our study was compromised by the lack of detailed qualitative data to fully explore issues related to why users sometimes ignored DDSS advice, or specific cases in which users found the DDSS useful. The comparison of system versus a panel gold standard had its own drawbacks – Isabel was provided variable amount of patient detail depending on the subject who used it, while the panel were provided detailed clinical information from medical notes. Changes in decision making were also assessed at one fixed point during the clinical assessment, preventing us from examining the impact of iterative use of the DDSS with evolving and sometimes rapidly changing clinical information. Due to the before-after design, it could also be argued that any improvement observed resulted purely from subjects rethinking the case; since all appropriate diagnoses included after system consultation were present in the DDSS advice, this seems unlikely. Subjects also spent negligible time between their initial assessment of cases and processing the system's diagnostic suggestions. Our choice of primary outcome focused on improvements in process, although we were also able to demonstrate a small but significant prevention of diagnostic error based on the discharge diagnosis. The link between improvements in diagnostic process and patient outcome may be difficult to illustrate, although model developed by Schiff et al suggests that avoiding process errors will lead to actual errors in some instances, as we have demonstrated in this study [37]. However, in our study design, it was not possible to test whether an ‘unsafe’ diagnostic workup would directly lead to patient harm. Finally, due to barriers associated with computer access and usage, we were not able to reach the target number of cases on whom complete medical data were available.

Conclusion
This clinical study demonstrates that it is possible for a stand-alone diagnostic system based on the reminder model to be used in routine practice to improve the process of diagnostic decision making among junior clinicians. Elimination of barriers to computer access is essential to fulfill the significant need for diagnostic assistance demonstrated in this study.

Competing interests
This study was conducted when the Isabel system was available free to users, funded by the Isabel Medical Charity. Dr Ramnarayan performed this research as part of his MD thesis at Imperial College London. Dr Britto was a Trustee of the medical charity in a non-remunerative post. Ms Fisher was employed by the Isabel Medical Charity as a Research Nurse for this study.

Since June 2004, Isabel has been managed by a commercial subsidiary of the Medical Charity called Isabel Healthcare. The system is now available only to subscribed users. Dr Ramnarayan now advises Isabel Healthcare on research activities on a part-time basis; Dr Britto is now Clinical Director of Isabel Healthcare. Both hold restricted shares in Isabel Healthcare. All other authors declare that they have no competing interests.

Authors’ contributions
PR conceived the study, contributed to the study design, analyzed the data and drafted the manuscript.

BJ assisted with data collection, validation of discharge diagnoses, and data analysis.

MC assisted with the study design, served as gold standard panel member, and revised the draft manuscript

VN assisted with study design, served as gold standard panel member, and helped with data analysis

RB assisted with the study design, served as gold standard panel member, and revised the draft manuscript

AW assisted with the study design, served as gold standard panel member, and revised the draft manuscript

HF assisted with data collection and analysis

PT assisted with study conception, study design and revised the draft manuscript
JW assisted with study conception, provided advice regarding study design, and revised the draft manuscript.

JB assisted with study conception, study design and data analysis.

All authors read and approved the final manuscript.

Acknowledgements
A number of clinicians and ward administrators were involved in assisting in study conduct at each participating centre. The authors would like to thank Nandan Thalange, Michael Bamford, and Elmo Thambapillai for their help in this regard. Statistical assistance was provided by Henry Potts.

Financial support: This study was supported by a research grant from the National Health Service (NHS) Research & Development Unit, London. The sponsor did not influence the study design; the collection, analysis, and interpretation of data; the writing of the manuscript; and the decision to submit the manuscript for publication.

References

Pre-publication history
The pre-publication history for this paper can be accessed here:
WHAT IS THE IMPACT ON CHANGE OF DIAGNOSIS DECISION MAKING BEFORE AND AFTER USING A DIAGNOSIS DECISION SUPPORT SYSTEM?

Assessment of the potential impact of a reminder system on the reduction of diagnostic errors: a quasi-experimental study

Padmanabhan Ramnarayan*1, Graham C Roberts2, Michael Coren3, Vasantha Nanduri4, Amanda Tomlinson5, Paul M Taylor6, Jeremy C Wyatt7 and Joseph F Britto5

Address: 1Children's Acute Transport Service (CATS), 44B Bedford Row, London, WC1H 4LL, UK, 2Department of Paediatric Allergy and Respiratory Medicine, Southampton University Hospital Trust, Tremena Road, Southampton, SO16 6YD, UK, 3Department of Paediatrics, St Mary's Hospital, Paddington, London, W2 1NY, UK, 4Department of Paediatrics, Watford General Hospital, Vicarage Road, Watford, WD18 0HB, UK, 5Isabel Healthcare Ltd, Po Box 244, Haslemere, Surrey, GU27 1WU, UK, 6Centre for Health Informatics and Multiprofessional Education (CHIME), Archway Campus, Highgate Hill, London, N19 5LW, UK and 7Health Informatics Centre, The Mackenzie Building, University of Dundee, Dundee, DD2 4BF, UK

Email: Padmanabhan Ramnarayan* - ramnarayan@msn.com; Graham C Roberts - g.c.roberts@soton.ac.uk; Michael Coren - michael.coren@st-marys.nhs.uk; Vasantha Nanduri - v_nanduri@hotmail.com; Amanda Tomlinson - mandy@isabelhealthcare.com; Paul M Taylor - p.taylor@chime.ucl.ac.uk; Jeremy C Wyatt - j.wyatt@chs.dundee.ac.uk; Joseph F Britto - joseph@isabelhealthcare.com

* Corresponding author

Abstract

Background: Computerized decision support systems (DSS) have mainly focused on improving clinicians’ diagnostic accuracy in unusual and challenging cases. However, since diagnostic omission errors may predominantly result from incomplete workup in routine clinical practice, the provision of appropriate patient- and context-specific reminders may result in greater impact on patient safety. In this experimental study, a mix of easy and difficult simulated cases were used to assess the impact of a novel diagnostic reminder system (ISABEL) on the quality of clinical decisions made by various grades of clinicians during acute assessment.

Methods: Subjects of different grades (consultants, registrars, senior house officers and medical students), assessed a balanced set of 24 simulated cases on a trial website. Subjects recorded their clinical decisions for the cases (differential diagnosis, test-ordering and treatment), before and after system consultation. A panel of two pediatric consultants independently provided gold standard responses for each case, against which subjects’ quality of decisions was measured. The primary outcome measure was change in the count of diagnostic errors of omission (DEO). A more sensitive assessment of the system’s impact was achieved using specific quality scores; additional consultation time resulting from DSS use was also calculated.

Results: 76 subjects (18 consultants, 24 registrars, 19 senior house officers and 15 students) completed a total of 751 case episodes. The mean count of DEO fell from 5.5 to 5.0 across all subjects (repeated measures ANOVA, p < 0.001); no significant interaction was seen with subject grade. Mean diagnostic quality score increased after system consultation (0.044; 95% confidence interval 0.032, 0.054). ISABEL reminded subjects to consider at least one clinically important diagnosis in 1 in 8 case episodes, and prompted them to order an important test in 1 in 10 case episodes. Median extra time taken for DSS consultation was 1 min (IQR: 30 sec to 2 min).

Conclusion: The provision of patient- and context-specific reminders has the potential to reduce diagnostic omissions across all subject grades for a range of cases. This study suggests a promising role for the use of future reminder-based DSS in the reduction of diagnostic error.
Background

A recent Institute of Medicine report has brought the problem of medical error under intense scrutiny [1]. While the use of computerized prescription software has been shown to substantially reduce the incidence of medication-related error [2,3], few solutions have demonstrated a similar impact on diagnostic error. Diagnostic errors impose a significant burden on modern healthcare: they account for a large proportion of medical adverse events in general [4-6], and form the second leading cause for malpractice suits against hospitals [7]. In particular, diagnostic errors of omission (DEO) during acute medical assessment, resulting from cognitive biases such as ‘premature closure’ and ‘confirmation bias’, lead to incomplete diagnostic workup and ‘missed diagnoses’ [8]. This is especially relevant in settings such as family practice [9], as well as hospital areas such as the emergency room and critical care [11,12]. 20% of patients discharged from emergency rooms raised concerns in a recent survey that their clinical assessment had been complicated by diagnostic error [12].

The use of clinical decision-support systems (DSS) has been one of many strategies proposed for the reduction of diagnostic errors in practice [13]. Consequently, a number of DSS have been developed over the past few years to assist clinicians during the process of medical diagnosis [14-16].

Even though studies of several diagnostic DSS have demonstrated improved physician performance in simulated (and rarely real) patient encounters [17,18], two specific characteristics may have contributed to their infrequent use in routine practice: intended purpose and design. Many general diagnostic DSS were built as ‘expert systems’ to solve diagnostic conundrums and provide the correct diagnosis during a ‘clinical dead-end’ [19]. Since true diagnostic dilemmas are rare in practice [20], and the initiative for DSS use had to originate from the physician, diagnostic advice was not sought routinely, particularly since clinicians prefer to store the patterns needed to solve medical problems in their heads [21]. There is, however, evidence that clinicians frequently underestimate their need for diagnostic assistance, and that the perception of diagnostic difficulty does not correlate with their clinical performance [22]. In addition, due to the demands of the information era [23], diagnostic errors may not be restricted to cases perceived as being difficult, and might occur even when dealing with common problems in a stressful environment under time pressure [24]. Further, most ‘expert systems’ utilized a design in which clinical data entry was achieved through a controlled vocabulary specific to each DSS. This process frequently took > 15 minutes, contributing to infrequent use in a busy clinical environment [25]. These ‘expert systems’ also provided between 20 and 30 diagnostic possibilities [26], with detailed explanations, leading to a lengthy DSS consultation process.

In order to significantly affect the occurrence of diagnostic error, it seems reasonable to conclude that DSS advice must therefore be readily available, and sought, during most clinical encounters, even if the perceived need for diagnostic assistance is minor. Ideally, real-time advice for diagnosis can be actively provided by integrating a diagnostic DSS into an existing electronic medical record (EMR), as has been attempted in the past [27,28]. However, the limited uptake of EMRs capable of recording sufficient narrative clinical detail currently in clinical practice indicates that a stand-alone system may prove much more practical in the medium term [29]. The key characteristic of a successful system would be the ability to deliver reliable diagnostic reminders rapidly following a brief data entry process in most clinical situations. ISABEL (ISABEL Healthcare, UK) is a novel Web-based pediatric diagnostic reminder system that suggests important diagnoses during clinical assessment [30,31]. The development of the system and its underlying structure have been described in detail previously [32,33]. The main hypotheses underlying the development of ISABEL were that the provision of diagnostic reminders generated following a brief data entry session in free text would promote user uptake, and lead to improvement in the quality of diagnostic decision making in acute medical settings. The reminders provided (a set of 10 in the first instance) aimed to remind clinicians of important diagnoses that they might have missed in the workup. Data entry is by means of natural language descriptions of the patient’s clinical features, including any combination of symptoms, signs and test results. The system’s knowledge base consists of natural language text descriptions of > 5000 diseases, in contrast to most ‘expert systems’ that use complex disease databases [34-36]. The advantages and trade-offs of these differences in system design have been discussed in detail elsewhere [37]. In summary, although the ability to rapidly enter patient features in natural language to derive a short-list of diagnostic suggestions may allow frequent use by clinicians during most patient encounters, variability resulting from the use of natural language for data entry, and the absence of probability ranking, may compromise the accuracy and usefulness of the diagnostic suggestions.

The overall evaluation of the ISABEL system was planned in systematic fashion in a series of consecutive studies [38].

a) An initial clinical performance evaluation: This would evaluate the feasibility of providing relevant diagnostic suggestions for a range of cases when data is entered in natural language. System accuracy, speed and relevance of suggestions were studied.
b) An assessment of the impact of the system in a quasi-experimental setting: This would examine the effects of diagnostic decision support on subjects using simulated cases.

c) An assessment of the impact of the system in a real life setting: This would examine the effects of diagnostic advice on clinicians in real patients in their natural environment.

In the initial performance evaluation, the ISABEL system formed the unit of intervention, and the quality of its diagnostic suggestions was validated against data drawn from 99 hypothetical cases and 100 real patients. Key findings from cases were entered into the system in free text by one of the developers. The system included the final diagnosis in 95% of the cases [39]. This design was similar to early evaluations of a number of other individual diagnostic DSS [40-42], as well as a large study assessing the performance characteristics of four expert diagnostic systems [43]. Since this step studied ISABEL in isolation, and did not include users uninvolved in the development of the system, it was vital to examine DSS impact on decision making by demonstrating in the subsequent step that the clinician-DSS combination functioned better than either the clinician or the system working in isolation [44,45]. Evaluation of impact is especially relevant to ISABEL: despite good system performance when tested in isolation, clinicians may not benefit from its advice either due to variability associated with user data entry leading to poor results, or the inability to distinguish between diagnostic suggestions due to the lack of ranking [46]. A previous evaluation of Quick Medical Reference (QMR) assessed a group of clinicians working a set of difficult cases, and suggested that the extent of benefit gained by different users varied with their level of experience [47].

In this study, we aimed to perform an impact evaluation of ISABEL in a quasi-experimental setting in order to quantify the effects of diagnostic advice on the quality of clinical decisions made by various grades of clinicians during acute assessment, using a mix of easy and difficult simulated cases drawn from all pediatric sub-specialties. Study design was based on an earlier evaluation of the impact of ILIAD and QMR on diagnostic reasoning in a simulated environment [48]. Our key outcome measure focused on appropriateness of decisions during diagnostic workup rather than accuracy in identifying the correct diagnosis. The validity of textual case simulations has previously been demonstrated in medical education exercises [49], and during the assessment of mock clinical decision making [50,51].

Methods
The simulated field study involved recording subjects’ clinical decisions regarding diagnoses, test-ordering and treatment for a set of simulated cases, both before and immediately after DSS consultation. The impact of diagnostic reminders was determined by measuring changes in the quality of decisions made by subjects. In this study, the quality of ISABEL’s diagnostic suggestion list per se was not examined. The study was coordinated at Imperial College School of Medicine, St Mary’s Hospital, London, UK between February and August 2002. The study was approved by the Local Research Ethics Committee.

Subjects
A convenience sample consisting of pediatricians of different grades (senior house officers [interns], registrars [residents] and consultants [attending physicians] from different geographical locations across the UK), and final year medical students, was enrolled for the study. All students were drawn from one medical school (Imperial College School of Medicine, London, UK). Clinicians were recruited by invitation from the ISABEL registered user database which consisted of a mixture of regular users as well as pediatricians who had never used the system after registration. After a short explanation of the study procedure, all subjects who consented for the study were included within the sample.

Cases
Cases were drawn from a pool of 72 textual case simulations, constructed by one investigator, based on case histories of real children presenting to emergency departments (data collected during earlier evaluation). Each case was limited to between 150 and 200 words, and only described the initial presenting symptoms, clinical signs and basic laboratory test results in separate sections. Since the clinical data were collected from pediatric emergency rooms, the amount of clinical information available at assessment was limited but typical for this setting. Ample negative features were included in order to prevent the reader from picking up positive cues from the text. These cases were then classified into one of 12 different pediatric sub-specialties (e.g. cardiology, respiratory) and to one of 3 case difficulty levels within each specialty (1- unusual, 2-not unusual, and 3-common clinical presentation, with reference to UK general hospital pediatric practice) by the author. This allocation process was duplicated by a pediatric consultant working independently. Both investigators assigned 57 cases to the same sub-specialty and 42 cases to both the same sub-specialty and the same level of difficulty (raw agreement 0.79 and 0.58 respectively). From the 42 cases in which both investigators agreed regarding the allocation of both specialty and level of difficulty, 24 cases were drawn such that a pair of cases per sub-specialty representing two different levels of diffi-
culty (level 1 & 2, 1 & 3 or 2 & 3) was chosen for the final case mix. This process ensured a balanced set of cases representing all sub-specialties and comprising easy as well as difficult cases.

Data collection website

A customized, password protected version of ISABEL was used to collect data during the study. This differed from the main website in that it automatically displayed the study cases to each subject in sequence, assigned each case episode a unique study number, and recorded time data in addition to displaying ten diagnostic suggestions. Three separate text boxes were provided to record subjects’ clinical decisions (diagnoses, tests and treatment) pre- and post-DSS consultation. The use of the customized trial website ensured that subjects proceeded from one step to the next without being able to skip steps or revise clinical decisions already submitted.

Training

Training was intended only to familiarize subjects with the trial website. During training, all subjects were assigned unique log-in and passwords, and one sample case as practice material. Practice sessions involving medical students were supervised by one investigator in group sessions of 2–3 subjects each. Pediatricians (being from geographically disparate locations) were not supervised during training, but received detailed instructions regarding the use of the trial website by email. Context-specific help was provided at each step on the website for assistance during the practice session. All subjects completed their assigned practice case, and were recruited for the study.

Study procedure

Subjects were allowed to complete their assessments of the simulated cases from any computer connected to the Internet at any time (i.e. they were not supervised). After logging into the trial website, subjects were presented with text from a case simulation. They assessed the case, abstracted the salient clinical features according to their own interpretation of the case, and entered them into the designated search query box in free text. Following this, they keyed in their decisions regarding diagnostic workup, test-ordering and treatment into the designated textboxes. These constituted pre-DSS clinical decisions. See figure 1 for an illustration of this step of the study procedure. On submitting this information, a list of diagnostic suggestions was instantly presented to the subject based on the abstracted clinical features. The subjects could not read the case text again at this stage, preventing them from processing the case a second time, thus avoiding ‘second-look’ bias. Diagnostic suggestions were different for different users since the search query was unique for each subject, depending on their understanding of the case and how they expressed it in natural language. On the basis of the diagnostic suggestions, subjects could modify their pre-DSS clinical decisions by adding or deleting items: these constituted post-DSS clinical decisions. All clinical decisions, and the time taken to complete each step, were recorded automatically. See figure 2 for an illustration of this step of the study procedure. The text from one case, and the variability associated with its interpretation during the study, is depicted in figure 3.

Each subject was presented with 12 cases such that one of the pair drawn from each sub-specialty was displayed. Cases were presented in random order (in no particular order of sub-specialty). Subjects could terminate their session at any time and return to complete the remainder of cases. If a session was terminated midway through a case, that case was presented again on the subject’s return. If the website detected no activity for > 2 hours, the subject was automatically logged off, and the session was continued on their return. All subjects had 3 weeks to complete their assigned 12 cases. Since each case was used more than once, by different subjects, we termed each attempt by a subject at a case as a ‘case episode’.

Scoring metrics

We aimed to assess if the provision of key diagnostic reminders would reduce errors of omission in the simulated environment. For the purposes of this study, a subject was defined to have committed a DEO for a case episode if they failed to include all ‘clinically important diagnoses’ in the diagnostic workup (rather than failing to include the ‘correct diagnosis’). A diagnosis was judged ‘clinically important’ if an expert panel working the case independently decided that the particular diagnosis had to be included in the workup in order to ensure safe and appropriate clinical decision making, i.e. they would significantly affect patient management and/or course, and failure to do so would be construed clinically inadequate. The expert panel comprised two general pediatricians with > 3 years consultant level experience. ‘Clinically important’ diagnoses suggested by the panel thus included the ‘most likely diagnosis/es’ and other key diagnoses; they did not constitute a full differential containing all plausible diagnoses. This outcome variable was quantified by a binary measure (for each case episode, a subject either committed a DEO or not). Errors of omission were defined for tests and treatments in similar fashion.

We also sought a more sensitive assessment of changes in the quality of clinical decisions made by subjects in this study. Since an appropriate and validated instrument was essential for this purpose, a measurement study was first undertaken to develop and validate such an instrument. The measurement study, including the development and validation of a diagnostic quality score and a management
plan quality score, has previously been reported in detail [52]. The scoring process, tested using a subset of cases worked on by clinicians during this study (190 case episodes), was reliable (intraclass correlation coefficient 0.79) and valid (face, construct and concurrent validity). During the scoring process, the expert panel was provided an aggregate list of decisions drawn from all subjects (pre- and post-DDSS consultation) for each case. They provided

---

**Figure 1**

**Screenshot of ISABEL simulated study procedure – step 1.** This figure shows how one subject was presented with the text of a case simulation, how he could search ISABEL by using summary clinical features, and record his clinical decisions prior to viewing ISABEL’s results. For this case, clinically important diagnoses provided by the expert panel are: nasopharyngitis (OR viral upper respiratory tract infection) and meningitis/encephalitis. This subject has committed a DEO (failed to include both clinically important diagnoses in his diagnostic workup).
measurements of quality for each of the clinical decisions in addition to identifying ‘clinically important’ decisions for each case. Prior to scoring, one investigator (PR) mapped diagnoses proposed by subjects and the expert panel to the nearest equivalent diagnoses in the ISABEL database. Quality of each diagnostic decision was scored.
A 6-week old infant is admitted with a week's history of poor feeding. Whereas previously the infant had been growing along the 25th centile, he has now fallen below the 10th centile. In the past week, his parents also think he is breathing quite quickly. He has a cough and is vomiting most feeds.

On examination, he is tachypnoeic, has moderate respiratory distress, and a cough. A grade 3 murmur is heard all over his precordium. He also has a 3 cm liver palpable.

Initial lab results show Haemoglobin 9 g/dL, White cell count 12.4 x 10^6/µL, Platelets 180 x 10^6/µL.

*This case was assigned to Cardiology and case level 2 by both investigators*

| User A | failure to thrive, growth below 10th centile, previous growth on 25th centile, tachypnoeic, moderate respiratory distress, cough, precordial grade 3 murmur, palpable liver 3cm, neutrophilia, thrombocytopenia |
| User B | poor feeding for 1/52, FTT, DIB + COUGH + vomiting, HEART MURMUR, liver palpable |
| User C | Weight loss, cough, vomiting, respiratory distress, heart murmur, hepatomegaly |
| User D | poor feeding, failure to thrive, tachypnoeic, cough, vomiting, respiratory distress, heart murmur, large liver, anaemia |
| User E | poor feeding, wt loss, tachypneic, cardiac murmur |

<table>
<thead>
<tr>
<th>‘Most likely diagnosis’</th>
<th>‘Clinically important’ diagnoses (panel)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ventricular septal defect</td>
<td>Ventricular septal defect (or other congenital heart disease with left to right shunt)</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngitis (or Bronchiolitis)</td>
</tr>
</tbody>
</table>

Figure 3
Example of one simulated case used in study*, the variability in clinical features as abstracted by five different users (verbatim), and clinically important diagnoses as judged by panel.

for the degree of plausibility, likelihood in the clinical setting, and its impact on further patient management. These measurements were used to derive scores for each set of subjects' clinical decisions (diagnostic workup, tests and treatments). As per the scoring system, subjects' decision plans were awarded the highest score (score range: 0 to 1) only if they were both comprehensive (contained all important clinical decisions), and focused (contained only important decisions). Scores were calculated for each subject's diagnostic, test-ordering and treatment plans both
pre- and post-ISABEL consultation. Figure 4 provides a schematic diagram of the complete scoring procedure.

Primary outcome
1. Change in the number of diagnostic errors of omission among subjects.

Secondary outcomes
1. Mean change in subjects' diagnostic, test-ordering and treatment plan quality scores.
2. Change in the number of irrelevant diagnoses contained within the diagnostic workup.
3. Proportion of case episodes in which at least one additional 'important' diagnosis, test or treatment decision was considered by the subject after DSS consultation.
4. Additional time taken for DSS consultation.

Analysis
Subjects were used as the unit of analysis for the primary outcome measure. For each subject, the total number of DEOs was counted separately for pre- and post-DSS diagnostic workup plans; only subjects who had completed all assigned cases were included in this calculation. Statistically significant changes in DEO count following DDSS consultation and interaction with grade was assessed by two-way mixed-model analysis of variance (grade being between-subjects factor and time being within-subjects factor). Mean number of DEOs was calculated for each subject grade, and DEOs were additionally analyzed according to level of case difficulty. Statistical significance was set at a p value of 0.05.

Subjects were used as the unit of analysis for the change in mean quality scores (the development of quality scores and their validation has been previously described; however, the scores have never been used as an outcome measure prior to this evaluation). In the first step, subjects' quality score (pre- and post-DSS) was calculated for each case episode. For each subject, a mean quality score across all 12 cases was computed. Only case episodes from subjects who completed all 12 assigned cases were used during this calculation. A two-way mixed model ANOVA (grade as between-subjects factor; time as within-subjects factor) was used to examine statistically significant differences in quality scores. This analysis was performed for diagnostic quality scores as well as test ordering and treatment plan scores. Data from a pilot study suggested that data from 64 subjects were needed to demonstrate a mean diagnostic quality score change of 0.03 (standard deviation 0.06, power 80%, level of significance 5%).

Using subjects as the unit of analysis, the mean count of diagnoses (and irrelevant diagnoses) included in the workup was calculated pre- and post-DSS consultation for each subject as an average across all case attempts. Only subjects who attempted all assigned cases were included in this analysis. Using this data, a mean count for diagnoses (and irrelevant diagnoses) was calculated for each subject grade. A two-way mixed model ANOVA was used to assess statistically significant differences in this outcome with respect to grade as well as occasion. Using case episodes as the unit of analysis, the proportion of case episodes in which at least one additional 'important' diagnosis, test or treatment was prompted by ISABEL was determined. The proportion of case episodes in which at least one clinically significant decision was deleted, and at least one inappropriate decision was added, after system consultation, was also computed. All data were analyzed separately for the subjects' grades.

Two further analyses were conducted to enable the interpretation of our results. First, in order to provide a direct comparison of our results with other studies, we used case episodes as the unit of analysis and examined the presence of the 'most likely diagnosis' in the diagnostic workup. The 'most likely diagnosis' was part of the set of 'clinically important' diagnoses provided by the panel, and represented the closest match to a 'correct' diagnosis in our study design. This analysis was conducted separately for
each grade. Second, since it was important to verify whether any reduction of omission errors was directly prompted by ISABEL, or simply by subjects re-thinking about the assigned cases, all case episodes in which at least one additional significant diagnosis was added by the user were examined. If the diagnostic suggestion added by the user had been displayed in the DSS list of suggestions, it strongly suggested that the system, rather than subjects’ re-thinking, prompted these additions.

Results
The characteristics of subjects, cases and case episodes are summarized in table 1. Ninety seven subjects were invited for the study. Although all subjects consented and completed their training, only seventy six subjects attempted at least one case (attempters) during their allocated three weeks. This group consisted of 15 medical students, 19 SHOs, 24 registrars and 18 consultants. There was no significant difference between attempters and non-attempters with respect to grade (Chi square test, p 0.07). Only 6/76 subjects had used ISABEL regularly (at least once a week) prior to the study period (3 SHOs, 1 Registrar and 2 Consultants); all the others had registered for the service, but never used the DSS previously. A total of 751 case episodes were completed by the end of the study period. Fifty two subjects completed all assigned 12 cases to produce 624 case episodes (completers); 24 other subjects did not complete all their assigned cases (non-completers). Completers and non-completers did not differ significantly with respect to grade (Chi square test, p 0.06). However, more subjects were trained remotely in the non-completers group (Chi square test, p 0.003). The majority of non-completers had worked at least two cases (75%); slightly less than half (42%) had worked at least 6 cases. Forty-seven diagnoses were considered ‘clinically important’ by the panel across all 24 cases (average ~ 2 per case). For 21/24 cases, the panel had specified a single ‘most likely diagnosis’; for 3 cases, two diagnoses were included in this definition.

Diagnostic errors of omission
624 case episodes generated by 52 subjects were used to examine DEO. During the pre-DSS consultation phase, all subjects performed a DEO in at least one of their cases, and 21.1% (11/52) in more than half their cases. This gradient was maintained post-ISABEL consultation (6.6 vs. 4.4); this gradient was maintained post-ISABEL consultation (5.9 vs. 0.4). Overall, 5.5 DEO were noted per subject pre-DSS consultation; this reduced to 5.0 DEO after DSS advice (p < 0.001). No significant interaction was noticed with grade (F3, 48 = 0.71, p = 0.55). Reduction in DEO following DSS advice within each grade is shown in table 2. Overall, more DEOs were noted

Table 1: Study participants, cases and case episodes*

<table>
<thead>
<tr>
<th>Grade of subject</th>
<th>Consultant (%)</th>
<th>Registrar (%)</th>
<th>SHO (%)</th>
<th>Student (%)</th>
<th>Total Case episodes</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects invited to participate</td>
<td>27 (27.8)</td>
<td>33 (34)</td>
<td>20 (20.6)</td>
<td>17 (17.5)</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>Subjects who attempted at least one case (attempters)</td>
<td>18 (23.7)</td>
<td>24 (31.6)</td>
<td>19 (25)</td>
<td>15 (19.7)</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>Subjects who attempted at least six cases</td>
<td>16 (25.8)</td>
<td>18 (29)</td>
<td>15 (24.2)</td>
<td>13 (20.9)</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Subjects who completed all 12 cases (completers)</td>
<td>15 (28.8)</td>
<td>14 (26.9)</td>
<td>10 (19.2)</td>
<td>13 (25)</td>
<td>52</td>
<td></td>
</tr>
</tbody>
</table>

* Since each case was assessed more than once, each attempt by a subject at a case was termed as a ‘case episode’.

Table 2: Mean count of diagnostic errors of omission (DEO) pre-ISABEL and post-ISABEL consultation

<table>
<thead>
<tr>
<th>Grade of subject</th>
<th>DEO pre-ISABEL (SD)</th>
<th>DEO post-ISABEL (SD)</th>
<th>Reduction (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consultant</td>
<td>5.13 (1.3)</td>
<td>4.6 (1.4)</td>
<td>0.53 (0.7)</td>
</tr>
<tr>
<td>Registrar</td>
<td>5.64 (1.5)</td>
<td>5.14 (1.6)</td>
<td>0.5 (0.5)</td>
</tr>
<tr>
<td>SHO</td>
<td>4.4 (1.6)</td>
<td>4.1 (1.6)</td>
<td>0.3 (0.5)</td>
</tr>
<tr>
<td>Medical student</td>
<td>6.61 (1.3)</td>
<td>5.92 (1.4)</td>
<td>0.69 (0.7)</td>
</tr>
</tbody>
</table>

Mean DEO across all subjects (n = 52)*  
5.50 (1.6)  
4.98 (1.5)  
0.52 (0.6)  

*Total number of subjects who completed all 12 assigned cases
for easy cases compared to difficult cases pre- and post-DSS advice (2.17 vs. 2.05 and 2.0 vs. 1.8); however, this was not true for medical students as a subgroup (2.5 vs. 2.9). Improvement following DSS advice seemed greater for difficult cases for all subjects, although this was not statistically significant. These findings are summarized in table 3.

**Mean quality score changes**

624 case episodes from 52 subjects who had completed all assigned 12 cases were used for this analysis. Table 4 summarizes mean diagnostic quality scores pre- and post-ISABEL consultation, and the change in mean quality score for diagnoses, for each grade of subject. There was a significant change in the weighted mean of the diagnostic quality score (0.044; 95% confidence interval: 0.032, 0.054; \(p < 0.001\)). No significant interaction between grade and occasion was demonstrated. In 9/52 subjects (17.3%), the pre-DSS score for diagnostic quality was higher than the post-DSS score, indicating that subjects had lengthened their diagnostic workup without substantially improving its quality. Overall, the mean score for test-ordering plans increased significantly from 0.345 to 0.364 (an increase of 0.019, 95% CI 0.011–0.027, \(t_{51} = 4.91, p < 0.001\)); this increase was smaller for treatment plans (0.01, 95% CI 0.007–0.012, \(t_{51} = 7.15, p < 0.001\)).

**Number of irrelevant diagnoses**

624 case episodes from 52 subjects were used for this analysis. The results are illustrated in table 5. Overall, the mean count of diagnoses included by subjects in their workup pre-DSS advice was 3.9. This increased to 5.7 post-DSS consultation (an increase of 1.8 diagnoses). The increase was largest for medical students (a mean increase of 2.6 diagnoses) and least for consultants (1.4 diagnoses). The ANOVA showed significant interaction between grade and occasion (\(F_{3, 58} = 3.14, p = 0.034\)). The number of irrelevant diagnoses in the workup changed from 0.7 pre-DSS to 1.4 post-DSS advice (an increase of 0.7 irrelevant diagnoses, 95% CI 0.5–0.75). There was a significant difference in this increase across grades (most for medical students and least for consultants; 1.1 vs. 0.3 irrelevant diagnoses, \(F_{3, 48} = 6.33, p < 0.01\)). The increase in irrelevant diagnoses did not result in a corresponding increase in the number of irrelevant or deleterious tests and treatments (an increase of 0.09 tests and 0.03 treatment decisions).

