Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 36 (2003) 61–69
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Medical errors as a result of specialization

Ahmad Hashem,[a,*] Michelene T.H. Chi,[b] and Charles P. Friedman[c]

[a] Healthcare Industry Solutions Group, Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA
[b] Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260, USA
[c] Center for Biomedical Informatics, University of Pittsburgh, Forbes Tower, Suite 8084 200 Lothrop Street, Pittsburgh, PA 15213-2582, USA

Received 8 July 2003

Abstract

Errors in medicine result in over 44,000 preventable deaths annually. Some of these errors are made by specialized physicians at the time of diagnosis. Building on error frameworks proposed in the literature, we tested the experimental hypothesis that physicians within a given specialty have a bias in diagnosing cases outside their own domain as being within that domain. Thirty-two board-certified physicians from four internal medicine subspecialties worked four patient cases each. Verbal protocol analysis and general linear modeling of the numerical data seem to confirm the experimental hypothesis, indicating that specialists try to "pull" cases toward their specialty. Specialists generate more diagnostic hypotheses within their domain than outside, and assign higher probabilities to diagnoses within that domain.

Keywords: Medical errors; Diagnostic errors; Cost of expertise

1. Introduction

Betsy Lehman, a *Boston Globe* health reporter, died from an overdose during chemotherapy. Willie King had the wrong leg amputated. Ben Kolb was eight years old when he died during surgery due to a drug mixup [1,2]. And these are just the "tip of the iceberg" of medical errors [1,2].

One important type of medical error occurs at the time of diagnosis. The popular press cites cases such as that of Dr. Franklin K. Yee, whose abdominal pain was diagnosed as viral gastroenteritis by a gastroenterologist, caused him to be admitted to a coronary care unit by a cardiologist, was suspected by a nephrologist to be the result of kidney stones, and eventually was found on abdominal surgery to be the result of a ruptured appendix [3]. This phenomenon of different specialists projecting their specialties on a patient has not been studied systematically. In the present paper, the role of medical specialization in inducing biases that may underlie some diagnostic medical errors is investigated empirically.

2. The cost of expertise

It may seem strange to talk about the costs of being an expert, but there is increasing experimental evidence that the benefits of expertise are not without costs. The costs of expertise can be divided into two main categories: Those related to accuracy of recall, and those related to inflexibility.

2.1. Accuracy of recall

Experts may outperform novices in recalling the details of a problem or text. But when the domain knowledge of experts cannot be utilized, experts tend to underperform novices. For example, in a study on recall of random chessboard positions, the performance of chess experts was slightly worse than that of novices [4]. Similarly, in a study on memory for baseball texts, participants with high baseball knowledge recalled significantly less baseball-irrelevant propositions from a

* Corresponding author. Fax: 1-425-936-7329.
E-mail address: ahmad@hashem.net (A. Hashem).

1532-0464/$ - see front matter © 2003 Published by Elsevier Inc.
doi:10.1016/S1532-0464(03)00057-1

text passage describing part of a baseball game than did participants with low baseball knowledge [5].

2.2. Reduced flexibility

Some of the early work on the costs of expertise was done by Ggestalt psychologists [6–8]. Luchins argued that experience produced an *Einstellung*, or mental set, that limited the search space of subsequent problem solving. More recent studies have shown similar results. For example, Shiffrin and Schneider [9] reported that people trained for several thousand trials to detect visual targets among distractors in a consistent-mapping condition were at a serious disadvantage when the target and distractor sets were reversed. Wiley [10] showed that problem solvers with a large amount of domain knowledge are confined by their knowledge to one area of the search space—a condition that is efficient if the solution happens to fall in this area but which backfires if it does not.

3. Characteristics of medical expertise

There is no reason to suspect that expertise in medicine is different from expertise in any other domain. In fact, research results have emphasized similarities (e.g. [11–15]). The general characteristics of expertise are summarized elsewhere [16,17] and will not be repeated here. Some characteristics of expertise in the medical domain that bear relevance to the present study will be discussed.

3.1. Diagnostic reasoning mode

According to Patel and Groen [14], expert clinicians confronted with routine cases use a data-driven (forward) approach, in which diagnoses are generated from data by applying a small set of if/then production rules without generating intermediate hypotheses and evaluating them. Less expert clinicians, such as medical students or residents, on the other hand, tend to use a hypothesis-driven (backward) approach, in which reasoning occurs backwards from a hypothesis in an attempt to find data that elucidates it [18]. This assertion is consistent with findings in other domains such as physics [19] and mathematics [20], and with the general notion that in routine situations experts tend to use highly specific problem-solving structures [21].

