Correlates of Diagnostic Accuracy in Patients with Nonspecific Complaints

Ralph Hertwig, PhD, Nathalie Meier, MSc, Christian Nickel, MD, Pia-Cristina Zimmermann, MD, Selina Ackermann, MSc, Jan K. Woike, PhD, Roland Bingisser, MD

Objective. *To investigate diagnostic accuracy in patient histories involving nonspecific complaints and the extent to which characteristics of physicians and structural properties of patient histories are associated with accuracy.* ***Methods***. *Six histories of patients presenting to the emergency department (ED) with nonspecific complaints were provided to 112 physicians: 36 ED physicians, 50 internists, and 26 family practitioners. Physicians listed the 3 most likely diagnoses for each history and indicated which cue(s) they considered crucial. Four weeks later, a subset of 20 physicians diagnosed the same 6 histories again. For each history, experts had previously determined the correct diagnoses and the diagnostic cues.* ***Results***. *Accuracy ranged from 14% to 64% correct diagnoses (correct diagnosis listed as the most likely) and from 29% to 87% correct differential diagnoses (correct diagnosis listed in the differential). Acute care physicians (ED physicians and internists) included the correct diagnosis in the differential in, on average, 3.4 histories, relative to 2.6 for the family practitioners (*P = *0.001,* d = *.75). Diagnostic performance was fairly reliable (r = .61,* P < *0.001). Clinical experience was negatively correlated with diagnostic accuracy (*r = *–.25,* P = *0.008). Two structural properties of patient histories—cue consensus and cue substitutability—were significantly associated with diagnostic accuracy, whereas case difficulty was not. Finally, prevalence of diagnosis also proved significantly correlated with accuracy.* ***Conclusions***. *Average diagnostic accuracy in cases with nonspecific complaints far exceeds chance performance, and accuracy varies with medical specialty. Analyzing cue properties in patient histories can help shed light on determinants of diagnostic performance and thus suggest ways to enhance physicians' ability to accurately diagnose cases with nonspecific complaints.* ***Key words:*** *nonspecific complaints; diagnostic decision making; experience; emergency department physicians; internists; family practitioners.* ***(Med Decis Making 2013;33:533–543)***

Patients presenting to the emergency department (ED) with nonspecific complaints, such as weakness, fatigue, or dizziness, pose a challenge to emergency physicians' diagnostic decision-making process. For instance, researchers involved in the Basel Non-Specific Complaints (BANC) Study[1] observed in unpublished data that in the ED, the misdiagnosis rate in cases involving nonspecific complaints is about 53%, relative to an overall rate of less than 10%. This high rate of errors matters because nonspecific complaints can be associated with life-threatening conditions that require prompt intervention to prevent further deterioration of the patient's health status.[1] Moreover, according to a large study, up to 20% of elderly patients presenting to the ED report nonspecific complaints.[2]

A key component in the process of diagnosing patients with nonspecific complaints is the patient history.[3] The information encapsulated therein guides the diagnostician's initial decision-making process. To investigate the properties of patient histories that affect diagnosticians' judgment, we presented original patient histories, as recorded by the admitting emergency physician,[4] to physicians with various medical specialties. We aimed to investigate 3 questions: First, is diagnosis of nonspecific complaints presenting at the ED better than chance? Second, does diagnostic accuracy relate to physicians'

Received 29 August 2011 from Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany (RH, NM, JKW), and Department of Emergency Medicine, University Hospital, Basel, Switzerland (CN, PZ, SA, RB). Financial support from the Scientific Fund of the Emergency Department, University Hospital, Basel, Switzerland. Revision accepted for publication 25 October 2012.

Address correspondence to Ralph Hertwig, Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany; e-mail: hertwig@mpib-berlin.mpg.de.

DOI: 10.1177/0272989X12470975

specialty and other physician characteristics? Third, what structural properties of the clinical case determine diagnostic accuracy?

THE PROBABILISTIC NATURE OF DIAGNOSTIC INFERENCE

To appreciate the importance of the structural properties of patient histories, consider the following conceptualization of diagnostic inference. Much of human perception and cognition can be understood as a probabilistic inference process.[5] For instance, a twitching foot might commonly suggest that a person is nervous, yet this cue can be uninformative or, worse, misleading because people sometimes twitch their feet for other reasons (e.g., because they are excited) or for no particular reason at all.[6] Because cognition and perception are probabilistic and based on imperfect cues, there is a natural limit to how accurate they can be. Inevitable though errors may be, they do not reflect a failure of the inferential system but a probabilistic environment that is not perfectly predictable from the available cues.[7]

Diagnostic decision making can also be modeled as a probabilistic inference. By extension, nonspecific complaints such as feeling dizzy or fatigued can be thought of as probabilistic cues, except that their predictive accuracy—that is, the relationship between the cue (symptom) and the target (correct diagnosis)—is likely to be weaker than that between specific symptoms and the correct diagnosis. The reason is that a nonspecific symptom such as fatigue is likely to arise in a wider spectrum of diagnoses than, for instance, chest pain. Consequently, the natural upper limit on the accuracy of diagnostic inferences involving nonspecific complaints is likely to be lower than that in specific complaints.