**Additional diagnoses, tests and treatment decisions**

At least one ‘clinically important’ diagnosis was added by the subject to their differential diagnosis after ISABEL consultation in 94/751 case episodes (12.5%, 95% CI 10.1–14.9%). 47/76 (61.8%) subjects added at least one ‘clinically important’ diagnosis to their diagnostic workup after consultation. Overall, 130 ‘clinically important’ diagnoses were added after DSS advice during the experiment. In general, students were reminded to consider many more important diagnoses than consultants, although this was not statistically significant (44 vs. 26, Chi square \(p > 0.05\)); a similar gradient was seen for difficult cases, but DSS consultation seemed helpful even for easy cases. Similar proportions for tests and treatment items were smaller in magnitude (table 6). No clinically significant diagnoses were deleted after consultation. Important tests included by subjects in their pre-DSS plan were sometimes deleted from the post-DSS plan (64 individual items from 44 case episodes). An inappropriate test was added to the post-ISABEL list in 7/751 cases. 751 case episodes were used to examine the presence of the ‘most likely diagnosis’. Overall, the ‘most likely diagnosis/es’ were included in the pre-DSS diagnostic workup by subjects in 507/751 (67.5%) case episodes. This

### Table 3: Mean DEO count analyzed by level of case and subject grade

<table>
<thead>
<tr>
<th>Grade</th>
<th>Difficult cases</th>
<th>Easy cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-DSS</td>
<td>Post-DSS</td>
</tr>
<tr>
<td>Consultant</td>
<td>1.66</td>
<td>1.47</td>
</tr>
<tr>
<td>Registrar</td>
<td>2.21</td>
<td>1.93</td>
</tr>
<tr>
<td>SHO</td>
<td>1.3</td>
<td>1.2</td>
</tr>
<tr>
<td>Medical student</td>
<td>2.92</td>
<td>2.54</td>
</tr>
</tbody>
</table>

### Table 4: Mean quality scores for diagnoses broken down by grade of subject

<table>
<thead>
<tr>
<th>Grade</th>
<th>Mean pre-ISABEL score</th>
<th>Mean post-ISABEL score</th>
<th>Mean score change*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consultant</td>
<td>0.39</td>
<td>0.43</td>
<td>0.044</td>
</tr>
<tr>
<td>Registrar</td>
<td>0.40</td>
<td>0.44</td>
<td>0.038</td>
</tr>
<tr>
<td>SHO</td>
<td>0.45</td>
<td>0.46</td>
<td>0.032</td>
</tr>
<tr>
<td>Medical student</td>
<td>0.31</td>
<td>0.37</td>
<td>0.059</td>
</tr>
</tbody>
</table>

* Weighted average (all subjects)† | 0.383 | 0.426 | 0.044

* There was no significant difference between grades in terms of change in diagnosis quality score (one-way ANOVA \(p > 0.05\))
increased to 561/751 (74.7%) case episodes after DSS advice. The improvement was fully attributable to positive consultation effects (where the ‘most likely diagnosis’ was absent pre-DSS but was present post-DSS); no negative consultations were observed. Diagnostic accuracy pre-ISABEL was greatest for consultants (73%) and least for medical students (57%). Medical students gained the most after DSS advice (an absolute increase of 10%). Analysis performed to elucidate whether ISABEL was responsible for the changes seen in the rate of diagnostic error indicated that all additional diagnoses were indeed present in the system’s list of diagnostic suggestions.

**Time intervals**

Reliable time data was available for 633/751 episodes (table 7). Median time taken for subjects to abstract clinical features and record their initial clinical decisions on the trial website was 6 min (IQR 4–10 min); median time taken to examine ISABEL’s suggestions and make changes to clinical decisions was 1 min (IQR 30 sec-2 min). Time taken for ISABEL to display its suggestions was less than 2 sec on all occasions.

**Discussion**

We have shown in this study that errors of omission occur frequently during diagnostic workup in an experimental setting, including in cases perceived as being common in routine practice. Such errors seem to occur in most subjects, irrespective of their level of experience. We have also demonstrated that it is possible to influence clinicians’ diagnostic workup and reduce errors of omission using a stand-alone diagnostic reminder system. Following DSS consultation, the quality of diagnostic, test-ordering and treatment decisions made by various grades of clinicians improved for a range of cases, such that a clinically important alteration in diagnostic decision-making resulted in 12.5% of all consultations (1 in 8 episodes of system use).

In a previous study assessing the impact of ILIAD and QMR, in which only diagnostically challenging cases were used in an experimental setting, Friedman et al showed that the ‘correct diagnosis’ was prompted by DSS use in approximately 1 in 16 consultations [48]. Although we used a similar experimental design, we used a mix of easy as well as difficult cases to test the hypothesis that incomplete workup was encountered in diagnostic conundrums as well as routine clinical problems. Previous evaluations of expert systems used the presence of the ‘correct’ diagnosis as the main outcome. We focused on clinical safety as the key outcome, preferring to use the inclusion of all ‘clinically important’ diagnoses in the workup as the main variable of interest. In acute settings such as emergency rooms and primary care, where an incomplete and evolving clinical picture results in considerable diagnostic uncertainty at assessment, the ability to generate a focused and ‘safe’ workup is a more clinically relevant outcome, and one which accurately reflects the nature of decision making in this environment [53]. Consequently, we defined diagnostic errors of omission at assessment as the ‘failure to consider all clinically important diagnoses (as judged by an expert panel working the same cases)’. This definition resulted in the ‘correct’ diagnosis, as well as other significant diagnoses, being included within the ‘minimum’ workup. Further, changes in test-ordering and treatment decisions were uniquely measured in this study as a more concrete marker of the impact of diagnostic decision support on the patient’s clinical management; we
were able to demonstrate an improvement in test-ordering in 1 in 10 system consultations, indicating that diagnostic DSS may strongly influence patient management, despite only offering diagnosis-related advice. Finally, the time expended during DSS consultation is an important aspect that has not been fully explored in previous studies. In our study, subjects spent a median of 6 minutes for clinical data entry (including typing in their unaided decisions), and a median of 1 minute to process the advice provided and make changes to their clinical decisions.

The research design employed in this study allowed us to confirm a number of observations previously reported, as well as to generate numerous unique ones. These findings relate to the operational consequences of providing diagnostic assistance in practice. In keeping with other DSS evaluations, different subject grades processed system advice in different ways, depending on their prior knowledge and clinical experience, leading to variable benefit. Since ISABEL merely offered diagnostic suggestions, and allowed the clinician to make the final decisions (acting as the 'learned intermediary') [54], in some cases, subjects ignored even important advice. In some other cases, they added irrelevant decisions or deleted important decisions after DSS consultation, leading to reduced net positive effect of the DDSS on decision making. For some subjects whose pre-DSS performance was high, a ceiling effect prevailed, and no further improvement could be demonstrated. These findings complement the results of our earlier system performance evaluation which solely focused on system accuracy and not on user interaction with DDSS. One of the main findings from this study was that consultants tended to generate shorter diagnostic workup lists containing the 'most likely' diagnoses, with a predilection to omit other 'important' diagnoses that might account for the patient's clinical features, resulting in a high incidence of DEO. Medical students generated long diagnostic workup lists, but missed many key diagnoses leading to a high DEO rate. Interestingly, all subject grades gained from the use of ISABEL in terms of a reduction in the number of DEO, although to varying degrees. Despite more DEOs occurring in cases considered to be routine in practice than in rare and difficult ones in the pre-DSS consultation phase, ISABEL advice seemed to mainly improve decision making for difficult cases, with a smaller effect on easy cases. The impact of DSS advice showed a decreasing level of beneficial effect from diagnostic to test-ordering to treatment decisions. Finally, although the time taken to process cases without DSS advice in this study compared favorably with the Friedman evaluation of QMR and ILIAD (6 min vs. 8 min), the time taken to generate a revised workup with DSS assistance was dramatically shorter (1 min vs. 22 min).

We propose a number of explanations for our findings. There is sufficient evidence to suggest that clinicians with more clinical experience resort to established pattern-recognition techniques and the use of heuristics while making diagnostic decisions [55]. While these shortcuts enable quick decision making in practice, and work successfully on most occasions, they involve a number of cognitive biases such as 'premature closure' and 'confirmation bias' that may lead to incomplete assessment on some occasions. On the other hand, medical students may not have developed adequate pattern-recognition techniques or acquired sufficient knowledge of heuristics to make sound diagnostic decisions. It may well be that grades at an intermediate level are able to process cases in an acute setting with a greater emphasis on clinical safety. This explanation may also account for the finding that subjects failed to include 'important' diagnoses during the assessment of easy cases. Recognition that a case was unusual may trigger a departure from the use of established pattern-recognition techniques and clinical shortcuts to a more considered cognitive assessment, leading to fewer DEO in these cases. We have shown that it is possible to reduce DEOs by the use of diagnostic reminders, including in easy cases, although subjects appeared to be more willing to revise their decisions for difficult cases on the basis of ISABEL suggestions. It is also possible that some subjects ignored relevant advice because the system's explanatory capacity was inadequate and did not allow subjects to sufficiently discriminate between the suggestions offered. User variability in summarizing cases may also explain why variable benefits were derived from ISABEL usage – subjects may have obtained different results depending on how they abstracted and entered clinical features. This user variability during clinical data entry has been demonstrated even with use of a controlled vocabulary in QMR [56]. We observed marked differences between users’ search terms for the same textual case; however, diagnostic suggestions did not seem to vary noticeably. This observation could be partially explained by the enormous diversity associated with various natural language disease descriptions contained within the ISABEL database, as well as by the system’s use of a thesaurus

Table 7: Time taken to process case simulations broken down by grade of subject

<table>
<thead>
<tr>
<th>Grade of Subject</th>
<th>Median time pre-ISABEL</th>
<th>Median time post-ISABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consultant</td>
<td>5 min 5 sec</td>
<td>42 sec</td>
</tr>
<tr>
<td>Registrar</td>
<td>5 min 45 sec</td>
<td>57 sec</td>
</tr>
<tr>
<td>SHO</td>
<td>5 min 54 sec</td>
<td>53 sec</td>
</tr>
<tr>
<td>Medical student</td>
<td>8 min 36 sec</td>
<td>3 min 42 sec</td>
</tr>
<tr>
<td>Overall</td>
<td>6 min 2 sec (IQR: 4:03 – 9:47)</td>
<td>1 min (IQR: 30 sec – 2:04)</td>
</tr>
</tbody>
</table>
that converts medical slang into recognized medical terms.

The diminishing level of impact from diagnostic to test-ordering to treatment decisions may be a result of system design – ISABEL does not explicitly state which tests and treatments to perform for each of its diagnostic suggestions. This advice is usually embedded within the textual description of the disease provided to the user. Study and system design may both account for the differences in time taken to process the cases. In previous evaluations, subjects processed cases without using the DSS in the first instance; in a subsequent step, they used the DSS to enter clinical data, record their clinical decisions, and processed system advice to generate a second diagnostic hypothesis list. In our study, subjects processed the case and recorded their own clinical decisions while using the DSS for clinical data entry. The second stage of the procedure only involved processing ISABEL advice and modifying previous clinical decisions. As such, direct comparison between the studies can be made only by the total time involved per case (30 min vs. 7 min). This difference could be explained by features in the system’s design that resulted in shorter times to enter clinical data and to easily process the advice provided.

The findings from this study have implications specifically for ISABEL as well as other diagnostic DSS design, evaluation and implementation. It is well recognized that the dynamic interaction between user and DSS plays a major role in their acceptance by physicians [57]. We feel that adoption of the ISABEL system during clinical assessment in real time is possible even with current computer infrastructure, providing an opportunity for reduction in DEO. Its integration into an EMR would allow further control on the quality of the clinical input data as well as provision of active decision support with minimum extra effort. Such an ISABEL interface has currently been developed and tested with four commercial EMRs [58]; this integration also facilitates iterative use of the system during the evolution of a patient’s condition, leading to increasingly specific diagnostic advice. The reminder system model aims to enable clinicians to generate ‘safe’ diagnostic workups in busy environments at high risk for diagnostic errors. This model has been successfully used to alter physician behavior by reducing errors of omission in preventive care [59]. It is clear from recent studies that diagnostic errors occur in the emergency room for a number of reasons. Cognitive biases, of which ‘premature closure’ and faulty context generation are key examples, contribute significantly [60]. Use of a reminder system may minimize the impact of some of these cognitive biases. When combined with cognitive forcing strategies during decision making, DDSS may act as ‘safety nets’ to reduce the incidence of omission errors in practice [61]. Reminders to perform important tests and treatment steps may also allow a greater impact on patient outcome [62].

The limitations of this study stem mainly from its experimental design. The repeated measures design raises the possibility that some of the beneficial effects seen in the study are a result of subjects ‘rethinking’ the case, or the consequence of a reflective process [63]. Consequently, ISABEL’s effects in practice could be related to the extra time taken by users in processing cases. We believe that any such effects are likely to be minimal since subjects did not actually process the cases twice during the study – a summary of the clinical features was generated by subjects when the case was displayed for the first time, and subjects could not review the cases while processing ISABEL suggestions in the next step. Subjects also spent negligible time between their first assessment of the cases and processing the diagnostic suggestions from the DSS. The repeated measures design provided the power to detect differences between users with minimal resources; a randomized design using study and control groups of subjects would have necessitated the involvement of over 200 subjects. The cases used in our study contained only basic clinical data gained at the time of acute assessment, and may have proved too concise or easy to process. However, this seems unlikely since subjects only took an average of 8 min to process even diagnostic conundrums prior to DSS use when ‘expert systems’ were tested. Our cases pertained to emergency assessments, making it difficult to generalize the results to other ambulatory settings. The ability to extract clinical features from textual cases may not accurately simulate a real patient encounter where missed data or ‘red herrings’ are quite common. The inherent complexity involved in patient assessment and summarizing clinical findings in words may lead to poorer performance of the ISABEL system in real life, since its diagnostic output depends on the quality of user input. As a corollary, some of our encouraging results may be explained by our choice of subjects: a few were already familiar with summarizing clinical features into the DSS. Subjects were not supervised during their case exercises since they may have performed differently under scrutiny,
raising the prospect of a Hawthorne effect [64]. The use of a structured website to explicitly record clinical decisions may have invoked the check-list effect, as illustrated in the Leeds abdominal pain system study [65]. The check list effect might also be invoked during the process of summarizing clinical features for ISABEL input; this may have worked in conjunction with ‘rethinking’ to promote better decision making pre-ISABEL. We also measured decision making at a single point in time, making it difficult to assess the effects of iterative usage of the DSS on the same patient. Finally, our definition of diagnostic error aimed to identify inadequate diagnostic workup at initial assessment that might result in a poor patient outcome. We recognize the absence of an evidence-based link between omission errors and diagnostic adverse events in practice, although according to the Schiff model [53], it seems logical to assume that avoiding process errors will prevent actual errors at least in some instances. In the simulated setting, it was not possible to test whether inadequate diagnostic workup would directly lead to a diagnostic error and cause patient harm. Our planned clinical impact assessment in real life would help clarify many of the questions raised during this experimental study.

Conclusion
This experimental study demonstrates that diagnostic omission errors are common during the assessment of easy as well as difficult cases. The provision of patient- and context-specific diagnostic reminders has the potential to reduce these errors across all subject grades. Our study suggests a promising role for the use of future reminder-based DSS in the reduction of diagnostic error. An impact evaluation, utilizing a naturalistic design and conducted in real life clinical practice, is underway to verify the conclusions derived from this simulation.

Competing interests
This study was conducted when the ISABEL system was available free to users, funded by the ISABEL Medical Charity (2001–2002). Dr Ramnarayan performed this research as part of his MD thesis at Imperial College London. Dr Britto was a Trustee of the ISABEL medical charity (non-remunerative post). Ms Tomlinson was employed by the ISABEL Medical Charity as a Research Nurse.

Since June 2004, ISABEL is managed by a commercial subsidiary of the Medical Charity called ISABEL Healthcare. The system is now available only to subscribers. Dr Ramnarayan now advises ISABEL Healthcare on research activities on a part-time basis; Dr Britto is now Clinical Director of ISABEL Healthcare; Ms Tomlinson is responsible for Content Management within ISABEL Healthcare. All three hold stock options in ISABEL Healthcare. All other authors declare that they have no competing interests.

Authors' contributions
PR conceived the study, contributed to the study design, analyzed the data and drafted the manuscript.
GR assisted with the design of the study and data analysis.
MC assisted with the study design, served as gold standard panel member, and revised the draft manuscript
VN assisted with study design, served as gold standard panel member, and helped with data analysis
MT assisted with data collection and analysis
PT assisted with study conception, study design and revised the draft manuscript
JW assisted with study conception, provided advice regarding study design, and revised the draft manuscript
JB assisted with study conception, study design and data analysis.

All authors read and approved the final manuscript.

Acknowledgements
The authors would like to thank Tina Sajjanhar for her help in allocating cases to specialties and levels of difficulty; Helen Fisher for data analysis; and Jason Maude of the ISABEL Medical Charity for help rendered during the design of the study.

Financial support: This study was supported by a research grant from the National Health Service (NHS) Research & Development Unit, London. The sponsor did not influence the study design; the collection, analysis, and interpretation of data; the writing of the manuscript; and the decision to submit the manuscript for publication.

References
1. Institute of Medicine: To Err is Human; Building a Safer Health System Washington DC: National Academy Press; 1999.


Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1472-6947/6/22/prepub
HOW ACCURATELY DOES A WEB-BASED DIAGNOSIS DECISION SUPPORT SYSTEM SUGGEST THE CORRECT DIAGNOSIS IN 50 ADULT NEJM CPC CASES?

Graber ML, Mathews A. 
Performance of a web-based clinical diagnosis support system for internists. 
Performance of a Web-Based Clinical Diagnosis Support System for Internists

Mark L. Graber, MD¹ ² and Ashlei Mathew¹ ²

¹Medical Service-111, VA Medical Center, Northport, NY 11768, USA; ²Department of Medicine, SUNY at Stony Brook, Stony Brook, NY, USA.

BACKGROUND: Clinical decision support systems can improve medical diagnosis and reduce diagnostic errors. Older systems, however, were cumbersome to use and had limited success in identifying the correct diagnosis in complicated cases.

OBJECTIVE: To measure the sensitivity and speed of “Isabel” (Isabel Healthcare Inc., USA), a new web-based clinical decision support system designed to suggest the correct diagnosis in complex medical cases involving adults.

METHODS: We tested 50 consecutive Internal Medicine case records published in the New England Journal of Medicine. We first either manually entered 3 to 6 key clinical findings from the case (recommended approach) or pasted in the entire case history. The investigator entering key words was aware of the correct diagnosis. We then determined how often the correct diagnosis was suggested in the list of 30 differential diagnoses generated by the clinical decision support system. We also evaluated the speed of data entry and results recovery.

RESULTS: The clinical decision support system suggested the correct diagnosis in 48 of 50 cases (96%) with key findings entry, and in 37 of the 50 cases (74%) if the entire case history was pasted in. Pasting took seconds, manual entry less than a minute, and results were provided within 2–3 seconds with either approach.

CONCLUSIONS: The Isabel clinical decision support system quickly suggested the correct diagnosis in almost all of these complex cases, particularly with key finding entry. The system performed well in this experimental setting and merits evaluation in more natural settings and clinical practice.

KEY WORDS: clinical diagnosis support systems; Isabel; internal medicine; diagnostic error; Google; decision support.

INTRODUCTION

The best clinicians excel in their ability to discern the correct diagnosis in perplexing cases. This skill requires an extensive knowledge base, keen interviewing and examination skills, and the ability to synthesize coherently all of the available information. Unfortunately, the level of expertise varies among clinicians, and even the most expert can sometimes fail. There is also a growing appreciation that diagnostic errors can be made just as easily in simple cases as in the most complex. Given this dilemma and the fact that diagnostic error rates are not trivial, clinicians are well-advised to explore tools that can help them establish correct diagnoses.

Clinical diagnosis support systems (CDSS) can direct physicians to the correct diagnosis and have the potential to reduce the rate of diagnostic errors in medicine.¹ ² The first-generation computer-based products (e.g., QMR—First Databank, Inc; Iliad—University of Utah; DXplain—Massachusetts General Hospital, Boston, MA) used precompiled knowledge bases of syndromes and diseases with their characteristic symptoms, signs, and laboratory findings. The user would enter findings from their own patients selected from a menu of choices, and the programs would use Bayesian logic or pattern-matching algorithms to suggest diagnostic possibilities. Typically, the suggestions were judged to be helpful in clinical settings, even when used by expert clinicians.³ These diagnosis support systems were also useful in teaching clinical reasoning.⁴ ⁵

Surprisingly and despite their demonstrated utility in experimental settings, none of these earlier systems gained widespread acceptance for clinical use, apparently related to the considerable time needed to input clinical data and their somewhat limited sensitivity and specificity.⁶ ⁷ A study of Iliad and QMR in an emergency department setting, for example, found that the final impression of the attending physician was found among the suggested diagnoses only 72% and 52% of the time, respectively, and data input required 20 to 40 minutes for each case.⁷

In this study, we evaluated the clinical performance of “Isabel” (Isabel Healthcare Inc, USA), a new, second generation, web-based CDSS that accepts either key findings or whole-text entry and uses a novel search strategy to identify candidate diagnoses from the clinical findings. The clinician first enters the key findings from the case using free-text entry (see Fig 1). There is no limit on the number of terms entered, although excellent results are typically obtained with entering just a few key findings. The program includes a thesaurus that facilitates recognition of terms. The program then uses natural language processing and search algorithms to compare these terms to those used in a selected reference library. For Internal Medicine cases, the library includes 6 key textbooks anchored around the Oxford Textbooks of Medicine, 4th Edition (2003) and the Oxford Textbook of Geriatric Medicine and 46 major journals in general and subspecialty medicine and toxicology. The search domain and results are filtered to take into account the patient’s age, sex, geographic location,
pregnancy status, and other clinical parameters that are either preselected by the clinician or automatically entered if the system is integrated with the clinician’s electronic medical record. The system then displays a total of 30 suggested diagnoses, with 10 diagnoses presented on each web page (see Fig. 2). The order of listing reflects an indication of the matching between the findings selected and the reference materials searched but is not meant to suggest a ranked order of clinical probabilities. As in the first generation systems, more detailed information on each diagnosis can be obtained by links to authoritative texts.

The Isabel CDSS was originally developed for use in pediatrics. In an initial evaluation, 13 clinicians (trainees and staff) at St Mary’s Hospital, London submitted a total of 99 case scenarios of hypothetical case presentations for different diagnoses, and Isabel displayed the expected diagnosis in 91% of these cases. Out of 100 real case scenarios gathered from 4 major teaching hospitals in the UK, Isabel suggested the correct diagnosis in 83 of 87 cases (95%). In a separate evaluation of 24 case scenarios in which the gold standard differential diagnosis was established by two senior academic pediatricians, Isabel decreased the chance of clinicians (a mix of trainees and staff clinicians) omitting a key diagnosis by suggesting a highly relevant new diagnosis in 1 of every 8 cases. In this study, the time to enter data and obtain diagnostic suggestions averaged less than 1 minute.

The Isabel clinical diagnosis support system has now been adapted for adult medicine. The goal of this study was to evaluate the speed and accuracy of this product in suggesting the correct diagnosis in a series of complex cases in Internal Medicine.

**METHOD**

We considered 61 consecutive “Case Records of Massachusetts General Hospital” (New England Journal of Medicine, vol. 350:166–176, 2004 and 353:189–198, 2005). Each case had an anatomical or final diagnosis, which was considered to be correct by the discussants. We excluded 11 cases (patients under the age of 10 and cases that focused solely on management issues). The 50 remaining case histories were copied and pasted into the Isabel data-entry field. The pasted material typically included the history, physical examination findings, and laboratory test results, but data from tables and figures were not submitted. Beyond entering the patient’s age, sex, and nationality, the investigators did not attempt to otherwise tailor the search strategy. Findings were compared to the recommended (but slightly more time consuming) strategy of entering discrete key findings, as compiled by a senior internist (MLG). Because the correct diagnosis is presented at the end of each case, data entry was not blinded.

**RESULTS**

Using the recommended method of manually entering key findings, the list of diagnoses suggested by Isabel contained the correct diagnosis in 48 of the 50 cases (96%). Typically 3–6 key findings from each case were used. The 2 diagnoses that were not suggested (progressive multifocal encephalopathy and nephrogenic fibrosing dermopathy) were not included in the Isabel database at the time of the study; thus, these 2 cases would never have been suggested, even with different keywords.
Using the copy/paste method for entering the whole text, the list of diagnoses suggested by Isabel contained the correct diagnosis in 37 of the 50 cases (76%). Isabel presented 10 diagnoses on the first web page and 10 additional diagnoses on subsequent pages up to a total of 30 diagnoses. Because users may tend to disregard suggestions not shown on later web pages, we tracked this parameter for the copy/paste method of data entry: The correct diagnosis was presented on the first page in 19 of the 37 cases (51%) or first two pages in 28 of the 37 cases (77%). Similar data were not collected for manual data entry because the order of presentation depended on which key findings were entered.

Both data entry approaches were fast: Manually entering data and obtaining diagnostic suggestions typically required less than 1 minute per case, and the copy/paste method typically required less than 5 seconds.

**DISCUSSION**

Diagnostic errors are an underappreciated cause of medical error,\(^\text{10}\) and any intervention that has the potential to produce correct and timely medical diagnosis is worthy of serious consideration. Our recent analysis of diagnostic errors in Internal Medicine found that clinicians often stop thinking after arriving at a preliminary diagnosis that explains all the key findings, leading to context errors and ‘premature closure’, where further possibilities are not considered.\(^\text{11}\) These and other errors contribute to diagnoses that are wrong or delayed, causing substantial harm in the patients affected. Systems that help clinicians explore a more complete range of diagnostic possibilities could conceivably reduce these types of error.

Many different CDSSs have been developed over the years, and these typically matched the manually entered features of the case in question to a database of key findings abstracted from experts or the clinical literature. The sensitivity of these systems was in the range of 50%–60%, and the time needed to access and query the database was often several minutes.\(^\text{3}\) More recently, the possibility of using Google to search for clinical diagnoses has been suggested. However, a formal evaluation of this approach on a subset of the same “Case Records” cases used in our study found a sensitivity of 58%,\(^\text{12}\) in the range of the first-generation CDSSs and unacceptably low for clinical use.

The findings of our study indicate that CDSS products have evolved substantially. Using the Isabel CDSS, we found that data entry takes under 1 minute, and the sensitivity in a series of highly complex cases approached 100% using entry of key findings. Entry of entire case histories using copy/paste functionality allowed even faster data entry but reduced sensitivity. The loss of sensitivity seemed primarily related to negative findings included in the pasted history and physical (e.g., “the patient denies chest pain”), which are treated as positive findings (chest pain) by the search algorithm.

There are several relevant limitations of this study that make it difficult to predict how Isabel might perform as a diagnostic aid in clinical practice. First, the results obtained here reflect the theoretical upper limit of performance, given that an investigator who was aware of the correct diagnosis entered the key findings. Entry of entire case histories using copy/paste functionality allowed even faster data entry but reduced sensitivity. The loss of sensitivity seemed primarily related to negative findings included in the pasted history and physical (e.g., “the patient denies chest pain”), which are treated as positive findings (chest pain) by the search algorithm.

There are several relevant limitations of this study that make it difficult to predict how Isabel might perform as a diagnostic aid in clinical practice. First, the results obtained here reflect the theoretical upper limit of performance, given that an investigator who was aware of the correct diagnosis entered the key findings. Entry of entire case histories using copy/paste functionality allowed even faster data entry but reduced sensitivity. The loss of sensitivity seemed primarily related to negative findings included in the pasted history and physical (e.g., “the patient denies chest pain”), which are treated as positive findings (chest pain) by the search algorithm.

There are several relevant limitations of this study that make it difficult to predict how Isabel might perform as a diagnostic aid in clinical practice. First, the results obtained here reflect the theoretical upper limit of performance, given that an investigator who was aware of the correct diagnosis entered the key findings. Entry of entire case histories using copy/paste functionality allowed even faster data entry but reduced sensitivity. The loss of sensitivity seemed primarily related to negative findings included in the pasted history and physical (e.g., “the patient denies chest pain”), which are treated as positive findings (chest pain) by the search algorithm.
will also clearly depend on the clinical expertise of the user and their facility in working with Isabel. A serious existential concern is whether presenting a clinician with dozens of diagnostic suggestions might be a distraction or lead to unnecessary testing. We have previously identified these trade-offs as an unavoidable cost of improving patient safety: The price of improving the odds of reaching a correct diagnosis is the extra time and resources consumed in using the CDSS and considering alternative diagnoses that might turn out to be irrelevant.\textsuperscript{14}

In summary, the Isabel CDSS performed quickly and accurately in suggesting correct diagnoses for complex adult medicine cases. However, the test setting was artificial, and the CDSS should be evaluated in more natural environments for its potential to support clinical diagnosis and reduce the rate of diagnostic error in medicine.

\textbf{Acknowledgements:} The administrative and library-related support of Ms. Grace Garey and Ms. Mary Lou Glazer is gratefully acknowledged. Funding was provided by the National Patient Safety Foundation. This study was approved by the Institutional Review Board (IRB).

\textbf{Conflict of interest statement:} None disclosed.

\textbf{Corresponding Author:} Mark L. Graber, MD; Medical Service-111, VA Medical Center, Northport, NY 11768, USA (e-mail: mark.graber@va.gov).

\textbf{REFERENCES}

HOW DOES A WEB-BASED DIAGNOSIS DECISION SUPPORT SYSTEM IMPROVE DIAGNOSIS ACCURACY IN A CRITICAL CARE SETTING?

Thomas NJ, Ramnarayan P, Bell P Et al.
An international assessment of a web-based diagnostic tool in critically ill children

Neal J. Thomas* a, Padmanabhan Ramnarayan b, Michael J. Bell c, Prabhat Maheshwari d, Shaun Wilson e, Emily B. Nazarian f, Lorri M. Phipps a, David C. Stockwell c, Michael Engel c, Frank A. Maffei f, Harish G. Vyas e and Joseph Britto g

a Penn State Children’s Hospital and The Pennsylvania State University College of Medicine, Hershey, PA, USA
b Children’s Acute Transport Service, London, UK
c Pediatric Critical Care Medicine, Children’s National Medical Center, Washington DC, USA
d Pediatric Intensive Care Unit, St Mary’s Hospital, Paddington, London, UK
e Pediatric Intensive Care Unit, Queen’s Medical Centre, Nottingham, UK
f Pediatric Critical Care Medicine, Golisano Children’s Hospital, University of Rochester at Strong, Rochester, NY, USA
g Isabel Healthcare Inc., Reston, VA, USA

Received 5 September 2007
Revised / Accepted 30 October 2007

Abstract. Improving diagnostic accuracy is essential. The extent of diagnostic uncertainty at patient admission is not well described in critically ill children. Therefore, we studied the extent that pediatric trainee diagnostic performance could be improved with the aid of a computerized diagnostic tool. Data regarding patient admissions to five Pediatric Intensive Care Units were collected. Information included patients’ clinical details, admitting team’s diagnostic workup and discharge diagnosis. An attending physician assessed each case independently and suggested additional diagnostic possibilities. Diagnostic accuracy was calculated using the discharge diagnosis as the gold standard. 206 out of 927 patients (22.2%) admitted to the PICUs did not have an established diagnosis at admission. The trainee teams considered a median of three diagnoses in their workup (IQR 3–5) and made an accurate diagnosis in 89.4% cases (95% CI 84.6%–94.2%). Diagnostic accuracy improved to 92.5% with use of the diagnostic tool alone, and to 95% with the addition of attending physicians’ diagnostic suggestions. We conclude that a modest proportion of admissions to these PICUs were characterized by diagnostic uncertainty during initial assessment. Although there was a relatively high accuracy rate of initial assessment in our clinical setting, it was further improved by both the diagnostic tool and the physicians’ diagnostic suggestions. It is plausible that the tool’s utility would be even greater in clinical settings with less expertise in critical illness assessment, such as community hospitals, or emergency departments of non-training institutions. The role of diagnostic aids in the care of critically ill children merits further study.

Keywords: Diagnosis, diagnostic reminder system, internet, pediatrics

*Address for correspondence: Neal J. Thomas, M.D., M.Sc, Pediatric Critical Care Medicine, Penn State Children’s Hospital, 500 University Drive, MC H085, Hershey, PA 17033, USA. Tel.: +1 717 531 5337; Fax: +1 717 531 0809; E-mail: nthomas@psu.edu.
1. Introduction

Accurate diagnosis of medical conditions has been of paramount importance for centuries and faulty diagnostic evaluation can lead to patient harm and poor medical management. Three etiologies of diagnostic errors have been recognized within the medical community: no fault errors, system-related errors and cognitive errors [20]. No-fault errors occur either when disease presentation is unusual and unpredictable or when patients are uncooperative. System-related errors occur as a result of either organizational flaws or mechanical problems within the institution. Cognitive errors occur due to lack of adequate medical knowledge or inadequate data collection.

With the explosion of medical literature at exponential rates, avoidance of cognitive errors using emerging technology may be the most efficient method for improving diagnostic skills. A large body of evidence suggests that diagnostic error by clinicians poses a significant problem across a wide variety of settings, comprising 10–30% of adverse medical events (AME) [9,22]. One recent observational study suggested that 10% of AME in adult critical care patients were related to misdiagnosis [16]. Moreover, autopsy studies have demonstrated that major diagnoses are missed in critically ill adults and children who die in intensive care units [2,4,8,17].

We hypothesized that a web-based diagnostic reminder system would produce additional differential diagnosis possibilities for children admitted to five international PICUs who were admitted without an established diagnosis. We further hypothesized that the web-based tool would increase diagnostic accuracy by input of the findings recorded in the medical record admission history and physical examination. To do this, we compared the accuracy of determining the discharge diagnosis for (i) the admitting team, (ii) an unbiased board-certified/board eligible pediatric intensivist and (iii) ISABEL, a web-based diagnostic reminder system (www.isabelhealthcare.com) [14,21].

