Predictors of Diagnostic Accuracy and Safe Management in Difficult Diagnostic Problems in Family Medicine

Olga Kostopoulou, PhD, Jurriaan Oudhoff, PhD, Radhika Nath, PhD, Brendan C. Delaney, MD, Craig W. Munro, MBChB, Clare Harries, PhD, Roger Holder, BSc

Objective. *To investigate the role of information gathering and clinical experience on the diagnosis and management of difficult diagnostic problems in family medicine.* ***Method.*** *Seven diagnostic scenarios including 1 to 4 predetermined features of difficulty were constructed and presented on a computer to 84 physicians: 21 residents in family medicine, 21 family physicians with 1 to 3 y in practice, and 42 family physicians with ≥10 y in practice. Following the Active Information Search process tracing approach, participants were initially presented with a patient description and presenting complaint and were subsequently able to request further information to diagnose and manage the patient. Evidence-based scoring criteria for information gathering, diagnosis, and management were derived from the literature and a separate study of expert opinion.* ***Results.*** *Rates of misdiagnosis were in accordance with the number of features of difficulty. Seventy-eight percent of incorrect diagnoses were followed by inappropriate management and 92% of correct diagnoses by appropriate management. Number of critical cues requested (cues diagnostic of any relevant differential diagnoses in a scenario) was a significant predictor of accuracy in 6 scenarios: 1 additional critical cue increased the odds of obtaining the correct diagnosis by between 1.3 (95% confidence interval [CI], 1.0–1.8) and 7.5 (95% CI, 3.2–17.7), depending on the scenario. No effect of experience was detected on either diagnostic accuracy or management. Residents requested significantly more cues than experienced family physicians did.* ***Conclusions.*** *Supporting the gathering of critical information has the potential to improve the diagnosis and management of difficult problems in family medicine.*
Key words: *diagnosis; error; information gathering; experience; process tracing; family practice.* ***(Med Decis Making 2008;28:668–680)***

Prompt diagnosis of poorly differentiated and potentially serious disease is one of the core clinical competencies of family physicians. Most medicolegal claims against family physicians are about delay in diagnosis or misdiagnosis (63%–66%),[1,2] yet little empirical research has been done on the causes of diagnostic error. Retrospective reviews of litigation cases in UK primary care[3] and US ambulatory care[4] and a multimethod study in US internal medicine[5] suggest that diagnostic errors are caused by both cognitive and system factors. The system is the usual focus of patient safety and quality improvement initiatives, whereas cognitive factors are relatively ignored.

This study concentrated on cognitive factors, specifically, aspects of the diagnostic process that might predict accuracy of diagnosis in difficult cases in family medicine. We assumed that difficult cases are more likely to be misdiagnosed and hence focused on these rather than a representative sample of the family physician's workload. Difficult cases need not be rare diseases that physicians may see once in their careers and thus fail to consider. Common conditions such as ischemic heart disease may be difficult to diagnose as they do not always present with typical features.

Beyond the influence of the system on cognition, through, for example, workload, availability of information, and interface design,[6] the likelihood of error also depends on the interaction between cognition

Received 3 January 2008 from the Department of Primary Care, University of Birmingham, United Kingdom (OK, JO, RN, BCD, CWM, RH), and the Department of Psychology, University College London, United Kingdom (CH). This work was presented at the 29th annual meeting of the Society for Medical Decision Making, Pittsburgh, Pennsylvania. Revision accepted for publication 14 March 2008.

Address correspondence to Olga Kostopoulou, PhD, Primary Care Clinical Sciences, University of Birmingham, Birmingham B15 2TT, UK; phone: +44(0)121 414 5390; fax: +44(0)121 414 3759; e-mail: o.kostopoulou@bham.ac.uk.

DOI: 10.1177/0272989X08319958

and diagnostic problem. Diagnostic problem characteristics include the profile of symptoms and signs (specificity, typicality, consistency), the number of diagnostic alternatives and extent of presentation overlap, the prevalence and severity of the diagnostic alternatives, and the availability of an explanation for the symptoms. These problem characteristics can influence the likelihood of error as a result of the way people naturally think (i.e., their reasoning tendencies). For example, people tend to look for and recognize familiar patterns in the presented information while ignoring or explaining away information that does not fit a recognized pattern.[7] They also tend to assume only 1 cause for the presenting symptoms.[8] Even though both these tendencies can be considered rational adaptations to complex task environments,[9] they can lead to diagnostic error,[7,8,10] in situations with multiple coexisting causes or multiple alternatives with partially overlapping features. Confirmation bias (i.e., the tendency to confirm rather than disconfirm one's working diagnosis),[11,12] has the potential to lead to error, especially in nonspecific presentations with only a few typical features, in which diagnostic alternatives cannot be excluded or confirmed with certainty. Clinicians have been found to rely on base rates rather than to neglect them.[13–15] This tendency may be more prominent in primary care, in which the prevalence of serious disease is low, and may result occasionally in missing less prevalent diagnoses. Finally, an existing diagnostic label,[16] a coexisting disease,[17] or a readily available explanation for the patient's symptoms[18,19] have all been found to influence the ability to restructure the diagnostic problem and look for alternative explanations. This is similar to diagnostic overshadowing found in the mental health[20] and intellectual disability[21,22] literatures (i.e., the misattribution of signs and symptoms to an existing mental illness, leading to underdiagnosis and mistreatment of comorbidities).