3.2. Experts working outside their domain

Cognitive literature holds abundant evidence that experts excel only at their domain of expertise (see [17] for an overview). Some of the very few studies that have examined the performance of subspecialist physicians on problems outside their area of specialty were done by Patel and colleagues [22,23]. Cardiologists and endocrinologists were asked to read cardiology and endocrinology cases and to think-aloud as they were reading them, or to recall case information and explain the underlying pathophysiology. The general finding is that experts working within their subdomain tend to use forward strategy more, and to rely on pathophysiological knowledge less, than experts working outside their subdomain. No significant difference in diagnostic accuracy was found, but it is hard to make any meaningful claim on this issue due to the small sample sizes (typically less than 10 participants total, working on one to two cases).

4. The present work

Based on the above discussion, it can be said that prior knowledge plays a crucial role in diagnostic reasoning. Expert performance is a function of the organization, structure, and quality of this prior knowledge. Further, the prior knowledge of the expert is automatically and unintentionally activated—experts cannot help being influenced by it. Therefore, it is plausible to suspect that what one already knows may bias the way one structures a problem and goes about solving it. This "bias" may be beneficial in some cases but costly in others. The goal of the present work is to investigate experimentally whether the cost of this bias is manifested by subspecialists working on problems that are outside their subspecialty area. Specifically, this study is concerned with whether physicians within a given specialty have a bias in diagnosing cases outside their own domain as being within that domain.

5. Methods

5.1. Participants

Thirty-two board-certified physicians practicing in the Pittsburgh area were recruited for this study: eight from each of the internal medicine subspecialties—cardiology, hematology, and infectious diseases (ID)—and eight internal medicine general practitioners (generalists) who did not subspecialize. (In that which follows, we refer to the subspecialties of cardiology, hematology, and infectious diseases as the "subspecialties of interest.") The participants' years of experience, after finishing all formal training, averaged 15.8 years (the standard deviation was 9.9 years and the median was 15.5 years). Participants volunteered their time and did not receive monetary compensation for participation in the study.

5.2. Material

The material used in this study consists of four patient cases, chosen from a total of 36 challenging cases, and used in studies reported elsewhere [24,25]. Each of these four patient cases is represented in a two- to five-page abstract prepared by an expert clinician based on the actual patient charts. The abstracts include all the salient history, physical examination findings, laboratory results, and radiological and other diagnostic studies. They also include ample non-salient data to avoid cueing. However, the abstracts do not include the findings that were judged by the expert abstractor to be gold standard or definitive findings, such as a positive biopsy, because including such findings would render the diagnosis trivial for clinicians.

The four challenging cases that were used in the present study satisfy three conditions: (1) their correct diagnoses belong, respectively, to the three subspecialties of interest, plus a fourth subspecialty not represented among the experts, (2) no case is inherently misleading, and (3) no case is inherently easily diagnosed into one specialty exclusively.

5.3. Design and procedure

Participants read the four patient cases on paper, one after the other. The presentation sequence of the cases was counterbalanced. For each case, the participants viewed its abstract in three consecutive segments corresponding to: chief complaint and history, physical examination, and laboratory data. Because the cases were presented on paper, participants were able to go back to view information from prior segments, but the instructions they were given asked them not to change the responses they provided in previous segments.

At the end of each segment, the participants were asked to give a set of up to six differential diagnoses along with an associated degree of belief (ranging from 0 to 100) that each diagnosis was the correct one. The degree of belief was taken as a measure for a participant's confidence in the correctness of his or her answer, and is a probability expressed as a percentage. The instructions specified that the numbers should add up to 100, but some participants did not follow this requirement. Therefore, the degrees of belief were later proportionately normalized so that their sum in a differential diagnosis set (after seeing one segment of one case) added up to 100. The scale was also transformed from 100 back to 1, consistent with the standard representation of probabilities.

Participants were asked to think-aloud as they worked the cases, and their verbal protocols were tape-recorded and later transcribed for analysis. The experimenter prompted participants to think loudly or to verbalize his/her thoughts whenever there was a period of several seconds of silence.

The independent variables in the study were:

a. The specialty of the participant: cardiology, hematology, infectious diseases, and general medicine.
b. The specialty of the case: cardiology, hematology, infectious diseases, and gastroenterology.
c. The amount of case information revealed to participants before they were asked to give a set of differential diagnoses: chief complaint and history only; the above plus physical examination; and the above plus lab data.
d. Whether the participant worked on a case that matched his/her specialty (same_domain). While this variable is subsumed under the first two variables, we are treating it here as a separate variable to facilitate analysis.