2. Materials and methods

This prospective study was conducted during a three-month period at five pediatric intensive care units (PICUs), two from the United Kingdom (UK) and three from the United States (USA). The selection of ICUs was based on working relationships between the five primary investigators in each site, with a goal of studying institutions that had diverse patient populations, different patient volumes, and a diverse culture. All patient data collection was approved by the research ethics committees (institutional review boards) at the participating centers prior to initiation of the study. The requirement for informed consent was waived. All admissions to the PICU during the study period were screened by a designated study investigator at each center. Medical admissions to the PICU without an established diagnosis were eligible for study. All surgical admissions and medical admissions with known primary diagnoses (such as culture-proven sepsis, diabetic ketoacidosis and others) were excluded. All PICUs utilize various medical personnel (including residents from pediatrics or pediatric subspecialties, clinical fellows, advanced practice nurses (APN), physician assistants and pediatric intensivists) as their daily clinical team.

2.1. Diagnosis by admitting team

At each PICU, the screening investigator (pediatric critical care medicine fellow, pediatric resident or pediatric APN) reviewed all admissions and enrolled eligible patients. Importantly, these investigators neither admitted the child to the PICU nor participated in the clinical care given to the child so that
the utility of the tool itself could be tested without bias from difficulties with the system or changes in routine practice. The medical records of eligible patients were reviewed and presenting symptoms and physical findings and the differential diagnosis (including all diagnostic possibilities) was extracted. The screening investigator further determined the discharge diagnosis based on review of the medical record at discharge.

2.2. Diagnosis by ISABEL

The screening investigator at each center used a customized study version of ISABEL to securely log in, enter, and store data on eligible patients. Each patient was assigned a unique study number, and the date and time of electronic access was logged. Clinical data from the chart were entered and was used to drive the diagnostic tool’s algorithm for production of diagnostic possibilities.

ISABEL is a Web based diagnostic decision support system (DSS). Computerized decision support tools utilize two or more pieces of clinical data to provide patient-specific recommendations; assist in diverse clinical processes such as diagnosis, prescribing or clinical management; and consist of an underlying knowledge base, user interface for data input, and an inference engine [23]. The ISABEL knowledge base consists of > 100,000 raw text documents describing > 10,000 diseases. Documents are drawn from multiple resources such as reputed textbooks (e.g. Oxford Textbook of Medicine) and review articles in journals. The user interface permits data entry in the form of a free text description of clinical findings, without the need for a specific terminology. The inference engine utilizes statistical natural language processing software to search through the entire database of text documents and returns all disease descriptions that match the clinical features entered. This process is similar to a search engine, although much more sensitive and specific. The software indexes each document in the database daily, and extracts relevant concepts based on frequency of their occurrence, uniqueness and proximity to other concepts. Each diagnosis in the database is assigned clinical weighting scores, derived from expert opinion, to reflect its prevalence by age group (e.g. child), gender and geographical region (e.g. North America). In the current study, when investigators entered historical data and physical findings noted at admission for each patient as search terms, and applied appropriate age, gender and region filters, 10–12 diagnoses that were matched within the database were displayed, arranged by body system (e.g. cardiovascular, respiratory) rather than by clinical probability. Hyperlinks from each diagnostic suggestion led to text documents that matched the search query. There was no limit on the number of search terms that could be entered, although the tool performed well even with 3–5 terms.

A number of different diagnostic DSS have been studied in the past. Tools such as De Dombal’s abdominal pain system were developed using data from > 6000 patients with proven surgical causes of abdominal pain and their clinical findings. Using probabilistic techniques, this tool assisted clinicians in differentiating surgical from non-surgical causes of abdominal pain [5]. Other DSS such as Dxplain, Quick Medical Reference, and ILIAD utilized a complex database of relationships between hundreds of diseases and thousands of clinical findings (e.g. rheumatoid arthritis and fever) derived from expert opinion to provide diagnostic hypotheses (i.e. quasi-probabilistic systems) [1]. Keeping such knowledge bases up to date was a challenge - an expert panel needed to frequently appraise new information, and make significant changes to existing relationships within the database. The ISABEL system utilizes a novel mechanism to keep its knowledge base current. When updated textbooks or reviews became available on a topic, old documents are simply replaced with new text in the database without any expert input. A small content management team searches monthly for newer editions of textbooks and reviews and updates the database on a quarterly basis. Expert input is only necessary to modify clinical
weighting scores, and is undertaken as soon as important new information is available (e.g. SARS virus outbreak in Hong Kong). Since the tool is web-based, updates are instantly available to users. A more detailed description of the methodology involved in the development and maintenance of this tool is available [13].

2.3. Diagnosis by board certified/board eligible pediatric intensivist

At each center, a pediatric intensivist independently examined the clinical data available at the time of admission for each patient and the admitting PICU team’s diagnostic workup; the discharge diagnosis was withheld from the attending physician during this process. They then highlighted further diagnoses from within the computerized tool’s list that they considered clinically relevant or significant for patient assessment. This judgment was based on whether consideration of that particular diagnosis in the workup would have triggered a specific action (either performing further focused clinical examination or a particular test or treatment). If two closely related (or nearly synonymous) diagnoses were displayed in the list, only one was taken into account. This investigator was then able to propose additional diagnoses they felt were relevant to generate an additional list of potential diagnoses.

2.4. Diagnostic accuracy assessment

Diagnostic accuracy was examined by matching the discharge diagnosis with lists generated by (i) the admitting team, (ii) ISABEL and (iii) the study intensivist. The diagnosis was considered accurate if the discharge diagnosis was located within the list of diagnoses generated by these three methods. An a priori calculation demonstrated that 250 patients would be required to identify the frequency of additional significant diagnoses in the diagnostic tool’s list assuming 20% of the patients were without an established diagnosis (with 80% power, type I error 5%). Data collection was planned for three months at each unit, based on an estimated annual admission rate of approximately 4000 patients. Diagnostic accuracy rates of admission teams, ISABEL and attending physicians were calculated based on the discharge diagnosis. Results were calculated as proportions, and expressed as percentages with 95% confidence interval limits. The Chi-square test was used to detect statistically significant differences between proportions (defined as p value $< 0.05$). Factors influencing diagnostic accuracy were analyzed by entering the variables into a multiple logistic regression model.

3. Results

Data were collected from all five participating centers during the study. All centers admitted a general mix of medical and surgical pediatric patients. The characteristics of each of the participating centers are shown in Table 1. A total of 927 patients were admitted to the five PICUs between May and July 2003. The majority of patients was either surgical/post-surgical or had an established diagnosis by the time they were admitted to the PICU (721/927, 77.8%). A total of 206 patients (22.2%) had no established diagnosis at the time of PICU admission, and were therefore eligible for study; 69 were from UK units and 137 from the US units. Admission data for each center are presented in Table 2. Of these, 45 lacked discharge diagnoses and were thereby excluded. The remaining 161 patients were therefore analyzed. There were significant differences between the various centers. The number of patients without an established diagnosis at admission varied between center, ranging from 19.6% to 77.8% (mean 38.9%;
Table 1
Key baseline characteristics of the participating critical care units

<table>
<thead>
<tr>
<th>Country</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>UK</td>
<td>UK</td>
<td>USA</td>
<td>USA</td>
<td>USA</td>
</tr>
<tr>
<td>Total number of beds</td>
<td>6</td>
<td>10</td>
<td>12</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>Medical admissions in 2003 (%)</td>
<td>143 (35)</td>
<td>363 (77)</td>
<td>289 (37)</td>
<td>402 (62)</td>
<td>1001 (67)</td>
</tr>
<tr>
<td>Surgical admissions in 2003 (%)</td>
<td>267 (65)</td>
<td>106 (23)</td>
<td>483 (63)</td>
<td>248 (38)</td>
<td>492 (33)</td>
</tr>
<tr>
<td>Total annual admissions*</td>
<td>410</td>
<td>469</td>
<td>772</td>
<td>650</td>
<td>1493</td>
</tr>
</tbody>
</table>

* Total annual admissions at all participating centers: 3794. (C = center; UK = United Kingdom; US = United States).

Table 2
Breakdown of patient admissions according to diagnostic groups: May-July 2003

<table>
<thead>
<tr>
<th>Known surgical diagnosis (%)</th>
<th>Known medical diagnosis (%)</th>
<th>Unknown diagnosis (%)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>67 (65.0)</td>
<td>8 (7.8)</td>
<td>28 (27.2)</td>
</tr>
<tr>
<td>C2</td>
<td>22 (22.7)</td>
<td>33 (34)</td>
<td>42 (43.3)</td>
</tr>
<tr>
<td>C3</td>
<td>117 (61.3)</td>
<td>30 (15.7)</td>
<td>44 (23)</td>
</tr>
<tr>
<td>C4</td>
<td>60 (38.2)</td>
<td>78 (49.7)</td>
<td>19 (12.1)</td>
</tr>
<tr>
<td>C5</td>
<td>132 (34.8)</td>
<td>174 (45.9)</td>
<td>73 (19.2)</td>
</tr>
<tr>
<td>Total</td>
<td>398 (42.9)</td>
<td>323 (34.8)</td>
<td>206 (22.2)</td>
</tr>
</tbody>
</table>

* Difference between centers was statistically significant (chi-square test, \( p < 0.05 \)). (C = center).

Chi-square test, \( p < 0.05 \). The proportion of surgical admissions on the units ranged from 23.7% to 65% (mean 42.9%).

The admitting team’s differential list contained a median of three diagnoses per case (IQR 3–5) with a range of 1–12 diagnoses and contained the discharge diagnosis in 144/161 cases (89.4%). In five of the remaining 17 cases in which the admitting team did not include the discharge diagnosis in their initial workup, the diagnostic tool displayed the final diagnosis in the list of suggestions. ISABEL generated a differential list including between 10 and 12 diagnoses and this list contained the final diagnosis in 149/161 cases (92.5%). The attending intensivist accurately provided the final diagnosis in a further four cases, leading to an overall accuracy rate of 92% (when combined with the admitting team) and 95% (combined with the admitting team and the diagnostic tool). Despite the fact that the diagnostic tool displayed the discharge diagnosis in five patients where the admitting team had failed to consider it, the attending physician was able to identify only two of these as being clinically relevant to the case. The admitting team’s diagnostic accuracy appeared to be a function of the age of the patient, but not the number of diagnoses considered at admission or the diagnostic group.

4. Discussion

In summary, we have demonstrated that within tertiary care PICUs, approximately 20% of critically ill children will be admitted without an established diagnosis and that the clinical team makes an accurate preliminary diagnosis almost 90% of the time. This rate is improved by use of the ISABEL tool as well as a higher degree of clinical training and experience from an intensivist. As we chose to examine every PICU admission during the study period, this tool may have some added utility in a variety of patient populations, particularly those that are complex diagnostic challenges.
In general, it is believed that improvements in technology will ultimately lead to improved outcomes for patients in the health care system of all countries. While this might be true in some instances, such as the prevention of medication errors [11], there has also been some evidence that increased need for technology may also lead to increased patient harm [7]. To our knowledge, ISABEL is a unique system in that it collates data from multiple sources including textbooks, journal articles and reviews to generate a potential diagnostic list for the user. However, its ability to improve outcome has yet to be demonstrated, and the direct impact on patient outcome related to use of the system is an area of research that requires further clinical study. Estimates of diagnostic uncertainty in critical care have come from various sources. In extreme circumstances, persistent diagnostic uncertainty is often only resolved by post mortem studies. A number of autopsy studies have concluded that the rate of clinically significant missed diagnoses in intensive care patients ranges from 2.1% to 21.2% [10,15,19]. In addition, it is reported that in 17% of pediatric autopsies in critical care, knowledge of the diagnosis pre-mortem would have influenced survival by affecting decision making [3]. Therefore, it is crucial to suspect and confirm the correct diagnosis early in order to improve outcome in these children. The rate of missed diagnoses directly leading to death is higher in patients with a shorter stay in critical care, while unexpected minor findings are common in patients with a longer ICU stay, presumably due to the addition of new medical problems while undergoing critical care treatments, either nosocomial or iatrogenic [6]. These post-mortem studies suggest a significant burden of missed diagnoses, and imply that diagnosis constitutes an important part of critical care decision making. We failed to find such a critical burden of missed diagnoses in our study and it did not appear that the missed diagnoses in our patients significantly altered patient care or outcome. However, prospective evaluation of diagnostic errors in the critical care environment suggests that they constitute only a small proportion of medical error; medication-related errors are the most common cause for AME in this setting [16]. A more rigorous evaluation of the ISABEL tool in a variety of clinical situations may lead to improved knowledge about missed diagnoses as well as a more thorough evaluation of ISABEL’s usefulness.

We aimed to explore the role of a diagnostic tool in critical illness as part of this preliminary study. Recently, ISABEL has been shown to be an effective reminding system in adult emergency departments, with 95% accuracy in presenting the user with the final diagnosis and 90% accuracy in obtaining “must-not-miss” diagnoses [12]. We undertook the current study in the Pediatric Intensive Care Units of the five centers for several reasons. First, we wanted to demonstrate the utility of ISABEL in a complex population of children with complex medical conditions, thereby collecting information on a very challenging population for study. Second, we believe it is important to demonstrate the utility of a tool in diverse patient groups. As three of the PICUs were in the USA and two were in the UK and all units showed similar ISABEL performance, we believe that this highlights the potential utility and generalizability of the tool. Moreover, the patient volumes at the five sites differed by more than 300%, again arguing for the utility of the tool in a diverse population.

The most interesting question that can be asked regarding our data and device is whether a clinician can improve his/her own diagnostic ability in collaboration with the use of ISABEL. It is not intended that ISABEL is to become a substitute for clinical acumen, but rather an ancillary tool to establish potential diagnoses. We believe that this is consistent with others who believe that the performance of the intended user working in combination with the tool is greater than the performance of the user alone [18]. Obviously, this necessitates that the user-computer interaction must be seamless and user-friendly. None of the investigators in our subset of more than a dozen had difficulty in utilizing the ISABEL tool to generate useful diagnostic lists. Again, further studies with more diverse clinician populations may also demonstrate whether this interface is sufficient for all users’ needs.
This study has a number of limitations. Patient selection of only cases without an established diagnosis at admission to ICU may not accurately indicate the requirement or need for diagnostic decision making. Since diagnoses are being constantly made and adjusted within the first hours of PICU admission, it is possible that ISABEL might have other uses than the method we tested. Our methods also did not allow for testing in other settings, such as the emergency department, where less data are available and diagnostic uncertainty might be considerably greater. Second, we only studied this tool in tertiary care facilities with well-trained clinicians and highly-productive trainees. The results of a similar study to this in a smaller unit are unpredictable at this time. It is possible that the experience of the trainees masked the true utility of the ISABEL device. Conversely, it is possible that in a smaller unit the diagnoses might be less challenging and this might make ISABEL’s utility even less. Only future studies can answer this question. Lastly, we intentionally excluded surgical cases in our study population even though they made up almost over 50% of total admissions to the PICUs. Therefore, it is not possible to assess the utility of ISABEL in diagnosing common pediatric surgical emergencies such as ruptured appendix or other disorders.

In conclusion, ISABEL improved the diagnostic accuracy of the clinical team caring for children without an established diagnosis upon PICU arrival. Future studies demonstrating ISABEL’s utility in other patient populations will be required to fully assess whether web-based tools can improve children’s outcome after critical illness. It will be important to determine if implementation of tools such as ISABEL can be seamlessly integrated within clinical practice and if subsequent versions of the tool can be sufficiently rigorous to advance as the field of pediatrics evolves and changes. This challenge must be met by constant update and maintenance of diagnostic tools, as well as ongoing evaluation of the efficacy of these tools in practice.

Acknowledgements

The computerized diagnostic tool studied was provided free for registered users by the Isabel Medical Charity at the time of this study. It is now managed by a commercial entity called Isabel Healthcare and is available only to subscribers. Dr P Ramnarayan currently advises Isabel Healthcare on research activities on a part-time basis. Dr J Britto is Clinical Director of Isabel Healthcare Inc, USA. All the other authors have no competing interests.

References

Y. Y. Han, J. A. Carcillo, S. T. Venkataraman, R. S. Clark, R. S. Watson, T. C. Nguyen et al., Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system, *Pediatrics* **116** (2005), 1506–1512.


WHAT IS THE MAGNITUDE OF DIAGNOSIS ERROR?

Gordon D. Schiff, Seijeoung Kim, Richard Abrams et al.

Diagnosing Diagnosis Errors: Lessons from a Multi-institutional Collaborative Project.
Advances in Patient Safety 2005; 2:255-278
Diagnosing Diagnosis Errors: Lessons from a Multi-institutional Collaborative Project


Abstract

Background: Diagnosis errors are frequent and important, but represent an underemphasized and understudied area of patient safety. Diagnosis errors are challenging to detect and dissect. It is often difficult to agree whether an error has occurred, and even harder to determine with certainty its causes and consequence. The authors applied four safety paradigms: (1) diagnosis as part of a system, (2) less reliance on human memory, (3) need to open “breathing space” to reflect and discuss, (4) multidisciplinary perspectives and collaboration. Methods: The authors reviewed literature on diagnosis errors and developed a taxonomy delineating stages in the diagnostic process: (1) access and presentation, (2) history taking/collection, (3) the physical exam, (4) testing, (5) assessment, (6) referral, and (7) followup. The taxonomy identifies where in the diagnostic process the failures occur. The authors used this approach to analyze diagnosis errors collected over a 3-year period of weekly case conferences and by a survey of physicians. Results: The authors summarize challenges encountered from their review of diagnosis error cases, presenting lessons learned using four prototypical cases. A recurring issue is the sorting-out of relationships among errors in the diagnostic process, delay and misdiagnosis, and adverse patient outcomes. To help understand these relationships, the authors present a model that identifies four key challenges in assessing potential diagnosis error cases: (1) uncertainties about diagnosis and findings, (2) the relationship between diagnosis failure and adverse outcomes, (3) challenges in reconstructing clinician assessment of the patient and clinician actions, and (4) global assessment of improvement opportunities. Conclusions and recommendations: Finally the authors catalogue a series of ideas for change. These include: reengineering followup of abnormal test results; standardizing protocols for reading x-rays/lab tests, particularly in training programs and after hours; identifying “red flag” and “don’t miss” diagnoses and situations and use of manual and automated check-lists; engaging patients on multiple levels to become “coproducers” of safer medical diagnosis practices; and weaving “safety nets” to mitigate harm from uncertainties and errors in diagnosis. These change ideas need to be tested and implemented for more timely and error-free diagnoses.
Introduction

Diagnosis errors are frequent and important, but represent an underemphasized and understudied area of patient-safety.1–8 This belief led us to embark on a 3-year project, funded by the Agency for Healthcare Research and Quality (AHRQ), to better understand where and how diagnosis fails and explore ways to target interventions that might prevent such failures. It is known that diagnosis errors are common and underemphasized, but they are also challenging to detect and dissect. It is often difficult even to agree whether or not a diagnosis error has occurred.

In this article we describe how we have applied patient safety paradigms (blame-free reporting/reviewing/learning, attention to process and systems, an emphasis on communication and information technology) to better understand diagnosis error.2, 7–9

We review evidence about the types and importance of diagnosis errors and summarize challenges we have encountered in our review of more than 300 cases of diagnosis error. In the second half of the article, we present lessons learned through analysis of four prototypical cases. We conclude with suggested “change ideas”—interventions for improvement, testing, and future research.

Although much of the patient safety spotlight has focused on medication errors, two recent studies of malpractice claims revealed that diagnosis errors far outnumber medication errors as a cause of claims lodged (26 percent versus 12 percent in one study;10 32 percent versus 8 percent in another study11). A Harris poll commissioned by the National Patient Safety Foundation found that one in six people had personally experienced a medical error related to misdiagnosis.12 Most medical error studies find that 10–30 percent (range = 0.6–56.8 percent) of errors are errors in diagnosis (Table 1).1–3, 5, 11, 13–21 A recent review of 53 autopsy studies found an average rate of 23.5 percent major missed diagnoses (range = 4.1–49.8 percent). Selected disease-specific studies (Table 2),6, 22–32 also show that substantial percentages of patients (range = 2.1 – 61 percent) experienced missed or delayed diagnoses. Thus, while these studies view the problem from varying vantage points using heterogeneous methodologies (some nonsystematic and lacking in standardized definitions), what emerges is compelling evidence for the frequency and impact of diagnosis error and delay.

Of the 93 safety projects funded by AHRQ, only 1 is focused on diagnosis error, and none of the 20 evidence-based AHRQ Patient Safety Indicators directly measures failure to diagnose.33 Nonetheless, for each of AHRQ’s 26 “sentinel complications” (e.g., decubitus ulcer, iatrogenic pneumothorax, postoperative septicemia, accidental puncture/laceration), timely diagnosis can be decisive in determining whether patients experience major adverse outcomes. Hence, while diagnosis error remains more in the shadows than in the spotlight of patient safety, this aspect of clinical medicine is clearly vulnerable to well-documented failures and warrants an examination through the lens of modern patient safety and quality improvement principles.
Table 1. General medical error studies that reported errors in diagnosis

<table>
<thead>
<tr>
<th>Author, Year</th>
<th>Context/Design</th>
<th>Diagnosis errors/Total errors (%)</th>
<th>Comments/Types of Diagnostic Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flannery FT (1991)</td>
<td>Physicians and surgeons update, St. Paul, MN. Malpractice claims data reviewed</td>
<td>1126/7233 (27.4%) failure to diagnose</td>
<td>Top 5 diagnoses (cancer, circulatory/thrombosis, fracture/dislocation, lack of attendance, infection)</td>
</tr>
<tr>
<td>Leape LL (1991)</td>
<td>AE in hospitalized patients</td>
<td>168/1276 (13.8%)</td>
<td>Failure to use indicated tests, act on test, appropriate test, delay in diagnosis, practicing outside area of expertise</td>
</tr>
<tr>
<td>Bogner M (1994)</td>
<td>Medical practice study, 1984, AE in NY hospitals (Leape)</td>
<td>11731/68645 (17.1%)</td>
<td>17.1% preventable errors due to diagnostic errors. 71% of them were negligent.</td>
</tr>
<tr>
<td>Bhasale AL (1998)</td>
<td>Austrian GP, Diagnostic incidents</td>
<td>275/805 (34.2%)</td>
<td>Calculated rates (N = 275): missed diagnosis (40 per 100), delayed (34), misdiagnosed (23), and diagnostic procedural (14)</td>
</tr>
<tr>
<td>Wilson RM (1999)</td>
<td>The Quality in Australian Health Care Study</td>
<td>267/470 (56.8%) of AEs due to delays were delays in diagnosis, 191 (40.6%) treatment delay. 198/252 (78.6%) investigation category were investigation not performed, 39 (15.5%) not acted on, 9 (3.6%) inappropriate.</td>
<td>1922/2351 AEs (81.8%) associated w/ human errors, 470 (20%) delays in diagnosis or treatment, 460 (19.6%) treatment, and 252 (10.7%) investigation. 1922 AEs w/ 2940 causes identified. Causes of human errors: 465/2940 (15.6%) failure to synthesis/decide/act on information, 346 (11.8%) failure to request, arrange an investigation.</td>
</tr>
<tr>
<td>Weingart S (2000)</td>
<td>Potential adverse events</td>
<td>29/110 (26.4%) process of care problems associated w/ diagnosis</td>
<td>18/110 (16.4%) inadequate evaluation, 7 (6.4%) diagnostic error, 4 (3.6%) delayed consultation.</td>
</tr>
<tr>
<td>Neale G (2001)</td>
<td>AE in England hospitals</td>
<td>5/118 (0.6%)</td>
<td>118/840 (14%) had AE. 57% of all AE, cognitive. 5/118 (0.6%) of admissions were associated with incorrect diagnoses. 2 missed heart failures, 2 incorrect assessment of abdominal pain, and 1 missed fracture.</td>
</tr>
<tr>
<td>Baker GR (2002)</td>
<td>Mailed survey questionnaires regarding patient safety issues, Canadian health care facilities, and colleges and associations</td>
<td>1/25 (4%) of health care errors in health care facilities, 9/21 (39%) in colleges/associations</td>
<td>Human factors, 3 (12%), 9 (39%); competency, 5 (20%), 2 (9%)</td>
</tr>
<tr>
<td>Author, Year</td>
<td>Context/Design</td>
<td>Diagnosis errors/Total errors (%)</td>
<td>Comments/Types of Diagnostic Errors</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td>-----------------------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>JCAHO Sentinel event advisory group (2002)</td>
<td>ED sentinel event</td>
<td>55/23 (42%)</td>
<td>55 delays in treatment, due to misdiagnosis (42%), test results availability (13%), delayed initial assessment (7%). Most frequent missed diagnosis: meningitis 7/23 (30%), cardiac disease, PE, trauma, asthma, neurologic disorder</td>
</tr>
<tr>
<td>Makeham MA (2002)</td>
<td>GP in 6 countries</td>
<td>17/104 (13%) in Austria, 55/236 (19%) in other countries</td>
<td>Process errors 104/134 (78%) in Austria, 235/301 (78%), Of the process errors, investigation errors (lab errors, diagnostic imaging errors, and others) were 13%, 19%</td>
</tr>
<tr>
<td>Medical malpractice lawyers and attorneys online (2002)</td>
<td>Medical malpractice</td>
<td>40%</td>
<td>Failure-to-diagnose, failed to diagnosis in timely fashion</td>
</tr>
<tr>
<td>Chaudhry S (2003)</td>
<td>Error detection by attending physicians, general medicine inpatients</td>
<td>12/63 (19.1%) of all errors, 8/39 (20.5%) of near misses, and 4/24 (16.7%) of AEs were related with diagnostic errors</td>
<td>55/528 (10.4%) patients admitted to the hospitalists had at least 1 error</td>
</tr>
<tr>
<td>Kravitz RL (2003)</td>
<td>Malpractice claims data, 4 specialties</td>
<td>3–8%</td>
<td>1371 claims, Failure to appropriate diagnostic testing/monitoring</td>
</tr>
<tr>
<td>Phillips R (2004)</td>
<td>Malpractice claims in primary care</td>
<td>26126 claims peer reviewed, 592/26126 (23%) related with negligent</td>
<td>2003/5921 (34%) associated with diagnosis error, 16% failure to monitor case, 4% failure/delay in referral, 2% failure to recognize a complication</td>
</tr>
</tbody>
</table>

AE = adverse events; GP = general practitioners; JCAHO = Joint Commission for Accreditation of Healthcare Organizations; ED = emergency department
Table 2. Illustrative disease-specific studies of diagnosis errors

<table>
<thead>
<tr>
<th>Author</th>
<th>Disease/Design</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steere AC (1993)</td>
<td>Lyme disease</td>
<td>Overdiagnosis of Lyme disease: patients given the diagnosis but on review did not meet criteria for diagnosis. - 452/788 (57%)</td>
</tr>
<tr>
<td>Cravan ER (1994)</td>
<td>Glaucoma</td>
<td>Diagnosis error in glaucoma claims/lawsuits - 42/194 (21.7%)</td>
</tr>
<tr>
<td>Lederle FA (1994)</td>
<td>Ruptured abdominal aortic aneurysm</td>
<td>Ruptured aneurysm diagnosis initially missed - 14/23 (61%)</td>
</tr>
<tr>
<td>Mayer PL (1996)</td>
<td>Symptomatic cerebral aneurysm</td>
<td>Patients initially misdiagnosed - 54/217 (25%, ranging from 13% to 35% at 4 sites); Of these, 26/54 (48%) deteriorated before definite treatment; erroneous working diagnoses in 54 pts; 8 (15%) viral meningitis, 7 (13%) migraine, 7 (13%) headache of uncertain etiology</td>
</tr>
<tr>
<td>Williams V (1997)</td>
<td>Brain and spinal cord biopsies</td>
<td>Errors based on &quot;second opinion&quot; reviews - 214/500 (43%); 44/214 (20%) w/ serious complications, 96 (45%) w/ substantial, 50 (23%) w/ minor errors</td>
</tr>
<tr>
<td>Clark S (1998)</td>
<td>Pyrogenic spinal infection</td>
<td>Delay in diagnosis - 41/69 (59.4%); more than 1 month of back/neck pain before specialist referral</td>
</tr>
<tr>
<td>Arbiser ZK (2000)</td>
<td>Soft tissue pathology</td>
<td>Second option reviews: minor discrepancy - 20/266 (7.5%); major discrepancy - 65 (25%)</td>
</tr>
<tr>
<td>Pope JH (2000)</td>
<td>Acute cardiac ischemia in ED</td>
<td>Mistakenly discharged from ED; MI - 894/889 (2.1%);</td>
</tr>
<tr>
<td>Edelman D (2002)</td>
<td>Diabetes in an outpatient clinic</td>
<td>Blood glucoses meeting criteria for DM with no diagnosis of diabetes in records - 258/1426 (18%)</td>
</tr>
<tr>
<td>Goodson WH (2002)</td>
<td>Breast Cancer</td>
<td>Inappropriately reassured to have benign lesions - 21/435 (5%); 14 (3%) misread mammogram, 4 (1%) misread pathologic finding, 5 (1%) missed by poor fine-needle biopsy</td>
</tr>
<tr>
<td>Kawalski RG (2004)</td>
<td>Aneurysmal subarachnoid hemorrhage (SAH)</td>
<td>Patients w/ SAH initially misdiagnosed - 56/482 (12%); migraine/tension headache most common incorrect dx – 20/56 (36%); failure to obtain a CT was most common error - 41/56 (73%)</td>
</tr>
</tbody>
</table>

Traditional and innovative approaches to learning from diagnosis error

Traditionally, studying missed diagnoses or incorrect diagnoses had a central role in medical education, research, and quality assurance in the form of autopsies. Other traditional methods of learning about misdiagnosed cases include malpractice litigation, morbidity and mortality (M&M) conferences, unsystematic feedback from patients, other providers, or simply from patients’ illnesses as they evolved over time. Beyond the negative aspects of being grounded in patients’ deaths or malpractice accusations, there are other limitations of these historical approaches, including
• Lack of systematic approaches to surveillance, reporting, and learning from errors, with nonrandom sample of cases subjected to such review\textsuperscript{37, 39}
• Lack of timeliness, with cases often reviewed months or years after the event\textsuperscript{38}
• Examinations that rarely dig to the root of problems: not focused on the “Five Whys”\textsuperscript{40}
• Postmortems that seldom go beyond the case-at-hand, with minimal linkages to formal quality improvement activities\textsuperscript{41}
• Atrophy of the value of even these suboptimal approaches, with autopsy rates in the single digits (in many hospitals, zero), many malpractice experiences sealed by nondisclosure agreements, and shorter hospitalizations limiting opportunities for followup to ultimate diagnosis\textsuperscript{34, 41, 42}

What is needed to overcome these limitations is not only a more systematic method for examining cases of diagnosis failure, but also a fresh approach. Therefore, our team approached diagnosis error with the following perspectives:

**Diagnosis as part of a system.** Diagnostic accuracy should be viewed as a system property rather than simply what happens between the doctor’s two ears.\textsuperscript{2, 43–45} While cognitive issues figure heavily in the diagnostic process, a quite from Don Berwick summarizes\textsuperscript{46} a much lacking and needed perspective: “Genius diagnosticians make great stories, but they don’t make great health care. The idea is to make accuracy reliable, not heroic.”

**Less reliance on human memory.** Relying on clinicians’ memory—to trigger consideration of a particular diagnosis, recall a disease’s signs/symptoms/pattern from a textbook or experience—or simply to remember to check on a patient’s lab result—is an invitation to variations and failures. This lesson from other error research resonates powerfully with clinicians, who are losing the battle to keep up to date.\textsuperscript{9, 45, 47}

**Need for “space” to allow open reflection and discussion.** Transforming an adversarial atmosphere into one conducive to honest reflection is an essential first step.\textsuperscript{48, 49} However, an equally important and difficult challenge is creating venues that allow clinicians (and patients) to discuss concerns in an efficient and productive manner.\textsuperscript{37} Cases need to be reviewed in sufficient detail to make them “real.” Firsthand clinical information often radically changes our understanding from what the more superficial “first story” suggested. As complex clinical circumstances are better understood, new light is often shed on what at first appeared to be indefensible diagnostic decisions and actions. Unsuspected additional errors also emerge. Equally important is not to get mired in details or making judgments (whether to label a case as a diagnosis error). Instead, it is more valuable to focus on generalizable lessons of how to ensure better treatment of similar future patients.\textsuperscript{16}
Adopting multidisciplinary perspectives and collaboration. A broad range of skills and vantage points are valuable in understanding the complex diagnostic problems that we encountered. We considered input from specialists and primary care physicians to be essential. In addition, specialists in emergency medicine (where many patients first present) offered a vital perspective, both for their diagnostic expertise and their pivotal interface with system constraints (resource limits mean that not every patient with a confusing diagnosis can be hospitalized). Even more valuable has been the role of non-MDs, including nursing quality specialists, information scientists, and social scientists (cognitive psychologist, decision theory specialist) in forging a team to broadly examine diagnosis errors.

Innovative screening approaches. Developing new ways to uncover errors is a priority. We cannot afford to wait for a death, lawsuit, or manual review. Approaches we have been exploring include electronic screening that links pharmacy and lab data (e.g., to screen for abnormal results, such as elevated thyroid stimulating hormone [TSH], unaddressed by thyroxin therapy), trajectory studies (retrospectively probing delays in a series of cases with a particular diagnosis), and screening for discrepancies between admitting and discharge diagnoses. A related approach is to survey specialists (who are poised to see diagnoses missed in referred patients), primary care physicians (about their own missed diagnoses), or patients themselves (who frequently have stories to share about incorrect diagnoses), in addition to various ad hoc queries and self-reports.

