Memory & Cognition
1992, 20 (4), 344-355

The correlation of feature identification and category judgments in diagnostic radiology

GEOFFREY R. NORMAN, LEE R. BROOKS, CRAIG L. COBLENTZ,
and CATHERINE J. BABCOOK
McMaster University, Hamilton, Ontario, Canada

Expert and novice radiologists were given films accompanied by clinical histories that supported a diagnosis either of bronchiolitis or of normal. To provide a plausible task context, some films were radiologically unambiguous and were accompanied by histories consistent with them. For a set of radiologically difficult films from confirmed normal or bronchiolitis patients, fictitious normal or abnormal histories were counterbalanced with the films. The clinical histories affected ratings both of diagnosis and of features present on the difficult films. Thus, uncertainty about individual features evidently was affected by history, and features did not act as an independent source of information. The dependence of feature calls on an overall judgment was also suggested by intra-observer agreement in another study in which an explicit diagnosis was not requested. It is unclear whether the history increased discrimination between normal and abnormal films, or indiscriminately added evidence for or against the disease. Factors are discussed that make it appropriate for feature identification to be partially dependent on category identification.

The distinction between diagnoses and the signs and features used to support those diagnoses is prominent in medical discourse. Textbooks and journals describe disorders in terms of the features that are normally present, and opinions about particular cases are commonly justified in terms of the features that are seen to be present. The form of these verbal descriptions, right down to the syntax, encourages thinking of the features as evidence that was collected independently of the final diagnostic decision. For example, a listing of most of the features is commonly given before mention of the disorder, and phrases such as "so I conclude" and "therefore" often precede the diagnostic conclusion. In fact, independence between feature identification and diagnosis is assumed in prominent models of medical decision making. Bayesian decision models (see, e.g., Fischhoff & Beyth-Marom, 1983), regression models (Slovic, Rorer, & Hoffman, 1971; Wigton, 1988), and computer-based decision aids (Guppy et al., 1989) normally focus on the way in which the available evidence is combined to produce a diagnostic decision, without allowing for the possibility that provisional diagnostic decisions can influence decisions about what features are present. These assumptions may be viable in domains such as laboratory medicine, where a plausible argument for independence of features and diagnoses might be made, but these models are also a basis for much research in such perceptually rich and ambiguous specialties as radiology (Lusted, 1968; Slovic et al., 1971).

There is good reason to question whether feature identification is actually done independently of diagnostic category identification. The reported features of a clinical problem are not perceptual first impressions, but rather decisions that are produced under no particular time pressure and are meant to be open to public scrutiny. The first mention of a feature can occur well into a diagnostic session, and after the first mention of the diagnosis that will ultimately be supported (e.g., Barrows, Norman, Neufeld, & Feightner, 1982). Consequently, the temporal relations alone provide grounds to suspect that information about the diagnostic category can provide a context in which decisions about some of the reported features are made.

There are also good normative reasons for the diagnostician to consider the diagnostic context when deciding on the presence of a feature. Most features are associated with minitheories about variables that can produce large changes in the feature's expected value, but that are irrelevant to the clinical disorder being contemplated. For example, serum creatinine, a body chemical that can reflect underlying kidney disease, depends, among other things, on body mass, protein intake, and pregnancy. The same value, then, could be thought of as normal, borderline, or seriously abnormal, depending on other aspects of the patient. In radiology, an apparently enlarged heart could be a normal heart casting a large shadow because the patient is in a rotated position. Some part of this background variability can be viewed as sloppy technique (e.g., allowing fluid samples to age too long before processing, or not exposing a chest radiograph with good inhalation or proper positioning), or even as outright lab

This research was carried out under the support of the Natural Sciences and Engineering Research Council of Canada through a grant to the first and second authors. Address reprint requests to Geoffrey Norman, Department of Clinical Epidemiology and Biostatistics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S 3Z5. The authors wish to thank Mary Lou Schmuck for her assistance in the analyses.

error. But in most cases, the variable interpretations are a result of understandable variation in the patient that radically changes the clinical significance of a finding. Clearly, a borderline abnormal sign would be considered differently if it were in the context of other information suggesting a known disorder, since such information would help to discount some of the many possible sources of normal variation. It is possible that this necessity of inspecting the "integrity of the data" results in expert nephrologists' subsequently being able to recall more of the data than medical students can, even when the data are irrelevant to the diagnosis (Norman, Brooks, & Allen, 1989). That is, the effect of expertise in this field does not seem to be to restrict attention to the narrowly construed relevant data.

The net effect of these considerations is that an important skill of an experienced diagnostician is to decide which variations are "real" and which are not. Obvious signs of expertise are comments showing a detailed understanding of the variety of factors that can influence the appearance of potentially significant features. Often, however, this assessment is done implicitly and is made explicit only in response to direct questioning. As a consequence, overt reports of features should be taken as the output of an evaluative process operating in a diagnostic context, rather than as an independent assessment of the features detected.

This is exactly the picture that is provided by Lesgold et al. (1988) for radiology:

> Radiologists report that it is not unusual to find disease, such as a tumor, in an X-ray film, and then once having seen it, to be able to see the beginnings of that disease process in an earlier film that was previously judged normal. (p. 331)

In their view, this expert performance stands in contrast to that produced by less experienced diagnosticians:

> There appeared to be little decoupling between the manifestations of chest structures and the [residents'] internal representations of the chest and its structures.... when a complex case arises, there is need for the representation of the patients' medical condition to be decoupled from film features. (Lesgold et al., 1988, p. 329)

On the other hand, there are factors that clearly suggest the need for restraint in deciding the "reality" of a feature in terms of the overall picture. Anyone who has thought about the role of evidence in science is very aware of the problems presented by rationalizing away any unexpected variation of data through recourse to the overall expected pattern. Consideration of the overall pattern can act only as a bias that masks observation of legitimate variation in the evidence, or it can prompt closer examination that may eliminate genuinely extraneous factors. If the public data is to have any value that is independent of the immediate hypothesis, it must be treated as if it is of interest in its own right. Medical experts must be concerned about such possibilities.