In the search for predictors of diagnosis and management, the study investigated the role of clinical experience and information gathering, using realistic diagnostic problems and evidence-based criteria for assessing performance. According to the script theory,[23] clinicians acquire more flexible scripts with increasing experience. Flexible scripts can account for greater variability in the presentation of disease and less typical features. Changes in reasoning and organization of knowledge have indeed been demonstrated consistently in the transition from medical student to doctor.[24] The effect of length of practice per se is rarely examined, and a clear relationship with diagnostic accuracy has not been found.[25–27] On the other hand, a possible inverse relationship between length of practice and appropriateness of patient management has recently been suggested.[28]

The gathering of clinical information is the observable part of the diagnostic process in terms of history taking, physical examination, ordering of tests, and referrals. The appropriateness of clinical information gathering and its impact on accuracy have not been examined using realistic diagnostic problems and evidence-based scoring criteria. Difficulties in information gathering (selecting the most diagnostic symptom) have been shown in a simple choice task involving 2 abstract diseases (diseases A and B) and 2 symptoms, but performance influences were unclear.[29] The authors suggested that information gathering is probably more problematic than information integration. However, the lack of context precluded the use of relevant clinical knowledge to guide hypothesis generation and testing. Generalizability of their finding to real-life clinical diagnosis is therefore questionable. Using realistic, context-rich problems, Elstein and colleagues[13] found that the gathering of diagnostic information did not influence accuracy, whereas thoroughness did to a small extent. They attributed diagnostic errors mainly to the overinterpretation of nondiagnostic information.

Our study used patient scenarios (vignettes) that were carefully constructed to reflect real practice; that is, they were interactive, were sufficiently complex, and imposed time constraints. Evidence-based scoring criteria were derived and employed. These conditions have been shown to establish the validity of vignettes as a measure of diagnostic and management performance.[30,31]

METHODS

Materials

We used 2 reviews of litigation cases,[2,3] conducted a systematic review of the clinical literature (manuscript in preparation), and performed interviews with 11 family physicians to survey difficult diagnoses in primary care. On the basis of this and the clinical reasoning literature,[13,32,33] we derived 10 features of diagnostic problems that induce difficulty and increase the likelihood of error. Even though each feature increases difficulty by itself, misdiagnosed cases documented in the literature and discussed in the interviews contained combinations of features. For the purposes of this study, 7 diagnostic scenarios were

thus designed to include 1 to 4 features of difficulty (Table 1).

Scenario building was an iterative, multistage process. First, a decision was made about the specific disease that each scenario would depict. The features of difficulty to be included were then determined, considering the real-life difficulty in making the specific diagnosis. Patient presentations were constructed consistent with the diagnosis, including sufficient detail to make the patients seem realistic and plausible. Patient descriptions were produced in a standardized format to include information that family physicians could access from the patient record and any relevant visual features (see the example in Box 1).

For each scenario, the relevant differential diagnoses were determined by considering what other illnesses present similarly in patients of the same age and gender (Table 1). Scenarios were designed to contain both diagnostic and nondiagnostic cues, reflecting real-life clinical consultations. Cue presence or absence was manipulated to have a leading diagnosis, the one that the scenario was intended to depict. Nondiagnostic cues were cues known to be used in clinical practice but whose diagnostic value is either unknown or too small (e.g., anxiety in irritable bowel syndrome). Diagnostic cues (critical cues) were identified from the literature and from expert opinion, as described below.

For each scenario, base rates of differential diagnoses and likelihood ratios (LRs) of cues were identified from Medline searches. We considered critical those cues with $LR+ > 1.5$ or $LR- < 0.67$. Only a few LRs were identified and only in relation to some of the differential diagnoses. Expert opinion was therefore sought to 1) elicit further LRs and 2) confirm the most likely diagnosis in each scenario. These were the aims of a separate Web-based study, the Expert Opinion Survey (EOS; manuscript in preparation). Experts were family physicians with a specialist interest in a clinical area relevant to the scenarios, who were invited to participate via their respective primary care specialist societies. Family physicians with a specialist interest undergo additional training and accreditation. Respondents were sequentially presented with all the cues available in a scenario. To minimize order effects, 3 different cue orders were employed per scenario. At each cue presentation, respondents were asked to estimate the likelihood of each differential diagnosis using sliders from 0 (*no chance*) to 10 (*certain*). A hierarchical linear model was created for each cue order that assessed the change in the likelihood ratings over the sequence of cues presented. Any cue that shifted ratings from the previous cue significantly was considered critical and complemented those derived from the literature. The number of critical cues included in each scenario is shown in Figure 1. The differential diagnoses per scenario shown in Table 1 are ranked according to the final mean likelihood estimates from the EOS.

Scenarios were piloted with family physicians and were iteratively reviewed by the research team for plausibility, consistency, and completeness of information. All cues with abnormal values in accordance with the patient's condition were included in a scenario, as were cues with normal values that participants might request. Finally, a number of generic "no" or "normal" responses were created to ensure that all information requests could be answered.

Participants

All 778 family physicians in Birmingham and Solihull, UK, were invited to participate. They were offered written feedback on their performance as compared with best practice and financial recompense for the time taken to participate in the study. Two hundred one family physicians responded (response rate = 26%), and 130 expressed willingness to participate. We applied the selection criterion (years of experience) to the positive responses and recruited 84 participants: 21 residents, 21 family physicians with 1 to 3 y in practice (intermediates), and 42 family physicians with $\geq$10 y in practice (experienced; range, 10–31 y, mean 17 y). The experienced group included 21 family physicians who trained residents (trainers) and 21 who did not (nontrainers), matched for years in practice. The mean age was 31 y for the residents (range, 26–42 y) and 44 y for the family physicians (range, 28–65 y); 60% were men. The mean age of the population of family physicians in the study area was 50 y, with 63% being men, making the study sample representative but slightly younger.