The dependent variables were:

a. The probability assigned to the correct case specialty (P(CS)): The sum of the probabilities assigned by a participant, after seeing a segment of a case, to all diagnoses that belonged to the correct specialty of the case. For example, if a participant saw the chief complaint and history of case 1 (a cardiology case) and offered a differential diagnosis of "aortic dissection, 20%" (which is a cardiology diagnosis), "aortic stenosis, 50%" (a cardiology diagnosis), and "hemolytic anemia, 30%" (a hematology diagnosis), then P(CS) would have been 70 to represent all the cardiology diagnoses in the differential.
b. The probability assigned to the correct diagnosis (P(CD)): The probability assigned by a participant to the correct diagnosis of a patient case.
c. The probability assigned to the participant's own specialty (P(OS)): The sum of the probabilities assigned by a participant, after seeing a segment of a case, to all diagnoses that belonged to the participant's specialty. For example, if the differential diagnosis above was provided by a cardiologist, then P(OS) would be 70. If it was provided by a hematologist, then P(OS) would be 30. This variable is not defined or used for the generalists.

The first two variables reflect diagnostic accuracy, with the first being less strict than the second, as it indicates the probability of being in the "general ballpark." The last variable reflects the degree of bias towards one's own specialty.

To facilitate the calculation of the first and last dependent variables (P(CS) and P(OS)), each diagnosis provided by participants was mapped to its corresponding medical specialty, such as mapping aortic stenosis to cardiology. This was done by a physician, relying on a standard medical textbook [26]. This mechanical task boils down to noting the title of the chapter in which the diagnosis is discussed in the textbook.

The hypothesis being tested in this study is: the participant's specialty biases the participant's answers by influencing the probability assigned to the participant's own specialty. Specifically, participant specialty is related to P(OS), even when controlled for the case. In addition, participants assign a higher collective probability to hypotheses within their domain of expertise than hypotheses outside their domain.

6. Results

In the following sections, we present (a) data from the verbal protocols, (b) ANOVA results of the probabilities assigned to different specialties, and (c) analyses using a generalized linear model.

6.1. Verbal protocol analyses

6.1.1. Generation of diagnostic hypotheses

Using the verbal protocols that participants generated as they worked through the experimental cases, we extracted all the diagnostic hypotheses mentioned in these protocols and mapped them to their specialties using a standard medical textbook [26]. This mechanical task boils down to noting the title of the chapter in which the diagnosis is discussed in the textbook. We then performed a simple counting of these hypotheses. If the same hypothesis was mentioned repeatedly by a participant for the same case, only the first utterance was counted. Fig. 1 shows the number of diagnostic hypotheses from each medical specialty that was generated by each group of specialists in the study while working through the four cases.

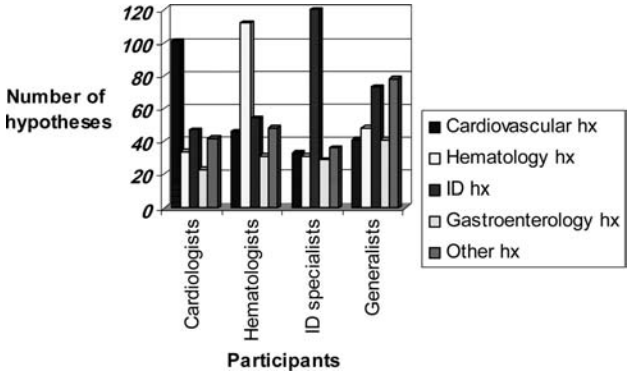

Fig. 1. The number of diagnostic hypotheses generated by participants in the verbal protocols for all cases.

Note the peak that indicates a possible correlation between the domain of the participant and the domain of the diagnostic hypotheses that were generated.

For each participant's specialty, we calculated the log odds of a hypothesis within the participant's specialty vs outside the specialty, for each of the three opposing specialties. We also calculated the standard errors and p values for the hypothesis that the odds are even. For example, for cardiologists we calculated the log odds of a hypothesis within cardiology vs within hematology, along with the standard error and p value for the hypothesis that the number of diagnoses in cardiology and in hematology is the same. We then repeated this for the other diagnostic domains and for the other participant domains. Table 1 shows the results of these calculations. Note that all p values are highly significant, indicating that participants generated more diagnostic hypotheses within their domain than outside their domain.