Taking as our starting point the probabilistic nature of diagnostic inference, we analyzed 4 structural properties of patient histories: positive cue validity, negative cue validity, cue consensus, and cue substitutability. Each piece of information in a patient history (henceforth cue) has 2 basic important characteristics: its *positive* and *negative validity*. The positive validity of a cue refers to its ability to predict the criterion (here the correct diagnosis). There exist various definitions of *positive cue validity*.[8] We define it pragmatically as follows: the number of times that a cue was identified as crucial by physicians who diagnosed the case correctly, divided by the total number of times this cue was identified as crucial. By extension, negative cue validity is as follows: the number of times that a cue was not identified as crucial by physicians who misdiagnosed the case, divided by the total number of times this cue was not identified as crucial.

Another property of a cue—cue consensus—refers to its ability to attract physicians' consensual endorsement. In many situations, knowledge that is shared by many people corresponds by and large to the truth.[9] Similarly, a cue that is identified as crucial by most physicians may also be more likely a valid cue than a cue identified as crucial by merely a few physicians. Common knowledge, however, does not always track truth; sometimes the majority of people get it wrong. By investigating cue consensus, we can find out whether the domain of nonspecific complaints is one in which common knowledge tracks truth (''kind environment'') or fails to track truth (''wicked environment'').[10] Cue consensus is defined as the number of physicians who selected a given cue as being crucial for their diagnosis divided by the total number of physicians.

Finally, cue substitutability (or vicarious functioning[11]) refers to the fact that different physicians can arrive at the same diagnostic judgment by using different subsets of cues (symptoms, clinical findings, etc.) or by attributing different degrees of importance to the same cues in a patient history. For instance, in a study of diagnosing streptococcal pharyngitis, some physicians based the diagnosis almost entirely on whether the patient had a fever and an inflamed throat, whereas others made no use of these symptoms and instead relied on swollen tonsils and lack of cough.[12] This and related observations[13,14] suggest that medical problems differ in the degree to which they provide interchangeable paths to the correct diagnosis. Metaphorically speaking, although not all roads lead to Rome, there may be more than one road that takes one there. Thus, cue substitutability—defined as the percentage of physicians who arrived at the correct diagnosis without considering any (or only some) of the *diagnostic* cues (as predefined by the consensus judgment of 2 experts) to be crucial—has the potential to foster diagnostic accuracy.

METHOD

Definition

The diagnostic value of a symptom diminishes with the number of its potential interpretations. Thus, a poorly defined symptom has little discriminative power in establishing a medical diagnosis, and if physicians are uncertain about the exact nature of

a symptom, they must take into account multiple competing interpretations of the same set of complaints.[15]

Material

We used 7 patient histories, 6 with nonspecific symptoms (target histories) and 1 with specific symptoms (control history). The selection of histories was made in 2 steps: Based on an analysis of a sample of 1210 patients with nonspecific complaints presenting to the ED at the University Hospital of Basel (Switzerland), we estimated the prevalence of each final diagnosis. Across this sample, several dozen diagnoses were observed, but 12 diagnostic groups accounted for more than 50% of all patients. Of those 12, we selected 5 diagnostic groups that accounted for 32% of the total prevalence: urinary tract infection, pneumonia, congestive heart failure, frailty, and valium intoxication. Including more target diagnoses would have overtaxed participating physicians' precious time. Having thus identified the target diagnoses, we next turned to the 686 original patient histories from the BANC-cohort database[1] and selected 6 histories representing the 5 target diagnoses. The diagnosis of congestive heart failure was represented by 2 patient histories, whereas each of the other diagnoses was represented by 1 history. Each history was identical to the original, electronically stored patient history and was presented in written form to the participating physicians. In addition, the history of a patient presenting with a specific symptom—namely, chest pain (with a final diagnosis of myocardial infarction)—was included. This control history was less demanding than the target histories, allowing us to gauge participants' level of motivation. All histories, translated from the original German, are listed in Table 1.

For each history, 2 physicians certified in internal medicine had determined the final diagnosis based on written 30-day follow-up data from the presenting patients' primary care physicians and hospital discharge reports.[4] These experts—selected for their extensive experience in emergency medicine (>10 years) and their involvement as outcome evaluators in the follow-up of 1210 case histories with nonspecific complaints—also identified the *diagnostic* cues for the correct diagnosis. The diagnostic cues are the pieces of information in the patient history that, according to these experts, are indicative of the correct diagnosis. The experts first determined the diagnostic cues independently and then resolved their judgment differences in a joint discussion. The final sets of diagnostic cues also conform to those reported in established emergency medicine textbooks (e.g., Tintinalli and others,[16] pp. 345, 608, 366, 448, and 1904).

Participants

We advertised the study within the University Hospital of Basel, through the Swiss Society for Emergency Medicine, and through an existing network of local family practitioners. A total of 112 physicians (66 male and 46 female) participated. Physicians received a small token of appreciation (a 25% chance to win a gift certificate worth 20 Swiss francs [about $23]). They were also offered feedback regarding the study's aggregate results.

Study Procedure

ED physicians and internists completed the questionnaires in the hospital. Family practitioners completed it off site and returned it by mail. Participants were informed that the goal was to investigate diagnostic inference in patients with nonspecific complaints and were assured that their data would be anonymized. Four different randomized presentation orders of the patient histories were created. For each history, physicians wrote down 1) what they believed to be the three most likely diagnoses (i.e., the differential diagnoses), ranked according to their likelihood, and 2) the cues they considered *crucial* (separately for each of the 3 most likely diagnoses). The crucial cues for the most likely diagnosis were extracted and entered in a spreadsheet. Other aspects of the histories and all cues for the differential diagnoses were also recorded but are not included in the following analyses.