Where does the diagnostic process fail?

One of the most powerful heuristics in medication safety has been delineation of the steps in the medication-use process (prescribing, transcribing, dispensing, administering, and monitoring) to help localize where an error has occurred. Diagnosis, while more difficult to neatly classify (because compared to medications, stages are more concurrent, recurrent, and complex), nonetheless can be divided into seven stages: (1) access/presentation, (2) history taking/collection, (3) the physical exam, (4) testing, (5) assessment, (6) referral, and (7) followup. We have found this framework helpful for organizing discussions, aggregating cases, and targeting areas for improvement and research. It identifies what went wrong, and situates where in the diagnostic process the failure occurred (Table 3). We have used it for a preliminary analysis of several hundred diagnosis error cases we collected by surveying physicians.

This taxonomy for categorizing diagnostic “assessment” draws on work of Kassirer and others, highlighting the two key steps of (a) hypothesis generation, and (b) differential diagnosis or hypothesis weighing/prioritization. We add another aspect of diagnostic assessment, one that connects to other medical and iatrogenic error work—the need to recognize the urgency of diagnoses and complications. This addition underscores the fact that failure to make the exact diagnosis is often less important than correctly assessing the urgency of the patient’s illness. We divide the “testing” stage into three components—ordering, performing, and clinician processing (similar but not identical to the laboratory literature classification of the phases of lab testing as preanalytic, analytic, and
Table 3. Taxonomy of where and what errors occurred

<table>
<thead>
<tr>
<th>Where in Diagnostic Process (~Anatomic localization)</th>
<th>What Went Wrong (~Lesion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Access/presentation</td>
<td>Denied care</td>
</tr>
<tr>
<td></td>
<td>Delayed presentation</td>
</tr>
<tr>
<td>2. History</td>
<td>Failure/delay in eliciting critical piece of history data</td>
</tr>
<tr>
<td></td>
<td>Inaccurate/misinterpretation “</td>
</tr>
<tr>
<td></td>
<td>Suboptimal weighing “</td>
</tr>
<tr>
<td></td>
<td>Failure/delay to followup “</td>
</tr>
<tr>
<td>3. Physical exam</td>
<td>Failure/delay in eliciting critical physical exam finding</td>
</tr>
<tr>
<td></td>
<td>Inaccurate/misinterpreted “</td>
</tr>
<tr>
<td></td>
<td>Suboptimal weighing “</td>
</tr>
<tr>
<td></td>
<td>Failure/delay to followup “</td>
</tr>
<tr>
<td>4. Tests (lab/radiology)</td>
<td>Failure/delay in ordering needed test(s)</td>
</tr>
<tr>
<td>Ordering</td>
<td>Performance</td>
</tr>
<tr>
<td></td>
<td>Sample mix-up/mislabeled (e.g., wrong patient)</td>
</tr>
<tr>
<td></td>
<td>Technical errors/poor processing of specimen/test</td>
</tr>
<tr>
<td></td>
<td>Erroneous lab/radiol reading of test</td>
</tr>
<tr>
<td></td>
<td>Failed/delayed transmission of result to clinician</td>
</tr>
<tr>
<td>Clinician processing</td>
<td>Failed/delayed followup action on test result</td>
</tr>
<tr>
<td></td>
<td>Erroneous clinician interpretation of test</td>
</tr>
<tr>
<td>5. Assessment</td>
<td>Failure/delay in considering the correct diagnosis</td>
</tr>
<tr>
<td>Hypothesis generation</td>
<td>Suboptimal weighing/prioritizing</td>
</tr>
<tr>
<td></td>
<td>Too much weight to low(er) probability/priority dx</td>
</tr>
<tr>
<td></td>
<td>Too little consideration of high(er) probability/priority dx</td>
</tr>
<tr>
<td></td>
<td>Too much weight on competing diagnosis</td>
</tr>
<tr>
<td>Recognizing urgency/complications</td>
<td>Failure to appreciate urgency/acuity of illness</td>
</tr>
<tr>
<td></td>
<td>Failure/delay in recognizing complication(s)</td>
</tr>
<tr>
<td>6. Referral/consultation</td>
<td>Failure/delay in ordering needed referral</td>
</tr>
<tr>
<td></td>
<td>Inappropriate/unneeded referral</td>
</tr>
<tr>
<td></td>
<td>Suboptimal consultation diagnostic performance</td>
</tr>
<tr>
<td></td>
<td>Failed/delayed communication/followup of consultation</td>
</tr>
<tr>
<td>7. Followup</td>
<td>Failure to refer to setting for close monitoring</td>
</tr>
<tr>
<td></td>
<td>Failure/delay in timely followup/rechecking of patient</td>
</tr>
</tbody>
</table>
For each broad category, we specified the types of problems we observed.

A recurring theme running through our reviews of potential diagnosis error cases pertains to the relationship between errors in the diagnostic process, delay and misdiagnosis, and adverse patient outcomes. Bates has promulgated a useful model for depicting the relationships between medication errors and outcomes. Similarly, we find that most errors in the diagnostic process do not adversely impact patient outcomes. And, many adverse outcomes associated with misdiagnosis or delay do not necessarily result from any error in the diagnostic process—the cancer may simply be undiagnosable at that stage, the illness presentation too atypical, rare, or unlikely even for the best of clinicians to diagnose early. These situations are often referred to as “no fault” or “forgivable” errors—terms best avoided because they imply fault or blame for preventable errors (Figure 1).

While deceptively simple, the model raises a series of extremely challenging questions—questions we found ourselves repeatedly returning to in our weekly discussions. We hope these questions can provide insights into recurring themes and challenges we faced, and perhaps even serve as a checklist for others to structure their own patient care reviews. While humbled by our own inability to provide more conclusive answers to these questions, we believe researchers and practitioners will be forced to grapple with them before we can make significant progress.

Questions for consideration by diagnosis error evaluation and research (DEER) investigators in assessing cases

Uncertainties about diagnosis and findings

1. What is the correct diagnosis? How much certainty do we have, even now, about what the correct diagnosis is?
2. What were the findings at the various points in time when the patient was being seen; how much certainty do we have that a particular finding and diagnosis was actually present at the time(s) we are positing an error?

Relationship between diagnosis failure and adverse outcomes

3. What is the probability that the error resulted in the adverse outcome? How treatable is the condition, and how critical is timely diagnosis and treatment for impacting on the outcome—both in general and in this case?
4. How did the error in the diagnostic process contribute to making the wrong diagnosis and wrong treatment?
Figure 1. Relationships between diagnostic process errors, misdiagnosis, and adverse events

*Delayed, missed, or misdiagnosis

Caption

Group A = Errors in diagnostic process (blood sample switched between two patients, MD doesn't do a physical exam for patient with abdominal pain)

Group B = Diagnostic process error with resulting misdiagnosis (patient given wrong diagnosis because blood samples switched)

Group C = Adverse outcome resulting from error-related misdiagnosis (Patient is given toxic treatment and has adverse effect as result of switched samples. Fail to diagnose appendicitis because of failure to examine abdomen, and it ruptures and patient dies)

Group D = Harm from error in diagnostic process (colon perforation from colonoscopy done on wrong patient)

Group E = Misdiagnosis, delayed diagnosis or missed diagnosis, but no error in care or harm (incidental prostate cancer found on autopsy)

Group F = Adverse event due to misdiagnosis but no identifiable process error (death from acute MI but no chest pain or other symptoms that were missed)

Group G = Adverse events but not related to misdiagnosis, delay, or error in diagnostic process, e.g., death from correctly diagnosed disease complication, or nonpreventable drug reaction (PCN anaphylaxis in patient never previously exposed)
Clinician assessment and actions

5. What was the physician’s diagnostic assessment? How much consideration was given to the correct diagnosis? (This is usually difficult to reconstruct because differential diagnosis often is not well documented.)

6. How good or bad was the diagnostic assessment based on evidence clinicians had on hand at that time (should have been obvious from available data vs. no way anyone could have suspected)?

7. How erroneous was the diagnostic assessment, based on the difficulty in making the diagnosis at this point? (Was there a difficult “signal-to-noise” situation, a rare low-probability diagnosis, or an atypical presentation?)

8. How justifiable was the failure to obtain additional information (i.e., history, tests) at a particular point in time? How can this be analyzed absolutely, as well as relative to the difficulties and constraints in obtaining this missing data? (Did the patient withhold or refuse to give accurate/additional history; were there backlogs and delays that made it impossible to obtain the desired test?)

9. Was there a problem in diagnostic assessment of the severity of the illness, with resulting failure to observe or follow up the patient more closely? (Again, both absolutely and relative to constraints.)

Global assessment of improvement opportunities

10. To what extent did the clinicians’ actions deviate from the standard-of-care (i.e., was there negligent care with failure to follow accepted diagnostic guidelines and expected practices, or to pursue abnormal finding that should never be ignored)?

11. How preventable was the error? How ameliorable or amenable to change are the factors/problems that contributed to the error? How much would such changes, designed to prevent this error in the future, cost?

12. What should we do better the next time we encounter a similar patient or situation? Is there a general rule, or are there measures that can be implemented to ensure this is reliably done each time?

Diagnosis error case vignettes

Case 1

A 25-year-old woman presents with crampy abdominal pain, vaginal bleeding, and amenorrhea for 6 weeks. Her serum human choriogonadotropin (HCG) level is markedly elevated. A pelvic ultrasound is read by the on-call radiology chief resident and obstetrics (OB) attending physician as showing an empty uterus,
suggesting ectopic pregnancy. The patient is informed of the findings and treated with methotrexate. The following morning the radiology attending reviews the ultrasound and amends the report, officially reading it as “normal intrauterine pregnancy.”

**Case 2**

A 49-year-old, previously healthy man presents to the emergency department (ED) with nonproductive cough, “chest congestion,” and dyspnea lasting 2 weeks; he has a history of smoking. The patient is afebrile, with pulse = 105, respiration rate (RR) = 22, and white blood count (WBC) = 6.4. Chest x-ray shows “marked cardiomegaly, diffuse interstitial and reticulonodular densities with blunting of the right costophrenic angle; impression—congestive heart failure (CHF)/pneumonia. Rule out (R/O) cardiomyopathy, valve disease or pericardial effusion.” The patient is sent home with the diagnosis of pneumonia, with an oral antibiotic.

The patient returns 1 week later with worsening symptoms. He is found to have pulsus paradoxicus, and an emergency echocardiogram shows massive pericardial effusion. Pericardiocentesis obtains 350 cc fluid with cytology positive for adenocarcinoma. Computed tomography of the chest suggests “lymphangitic carcinomatosis.”

**Case 3**

A 50-year-old woman with frequent ED visits for asthma (four visits in the preceding month) presents to the ED with a chief complaint of dyspnea and new back pain. She is treated for asthma exacerbation and discharged with nonsteroidal anti-inflammatory drugs (NSAID) for back pain.

She returns 2 days later with acutely worsening back pain, which started when reaching for something in her cupboard. A chest x-ray shows a “tortuous and slightly ectatic aorta,” and the radiologist’s impression concludes, “If aortic dissection is suspected, further evaluation with chest CT with intravenous (IV) contrast is recommended.” The ED resident proceeds to order a chest CT, which concludes “no evidence of aneurysm or dissection.” The patient is discharged.

She returns to the ED 3 days later, again complaining of worsening asthma and back pain. While waiting to be seen, she collapses in the waiting room and is unable to be resuscitated. Autopsy shows a ruptured aneurysm of the ascending aorta.

**Case 4**

A 50-year-old woman with a past history diabetes and alcohol and IV drug abuse, presents with symptoms of abdominal pain and vomiting and is diagnosed as having “acute chronic pancreatitis.” Her amylase and lipase levels are normal. She is admitted and treated with IV fluids and analgesics. On hospital day 2 she begins having spiking fevers and antibiotics are administered. The next day, blood cultures are growing gram negative organisms.
At this point, the service is clueless about the patient’s correct diagnosis. It only becomes evident the following day when (a) review of laboratory data over the past year shows that patient had four prior blood cultures, each positive with different gram negative organisms; (b) a nurse reports patient was “behaving suspiciously,” rummaging through the supply room where syringes were kept; and (c) a medical student looks up posthospital outpatient records from 4 months earlier and finds several notes stating that “the patient has probable Munchausen syndrome rather than pancreatitis.” Upon discovering these findings, the patient’s IVs are discontinued and sensitive, appropriate followup primary and psychiatric care are arranged.

A postscript to this admission: 3 months later, the patient was again readmitted to the same hospital for “pancreatitis” and an unusual “massive leg abscess.” The physicians caring for her were unaware of her past diagnoses and never suspected or discovered the likely etiology of her abscess (self-induced from unsterile injections).

Lessons and issues raised by the diagnosis error cases

Difficulties in sorting out “don’t miss” diagnoses

Before starting our project, we compiled a list of “don’t miss” diagnoses (available from the authors). These are diagnoses that are considered critical, but often difficult to make—critical because timely diagnosis and treatment can have major impact (for the patient or the public’s health, or both), and difficult because they either are rare or pose diagnostic challenges. Diagnoses such as spinal epidural abscess (where paraplegia can result from delayed diagnosis), or active pulmonary tuberculosis (TB) (where preventing spread of infection is critical) are examples of “don’t miss” diagnoses. While there is a scant evidence base to definitively compile and prioritize such a list, three of our cases—ectopic pregnancy, dissecting aortic aneurysm, and pericardial effusion with tamponade—are diagnoses that would unquestionably be considered as life-threatening diagnoses that ought not be delayed or missed.25

Although numerous issues concerning diagnosis error are raised by these cases, they also illustrate problems relating to uncertainties, lack of gold standards (for both testing and standard of care), and difficulties reaching consensus about best ways to prevent future errors and harmful delays. Below we briefly discuss some of these issues and controversies.

Diagnostic criteria and strategies for diagnosing ectopic pregnancy are controversial, and our patient’s findings were particularly confusing. Even after careful review of all aspects of the case, we were still not certain who was “right”—the physicians who read the initial images and interpreted them as consistent with ectopic pregnancy, or the attending physician rereading the films the next day as normal. The literature is unclear about criteria for establishing this
diagnosis.\textsuperscript{54–57} In addition, there is a lack of standards in performance and interpretation of ultrasound exams, plus controversies about timing of interventions. Thus this “obvious” error is obviously more complex, highlighting a problem-prone clinical situation.

The patient who was found to have a malignant pericardial effusion illustrates problems in determining the appropriate course of action for patients with unexplained cardiomegaly, which the ED physicians failed to address on his first presentation: Was he hemodynamically stable at this time? If so, did he require an urgent echocardiogram? What criteria should have mandated this be done immediately? How did his cardiomegaly get “lost” as his diagnosis prematurely “closed” on pneumonia when endorsed from the day to the night ED shift? How should one assess the empiric treatment for pneumonia given his abnormal chest x-ray? Was this patient with metastatic lung cancer harmed by the 1 week delay?

The diagnosis of aneurysms (e.g., aortic, intracranial) arises repeatedly in discussions of misdiagnosis. Every physician seems to recall a case of a missed aneurysm with catastrophic outcomes where, in retrospect, warnings may have been overlooked. A Wall Street Journal article recently won a Pulitzer Prize for publicizing such aneurysm cases.\textsuperscript{58} Our patient’s back pain was initially dismissed. Because of frequent visits, she had been labeled as a “frequent flyer”—and back pain is an extremely common and nonspecific symptom. A review of literature on the frequency of dissecting aortic aneurysm reveals that it is surprisingly rare, perhaps less than 1 out of 50,000 ED visits for chest pain, and likely an equally rare cause for back pain.\textsuperscript{59–62} She did ultimately undergo a recommended imaging study after a suspicious plain chest x-ray, however it was deemed “negative.”

Thus, in each case, seemingly egregious and unequivocal errors were found to be more complex and uncertain.

**Issues related to limitations of diagnostic testing**

Even during our 3 years of diagnosis case reviews, clinicians have been confronted with rapid changes in diagnostic testing. New imaging modalities, lab tests, and testing recommendations have been introduced, often leaving clinicians confused about which tests to order and how to interpret their confusing and, at times, contradictory (from one radiologist to the next) results.\textsuperscript{63}

If diagnosis errors are to be avoided, clinicians must be aware of the limitations of the diagnostic tests they are using. It is well known that a normal mammogram in a woman with a breast lump does not rule out the diagnosis of breast cancer, because the sensitivity of test is only 70 to 85 percent.\textsuperscript{13, 26, 64, 65} A recurring theme in our cases is failure to appreciate pitfalls in weighing test results in the context of the patient’s pretest disease probabilities. Local factors, such as the variation in quality of test performance and readings, combined with communication failures between radiology/laboratory and ordering physicians (either no direct communication or interactions where complex interpretations get
reduced to “positive” or “negative,” overlooking subtleties and limitations) provide further sources of error.66

The woman with the suspected ectopic pregnancy, whose emergency ultrasound was initially interpreted as being “positive,” illustrates the pitfalls of taking irreversible therapeutic actions without carefully weighing test reading limitations. Perhaps an impending rupture of an ectopic pregnancy warrants urgent action. However, it is also imperative for decisions of this sort that institutions have fail-safe protocols that anticipate such emergencies and associated test limitations.

The patient with a dissecting aortic aneurysm clearly had a missed diagnosis, as confirmed by autopsy. This diagnosis was suspected premortem, but considered to be “ruled out” by a CT scan that did not show a dissection. Studies of the role of chest CT, particularly when earlier CT scanning technology was used, show sensitivity for dissecting aneurysm of only 83 percent.67 When we reexamined the old films for this patient, several radiologists questioned the adequacy of the study (quality of the infusion plus question of motion artifact). Newer, faster scanners reportedly are less prone to these errors, but the experience is variable. We identified another patient where a “known” artifact on a spiral CT nearly led to an unnecessary aneurysm surgery; it was only prevented by the fact that she was a Jehovah’s Witness and, because her religious beliefs precluded transfusions, surgery was considered too risky.62

The role of information transfer and the communication of critical laboratory information

Failure of diagnosis because of missing information is another theme in our weekly case reviews and the medical literature. Critical information can be missed because of failures in history-taking, lack of access to medical records, failures in the transmission of diagnostic test results, or faulty records organization (either paper or electronic) that created problems for quickly reviewing or finding needed information.

For the patient with the self-induced illness, all of the “missing” information was available online. Ironically, although many patients with a diagnosis of Munchausen’s often go to great lengths to conceal information (i.e., giving false names, using multiple hospitals), in our case, there was so much data in the computer from previous admissions and outpatient visits that the condition was “lost” in a sea of information overload—a problem certain to grow as more and more clinical information is stored online. While this patient is an unusual example of the general problems related to information transfer, this case illustrates important principles related to the need for conscientious review, the synthesizing of information, and continuity (both of physicians and information) to avoid errors.

Simply creating and maintaining a patient problem list can help prevent diagnosis errors. It can ensure that each active problem is being addressed, helping all caregivers to be aware of diagnoses, allergies, and unexplained
findings. Had our patient with “unexplained cardiomegaly” been discharged listing this as one of his problems, instead of only “pneumonia,” perhaps this problem would not have been overlooked. However, making this seemingly simple documentation tool operational has been unsuccessful in most institutions, even ones with advanced electronic information systems, and thus represents a challenge as much as a panacea.

One area of information transfer, the followup of abnormal laboratory test results, represents an important example of this information transfer paradigm in diagnostic patient safety.68–73 We identified failure rates of more than 1 in 50 for both followup abnormal thyroid tests (where the diagnosis of hypothyroidism was missed in 23 out of 982, or 2.3 percent, of patients with markedly elevated TSH results), and our earlier study on failure to act on elevated potassium levels (674 out of 32,563, or 2.0 percent, of potassium prescriptions were written for hyperkalemic patients).74 Issues of communication, teamwork, systems design, and information technology stand as areas for improvement.75–77 Recognizing this, the Massachusetts Safety Coalition has launched a statewide initiative on Communicating Critical Test Results.78

Physician time, test availability, and other system constraints

Our project was based in two busy urban hospitals, including a public hospital with serious constraints on bed availability and access to certain diagnostic tests. An important recurring theme in our case discussions (and in health care generally) is the interaction between diagnostic imperatives and these resource limitations.

To what extent is failure to obtain an echocardiogram, or even a more thorough history or physical exam, understandable and justified by the circumstances under which physicians find themselves practicing? Certainly our patient with the massive cardiomegaly needed an echocardiogram at some time. Was it reasonable for the ED to defer the test (meaning a wait of perhaps several months in the clinic), or would a more “just-in-time” approach be more efficient, as well as safer in minimizing diagnosis error and delay?79 Since, by definition, we expect ED physicians to triage and treat emergencies, not thoroughly work up every problem patients have, we find complex trade-offs operating at multiple levels.

Similar trade-offs impact whether a physician had time to review all of the past records of our factitious-illness patient (only the medical student did), or how much radiology expertise is available around-the-clock to read ultrasound or CT exams, to diagnose ectopic pregnancy or aortic dissection. This is perhaps the most profound and poorly explored aspect of diagnosis error and delay, but one that will increasingly be front-and-center in health care.
Cognitive issues in diagnosis error

We briefly conclude where most diagnosis error discussions begin, with cognitive errors.\textsuperscript{45, 53, 80–84}

Hindsight bias and the difficulty of weighing prior probabilities of the possible diagnoses bedeviled our efforts to assess decisions and actions retrospectively. Many “don’t miss” diagnoses are rare; it would be an error to pursue each one for every patient. We struggled to delineate guidelines that would accurately identify high-risk patients and to design strategies to prevent missing these diagnoses.

Our case reviews and firsthand interviews often found that each physician had his or her own individual way of approaching patients and their problems. Such differences made for lively conference discussions, but have disturbing implications for developing more standardized approaches to diagnosis.

The putative dichotomy between “cognitive” and “process” errors is in many ways an artificial distinction.\textsuperscript{7, 8} If a physician is interrupted while talking to the patient or thinking about a diagnosis and forgets to ask a critical question or consider a critical diagnosis, is this a process or cognitive error?

Conclusion

Because of their complexity, there are no quick fixes for diagnosis errors. As we review what we learned from a variety of approaches and cases, certain areas stood out as ripe for improvement—both small-scale improvements that can be tested locally, as well as larger improvements that need more rigorous, formal research. Table 4 summarizes these change ideas, which harvest the lessons of our 3-year project.

As outlined in the table, there needs to be a commitment to build learning organizations, in which feedback to earlier providers who may have failed to make a correct diagnosis becomes routine, so that institutions can learn from this aggregated feedback data. To better protect patients, we will need to conceptualize and construct safety nets to mitigate harm from uncertainties and errors in diagnosis. The followup of abnormal test results is a prime candidate for reengineering, to ensure low “defect” rates that are comparable to those achieved in other fields. More standardized and reliable protocols for reading x-rays and laboratory tests (such as pathology specimens), particularly in residency training programs and “after hours,” could minimize the errors we observed. In addition, we need to better delineate “red flag” and “don’t miss” diagnoses and situations, based on better understanding and data regarding pitfalls in diagnosis and ways to avoid them.
Table 4. Change ideas for preventing and minimizing diagnostic error

<table>
<thead>
<tr>
<th>Change Idea</th>
<th>Rationale/Description</th>
<th>Challenges</th>
</tr>
</thead>
</table>
| **Upstream feedback** to earlier providers who have may have failed to make correct diagnosis | • Promotes culture of safety, accountability, continuous and blame-free learning, and communication  
• “Hard-wires” organizational and practitioner learning from diagnosis evolution and delays  
• Feedback from specialists poised to see missed diagnosis could be especially useful  
• Build in “feedback from the feedback” to capture reflective practitioner assessment of why errors may have occurred and considerations for future prevention  
• Permits aggregation for tracking, uncovering patterns, learning across cases, elucidating pitfalls, measuring improvements | • Logistical requirements for implementation and surveillance screening to identify errors  
• Avoiding “tampering,” from availability bias that neglects base rates (ordering aortogram on every chest pain patient to “rule out” dissection)  
• Protecting confidentiality, legal liabilities, blame-free atmosphere  
• Ways to extend to previous hospitals and physicians outside of own institution  
• To be highest leverage needs to be coupled with reporting, case conferences |
| **Safety nets to mitigate harm from diagnostic uncertainty and error** | • Well designed observation and followup venues and systems (e.g., admit for observation, followup calls or e-mail from MD in 48 hours, automated 2 wk phone followup to ensure tests obtained) for high-risk, uncertain diagnoses  
• Educating and empowering patients to have lower threshold for seeking followup care or advice, including better defining and specifying particular warning symptoms | • Logistics  
• Resource constraints (bed, test availability, clinician time)  
• Avoiding false positive errors, inappropriate use of scarce/costly resources  
• Not creating excessive patient worry/anxiety |
<table>
<thead>
<tr>
<th>Change Idea</th>
<th>Rationale/Description</th>
<th>Challenges</th>
</tr>
</thead>
</table>
| **Timely and reliable abnormal test result followup systems** | • Fail-safe, prospectively designed systems to identify which results are critical abnormals (panic and routine), who to communicate result to, how to deliver  
• Uniform approach across various test types and disciplines (lab, pathology, radiology, cardiology)  
• Leverage information technologies to automate and increase reliability  
• Involve patients to ensure timely notification of results, and contacting provider when this fails  
• Modeled on Massachusetts Patient Safety Coalition program 78 | • Achieving consensus on test types and cut-off thresholds  
• On-call, cross-coverage issues for critical panic results  
• Defining responsibilities: e.g., for ED patients, for lab personnel  
• Practitioner efficiency issues: avoiding excess duplicate work; efficient documentation |
| **Fail-safe protocols for preliminary/resident and definitive readings of tests** | • Must be well-defined and organized system for supervision, creating and communicating final reports  
• Need for system for amending and alerting critical changes to clinicians  
• “After-hours” systems for readings, amending | • How to best recognize and convey variations in expertise of attendings who write final reports  
• Quality control poses major unmet challenges |
| **Prospectively defining red flag diagnoses and situations and instituting prospective readiness** | • Create “pull” systems for patients with particular medical problems to ensure standardized, expedited diagnostic evaluations (so don't have to “push” to quickly obtain)  
• Like AMI thrombolytic “clot box,” in-place for ready activation the moment patient first presents  
• Embody/require coordinated multidisciplinary approach (e.g., pharmacy, radiology, specialists) | • Difficulties in evidence-based delineation of diagnoses, situations, patient selection, criteria, and standardized actions  
• Avoiding excessive work-up, diverting resources from other problems |
Table 4. Change ideas for preventing and minimizing diagnostic error, cont.

<table>
<thead>
<tr>
<th>Change Idea</th>
<th>Rationale/Description</th>
<th>Challenges</th>
</tr>
</thead>
</table>
| Automated screening **check lists** to avoid missing key history, physical, lab data | • Less reliance on human memory for more thorough questioning  
• Queries triggered by individual patient features  
• Could be customized based on presenting problem (i.e., work exposures for lung symptoms, travel history for fever) | • Evidence of efficacy to-date unconvincing; unclear value of unselective “review of systems”  
• Sorting out, avoiding false positive errors from data with poor signal:noise ratio  
• Effectively implementing with teamwork and integrated information technology |
| High-level **patient education and engagement** with diagnosis probabilities and uncertainties | • Since diagnosis often so complex and difficult, this needs to be shared with patients in ways to minimize disappointments and surprises  
• Support and enhance patients taking initiative to question diagnosis, particularly if not responding as expected | • Potentially more time-consuming for practitioners  
• Avoiding unnecessary testing  
• Doing in way that balances need for preserving patient confidence in their physicians and advice, with education and recognition of diagnosis fallabilities |
| Test and leverage **information technology tools** to avoid known cognitive and care process pitfalls | • Better design of ways to streamline documentation (including differential diagnosis) and access/display of historical data  
• Easing documentation time demands to give practitioners more time to talk to patients and think about their problems  
• Facilitating real-time access to medical knowledge sources  
• Sophisticated decision support tools that use complex rules and individualized patients data  
• Prompts to suggest consideration of medication effects in differential, based on linkages to patient’s medication profile, lab results | • Shortcomings of first generation of “artificial intelligence” diagnosis software  
• Challenges coupling knowledge bases with individual patient characteristics  
• Paucity of standardized, accepted, sharable clinical alerts/rules  
• New errors and distractions introduced by intrusion of computer into clinical encounter  
• Alleged atrophy of unaided cognitive skills |
To achieve many of these advances, automated (and manual) checklists and reminders will be needed to overcome current reliance on human memory. But information technology must also be deployed and reengineered to overcome growing problems associated with information overload. Finally, and most importantly, patients will have to be engaged on multiple levels to become “coproducers” in a safer practice of medical diagnosis. It is our hope that these change ideas can be tested and implemented to ensure safer treatment based on better diagnoses—diagnosis with fewer delays, mistakes, and process errors.

Acknowledgments

This work was supported by AHRQ Patient Safety Grant #11552, the Cook County–Rush Developmental Center for Research in Patient Safety (DCERPS) Diagnostic Error Evaluation and Research (DEER) Project.

Author affiliations

Cook County John H. Stroger Hospital and Bureau of Health Services, Chicago (GDS, KC, MFW). Rush University Medical Center, Chicago (RA, SH, RO, RAM). Hektoen Research Institute, Chicago (SK, NK). University of Illinois at Chicago, College of Pharmacy (BL). University of Illinois at Chicago Medical School (ASE).

Address correspondence to: Gordon D. Schiff; Cook County John H. Stroger Hospital and Bureau of Health Services, 1900 W. Polk Street, Administration Building #901, Chicago, IL 60612. Phone: 312-864-4949; fax: 312-864-9594; e-mail: gdschiff@aol.com.

References


WHAT ARE THE PRIMARY & SECONDARY OUTCOMES OF DIAGNOSTIC PROCESS - THE LEADING PHASE OF WORK IN PATIENT CARE MANAGEMENT PROBLEMS IN EMERGENCY DEPARTMENT?

Characteristics of Patient Care Management Problems Identified in Emergency Department Morbidity and Mortality Investigations During 15 Years

Karen S. Cosby, MD  
Rebecca Roberts, MD  
Lisa Palivos, MD  
Christopher Ross, MD  
Jeffrey Schaider, MD  
Scott Sherman, MD  
Isam Nasr, MD  
Eileen Couture, DO  
Moses Lee, MD  
Shari Schabowski, MD  
Ibrar Ahmad, BS  
R. Douglas Scott II, PhD

From the Department of Emergency Medicine, Cook County Hospital, Rush Medical School, Chicago, IL.

Study objective: We describe cases referred for physician review because of concern about quality of patient care and identify factors that contributed to patient care management problems.

Methods: We performed a retrospective review of 636 cases investigated by an emergency department physician review committee at an urban public teaching hospital over a 15-year period. At referral, cases were initially investigated and analyzed, and specific patient care management problems were noted. Two independent physicians subsequently classified problems into 1 or more of 4 major categories according to the phase of work in which each occurred (diagnosis, treatment, disposition, and public health) and identified contributing factors that likely affected outcome (patient factors, triage, clinical tasks, teamwork, and system). Primary outcome measures were death and disability. Secondary outcome measures included specific life-threatening events and adverse events. Patient outcomes were compared with the expected outcome with ideal care and the likely outcome of no care.

Results: Physician reviewers identified multiple problems and contributing factors in the majority of cases (92%). The diagnostic process was the leading phase of work in which problems were observed (71%). Three leading contributing factors were identified: clinical tasks (99%), patient factors (61%), and teamwork (61%). Despite imperfections in care, half of all patients received some benefit from their medical care compared with the likely outcome with no care.

Conclusion: These reviews suggest that physicians would be especially interested in strategies to improve the diagnostic process and clinical tasks, address patient factors, and develop more effective medical teams. Our investigation allowed us to demonstrate the practical application of a framework for case analysis. We discuss the limitations of retrospective cases analyses and recommend future directions in safety research. [Ann Emerg Med. 2008;51:251-261.]

0196-0644/$-see front matter  
Copyright © 2008 by the American College of Emergency Physicians.  
doi:10.1016/j.annemergmed.2007.06.483
Editor’s Capsule Summary

What is already known on this topic
Morbidity and mortality conferences and other forms of quality review are commonplace. Their content has not been well studied.

What question this study addressed
What types of events and what causal and contributing factors are common in morbidity and mortality reviews?

What this study adds to our knowledge
The diagnostic process was judged the most common locus of failure in more than 600 cases, spanning 15 years. Despite imperfections in care, more than half the patients still received some benefit compared with the likely outcome with no care at all.

How this might change clinical practice
Detailed case reviews provide useful information for practice improvement but have selection bias in the choice of cases and biases produced by the retrospective interpretation of incomplete information.

SEE EDITORIAL, P. 262.

INTRODUCTION
Background
Several major studies have reported the incidence of medical error and risks associated with health care.1-6 These studies have raised awareness of imperfections in care but offer little understanding of the nature of the work or solutions to improve safety. Although the focus on patient safety is relatively new, processes for improvement have existed for decades. We suggest that cumulative reviews of existing data, available in most health care organizations, can be used to guide current efforts to improve safety.

Our study was initially aimed at detecting “medical errors.” Although the concept of error has been widely disseminated in the medical literature during the last decade, the term “error” carries with it connotations of carelessness, implies blame, and often focuses unduly on humans as the source of harm. As our understanding of these problems grew, we modified our study to describe events once labeled errors as “patient care management problems.” This subtle change is intended to promote a healthier perspective and constructive analysis. The term “problem” implies something to be solved and states more directly our intent to seek strategies and solutions for improvement.