6.1.2. Cues used in hypothesis generation

For every new diagnostic hypothesis generated by specialists, we noted whether the generation was based on a single cue in the patient case or multiple cues. Examples about generating a hypothesis based on a single cue include: "He's got a very low MCV, which is 72. [...] You see that with um, iron deficiency anemia," and "Weight loss always makes you concerned about cancer." Examples about generating hypotheses based on multiple cues include: "With him hemolyzing and with normal coags, and having thrombocytopenia, um, I'd wonder about the possibility of um, TTP," and "In terms of diagnosis probably ah, with the history of marijuana use, with a murmur, with a temperature you are now thinking in terms of endocarditis." There was

Table 1
Log odds, standard error, and p values for the hypothesis that the number of diagnoses in the participant's domain and that in the other domains is similar

Domain of participants	Domain of diagnoses	Log odd	Standard error	p value
Cardiology	Hematology	1.089	0.198	<0.001
Cardiology	Infectious diseases	0.765	0.177	<0.001
Cardiology	Gastroenterology	1.480	0.231	<0.001
Hematology	Cardiology	0.890	0.175	<0.001
Hematology	Infectious diseases	0.730	0.166	<0.001
Hematology	Gastroenterology	1.285	0.203	<0.001
Infectious diseases	Cardiology	1.291	0.197	<0.001
Infectious diseases	Hematology	1.354	0.201	<0.001
Infectious diseases	Gastroenterology	1.420	0.207	<0.001

Table 2
The number of cues as a basis for generating hypotheses

Basis of generating new hypotheses	Number of hypotheses in cases matching participants' specialty	Number of hypotheses in cases not matching participants' specialty
Single cue	29	190
Multiple cue	61	45
Can't Determine	128	334
Total	218	569

enough information in the protocols to allow for classifying 42% of the hypotheses into these two cue categories, and the remaining hypotheses were classified into a "Can't Determine" category. Table 2 shows the results of this classification.

We calculated the log odds of a single-cue hypothesis within the participants' specialties vs a multiple-cue hypothesis, along with the standard error and p value for the experimental hypothesis that the odds are even. We repeated this for the single-cue vs multiple-cue hypotheses in cases outside the participants' specialties. The results indicated that participants are more likely to base hypotheses on multiple cues when working on cases within their specialty (log odds $= 0.744$, $p < 0.001$), and to base them on single cues when working on cases outside their specialty (log odds $= 1.44$, $p < 0.001$).

6.1.3. Anomalies

For every new hypothesis generated by specialists, we noted whether there was an indication in the verbal protocols that the specialist recognized "an anomaly," or recognized that the hypothesis generated did not quite fit the patient case being examined. Examples about anomaly recognition include: "She could have a urinary tract infection, but it sounds more like this is not a clean catch," and "He could have bacterial meningitis. It doesn't sound like it, but he could." We then noted, for each instance of an anomaly, whether the specialist kept the anomalous hypothesis in the final differential diagnosis list or not. The results can be found in Table 3.

Chi-squared analysis indicated that specialists are more likely to recognize anomalies in the hypotheses they generated when working on cases matching their specialty than they are when working on cases outside their specialty ($\chi^2(1) = 16.31$, $p < 0.001$). However, once recognized, an anomalous hypothesis is equally likely to be kept in the final differential diagnosis list for cases matching or not matching the participants' specialty ($\chi^2(1) = 0.01$, $p = 0.92$).

6.1.4. Summary of protocol analysis results

In summary, the protocol analysis showed that participants generated more diagnostic hypotheses within their domain of expertise than outside. It also showed that when specialists work on cases within their domain of expertise, they are more likely to base their hypotheses on multiple cues as opposed to single cues, and are more likely to recognize anomalies in the hypotheses they generate. However, an anomaly recognized does not necessarily translate into a hypothesis rejected.

6.2. ANOVA results for the probabilities assigned to different specialties

We calculated the probabilities assigned by each group of specialists to each of the subspecialties of interest plus gastroenterology. For example, if a cardiologist provided six diagnoses after seeing a patient case, two of which were in the subspecialty of hematology, then the probabilities assigned to these two diagnoses were added to generate one probability assigned by this participant to hematology. The averages of these probabilities are shown in Table 4. Table 4 also shows the ANOVA results for the hypothesis that the probabilities assigned to the different specialties by each group of specialists are equal. The last column in each row shows the minimum difference per Scheffé's test at the 0.05 level for two probabilities in that row to be significantly different.