We calculated cue consensus, positive cue validity, negative cue validity, and cue substitutability. When analyzing cue substitutability, we also examined the extent to which physicians relied on diagnostic cues and how their reliance determined accuracy. We recorded physicians' age, sex, specialty, years of clinical experience, involvement in research, board certification, and years spent working in internal medicine and in emergency medicine. On average, physicians took about 45 minutes to complete the questionnaire.

An average of 4 weeks after participating in the initial study, a randomly selected subset of 20 participants was asked again to diagnose the control history and the 6 target histories. In this retest, a random presentation order of the histories was generated for each participant.

Table 1 Six Patient Histories Involving Nonspecific Complaints and 1 Control History Involving Specific Complaints

Final Diagnosis (Prevalence)	Patient History
Urinary tract infection (9%)	An 88-year-old female patient living in a senior residence, normally very active and at times expressing discontent, had been subdued for several days. Furthermore, she had almost stopped eating, and when she ate, she frequently choked. For this reason, her medication (pipamperone and risperidone) was stopped. In addition, she had intermittent low-grade fever. No cough, no headache, no abdominal pain. At night, restless and often wandering in the hallway. Day/night reversal.
Congestive heart failure 1 (6%)	A 74-year-old female patient collapsed because of weakness in both legs. She did not lose consciousness but was unable to get up from the floor. The daughter found her mother lying on the floor and brought her to the emergency department. No shortness of breath. Recently, her thirst had increased and she had become more tired. She also noticed pitting edema, despite taking torsemide. She had not checked her weight recently. Chest discomfort once daily, duration of a few minutes, no radiation. According to the daughter, the patient had been suffering from the flu with a cough recently. Her general practitioner had prescribed 5 days of moxifloxacin. Furthermore, the patient had difficulty walking. However, she has no problems moving about in her own home.
Pneumonia (6%)	A 78-year-old male patient was brought to the emergency department by ambulance, referred by his general practitioner. Since the previous weekend, the patient noticed an increasing weakness in his legs. Furthermore, frequent hiccups. Lives at home with his wife, who is in Germany at present, and with a female caretaker. Intake of medication this morning not certain. No pains. History of cough last weekend, but not right now. No expectoration.
Frailty (7%)	Referral of an 84-year-old female patient by the general practitioner because of progressive decline in mobility with virtual immobility despite fully developed help from Spitex[a] and devoted help from her husband (caring for his wife for 30 years). According to the husband, his wife's spastic paralysis had worsened during the previous 2 weeks. Preexisting hemi-syndrome on the right side following a stroke in 1978. After 30 years, he is no longer able to provide care for her on his own, despite the help from Spitex. According to her husband, no pains, no falls, no shortness of breath. Previous history: arterial hypertension; stroke with right-sided paralysis in 1978.
Valium intoxication (4%)	Self-referral of a 55-year-old female patient because of tiredness and general weakness, first onset 6 months ago. Moreover, her mouth is very dry and she fears suffocation. The patient reports that she has trouble with her neighbors about a clothesline, which is located between the two properties. At the beginning, diazepam tablets helped to resolve the stress, but now she feels tired. For some time, she has had a prescription for medications, whose names she is unable to remember. No previous illness is known.
Congestive heart failure 2 (6%)	Referral of a 78-year-old female patient by a general practitioner because of deterioration of general health. She does not know exactly why she is in the emergency department. Her husband reports that she has been getting worse over the past 2 months. She wants neither to eat nor to drink. Her husband has noticed a gait disturbance. She sleeps a lot and is generally tired. She is unable to recall how much weight she has lost. Occasional diarrhea, occasional cough, dyspnea on exertion, feels depressed, no pain, no falls. The husband (homeopathic practitioner) reports a steady deterioration over the past 10 years. Loss of weight.
Myocardial infarction (control history)	Admission by ambulance. The 66-year-old male patient reports left-sided chest pain with onset 30 minutes previously. Visual analog scale (VAS) pain intensity of 8/10, of burning nature, radiating to the jaw and left upper arm. Pain is now tolerable lying on the stretcher. Feeling of impending doom.

[a]Spitex is an organization in Switzerland providing home care, nursing, and general help for patients and their caretakers.

RESULTS

Demographics

Of the 112 physicians, 36, 50, and 26 were emergency physicians, internists, and family practitioners, respectively. Their average age was 41 years, their average clinical experience was 13 years, 26% were involved in clinical research, 66% were board-certified specialists, and their average postgraduate experience in hospital-based internal

medicine and emergency medicine was 4.1 and 1.3 years, respectively.

Diagnostic Accuracy

Two measures of diagnostic accuracy were employed—namely, how often the correct diagnosis was listed as the most likely one (*correct diagnosis*) and how often the correct diagnosis was listed in the differential (*correct differential diagnosis*). All but one of the physicians (99% of the sample) correctly diagnosed the control problem, suggesting that they were motivated. Because the physician who failed to solve the control problem correctly diagnosed 4 of 6 nonspecific histories, we did not exclude this physician from further analyses.

Table 2 reports the 2 measures of accuracy across the 6 patient histories involving nonspecific complaints. The percentage of correct diagnoses ranged from 14% to 64%, with an average of 34%. The percentage of correct differential diagnoses ranged from 29% to 87%, with an average of 53%. The difference between the percentage of correct diagnoses and the percentage of correct differential diagnoses for each history ranged from 11% (frailty) to 31% (pneumonia).