Importance
Demand for improvement has led to changes throughout health care, including new safety regulations for health care organizations, proposed legislation for error reporting and malpractice reform, and advancements in information technology applications for medical settings.7-10 Reforms in medical education have led to restrictions in house staff working hours, development of curriculums for patient safety, increased rigor in requirements for lifelong learning for physicians, and new emphasis on competency standards.11,12 Although we are anxious to improve, recommended changes should be founded on a thorough understanding of the types of patient care management problems that occur in medicine and factors that may increase risk.

Goals of This Investigation
The objective of this study was to characterize the types of cases referred to a physician review committee of an urban emergency department (ED) and identify the phase of work in which problems were detected and specific factors that affected quality of patient care. Our long-term goal is to use our existing morbidity and mortality investigations to guide safety interventions and to develop valid methods for future study of patient safety targets. The ED has been described as a high-risk practice environment, particularly under conditions of crowding, and serves as a natural setting to study safety in health care.13,14

MATERIALS AND METHODS
Study Design and Setting
This was a retrospective study to characterize patient care management problems identified by routine mortality and morbidity surveillance at an urban public teaching hospital during the 15-year period spanning 1989 through 2003. The annual adult ED census during those years ranged from 109,000 to 128,000. All cases referred for emergency physician review between 1989 and 2003 were eligible. Our institution has separate reporting mechanisms for pediatric and trauma care; these populations are not included in our study. Case referrals came from ED staff, admitting services, consultants, or quality assurance managers. Two independent emergency physician reviewers assessed each case by using an existing framework described previously.15,16 The lead author designed the data collection form and performed review 1 in every case. Cases were randomly assigned to one of 8 coinvestigators for review 2. Whenever possible, reviewers were disqualified from cases in which they had direct involvement or firsthand knowledge. The study was exempted from review by our institutional review board.

Our department actively sought referrals from hospital staff of cases in which there were concerns about quality of care and encouraged reports of administrative and system problems, as well as questions about medical management. This process began as routine surveillance for morbidity and mortality cases but became a significant part of the quality process in our department. Cases underwent an investigation at referral that were judged the most common locus of failure in more than 600 cases, spanning 15 years. Despite imperfections in care, more than half the patients still received some benefit compared with the likely outcome with no care at all.
as they existed at the time. Unlike many retrospective medical record reviews, these investigations yielded thick files rich in context and content. Each case was typically presented for discussion at 1 or more forums, including physician review committees, morbidity and mortality conferences, and faculty meetings. Once investigations were complete, the original medical records, notes from interviews and committee discussions, formal recommendations, and final case summaries were placed in an investigations library. Throughout the 15-year period of case collection summarized in this study, a common philosophy guided case review: cases were scrutinized with the intent to detect all patient care management problems, their potential causes, and contributing factors, with the ultimate goal to find solutions and seek improvement. These records constitute the source data reviewed for this study.

Data Collection
At the start of this study, a research assistant deidentified the source data from the investigations library. Study investigators reviewed the case files, identified management problems, and then classified the problems by the phase of work in which they occurred, their contributing factors, and patient outcome. Investigators were allowed to use summary statements and conclusions from the initial reviews but were also encouraged to record any other relevant factors they identified during their reviews of file documents. To ensure internal validity throughout the study, specific criteria were established for each data field before data collection. Reviewers were asked to consider each study category. If they selected a category, they were required to check at least 1 of the category criteria to justify their selection. Reviewers participated in educational sessions and a pilot study, with feedback to verify consistent application of study definitions.

The framework that guided this investigation applies a variety of approaches to categorizing problems in care.15 We determine the stage at which a problem occurred and what factors likely contributed. Identified problems were classified into 1 or more broad categories according to the phase of work in which they occurred: diagnostic, treatment, disposition, and public health (Table 1). Reviewers were then asked to determine whether any of the following contributing factors added risk, contributed to a problem in care, or had a negative impact on patient outcome: patient factors, triage, clinical factors, teamwork, and system. Clinical factors included medical reasoning, specific interpretive and procedural skill sets, task-based problems, and affective influences. The concept of task-based problems has been described previously: they involve specific tasks that are so routine and basic to care that they are relegated to lower-order thinking, that is, they are done as a matter of habit and typically with little conscious thought.15,17 Problems in these basic tasks are observed as a failed human behavior but likely reflect a system weakness. Such events tend to occur when the system is overloaded, workers are distracted, or altered staffing patterns affect work flow. They can be used as a marker of system overload or disorganization. The category of affective influences includes the human tendency to be influenced by factors other than objective facts; it includes the sometimes unconscious effects of personality, external pressures, and conflict on reasoning.18 Individual assessments were made for each team of clinicians. System problems included any process, service, equipment, or supplies that failed or were unavailable, broken, or faulty. System problems were further classified by location: ED microsystem, hospital-wide failure, administrative factors, and community resources. Specific problems were also characterized according to Reason’s17 scheme as problems in “planning,” “execution,” or both. This classification distinguishes problems that are primarily cognitive from those that are primarily related to process failure; this distinction offers additional insight into potential solutions. None of these descriptors or classifications was mutually exclusive. Judgments about care were based on whether actions and decisions were accurate, timely, and effective. An example of a training case analysis is demonstrated in Figure 1.

Outcome Measures
Primary patient outcome was categorized as death, permanent disability, long-term disability (between 30 days and 12 months), short-term disability (up to 30 days), or no disability. These outcomes were mutually exclusive. Final outcomes, however, do not always reflect significant events. Thus, we added 2 additional secondary outcome measures to capture events that deserve attention: acute but short-lived life-threatening events and adverse events (harm caused by medical management itself, independent of the patient’s disease). These secondary outcome measures were not mutually exclusive and could also overlap with the primary outcomes. Finally, each reviewer was asked to compare the actual patient outcome for each patient in this series to the outcome expected for optimal care and the likely outcome of no care to determine to what extent our management met expectations or caused harm.

Primary Data Analysis
The results for both reviewers were entered in a SAS data set. To increase data entry accuracy, 2 personnel collaborated: one to read aloud results from the data collection forms and validate the keystroke entry and the other to perform keystrokes. Observer agreement between reviewers 1 and 2 was quantified using the $\kappa$ statistic. For rates and proportions, 95% confidence intervals were calculated. All analyses were performed with SAS version 7 (SAS Institute, Cary, NC).

RESULTS
Of the original library of 673 cases, 37 were excluded from the study because of incomplete or illegible records, inadequate 30-day follow-up, incomplete investigations, or inconclusive summaries (4 unexplained deaths). The final study group of 636 patients includes 353 male patients, 276 female patients, and 7 with undesignated sex. (Seven cases had sex designation inadvertently removed from the original source data.)
There was significant overlap between the main categories of work in which problems were noted, with diagnosis being the most commonly identified classification (451; 71%), followed by disposition (280; 44%) and then treatment (265; 42%). Public health decisions comprised the smallest fraction of cases (23; 4%). Multiple categories were selected in more than half the cases (322; 51%) (Figure 2).

Problems were most likely to occur in the planning stage (591; 93%) compared with the execution stage (170; 27%), although some events had problems in both (144, 23%).
Problems in specific clinical tasks were the most common contributing factor identified (632; 99%) (Table 2). Problems in clinical tasks most commonly involved medical reasoning (595; 94%) but also included specific skill sets (212; 33%), routine task-based problems (173; 27%), and destructive affective influences (38; 6%). In almost half of all cases, more than 1 group of clinicians participated in specific management problems (289; 45%). Thus, many clinical problems did not occur in isolation and were often not the result of a single person, team, or process. Other leading factors that affected outcome included patient factors (391; 61%) and teamwork (387; 61%). System and triage factors were selected to a lesser extent (41% and 16%, respectively). The distribution of contributing factors was similar across all phases of work. More than 1 contributing factor was identified for most cases. In fact, 2 or more factors were identified in all but 56 cases (92%); 420 cases (66%) had 3 or more factors identified; 113 cases (18%) had 5 or more factors identified. Specific problems in care were typically not the result of failure by 1 factor (person or process) alone but rather the result of the complex interplay of multiple factors that ultimately influenced clinical work.

Two thirds (66%) of patients in this series had death, major permanent disability, or disability exceeding 30 days; 34%
LIMITATIONS

This study is a retrospective review of investigations completed on cases referred because of poor outcome, flawed processes, or the perception of imperfections in care. There are several limitations inherent to the study design.

First, the study has selection bias. Because our cases came largely from physician referrals, our data are weighted toward the types of problems that physicians find interesting or important and do not necessarily reflect a broad sampling of all types of cases. This does not mean that they represent the most common or even the “worst” management problems; our cases probably represent the most visible events or those most likely to have immediate impact on patient outcome. There is no control group of randomly selected patients during the same study period. Thus, many of the same factors identified in the study sample may have been present in other cases without harm. Cases with significant problems or harm may have occurred during the same period but gone unreported.

Second, post hoc critiques of care are inevitably subject to bias. Our cases were typically referred after outcome was known; thus, they have outcome bias, the tendency to judge similar events more harshly when the outcome is poor. In addition, our investigations were challenged by our ability to judge the events as they actually occurred, not as we might imagine them. Hindsight bias, the reconstruction of events in a way that makes sense to the investigator, can mislead judgments about events. Even clinicians present at the time, in an effort to explain their actions, may reconstruct memories that are only partially true. We attempted to minimize the influence of bias by encouraging staff to record their impressions and recall specific information they had or lacked at the time of their decisions, what factors influenced their decisions, and what the ambient conditions were in the ED at the time. Timelines of actions were generated from electronic records and added to information from interviews to reconstruct, as much as possible, the actual events. The thick descriptions in our study files permit detailed analysis aimed at providing contextually rich detail to explain the circumstances surrounding care.

There are several other limitations to our study. The type of factors identified is constrained by the expertise of the reviewers and the quality of information they preserved. Although we observe the presence of factors that may have contributed to outcome, we do not make conclusions of causality.

Although our study attempted to define discrete phases of work in which problems occurred, we found overlap between the major categories, which suggests that our classifications (and perhaps the events themselves) were not independent of one another; clinical decisions in one phase of work likely affect other phases of work. Failure to discriminate between these categories also reflects several natural limitations in quality reviews. First, it is difficult to separate a single flawed step in the continuum of care. Second, a critical appraisal of many cases, with good or bad outcome, can often find multiple flaws. Last, such distinctions require judgments about what was in the mind of the clinician, facts that may be difficult for the involved clinicians to recall accurately and even more difficult for an independent reviewer remote from the event to determine.

The language of safety science and models of causation are controversial, particularly when applied to health care. Terms

Figure 2. Major categories of work in which problems were observed. This Venn diagram illustrates the number of events identified in each phase of work and their overlap. Numbers in parentheses indicate total numbers for each major heading. Public health decisions, not shown in this diagram, accounted for an additional 23 events (4% of all events).
such as “error” tend to evoke accusations of blame, whereas terms such as “problems” or “failures” may be vague. Judgments about quality and causality are flawed by outcome knowledge. From a simple human perspective, we want patients to have the best possible outcome. When they do not, there is a natural tendency to cite any imperfection in their care as a contributing (if not causal) factor. No doubt, similar problems in the care of patients with satisfactory results are more easily forgiven and dismissed. This study is ultimately a summary of the aspects of clinical work that our reviewers found to be most prone to problems and thus most likely to contribute to risk. The development of safety systems in other fields has started by analysis of accidents or events and then progressed to more sophisticated, systematic, and prospective design. Studies such as ours provide a framework from which to begin.

**DISCUSSION**

Our data demonstrate that cases referred because of patient care management problems are often complex and not the result of the failure of any single individual or process at any one moment in time. The typical course of a “medical error” has been described as a cascade, in which one failure leads to another.22 However, our study went beyond simply recognizing complexity to identifying specific factors that can be targeted to improve safety.

Our reviewers selected “diagnosis” as the leading phase of work in which problems were noticed, consistent with results from other major studies,2,4,23-25 which is not unexpected, because diagnosis is at the heart of clinical work and is the foundation on which all other actions are predicated. That judgment is likely influenced by the perspective of our reviewers, who may be inclined to focus on their own expertise and reluctant to judge factors outside their usual sphere of influence. Our results provide some clues about why the diagnostic process is difficult and at times imprecise. Two major contributing factors were consistently identified in most cases in our study and likely created an environment in which rapid decisions may be problematic: patient factors and teamwork.

**Table 2a.** Association between contributing factors and primary outcome measures.

<table>
<thead>
<tr>
<th>Contributing Factors</th>
<th>All Cases</th>
<th>Death</th>
<th>Permanent,</th>
<th>Long Term,</th>
<th>Short Term,</th>
<th>None,</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=636</td>
<td>n=226</td>
<td>n=24</td>
<td>n=171</td>
<td>n=178</td>
<td>n=37</td>
</tr>
<tr>
<td>Patient</td>
<td>391 (61)</td>
<td>170 (75)</td>
<td>16 (67)</td>
<td>86 (50)</td>
<td>102 (57)</td>
<td>17 (46)</td>
</tr>
<tr>
<td>Triage</td>
<td>103 (16)</td>
<td>47 (21)</td>
<td>2 (8)</td>
<td>20 (12)</td>
<td>31 (17)</td>
<td>3 (8)</td>
</tr>
<tr>
<td>Clinical (all subsets)</td>
<td>632 (99)</td>
<td>225 (99)</td>
<td>23 (96)</td>
<td>171 (100)</td>
<td>177 (99)</td>
<td>36 (97)</td>
</tr>
<tr>
<td>Reasoning</td>
<td>595 (94)</td>
<td>212 (94)</td>
<td>21 (88)</td>
<td>164 (96)</td>
<td>169 (95)</td>
<td>29 (78)</td>
</tr>
<tr>
<td>Skill set</td>
<td>212 (33)</td>
<td>77 (34)</td>
<td>12 (50)</td>
<td>47 (27)</td>
<td>58 (33)</td>
<td>18 (49)</td>
</tr>
<tr>
<td>Task based</td>
<td>173 (27)</td>
<td>91 (40)</td>
<td>7 (29)</td>
<td>29 (17)</td>
<td>37 (21)</td>
<td>9 (24)</td>
</tr>
<tr>
<td>Affective influences</td>
<td>38 (6)</td>
<td>11 (5)</td>
<td>2 (8)</td>
<td>9 (5)</td>
<td>14 (8)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>Teamwork</td>
<td>387 (61)</td>
<td>155 (66)</td>
<td>13 (54)</td>
<td>92 (54)</td>
<td>100 (56)</td>
<td>27 (73)</td>
</tr>
<tr>
<td>System</td>
<td>261 (41)</td>
<td>103 (46)</td>
<td>9 (38)</td>
<td>79 (46)</td>
<td>60 (34)</td>
<td>10 (27)</td>
</tr>
</tbody>
</table>

*Contributing factors are not mutually exclusive; thus, percentages do not sum to 100.

**Table 2b.** Association between contributing factors and secondary outcome measures.

<table>
<thead>
<tr>
<th>Contributing Factors</th>
<th>All Cases, No. (%), n=636</th>
<th>Life Threats, n=66</th>
<th>Adverse Events, n=86</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>391 (61)</td>
<td>47 (71)</td>
<td>59 (69)</td>
</tr>
<tr>
<td>Triage</td>
<td>103 (16)</td>
<td>11 (17)</td>
<td>10 (12)</td>
</tr>
<tr>
<td>Clinical (all subsets)</td>
<td>632 (99)</td>
<td>66 (100)</td>
<td>84 (98)</td>
</tr>
<tr>
<td>Reasoning</td>
<td>595 (94)</td>
<td>60 (91)</td>
<td>69 (80)</td>
</tr>
<tr>
<td>Skill set</td>
<td>212 (33)</td>
<td>28 (42)</td>
<td>49 (57)</td>
</tr>
<tr>
<td>Task based</td>
<td>173 (27)</td>
<td>23 (35)</td>
<td>30 (35)</td>
</tr>
<tr>
<td>Affective influences</td>
<td>38 (6)</td>
<td>7 (11)</td>
<td>5 (6)</td>
</tr>
<tr>
<td>Teamwork</td>
<td>387 (61)</td>
<td>40 (61)</td>
<td>50 (58)</td>
</tr>
<tr>
<td>System</td>
<td>261 (41)</td>
<td>22 (33)</td>
<td>30 (35)</td>
</tr>
</tbody>
</table>

*Secondary outcomes note specific events that are a subset of all cases; thus, percentages do not sum to 100.

*Contributing factors are not mutually exclusive; thus, percentages do not sum to 100.
history, which contributes to the “information gap” described by Stiell et al,26 in which physicians must act despite a paucity of reliable facts. Some of these factors are inherent to the practice of emergency medicine and contribute substantially to risk.

Other acute care specialties have recognized the importance of teamwork; our study confirmed that problems in care often involve ineffective team functioning.27,28 ED care tends to be punctuated by numerous transitions that can lead to dropped tasks and lost information. In addition, handoffs in care may hamper the ability of clinicians to recognize dynamic changes in a patient’s condition.29

There are implications from this study that can be used to design safer care. First, clinicians need reliable access to information and strategies for making decisions in the face of uncertainty. Cognitive psychologists can offer specific training in cognition, critical thinking, and decisionmaking.30-38 Clinical decision support systems and cognitive aids can reduce cognitive load and standardize clinical protocols.39,40 Medical simulation can provide opportunities to practice and refine performance.41

Second, our study reveals that patient factors contribute substantially to management problems. Vincent et al16,42 have observed that patient factors may contribute to the likelihood of an adverse event; their framework for analyzing clinical events includes consideration of patient factors. Patient factors are often beyond our immediate control; harm caused by some patient factors has even been labeled “no fault.”38,43 Although a no-fault judgment may be fair to those present at the time, it prematurely limits discussions about how to improve care. We argue that all information from investigations can be used to improve care, even if problems are judged to be unavoidable at the time. Creative use of resources can address common patient factors, such as providing 24-hour access to interpreter services, improving electronic medical record databases for access to reliable medical information, and using ultrasonographic guidance to improve our ability to care for patients with difficult venous access, to name just a few.

Third, organization and teamwork in acute care medicine present specific challenges. Reliable and consistent processes for transitions and exchange of patient care responsibility and accurate communication of test results and patient information are fundamental to the design of safe health care systems, particularly in the dynamic setting of the ED.

The fact that system factors were not identified as a leading contributor to problems identified in this study contrasts sharply with current views in safety science.8,44 Our study was

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Cases</th>
<th>Total, % (95% CI)</th>
<th>Observer Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phases of work in which problems were observed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnosis</td>
<td>451</td>
<td>71 (67–75)</td>
<td>98</td>
</tr>
<tr>
<td>Treatment</td>
<td>265</td>
<td>42 (38–46)</td>
<td>93</td>
</tr>
<tr>
<td>Disposition</td>
<td>280</td>
<td>44 (40–48)</td>
<td>86</td>
</tr>
<tr>
<td>Public health</td>
<td>23</td>
<td>4 (2–6)</td>
<td>100</td>
</tr>
<tr>
<td>Reason’s classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Planning</td>
<td>591</td>
<td>93 (96–98)</td>
<td>92</td>
</tr>
<tr>
<td>Execution</td>
<td>170</td>
<td>27 (24–30)</td>
<td>79</td>
</tr>
<tr>
<td>Planning and execution</td>
<td>144</td>
<td>23 (20–26)</td>
<td>79</td>
</tr>
<tr>
<td>Contributing factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient</td>
<td>391</td>
<td>61 (57–65)</td>
<td>77</td>
</tr>
<tr>
<td>Triage</td>
<td>103</td>
<td>16 (13–19)</td>
<td>93</td>
</tr>
<tr>
<td>Clinical (all subsets)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reasoning</td>
<td>595</td>
<td>94 (92–96)</td>
<td>99</td>
</tr>
<tr>
<td>Specific skill set</td>
<td>212</td>
<td>33 (29–37)</td>
<td>90</td>
</tr>
<tr>
<td>Task based</td>
<td>173</td>
<td>27 (24–30)</td>
<td>82</td>
</tr>
<tr>
<td>Affective influences</td>
<td>38</td>
<td>6 (4–8)</td>
<td>92</td>
</tr>
<tr>
<td>Teamwork</td>
<td>387</td>
<td>61 (57–65)</td>
<td>72</td>
</tr>
<tr>
<td>System</td>
<td>261</td>
<td>41 (37–45)</td>
<td>83</td>
</tr>
<tr>
<td>Primary outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death</td>
<td>226</td>
<td>36 (32–40)</td>
<td>100</td>
</tr>
<tr>
<td>Permanent disability</td>
<td>24</td>
<td>4 (2–5)</td>
<td>100</td>
</tr>
<tr>
<td>Long-term disability</td>
<td>171</td>
<td>27 (23–30)</td>
<td>100</td>
</tr>
<tr>
<td>Short-term disability</td>
<td>178</td>
<td>28 (24–31)</td>
<td>100</td>
</tr>
<tr>
<td>No disability</td>
<td>37</td>
<td>6 (4–8)</td>
<td>100</td>
</tr>
<tr>
<td>Secondary outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life-threatening events</td>
<td>66</td>
<td>10 (8–12)</td>
<td>100</td>
</tr>
<tr>
<td>Adverse events</td>
<td>86</td>
<td>14 (11–17)</td>
<td>100</td>
</tr>
</tbody>
</table>

CI, Confidence interval.

*Categories are not mutually exclusive; thus, percentages do not sum to 100.
performed by physicians, not safety scientists; thus, their reviews focus on medical judgments, not system design. Human tasks and the systems that support them are inextricably linked. The fact that our reviewers consistently selected human aspects of tasks over system factors may be due to fundamental attribution error, the tendency for organizational actors to attribute undesirable outcomes to the perceived character flaws of people rather than to the processes engendered by the organization’s structure. This tendency is reinforced by cultural pressures that promote an overriding sense of professional accountability but create “system blindness,” that is, clinicians are largely unaware of the potential for system design to improve care. The failure to detect a similar proportion of human and system factors may also reflect an imbalance in health care design that encourages reliance on individuals over development of support systems. If so, the recent focus on design and development of safe systems to support clinical work offers significant promise for improvement.

Future safety research will likely focus on prospective error reporting systems. The development of a national error reporting mechanism in the United States is under way but faces challenges to define process: how to collect case information, what information to collect, and how to analyze and then disseminate useful results. Once in place, reporting systems are often slow to produce actionable items. Meanwhile, studies such as ours can use historical cases to identify major problems needing correction and identify the types of factors that are worth tracking prospectively.

In the course of this study, we learned several lessons that can affect the success of future error reporting systems. During the early phases of the study, we attempted to identify a variety of potential contributing factors based on safety research, including cognitive bias, human factors, ergonomics, and ED design. Our efforts failed, primarily because physician reviewers were not familiar with the concepts and routine clinical investigations do not typically address them. We recommend that future studies and error reporting systems engage nonclinician experts from safety disciplines to provide fresh insight and perspective.

We also found that observer agreement was best for categories that were described by detailed checklists and definitions. Reviewers were less likely to agree when asked to make subjective assessments. Large-scale reporting systems need a workable reporting mechanism and may struggle between the need for uniformity and objective definitions and the desire for free-text reports. An ideal system may need both and may need to be modified periodically to allow for evolving standards of care and improved understanding of causes of medical failure.

For future work, there is a variety of sources of data that can be used to probe problems in patient care. Dramatic, visible, and highly publicized accounts of individual cases have been used successfully to drive safety initiatives. In contrast, aggregate data from existing databases of cases may provide a broad overview of risk and help define particular tasks and moments in patient care that are most vulnerable and most in need of system support. Several studies have used closed malpractice claims and risk management files to identify causes and contributing factors to adverse events. We suggest that collective reviews of existing mortality and morbidity investigations and internal review processes, available in most health care organizations, offer an additional source of data to guide improvement (Appendix E1, available online at www.annemergmed.com). Finally, we suggest that collaboration with safety experts in other disciplines can expand our perspective on safety concepts and the potential for system improvements in health care.

In summary, the primary goal for understanding patient care management problems is to prevent them or at least mitigate harm. However, our efforts to improve care have been frustrated by the realization that health care and patient care processes are complex. Our data confirm that most problems are multifactorial. The leading factors that contributed to problems detected by physician reviewers in this study include medical reasoning, patient factors, and teamwork, although these problems occurred in a background of system imperfections. Efforts to improve safety in medicine should be directed at improving access to accurate medical information and communication across the continuum of medical care; coordination of care among health care teams; and improved processes, education, and support for decisionmaking in the uncertain environment of acute care medicine. Advances in system design to support these aspects of clinical care offer opportunities for improvement in health care safety.

We wish to thank Mr. Geoffrey Andrade for assistance with data entry and Mr. Donald Bockenfeld, BS, for assistance with article preparation.
funded in part by the Agency for Healthcare Research and Quality, grant number 5 P20 HS011552, and the Department of Emergency Medicine, Cook County Hospital (Stroger).

Publication dates: Received for publication November 18, 2006. Revisions received March 8, 2007; May 21, 2007; and June 14, 2007. Accepted for publication June 25, 2007. Available online October 15, 2007.

Address for reprints: Karen Cosby, MD, Department of Emergency Medicine, Cook County Hospital (Stroger), 10th floor Administration Building, 1900 W Polk St, Chicago, IL 60612; 312-864-1986 or 312-864-0060, fax 312-864-9656; E-mail kcosby@ccbh.org.

REFERENCES


IMAGES IN EMERGENCY MEDICINE
(continued from p. 230)

DIAGNOSIS:

Emphysematous cystitis. Emphysematous cystitis is a rare, necrotizing infection characterized by gas collection in the urinary bladder wall and lumen, resulting from gas-producing pathogen infection. The risk factors are diabetes (up to 80%), bladder outlet obstruction, recurrent urinary tract infection, urinary stasis, neurogenic bladder, immunosuppression, female sex, and being a transplant recipient. The mechanism and pathogenesis of emphysematous cystitis are still unknown. The gas is suggested to be produced by the infected organism by the fermentation of albumin or glucose in urine. The most common organisms are Escherichia coli, Enterobacter aerogenes, and Klebsiella pneumoniae. Emphysematous cystitis has nonspecific clinical features and is often misdiagnosed. Clinically, emphysematous cystitis is often diagnosed by the unanticipated imaging findings. Plain abdominal radiograph usually makes the diagnosis, with high sensitivity (97.4%), but abdominal computed tomography scan was the most sensitive and specific diagnostic tool.

About 18.8% of emphysematous cystitis cases have complicated courses. Emphysematous cystitis demands prompt diagnosis and intervention, including aggressive parenteral antibiotics and even bladder drainage. Generally, emphysematous cystitis has favorable prognosis, whereas delays in diagnosis and treatment may contribute to high mortality rate, which approaches 20%.

A favorable prognosis may be achieved by early recognition of emphysematous cystitis, by clinical and radiologic assessment, by appropriate antibiotic use, and by timely surgical intervention when indicated.

This patient was administrated empiric antibiotic and promoted surgical drainage. Escherichia coli was isolated subsequently from both urine and drainage pus cultures. The patient was discharged after a 2-week hospitalization.

REFERENCES
APPENDIX E1.
Examples of actions resulting from morbidity and mortality investigations.

Uncommon Diagnosis Requiring Complex Evaluations and Multispecialty Interventions

The Problem: After observing difficulties in the timely diagnosis and treatment of aortic dissections, we reviewed a series of cases of aortic dissections from our morbidity and mortality database and identified key problem areas. The diagnosis was observed to be difficult, often overlapping more common chest pain entities. The evaluation for dissection could involve a variety of approaches (echocardiography, computed tomography, or angiography) and involved multiple consultants (chest surgeons, cardiologists, vascular surgeons, intensivists, and interventional radiologists). Because of the variation in practice, there was often conflict between specialists and services about the optimal testing and management of these cases; care was often delayed by conflict and indecision.

Actions Taken: After review of a series of problematic cases, the ED staff convened a multidisciplinary conference with all the involved specialties, eventually developing a standard approach agreed on by all disciplines. Once the policy was developed, hospital staff worked to ensure that services were available to follow the protocol. The result was a much more simplified and direct approach to caring for similar cases.

Cases With Evolving Standards of Care and Controversy About Optimal Management: Ectopic Pregnancy

The Problem: Multiple problems were observed in the timely diagnosis of ectopic pregnancy and in management of patients with first trimester bleeding. Because of evolving standards in the care of patients with first trimester pregnancy, the ED staff experienced conflict and inconsistent standards within their own group, as well as consultants. In the process, the diagnosis of several ectopic pregnancies was delayed.

Actions Taken: The ED organized a literature review and grand rounds conference on ectopic pregnancy, developed a treatment protocol for first trimester pregnancy, and then met with a working group with obstetrics to standardize criteria for consultation, admission, and follow-up. Meanwhile, the ED improved its proficiency in bedside ultrasonography and began more liberal screening of all first trimester pregnancy bleeding.

Dealing With a Local Infectious Disease Threat: Tuberculosis

The Problem: A resurgence of tuberculosis in our city led to an increase in number of patients with active tuberculosis in our ED. There were a limited number of isolation beds available within the hospital, and medical staff had too few beds to isolate all potential cases. A number of patients admitted to general medical wards were found to have active tuberculosis; at the same time, there was an increase in the number of house staff who converted to positive tuberculin skin tests.

The Action: Cases with positive tuberculosis culture results were identified and their ED records tracked. The ED convened a multidisciplinary conference with the Infectious Disease and Radiology Departments to review the cases. Ultimately, we found that many active cases of tuberculosis could not be predicted according to clinical and radiographic criteria. Eventually, the medical staff obtained additional isolation rooms in the ED and in the hospital, applied new screening at triage to prioritize chest radiographs in patients with respiratory symptoms, prioritized movement of patients with suspected cases from triage to isolation beds in the ED, and successfully argued for increased resources to handle the challenge.
WHAT ARE THE CAUSES OF MISSED AND DELAYED DIAGNOSES IN THE AMBULATORY SETTING?

Gandhi TK, Kachalia A, Thomas EJ et al. 
Missed and Delayed Diagnoses in the Ambulatory Setting: A Study of Closed Malpractice Claims

Tejal K. Gandhi, MD, MPH; Allen Kachalia, MD, JD; Eric J. Thomas, MD, MPH; Ann Louise Puopolo, BSN, RN; Catherine Yoon, MS; Troyen A. Brennan, MD, JD; and David M. Studdert, LLB, ScD

Background: Although missed and delayed diagnoses have become an important patient safety concern, they remain largely unstudied, especially in the outpatient setting.

Objective: To develop a framework for investigating missed and delayed diagnoses, advance understanding of their causes, and identify opportunities for prevention.

Design: Retrospective review of 307 closed malpractice claims in which patients alleged a missed or delayed diagnosis in the ambulatory setting.

Setting: 4 malpractice insurance companies.

Measurements: Diagnostic errors associated with adverse outcomes for patients, process breakdowns, and contributing factors.

Results: A total of 181 claims (59%) involved diagnostic errors that harmed patients. Fifty-nine percent (106 of 181) of these errors were associated with serious harm, and 30% (55 of 181) resulted in death. For 59% (106 of 181) of the errors, cancer was the diagnosis involved, chiefly breast (44 claims [24%]) and colorectal (13 claims [7%]) cancer. The most common breakdowns in the diagnostic process were failure to order an appropriate diagnostic test (100 of 181 [55%]), failure to create a proper follow-up plan (81 of 181 [45%]), failure to obtain an adequate history or perform an adequate physical examination (76 of 181 [42%]), and incorrect interpretation of diagnostic tests (67 of 181 [37%]). The leading factors that contributed to the errors were failures in judgment (143 of 181 [79%]), vigilance or memory (106 of 181 [59%]), knowledge (86 of 181 [48%]), patient-related factors (84 of 181 [46%]), and handoffs (36 of 181 [20%]). The median number of process breakdowns and contributing factors per error was 3 for both (interquartile range, 2 to 4).

Limitations: Reviewers were not blinded to the litigation outcomes, and the reliability of the error determination was moderate.

Conclusions: Diagnostic errors that harm patients are typically the result of multiple breakdowns and individual and system factors. Awareness of the most common types of breakdowns and factors could help efforts to identify and prioritize strategies to prevent diagnostic errors.


Missed and delayed diagnoses in the ambulatory setting are an important patient safety problem. The current diagnostic process in health care is complex, chaotic, and vulnerable to failures and breakdowns. For example, one third of women with abnormal results on mammography or Papanicolaou smears do not receive follow-up care that is consistent with well-established guidelines (1, 2), and primary care providers often report delays in reviewing test results (3). Recognition of systemic problems in this area has prompted urgent calls for improvements (4).

However, this type of error remains largely unstudied (4). At least part of the reason is technical: Because omissions characterize missed diagnoses, they are difficult to identify; there is no standard reporting mechanism; and when they are identified, documentation in medical records is usually insufficiently detailed to support detailed causal analyses. The result is a relatively thin evidence base from which to launch efforts to combat diagnostic errors. Moreover, conceptions of the problem tend to remain rooted in the notion of physicians failing to be vigilant or up-to-date. This is a less nuanced view of error causation than careful analysis of other major patient safety problems, such as medication errors (5, 6), has revealed.