Diagnosticians also have independent reasons to be interested in the features, since some of those features justify therapeutic decisions regardless of cause. This is common in critical-care medicine, where much therapy is directed at holding within normal range physiological parameters such as respiration rate, blood gases, or serum potassium, regardless of the cause of the deviation. Also, for some disorders, the distinction between a feature and a diagnostic category is fuzzy at best. In radiology, for example, pulmonary nodules (see, e.g., Swensson, 1980) or bowel polyps (Markus, Somers, O'Malley, & Stephenson, 1989) are discussed as features, but are themselves disorders that are candidates for direct treatment.

In sum, there is prior reason to suspect that in many common diagnostic situations, uncertainty about the clinical significance of labeled features is resolved together with the probable diagnostic category. Most medical tests are influenced by such a variety of extraneous, but lawful, factors that apparently abnormal features cannot be taken at face value. On the other hand, consideration of the desirability of independence between evidence and conclusions, as well as the separate importance of many of the features, makes it unlikely that skilled practice allows the process of feature and diagnostic category identification to be completely dependent on one another.

The Effect of Prior Information

Radiologists are not unfamiliar with the potential biasing effect of cues that are extraneous to the characteristics of the film itself. One issue that regularly confronts the practicing radiologist is whether to read the referring note, which often contains some clinical information and a tentative diagnosis, prior to examining the film. This dilemma has led to a number of studies of the impact of prior information—a suggestive history, a tentative diagnosis, or a directed instruction—on diagnosis.

Berbaum et al. (1986) showed that the provision of a tentative diagnosis resulted in an improvement in true-positive rates for detection of diverse chest lesions. Schreiber (1963) also demonstrated that the provision of a clinical history with chest films improved true-positive rates, although at the cost of a small increase in false-positive rates. Berbaum et al. (1988) also demonstrated an improvement in true-positive rates, with no change in false-positive rates, for detection of fractures. Finally, Doubilet and Herman (1981) also showed that a suggestive history improved true-positive rates on reading chest films, with a small increase in false positives. One exception to this trend is the study of Good, Cooperstein, and deMarino (1990), who found no difference in accuracy, as measured by the area under the receiver-operating characteristic (ROC) curve.

Signal detection theory is one perspective from which to interpret the findings of these studies. The history was consistent with the final diagnosis in all of the studies listed above, so that the radiologists received more consistent information, primarily for the abnormal films. This can

be viewed as a bias in favor of the abnormal diagnosis for positive films and as a bias toward normal for the negative films. The consequence is that there is a higher true-positive rate and true-negative rate, and a positive change in d'. It is unclear whether this results from an increase in detection of abnormalities on abnormal films and from discounting on normal films, or simply from the incorporation of the additional information into the overall judgment. Even more surprising is that there is *any* increase in false-positive rates, which amounts to a greater likelihood of calling a normal film positive, even though the history is normal or noncontributory, simply because there are other positive histories around.

Disaggregating the effect of history in contributing information to the overall judgment, as opposed to biasing the search for cues on the film, requires two conditions: a careful separation between features and the diagnostic judgment, and a crossover of the prior information with the film. In all of the studies, no clear distinction has been made between the features visible on the film and the diagnostic interpretation made from those features, or even between the specific diagnosis and whether the film was simply normal or abnormal.

In this regard, the studies of pulmonary nodules are informative. Berbaum et al. (1986) have demonstrated that categorical prompts, which led to overall improvement in other studies, did not result in improvement in the search for nodules. Their explanation is that detection of a pulmonary nodule (particularly a simulated one) is an issue of recognition of a single feature, but this is quite different from the integration of a number of separate features required in order to diagnose many other chest conditions or other abnormalities. Since categorical prompts do not appear to affect discrimination for nodule search, it may be that the prompts are acting at the level of aggregation of information, rather than biasing the search for features. But this interpretation is hardly based on direct evidence.

The experiments in this paper are designed to directly assess the co-determination of feature and category judgments in radiology by varying the clinical histories that accompany the films. In contrast with the design of the previous studies, history information is crossed with film, so that the specific magnitude of the bias can be determined. In addition, the effect of this category-level bias both on the judgment of diagnostic likelihood and in the identification of the features present is examined.

We selected a diagnostic category that, in the view of the two radiologists among us (C.L.C. and C.J.B.), should be particularly vulnerable to biasing information. This condition, bronchiolitis, is a viral illness resulting in inflammation of the small bronchioles. It occurs in very young children, and is particularly serious for those with congenital heart disease, prematurity, or previous lung problems. It may be accompanied by severe clinical manifestations—rapid breathing, high fever, coughing, and wheezing. Five radiological features are used to verify the disorder: hyperinflation, bronchial wall thickening, perihilar linear opacities (linear streaks around the hilum, a central structure of the lungs), consolidation (fluid or pus filling the air spaces of the lungs), and atelectasis (loss of lung volume). Any of these features may be present or absent. Bronchiolitis is rarely fatal, but it is sometimes so severe that a child may become cyanotic and require hospitalization. The treatment of a very ill child suspected of having bronchiolitis routinely includes a chest X ray, for several reasons: (1) Diagnosis: Bronchiolitis can present itself similarly to other conditions that have very different causes and treatment, such as congestive heart failure or a foreign body in the lung (carrots, peanuts, etc.). The radiograph can enable the identification of bronchiolitis itself. (2) Management: Management depends on the diagnosis. In addition, the radiograph may give information about the severity of the condition, which may affect aggressiveness of management. (3) Course of treatment: The radiograph may give information about the initial state of the disease, so that response to therapy can be verified radiographically as well as clinically.

The diagnosis of bronchiolitis should be sufficiently labile to make it appropriate for investigating the effect of bias from clinical history. No features are unique to this disorder; singly, each of the features could be attributed to other disorders in the competitor set, such as asthma, pneumonia, a foreign body in the lung, congestive heart failure, or cystic fibrosis. Furthermore, because there is low direct cost of false alarm, our manipulations are not fighting against a strong response bias. Unlike bacterial pneumonia, whose diagnosis often prompts a course of antibiotics, the usual immediate response to bronchiolitis is observation in case the clinical symptoms should worsen. On the other hand, there are important medical reasons for considering the features of bronchiolitis separately, since some of the features may need treatment themselves. For example, collapse is a feature that could suggest specific management interventions such as physiotherapy.