Study Procedure

The study was carried out at the participants' office or home or at the university, according to their preference. A total of 1.5 h were allocated for the provision of standardized study information and instructions by a researcher, obtaining of written participant consent, practice on a training scenario, and completion of all 7 scenarios. The presentation order of the 7 scenarios was varied following a balanced Latin square design. Specifically, 7 scenario orders were created, so that a scenario never appeared in the same position in more than 1 order.

Table 1 The 7 Scenarios and Their Respective Features of Difficulty

Scenario Name	Pyrexial Child	Dyspnea 1	Abdominal Pain	Chest Pain	Dyspnea 2	Headache	Fatigue
Brief description of patient and main complaint	Toddler with fever (presenting 3 consecutive times to the family physician)	68-y-old man, smoker, presenting with dyspnea	Young woman presenting with 3-mo abdominal pain	60-y-old man presenting with chest pain, 1st felt while lifting washing machine	Elderly COPD patient presenting with episodes of dyspnea increasing in frequency	69-y-old woman presenting with headaches and other nonspecific symptoms	52-y-old man, previously diagnosed with depression, on antidepressants (lofepramine), presenting with fatigue and dry mouth
Differential diagnoses ranked according to EOS likelihood estimates (correct diagnoses in bold)	**Kawasaki disease**; Pharyngitis; URTI; tonsillitis; UTI, pneumonia; meningitis	**COPD and aortic stenosis**; COPD alone; CCF; MPE	**Celiac disease**; IBS; IBD; UTI; PID	Musculoskeletal pain; **new-onset angina**; GERD; Anxiety; PE; pneumonia	**Cor pulmonale**; acute COPD exacerbation; multiple PE; aortic stenosis	**Temporal arteritis**; tension headache; migraine; sinusitis; brain tumor	**Diabetes**; side effects of lofepramine
Features of difficulty							
1. Single, nonspecific symptom	Fever		Abdominal pain				
2. Multiple nonspecific symptoms that do not make a pattern (in addition to the difficulty deciding what symptom is important to pursue, such presentations may be attributed to somatization of emotional problems; as a result, organic signs may be missed)						Headaches, feeling weak and tired, heaviness in the arms and legs	

(continued)

Table 1 (continued)

Scenario Name	Pyrexial Child	Dyspnea 1	Abdominal Pain	Chest Pain	Dyspnea 2	Headache	Fatigue
3. Obvious, but not necessarily the correct etiology				Onset of pain during heavy lifting: "I thought I'd pulled a muscle," says patient			Fatigue can be attributed to the depression; dry mouth is a known side effect of the lofepramine
4. Uncommon critical cues or cue-disease associations	Strawberry tongue, peeling fingers						
5. Atypical presentations (presentations with a shortage of typical features or including cues with unexpected values)		Patient has dyspnea but no chest pain or syncope (typical features of aortic stenosis)		Right arm radiation of chest pain (rather than left)			
6. Critical cue that is necessary for diagnosis		Soft systolic murmur with neck radiation	Microcytic anemia		Resting ECG shows heart rate of 98, right axis deviation, and p-pulmonale		
7. Higher prevalence of the main diagnostic competitor(s) (less prevalent diseases may not be considered seriously)	Pharyngitis; tonsillitis; URTI		IBS	Musculoskeletal pain		Tension headache	

(continued)

Table 1 (continued)

Scenario Name	Pyrexial Child	Dyspnea 1	Abdominal Pain	Chest Pain	Dyspnea 2	Headache	Fatigue
8. Rare disease (in addition to no. 7 above, the clinician may not know or have forgotten a rare disease)	Kawasaki disease						
9. Causally interacting coexisting diseases with partially overlapping features (in which one disease has caused the other or both diseases have been caused by a common factor)					Preexisting COPD can cause cor pulmonale but can also explain the presenting complaint		
10. Causally independent coexisting diseases with partially overlapping features		COPD can be easily diagnosed in this patient and explains most symptoms; the clinician may stop searching and miss aortic stenosis					
Number of features of difficulty	4	3	3	3	2	2	1

Note: COPD = chronic obstructive pulmonary disease; EOS = Expert Opinion Survey; URTI = upper respiratory tract infection; UTI = urinary tract infection; CCF = congestive cardiac failure; MPE = malignant pleural effusion; IBS = irritable bowel syndrome, IBD = inflammatory bowel disease; PID = pelvic inflammatory disease; GERD = gastroesophageal reflux disease; PE = pulmonary embolism; ECG = electrocardiogram.