Several considerations highlight malpractice claims as a potentially rich source of information about missed and delayed diagnoses. First, misdiagnosis is a common allegation. Over the past decade, lawsuits alleging negligent misdiagnoses have become the most prevalent type of claim in the United States (7, 8). Second, diagnostic breakdowns that lead to claims tend to be associated with especially severe outcomes. Third, relatively thorough documentation on what happened is available in malpractice insurers’ claim files. In addition to the medical record, these files include depositions, expert opinions, and sometimes the results of internal investigations.

Previous attempts to use data from malpractice claims to study patient safety have had various methodologic constraints, including small sample size (9, 10), a focus on single insurers (11) or verdicts (9, 10) (which constitute <10% of claims), limited information on the claims (8–11), reliance on internal case review by insurers rather than by independent experts (8, 11), and a general absence of

See also:

Print
Editors’ Notes 489
Editorial comment 547
Summary for Patients I-12

Web-Only
Appendix
Appendix Tables
Appendix Figures
Conversion of tables and figures into slides
robust frameworks for classifying types and causes of failures. To address these issues, we analyzed data from closed malpractice claims at 4 liability insurance companies. Our goals were to develop a framework for investigating missed and delayed diagnoses, advance understanding of their causes, and identify opportunities for prevention.

**Methods**

**Study Sites**

Four malpractice insurance companies based in 3 regions (northeastern, southwestern, and western United States) participated in the study. Collectively, the participating companies insured approximately 21000 physicians, 46 acute care hospitals (20 academic and 26 nonacademic), and 390 outpatient facilities, including a wide variety of primary care and outpatient specialty practices. The ethics review boards at the investigators’ institutions and at each review site approved the study.

**Claims Sample**

Data were extracted from random samples of closed claim files from each insurer. A claim is classified as closed when it has been dropped, dismissed, paid by settlement, or resolved by verdict. The claim file is the repository of information accumulated by the insurer during the life of a claim. It captures a wide variety of data, including the statement of claim, depositions, interrogatories, and other litigation documents; reports of internal investigations, such as risk management evaluations and sometimes root-cause analyses; expert opinions from both sides; medical reports detailing the plaintiff’s preevent and postevent condition; and, while the claim is open, medical records pertaining to the episode of care at issue. We reacquired the relevant medical records for sampled claims.

Following previous studies, we defined a claim as a written demand for compensation for medical injury (12, 13). Claims involving missed or delayed diagnoses were defined as those alleging an error in diagnosis or testing that caused a delay in appropriate treatment or a failure to act or follow up on results of diagnostic tests. We excluded allegations related to pregnancy and those pertaining to care rendered solely in the inpatient setting.

We reviewed 429 diagnostic claims alleging injury due to missed or delayed diagnoses. Insurers contributed to the study sample in proportion to their annual claims volume (Appendix, available at www.annals.org). The claims were divided into 2 main categories based on the primary setting of the outpatient care involved in the allegation: the emergency department (122 claims) and all other locations (for example, physician’s office, ambulatory surgery, pathology laboratory, or radiology suites) (307 claims). The latter group, which we call ambulatory claims, is the focus of this analysis.

**Study Instruments and Claim File Review**

Physicians who were board-certified attendings, fellows, or third-year residents in internal medicine reviewed sampled claim files at the insurers’ offices or insured facilities. Physician-investigators trained the reviewers in the content of claim files, use of the study instruments, and confidentiality procedures in 1-day sessions at each site. The reviewers also used a detailed manual. Reviews took on average 1.4 hours per file. To test review reliability, a second reviewer reviewed a random selection of 10% (42 of 429) of the files. Thirty-three of the 307 ambulatory claims that are the focus of this analysis were included in the random blinded re-review. A sequence of 4 instruments guided the review. For all claims, insurance staff recorded administrative details of the case (Appendix Figure 1, available at www.annals.org), and clinical reviewers recorded details of the adverse outcome the patient experienced, if any (Appendix Figure 2, available at www.annals.org). Reviewers scored adverse outcomes on a 9-point severity scale ranging from emotional injury only to death. This scale was developed by the National Association of Insurance Commissioners (14) and has been used in previous research (15). If the patient had multiple adverse outcomes, reviewers scored the most severe outcome. To simplify presentation of our results, we grouped scores on

**Context**

Efforts to reduce medical errors and improve patient safety have not generally addressed errors in diagnosis. As with treatment, diagnosis involves complex, fragmented processes within health care systems that are vulnerable to failures and breakdowns.

**Contributions**

The authors reviewed malpractice claims alleging injury from a missed or delayed diagnosis. In 181 cases in which there was a high likelihood that error led to the missed diagnosis, the authors analyzed where the diagnostic process broke down and why. The most common missed diagnosis was cancer, and the most common breakdowns were failure to order appropriate tests and inadequate follow-up of test results. A median of 3 process breakdowns occurred per error, and 2 or more clinicians were involved in 43% of cases.

**Cautions**

The study relied on malpractice claims, which are not representative of all diagnostic errors that occur. There was only moderate agreement among the authors in their subjective judgments about errors and their causes.

**Implications**

Like other medical errors, diagnostic errors are multifactorial. They arise from multiple process breakdowns, usually involving multiple providers. The results highlight the challenge of finding effective ways to reduce diagnostic errors as a component of improving health care quality.

—The Editors
this scale into 5 categories (emotional, minor, significant, major, and death).

Next, reviewers considered the potential role of a series of contributing factors (Appendix Figure 3, available at www.annals.org) in causing the adverse outcome. The factors covered cognitive-, system-, and patient-related causes that were related to the episode of care as a whole, not to particular steps in the diagnostic process. The factors were selected on the basis of a review of the patient safety literature performed in 2001 by 5 of the authors in consultation with physician-collaborators from surgery and obstetrics and gynecology.

Reviewers then judged, in light of available information and their decisions about contributing factors, whether the adverse outcome was due to diagnostic error. We used the Institute of Medicine’s definition of error, namely, “the failure of a planned action to be completed as intended (i.e. error of execution) or the use of a wrong plan to achieve an aim (i.e. error of planning)” (16). Reviewers recorded their judgment on a 6-point confidence scale ranging from “1. Little or no evidence that adverse outcome resulted from error/errors” to “6. Virtually certain evidence that adverse outcome resulted from error/errors.” Claims that scored 4 (“More likely than not that adverse outcome resulted from error/errors; more than 50-50 but a close call”) or higher were classified as having an error. The confidence scale and cutoff point were adapted from instruments used in previous studies of medical injury (15, 17).

Reviewers were not blinded to the litigation outcomes but were instructed to ignore them and rely on their own clinical judgment in making decisions about errors. Training sessions stressed that the study definition of error is not synonymous with the legal definition of negligence and that a mix of factors extrinsic to merit influence whether claims are paid during litigation.

Finally, for the subset of claims judged to involve errors, reviewers completed an additional form (Appendix Figure 4, available at www.annals.org) that collected additional clinical information about the missed diagnosis. Specifically, reviewers considered a defined sequence of diagnostic steps (for example, history and physical examination, test ordering, and creation of a follow-up plan) and were asked to grade their confidence that a breakdown had occurred at each step (5-point Likert scale ranging from “highly unlikely” to “highly likely”). If a breakdown was judged to have been at least “somewhat likely” (score of ≥3), the form elicited additional information on the particular breakdown, including a non–mutually exclusive list of reasons for the breakdown.

Statistical Analysis

The primary unit of analysis was the sequence of care in claims judged to involve a diagnostic error that led to an adverse outcome. For ease of exposition, we henceforth refer to such sequences as errors. The hand-filled data forms were electronically entered and verified by a professional data entry vendor and sent to the Harvard School of Public Health for analysis. Additional validity checks and data cleaning were performed by study programmers. Analyses were conducted by using SAS, version 8.2 (SAS Institute, Cary, North Carolina) and Stata SE, version 8.0 (Stata Corp., College Station, Texas). We examined characteristics of the claims, patients, and injuries in our sample and the frequency of the various contributing factors. We compared characteristics of error subgroups using Pearson chi-square tests and used percentage agreement and κ scores (18) to measure interrater reliability of the injury and error determinations.

Role of the Funding Sources

This study was funded by grants from the Agency for Healthcare Research and Quality (HS011886-03) and the Harvard Risk Management Foundation. These organizations did not play a role in the design, conduct, or analysis of the study, and the decision to submit the manuscript for publication was that of the authors.

RESULTS

The 307 diagnosis-related ambulatory claims closed between 1984 and 2004. Eighty-five percent (262 of 307) of the alleged errors occurred in 1990 or later, and 80% (245 of 307 claims) closed in 1997 or later. In 2% (7 of 307) of claims, no adverse outcome or change in the patient’s clinical course was evident; in 3% (9 of 307), the reviewer was unable to judge the severity of the adverse outcome from the information available; and in 36% (110 of 307), the claim was judged not to involve a diagnostic error. The remaining group of 181 claims, 59% (181 of 307) of the sample, were judged to involve diagnostic errors that led to adverse outcomes. This group of errors is the focus of further analyses.

In 40 of the 42 re-reviewed claims, reviewers agreed about whether an adverse outcome had occurred (95% agreement). The reliability of the determination of whether an error had occurred (a score <4 vs. ≥4 on the confidence scale) was moderate (72% agreement; κ = 0.42 [95% CI, −0.05 to 0.66]).

Errors and Diagnoses

Fifty-nine percent (106 of 181) of the errors were associated with significant or major physical adverse outcomes and 30% (55 of 181) were associated with death (Table 1). For 59% (106 of 181) of errors, cancer was the diagnosis missed, chiefly breast (44 of 181 [24%]), colorectal (13 of 181 [7%]), and skin (8 of 181 [4%]) cancer. The next most commonly missed diagnoses were infections (9 of 181 [5%]), fractures (8 of 181 [4%]), and myocardial infarctions (7 of 181 [4%]). Most errors occurred in physicians’ offices (154 of 181 [85%]), and primary care physicians were the providers most commonly involved (76 of 181 [42%]). The mean interval between when diagnoses should have been made (that is, in the absence of error)
and when they actually were made was 465 days (SD, 571 days), and the median was 303 days (interquartile range, 36 to 681 days).

Appendix Table 1 (available at www.annals.org) outlines several examples of missed and delayed diagnoses in the study sample.

**Breakdowns in the Diagnostic Process**

The leading breakdown points in the diagnostic process were failure to order an appropriate diagnostic test (100 of 181 [55%]), failure to create a proper follow-up plan (81 of 181 [45%]), failure to obtain an adequate history or to perform an adequate physical examination (76 of 181 [42%]), and incorrect interpretation of a diagnostic test (67 of 181 [37%] (Table 2). Missed cancer diagnoses were significantly more likely to involve diagnostic tests being performed incorrectly (14 of 106 [13%] vs. 1 of 75 [1%]; P < 0.004) and being interpreted incorrectly (49 of 106 [46%] vs. 18 of 75 [24%]; P = 0.002), whereas missed noncancer diagnoses were significantly more likely to involve delays by patients in seeking care (4 of 106 [4%] vs. 12 of 75 [16%]; P < 0.001), inadequate follow-up plan (24 of 106 [23%] vs. 52 of 75 [69%]; P < 0.001), and failure to refer (19 of 106 [18%] vs. 28 of 75 [37%]; P = 0.003).

Table 3 details the 4 most common process breakdowns. Among failures to order diagnostic tests, most tests involved were from the imaging (45 of 100 [45%]) or other test (50 of 100 [50%]) categories. With respect to the specific tests involved, biopsies were most frequently at issue (25 of 100 [25%]), followed by computed tomography scans, mammography, ultrasonography, and colonoscopy (11 of 100 for each [11%]). The most common explanation for the failures to order was that the physician seemed to lack knowledge of the appropriate test in the clinical circumstances. Among misinterpretations of test results, 27% (18 of 67) related to results on mammography and 15% (10 of 67) related to results on radiography. The main reasons for inadequate follow-up plans were that the physician did not think follow-up was necessary (32 of 81 [40%]), selected an inappropriate follow-up interval (29 of

### Table 1. Characteristics of Patients, Involved Clinicians, Adverse Outcomes, and Missed or Delayed Diagnoses among 181 Diagnostic Errors in Ambulatory Care

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Errors, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female patient</strong></td>
<td>110 (61)</td>
</tr>
<tr>
<td><strong>Patient age (mean, 44 y [SD, 17])</strong></td>
<td></td>
</tr>
<tr>
<td>&lt;1 y</td>
<td>5 (3)</td>
</tr>
<tr>
<td>1–18 y</td>
<td>9 (5)</td>
</tr>
<tr>
<td>18–34 y</td>
<td>34 (19)</td>
</tr>
<tr>
<td>35–49 y</td>
<td>64 (35)</td>
</tr>
<tr>
<td>50–64 y</td>
<td>50 (28)</td>
</tr>
<tr>
<td>&gt;64 y</td>
<td>19 (10)</td>
</tr>
<tr>
<td><strong>Health insurance (n = 117)</strong></td>
<td></td>
</tr>
<tr>
<td>Private</td>
<td>103 (88)</td>
</tr>
<tr>
<td>Medicaid</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Uninsured</td>
<td>7 (6)</td>
</tr>
<tr>
<td>Medicare</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Other</td>
<td>3 (3)</td>
</tr>
<tr>
<td><strong>Clinicians involved in error</strong></td>
<td></td>
</tr>
<tr>
<td>Specialty</td>
<td></td>
</tr>
<tr>
<td>Primary care†</td>
<td>88 (49)</td>
</tr>
<tr>
<td>Radiology</td>
<td>31 (17)</td>
</tr>
<tr>
<td>General surgery</td>
<td>23 (13)</td>
</tr>
<tr>
<td>Pathology</td>
<td>13 (7)</td>
</tr>
<tr>
<td>Physician’s assistant</td>
<td>13 (7)</td>
</tr>
<tr>
<td>Registered nurse or nurse practitioner</td>
<td>14 (8)</td>
</tr>
<tr>
<td>Trainee (resident, fellow, or intern)</td>
<td>20 (11)</td>
</tr>
<tr>
<td><strong>Setting</strong></td>
<td></td>
</tr>
<tr>
<td>Physician’s office</td>
<td>154 (85)</td>
</tr>
<tr>
<td>Ambulatory surgery facility</td>
<td>8 (4)</td>
</tr>
<tr>
<td>Pathology or clinical laboratory</td>
<td>8 (4)</td>
</tr>
<tr>
<td>Radiology suite</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Other</td>
<td>6 (3)</td>
</tr>
<tr>
<td><strong>Missed or delayed diagnosis</strong></td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>106 (59)</td>
</tr>
<tr>
<td>Breast</td>
<td>44 (24)</td>
</tr>
<tr>
<td>Colorectal</td>
<td>13 (7)</td>
</tr>
<tr>
<td>Skin</td>
<td>8 (4)</td>
</tr>
<tr>
<td>Hematologic</td>
<td>7 (4)</td>
</tr>
<tr>
<td>Gynecologic</td>
<td>7 (4)</td>
</tr>
<tr>
<td>Lung</td>
<td>6 (3)</td>
</tr>
<tr>
<td>Brain</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Prostate</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Liver or gastric</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Other</td>
<td>9 (5)</td>
</tr>
<tr>
<td>Infection</td>
<td>9 (5)</td>
</tr>
<tr>
<td>Fracture</td>
<td>8 (4)</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>7 (4)</td>
</tr>
<tr>
<td>Embolism</td>
<td>6 (3)</td>
</tr>
<tr>
<td>Appendicitis</td>
<td>5 (3)</td>
</tr>
<tr>
<td>Cerebral vascular disease</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Other neurologic condition</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Aneurysm</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Other cardiac condition</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Gynecologic disease (noncancer)</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Endocrine disorder</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Birth defect</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Ophthalmologic disease</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Peripheral vascular disease</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Abdominal disease</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Psychiatric illness</td>
<td>2 (1)</td>
</tr>
<tr>
<td>Other</td>
<td>8 (4)</td>
</tr>
<tr>
<td><strong>Adverse outcome</strong></td>
<td></td>
</tr>
<tr>
<td>Psychological or emotional</td>
<td>9 (5)</td>
</tr>
</tbody>
</table>

* Percentages do not sum to 100% because multiple physicians were involved in some errors.
† Includes 12 pediatricians, therefore pediatrics accounts for an absolute 7% of the 49%.
‡ Categories correspond to the following scores on the National Association of Insurance Commissioners’ 9-point severity scale: psychiatric or emotional (score 1), minor physical (scores 2 and 3), significant physical (scores 4 to 6), major physical (scores 7 and 8), and death (score 9). For details, see Appendix Figure 2, available at www.annals.org.
TABLE 2. Breakdown Points in the Diagnostic Process

<table>
<thead>
<tr>
<th>Process Breakdown</th>
<th>All Missed Diagnoses (n = 181), n (%)</th>
<th>Missed Cancer Diagnoses (n = 106), n (%)</th>
<th>Missed Noncancer Diagnoses (n = 75), n (%)</th>
<th>P Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial delay by the patient in seeking care</td>
<td>16 (9)</td>
<td>4 (4)</td>
<td>12 (16)</td>
<td>0.004</td>
</tr>
<tr>
<td>Failure to obtain adequate medical history or physical examination</td>
<td>76 (42)</td>
<td>24 (23)</td>
<td>52 (69)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Failure to order appropriate diagnostic or laboratory tests</td>
<td>100 (55)</td>
<td>63 (59)</td>
<td>37 (49)</td>
<td>0.178</td>
</tr>
<tr>
<td>Adequate diagnostic or laboratory tests ordered but not performed</td>
<td>17 (9)</td>
<td>10 (9)</td>
<td>7 (9)</td>
<td>0.98</td>
</tr>
<tr>
<td>Diagnostic or laboratory tests performed incorrectly</td>
<td>15 (8)</td>
<td>14 (13)</td>
<td>1 (1)</td>
<td>0.004</td>
</tr>
<tr>
<td>Incorrect interpretation of diagnostic or laboratory tests</td>
<td>67 (37)</td>
<td>49 (46)</td>
<td>18 (24)</td>
<td>0.002</td>
</tr>
<tr>
<td>Responsible provider did not receive diagnostic or laboratory test results</td>
<td>23 (13)</td>
<td>17 (16)</td>
<td>6 (8)</td>
<td>0.110</td>
</tr>
<tr>
<td>Diagnostic or laboratory test results were not transmitted to patient</td>
<td>22 (12)</td>
<td>15 (14)</td>
<td>7 (9)</td>
<td>0.33</td>
</tr>
<tr>
<td>Inappropriate or inadequate follow-up plan</td>
<td>81 (45)</td>
<td>51 (48)</td>
<td>30 (40)</td>
<td>0.28</td>
</tr>
<tr>
<td>Failure to refer</td>
<td>47 (26)</td>
<td>19 (18)</td>
<td>28 (37)</td>
<td>0.003</td>
</tr>
<tr>
<td>Failure of a requested referral to occur</td>
<td>9 (5)</td>
<td>8 (8)</td>
<td>1 (1)</td>
<td>0.058</td>
</tr>
<tr>
<td>Failure of the referred-to clinician to convey relevant results to the referring clinician</td>
<td>3 (2)</td>
<td>2 (2)</td>
<td>1 (1)</td>
<td>0.77</td>
</tr>
<tr>
<td>Patient nonadherence to the follow-up plan</td>
<td>31 (17)</td>
<td>21 (20)</td>
<td>10 (13)</td>
<td>0.25</td>
</tr>
</tbody>
</table>

* Pearson chi-square test for differences between cancer and noncancer categories.

81 [36%]), or did not document the plan correctly (22 of 81 [27%]).

**Contributing Factors**

The leading factors that contributed to errors were failures in judgment (143 of 181 [79%]), vigilance or memory (106 of 181 [59%]), knowledge (86 of 181 [48%]), patient-related factors (84 of 181 [46%]), and handoffs (36 of 181 [20%]) (Table 4). The patient-related factors included nonadherence (40 of 181 [22%]), atypical clinical presentation (28 of 181 [15%]), and complicated medical history (18 of 181 [10%]). There were no significant differences in the prevalence of the various contributing factors when missed cancer and missed noncancer diagnoses were compared, with the exception of lack of supervision, which was significantly more likely to have occurred in missed noncancer diagnoses (10 of 75 [13%] vs. 5 of 106 [5%]; P = 0.038). The higher prevalence of lack of supervision as a contributing factor to missed noncancer diagnoses was accompanied by greater involvement of trainees in these cases (13 of 75 [17%] vs. 7 of 106 [7%]; P = 0.023).

**Multifactorial Nature of Missed or Delayed Diagnoses**

The diagnostic errors were complex and frequently involved multiple process breakdowns, contributing factors, and contributing clinicians (Table 5). In 43% (78 of 181) of errors, 2 or more clinicians contributed to the missed diagnosis, and in 16% (29 of 181), 3 or more clinicians contributed. There was a median of 3 (interquartile range, 2 to 4) process breakdowns per error; 54% (97 of 181) of errors had 3 or more process breakdowns and 29% (52 of 181) had 4 or more. Thirty-five diagnostic errors (35 of 181 [19%]) involved a breakdown at only 1 point in the care process (single-point breakdowns). Twenty-four (24 of 35 [69%]) of these single-point breakdowns were either failures to order tests (n = 10) or incorrect interpretations of tests (n = 14); only 3 involved trainees.

The median number of contributing factors involved in diagnostic errors was 3 (interquartile range, 2 to 4); 59% (107 of 181) had 3 or more contributing factors, 27% (48 of 181) had 4 or more, and 13% (23 of 181) had 5 or more. Thus, although virtually all diagnostic errors were linked to cognitive factors, especially judgment errors, cognitive factors operated alone in a minority of cases. They were usually accompanied by communication factors, patient-related factors, or other system factors.

Specifically, 36% (66 of 181) of errors involved cognitive factors alone, 16% (29 of 181) involved judgment or vigilance and memory factors alone, and 9% (16 of 181) involved only judgment factors. The likelihood that cognitive factors alone led to the error was significantly lower among errors that involved an inadequate medical history or physical examination (21 of 66 [32%] vs. 55 of 115 [48%]; P = 0.036), lack of receipt of ordered tests by the responsible provider (3 of 66 [5%] vs. 20 of 115 [17%]; P = 0.013), and inappropriate or inadequate follow-up planning (21 of 66 [32%] vs. 60 of 115 [52%]; P = 0.008). In other words, the role of communication and other systems factors was especially prominent in these 3 types of breakdowns.

**DISCUSSION**

Our study of closed malpractice claims identified a group of missed diagnoses in the ambulatory setting that were associated with dire outcomes for patients. Over half of the missed diagnoses were cancer, primarily breast and colorectal cancer; no other diagnosis accounted for more
than 5% of the sample. The main breakdowns in the diagnostic process were failure to order appropriate diagnostic tests, inappropriate or inadequate follow-up planning, failure to obtain an adequate medical history or perform an adequate physical examination, and incorrect interpretation of diagnostic test results. Cognitive factors, patient-related factors, and handoffs were the most prominent contributing factors overall. However, relatively few diagnostic errors could be linked to single-point breakdowns or lone contributing factors. Most missed diagnoses involved manifold breakdowns and a potent combination of individual and system factors. The resultant delays tended to be long, setting diagnosis back more than 1 year on average.

The threshold question of what constitutes a harmful diagnostic error is extremely challenging. Indeed, the nebulous nature of missed diagnoses probably helps to explain why patient safety research in this area has lagged. Our approach was 2-dimensional. We considered the potential role of a range of contributing factors drawn from the fields of human factors and systems analysis (16, 19) in causing the patient’s injury. We also disaggregated the diagnostic process into discrete steps and then examined the prevalence of problems within particular steps. Trained physician-reviewers had the benefit of information from the claim file and from the medical record in deciding whether diagnoses were missed. Even so, agreement among reviewers was only fair, highlighting how difficult and subjective clinical judgments regarding missed diagnoses are.

### Table 3. Details of the 4 Most Frequent Process Breakdowns*

<table>
<thead>
<tr>
<th>Process Breakdown</th>
<th>Leading Reasons for Breakdown</th>
<th>Value, n</th>
<th>Within Process Breakdown Category, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure to order appropriate diagnostic or laboratory tests (n = 100)</td>
<td>Provider lacked knowledge of appropriate test</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Poor documentation</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Failure of communication</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Among providers</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Between provider and patient</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Patient did not schedule or keep appointment</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Patient declined tests</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Type of test not ordered</td>
<td>Imaging</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Computed tomography scanning</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Mammography</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Ultrasonography</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Blood</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Biopsy</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Colonoscopy or flexible sigmoidoscopy</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Inappropriate or inadequate follow-up plan (n = 81)</td>
<td>Provider did not think follow-up was necessary</td>
<td>32</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Provider selected inappropriate follow-up interval</td>
<td>29</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>Follow-up plan not documented</td>
<td>22</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Follow-up appointment not scheduled</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Miscommunication between patient and provider</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Miscommunication between providers</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Failure to obtain adequate medical history or physical examination (n = 76)</td>
<td>Incomplete physical examination</td>
<td>39</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Failure to elicit relevant information</td>
<td>39</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Poor documentation</td>
<td>20</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Patient provided inaccurate history</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>Incorrect interpretation of diagnostic or laboratory tests (n = 67)</td>
<td>Error in clinical judgment</td>
<td>52</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>Failure of communication among providers</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Inexperience</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Whose misinterpretation?</td>
<td>27</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Radiologist</td>
<td>19</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Primary care physician</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Type of test interpreted incorrectly</td>
<td>Imaging</td>
<td>36</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Mammography</td>
<td>18</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>Radiography</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Ultrasonography</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Blood</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Other‡</td>
<td>21</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Biopsy</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

* Category and subcategory totals may exceed 100% because of non–mutually exclusive reasons and multiple tests. Only leading reasons and tests are presented.
† The tests not shown in this category are echocardiography (n = 3), cardiac catheterization (n = 3), electrocardiography (n = 2), treadmill (n = 2), urinalysis (n = 1), esophagogastroduodenoscopy/barium swallow (n = 1), colposcopy (n = 1), urine pregnancy (n = 1), and joint aspirate (n = 1).
‡ The tests not shown in this category are electrocardiography (n = 4), urinalysis (n = 1), echocardiography (n = 1), stool guaiac (n = 2), Papanicolaou smear (n = 3), cystoscopy (n = 1), and treadmill (n = 1).
### Missed or Delayed Diagnoses

In general, our findings reinforce the need for systems interventions that mitigate the potential impact of cognitive errors by reducing reliance on memory, forcing consideration of alternative diagnostic plans or second opinions, and providing clinical decision support systems (21, 22). However, cognitive factors rarely appear alone, so interventions must also address other contributing factors, such as handoffs and communication. To be clinically use-
ful and maximally effective, such interventions must be easy to use and well integrated into clinicians’ workflows (25). For example, incorporating clinical decision support into the electronic medical record of a patient with breast symptoms should help guard against failures in history taking or test ordering by bolstering physicians’ knowledge and reducing their reliance on vigilance and memory.

An important feature of interventions is that their use should not rely on voluntary decisions to access them, because physicians are often unaware of their need for help (26). Rather, the use of an intervention should be automatically triggered by certain predetermined and explicit characteristics of the clinical encounter. For example, strategies to combat misinterpretation could mandate second reviews of test results in designated circumstances or require rapid expert reviews when physicians interpret test results outside of their areas of expertise.

After follow-up plans are made, improvements in scheduling procedures, tickler systems, and test result tracking systems could help keep patients and physicians on track. Nationwide attention is being given to the issue of follow-up (27, 28). Research and quality improvement efforts that equip physicians with tools to reliably perform follow-up must be a priority.

Selecting just 1 of the interventions we have outlined may not be sufficient. The multifactorial and complex nature of diagnostic errors suggests that meaningful reductions will require prevention strategies that target multiple levels in the diagnostic process and multiple contributing factors. Nevertheless, the resource constraints most health care institutions face demand priorities. Attention to the 3 vulnerable points we identified—ordering decisions, test interpretation, and follow-up planning—is a useful starting point and promises high yields, especially in reducing the number of missed cancer diagnoses.

Study Limitations

Our study has several limitations. First, unlike prospective observational studies or root-cause analyses, retrospective review of records, even the detailed records found in malpractice claim files, will miss certain breakdowns (for example, patient nonadherence) and contributing factors (for example, fatigue and workload), unless they emerged as issues during litigation. This measurement problem means that prevalence findings for such estimates will be lower bounds, and the multifactorial causality we observed probably understates the true complexity of diagnostic errors.

An additional measurement problem relates to the process breakdowns. Although reviewers considered breakdowns as independent events, in some situations breakdowns may have been prompted or influenced by earlier breakdowns in the diagnostic sequence. For example, a failure to order appropriate tests may stem from oversights in the physical examination. Such interdependence would tend to inflate the frequency of some breakdowns.

Second, awareness of the litigation outcome may have biased reviewers toward finding errors in claims that received compensation and vice versa (29, 30). Several factors militate against this bias: Reviewers were instructed to ignore the litigation outcome; physicians, who as a group tend to be skeptical of the malpractice system, may have been disinclined to credit the system’s findings; and, in fact, one quarter of error judgments diverged from the litigation outcomes.

Third, the reliability of the error determination was not high, and the CIs around the \( \kappa \) score are statistically compatible with poor or marginal reliability. Diagnostic errors are one of the most difficult types of errors to detect reliably (31). Although reviewers may have included some episodes of care in the study sample that were not “true” errors, and may have overlooked some that were, we know of no reason why the characteristics of such false-positive results and false-negative results would differ systematically from those we analyzed.

Fourth, malpractice claims data generally, and in our sample in particular, have several other biases. Severe injuries and younger patients are overrepresented in the subset of patients with medical injuries that trigger litigation (32, 33). It is possible that the factors that lead to errors in litigated cases may differ systematically from the factors that lead to errors in nonlitigated cases, although we know of no reason why they would. In addition, outpatient clinics associated with teaching hospitals are overrepresented in our sample, so the diagnostic errors we identified may not be generalizable outside this setting.

Conclusions

Our findings highlight the complexity of diagnostic errors in ambulatory care. Just as Reason’s “Swiss cheese” model of accident causation suggests (19) diagnostic errors that harm patients seem to result from the alignment of multiple breakdowns, which in turn stem from a confluence of contributing factors. The task of effecting meaningful improvements to the diagnostic process—with its numerous clinical steps, stretched across multiple providers and months or years, and the heavy reliance on patient initiative—looms as a formidable challenge. The prospects for “silver bullets” in this area seem remote.

Are meaningful gains achievable in the short to medium term? The answer will probably turn on whether simple interventions that target 2 or 3 critical breakdown points are sufficient to disrupt the causal chain or whether interventions at a wider range of points are necessary to avert harm. In this sense, our findings are humbling, and they underscore the need for continuing efforts to develop the “basic science” of error prevention in medicine (34), which remains in its infancy.

From Brigham and Women’s Hospital and Harvard School of Public Health, Boston, Massachusetts; University of Texas Health Science Center, Houston, Texas; and the Harvard Risk Management Foundation, Cambridge, Massachusetts.
Note: Drs. Gandhi, Kachalia, and Studdert had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Grant Support: This study was supported by grants from the Agency for Healthcare Research and Quality (HS011886-03) and the Harvard Risk Management Foundation. Dr. Studdert was also supported by the Agency for Healthcare Research and Quality (KO2HS11285).

Potential Financial Conflicts of Interest: Employment: T.A. Brennan (Aetna); Stock ownership or options (other than mutual funds): T.A. Brennan (Aetna); Expert testimony: T.A. Brennan.

Requests for Single Reprints: David M. Studdert, LLB, ScD, Department of Health Policy and Management, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115; e-mail, studdert@hsph.harvard.edu.

Current author addresses and author contributions are available at www.annals.org.

References


Current Author Addresses: Drs. Gandhi and Kachalia and Ms. Yoon: Division of General Internal Medicine, Brigham and Women’s Hospital, 75 Francis Street, Boston, MA 02114.
Dr. Thomas: University of Texas Medical School at Houston, 6431 Fannin, MSB 1.122, Houston, TX 77030.
Dr. Brennan: 151 Farmington Avenue, RC5A, Hartford, CT 06156.
Dr. Studdert: Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115.

Author Contributions: Conception and design: T.K. Gandhi, E.J. Thomas, T.A. Brennan, D.M. Studdert.
Critical revision of the article for important intellectual content: T.K. Gandhi, A. Kachalia, E.J. Thomas, T.A. Brennan, D.M. Studdert.
Provision of study materials or patients: A.L. Puopolo.
Statistical expertise: D.M. Studdert.
Obtaining of funding: D.M. Studdert.
Administrative, technical, or logistic support: A. Kachalia, A.L. Puopolo, D.M. Studdert.
WHAT ARE THE FACTORS THAT CONTRIBUTE TO DIAGNOSIS ERROR?

Mark Graber, MD et al.
Diagnostic Error in Internal Medicine.
Arch Intern Med. 2005 Jul 11; 165(13):1493-9
Diagnostic Error in Internal Medicine

Mark L. Graber, MD; Nancy Franklin, PhD; Ruthanna Gordon, PhD

Background: The goal of this study was to determine the relative contribution of system-related and cognitive components to diagnostic error and to develop a comprehensive working taxonomy.

Methods: One hundred cases of diagnostic error involving internists were identified through autopsy discrepancies, quality assurance activities, and voluntary reports. Each case was evaluated to identify system-related and cognitive factors underlying error using record reviews and, if possible, provider interviews.