Experiment 1 provided initial evidence of the interdependence of feature and diagnostic identification and also served the purpose of calibrating the material for the subsequent work. In Experiment 2, equivocal films were paired at random with either a normal clinical history or a history that would be consistent with a diagnosis of bronchiolitis. We observed the effect of this history on ratings of the probability of bronchiolitis and on ratings of the presence of the five cardinal radiological features of bronchiolitis. We were particularly interested in whether the history would affect just the rating of the disease, which from some points of view would be normatively proper, or would also affect the number of features observed. If an effect of history should be observed, the question would become whether this effect is purely a bias toward resolving uncertainty in line with the clinical history, or whether the history occasioned greater discrimination between the normal and the abnormal films. Either alternative was possible, given the considerations regarding feature identification discussed above. In Experiment 3, we extended this design by testing diagnosticians

with much less expertise than that of the subjects of Experiment 2.

EXPERIMENT 1
Interobserver Variation

Experiment 1 was originally designed to address questions of interest to radiologists—in particular, the degree of variability between and within observers in the identification of the features of bronchiolitis. With respect to the present paper, Experiment 1 was primarily of interest because it provided initial evidence of co-determination among the identification of the various features. In addition, it allowed us to select stimuli that were equivocal for the disorder, which is important for the design of the subsequent studies. The subjects in Experiment 1 were not required to state or defend a diagnosis, and consequently they were not under overt pressure to reconcile the feature calls with an overall diagnosis.

Method

Subjects. The subjects were 3 expert pediatric radiologists from the Hospital for Sick Children in Toronto. This hospital is a national referral center for pediatric problems, and the radiologists involved in the study are acknowledged as being among the preeminent pediatric radiologists in Canada. Each subject was approached individually, and each one participated on a volunteer basis.

Materials. Chest radiographs from 25 patients with known bronchiolitis were randomized with 25 normal chest radiographs. The bronchiolitis radiographs were obtained from patients diagnosed as having bronchiolitis, on the basis of the following clinical inclusion criteria: age 2 months to 2 years inclusive, fever >38°C, tachypnea (rapid breathing), and discharge diagnosis of bronchiolitis. Under ideal circumstances, it would be optimal to include viral titers in the diagnostic gold standard, but since the study was done retrospectively and titers are not routinely taken, this was not possible. In any case, it has been demonstrated that a positive viral titer occurred in only 65%-85% of confirmed bronchiolitis cases (Holdaway, Rome, & Gardner, 1967) and consequently is not a gold standard in itself. The use of the discharge diagnosis is generally considered a reasonable inclusion criterion, since it is obtained following a period of hospitalization, in which the clinical course, response to therapy, and so forth are observed to be consistent with the bronchiolitis diagnosis. There was no effort to select unusually easy or unusually difficult cases. The cases that were used were the first 25 found to meet the criteria.

The control radiographs were obtained from children with a positive TB skin test, and from patients with innocent cardiac murmurs who, upon follow-up, were found to be free of disease. Patients with a history of respiratory distress in the newborn period were excluded from the study.

We recognize that restricting the radiographs to a single condition or normal, and designing the reporting forms to elicit the features of only one condition, does not represent some aspects of clinical practice adequately. Nevertheless, in many circumstances in clinical radiology, this restriction is not an issue. For example, in screening tests such as mammography or radiological assessment for pneumoconiosis, there is only one diagnosis actively under consideration. In some diagnostic situations, the request from the clinician may be to rule in, or out, a particular condition. Finally, there are other situations in which the primary judgment is one of the disease's severity (e.g., in emphysema or pneumoconiosis), once the presence of the disease has been established.

Procedure. The radiologists independently reviewed the radiographs twice, with 1 week between sessions. The radiographs were randomized in different orders for the first and second series to minimize order bias. No clinical history or other information about diagnostic category was provided. The following radiological features were assessed, in checklist form, as being present, absent, or equivocal: hyperinflation, bronchial wall thickening, perihilar linear opacities, collapse, and consolidation.

Analysis. Agreement statistics were calculated using weighted kappa statistics with quadratic weights, which permits the three levels of assessment (absent, equivocal, and present). Kappa is a measure of agreement for nominal and ordinal scales, based on agreement adjusted for unequal marginals. It has been shown to be mathematically equivalent to the more common intraclass correlation (Cohen, 1968), and it is the usual measure of agreement for nominal categories in medicine.

Results

The interobserver agreement coefficients, shown as weighted kappas in Table 1, were in the middle range, between 0.4 and 0.65, with the best agreement for hyperinflation. Not surprisingly, the intra-observer kappas were consistently higher. These kappas are in a range similar to those found by other authors in a range of radiological diagnoses (reviewed in Coblentz, Babcook, Alton, Riley, & Norman, 1991).

The most frequently detected features were hyperinflation and perihilar linear opacities, which were rated as present in about 55% of the bronchiolitis radiographs. Bronchial wall thickening followed closely at 53%. Collapse and consolidation were detected relatively infrequently at 22% and 30%, respectively. Features were also reported in the normal films, with hyperinflation, thickening, and opacities reported most frequently—up to 30%

Table 1
Inter- and Intraobserver Agreement for Each of the Five Diagnostic Features for Bronchiolitis, Measured in Kappa

	Weighted Kappa Coefficients			
Features	Observer 1	Observer 2	Observer 3	Interobserver
Hyperinflation	.62	.56	.65	.48
Bronchial thickening	.46	.34	.64	.35
Perihilar opacities	.32	.34	.67	.28
Collapse	.51	.47	.57	.41
Consolidation	.60	.32	.31	.33

Note—Kappa coefficients measure the proportions of agreement above chance, divided by the possible proportions above chance.

if equivocal responses are counted. As we explained, these radiographs were selected so that there could be reasonable assurance that the children would show no radiological abnormalities. The fact that some features were found on normal films is not in itself remarkable, although the rate seemed suspiciously high to the radiologists among us (C.L.C. and C.J.B.), which encouraged further investigation.

One possibility for this high rate of features in normal films is that the reader was induced by the task (to find the presence of features of bronchiolitis) and materials (a set of films, some of which showed clear evidence of bronchiolitis) to consider each film in terms of bronchiolitis, even though an explicit diagnosis was not requested. The reader might thus have searched for features consistent with bronchiolitis and tended, as a result, to resolve uncertainty in a direction consistent with an overall diagnosis. Some evidence consistent with this possibility is provided by the subsequent analysis.