The analysis indicates that, with the exception of hematologists, participants are inclined to assign higher probabilities to their own specialty than other specialties, and that this inclination decreases as more case information is revealed.

Table 3
Instances of anomaly recognition

	Cases matching participants' specialty	Cases not matching participants' specialty
Total number of hypotheses	218	569
Instances of anomaly recognition	27	25
Hypotheses kept despite anomaly recognition	18	17

Table 4
The average probabilities assigned to different specialties by each group of participants after each case segment

Participant specialty	Case segment	Average probability assigned to cardiology	Average probability assigned to hematology	Average probability assigned to infectious diseases	Average probability assigned to gastroenterology	F	p value	Scheffé's critical difference at the 0.05 level
Cardiology	1	0.44	0.04	0.16	0.14	14.68	<0.001	0.09
Cardiology	2	0.48	0.03	0.20	0.09	16.64	<0.001	0.10
Cardiology	3	0.26	0.25	0.19	0.14	1.13	0.341	0.10
Hem	1	0.20	0.22	0.16	0.13	0.94	0.425	0.08
Hem	2	0.20	0.20	0.17	0.16	0.18	0.910	0.09
Hem	3	0.12	0.35	0.13	0.18	4.99	0.003	0.10
ID	1	0.19	0.03	0.39	0.17	9.79	<0.001	0.09
ID	2	0.17	0.07	0.43	0.12	11.77	<0.001	0.09
ID	3	0.09	0.28	0.31	0.17	2.93	0.037	0.11

The table also shows the ANOVA F and p values for the hypothesis that the probabilities in each line are equal. Scheffé's critical difference between the averages appears in the last column.

6.3. Generalized linear model

We built a generalized linear model to fit the data using quasi-likelihood. The goal of this task was to find a model that accurately simulates the experimental data and helps to identify the significant predictors for each of the three dependent variables in the study: the probability assigned to the correct case specialty (P(CS)), the probability assigned to the correct diagnosis (P(CD)), and the probability assigned to the participant's own specialty (P(OS)). The model will serve as a tool to establish relationships between these dependent variables and other variables.

Because the dependent variables were the assigned probabilities, it was appropriate to treat the data as pseudo-binomial observations and to perform a linear logistic analysis. However, the data were found to be significantly under-dispersed. Hence, it was necessary to perform quasi-likelihood estimation with binomial family link functions and estimate the dispersion parameters [27]. First, for the task of variable selection, exploratory univariate analyses were performed. The Kruskal–Wallis rank sum test [28] was used to test for potential effects on the three dependent variables by the following variables: specialty of the case, specialty of the participant, the sequential order in which a case was presented to participants, the amount of case information revealed to participants, and whether the participant was working on a case that matched his or her specialty. This test was used because the distributions of the three dependent variables were significantly non-normal. Kendall's rank correlation tau [29] was used to test for significant correlation between the participants' years of experience and the three dependent variables. The statistical package S-Plus was used to perform all the analyses.

The univariate analysis highlighted four potential covariates: amount of information revealed, case specialty, participant specialty, and whether the participant's specialty and the case specialty match. Next forward and backward stepwise variable selection procedures were employed to generate plausible multivariate models for the three dependent variables. Interaction and quadratic terms were added to the variable set to check the linearity and additivity model assumptions.

The following model fits the experimental data adequately, as confirmed by deviance residual inspection and leverage and influence diagnostics

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i, \tag{1}$$

where:

- π is the particular probability, or dependent variable, being modeled.
- x_i are the covariates. They take a value of 1 when the circumstances being modeled match the entry in the "parameter" column in Tables 5, 6, or 7, for P(CS), P(CD), or P(OS), respectively. Otherwise, they take a value of 0.
- i represents a line number in Tables 5, 6, or 7. p represents the largest i in these tables.
- β_0 (the intercept) and β_i are parameters for the model whose estimates can be found in Tables 5–7.

The following subsections provide more details about how the model applies to the three dependent variables. The important message to take from these subsections concerns which variables are significant predictors for each of the dependent variables, as this will be taken as evidence for a relationship between the predictor and the predicted variable. We have pointed out these predictors in the corresponding subsections, and the implications are addressed in Section 7.