Results: Ninety cases involved injury, including 33 deaths. The underlying contributions to error fell into 3 natural categories: “no fault,” system-related, and cognitive. Seven cases reflected no-fault errors alone. In the remaining 93 cases, we identified 548 different system-related or cognitive factors (5.9 per case). System-related factors contributed to the diagnostic error in 65% of the cases and cognitive factors in 74%. The most common system-related factors involved problems with policies and procedures, inefficient processes, teamwork, and communication. The most common cognitive problems involved faulty synthesis. Premature closure, ie, the failure to continue considering reasonable alternatives after an initial diagnosis was reached, was the single most common cause. Other common causes included faulty context generation, misjudging the salience of findings, faulty perception, and errors arising from the use of heuristics. Faulty or inadequate knowledge was uncommon.

Conclusions: Diagnostic error is commonly multifactorial in origin, typically involving both system-related and cognitive factors. The results identify the dominant problems that should be targeted for additional research and early reduction; they also further the development of a comprehensive taxonomy for classifying diagnostic errors.

Arch Intern Med. 2005;165:1493-1499

Once we realize that imperfect understanding is the human condition, there is no shame in being wrong, only in failing to correct our mistakes.

George Soros

In his classic studies of clinical reasoning, Elstein1 estimated the rate of diagnostic error to be approximately 15%, in reasonable agreement with the 10% to 15% error rate determined in autopsy studies.2-4 Considering the frequency and impact of diagnostic errors, one is struck by how little is known about this type of medical error.2 Data on the types and causes of errors encountered in the practice of internal medicine are scant, and the field lacks both a standardized definition of diagnostic error and a comprehensive taxonomy, although preliminary versions have been proposed.6-12

According to the Institute of Medicine, the most powerful way to reduce error in medicine is to focus on system-level improvements,13,14 but these interventions are typically discussed in regard to patient treatment issues. The possibility that system-level dysfunction could also contribute to diagnostic errors has received little attention. Typically, diagnostic error is viewed as a cognitive failing.7,15-17 Diagnosis reflects the clinician’s knowledge, clinical acumen, and problem-solving skills.1 In everyday practice, clinicians use expert skills to arrive at a diagnosis, often taking advantage of various mental shortcuts known as heuristics.18-21 These strategies are highly efficient, relatively effortless, and generally accurate, but they are not infallible.

The goal of this study was to clarify the basic etiology of diagnostic errors in internal medicine and to develop a working taxonomy. To understand how these errors arise and how they might be prevented in the future, we systematically examined the etiology of error using root cause analysis to classify both system-related and cognitive components.

METHODS

Based on a classification used by the Australian Patient Safety Foundation, we defined diagnostic error operationally as a diagnosis that was unintentionally delayed (sufficient information was available earlier), wrong (another diagnosis was made before the correct one), or missed (no diagnosis was ever made), as judged from the eventual appreciation of more definitive information.
Cases of suspected diagnostic error were collected from 5 large academic tertiary care medical centers over 5 years. To obtain a broad sampling of errors, we reviewed all eligible cases from 3 sources:

1. Performance improvement and risk management coordinators and peer review committees.
2. Voluntary reports from staff physicians and resident trainees.
3. Discrepancies between clinical impressions and autopsy findings.

Cases were included if internists (staff specialists or generalists or trainees) were primarily responsible for the diagnosis and if sufficient details about the case and the decision-making process could be obtained to allow analysis. In all cases, details were gathered from a review of the medical record, from fact-finding information obtained in the course of quality assurance activities when available, and in 42 cases, involved practitioners were interviewed, typically within 1 month of error identification. To minimize hindsight bias, reviews of medical records and interviews used a combination of open-ended queries and a root cause checklist developed by the Veterans Health Administration (VHA). The VHA instrument identifies specific flaws in the standard dimensions of organizational performance, and is well suited to exploring system-related factors. We developed the cognitive factors portion of the taxonomy by incorporating and expanding on the categories suggested by Chimonavitz et al., Kassirer and Kopelman, and Bordage. These categories differentiate flaws in the clinician’s knowledge and skills, ability to gather data, and ability to synthesize all available information into verifiable hypotheses. Criteria and definitions for each category were refined, and new categories added, as the study progressed.

Case histories were redacted of identifying information and analyzed as a team to the point of consensus by 1 internist and 2 cognitive psychologists to confirm the existence of a diagnostic error and to assign the error type (delayed, wrong, or missed) and both the system-related and the cognitive factors contributing to the error. To identify 100 usable cases, 129 cases of suspected error were reviewed, 29 of which were rejected. Definitive confirmation of an error was lacking in 19 cases. In 6 cases, the diagnosis was somewhat delayed but judged to have been made within an acceptable time frame. In 3 cases, the data were inadequate for analysis, and 1 case was rejected because the error reflected an intentional act that violated local policies (wrong diagnosis of hypotension from blood drawn above an intravenous line).

Impact was judged by an internist using a VHA scale that multiplies the likelihood of recurrence (1, remote; up to 4, frequent) by the severity of harm (1, minor injury; up to 4, catastrophic injury). A minor injury with a remote chance of recurrence received an impact score of 1, and a catastrophic event with frequent recurrence received an impact score of 16. Close-call errors were assigned an impact score of 0, and psychological impact was similarly discounted. The relative frequency of both system-related and cognitive factors varied with the source of the case and the type of error involved. Cases identified from quality assurance reports and from voluntary reports had a similar prevalence of system-related factors (72% and 76%, respectively) and cognitive factors (65% and 85%, respectively). In contrast, cases identified from autopsy discrepancies involved cognitive factors 90% of the time (P > .50) and system-related factors only 10% of the time (P < .001). Cases of delayed diagnosis had relatively more

Impact

Ten cases were classified as close calls, and 90 cases involved some degree of harm, including 33 deaths. The clinical impact averaged 3.80 ± 0.28 (mean ± SEM) on the VHA impact scale, indicating substantial levels of harm, on average. The impact score tended to be lower in cases of delayed diagnosis than in cases that were missed or wrong (3.79 ± 0.52 vs 4.76 ± 0.40 and 4.47 ± 0.46; P = .34). Cases with solely cognitive factors or with mixed cognitive and system-related factors had significantly higher impact scores than cases with only system-related factors (4.11 ± 0.46 and 4.27 ± 0.47 vs 2.54 ± 0.55; P = .03 for both comparisons). These 2 effects may be related, as delays were the type of error most likely to result from system-related factors alone.

ETIOLOGY OF DIAGNOSTIC ERROR

Our results suggested that diagnostic error in medicine could best be described using a taxonomy that includes no-fault, system-related (Table 1), and cognitive (Table 2) factors.

<table>
<thead>
<tr>
<th>No-fault errors</th>
<th>System-related errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faulty synthesis</td>
<td>Technical failure and equipment problems</td>
</tr>
<tr>
<td>Masked or unusual presentation of disease</td>
<td>Organizational flaws</td>
</tr>
<tr>
<td>Patient-related error (uncooperative, deceptive)</td>
<td>Cognitive errors</td>
</tr>
<tr>
<td>System-related errors</td>
<td>Faulty knowledge</td>
</tr>
<tr>
<td>Technical failure and equipment problems</td>
<td>Faulty data gathering</td>
</tr>
<tr>
<td>System-related errors</td>
<td>Faulty synthesis</td>
</tr>
</tbody>
</table>

In 46% of the cases, both system-related and cognitive factors contributed to diagnostic error. Cases involving only cognitive factors (28%) or only system-related factors (19%) were less common, and 7 cases were found to reflect solely no-fault factors, without any other system-related or cognitive factors. Combining the pure and the mixed cases, system-related factors contributed to the diagnostic error in 65% of the 100 cases and cognitive factors contributed in 74% (Figure). Overall, we identified 228 system-related factors and 320 cognitive factors, averaging 5.9 per case.

The relative frequency of both system-related and cognitive factors varied with the source of the case and the type of error involved. Cases identified from quality assurance reports and from voluntary reports had a similar prevalence of system-related factors (72% and 76%, respectively) and cognitive factors (65% and 85%, respectively). In contrast, cases identified from autopsy discrepancies involved cognitive factors 90% of the time (P > .50) and system-related factors only 10% of the time (P < .001). Cases of delayed diagnosis had relatively more
system-related errors (89%) and fewer cognitive errors (36%) on average, and cases of wrong diagnosis involved more cognitive errors (92%) and fewer system-related errors (50%, \( P < .01 \) for both pairs).

**NO-FAULT ERRORS**

No-fault factors were identified in 44 of the 100 cases and constituted the sole explanation in 7 cases. Eleven of these cases involved patient-related factors, including 2 instances of deception (surreptitious self-injection of saliva, mimicking sepsis and denial of high-risk sexual activity, which delayed diagnosis of *Pneumocystis carinii* pneumonia) and 9 cases involving delayed diagnoses related to missed appointments or instances in which patient statements were unintentionally misleading or incomplete. By far, the most common no-fault factor was an atypical or masked disease presentation, encountered in 33 cases.

**SYSTEM-RELATED CONTRIBUTIONS TO ERROR**

In 65 cases, system-related factors contributed to diagnostic error (Table 1). The vast majority of these (215 instances) were related to organizational problems, and a small fraction (13 instances) involved technical and equipment problems. The factors encountered most often related to policies and procedures, inefficient processes, and difficulty with teamwork and communication, especially communication of test results. Many error types were encountered more than twice in the same institution, an event we referred to as clustering.

**COGNITIVE CONTRIBUTIONS TO ERROR**

We identified 320 cognitive factors in 74 cases (Table 2). The most common category of factors was faulty synthesis (264 instances), or flawed processing of the available information. Faulty data gathering was identified in 45 instances. Inadequate or faulty knowledge or skills were identified in only 11 instances.

**FAULTY KNOWLEDGE OR SKILLS**

Inadequate knowledge was identified in only 4 cases, each concerning a rare condition: (1) a case of missed Fournier gangrene; (2) a missed diagnosis of calciphylaxis in a patient undergoing dialysis with normal levels of serum calcium and phosphorus; (3) a case of chronic thrombotic thrombocytopenic purpura; and (4) a wrong diagnosis of disseminated intravascular coagulation in a patient ultimately thought to have...
The 7 cases involving inadequate skills involved misinterpretations of x-ray studies and electrocardiograms by nonexperts.

### Table 2. Cognitive Contributions to Diagnostic Error

<table>
<thead>
<tr>
<th>No. of Encounters</th>
<th>Type</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Knowledge base inadequate or deficient</td>
<td>Insufficient knowledge of relevant condition</td>
<td>Providers not aware of Fournier gangrene</td>
</tr>
<tr>
<td>7</td>
<td>Skills inadequate or deficient</td>
<td>Insufficient diagnostic skill for relevant condition</td>
<td>Missed diagnosis of complete heart block: clinician misread electrocardiogram</td>
</tr>
<tr>
<td>24</td>
<td>Ineffective, incomplete, or faulty workup</td>
<td>Problems in organizing or coordinating patient's tests and consultations</td>
<td>Delayed diagnosis of drug-related lupus: failure to consult patient's old medical records</td>
</tr>
<tr>
<td>10</td>
<td>Ineffective, incomplete, or faulty history and physical examination</td>
<td>Failure to collect appropriate information from the initial interview and examination</td>
<td>Delayed diagnosis of abdominal aortic aneurysm: incomplete history questioning</td>
</tr>
<tr>
<td>7</td>
<td>Faulty test or procedure techniques</td>
<td>Standard test/procedure is conducted incorrectly</td>
<td>Wrong diagnosis of myocardial infarction: electrocardiographic leads reversed</td>
</tr>
<tr>
<td>3</td>
<td>Failure to screen (prehypothesis)</td>
<td>Failure to perform indicated screening procedures</td>
<td>Missed prostate cancer: rectal examination and PSA testing never performed in a 55-year-old man</td>
</tr>
<tr>
<td>1</td>
<td>Poor etiquette leading to poor data quality</td>
<td>Failure to collect required information owing to poor interaction with patient</td>
<td>Missed CNS contusion after very abbreviated history; perjorative questioning</td>
</tr>
<tr>
<td>26</td>
<td>Faulty context generation</td>
<td>Lack of awareness/consideration of aspects of patient's situation that are relevant to diagnosis</td>
<td>Missed perforated ulcer in a patient presenting with chest pain and laboratory evidence of myocardial infarction</td>
</tr>
<tr>
<td>25</td>
<td>Overestimating or underestimating usefulness or salience of a finding</td>
<td>Clinician is aware of symptom but either focuses too closely on it to the exclusion of others or fails to appreciate its relevance</td>
<td>Wrong diagnosis of sepsis in a patient with stable leukocytosis in the setting of myelodysplastic syndrome</td>
</tr>
<tr>
<td>25</td>
<td>Faulty detection or perception</td>
<td>Symptom, sign, or finding should be noticeable, but clinician misses it</td>
<td>Missed pneumothorax on chest radiograph</td>
</tr>
<tr>
<td>23</td>
<td>Failed heuristics</td>
<td>Failure to apply appropriate rule of thumb, or overapplication of such a rule under inappropriate/typical circumstances</td>
<td>Wrong diagnosis of bronchitis in a patient later found to have pulmonary embolism</td>
</tr>
<tr>
<td>15</td>
<td>Failure to act sooner</td>
<td>Delay in appropriate data-analysis activity</td>
<td>Missed diagnosis of ischemic bowel in a patient with a 12-week history of bloody diarrhea</td>
</tr>
<tr>
<td>14</td>
<td>Faulty triggering</td>
<td>Clinician considers inappropriate conclusion based on current data or fails to consider conclusion reasonable from data</td>
<td>Wrong diagnosis of pneumonia in a patient with hemiplegia: never considered the eventual diagnosis of vasculitis</td>
</tr>
<tr>
<td>11</td>
<td>Misidentification of a symptom or sign</td>
<td>One symptom is mistaken for another</td>
<td>Missed cancer of the pancreas in a patient with pain radiating to the back, attributed to GERD</td>
</tr>
<tr>
<td>10</td>
<td>Distraction by other goals or issues</td>
<td>Other aspects of patient treatment (eg, dealing with an earlier condition) are allowed to obscure diagnostic process for current condition</td>
<td>Wrong diagnosis of panic disorder: patient with a history of schizophrenia presenting with abnormal mental status and found to have CNS metastases</td>
</tr>
<tr>
<td>10</td>
<td>Faulty interpretation of a test result</td>
<td>Test results are read correctly, but incorrect conclusions are drawn</td>
<td>Missed diagnosis of <em>Clostridium difficile</em> enteritis in a patient with a negative stool test result</td>
</tr>
<tr>
<td>0</td>
<td>Reporting or remembering findings not gathered</td>
<td>Symptoms or signs reported that do not exist, often findings that are typically present in the suspected illness</td>
<td>None encountered</td>
</tr>
<tr>
<td>39</td>
<td>Premature closure</td>
<td>Failure to consider other possibilities once an initial diagnosis has been reached</td>
<td>Wrong diagnosis of musculoskeletal pain after a car crash: ruptured spleen ultimately found</td>
</tr>
<tr>
<td>18</td>
<td>Failure to order or follow up on appropriate test</td>
<td>Clinician does not use an appropriate test to confirm a diagnosis or does not take appropriate next step after test</td>
<td>Wrong diagnosis of urosepsis in a patient: bedside urinalysis never performed</td>
</tr>
</tbody>
</table>

### Faulty Data Gathering

The dominant cause of error in the faulty data-gathering category lay in the subcategory of ineffective,
incomplete, or faulty workup (24 instances). For example, the diagnosis of subdural hematoma was missed in a patient who was seen after a motor vehicle crash because the physical examination was incomplete. Problems with ordering the appropriate tests and interpreting test results were also common in this group.

**FAULTY INFORMATION SYNTHESIS**

Faulty information synthesis, which includes a wide range of factors, was the most common cause of cognitive-based errors. The single most common phenomenon was premature closure: the tendency to stop considering other possibilities after reaching a diagnosis. Other common synthesis factors included faulty context generation, misjudging the salience of a finding, faulty perception, and failed use of heuristics. Faulty context generation and misjudging the salience of a finding often occurred in the same case (15 of 25 instances). Perceptual failures most commonly involved incorrect readings of x-ray studies by internists and emergency department staff before official reading by a radiologist. Of the 23 instances related to heuristics, 14 reflected the bias to assume that all findings were related to a single cause when a patient actually had more than 1 condition. In 7 cases, the most common condition was chosen as the likely diagnosis, although a less common condition was responsible.

**COVARIATION AMONG FACTORS**

Cognitive and system-related factors were found to often co-occur, and these factors may have led, directly or indirectly, to each other. For example, a mistake relatively early on (eg, an inadequate history or physical examination) is likely to lead to subsequent mistakes (eg, in interpreting test results, considering appropriate candidate diagnoses, or calling in appropriate specialists). We examined the patterns of factors identified in these 100 cases to identify clusters of cognitive factors that tended to co-occur. Using Pearson r tests and correcting for the use of multiple pairwise analyses, we found several such clusters of cognitive factors. The more common clusters of 3 factors, all of which have significant pairwise correlations within a cluster, were as follows:

- Incomplete/faulty history and physical examination; failure to consider the correct candidate diagnosis; and premature closure
- Incomplete/excessive data gathering; bias toward a single explanation; and premature closure
- Underestimating the usefulness of a finding; premature closure; and failure to consult

<table>
<thead>
<tr>
<th>No. of Encounters</th>
<th>Type</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Failure to consult</td>
<td>Appropriate expert is not contacted</td>
<td>Hyponatremia inappropriately ascribed to diuretics in a patient later found to have lung cancer; no consultations requested</td>
</tr>
<tr>
<td>10</td>
<td>Failure to periodically review the situation</td>
<td>Failure to gather new data to determine whether situation has changed since initial diagnosis</td>
<td>Missed colon cancer in a patient with progressively declining hematocrit attributed to gastritis</td>
</tr>
<tr>
<td>10</td>
<td>Failure to gather other useful information to verify diagnosis</td>
<td>Appropriate steps to verify diagnosis are not taken</td>
<td>Wrong diagnosis of osteoarthritis in a patient found to have drug-induced lupus after ANA testing</td>
</tr>
<tr>
<td>8</td>
<td>Overreliance on someone else’s findings or opinion</td>
<td>Failure to check previous clinician’s diagnosis against current findings</td>
<td>Outpatient followed up with diagnosis of CHF, admitted with increased shortness of breath, later found to have lung cancer as the cause</td>
</tr>
<tr>
<td>5</td>
<td>Failure to validate findings with patient</td>
<td>Clinician does not check with patient concerning additional symptoms that might confirm/disconfirm diagnosis</td>
<td>Wrong diagnosis of bone metastases in a patient with many prior broken ribs</td>
</tr>
<tr>
<td>1</td>
<td>Confirmation bias</td>
<td>Tendency to interpret new results in a way that supports one’s previous diagnosis</td>
<td>Wrong diagnosis of pulmonary embolism: positive results on D-dimer test taken to support this diagnosis in a patient with respiratory failure due to ARDS and gram-negative sepsis</td>
</tr>
</tbody>
</table>

Abbreviations: ANA, antinuclear antibody; ARDS, adult respiratory distress syndrome; CHF, congestive heart failure; CNS, central nervous system; GERD, gastroesophageal reflux disease; PSA, prostate-specific antigen.
In classifying the underlying factors contributing to error, 3 natural categories emerged: no fault, system-related, and cognitive. This classification validates the cognitive and no-fault distinctions described by Chimowitz et al., Kassirer and Kopelman, and Bordage and adds a third major category of system-level factors.

A second objective was to assess the relative contributions of system-related and cognitive root cause factors. The results allow 3 major conclusions regarding diagnostic error in internal medicine settings.

**DIAGNOSTIC ERROR IS TYPICALLY MULTIFACTORIAL IN ORIGIN**

Excluding the 7 cases of pure no-fault error, we identified an average of 5.9 factors contributing to error in each case. Reason’s “Swiss cheese” model of error suggests that harm results from multiple breakdowns in the series of barriers that normally prevent injury. This phenomenon was identified in many of our cases, in which the ultimate diagnostic failure involved separate factors at multiple levels of both the system-related and the cognitive pathways.

A second reason for encountering multiple factors in a single case is the tendency for one type of error to lead to another. For example, a patient with retrosternal and upper epigastric pain was given a diagnosis of myocardial infarction on the basis of new Q waves in his electrocardiogram and elevated levels of troponin. The clinicians missed a coexisting perforated ulcer, illustrating that if a case is viewed in the wrong context, clinicians may miss relevant clues and may not consider the correct diagnosis.

**SYSTEM FLAWS CONTRIBUTE COMMONLY TO DIAGNOSTIC ERROR**

System-related factors were identified in 65% of cases. This finding supports a previous study linking diagnostic errors to system issues but contrasts with the prevailing belief that diagnostic errors overwhelmingly reflect defective cognition. The system flaws identified in our study reflected far more organizational issues than technical problems. Errors related to suboptimal supervision of trainees occurred, but uncommonly.

**COGNITIVE ERRORS ARE A COMMON CAUSE OF DIAGNOSTIC ERROR AND Predominantly REFLECT PROBLEMS WITH SYNTHESIS OF THE AVAILABLE INFORMATION**

Faulty data gathering was much less commonly encountered, and defective knowledge was rare. These results are consistent with conclusions from earlier studies and from autopsy data almost 50 years ago: “. . . mistakes were due not so much to lack of knowledge of factual data as to certain deficiencies of approach and judgment.” This finding may distinguish medical diagnosis from other types of expert decision making, in which knowledge deficits are more commonly encountered as the cause of error.

As predicted by other authors, premature closure was encountered more commonly than any other type of cognitive error. Simon described the initial stages of problem solving as a search for an explanation that best fits the known facts, at which point one stops searching for additional explanations, a process he termed satisfying. Experienced clinicians are as likely as more junior colleagues to exhibit premature closure, and elderly physicians may be particularly predisposed.

**STUDY LIMITATIONS**

This study has a variety of limitations that restrict the generality of the conclusions. First, because the types of error are dependent on the source of the cases, a different spectrum of case types would be expected outside internal medicine. Also, selection bias might be expected in cases that are reported voluntarily. Distortions could also result from our nonrandomized method of case selection if they are not representative of the errors that actually occurred.

A second limitation is the difficulty in discerning exactly how a given diagnosis was reached. Clinical reasoning is hidden from direct examination, and may be just as mysterious to the clinician involved. A related problem is our limited ability to identify other factors that likely affect many clinical decision-making situations, such as stress, fatigue, and distractions. Clinicians had difficulty recalling such factors, which undoubtedly existed. Their recollections might also be distorted because of the unavoidable lag time between the experience and the interview, and by their knowledge of the clinical outcomes.

A third weakness is the subjective assignment of root causes. The field as it evolves will benefit from further clarification and standardization for each of these causes. A final concern is the inevitable bias that is introduced in a retrospective analysis in which the outcomes are known. With this in mind, we did not attempt to evaluate the appropriateness of care or the preventability of adverse events, judgments that are highly sensitive to hindsight bias.

**STRATEGIES TO DECREASE DIAGNOSTIC ERROR**

Although diagnostic error can never be eliminated, our results identify the common causes of diagnostic error in medicine, ideal targets for future efforts to reduce the incidence of these errors. The high prevalence of system-related factors offers the opportunity to reduce diagnostic errors if health care institutions accept the responsibility of addressing these factors. For example, errors could be avoided if radiologists were reliably available to interpret x-ray studies and if abnormal test results were reliably communicated. Institutions should be especially sensitive to clusters of errors of the same type. Although these institutions may simply excel at error detection, clustering could also indicate misdirected resources or a culture of tolerating suboptimal performance.
Devising strategies for reducing cognitive error is a more complex problem. Our study suggests that internists generally have sufficient medical knowledge and that errors of clinical reasoning overwhelmingly reflect inappropriate cognitive processing and/or poor skills in monitoring one’s own cognitive processes (metacognition). Croskerry and others have argued that clinicians who are oriented to the common pitfalls of clinical reasoning would be better able to avoid them. High-fidelity simulations may be one way to provide this training. Elstein has suggested the value of compiling a complete differential diagnosis to combat the tendency to premature closure, the most common cognitive factor we identified. A complementary strategy for considering alternatives involves the technique of prospective hindsight: the crystal ball experience: The clinician would be told to assume that his or her working diagnosis is incorrect, and asked, “What alternatives should be considered?” A final strategy is to augment a clinician’s inherent metacognitive skills by using expert systems, an approach currently under active research and development.

Accepted for Publication: February 21, 2005.
Correspondence: Mark L. Graber, MD, Medical Service 111, Veterans Affairs Medical Center, Northport, NY 11768 (mark.graber@med.va.gov).

Funding/Support: This work was supported by a research support grant honoring James S. Todd, MD, from the National Patient Safety Foundation, North Adams, Mass.

Acknowledgment: We thank Grace Carey and Kathy Kessel for their assistance with the manuscript and references.

WHAT ARE THE FACTORS THAT CONTRIBUTE TO DIAGNOSIS ERROR?

Mark Graber MD.

Diagnostic errors in medicine: a case of neglect, Jt Comm J Qual Patient Saf.
2005 Feb;31(2):106-13
Diagnostic Errors in Medicine: A Case of Neglect

Mark Graber, M.D.

More than 90 years have passed since Harvard’s Dr. Richard Cabot illustrated the value of autopsies to reveal diagnostic error.1 At his weekly meetings with medical students, Cabot would compare the clinical impressions before death with the findings at autopsy, creating the powerful and enduring teaching format known as the “CPC” (clinical-pathological correlation). From his experience with more than 3,000 cases, Cabot estimated that the clinical diagnosis was correct less than half the time in patients dying of thoracic aneurysms, cirrhosis, pericarditis, nephritis, and a variety of other conditions. The reaction at the time was one of incredulity.2 Dr. Alfred Croftan, a Chicago physician, responded that “the overwhelming majority of cases are today diagnosed correctly…” and that the errors noted by Dr. Cabot could only reflect the “unpardonably careless work” by the physicians of Boston!3(p. 145)

Modern clinicians still struggle to recognize and accept the possibility of diagnostic error. Although many of the diseases on Dr. Cabot’s list can now be diagnosed with routine blood tests or imaging, diagnostic error still exists and always will.4 Medical diagnoses that are wrong, missed, or delayed make up a large fraction of all medical errors and cause substantial suffering and injury. Compared with other types of medical error, however, diagnostic errors receive little attention—a major factor in perpetuating unacceptable rates of diagnostic error. Diagnostic errors are fundamentally obscure, health care organizations have not viewed them as a system problem, and physicians responsible for making medical decisions seldom perceive their own error rates as problematic. The safety of modern health care can be improved if these three issues are understood and addressed.

Background: Medical diagnoses that are wrong, missed, or delayed make up a large fraction of all medical errors and cause substantial suffering and injury. Compared with other types of medical error, however, diagnostic errors receive little attention—a major factor in perpetuating unacceptable rates of diagnostic error. Diagnostic errors are fundamentally obscure, health care organizations have not viewed them as a system problem, and physicians responsible for making medical decisions seldom perceive their own error rates as problematic. The safety of modern health care can be improved if these three issues are understood and addressed.

Solutions: Opportunities to improve the visibility of diagnostic errors are evident. Diagnostic error needs to be included in the normal spectrum of quality assurance surveillance and review. The system properties that contribute to diagnostic errors need to be systematically identified and addressed, including issues related to reliable diagnostic testing processes. Even for cases entirely dependent on the skill of the clinician for accurate diagnosis, health care organizations could minimize errors by using system-level interventions to aid the clinician, such as second readings of key diagnostic tests and providing resources for clinical decision support. Physicians need to improve their calibration by getting feedback on the diagnoses they make. Finally, clinicians need to learn about overconfidence and other innate cognitive tendencies that detract from optimal reasoning and learning.

Conclusion: Clinicians and their health care organizations need to take active steps to discover, analyze, and prevent diagnostic errors.
Neglected Errors

Although clinicians routinely arrive at most of their diagnoses with ease and accuracy, diagnostic errors occur at non-negligible rates. Diagnostic errors are encountered in every specialty and every type of practice. It is estimated that mistakes are made in roughly 5% of radiology and pathology diagnoses. Error rates in clinical practice are not known with certainty but might be as high as 12%, for example in the emergency department where complex decision making is involved in settings of above-average uncertainty and stress. Diagnostic errors are easily demonstrated when physicians are tested with standardized patients or scenarios. There is excess variation between providers who analyze the same case, and physicians at times even disagree with themselves when re-presented with a case they have previously diagnosed.

Medical residents report that diagnostic errors are the most common medical errors they encounter, and only a minority of these are ever discussed with their mentors. In the Harvard Medical Practice Study, diagnostic errors were the second leading cause of adverse events. Autopsies are considered the gold standard for detecting diagnostic errors, and, indeed, discrepancies from the clinical diagnosis are found in approximately one quarter of cases. Not all these discrepancies have clinical relevance, but errors that potentially could have changed the outcome are found in 5% to 10% of all autopsies.

Despite this evidence that diagnostic errors are a major issue in patient safety, they receive scant attention. This may reflect the general tendency for clinicians to be unaware of medical error, but even within the patient safety community diagnostic errors are barely on the radar screen. Clinical decision making is an underfunded research area, and the topic is generally neglected in patient safety symposia and in medical school curricula. If we are ever to optimize patient safety, it will require that we minimize diagnostic errors. As a start, we need to understand why these errors are neglected.

Fundamental Issues

Although clinical decision making has been described in general terms, at a fundamental level the process is complicated, hidden, and only partially understood. Making a diagnosis is primarily a cognitive process and is subject to influence by the affective state of the clinician. These processes are difficult to study and quantify and to understand. Even the decision maker may not be aware of how or why a given diagnosis was reached. Experts in this field work in the area of cognitive psychology and discuss specialized concepts such as “metacognition,” “heuristics,” and “debiasing.” It is little wonder that practicing clinicians and patient safety staff have difficulty acquiring a comprehensive understanding of clinical decision making, the first step in trying to understand diagnostic errors. Root cause analysis, so powerful in understanding other types of medical error, is less easily applied when the root causes are cognitive.

Diagnostic errors often escape detection. It has been estimated that for every error we detect, scores are missed. Compared with adverse events related to surgery or other “process” and treatment errors, which are often glaring and obvious, diagnostic errors tend to be more subtle and difficult to pinpoint in time and place. Detecting unacceptable delays in diagnosis is especially difficult, because clear standards are lacking, and the perceptions of the clinician regarding the existence of a delay may differ substantially from the impressions of an affected patient.

Physicians are uncomfortable discussing diagnostic error. It is unsettling to consider the possibility that one’s own diagnostic capabilities might be in question. This reflects, in part, the real concern that diagnostic errors can lead to career-threatening malpractice suits. A substantial proportion of these suits are provoked by diagnostic errors. In the Veterans Health Administration, tort claims related to diagnostic errors were twice as common as claims related to medication errors. Malpractice suits related to diagnostic errors involve all specialties and are the most difficult to defend.
System Issues

Health care organizations view diagnostic error as physician failure rather than as an institutional problem. The available data, however, suggests that system-related factors commonly contribute to diagnostic error. Emergency departments are particularly predisposed to system-related diagnostic errors relating to workload-related stress, constant distractions, nonstandardized processes, and a variety of associated issues. Similarly, when diagnostic errors involve internists, both cognitive and latent system flaws are typically identified as root causes. The system-level processes most commonly found in these cases involve cultural acceptance of suboptimal systems, inadequate policies and procedures, inefficient processes, and problems with teamwork and communication.

Diagnostic errors are simply not a priority for health care organizations. No one has told them they should be. Patient safety is an enormous universe, and many organizations look to expert advisory groups for advice on which areas need attention. Unfortunately, the guidance provided by advisory organizations has yet to identify diagnostic error as an area in need of attention. The Joint Commission on Accreditation of Healthcare Organizations encourages facilities to develop their own patient safety priorities using the tools of root cause analysis to investigate serious injuries. In practice, however, organizations tend to focus on the initial seven National Patient Safety Goals, none of which directly involves diagnostic error (but see page 110). The Joint Commission also requires organizations to study and report on 13 categories of sentinel event, but none of these address diagnostic accuracy, nor do the 27 reportable events advocated by the National Quality Forum. Similarly, the Leapfrog Group, a consortium of 145 organizations providing healthcare benefits, has issued patient safety priorities for healthcare organizations, none directly involving medical diagnosis. The Agency for Healthcare Research and Quality recently compiled evidence-based patient safety practices. Of the 25 measures with the strongest evidence base, none are measures to improve diagnostic accuracy or timeliness.

Health care organizations have many incentives to bypass thorny issues, such as diagnostic error, and focus instead on “low-hanging fruit,” such as medication errors and wrong-site surgery. These errors are easier to understand, interventions are more obvious, and the organization can show progress towards achieving patient safety goals.

Provider Issues

Although the act of making a medical diagnosis is a defining skill of physicians, clinicians seldom think about how well they carry out this function. Two questions immediately arise: (1) Are clinicians aware that diagnostic errors can occur? (2) Are they themselves susceptible to this problem? The answer to the first question is unequivocally yes: Physicians are keenly aware that diagnostic errors are made. Throughout their training and on a regular basis in their practice years, physicians learn of diagnostic errors via quality assurance proceedings and malpractice suits. The fear of a malpractice suit relating to diagnostic error is a major impetus for the practice of defensive medicine—excessive referrals to specialists and the use of expensive and redundant tests and procedures. Defensive practice is ubiquitous, and some estimate that it accounts for 5% to 9% of health care expenditures in the United States. Of direct relevance, this malpractice pressure likely has more significant impact on diagnostic decisions than management decisions.