Our main interest in these data is to provide evidence relevant to whether the features were being read in the context of an opinion about an overall diagnosis. To analyze for this possibility, we examined the consistency of feature calls across the two readings that a given reader made of each film. There were 50 films read by each of 3 radiologists, for a total of 150 reader-film pairs. For each of these pairs, we counted the number of features out of five that were called *present* or *equivocally present* on one trial, and *absent* on the other. We then counted how many of these changes on the same film were directionally consistent with each other; that is, we counted how many of them were consistent with bronchiolitis on the first trial and consistent with normal health on the second. From the first to the second reading, zero to five features might change, either from normal to abnormal, or vice versa. If feature calls are consistent with an overall impression, all of the changes in features from one session to the next should be in the same direction. Conversely, if the readings of the features are independent, changes in the call from present to absent on one feature should not be directionally associated with changes in the other features. That is, if the feature changes are independent of one another, the total number of changed features indicating bronchiolitis from one trial to another should have a binomial distribution. These data are shown in Table 2.

Two aspects of these data are immediately apparent. First, the data are not distributed binomially. Evidently, in a large number of cases, the readers were either reading the films as showing bronchiolitis or as showing normal health and were resolving uncertainty about the features in a manner consistent with this presumptive diagnosis. This pattern of covariation was consistent across all 3 observers. With the exception of one cell for 1 observer, the two highest scores for each observer for each cell were at the opposite ends of the distribution. This evidence for covariation is more striking, given that the readers had not been asked to give an overt diagnosis but rather were supposed to be reading only the individual features. Of course, they were being asked to rate the presence of five features that they must have known to be the radiological criteria for bronchiolitis. But at least they were not in the position of feeling pressure to reconcile their feature calls with a publicly stated diagnosis. As indicated in the introduction, this pattern of covariation among the features is not necessarily normatively improper, but it is certainly not consistent with a model that treats the features as evidence that is read independently of a category decision. As is evident, kappa, or any coefficient that treats the features singly and ignores covariation, yields an optimistic assessment of the reliability in the assessment.

The second obvious feature of the data in Table 2 is that the overall reproducibility seems low. On 10% of the reader-film pairs, the reader changed calls on all five features, and on 50% of the reader-film pairs, calls were changed on over half of the features. In considering this, it is worth emphasizing that the 3 readers were preemi-

Table 2
The Number of Features of Films Called Present or Equivocally Present on One Trial and Absent on the Other, and the Number of Changes for the Same Films that Were Directionally Consistent

Changes Directionally Consistent	Features Present on One Trial and Absent on the Other					
	0	1	2	3	4	5
0		8	10	14	11	9
1		5	3	6	3	1
2			15	5	0	1
3				9	2	0
4					9	0
5						5
Total pairs of readings	34	13	28	34	25	16
χ^2			19.1	36.0	107.1	61.1

Note—If changes in feature readings by the same reader across the two sessions were independent of one another, the column frequencies should be distributed binomially. For each number of discrepant features above one per film, a chi-square test rejected a binomial distribution.

nent pediatric radiologists, and no information about clinical history or outcomes was given on either of the two reading sessions. As previously mentioned, bronchiolitis is a difficult diagnostic category, in that no feature is unique to this disorder. But the kappas shown in Table 1 for the reliability of individual features are by no means low for other radiological diagnoses (Coblentz et al., 1991). As will be shown in Experiment 2, some of the films in this 50-film set were diagnostically reasonably straightforward, and others were consistently borderline. However, since the films were not especially selected to be difficult, this rate of equivocal films seems to be a fact of life for this disorder.

If memory for the previous reading were high, reliability might appear to be higher than it would under more normal clinical conditions, in which the same film does not appear again. In a previous study of chest radiographs, it has been suggested that memory for individual cases is sufficiently high to make this a concern. Myles-Worsley, Johnston, and Simons (1988) showed that senior radiologists had a hit rate of .59 and a false-alarm rate of .25 for a series of 25 abnormal films, which is approximately what they scored for a set of faces presented under the same conditions. However, it is difficult to evaluate this performance in relation to the conditions of the present study. There were several aspects of Myles-Worsley et al.'s procedure that might be expected to have produced more evidence for memory: the subjects knew that they were in a memory experiment when they first saw the films, the test was an explicit old/new judgment rather than an incidental test of memory, the retention test was given immediately, and each abnormal film portrayed an abnormality not present in any other film in the series. On the other hand, in order to maximize an effect of expertise, the materials were only presented for 500 msec at a rate of one slide per second, which is clearly less favorable than the free pace used in the present study. In relation to this, the memory conditions in the present study are probably much poorer. In fact, they can be construed as being worse than those in many clinical conditions. Half of the slides in the present study were of exactly the same disorder, and they were presented under massed conditions. There was not the variety of disorders and queries that might be expected to make the coding episodes more easily retrievable under more normal clinical circumstances. The quality of the memory conditions is of interest, since in dermatology, memory for previous cases has been shown to influence accuracy by 10%-15% for subsequent encounters with the same or similar cases, even after a 2-week interval (Brooks, Norman, & Allen, 1991).

Again, however, the major focus of this paper is on the evidence for co-determination of feature and diagnostic ratings. In Experiment 1, we did not have any direct information about the diagnoses considered by subjects. Whether the identification of features would be influenced by opinions on the diagnosis could be tested more directly by experimentally manipulating clinical history and measuring its effect on both the diagnostic and the feature ratings. This manipulation was made in Experiments 2 and 3.

EXPERIMENT 2
Effect of History on Judgments of Experts

The films from Experiment 1 were used again. However, in Experiment 2, the subjects were explicitly asked to rate the estimated likelihood of bronchiolitis for each film. In addition, expectations about the diagnosis were manipulated through the use of a brief clinical history. The films that proved to be radiologically borderline in Experiment 1 were associated, across subjects, with both a history indicating bronchiolitis and a history indicating normal health. The bronchiolitis history (fever, cough, and tachypnea) is the standard set of presenting symptoms for bronchiolitis, and the features of this history are not associated singly with the individual radiological features. Since bronchiolitis was the only disorder clearly represented in the study, since the features being rated were those known to be diagnostic for bronchiolitis, and since the participants referred to it as "the bronchiolitis study," we suspected that this manipulation of history would most directly influence the decision about diagnostic category. Our immediate interest, however, was whether the influence of history would affect only the rating of the diagnostic category or would also change the number of radiological features rated as present, thus suggesting that the features are not fully independent sources of information.