Attributes of Physicians Associated with Diagnostic Accuracy

Given the scarcity of current knowledge and the fact that, to the best of our knowledge, the present study is the first to investigate diagnostic accuracy in patient histories with nonspecific complaints, our goal was to generate hypotheses rather than to test existing ones (because there are none). First, we assessed performance differences as a function of medical specialty. As Table 3 shows, we found an almost identical level of performance for emergency physicians and internists on both measures of accuracy. We therefore collapsed them into 1 group, which we henceforth refer to as *acute care physicians*. Acute care physicians' average number of correct diagnoses (2.2 out of 6) was higher than that of family practitioners (1.5; Δ = .64; 95% confidence interval [CI], 1.1–0.14; t = 2.5, df = 110, P = 0.01). This difference corresponds to d = .60 (standardized difference) and represents a medium to large effect (d = .2, .5, and .8 represent effects of small, medium, and large size, respectively).[17] The same pattern emerged on the second measure of accuracy: Acute care physicians' average number of correct differential diagnoses (3.4) exceeded that of family

Table 2 Percentage (Frequencies) of Correct Diagnoses and Average Values (in Percentages) for Cue Consensus, Positive Cue Validity, and Negative Cue Validity, Separately for Patient Histories and Separately for No, Medium, and Full Reliance on Diagnostic Cues (Cue Substitutability; See Text)

Patient Histories	Correct Diagnosis Given as the Most Likely	Correct Diagnosis Given in the Differential	Average Positive Cue Validity	Average Negative Cue Validity	Average Cue Consensus	Reliance on Diagnostic Cues		
						No	Medium	Full
Urinary tract infection	64 (72)	87 (97)	66	39	33	36[a]	74[b]	83[b] (73 v. 100)
Congestive heart failure 1	44 (49)	70 (78)	46	59	28	15[a]	64[b]	70[b] (67 v. 100)
Pneumonia	32 (36)	63 (71)	35	74	31	8[a]		79[b] (77 v. 88)
Frailty	30 (34)	41 (46)	36	65	32	20[a]		68[b] (40 v. 75)
Valium intoxication	18 (20)	30 (34)	22	83	26	19[a]	13[a]	36[a] (—)
Congestive heart failure 2	14 (16)	29 (32)	19	86	22	3[a]		77[b] (67 v. 100)
Myocardial infarction (control history)	99 (111)	99 (111)	100	2	58	100	99	100 (—)

Different superscripts ([a,b]) denote that a test of difference between proportions (2-proportion z test) found a significant difference between 2 groups (i.e., $P < 0.01$ after Bonferroni correction). Consider, for instance, the urinary tract infection history: 36% correct diagnoses is statistically different from 74%, but the latter value is not statistically different from 83% correct diagnoses. The numbers in parentheses in the "full" column represent the percentage of correct diagnoses for physicians considering both the diagnostic cues and others to be crucial, as well as those who exclusively consider the diagnostic cues to be crucial.

Table 3 Correct (Differential) Diagnoses across 6 Patient Histories with Nonspecific Complaints

	Correct Diagnosis			Correct Differential Diagnosis		
	Mean[a]	SD	95% CI	Mean[a]	SD	95% CI
ED physicians (n = 36)	2.2	1.1	1.8–2.6	3.4	1.3	2.9–3.8
Internists (n = 50)	2.1	1.2	1.8–2.4	3.4	1.0	3.2–3.7
Family practitioners (n = 26)	1.5	1.1	1.1–2.0	2.6	1.1	2.1–3.0

CI, confidence interval; ED, emergency department.
[a]The highest possible level of accuracy is 6.

practitioners (2.6; Δ = .80; 95% CI, 1.3–0.31; t = 3.3, df = 110, P = 0.001; d = .75).

There was also substantial variability in diagnostic accuracy within each medical specialty, with 3% of acute care physicians and 12% of family practitioners providing only 1 or no correct differential diagnosis (out of a possible 6) and 47% of acute care physicians and 23% of family practitioners providing 4 or 5 correct differential diagnoses. Finally, we found that across all 112 participants, 3 of the physicians' attributes correlated negatively with diagnostic accuracy: clinical experience (r = –.25, P = 0.007), board certification (r = –.22, P = 0.02), and age (Spearman rank correlation r = –.29, P = 0.002). Relatedly, practitioners were, on average, significantly older than acute care physicians (54.9 v. 36.8 years; Δ = 18.1; 95% CI, 14.9–21.2; t = 11.3, df = 110, $P < 0.001$; d = 6.6). Finally, we also found that the retest scores of the randomly selected subset of 20 physicians were correlated with their initial score (r = .61, $P <$ 0.001); that is, diagnostic performance was not a matter of chance.

Attributes of Patient Histories Associated with Diagnostic Accuracy

Beyond physician attributes, structural properties of patient histories may also account for diagnostic accuracy. We analyzed 4: positive cue validity, negative cue validity, cue consensus, and cue substitutability. Table 4 reports for each patient history the diagnostic cues (as predefined by the consensual judgment of 2 experts), the crucial cues (as chosen by at least 5% of participants), cue consensus, and the cues' positive and negative validities. The aggregated values are reported in Table 2. Mean positive cue validity ranged from 19% (congestive heart failure 2) to 66% (urinary tract infection) compared with 100% for the control history. Mean negative cue validity ranged from 39% (urinary tract infection) to 86% (congestive heart failure 2) compared with 2% for the control history. Table 2 shows that these average values are aligned with the number of correct diagnoses, which is not surprising given that this quantity is part of the definition of cue validity. Cue validities are still informative, however, as they tell us which cues (diagnostic and nondiagnostic) provide interchangeable paths to correct diagnosis.