The answer to the second question is more interesting: Although clinicians are aware of the possibility for diagnostic error and that they are personally at risk, they believe that their own diagnoses are correct. Errors are made by someone else. Clinicians can perhaps remember a diagnostic error or two they’ve made in their career, but when asked about their current patients and their current diagnoses, my personal observation is that most believe, like Dr. Croftan, that virtually all of these are accurate. Indeed only 29% of physicians reported encountering any medical error in the past year. Given the evidence that the true error rate is in the range of 5% to 15%, how is it possible that individual physicians believe that their own error rates approach zero? Inadequate calibration and overconfidence may help explain this contradiction.

Inadequate calibration. Across a wide spectrum of skills and professions, receiving feedback on one’s performance is an essential requirement for developing expertise. Feedback provides the follow-up information necessary to assess the accuracy of initial predications, a process referred to as calibration. Professionals who
are well calibrated agree with each other and correctly interpret “gold standard” cases.

Physicians occasionally receive feedback regarding some of their diagnoses as patients move through a diagnostic evaluation or return for follow-up after treatment. Feedback is too sporadic, however, and lack of feedback contributes to the physician’s sincere belief that the great majority (or all) of their diagnoses are correct. A variety of factors limit feedback. Some of the more important factors follow:

- Availability of definitive diagnostic information. The true diagnosis is not known in every case. This would include, for example, most diagnoses regarding less serious ailments; if a patient’s knee pain resolves after two weeks of treatment with aspirin, the clinician may never know if the correct diagnosis was arthritis, bursitis, or tendonitis.

- Communication barriers. If definitive diagnostic information does become available at some later time, the clinician making the initial diagnosis may no longer be in the information loop. The clinician in an outpatient setting may not be routinely informed if his or her initial diagnosis is later modified in the hospital, or vice versa. Even if the opportunity does arise to convey knowledge of an error, clinicians feel awkward informing colleagues their diagnoses were wrong.

- Paucity of planned feedback. The autopsy rate in the United States is now estimated to be less than 5%, depriving clinicians of a unique calibration tool. Although some medical specialty organizations provide formal programs that provide and encourage feedback, this is more the exception than the rule. Physicians are rarely expected to show evidence that they have tracked or processed feedback in any organized way.

**Overconfidence.** Lack of calibration contributes to overconfidence, a second factor causing clinicians to overestimate their diagnostic accuracy. There is substantial evidence that humans are overconfident in a variety of settings; we routinely overestimate our ability to function flawlessly. A classic example is eyewitness testimony—the correlation between the accuracy of eyewitness identification and the confidence level of the witness is generally less than 0.25. Another sobering statistic is that only 1% of drivers rate their driving skills below those of the average driver.

Just as we overestimate our skills and knowledge, we are overconfident that our decisions, such as medical diagnoses, are correct. Overconfidence in the setting of medical decision making, however, is particularly inappropriate given the general uncertainty that surrounds most diagnostic endeavors.

A chilling observation in studies of overconfidence is that the least skilled are the most overconfident. In tests of grammar, logic, and humor, the subjects scoring the lowest by objective criteria are substantially more overconfident of their performance than subjects who score well on these tests. Exactly the same phenomenon is seen when medical residents estimate their skill in communicating with patients. Those residents least skilled in communicating were exactly the ones least able to judge their skill level and the most likely to overestimate it. In the same vein, trainees dismissed from residency training programs believed that they rarely made any mistakes. This phenomenon fits well with current theories regarding the development of expertise. Experts not only possess superior knowledge of their field but are highly aware of their performance level and are less likely to exhibit overconfidence in regard to their performance. Superior metacognition and accurate calibration are the hallmarks of an expert.

Overconfidence and poor calibration become self-perpetuating if our (inappropriate) certainty about our diagnoses dissuades us from asking for autopsies. Is our diagnostic acumen sharp enough to predict that an autopsy will simply confirm our suspicions? The available evidence would suggest otherwise. For example, in a recent study, clinicians were asked about the certainty of the diagnosis; fatal but potentially treatable errors were found at autopsy in 10% of the cases regardless of whether the clinicians were certain or not.

Physicians are uncomfortable with uncertainty. Partly in response to pressures from our patients and partly to satisfy our own unease, we assign a “working” diagnosis. Typically, this is produced by a “cognitive disposition to respond” without always ensuring that the process is rigorous and the product is accurate. Physicians are expected to present a professional air of confidence and expertise. Diagnostic errors fall in our blind spot—we ignore them.
Solutions

Opportunities to improve the visibility of diagnostic errors are evident.

Fundamental Issues

Diagnostic error needs to be included in the normal spectrum of quality assurance surveillance and review. Staff responsible for organizational performance needs simple definitions and working rules to identify, classify, and study these errors. A simple working definition of diagnostic error is those diagnoses that are missed, wrong, or delayed, as detected by some subsequent definitive test or finding. The origins of these errors can be classified by considering the provider-specific (cognitive) elements, the system-related contributions, and "no fault" elements reflecting diseases that present atypically or involve excessive patient noncompliance.4,27 Medical schools need to teach principles of optimal decision making and cognitive debiasing.47,48

More research on the nature of clinical decision making is needed to understand how errors arise and how they can be prevented. Research specifically targeted at diagnostic errors is currently being sponsored by the Agency for Healthcare Research and Quality,49 as well as the National Patient Safety Foundation, and hopefully future research funding can be identified to build on this foundation.

System Issues

A host of system properties contributes to diagnostic errors, and these need to be systematically identified and addressed, including issues related to reliable processes related to diagnostic testing, which are addressed in this issue of the Joint Commission Journal on Quality and Patient Safety.

Health care organizations need to accept partial responsibility for diagnostic errors,39 and diagnostic accuracy should be a concern of regulatory groups and policy-guiding organizations. The Joint Commission has taken a positive step in this direction by adding two requirements to the 2005 National Patient Safety Goals which are directed at improving the timeliness of reporting abnormal laboratory results and ensuring that critical abnormalities are communicated to a licensed caregiver.31 These requirements address one of the more common system-related diagnostic errors—failure to appropriately communicate abnormal test results. Health care organizations would also be well advised to address another common system-related condition that predisposes to error: ensuring that specialty expertise is available when needed, at all times and on all days.

Mandatory second opinions and computer-assisted diagnosis have the potential to reduce errors in pathology and radiology.62-64 Health care organizations need to evaluate the costs versus benefits of these novel interventions.

Even for those cases entirely dependent on the skill of the clinician for accurate diagnosis, health care organizations could help minimize errors by using system-level interventions to aid the clinician. For example, second readings of key diagnostic tests improve diagnostic accuracy,17 and clinical decision support tools may be helpful in a variety of settings. Similarly, efforts to enhance feedback to clinicians regarding their diagnoses would be beneficial. Organizations also have responsibility for training their clinicians in both the fundamentals of their science and in the diagnostic pitfalls that contribute to errors.60

Provider Issues

Physicians need to improve their calibration by getting feedback on the diagnoses they make. Cognitive feedback improves judgment and decisions made under conditions of uncertainty.59 For example, feedback regarding their diagnostic accuracy in reading standardized cases is thought to be a major factor explaining how radiologists in the United Kingdom have reduced their rate of diagnostic error.60

For the perceptual specialties (radiology, pathology, and dermatology), feedback can be provided through voluntary or mandated participation in quality assurance procedures. For primary care physicians and specialists, the pathway to enhanced feedback is not so clear-cut, and creative ideas are needed to identify how to obtain and provide this information.

The value of an autopsy as a feedback tool needs to be rediscovered.61 Beyond identifying what diagnoses were missed or wrong, the autopsy is the best tool we have to combat the overconfidence that seems to be ubiquitous in modern medicine. If providers and organizations are unable to increase the autopsy rate voluntarily, autopsies
could be required as a condition of Medicare participation or for Joint Commission certification. A novel strategy just emerging is the use of postmortem magnetic resonance imaging examination as a supplement or alternative to autopsy. Grieving family members may view this noninvasive procedure as a more acceptable alternative to autopsy.

Although they are less likely to feature autopsy findings than in Cabot’s day, conferences dedicated to reconciling clinical and pathological findings (CPCs) continue to be warranted. Morbidity and mortality conferences, while ubiquitous today in virtually all medical schools, need to be better structured to distill and learn from diagnostic mistakes. An extension on this theme is the growing number of sections in leading medical journals featuring quality and safety discussions, and the creative, multimedia “Web M&M” conferences sponsored by the Agency for Healthcare Research and Quality, presenting monthly cases of medical error in internal medicine, surgery and anesthesiology, critical care, and emergency medicine. These new resources, like autopsies, can increase awareness of medical error. It remains to be seen, however, whether they are as effective a feedback tool, insofar as the autopsy focuses on one’s own case while the literature and Web resources focuses on the errors of others: “. . . the goal of autopsy is not to uncover clinician’s mistakes or judge them, but rather to instruct clinicians in the sense of errando discimus (to be taught by one’s own mistakes).”

Finally, clinicians need to learn about overconfidence and the many other innate cognitive tendencies that detract from optimal reasoning and learning. Principles and pitfalls of clinical decision making need to be incorporated into the medical school curriculum, and education must also reach the legions of practicing clinicians who have had inadequate exposure to these concepts.

A potential concern is that feedback, though likely to improve calibration, may not be well received. An interesting and relevant experiment is currently underway in Major League Baseball, exploring the possibility that umpires can improve the consistency of their strike calls by viewing videos of each pitch at the completion of a game. The umpires have filed suit to stop the experiment, viewing the process as an objectionable intrusion. Similarly, some radiologists have resisted mandatory self-assessment meant to improve the accuracy of reading mammograms. This tension between autonomy and the need to improve calibration via feedback challenges efforts to increase diagnostic accuracy through this approach, but perhaps new incentives can be offered to make clinicians more interested in self-improvement.

Conclusion

A bright light is now being focused on patient safety, but diagnostic errors lie in the shadow. Clinicians and their health care organizations both need to take active steps to discover and analyze these errors. Diagnostic error rates will fall as remediable problems are identified, but the first step is to make these errors visible.

The author thanks Pat Croskerry and Sam Campbell for suggesting this topic; Nancy Franklin, Pat Croskerry, and Gordon Schiff for editorial suggestions; and Grace Garey and Kathy Kessel for help obtaining and organizing the references. This work was supported by a grant from the National Patient Safety Foundation.

Mark Graber, M.D., is Chief, Medical Service, VA Medical Center, Northport, New York, and Professor and Vice Chair, Department of Medicine, State University of New York at Stony Brook, New York. Please address requests for reprints to Mark Graber, M.D., mark.graber@med.va.gov.
References

References, continued


DO PHYSICIANS KNOW WHEN THEIR DIAGNOSES ARE CORRECT?

Charles P. Friedman, PhD et al.

Do Physicians Know When Their Diagnoses Are Correct? Implications for Decision Support and Error Reduction.

Do Physicians Know When Their Diagnoses Are Correct?  
Implications for Decision Support and Error Reduction

Charles P. Friedman, PhD,1 Guido G. Gatti, MS,1 Timothy M. Franz, PhD,2 Gwendolyn C. Murphy, PhD,3 Fredric M. Wolf, PhD,4 Paul S. Heckerling, MD,5 Paul L. Fine, MD,7 Thomas M. Miller, MD,8 Arthur S. Elstein, PhD6

1Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA; 2Department of Psychology, St. John Fisher College, Rochester, NY, USA; 3Division of Community Health, Duke University, Durham, NC, USA; 4Department of Medical Education and Informatics, University of Washington, Seattle, WA, USA; Departments of 5Medicine and 6Medical Education, University of Illinois at Chicago, Chicago, IL, USA; 7Department of Medicine, University of Michigan, Ann Arbor, MI, USA; 8Department of Medicine, University of North Carolina, Chapel Hill, NC, USA.

OBJECTIVE: This study explores the alignment between physicians’ confidence in their diagnoses and the “correctness” of these diagnoses, as a function of clinical experience, and whether subjects were prone to over- or underconfidence.

DESIGN: Prospective, counterbalanced experimental design.

SETTING: Laboratory study conducted under controlled conditions at three academic medical centers.

PARTICIPANTS: Seventy-two senior medical students, 72 senior medical residents, and 72 faculty interns.

INTERVENTION: We created highly detailed, 2-to 4-page synopses of 36 diagnostically challenging medical cases, each with a definitive correct diagnosis. Subjects generated a differential diagnosis for each of 9 assigned cases, and indicated their level of confidence in each diagnosis.

MEASUREMENTS AND MAIN RESULTS: A differential was considered “correct” if the clinically true diagnosis was listed in that subject’s hypothesis list. To assess confidence, subjects rated the likelihood that they would, at the time they generated the differential, seek assistance in reaching a diagnosis. Subjects’ confidence and correctness were “mildly” aligned (k = .314 for all subjects, .285 for faculty, .227 for residents, and .349 for students). Residents were overconfident in 41% of cases where their confidence and correctness were not aligned, whereas faculty were overconfident in 36% of such cases and students in 25%.

CONCLUSIONS: Even experienced clinicians may be unaware of the correctness of their diagnoses at the time they make them. Medical decision support systems, and other interventions designed to reduce medical errors, cannot rely exclusively on clinicians’ perceptions of their needs for such support.

KEY WORDS: diagnostic reasoning; clinical decision support; medical errors; clinical judgment; confidence.

DOI: 10.1111/j.1525-1497.2005.30145.x
J GEN INTERN MED 2005; 20:000–000.

When making a diagnosis, clinicians combine what they personally know and remember with what they can access or look up. While many decisions will be made based on a clinician’s own personal knowledge, others will be informed by knowledge that derives from a range of external sources including printed books and journals, communications with professional colleagues, and, increasingly, a range of computer-based knowledge resources.1 In general, the more routine or familiar the problem, the more likely it is that an experienced clinician can “solve it” and decide what to do based on personal knowledge only. This method of decision making uses a minimum of time, which is a scarce and precious resource in health care practice.

Every practitioner’s personal knowledge is, however, incomplete in various ways, and decisions based on incorrect, partial, or outdated personal knowledge can result in errors. A recent landmark study2 has documented that medical errors are a significant cause of morbidity and mortality in the United States. Although these errors have a wide range of origins,3 many are caused by a lack of information or knowledge necessary to appropriately diagnose and treat.4 The exponential growth of biomedical knowledge and shortening half-life of any single item of knowledge both suggest that modern medicine will increasingly depend on external knowledge to support practice and reduce errors.5

Still, the advent of modern information technology has not changed the fundamental nature of human problem solving. Diagnostic and therapeutic decisions, for the foreseeable future, will be made by human clinicians, not machines. What has changed in recent years is the potential for computer-based decision support systems (DSSs) to provide relevant and patient-specific external knowledge at the point of care, assembling this knowledge in a way that complements and enhances what the clinician decision maker already knows.6,7 DSSs can function in many ways, ranging from the generation of alerts and reminders to the critiquing of management plans.8-14 Some DSSs “push” information and advice to clinicians whether they request it or not; others offer no advice until it is specifically requested.

The decision support process presupposes the clinician’s openness to the knowledge or advice being offered. Clinicians who believe they are correct, or believe they know all they need to know to reach a decision, will be unmotivated to seek additional knowledge and unreceptive to any knowledge or sug-
gestions a DSS presents to them. The literatures of psychology and medical decision making\textsuperscript{15–18} address the relationship between these subjective beliefs and objective reality. The well-established psychological bias of “anchoring”\textsuperscript{15} stipulates that all human decision makers are more loyal to their current ideas, and resistant to changing them, than they objectively should be in light of compelling external evidence.

This study addresses a question central to the potential utility and success of clinical decision support. If clinicians’ openness to external advice hinges on their confidence in their assessments based on personal knowledge, how valid are these perceptions? Conceptually, there are 4 possible combinations of objective “correctness” of a diagnosis and subjective confidence in it: 2 in which confidence and correctness are aligned and 2 in which they are not. The ideal condition is an alignment of high confidence in a correct diagnosis. Confidence and correctness can also be aligned in the opposing sense: low confidence in a diagnosis that is incorrect. In this state, clinicians are likely to be open to advice and disposed to consult an external knowledge resource. In the “underconfident” state of nonalignment, a clinician with low confidence in a correct diagnosis will be motivated to seek information that will likely confirm an intent to act correctly. However, it is also possible that a consultation with an external resource can talk a clinician out of a correct assessment.\textsuperscript{19} The other nonaligned state, of greater concern for quality of care, is high confidence in an incorrect diagnosis. In this “overconfident” state, clinicians may not be open or motivated to seek information that could point to a correct assessment.

This work addresses the following specific questions:

1. In internal medicine, what is the relationship between clinicians’ confidence in their diagnoses and the correctness of these diagnoses?
2. Does the relationship between confidence and correctness depend on clinicians’ levels of experience ranging from medical student to attending physician?
3. To the extent that confidence and correctness are mismatched, do clinicians tend toward overconfidence or underconfidence, and does this tendency depend on level of clinical experience?

One study similar to this one in design and intent,\textsuperscript{20} but limited to medical students as subjects, found that students were frequently unconfident about correct diagnostic judgments when classifying abnormal heart rhythms. Our preliminary study of this question has found the relationship between correctness and confidence, across a range of training levels, to be modest at best.\textsuperscript{21}

**METHODS**

**Experimental Design and Dataset**

To address these questions, we employed a large dataset originally collected for a study of the impact of decision support systems on the accuracy of clinician diagnoses.\textsuperscript{19} We developed for this study detailed written synopses of 36 diagnostically challenging cases from patient records at the University of Illinois at Chicago, the University of Michigan, and the University of North Carolina. Each institution contributed 12 cases, each with a firmly established final diagnosis. The 2-to 4-page case synopses were created by three coauthors who are experienced academic internists (PSH, PSF, TMM). The synopses contained comprehensive historical, examination, and diagnostic test information. They did not, however, contain results of definitive tests that would have made the correct diagnosis obvious to most or all clinicians. The cases were divided into 4 approximately equivalent sets balanced by institution, pathophysiology, organ systems, and rated difficulty. Each set, with all patient-and institution-identifying information removed, therefore contained 9 cases, with 3 from each institution.

We then recruited to the study 216 volunteer subjects from these same institutions: 72 fourth-year medical students, 72 second-and third-year internal medicine residents, and 72 general internists with faculty appointments and at least 2 years of postresidency experience (mean, 11 years). Recruitment was balanced so that each institution contributed 24 subjects at each experience level. Each subject was randomly assigned to work the 9 cases comprising 1 of the 4 case sets. Each subject then worked through each of the assigned cases first without, and then with, assistance from an assigned computer-based decision support system. On the first pass through each case, subjects generated a diagnostic hypothesis set with up to 6 items. After generating their diagnostic hypotheses, subjects indicated their perceived confidence in their diagnosis in a manner described below. On the second pass through the case, subjects employed a decision support system to generate diagnostic advice, and again offered a differential diagnosis and confidence ratings. After deleting cases with missing data, the final dataset for this work consisted of 1,911 cases completed by 215 subjects.

Results reported elsewhere\textsuperscript{19} indicated that the computer-based decision support systems engendered modest but statistically significant improvements in the accuracy of diagnostic hypotheses (overall effect size of .32). The questions addressed by this study, emphasizing the concordance between confidence and correctness under conditions of uncertainty, focus on the first pass through each case where the subjects applied only their personal knowledge to the diagnostic task.

**Measures**

To assess the correctness of each clinician’s diagnostic hypothesis set for each case, we employed a binary score (correct or incorrect). We scored a case as correct if the established diagnosis for that case, or a very closely related disease, appeared anywhere in the subject’s hypothesis set. Final scoring decisions, to determine whether a closely related disease should be counted as correct, were made by a panel comprised of coauthors PSF, PSH, and TMM. The measure of clinician confidence was the response to the specific question: “How likely is it that you would seek assistance in establishing a diagnosis for this case?” “Assistance” was not limited to that which might be provided by a computer. After generating their diagnostic hypotheses for each case, subjects responded to this question using an ordinal 1 to 4 response scale with anchor points of 1 representing “unlikely” (indicative of high confidence in their diagnosis) and 4 representing “likely” (indicative of low confidence). Because subjects did not receive feedback, they offered their confidence judgments for each case without any definitive knowledge of whether their diagnoses were, in fact, correct. Because they reflect only the subjects’ first pass through each case, these confidence judgments
were not confounded by any advice subjects might later have received from the decision support systems.

**Analysis**

In this study, each data point pairs a subjective confidence assessment on a 4-level ordinal scale with a binary objective correctness score. The structure of this experiment and the resulting data suggested two approaches to analyzing the results. Given that each subject in this study worked 9 cases, and offered confidence ratings on a 1 to 4 scale for each case, interpretations of the meanings of these scale points might be highly consistent for each subject but highly variable across subjects. Our first analytic approach therefore sought to identify an optimal threshold for each subject to distinguish subjective states of “confident” and “unconfident.” This approach addresses the “pooling” problem, identified by Swets and Picket,

that would tend to underestimate the magnitude of the relationship between confidence and correctness. Our second analytical approach took the assumption that all subjects made the same subjective interpretation of the confidence scale. This second approach entails a direct analysis of the 2-level by 4-level data with no within-subject thresholding. Qualitatively, the first approach approximates the upper bound on the relationship between confidence and correctness, while the second approach approximates the lower bound.

To implement the first approach, we identified, for each subject, the threshold value along the 1 to 4 scale that maximized the proportion of cases where confidence and correctness were aligned. With reference to Table 1, we sought to find the threshold value that maximized the numbers of cases in the on-diagonal cells. For 58 subjects (27%), we found that maximum alignment was achieved by classifying only ratings of 1 as confident and all other ratings as unconfident; for 105 subjects (49%), maximum alignment was achieved by classifying ratings of 1 or 2 as confident; and for the remaining 52 subjects (24%), maximum alignment was achieved by classifying ratings of 1, 2, or 3 as confident. This finding validated our assumption that subjects varied in their interpretations of the scale points. We then created a dataset for further analysis that consisted, for each case worked by each subject, of a binary correctness score and a binary confidence score calculated using each subject’s optimal threshold.

To address the first research question with the first approach, we computed Kendall’s $\tau_b$ and $x$ coefficients to characterize the relationship between subjects’ correctness and confidence levels. We then modeled statistically the proportions of cases correctly diagnosed, as a function of confidence (computed as a binary variable as described above), subjects’ level of training (faculty, resident, student), and the interaction of confidence and training level. To address the second question, we modeled the proportions of cases in which confidence and correctness were aligned, as a function of training level. To address the third research question, we focused only on those cases in which confidence and correctness were not aligned. We modeled the proportions of cases in which subjects were overconfident (high confidence in an incorrect diagnosis) as a function of training level.

All statistical models used the Generalized Linear Model (GzLM) procedure assuming diagnostic correctness, alignment, and overconfidence to be distributed as Bernoulli variables with a logit link and used naive empirical covariance estimates for the model effects to account for the clustering of cases within subjects. Wald statistics were employed to test the observed results against the null condition. Ninety-five percent confidence intervals were calculated by transforming logit scale Wald intervals using naive empirical standard error estimates into percent scale intervals. The SPSS for Windows (SPSS Inc., Chicago, IL) and SAS Proc GENMOD (SAS Institute Inc., Cary, NC) software were employed for statistical modeling and data analyses.

Our second approach offers a contrasting strategy to address the first and second research questions. To this end, we computed nonparametric correlation coefficients (Kendall’s $\tau_b$) between the 2-level variable of correctness and the 4 levels of confidence from the original data, without thresholding. We computed separate $\tau_b$ coefficients for subjects at each experience level, and for the sample as a whole. Correlations were computed with case as the unit of analysis after exploratory analyses correcting for the nesting of cases within subjects led to negligible changes in the results.

The power of the inferential statistics employed in this analysis was based on the two-tailed $t$ test, as the tests we performed are analogous to testing differences in means on a logit scale. Because our tests are based on a priori unknown marginal cell counts, we halved the sample size to estimate power. For the analyses addressing research question 1, which use all cases, power is greater than .96 to detect a small to

<table>
<thead>
<tr>
<th>Experience Level</th>
<th>Correctness of Diagnosis</th>
<th>High</th>
<th>Confidence Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>Correct</td>
<td>63 (55 to 71)</td>
<td>105 (88 to 125)</td>
<td>168 (146 to 192)</td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td>35 (27 to 43)</td>
<td>442 (422 to 459)</td>
<td>477 (453 to 499)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>98 (72 to 132)</td>
<td>547 (513 to 573)</td>
<td>645</td>
</tr>
<tr>
<td>Residents</td>
<td>Incorrect</td>
<td>140 (129 to 150)</td>
<td>141 (124 to 159)</td>
<td>281 (256 to 306)</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
<td>98 (88 to 109)</td>
<td>259 (241 to 276)</td>
<td>357 (332 to 382)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>238 (193 to 287)</td>
<td>400 (351 to 445)</td>
<td>638</td>
</tr>
<tr>
<td>Faculty</td>
<td>Incorrect</td>
<td>80 (69 to 92)</td>
<td>237 (221 to 253)</td>
<td>317 (299 to 346)</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
<td>270 (250 to 300)</td>
<td>381 (338 to 433)</td>
<td>628</td>
</tr>
<tr>
<td>All subjects</td>
<td>Incorrect</td>
<td>370 (352 to 398)</td>
<td>390 (357 to 425)</td>
<td>760 (713 to 808)</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
<td>213 (195 to 231)</td>
<td>1,328 (1,246 to 1,405)</td>
<td>1,911</td>
</tr>
</tbody>
</table>

Cells contain counts of cases, with 95% confidence intervals in parentheses.
RESULTS

Results with Threshold Correction

Table 1 displays the crosstabulation of correctness of diagnosis and binary levels of confidence (with 95% confidence interval) for all subjects and separately for each clinical experience level, using each subject’s optimal threshold to dichotomize the confidence scale. The difficulty of these cases is evident from Table 1, as 760 of 1,911 (40%) were correctly diagnosed by the full set of subjects. Diagnostic accuracy increased monotonically with subjects’ clinical experience. The difficulty of the cases is also reflected in the distribution of the confidence ratings, with subjects classified as confident for 583 (31%) of 1,911 cases, after adjustment for varying interpretations of the scale. These confidence levels revealed the same general monotonic relationship with clinical experience. Across the entire sample of subjects, confidence and correctness were aligned for 1,308 of 1,911 cases (68%), corresponding to Kendall’s τ₀ = .321 (P < .0001) and a k value of .314. Alignment was seen in 64% of cases for faculty (τ₀ = .291 [P < .0001]; k = .285), 63% for residents (τ₀ = .230 [P < .0001]; k = .227), and 78% for students (τ₀ = .369 [P < .0001]; k = .349).

Figure 1 offers a graphical portrayal, for each experience level, of the proportions of correct diagnoses as a function of confidence, with 95% confidence intervals. The relationship between correctness and confidence, at each level, is seen in the differences between these proportions.

Wald statistics generated by the statistical model reveal a significant alignment between diagnostic correctness and confidence across all subjects (χ² = 199.64, df = 1, P < .0001). Significant relationships are also seen between correctness and training level (χ² = 20.40, df = 2, P < .0001) and in the interaction between confidence and training level (χ² = 17.00, df = 2, P < .0002). Alignment levels for faculty and residents differ from those of the students (P < .05); and from inspection of Figure 1 it is evident that students’ alignment levels are higher than those of faculty or residents.

With reference to the third research question, Table 2 summarizes the case frequencies for which clinicians at each level were correctly confident—where confidence was aligned with correctness—as well as frequencies for the “nonaligned” cases where they were overconfident and underconfident. Students were overconfident in 25% of nonaligned cases, corresponding to 5% of cases they completed. Residents were overconfident in 41% of nonaligned cases, and 15% of cases overall. Faculty physicians were overconfident in 36% of nonaligned cases, and 13% of cases overall.

All subjects were more likely to be underconfident than overconfident (χ² = 29.05, P < .0001). Students were found to be more underconfident than residents (Wald statistics: χ² = 6.19, df = 2, P < .05). All other differences between subjects’ experience levels were not significant.

Results Without Threshold Correction

The second approach to analysis yielded Kendall τ₀ measures of association between the binary measure of correctness and the 4-level measure of confidence, computed directly from the study data, without any corrections. For all subjects and cases, we observed τ₀ = .106 (N = 1,911 cases; P < .0001). Separately for each level of training, Kendall coefficients are: faculty τ₀ = .103 (n = 628; P < .005), residents τ₀ = .041 (n = 638; NS), and students τ₀ = .121 (n = 645 cases; P < .001). The polarity of the relationship is as would be expected, associating correctness of diagnosis with higher confidence levels. The τ₀ values reported here can be compared with their counterparts, reported above, for the analyses that included threshold correction.

DISCUSSION

The assumption built into the first analytic strategy, that subjects make internally consistent but personally idiosyncratic interpretations of confidence, generates what may be termed upper-bound estimates of alignment between confidence and
correctness. Under the assumptions embedded in this analysis, the results of this study indicate that the correctness of clinicians’ diagnoses and their perceptions of the correctness of these diagnoses are, at most, moderately aligned. The correctness and confidence of faculty physicians and senior medical residents were aligned about two thirds of the time—and in cases where correctness and confidence were not aligned, these subjects were more likely to be underconfident than overconfident. While faculty subjects demonstrated tendencies toward greater alignment and less frequent overconfidence than residents, these differences were not statistically significant. Students’ results were substantially different from those of their more experienced colleagues, as their confidence and correctness were aligned about four fifths of the time and more highly skewed, when nonaligned, toward underconfidence. The alignment between “being correct” and “being confident”—within groups and for all subjects—would be qualitatively characterized as “fair,” as seen by $x$ coefficients of .2 to .4.\(^{27}\)

The more conservative second mode of analysis yielded smaller relationships between correctness and confidence, as seen in the $g_0$ coefficient for all subjects, which is smaller by a factor of three. For the residents, the relationship between correctness and confidence does not exceed chance expectations when computed without thresholding. Comparison across experience levels reveals the same trend seen in the primary analysis, with students displaying the highest level of alignment.

The greater apparent alignment for the students, under both analytic approaches, may be explained by the difficulty of the cases. The students were probably overmatched by many of these cases, perhaps guessing at diagnoses, and were almost certainly aware that they were overmatched. This is seen in the low proportions of correct diagnoses for students and the low levels of expressed confidence. These skewed distributions would generate alignment between correctness and confidence of 67% by chance alone. While students’ alignments exceeded these chance expectations, a better estimate of their concordance between confidence and correctness might be obtained by challenging the students with less difficult cases, making the diagnostic task as difficult for them as it was for the faculty and residents with the cases employed in this study. We do not believe it is valid to conclude from these results that the students are “more aware” than experienced clinicians of when they are right and wrong.

By contrast, residents and faculty correctly diagnosed 44% and 50% of these difficult cases, respectively, and generated distributions of confidence ratings that were less skewed than those of the students. In cases for which these clinicians’ correctness and confidence were not aligned, both faculty and residents showed an overall tendency toward underconfidence in their diagnoses. Despite the general tendency toward underconfidence, residents and faculty in this study were overconfident, placing credence in a diagnosis that was in fact incorrect, in 15% (98/938) and 12% (80/928) of cases, respectively. Because these two more experienced groups are directly responsible for patient care, and offered much more accurate diagnoses for these difficult cases, findings for these groups take on a different interpretation and perhaps greater potential significance.

In designing the study, we approached the measurement of “confidence” by grounding it in hypothetical clinical behavior. Rather than asking subjects directly to estimate their confidence levels in either probabilistic or qualitative terms, we asked them for the likelihood of their seeking help in reaching a diagnosis for each case. We considered this measure to be a proxy for “confidence.” Because our intent was to inform the design of decision support systems and medical error reduction efforts generally, we believe that this behavioral approach to assessment of confidence lends validity to our conclusions.

Limitations of this study include restriction of the task to diagnosis. Differences in results may be seen in clinical tasks other than diagnosis, such as determination of appropriate therapy for a problem already diagnosed. The cases, chosen to be very difficult and with definitive findings excluded, certainly generated lower rates of accurate diagnoses than are typically seen in routine clinical practice. Were the cases in this study more routine, this may have affected the measured levels of alignment between confidence and correctness. In addition, this study was conducted in a laboratory setting, using written case synopses, to provide experimental precision and control. While the case synopses contained very large amounts of clinical information, the task environment for these subjects was not the task environment of routine patient care. Clinicians might have been more, or less, confident in their assessments had the cases used in the study been real patients for whom these clinicians were responsible; and in actual practice, physicians may be more likely to consult on difficult cases regardless of their confidence level. While we employed volunteer subjects in this study, the sample sizes at each institution for the resident and faculty groups were large relative to the sizes at each institution of their respective populations, and thus unlikely to be skewed by sampling bias.

The relationships, of “fair” magnitude, between correctness and confidence were seen only after adjusting each subject’s confidence ratings to reflect differing interpretations of the confidence scale. The secondary analytic approach, which does not correct individuals’ judgments against their own optimal thresholds, results in observed relationships between correctness and confidence that are smaller. Under either set of assumptions, the relationship between confidence and correctness is such that designers of clinical decision support systems cannot assume clinicians to be accurate in their own assessments of when they do and do not require assistance from external knowledge resources.

This work was supported by grant R01-LM-05630 from the National Library of Medicine.

**REFERENCES**


16. Lichtenstein S, Fischhoff B. Do those who know more also know more about how much they know? Organ Behav Hum Perform. 1977;20:159–83.