Method

Subjects. The study involved 4 acknowledged expert pediatric chest radiologists—1 from McMaster University in Hamilton, and 3 from the Hospital for Sick Children in Toronto. Two of the 4 radiologists had also participated in Experiment 1. The time interval between studies was over 1 year, so we were not concerned with large prior memory effects.

Materials. The materials for the study were the set of 50 posterior-anterior (PA) chest radiographs of children used in Experiment 1. However, we used the results of the first experiment to identify the radiographs that, in the judgment of the investigators, were "definite" or "ambiguous" for each of the diagnoses bronchiolitis or normal. Specifically, we examined the data from the first experiment to locate films, both normal and abnormal, that resulted in relatively more disagreements among observers in the description of features. These films were then examined by the radiologists (C.L.C. and C.J.B.) to reach agreement on the degree of ambiguity.

We then created a standard history, which was attached to each radiograph envelope, stating either "fever, cough, and tachypnea" (a history consistent with bronchiolitis) or "preoperative screen in a well child" (consistent with a normal radiograph). Although this history is obviously brief and not individualized, it is not unrepresentative of the frequently brief and standardized instructions on radiological requisitions.

For the definite radiographs, the history was always consistent with the radiological diagnosis, since we presumed that the combination of a positive clinical history with a clearly normal film, or alternatively, the presence of obvious and serious chest signs in an allegedly healthy child, would lead the subjects to deduce that some experimental manipulation was present, which could seriously compromise the study.

A second strategy was used in the sequencing of the films, in order to establish the credibility of the study. The first seven films in the sequence were all definitely normal or abnormal films with a consistent history, so that the experimental intervention was begun only after subjects had experienced several consistent films in the task. All the subjects saw the films in the same sequence.

For the equivocal radiographs, two envelopes were prepared, one with each history. The radiographs and envelopes were then combined into two sets, so that an individual subject would see a total of 50 radiographs, of which about half were definitely normal or abnormal with a consistent history, and the remainder would be equivocally normal or abnormal. For the latter radiographs, half would have a history consistent with the actual diagnosis, and half would have the opposite history.

Thus, each radiologist saw 12 definitely normal films with a normal history, 12 definite bronchiolitis films with an abnormal history, 14 equivocally normal films (7 with a normal history and 7 with an abnormal history) and 12 equivocal bronchiolitis films (6 with an abnormal history and 6 with a normal history). Conversely, each definite radiograph was interpreted by 4 experts, all with a consistent history; each equivocal radiograph was interpreted by 2 experts with a consistent history and by 2 experts with the opposite history.

Procedure. Each subject interpreted the 50 radiographs in a single session and completed a structured form in which he/she was asked to state the presence or absence of the five features on three levels—present, equivocal, or absent. The subject then rated the likelihood of bronchiolitis on a 6-point scale, ranging from −3 = *definitely absent* to +3 = *definitely present*. The rating of 0 was omitted so that subjects were forced to commit to normal or abnormal.

Analysis. The primary analysis focused on the equivocal radiographs. The diagnosis ratings were analyzed using a repeated measures analysis of variance (ANOVA), with radiograph as the "subject" in the analysis. The radiographs were grouped into the two categories—normal and bronchiolitis. Ratings from 4 subjects on each radiograph were analyzed. Two subjects saw each radiograph with a positive history, and 2 saw each radiograph with a negative history. Since each rater observed half the films with a positive and half with a negative history, each rating was not associated with the same rater on all films. (The effect of this design is to include variance due to rater bias, if present into the error term, resulting in a possibly conservative test of the research hypotheses.) Thus, the ANOVA had three factors: one between-subject factor (bronchiolitis/normal), and two within-subject factors (positive/negative history and Reader 1/Reader 2). Because there were 26 films in the primary analysis (14 normal and 12 bronchiolitis), the degrees of freedom in the error term were (12−1) + (14−1) = 24.

The dependent variables were (1) the number of features rated as present or equivocal for each film, and (2) the diagnosis rating on the +3 to −3 scale. A secondary analysis included ratings and feature calls on the definitely bronchiolitis and normal films.

Results

The effect of history on the diagnosis ratings is shown in Figure 1. Restricting analysis to the equivocal films, it is apparent that there is an overall effect of history, amounting to a scale change between one half and one unit on the 6-point scale, which was confirmed by the ANOVA [$F(1,23) = 7.80$, $MS_e = 2.00$, $p < .01$; df = 23 because one diagnostic rating was not filled in]. The apparent interaction between film type and positive/negative history was not confirmed [$F(1,23) = 0.77$, $MS_e = 2.00$, $p = .38$]. The raters were apparently unable to differentiate between equivocally abnormal and equivocally normal films [$F(1,23) = 0.31$, $MS_e = 3.97$, $p = .58$]. No other effects were significant.

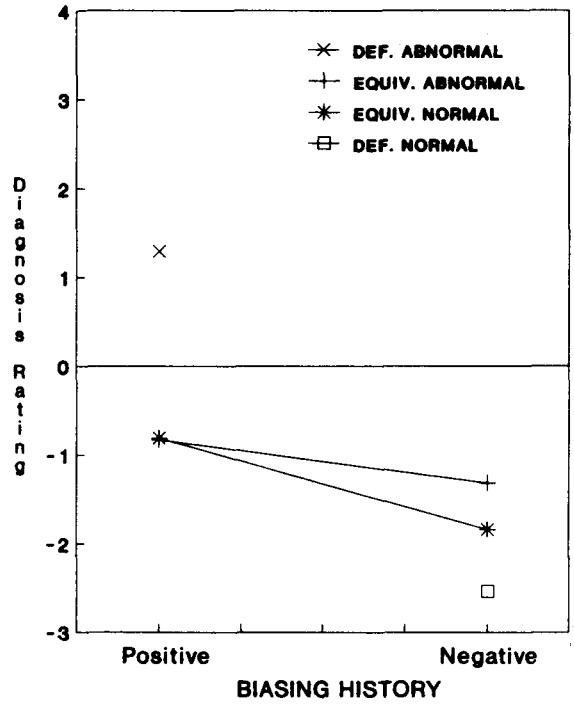

Figure 1. Mean diagnostic rating, ranging from 3 (*definitely bronchiolitis*) to −3 (*definitely normal*). To provide a plausible task context, definitely abnormal films were presented only with a clinical history positive for bronchiolitis, and definitely normal films were presented only with a history negative for bronchiolitis. For the equivocally abnormal and equivocally normal films, positive and negative histories were counterbalanced across subjects.