Box 1
Patient Description and Presenting Complaint for the Dyspnea 2 Scenario

NAME: Mabel Evans
AGE: 76 years old
ETHNICITY: Caucasian
HEIGHT: 1.60 m
WEIGHT: 62 kg (BMI = 24.2, measured last year)
SMOKING STATUS: Smoked 20 cigarettes per day from age 15, gave up last year (60 pack years)
LAST BP: 122/67, 6 weeks ago
PAST MEDICAL HISTORY: COPD 1998, hypertension 1996
MEDICATION: Combivent inhaler (via spacer) 2 puffs qds, bendroflumethazide 2.5 mg od
LAST CONSULTATION: 6 weeks ago, for exacerbation of COPD. Attended 3 times in previous 6 months for this reason.
APPEARANCE: On entering the room, you notice that her lips look blue.
Patient: "I've come to see you about my breathing again, doctor. This is the 4th time in the past 6 months that it's flared up and I'm getting worried about it. It came out of the blue about 3 days ago. My breathing is now awful, I get out of breath doing the slightest thing, even getting ready to come to the surgery had me all out of puff today. I've been wheezy too, much worse than usual."
Note: BMI = body mass index; BP = blood pressure; COPD = chronic obstructive pulmonary disease.

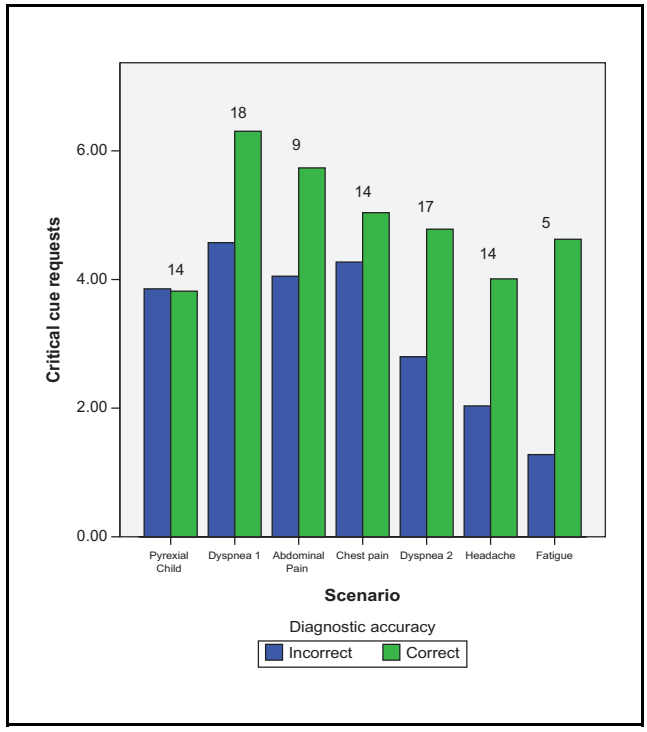

Figure 1 Number of critical cue requests per scenario shown separately for correct and incorrect diagnoses. The number of critical cue requests was standardized, that is, divided by the critical cues available per scenario (shown above the bars) and multiplied by 10.

Each order was presented to 12 participants: 3 residents, 3 intermediates, 3 trainers, and 3 nontrainers. The researcher administered the scenarios via a laptop computer and an additional monitor, using a computer program specifically written for the study. A process-tracing approach, Active Information Search,[34,35] was employed, as it enables information gathering to be traced without changing the nature of clinical consultation. On the participant's screen, the patient description and presenting complaint were shown initially. Participants could then request further information to diagnose and manage the patient. The researcher selected the cue relevant to the participant's question from a drop-down menu and displayed the answer on the participant's screen. If participants asked open questions, the researcher asked them to be more specific. If they asked for cues that had not been designed into the scenarios, the researcher selected appropriately from the set of generic responses, for example, "No," "No, I haven't," "No abnormality," "I haven't had any problems with that," and "The test is normal." The computer program logged the information requests in order and the time that each answer was displayed to the participant. The scenario ended when participants decided to end the consultation and initiate management or when they wished to refer to a specialist for diagnostic or management purposes. They were asked what they thought was wrong with the patient and how they would manage him or her. The researcher recorded all differential diagnoses and management decisions offered and proceeded to the next scenario. No feedback was provided about diagnostic accuracy at that stage. Following the diagnosis of all 7 scenarios, 3 scenarios were selected for stimulated recall.[36] Analysis of the stimulated recall data is ongoing and not reported here.

Outcome Measures

Outcome measures were the number of cue requests (total, critical, noncritical), the time taken to diagnose (from patient presentation to presentation of the last cue requested), diagnostic accuracy, and

appropriateness of management. Before scoring accuracy, the diagnostic terminology used by the participants was standardized, as different terms (e.g., *celiac disease* and *gluten sensitivity*) could be used for the same diagnosis. All diagnoses provided were mapped onto the list of differential diagnoses for each scenario. Any diagnoses that could not be mapped were recorded as they were given. The 2 clinicians from the research team (B.C.D. and C.W.M.) performed the mapping independently. Agreement across scenarios was 89% to 96%. Disagreements were resolved with discussion.

Diagnostic accuracy was scored as either correct (correct diagnosis included in the final list of differentials) or incorrect (correct diagnosis not given). In each scenario, the diagnosis that ranked at the top according to the EOS mean likelihood estimates was taken as the correct diagnosis (Table 1). In the chest pain scenario, the musculoskeletal cause was ranked 1st, with angina 2nd. However, angina could not be satisfactorily excluded given the patient features and was therefore considered a necessary differential diagnosis when scoring accuracy.

Management was scored as either appropriate or inappropriate in relation to the correct diagnosis. Appropriate management was informed by clinical guidelines for the diagnosis. Inappropriate management was likely to harm the patient. The 2 clinicians from the team scored all management decisions independently. Agreement ranged from 80% to 100% across scenarios. All disagreements were resolved with discussion.