As the number of correct diagnoses does not affect the definition of cue consensus, we also investigated whether the consensual endorsement of specific cues is predictive of accuracy. As Table 2 shows, average cue consensus ranged from 22% (congestive heart failure 2) to 33% (urinary tract infection) and averaged 28%. We observed a marginally significant correlation between average cue consensus and percentage of correct diagnoses across the 6 target histories (r = .74, P = 0.09). That is, the more physicians agreed on which cues are crucial, the more likely the problem was to be correctly diagnosed.

In terms of cue substitutability, we investigated the extent to which exclusive reliance on the diagnostic cues (identified by the experts) is necessary to arrive at the correct diagnosis or whether diagnosticians can make use of other cues and still diagnose accurately. Half of our target histories included 1 diagnostic cue, the other half 2 diagnostic cues. Table 2 reports the percentage of correct diagnoses separately for ''no reliance'' on these diagnostic cues (i.e., physicians failed to consider the diagnostic cue(s) to be crucial), ''medium reliance'' (i.e., physicians considered 1 of the 2 diagnostic cues to be crucial), and ''full reliance'' (i.e., physicians considered the diagnostic cue(s) to be crucial). Several results are noteworthy. First, the more physicians relied on diagnostic cues, the better, on average, was diagnostic accuracy (in 5 of the 6 histories, accuracy is significantly higher for full than for no reliance). Second, some histories were ''unforgiving'' when physicians failed to identify the diagnostic cue(s)—namely, the histories of congestive heart failure 2 and pneumonia (no reliance: 3% and 8% correct diagnoses, respectively). In contrast, the history of urinary tract infection allowed 36% of the physicians to arrive at correct

Table 4 Properties of Patient Histories Associated with Diagnostic Accuracy (See Text for Definitions)

Patient History/Correct Diagnosis	Cues[a]	Cue Consensus[b]	Positive Cue Validity[c]	Negative Cue Validity[c]	Most Frequently Named Wrong Diagnosis
Myocardial infarction (control history): 99%	**Left-sided chest pain**	82 (73%)	100 (82/82)	3.3 (1/30)	—
	Character of pain	79 (71%)	99 (78/79)	0 (0/33)	
	Fear of death	33 (29%)	100 (33/33)	1.3 (1/79)	
Urinary tract infection: 64%	**Fever**	75 (67%)	76 (57/75)	59 (22/37)	Pneumonia, cerebrovascular disease
	Choking	35 (31%)	20 (7/35)	16 (12/77)	
	Behavioral change (calmer)	32 (29%)	72 (23/32)	39 (31/80)	
	Restlessness	22 (20%)	82 (18/22)	40 (36/90)	
	Loss of appetite	22 (20%)	82 (18/22)	40 (36/90)	
Congestive heart failure 1: 44%	**Leg edema**	47 (42%)	85 (40/47)	86 (56/65)	Pneumonia, diabetes
	Weakness	47 (42%)	40 (19/47)	54 (35/65)	
	Severe flu	31 (28%)	13 (4/31)	44 (36/81)	
	Chest discomfort	28 (25%)	82 (23/28)	69 (58/84)	
	Coughing	28 (25%)	32 (9/28)	52 (44/84)	
	Increased thirst	24 (21%)	13 (3/24)	48 (42/88)	
	More tired	14 (13%)	57 (8/14)	58 (57/98)	
Pneumonia: 32%	Increasing weakness	55 (49%)	47 (26/55)	82 (47/57)	Cerebrovascular disease, neurological disease
	Coughing	38 (34%)	79 (30/38)	92 (68/74)	
	Hiccupping	25 (22%)	8 (2/25)	61 (53/87)	
	Leg weakness	22 (20%)	5 (1/22)	61 (55/90)	
Frailty: 30%	Rapid decline in past 2 weeks	59 (53%)	7 (4/59)	43 (23/53)	Cerebrovascular disease, infection
	Continuous decline past 30 years	25 (22%)	68 (17/25)	80 (70/87)	
	History of stroke	24 (21%)	33 (8/24)	70 (62/88)	
Valium intoxication: 18%	**Fatigue**	45 (40%)	16 (7/45)	81 (54/67)	Psychiatric disorder (depression, anxiety disorder), hypothyroidism
	Trouble with neighbors	35 (31%)	3 (1/35)	75 (58/77)	
	Very dry mouth	30 (27%)	40 (12/30)	90 (74/82)	
	Fear	29 (26%)	7 (2/29)	78 (65/83)	
	Intake of Valium	25 (22%)	28 (7/25)	85 (74/87)	
	New medication	19 (17%)	47 (9/19)	88 (82/93)	
	General weakness	17 (15%)	12 (2/17)	81 (77/95)	
Congestive heart failure 2: 14%	Deterioration of general health	52 (46%)	10 (5/52)	82 (49/60)	Dementia, malignancy
	Weight loss	26 (23%)	0 (0/26)	81 (70/86)	
	Does not know why in ED	19 (17%)	0 (0/19)	83 (77/93)	
	Inappetence	18 (16%)	11 (2/18)	85 (80/94)	
	Exertional dyspnea	17 (15%)	76 (13/17)	97 (92/95)	
	Very tired	16 (14%)	19 (3/16)	86 (83/96)	