An effect of history on diagnosis ratings is mainly interesting as an indication that the standardized clinical history clearly had a measurable and consistent effect on the diagnostic process. The more interesting issue is whether the history had a demonstrable effect on the call of features. The first analysis examined the number of features identified on each equivocal film under the various conditions, with identification based on a call of either present or equivocal. This is shown in Figure 2. Also included in Figure 2 is the number of features identified by the 3 radiologists in Experiment 1, in which there was no biasing history, although these were not included in the statistical analysis. Again, there is an overall and significant effect of history, amounting this time to an increase of about one half of a feature per film [$F(1,24) = 5.03$, $MS_e = 1.20$, $p < .05$]. There was again no difference between normal and abnormal films [$F(1,24) = 2.73$, $MS_e = 3.72$, $p = .11$]. Because of the lack of a clear baseline (the data from Experiment 1 are only suggestive), we cannot clearly determine the size or direction of the effect of history. However, there is no doubt that the provision of history did affect the feature calls, which was the direct point of interest.

Detailed examination of the data indicated a possible "basement" effect, in that the within-cell standard deviations were positively related to the cell means. Accord-

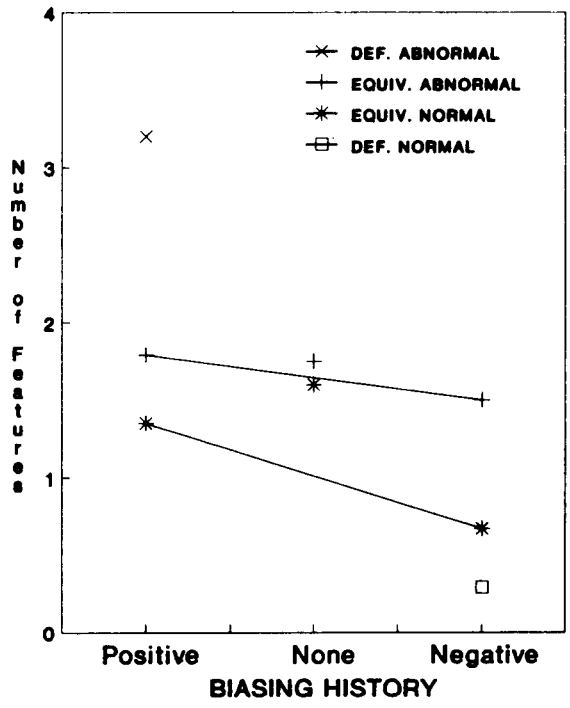

Figure 2. Mean number of features identified per film, out of five possible. Definitely abnormal films were only presented with a positive history for bronchiolitis, and definitely normal films were presented only with a negative history for bronchiolitis. For the equivocally abnormal and equivocally normal films, both positive and negative histories were presented.

ingly, the analysis was repeated, using log-transformed feature counts. The results were virtually identical [$F(1,24) = 4.93$, $p = .036$ with transformed data vs. $F(1,24) = 5.03$, $p = .034$ with the untransformed data for the main effect of history]. A second concern was that the effects resulted simply from changes in the "equivocal" rating of the features, amounting to a change in uncertainty, but no actual shift in the perception of features. Accordingly, the analysis was repeated, excluding any equivocal ratings from the statistics. The effects were similar, but the main effect of positive/negative history was marginal [$F(1,24) = 3.00$, $p = .09$]. In light of the reduced sample size of observations and the cruder scale, we believe that this result is consistent with the effects reported above.

No other main effects or interactions were significant. The lack of significant difference between equivocal bronchiolitis and equivocally normal films, both in the diagnosis ratings and in the features, can be interpreted favorably as a reflection of the appropriate selection of experimental materials, or unfavorably as an indication that the condition is intolerably indeterminate on the basis of radiological evidence.

If the apparent interaction between film type and history in Figures 1 and 2 had been significant, it would have suggested that the diagnosticians were using the history discriminatively rather than just as a response bias for resolving uncertainty about marginal features. Because the interaction is not significant, we cannot conclude that the history is being used discriminatively. However, we are also not in a position to affirm the null hypothesis and conclude that the effect of history is strictly setting a response bias. Post hoc analysis of the diagnostic ratings with the Tukey test demonstrated that the difference between the ratings for negative films was significant [$q(3,23) = 3.64$, $p = .05$], whereas the difference between ratings for the positive films was not [$q(3,23) = 1.75$, n.s.]. Similarly, Tukey tests of the feature calls showed that the difference between positive and negative history for normal films was significant [$q(3,23) = 3.77$, $p = .05$], whereas the difference for abnormal films was not [$q(3,23) = 1.54$, n.s.].

In an attempt to resolve this ambiguous outcome, we also performed analyses on the separate features, rather than on the total number of features. Figure 3 shows the proportion of equivocal films, under each condition, that were judged to contain each feature. It appears that the differences with the abnormal films are small, reflecting the small effect of the history on bronchiolitis films. By contrast, there was nearly a factor of 2 in the frequency of features from positive and negative history on the normal films. This interaction was significant for hyperinflation ($\chi^2 = 3.73$, $p < .05$), the most prevalent of the individual features, and it was in the direction of showing a larger difference due to history for normal than for abnormal films for each of the other features. This was formalized in a log-linear analysis, in which the partial association representing the film × history interaction was significant ($\chi^2 = 6.35$, $p = .01$).

Another perspective on the possible role of prior information comes from signal detection theory, where the potential effect is characterized as "bias," related to a change in the threshold at which a feature is declared present, or "discrimination," wherein the presence of the clinical information permits finer discrimination between normal variation and abnormal features. It is not possible to do a formal ROC analysis on individual features, since such analysis depends on systematically varying the threshold level, as was done by the 6-point rating scale on diagnosis. However, we can examine the change in the prevalence of individual features on equivocal bronchiolitis and normal films under the condition of positive and negative history. Examination of prevalence of features is not formally equivalent to a plot of true-positive and false-positive rates, since we have no prior reason to presume that the rate of occurrence of each feature on the bronchiolitis film was 100% or that the rate was 0% on normal films. Nevertheless, there should be a close relationship between this analysis and a formal ROC analysis.