6.3.1. Modeling the probability assigned to the correct specialty (P(CS))

Table 5 provides the parameter estimates for the probability assigned to the correct specialty (P(CS))

Table 5
Parameter estimates, standard errors, and t values for Wald test that parameter = 0 for the P(CS) model

i	Parameter	Estimate (β_i)	Standard error	t value	t value significant?
0	Intercept (β_0)	−0.126	0.212	−0.594	No
1	Case specialty = Gastroenterology	0.421	0.221	1.909	No
2	Case specialty = Hematology	−2.872	0.322	−8.932	Yes
3	Case specialty = Infectious diseases	−1.119	0.229	−4.883	Yes
4	Same_domain = YES	1.458	0.201	7.268	Yes
5	Information revealed = All	−0.566	0.271	−2.090	Yes
6	Participant specialty = Hematology	−0.123	0.200	−0.616	No
7	Participant specialty = Infectious diseases	−0.026	0.195	−0.126	No
8	Participant specialty = General medicine	0.411	0.205	2.003	Yes
9	Case specialty = Gastroenterology × information revealed = All	1.049	0.386	2.721	Yes
10	Case specialty = Hematology × information revealed = All	3.652	0.452	8.080	Yes
11	Case specialty = Infectious diseases × information revealed = All	0.876	0.390	2.248	Yes

Table 6
Parameter estimates, standard errors, and t values for the Wald test that parameter = 0 for the P(CD) model

i	Parameter	Estimate (β_i)	Standard error	t value	t value significant?
0	Intercept (β_0)	−4.521	0.399	−11.343	Yes
1	Case specialty = Gastroenterology	0.976	0.421	2.317	Yes
2	Case specialty = Hematology	−2.756	1.017	−2.710	Yes
3	Case specialty = Infectious diseases	−2.434	0.794	−3.065	Yes
4	Same_domain = YES	2.343	0.375	6.250	Yes
5	Participant specialty = Hematology	0.236	0.350	0.673	No
6	Participant specialty = Infectious diseases	0.795	0.360	2.209	Yes
7	Participant specialty = General medicine	1.546	0.344	4.491	Yes
8	Case specialty = Cardiology × information revealed = All	−0.473	0.564	−0.838	No
9	Case specialty = Gastroenterology × information revealed = All	2.054	0.313	6.563	Yes
10	Case specialty = Hematology × information revealed = All	4.650	0.987	4.712	Yes
11	Case specialty = Infectious diseases × information revealed = All	−0.120	1.264	−0.095	No

Table 7
Parameter estimates, standard errors, and t values for the Wald test that parameter = 0 for the P(OS) model

i	Parameter	Estimate (β_i)	Standard error	t value	t value significant?
0	Intercept (β_0)	−0.593	0.225	−2.632	Yes
1	Same_domain = YES	1.549	0.203	7.620	Yes
2	Information revealed = Most	0.734	0.177	4.146	Yes
3	Participant specialty = Hematology	−0.700	0.214	−3.268	Yes
4	Participant specialty = Infectious diseases	−0.215	0.204	−1.054	No
5	Case specialty = Gastroenterology	−0.592	0.254	−2.330	Yes
6	Case specialty = Hematology	−0.428	0.234	−1.828	No
7	Case specialty = Infectious diseases	−0.579	0.242	−2.395	Yes

model. It also includes the standard errors and the t values for Wald test that the parameter is zero.

ANOVA showed that the significant predictors for the probability assigned to the correct specialty were the case specialty, whether the case specialty and the participant specialty match (same_domain) and whether participants viewed the third portion of the case information (information revealed = all), and the participant specialty.

6.3.2. Modeling the probability assigned to the correct diagnosis ($P(CD)$)

Table 6 provides the parameter estimates for the probability assigned to the correct diagnosis (P(CD))

Table 8
A matrix showing the significant predictors for the three dependent variables

	Case specialty	Participant specialty	Same_domain	Amount of information revealed
P(CS)	X	X	X	X
P(CD)	X	X	X	X
P(OS)		X	X	X

model. It also includes the standard errors and the t values for Wald test that a parameter is zero.

ANOVA showed that the significant predictors for the probability assigned to the correct diagnosis were the case specialty, participant specialty, whether the case specialty and the participant specialty match (same_domain), and whether participants viewed the third portion of the case information (information revealed = all).

6.3.3. Modeling the probability assigned to the participant's own specialty (P(OS))

Table 7 provides the parameter estimates for probability assigned to the participant's own specialty (P(OS)) model. It also includes the standard errors and the t values for the Wald test that the parameter is zero.

ANOVA showed that the significant predictors for the probability assigned to the participant's own specialty were whether the case specialty and the participant specialty matched (same_domain), whether participants viewed the second portion of the case information (information revealed = most), and the participant's specialty. A weak predictor is the case specialty.