Analyses

Logistic regressions were performed on diagnostic accuracy and appropriateness of management following a stepwise forward variable selection process. The number of critical cue requests was standardized to give a range of 0 to 10 (Figure 1). Where significant interactions with the scenario were identified, logistic regressions on accuracy and Fisher's exact tests on management (due to small frequencies in some cells) were performed for each scenario separately. The effect of experience on cue requests and time taken was examined with 1-way analyses of variance, followed by post hoc comparisons (Dunnett C). To test for differences in requests for specific critical cues between accurate and inaccurate diagnoses, chi-square tests were performed on each scenario with Bonferroni adjustment for multiple comparisons.

The pyrexial child scenario: failure to manage Kawasaki disease appropriately can have serious implications for the patient, including the development of coronary artery aneurysms. These occur most commonly after day 10 of the illness in untreated children.[37] In the scenario, the child presents to the physician on days 4, 8, and 12 after symptoms began. It is essential that the child is sent to the hospital by day 8, to reduce the risk of complications. Therefore, all the statistical analyses were done with data from the 2nd presentation (day 8 of illness). This included data from 4 participants who admitted the child to the hospital at 1st presentation (day 4 of illness) and thus never received the subsequent presentations of the scenario. Summary accuracy and management data from the 3rd presentation (day 12 of illness) are presented in Table 2.

RESULTS

Diagnostic Accuracy

Average rates of accuracy ranged from 25% to 57% depending on the scenario and in accordance with the number of features of difficulty (Spearman's rho = 0.95, $P < 0.01$; Table 2). Thirty-five percent of the diagnoses made by residents were correct, compared with 43% by intermediates and 45% by experienced family physicians (44% by trainers and 46% by nontrainers). Experience was not a significant predictor of accuracy in the regression model ($P = 0.06$). A highly significant interaction between the number of critical cue requests and scenario predicted diagnostic accuracy ($P < 0.0001$) and explained 43% of the variation in accuracy (Nagelkerke $R^2 = 0.43$). More critical cue requests were associated with greater accuracy in all scenarios, except for the pyrexial child (Figure 1). The relationship was significant ($P < 0.05$) but weak for the chest pain scenario (odds ratio, 1.3; 95% confidence interval [CI], 1.0–1.7). In the rest of the scenarios, the odds of diagnosing correctly with 1 additional critical cue more than doubled: odds ratios of 2.3 for dyspnea 2 (95% CI, 1.6–3.3), 2.6 for abdominal pain (95% CI, 1.6–4.1), 2.7 for dyspnea 1 (95% CI, 1.7–4.3), 3.3 for fatigue (95% CI, 1.9–5.9), and 7.5 for headache (95% CI, 3.2–17.7). There were significant differences in the requests for specific critical cues between correct and incorrect diagnoses in the 6 scenarios in which critical cues predicted accuracy. Examples are provided in Table 3.

The total number of cue requests differed by experience ($F = 7.62$, $df_1 = 2$, $df_2 = 585$, $P < 0.01$): residents requested more information than experienced family

Table 2 Frequencies and Percentages of Diagnostic Accuracy, Appropriateness of Management, and Number of Features of Difficulty per Scenario[a]

	Diagnosis					Management				
	Correct		Incorrect		Features of	Appropriate		Inappropriate		
Scenario	Frequency	%	Frequency	%	Difficulty	Frequency	%	Frequency	%	Total *n*
Fatigue	48	57.1	36	42.9	1	46	54.8	38	45.2	84
Headache	44	52.4	40	47.6	2	44	52.4	40	47.6	84
Dyspnea 2	39	46.4	45	53.6	2	62	73.8	22	26.2	84
Chest pain	37	44	47	56	3	35	41.7	49	58.3	84
Abdominal pain	34	40.5	50	59.5	3	47	56	37	44	84
Dyspnea 1	23	27.4	61	72.6	3	26	31	58	69	84
Pyrexial child, day 8	21	25	63	75	4	44	52.4	40	47.6	84
Total	246	41.8	342	58.2		303	51.5	285	48.5	588
Pyrexial child, day 12	37	44	47	56		64	76.2	20	23.8	

a. Percentages are calculated on the basis of total frequencies in rows.

physicians (means 22.59 v. 19.29) but did not take significantly more time to diagnose ($P = 0.12$). Mean times were 9.30, 8.97, and 8.42 minutes for residents, intermediates, and experienced family physicians, respectively. The number of noncritical cue requests also differed by experience ($F = 9.37$, $df_1 = 2$, $df_2 = 585$, $P < 0.001$), with residents requesting more noncritical cues than both intermediates and experienced family physicians (means 16.33, 14.17, and 13.44, respectively). No effect of experience was detected on the number of critical cue requests ($P = 0.77$).

Appropriateness of Management

Seventy-eight percent of incorrect diagnoses were followed by inappropriate management, and 92% of correct diagnoses were followed by appropriate management. Experience was not a significant predictor of management in the regression model ($P = 0.80$). A highly significant interaction between diagnostic accuracy and scenario predicted appropriateness of management ($P < 0.0001$) and explained 63% of the variation in management (Nagelkerke $R^2 = 0.63$). The positive relationship between diagnostic accuracy and management was significant for all scenarios except dyspnea 2 ($P = 0.14$). In this scenario, incorrect diagnoses of chronic obstructive pulmonary disease exacerbation or left ventricular failure led to appropriate management (referral to pulmonologist, cardiologist, or hospital) most of the time (67%). Any referral to secondary care in this scenario was considered beneficial to the patient, as it was likely to lead to proper assessment and diagnosis of cor pulmonale.