We did find a differential effect of history. With a positive history, the average prevalence of individual features on bronchiolitis films was 32%, and with a negative history, it was 34% (paired t test $= -.64$, $df = 4$, n.s.).

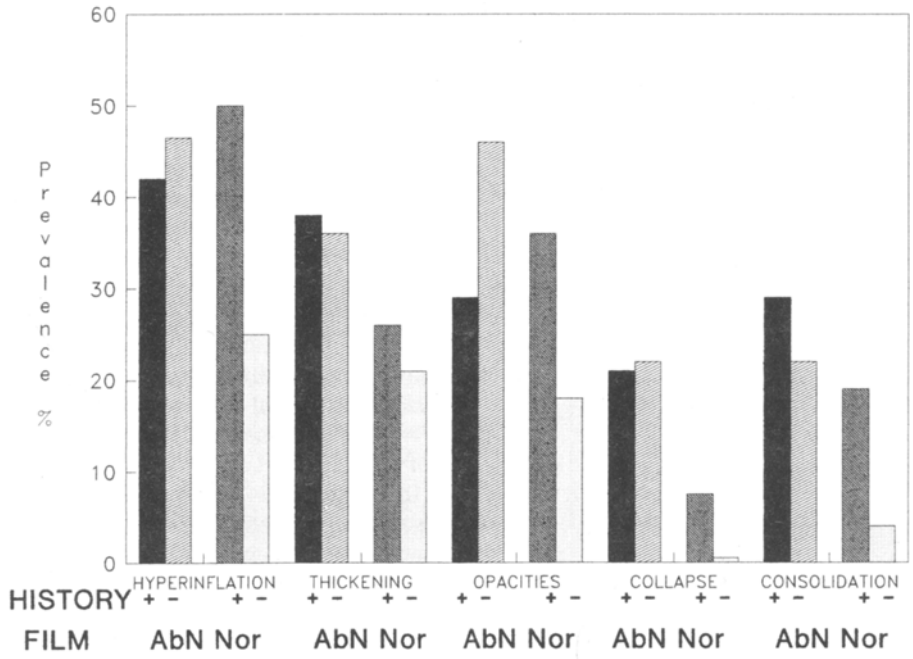

Figure 3. Judged presence of features of bronchiolitis in equivocally normal and equivocally abnormal films, as a function of clinical history being positive or negative for bronchiolitis.

By contrast, the prevalence of features on normal films changed significantly from 28% with a positive history to 14% with a negative history (paired t test = 3.82, df = 4, p = .02). Thus, this analysis demonstrated a significant effect of history on normal films and none on bronchiolitis films, consistent with the findings above. In a signal detection theory framework, the differential effect is consistent with a decreased discrimination with positive history.

The net effect of these analyses of the film type × history interaction is resolutely borderline. A significant effect is found when one considers the features individually, but not when one considers the total number read on each film. There is no prior or distributional consideration that would justify our relying on one rather than another of these statistics. This leaves us with no firm basis for concluding that the effect of histories acts primarily as evidence being added to that derived from the film, or as an occasion for extracting additional evidence from the film. Neither the testimony of the radiologists nor prior consideration of the task gives a clear reason to adopt either of these possibilities as the null hypothesis. However, as we have already described, there is evidence from the radiological literature that clinical history can add to the discriminative use of the evidence. The hypothesis that the history has the same effect in this experiment is thus not improbable, particularly in view of the low power resulting from the small number of subjects in the present experiment. As a result, we are clearly not justified in affirming the null hypothesis, even by default. Overall, although history definitely has a significant effect on feature calls, the borderline interaction leaves in doubt the issue of how the histories are producing their effect.

If we were to take this interaction between positive/negative film and positive/negative history at face value, it would appear that the effect is remarkably innocuous. The effect induced by history appears to occur primarily with negative history and normal films, amounting to a reduction in the number of features diagnosed when there is a normal history. There is no evidence that a positive history, as opposed to the absence of history, increased the number of false-positive calls. However, there is no reason to take the direction of these effects as general. As pointed out in Experiment 1, the distribution of items in the test phase is surely just as much a biasing condition on reading a film as is the clinical history that we provide. Half of the films were of bronchiolitis, many of them were radiologically definite, and the features being rated were restricted to those definitional for bronchiolitis. This could be expected to result in the radiologists' treating each film as being a possible case of bronchiolitis, unless indicated otherwise. Under conditions in which there are a variety of disorders, a history suggesting bronchiolitis might be expected to have more of an effect.

We are not proposing that the effects of history on feature calls and diagnosis ratings are independent. Indeed, there is a strong relationship between the diagnosis rating and the perception of positive features, with correlations ranging from 0.65 to 0.86, when one considers only ambiguous films. But the important demonstration is that a brief standardized history can affect both perception of features and judgment of the likelihood of diagnosis. The latter is almost self-evident; the former is not.

In summary, expert radiologists, who were at the top of their field, showed the effect of a brief standardized history on the features that they reported as present. The

effect amounted to an increase of about 25%-50% in the number of features identified on the film and a commensurate increase in the diagnosis rating. Both the analyses of the interaction between film type and history and the impact of history on individual features were consistently borderline, leaving no firm grounds to conclude that the history was acting either additively or interactively with the evidence from the film.

EXPERIMENT 3
Effect of Expertise

Experiment 2 demonstrated that the clinical history affected both the ratings of the diagnostic category and the identification of features on equivocal films by expert radiologists. It remained to be demonstrated whether radiology residents, who have substantially less experience, would be more or less susceptible to bias, and whether or not the bias would take the same form.

Method

The method was exactly the same as that for Experiment 2. Four readers were chosen from among first-year residents in the radiology program at McMaster University. Residents have completed the MD degree, and these individuals had acquired an average of about 8 months of specific training in radiology after graduation.

Results

First, to establish an effect of expertise, we analyzed the four types of films (definitely abnormal, equivocally abnormal, equivocally normal, and definitely normal) but restricted ourselves to only those data for which there was a consistent history (see Table 3). An analysis of the diagnostic ratings shows a large main effect of film type [$F(3,45) = 37.64, p < .0001$], a significant main effect of experience [$F(1,45) = 4.76, p < .05$], and a highly significant film × experience interaction [$F(3,45) = 4.50, p < .01$]. The main effect of experience is consistent with the expected advantage for expertise, although by a smaller margin than might have been hoped, considering the professional disparity between the two samples. The significant interaction suggests that the experts were more polarized in their ratings.