6.3.4. Summary of the generalized linear model results

Based on the previous analysis, we can build the following table (Table 8) to show which variables are significant predictors for the three dependent variables P(CS), P(CD), and P(OS).

It is important to note that each of these predictors was significant, even when controlling for the other predictors.

The generalized linear model, then, indicates that diagnostic accuracy (in the weak sense of identifying the correct case specialty and in the strong sense of identifying the correct diagnosis) is predicted by the case specialty, the participant specialty, the amount of information revealed, and whether the participant's specialty and case specialty match. The model also indicates that the tendency to identify a case as belonging to one's own specialty is predicted by the participant specialty, the amount of information revealed, and whether the participant's specialty and the case specialty match.

7. Discussion

The main point of this study was to examine whether physicians with a given specialty have a bias in diagnosing cases outside their own domain as being within that domain. The answer appears to be yes, and several pieces of evidence support this view:

a. As can be seen from Table 8, a participant's specialty is related to the probability assigned by that participant to his or her own specialty when diagnosing a case, even when we control for the case. This establishes a relationship between the specialty and the participant's answer.
b. From Table 4, we see that participants assign a higher collective probability to hypotheses within their domain of expertise than outside this domain (except for hematologists). Therefore, it appears that a participant's specialty biases the participant's answers to a case diagnosis by influencing the probability assigned to the participant's own specialty.
c. From Table 1, we can see that in the verbal protocols, participants generate many more diagnostic hypotheses within their domain than outside their domain, further indicating a bias toward one's specialty.

This confirms the experimental hypothesis, that there is a bias toward one's specialty.

Using Rasmussen's skill-rule-knowledge (SRK) framework for understanding human error [30], we can say that because of the automation that results from years of experience with cases within one's specialty, many rules or schemas at the intermediate level of performance are acquired to codify expertise. Many of these rules are primed to be quickly activated, even with minimal information, and sometimes prematurely [31]. This premature activation leads to the generation of hypotheses within one's specialty, as evidenced by the verbal protocols findings, and to assigning higher probabilities to these hypotheses.

The protocol analysis showed that when specialists work on cases within their domain of expertise, they are more likely to base their hypotheses on multiple cues as opposed to single cues, and are more likely to recognize anomalies in the hypotheses they generate. This points to the elaborate knowledge structures to represent an expert's domain and to the impoverished knowledge structures representing the other domains, and is consistent with findings in other fields such as chess [4,32] and physics [19,33]. The use of only single cues outside the specialties of expertise suggests that simple IF-THEN conditional rules are being utilized, whereas experts working within their specialties are likely using "second-order" interactive cues [33]. That is, for

specialists, cues are not directly related to diagnoses (or actions), but rather, their interactions may suggest second-order cues.

One can see from Table 8 that diagnostic accuracy is related to the participant's specialty, indicating that participants from some specialties are more diagnostically accurate than others. Looking at the line associated with $i = 8$ in Table 5, and $i = 7$ in Table 6, one can see that generalists are overall significantly better diagnosticians than the baseline. It is particularly interesting to note that generalists do better than the baseline, whereas specialists tend to be biased toward their own specialty.

This study had several limitations. The number of cases used (four) is relatively small. Although the findings in this study were present despite controlling for the case, it is possible that different cases would show different results. Also, participants were largely affiliated with an academic medical center, and therefore the results may not generalize to other practice settings. The case material used in this study, although comprehensive and based on real patient charts, was not the complete patient record and was not the same as the real patient. In practice, physicians typically have access to more information about their patients than the summaries provided for this study. In addition, if the participants were examining real patients under their care, their motivation, and subsequently their performance, might have been different.

References

[1] Cook R, Woods D, Miller C. A tale of two stories: contrasting views of patient safety. Chicago: National Patient Safety Foundation; 1998.

[2] Kohn LT, Corrigan JM, Donaldsons MS, editors. To err is human: building a safer health system. Washington, DC: National Academy Press; 2000.

[3] Altman LK. How tools of medicine get in the way? New York Times 2000;(May 12):C3.

[4] Chase WG, Simon HA. The mind's eye in chess. In: Chase WG, editor. Visual information processing. New York: Academic Press; 1973.

[5] Voss JF, Vesonder G, Spilich H. Text generation and recall by high-knowledge and low-knowledge individuals. J Verbal Learn Verbal Behav 1980;9:651–67.