[a]Cues identified as crucial by at least 5% of physicians. The cues in boldface represent the diagnostic ones according to consensus of 2 experts (see text).
[b]Absolute frequencies and percentages in parentheses.
[c]Percentage and absolute frequencies in parentheses.

diagnoses even if they made no use of the diagnostic cues. The cues in this history that provide interchangeable paths to the correct diagnosis are, for instance, loss of appetite and behavioral change (cues with high positive validity; Table 4). In contrast, the cues in the history of pneumonia, for instance, such as hiccupping and leg weakness, point to wrong diagnoses such as cerebrovascular and other neurological disease (note these cues' negative cue validity; Table 4). Third, some histories remained difficult even when the physicians identified all diagnostic cues. In the history of Valium intoxication, for instance, physicians who relied on both diagnostic cues (including the cue ''intake of Valium'') attained a 36% level of accuracy (Table 2). The reason is likely to be that in combination with symptoms such as fear and trouble with neighbors, Valium intake colludes to indicate a psychiatric disorder (Table 4; most frequently named wrong diagnosis).

Finally, when physicians reported all diagnostic cues to be crucial, their average performance was only 69% (Table 2). Why? This group includes 2 groups of diagnosticians, one that considered only the diagnostic cues to be crucial and another that considered both the diagnostic cues and the other cues to be crucial. The former group reached an average performance of 93%, the latter 65%; in each of the 5 histories in which the performance of these 2 groups differed, the former group achieved higher accuracy ($P = 0.03$, exact binomial test). In other words, the key to diagnostic performance in histories with nonspecific complaints is not just the ability to identify all diagnostic cues but also the ability to discard other cues (although sometimes there are interchangeable paths to the correct diagnosis).

Beyond the structural properties analyzed, other aspects may influence diagnostic accuracy. Therefore, we investigated 2 additional aspects: difficulty and disease prevalence. Specifically, we asked a group of 15 experts—ED physicians with daily exposure to patients with nonspecific symptoms and average experience of 8 years in the ED—to judge the diagnostic difficulty of our 6 target patient histories. Their judgments were uncorrelated with the percentage of correct diagnoses ($r = -.44$, $P = 0.38$), cue consensus ($r = .04$, $P = 0.94$), and cue substitutability ($r = -.71$, $P = 0.11$), respectively. Furthermore, their judgment of difficulty was not significantly correlated with how frequently they thought the respective diagnostic groups presented to the ED ($r = -.42$, $P = 0.41$). In contrast, disease prevalence (in the ''Material'' section, we describe how we arrived at prevalence) proved to be strongly associated with accuracy ($r = .82$, $P = 0.05$). Yet, one should not overrate this association as it is strongly influenced by the patient history of urinary tract infection, which was diagnosed accurately more frequently that any other patient history. Among our set of histories, it was also the most prevalent one. Once this history is removed, the correlation drops to $r = .45$ ($P = 0.45$).

DISCUSSION

Because of the weaker relationship between nonspecific complaints and diagnoses relative to that between specific complaints and diagnoses, the former represents an objectively difficult-to-predict environment.[18] Histories with nonspecific complaints proved to be substantially more difficult to diagnose than a control history. Yet the patient histories with nonspecific complaints were not invariably difficult to diagnose. We observed large variability, with some histories being correctly diagnosed by a majority and others by only few physicians (for a similar finding, see Funder[6]). A history of urinary tract infection, for instance, was correctly diagnosed by 64% of physicians, and 87% of physicians included this diagnosis in their differential diagnoses. About a third of physicians (30%) correctly diagnosed frailty (prevalence of 7%), and 41% included the correct diagnosis in their differential diagnoses. Even for the most difficult patient history, congestive heart failure 2, 14% of physicians gave the correct diagnosis and 29% included it in their differential diagnoses.

In our prevalence analysis of 1210 case histories with nonspecific complaints (see ''Material''), congestive heart failure proved to be only slightly less frequent (6%) than urinary tract infection (9%), with the latter being the most prevalent diagnosis overall. This suggests that the difficulty of a patient history cannot be simply reduced to the diagnosis' prevalence. A simple base-rate strategy (i.e., always predict the most prevalent diagnosis), for instance, would be wrong most of the time.

The level of performance we observed suggests that correctly diagnosing nonspecific complaints is not out of reach. Yet, it clearly is not a trivial task either. Across cases with nonspecific complaints, hundreds of diagnoses can be observed,[19,20] and in a previous study,[4] the diagnostic spectrum in this presentation extended over 16 chapters of the *International Classification of Diseases, Tenth Revision* (*ICD-10*). Finally, we also found that good performance was not a matter of luck. If it were, physicians' diagnostic reliability would be nil. In contrast, we

observed a retest reliability of $r = .61$ in the subset of physicians who diagnosed the same set of histories again about 4 weeks after the initial study.

These findings raise the question of what properties of physicians and patient histories can explain diagnostic accuracy.

What Physician Properties Foster Accurate Diagnostic Performance?