DISCUSSION

This study attempted to identify predictors of diagnostic accuracy on difficult cases in family medicine. It did not aim to gauge the extent of diagnostic error in family medicine and should not be taken as such. Family physicians would not expect to encounter all 7 difficult diagnostic problems in a single day, as the bulk of more routine complaints dominate. We chose scenarios that were difficult, as these increase the likelihood of diagnostic error and are more likely to benefit from future diagnostic interventions. We chose nontrivial diseases, only 1 of them rare, in which diagnostic accuracy matters: the scenario patients could have been harmed to different degrees by diagnostic error or delay.

Diagnostic accuracy was found to be in accordance with the number of features of difficulty designed into the scenarios, making this a valid indicator of difficulty that could be used in building further scenarios. Further work could establish how comprehensive the scheme is, the extent of differentiation between features, and any grading of features in terms difficulty. The systematic, painstaking approach followed during scenario design enabled us to determine evidence-based scoring criteria for information gathering, diagnosis, and management. Previous studies using hypothetical clinical scenarios to assess diagnostic performance determined the correct diagnosis on the basis of the opinion of 1 or a small number of senior doctors[13,38] or on the basis of what most doctors in the study diagnosed.[39] Furthermore, information gathering has not previously been scored against evidence-based criteria.

Table 3 Frequencies and Percentages of Requests for Specific Critical Cues Shown for Correct and Incorrect Diagnoses in 4 Scenarios

Scenario	Critical Cues (Answers to Cue Requests)	Diagnoses, Frequency (%)		*P* Values with Bonferroni Adjustment[a]
		Correct	Incorrect	
Dyspnea 1 (COPD and aortic stenosis)	Systolic murmur radiates to neck	5 (21.7)	0	<0.003
	Echocardiogram (shows aortic stenosis)	22 (95.7)	1 (1.6)	<0.003
	Total diagnoses	23	61	84
Abdominal pain (celiac disease)	Complete blood count (microcytic anemia)	33 (97.1)	21 (42)	<0.006
	Hematinics (consistent with iron deficiency)	12 (35.3)	5 (10)	<0.006
	Serological tests for celiac disease (positive)	33 (97.1)	0	<0.006
	Total diagnoses	34	50	84
Headache (temporal arteritis)	Jaw claudication (present)	10 (22.7)	0	<0.004
	Complete blood count (mild anemia, Hb = 10.8)	27 (61.4)	10 (25)	<0.004
	Examination of the temporal artery (prominent, tender, beaded, nonpulsatile)	35 (79.5)	0	<0.004
	Erythrocyte sedimentation rate (72)	39 (88.6)	3 (7.5)	<0.004
	Total diagnoses	44	40	84
Fatigue (new-onset diabetes)	Polydipsia (present)	27 (56.3)	1 (2.8)	<0.01
	Urinalysis, fingerprick glucose, fasting glucose (positive for diabetes)	48 (100)	0	<0.01
	Total diagnoses	48	36	84

Note: COPD = chronic obstructive pulmonary disease.
a. The *P* values refer to chi-square tests between correct and incorrect diagnoses for the specific cue requests.

No significant effect of experience was detected on diagnostic accuracy, in accordance with other studies.[26,27] Significant effects were detected on information gathering, with residents requesting the most information overall and the most noncritical information. This could be due to a more routinized approach to diagnostic workup, reflecting the residents' training. With practice, a lot of this noncritical information gathering can be bypassed, and clinicians become more efficient in their search.[40] The number of noncritical cue requests did not affect accuracy, and the effect of total cue requests disappeared as soon as critical cues were accounted for—critical cues being the only significant predictor of accuracy. These findings are in contrast to Elstein and colleagues' early findings that misdiagnoses were due to overinterpreting nondiagnostic information, critical cue gathering had no effect, and thoroughness of information gathering had some effect.[13] They used a smaller number of diagnostic problems (3) and doctors (24), which may have masked the importance of critical cue gathering, whereas critical cues were identified by 1 senior doctor. Nevertheless, 'case specificity,' an important finding of Elstein and colleagues, suggesting that diagnostic process and outcome are specific to the case and not generalizable to other cases in different domains, was supported in our study, as the relationship between the number of critical cue requests and diagnostic accuracy varied across scenarios.

The pyrexial child scenario was the most difficult, according to both the number of features of difficulty and rate of misdiagnoses. This was the only scenario in which the number of critical cue requests did not predict diagnostic accuracy. The scenario depicted a rare disease (Kawasaki disease), which family physicians may have forgotten, together with its typical features of presentation (high fever, conjunctivitis without exudate, strawberry tongue, pharyngitis without exudate, erythematous and swollen fingers, maculopapular rash). So even when these features were discovered (some were given in the presenting complaint, e.g., fever, conjunctivitis, rash), they could not be linked to Kawasaki disease. If this was the case, the role of diagnostic hypotheses was perhaps more important in this scenario than the gathering of critical information.

Beyond the general relationship observed between critical cues and accuracy, more detailed analyses per scenario raise questions about the potential role of diagnostic hypotheses in the gathering and interpretation of critical cues. Certain critical cues were requested significantly more frequently by participants who diagnosed correctly, for example, a complete blood count in the abdominal pain scenario (Table 3). Participants who did not request those cues and diagnosed incorrectly may have never considered the correct diagnosis as a hypothesis during the workup. Participants who requested those cues and still diagnosed incorrectly may not have interpreted them appropriately. Alternatively, they may have never considered the correct diagnosis but requested the cues routinely or to test some other hypothesis. Ongoing analyses of the stimulated recall data from the study aim to answer these questions and identify the specific reasons for misdiagnosing each scenario.