[6] Duncker K. On problem solving. Psychol Monogr 1945;58:270.

[7] Luchins AS. Mechanization in problem solving. Psychol Monogr 1942;54:6, Whole No. 246.

[8] Luchins AS, Luchins EH. Rigidity of behavior: a variational approach to the effects of Einstellung. Eugene, OR: University of Oregon Books; 1959.

[9] Shiffrin RM, Schneider W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychol Rev 1977;84:127–90.

[10] Wiley J. Expertise as mental set: the effects of domain knowledge in creative problem solving. Unpublished PhD Dissertation, University of Pittsburgh, 1996.

[11] Boshuizen HPA, Hobus PPM, Custers EJFM, Schmidt HG. Cognitive effects of practical experience. In: Evans DE, Patel VL, editors. Advanced models of cognition for medical training and practice, vol. 97. Berlin: Springer; 1991. p. 337–48.

[12] Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge, MA: Harvard University Press; 1978.

[13] Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: a ten-year retrospective. Eval Health Prof 1990;13/1:5–36.

[14] Patel VL, Groen GJ. Knowledge-based solution strategies in medical reasoning. Cogn Sci 1986;10:91–116.

[15] Patel VL, Groen GJ. Cognitive frameworks for clinical reasoning: applications for training and practice. In: Evans DE, Patel VL, editors. Advanced models of cognition for medical training and practice, vol. 97. Berlin: Springer; 1991. p. 193–211.

[16] Ericsson KA, editor. The road to excellence: the acquisition of expert performance in the arts and sciences, sports, and games. Mahwah, NJ: Erlbaum; 1996.

[17] Glaser R, Chi MTH. Overview. In: Chi MTH, Glaser R, Farr MJ, editors. The nature of expertis. Hillsdale, NJ: Erlbaum; 1988. p. xv–xxviii.

[18] Patel VL, Kaufman DR. Clinical reasoning and biomedical knowledge: implications for teaching. In: Higgs J, Jones M, editors. Clinical reasoning in the health professions. Oxford, UK: Butterworth–Heinemann Ltd; 1995. p. 117–28.

[19] Larkin JH, McDermott J, Simon HA, Simon DP. Expert and novice performances in solving physics problems. Science 1980;208:1335–42.

[20] Hinsley DA, Hayes JR, Simon HA. From words to equations: meaning and representation in algebra word problems. In: Just MA, Carpenter PA, editors. Cognitive processes in comprehension. Hillsdale, NJ: Erlbaum; 1977. p. 89–108.

[21] Ericsson KA, Smith J. Prospects and limits of the empirical study of expertise: an introduction. In: Ericsson KA, Smith J, editors. Toward a general theory of expertise: prospects and limits. New York: Cambridge University Press; 1991. p. 1–38.

[22] Joseph GM, Patel VL. Domain knowledge and hypothesis generation in diagnostic reasoning. Med Decis Making 1990;10:31–46.

[23] Patel VL, Groen GJ, Arocha JF. Medical expertise as a function of task difficulty. Mem Cognit 1990;18:394–406.

[24] Elstein AS, Friedman CP, Wolf FM, Murphy G, Miller J, Fine P, Heckerling P, Miller T, Sisson J, Barlas S, Biolsi K, Ng M, Mei X, Franz T, Capitano A. Effects of a decision support system on the diagnostic accuracy of users: a preliminary report. JAMIA 1996;3/6:422–8.

[25] Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, Fine PL, Miller TM, Abraham V. Enhancement of clinicians' diagnostic reasoning by computer-based consultation. JAMA 1999;282/19:1851–6.

[26] Cecil RL. Cecil's Textbook of Medicine. 20th ed. Philadelphia: W.B. Saunders; 1996.

[27] McCullagh P, Nelder JA. Generalized linear models. 2nd ed. London: Chapman & Hall; 1989.

[28] Hollander M, Wolfe DA. Nonparametric statistical methods. New York: Wiley; 1973.

[29] Conover WJ. Practical nonparametric statistics. 2nd ed. New York: Wiley; 1980.

[30] Rasmussen J. Information processing and human-machine interaction: an approach to cognitive engineering. New York: North-Holland; 1986.

[31] Reason JT. Human error. Cambridge, England: Cambridge University Press; 1990.

[32] de Groot AD. Thought and choice in chess. New York: Mouton; 1965.

[33] Chi MTH, Feltovich PJ, Glaser R. Categorization and representation of physics problems by experts and novices. Cogn Sci 1981;5:121–52.