Among physician properties, medical specialty proved to be indicative of diagnostic accuracy. Family practitioners' diagnostic performance was significantly lower than that of emergency physicians and hospital internists (Table 3). One possible explanation for this difference (of medium to large size[17]) is that family practitioners work in a medical environment in which they are less likely to be exposed to the kind of cases that are ultimately admitted to the hospital via the emergency department. Furthermore, family practitioners were, on average, significantly older than acute care physicians, and so their training may be less up-to-date than that of acute care physicians; indeed, across all physicians, we observed a negative correlation of accuracy with age. Importantly, it deserves to be pointed out that the patient histories were originally collected in an emergency department, by emergency physicians, and were adjudicated by emergency physicians (our experts). Therefore, the experimental design may have favored acute care physicians. Variation in performance between the medical specialties might have turned out quite differently if cases had been sampled from the population of patient histories involving nonspecific complaints that family practitioners typically experience.

We also observed that clinical experience (and board certification) proved to be negatively correlated with diagnostic performance. Intuitively, one might have expected the opposite—namely, that clinical experience (and thus learning opportunities) with cases of nonspecific complaints would allow diagnosticians to practice and fine-tune their skills. However, there is evidence that diagnostic accuracy does not necessarily improve with clinical experience,[3,21,22] and clinical experience may be even inversely correlated with quality of health care.[23] It could be that a physician's illness scripts, built up during training, are not sufficiently updated by later experience. One reason for insufficient updating might be a learning environment that is not conducive to accurate learning.[24] Specifically, due to the extremely heterogeneous diagnostic spectrum in this presentation,[4] the *n*s per diagnosis (and related outcome feedback) experienced by even a seasoned physician may simply be too small for him or her to hone his or her craft. But this explanation is speculative and needs to be explored further (as does the robustness of the observed negative correlation between clinical experience and diagnostic performance).

Cue Consensus and Cue Substitutability: Two Properties Correlated with Accuracy

Both cue consensus and cue substitutability were correlated with diagnostic accuracy. Cue consensus need not be associated with accuracy. Take, for instance, the patient history in which the correct final diagnosis is frailty. The cue that most physicians considered crucial was ''rapid decline in past 2 weeks''—this popular cue, however, led them toward a wrong diagnosis (cerebrovascular disease; Table 4). Cue consensus and accuracy can thus diverge, but across histories, we found a relatively high correlation ($r = .74$) between both.

Cue substitutability denotes the extent to which a patient history allows diagnosticians to rely on cues other than the diagnostic ones and still arrive at the correct diagnosis. Although our physicians who relied exclusively on the diagnostic cues attained by far the highest level of accuracy (Table 2), it was possible for them to arrive at the same (correct) diagnosis via different paths.[12–14] This is possible to the extent that some of the cues could substitute for one another—a phenomenon called vicarious functioning, where one can reach the same end by a variety of means.[25] Indeed, cue substitutability was highly correlated with diagnostic accuracy. For instance, in the most often correctly diagnosed patient history, urinary tract infection, 4 (of the total 5) cues had high positive cue validity (Table 4). Apart from the 2 cues deemed diagnostic by the experts, 2 other cues were positively associated with the correct diagnosis. In contrast, in the most difficult patient history, congestive heart failure 2, only 1 cue, exertional dyspnea, was associated with the correct diagnosis (positive cue validity = 76%). However, only a few physicians (17%) considered the diagnostic cue to be crucial, and those who failed to do so ended up misdiagnosing the patient history (negative cue validity of 97%; Table 4). This history was thus ''unforgiving,'' as no other cue afforded a pathway to the correct diagnosis (i.e., the other cues' positive cue validity was very low; Table 4).

The Limitations of This Exploratory Investigation

We calculated the predictive value of cues (i.e., their validity) by analyzing how physicians used (or failed to use) them. A more standard approach is to analyze a representative corpus of patient histories so as to determine how frequently single nonspecific symptoms (e.g., loss of appetite, dizziness) are associated with specific diagnoses (e.g., depression, urinary tract infections), thereby also gauging the cues' sensitivity and specificity. One could thus determine the predictive value of nonspecific symptoms independently of how physicians use and interpret such symptoms as cues. Due to the scarcity of such information in the literature, we determined positive cue validity instead by counting the number of physicians (in our sample) who indicated a cue as crucial for their diagnosis and whose diagnosis was correct. Our analysis of cue consensus, positive cue validity, and negative cue validity tells us, among other things, what cues attracted physicians' attention and to what extent the cues to which they attended enabled them to arrive at the correct diagnosis—or led them to the wrong one. As soon as a representative reference class of histories with nonspecific symptoms becomes available, however, future investigations should analyze validities using the standard approach.

A second limitation of our study is that although we found that cue consensus and cue substitutability are correlated with diagnostic performance, we cannot say to what extent this is a causal relationship. Informed by our correlation analysis, however, future studies can construct patient histories by varying both these properties to determine their causal impact on diagnostic accuracy. Such an approach could also investigate the extent to which physicians take advantage of combinations of nonspecific complaints rather than individual complaints.

A third limitation is that patient histories, drafted by admitting emergency physicians, were notably brief (Table 1). In preparing them, the attending physicians presumably selected the information they considered to be important and omitted what they thought to be irrelevant for the further diagnostic process. A brief, selective patient history represents good clinical practice and reflects the time constraints under which a busy urban emergency department is bound to operate. We chose to use these original (unedited) notes because they are the kind of histories that ED physicians work with every day. Admittedly, however, pondering such prefiltered histories, as our participants did, can only approximate but is not identical to the process through which the admitting emergency physician goes when sifting in real time through a patient history with nonspecific complaints.