No effect of experience was detected on appropriateness of management. This is not surprising, given that management strongly depended on diagnostic accuracy, which did not depend on experience. In a systematic review, most studies (14/19) found that physicians with more years in practice were less likely to adhere to standards of appropriate therapy, such as ordering specific investigations.[28] We assessed management as a single decision or set of decisions at the end of the consultation rather than as adherence to specific therapeutic standards during the consultation. Furthermore, only 1 of our participants was older than 60 y, making age-related performance decline[41] less likely in our sample.

Study limitations relate mainly to the ecological validity of the diagnostic task. First, neither patient photos nor visual information were presented to participants beyond the verbal descriptions. Clinicians do not necessarily notice typical clinical features on patient photographs, and their diagnostic accuracy was found to increase substantially when verbal descriptions were also provided.[42] By eliminating the variability that can result from interpreting visual cues, we were better able to measure and relate information gathering to accuracy. Second, participants were urged to ask specific rather than general "do you have any other symptoms?" types of questions. This too was a result of the need to measure information gathering precisely and identify its influence on accuracy. It is unknown if these manipulations had any influence—positive or negative—on accuracy.

Six of the simulated patients presented only once to the family physician. It could be argued that this does not capture the nature of primary care diagnosis, in which a patient may present several times before a diagnosis is made. The study included both cases in which diagnostic delay can have important consequences (e.g., chest pain) and cases in which this may be less so. Missing celiac disease, aortic

stenosis, cor pulmonale, temporal arteritis, or diabetes at 1st presentation may not harm the patient if the disease is promptly diagnosed and managed at a subsequent consultation. This requires that the patient consults again and that the doctor questions his or her earlier diagnosis. Neither is guaranteed. Patients may be reassured and delay seeking further help. An initial diagnosis is difficult to question later, and both clinicians and patients may interpret the continuing symptoms as a consequence of the diagnosed condition. The human mind has difficulty restructuring problems and devising new solutions. Research in problem solving has shown that people will persistently apply well-learned solution methods to new, similar problems for which they are inappropriate or suboptimal.[43,44] Participants were given a 2nd chance to diagnose correctly in the pyrexial child scenario. On the 3rd presentation of the patient to the family physician, the physicians still misdiagnosed 56% of the time and managed inappropriately (did not send patient to hospital) 24% of the time.

Our findings suggest that having the correct diagnosis in the differentials at the end of the consultation determines appropriate management, both in terms of referral to secondary care and within primary care. Accurate diagnosis is especially important given the current adoption of managed care pathways and dedicated assessment centers for conditions such as chest pain. Of course, there will always be situations resembling dyspnea 2, for which an appropriate referral can result from the wrong diagnosis. This, however, is a chance occurrence.

This study suggests that attempts to improve diagnosis should concentrate on supporting clinicians to ask for critical information. The best way of doing this requires further investigation, especially the role of information technology in prompting clinicians to gather critical information that has been omitted during the diagnostic process.

ACKNOWLEDGMENTS

The research was funded by a grant from the Patient Safety Research Programme, Department of Health, United Kingdom (grant number PS/027). The Primary Care Research Trust for Birmingham supported the costs of the participants' time. The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report. The views and opinions expressed herein do not necessarily reflect those of the Secretary of State for Health. The authors thank the members of the study steering group, Rob Ranyard, Maureen Baker, and Colin Graham, and 3 anonymous reviewers.

REFERENCES

1. The Medical Defence Union. Training and Education: Primary Care Development Programme—Risk Management and Delay in Diagnosis. London: The Medical Defence Union; 2004.

2. Silk N. What went wrong in 1000 negligence claims. Health Care Risk Rep. 2000.

3. Esmail A, Neale G, Elstein M, Firth-Cozens J, Davy C, Vincent C. Case Studies in Litigation: Claims Reviews in Four Specialties. Patient Safety: Lessons from Litigation. Manchester, UK: Manchester Centre for Healthcare Management, University of Manchester; 2004.

4. Gandhi T, Kachalia A, Thomas E, et al. Missed and delayed diagnoses in the ambulatory setting: a study of closed malpractice claims. Ann Intern Med. 2006;145(7):488–96.

5. Graber M, Franklin N, Gordon R. Diagnostic errors in internal medicine. Arch Intern Med. 2005;165:1493–9.

6. Kostopoulou O. From cognition to the system: developing a multilevel taxonomy of patient safety in general practice. Ergonomics. 2006;49(5–6 Special issue: Patient Safety):486–502.

7. Reason J. Human Error. Cambridge, UK: Cambridge University Press; 1990.

8. Patrick J, Gregov A, Halliday P, Handley J, O'Reilly S. Analysing operators' diagnostic reasoning during multiple events. Ergonomics. 1999;42(3):493–515.

9. Simon H. Invariants of human behavior. Ann Rev Psychol. 1990;41:1–19.

10. Fairweather D, Campbell A. Diagnostic-accuracy—the effects of multiple etiology and the degradation of information in old-age. J R Coll Physicians Lond. 1991;25(2):105–10.

11. Klayman J. Varieties of confirmation bias. In: Busemeyer J, Hastie R, Medin D, eds. The Psychology of Learning and Motivation, Vol. 32. San Diego, CA: Academic Press; 1995. p 385–418.