A final limitation concerns our use of only 1 control history (myocardial infarction), which obviously does not represent the whole universe of patient histories involving specific symptoms. Our comparisons between the control history and the target histories are therefore only tentative in nature, and the suggestive differences we observed need to be explored in more detail using more comprehensive sets of patient histories.

CONCLUSIONS

We identified 2 correlates of diagnostic accuracy in patient histories involving nonspecific complaints: cue consensus and cue substitutability. To take advantage of the latter, diagnosticians should be aware that, particularly in nonspecific complaints, valid cues might initially be overlooked because they seem insignificant. This can hamper diagnostic accuracy because it is difficult to foretell which combination of cues will provide a path to the correct diagnosis. Therefore, one tentative recommendation from our study is that, in a case involving nonspecific complaints, all possible cues should be acknowledged and the decision about which cues are crucial made only after the complete history is taken. In conjunction, several nonspecific cues can form an informative cluster of intercorrelated (redundant) cues. Looking for clusters of nonspecific cues that point in the same direction, rather than a ''silver bullet'' cue that may not exist in such patient histories, offers one possible route to diagnostic success.

ACKNOWLEDGMENTS

We are grateful to Valerie M. Chase and Laura Wiles for editing the manuscript.

REFERENCES

1. Nickel CH, Ruedinger J, Misch F, et al. Copeptin and peroxiredoxin-4 independently predict mortality in patients with non specific complaints presenting to the emergency department. Acad Emerg Med. 2011;18(8):851–9.

2. Vanpee D, Swine C, Vandenbossche P, Gillet JB. Epidemiological profile of geriatric patients admitted to the emergency department of a university hospital localized in a rural area. Eur J Emerg Med. 2001;8(4):301–4.

3. Kostopoulou O, Oudhoff J, Nath R, et al. Predictors of diagnostic accuracy and safe management in difficult diagnostic problems in family medicine. Med Decis Making. 2008;28(5):668–80.

4. Nemec M, Koller MT, Nickel CH, et al. Patients presenting to the emergency department with non-specific complaints: the Basel Non-specific Complaints (BANC) study. Acad Emerg Med. 2010; 17(3):284–92.

5. Hammond KR, Stewart TR, eds. The Essential Brunswik: Beginnings, Explications, Applications. New York: Oxford University Press; 2001.

6. Funder DC. On the accuracy of personality judgment: a realistic approach. Psychol Rev. 1995;102(4):652–70.

7. Dhami MK, Hertwig R, Hoffrage U. The role of representative design in an ecological approach to cognition. Psychol Bull. 2004;130(6):959–88.

8. Martignon L, Hoffrage U. Why does one-reason decision making work? A case study in ecological rationality. In: Gigerenzer G, Todd PM, the ABC Research Group, eds. Simple Heuristics That Make Us Smart. New York: Oxford University Press; 1999. p. 119–40.

9. Koriat A. When are two heads better than one and why? Science. 2012;336(6079):360–2.

10. Hertwig R. Psychology: tapping into the wisdom of the crowd—with confidence. Science. 2012;336(6079):303–4.

11. Doherty ME, Kurz EM. Social judgement theory. Think Reason. 1996;2(2/3):109–40.

12. Poses RM, Wigton RS, Cebul RD, Centor RM, Collins M, Fleischli GJ. Practice variation in the management of pharyngitis: the importance of variability in patients' clinical characteristics and in physicians' responses to them. Med Decis Making. 1993;13(4):293–301.

13. Wigton RS, Hoellerich VL, Patil KD. How physicians use clinical information in diagnosing pulmonary embolism: an application of conjoint analysis. Med Decis Making. 1986;6(1):2–11.

14. Wigton RS, Patil KD, Hoellerich VL. The effect of feedback in learning clinical diagnosis. J Med Educ. 1986;61(10):816–22.

15. Sonnenberg A, Gogel HK. Translating vague complaints into precise symptoms: the implications of a poor medical history. Eur J Gastroenterol Hepatol. 2002;14(3):317–21.

16. Tintinalli JE, Stapczynski JS, Ma OJ, Cline DM, Cydulka RK, Meckler GD, eds. Tintinalli's Emergency Medicine: A Comprehensive Study Guide. 7th ed. New York: McGraw-Hill; 2010.

17. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988.

18. Tetlock PE. Expert Political Judgment: How Good Is It? How Can We Know? Princeton, NJ: Princeton University Press; 2005.

19. Chew WM, Birnbaumer DM. Evaluation of the elderly patient with weakness: an evidence based approach. Emerg Med Clin North Am. 1999;17(1):265–78, x.

20. Gordon M. Differential diagnosis of weakness—a common geriatric symptom. Geriatrics. 1986;41(4):75–80.

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

22. Bordage G, Grant J, Marsden P. Quantitative assessment of diagnostic ability. Med Educ. 1990;24(5):413–25.

23. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med. 2005;142(4):260–73.

24. Hogarth RM. Educating Intuition. Chicago: University of Chicago Press; 2001.

25. Brunswik, E. The conceptual framework of psychology (1952). Reprinted in: Hammond KR, Stewart TR, eds. The Essential Brunswik. New York: Oxford University Press; 2001.