12. Nickerson R. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol. 1998;2(2):175–220.

13. Elstein AS, Shulman L, Sprafka S. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press; 1978.

14. Weber E, Bockenholt U, Hilton D, Wallace B. Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience. J Exp Psychol Learn Mem Cogn. 1993;19(5):1151–64.

15. Chapman G, Bergus G, Elstein A. Order of information affects clinical judgment. J Behav Decis Making. 1996;9(3):201–11.

16. Macdonald S, Macleod U, Campbell NC, Weller D, Mitchell E. Systematic review of factors influencing patient and practitioner delay in diagnosis of upper gastrointestinal cancer. Br J Cancer. 2006;94(9):1272–80.

17. Demesquita P, Gilliam W. Differential-diagnosis of childhood depression—using comorbidity and symptom overlap to generate multiple hypotheses. Child Psychiatry Hum Dev. 1994;24(3):157–72.

18. Goyal S, Roscoe J, Ryder WD, Gattamaneni HR, Eden TO. Symptom interval in young people with bone cancer. Eur J Cancer. 2004;40(15):2280–6.

19. Bouma J, Broer J, Bleeker J, van Sonderen E, Meyboom-de Jong B, DeJongste MJ. Longer pre-hospital delay in acute myocardial infarction in women because of longer doctor decision time. J Epidemiol Community Health. 1999;53(8):459–64.

20. Thornicroft G, Rose D, Kassam A. Discrimination in health care against people with mental illness. Int Rev Psychiatry. 2007;19(2):113–22.

21. Jopp DA, Keys CB. Diagnostic overshadowing reviewed and reconsidered. Am J Ment Retard. 2001;106(5):416–33.

22. White MJ, Nichols CN, Cook RS, Spengler PM, Walker BS, Look KK. Diagnostic overshadowing and mental-retardation—a meta-analysis. Am J Ment Retard. 1995;100(3):293–8.

23. Charlin B, Tardif J, Boshuizen H. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med. 2000;75(2):182–90.

24. Boshuizen H. Does practice make perfect? A slow and discontinuous process. In: Boshuizen H, Bromme R, Gruber H, eds. Professional Learning: Gaps and Transitions on the Way from Novice to Expert. Dordrecht, the Netherlands: Kluwer Academic; 2004. p 73–95.

25. Bordage G, Grant J, Marsden P. Quantitative assessment of diagnostic ability. Med Educ. 1990;24:413–25.

26. Fasoli A, Lucchelli S, Fasoli R. The role of clinical "experience" in diagnostic performance. Med Decis Making. 1998;18(2): 163–7.

27. Hauser S, Spada H, Rummel N. The effects of practical experience on expertise in clinical psychology and collaboration. Proceedings of the 29th Meeting of the Cognitive Science Society; Nashville, TN; August 2007.

28. Choudhry N, Fletcher R, Soumerai S. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med. 2005;142(4):260–73.

29. Gruppen L, Wolf F, Billi J. Information gathering and integration as sources of error in diagnostic decision-making. Med Decis Making. 1991;11(4):233–9.

30. Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. JAMA. 2000;283(13):1715–22.

31. Peabody J, Luck J, Glassman P, et al. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. Ann Intern Med. 2004;141(10):771–80.

32. Berner E, Webster GD, Shugerman AA, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994; 330(25):1792–6.

33. Bordage G. Diagnostic errors: poor reasoning habits or ill-structured knowledge. In: Schmidt H, De Volder M, eds. Tutorials in Problem-Based Learning: A New Direction in Teaching the Health Professions. Assen, the Netherlands: Van Gorcum; 1984. p 158–66.

34. Huber O, Wider R, Huber O. Active information search and complete information presentation in naturalistic risky decision tasks. Acta Psychol. 1997;95:15–29.

35. Williamson J, Ranyard R, Cuthbert L. A conversation-based process tracing method for use with naturalistic decisions: an evaluation study. Br J Psychol. 2000;91(pt 2):203–21.

36. Lyle J. Stimulated recall: a report on its use in naturalistic research. Br Educ Res J. 2003;29(6):861–78.

37. Newburger JW, Takahashi M, Gerber MA, et al. Diagnosis, treatment, and long-term management of Kawasaki disease—a statement for health professionals from the Committee on Rheumatic Fever, Endocarditis and Kawasaki Disease, Council on Cardiovascular Disease in the Young, American Heart Association. Circulation. 2004;110(17):2747–771.

38. Friedman CP, Elstein AS, Wolf FM, et al. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems. JAMA. 1999;282(19):1851–6.

39. Barrows H, Norman G, Neufeld V, Feightner J. The clinical reasoning of randomly selected physicians in general medical practice. Clin Invest Med. 1982;5(1):49–55.

40. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999;74(10):1129–34.

41. Eva K. The aging physician: changes in cognitive processing and their impact on medical practice. Acad Med. 2002;77(10): S1–S6.

42. Brooks L, LeBlanc V, Norman G. On the difficulty of noticing obvious features in patient appearance. Psychol Sci. 2000;11(2): 112–7.

43. Luchins A. Mechanization in problem solving: the effect of Einstellung. Psychol Monogr. 1942;54:248.

44. Cherubini P, Mazzocco A. From models to rules: mechanization of reasoning as a way to cope with cognitive overloading in combinatorial problems. Acta Psychol. 2004;116(3):223–43.