Similar results were found from an analysis of the number of features reported per film. There was a large effect of film type [$F(3,46) = 17.87, MS_e = 3.68, p < .001$], a significant main effect of experience [$F(1,46) = 7.47, MS_e = 1.05, p < .01$], and a marginal experience × film type interaction [$F(3,46) = 2.14, MS_e = 1.05, p = .10$], with experts again making more extreme feature calls. Overall, then, bronchiolitis is discriminable on the basis of the combination of clinical and radiological evidence, with the experts outperforming the residents.

Our main interest is in the equivocal films and the effect of manipulation of clinical history, shown in Table 4. It is apparent that there is an overall effect of history, amounting to a scale change between one half and one unit on the 6-point scale, which was confirmed by the ANOVA [$F(1,23) = 22.60, MS_e = 2.20, p < .0001$]. In this analysis, there was a significant difference between positive and negative films, with an average difference of about one half of a scale unit, again confirmed by the ANOVA [$F(1,23) = 6.42, MS_e = 1.71, p < .05$]. The marginally higher ratings of novices was not significant, nor was the apparent interaction between expert/novice and positive/negative history. No other effects were significant.

The effect of history on the number of features identified on each film under the various conditions is also shown in Table 4. Again, there is an overall and significant effect of history, amounting this time to an increase of about one half of a feature per film, a 25%-33% increase [$F(1,24) = 7.56, MS_e = 1.90, p = .01$]. There was also a marginal difference between normal and ab-

Table 3
Ratings of Certainty of Diagnosis and Number of Features Found

	Diagnostic Rating		No. Features	
Film Type	Expert	Resident	Expert	Resident
Definite bronchiolitis	1.46	0.76	3.21	3.08
Equivocal bronchiolitis	−0.83	0.27	1.79	2.29
Equivocally normal	−1.84	−1.57	0.67	0.89
Definitely normal	−2.54	−1.57	0.29	1.08

Note—Ratings are shown for four types of films and two levels of experience. Six-point scale used: +3 = definitely present; −3 = definitely not present.

Table 4
Effect of History and Level of Experience on Ratings of Certainty of Diagnosis and Number of Features Found

		Diagnostic Rating		No. Features	
Film Type	History	Expert	Resident	Expert	Resident
Equivocal bronchiolitis	positive	−0.83	0.27	1.79	2.29
	negative	−1.37	−1.12	1.50	1.79
Equivocally normal	positive	−0.80	−0.57	1.36	1.54
	negative	−1.84	−1.57	0.67	0.89

Note—Ratings are shown for two types of films and two levels of experience. Six-point scale used: +3 = definitely present; −3 = definitely not present.

normal films [$F(1,24) = 3.07$, $MS_e = 1.47$, $p < .10$]. There was a small significant effect of expertise [$F(1,24) = 5.02$, $MS_e = 5.45$, $p < .05$], with novices systematically reporting more features than did experts. No other main effects or interactions were significant.

The analysis was repeated, excluding all "equivocal" feature calls from the analysis. The results were substantially the same. There was a main effect of clinical history [$F(1,24) = 8.18$, $MS_e = 0.66$, $p < .01$], and a main effect of normal/abnormal film [$F(1,24) = 4.77$, $MS_e = 0.37$, $p < .05$]. In this analysis, however, novices reported fewer features than did experts. The apparent difference between novices and experts was, however, only marginally significant [$F(1,24) = 2.90$, $MS_e = 3.23$, $p = .10$]. No other effects were significant.

In the discussion of Experiment 2, the interaction between film type and history was examined for the equivocal films to determine whether the history was acting discriminatively on the reading of the features, or was only adding evidence for or against the hypothesis. There is no hint of this interaction of film type and history in the data of the residents. However, since the status of the interaction was borderline for the experts, we are hardly in a position to say that the pattern increased with experience.

GENERAL DISCUSSION

The concordance of changes in feature calls in Experiment 1 provided strong evidence that feature calls were being made in the context of an opinion about the overall diagnosis. This occurred despite the required task involving feature calls only, with no overt diagnosis being requested. The evidence for the dependence of feature and diagnostic judgments on prior biasing information was confirmed in Experiment 2, in which clinical history relevant to the overall diagnosis was experimentally manipulated for the equivocal films. The history affected both the ratings of the diagnostic category and the identification of features for both experts and residents. It is not clear whether the effect of diagnostic decisions (or consistency among features) on feature identification occurred because the general information was being added to the information on the film, or because it set the occasion for extracting differential information from the film. Nor is it clear whether this finding is generalizable to apply to other disease states, or to other experimental conditions in which multiple possibilities are considered. Nevertheless, since the experiments reported here contained many aspects that could influence judgment toward a diagnosis of bronchiolitis independent of the experimental manipulation (the rating form in particular), it is possible that the possible influence of prior information was underestimated. Residents performed less discriminatively than did the experts, both on feature calls and on diagnostic ratings, but this effect of experience was limited to the diagnosis of definite films.

There have been several theories in the radiology literature to explain the process of diagnosis. Common to all of them are a perceptual component, which rapidly recognizes patterns in the data, and a cognitive decision-making component, which evaluates the output of the perceptual stage and seeks further data when appropriate. However, the separation of feature identification from category judgments is rarely made explicit, although some authors have identified this as an issue. For example, Berbaum et al. (1990) did a study in which they imposed a simulated pulmonary nodule on normal and abnormal chest films. The presence of the simulated nodule resulted in a lower detection of abnormal features, a phenomenon Berbaum calls "satisfaction of search." In Berbaum's words: "Concluding that certain image features are indicative of one diagnostic category might make it difficult to detect features of another category" (p. 139).

Similarly, Lesgold et al. (1988) have described one aspect of expertise as the ability to recognize normal variations and to reinterpret and refine initial perceptual judgments in light of new information. Swanson, Feltovich, and Johnson (1977) also discussed a model of medical diagnosis involving the idea of the initial features triggering a diagnostic schema, which then guides search for additional features to ascertain the appropriateness of the schema. However, in the present studies, there is only one relevant schema, so it is difficult to see how such an explanation would account for our findings.

Treating the identification of features and the identification of larger units as interdependent decisions is a property of several models in psychology. For example, McClelland and Rumelhart's (1981) interactive activation model provides a mechanism in which activity on the word level is influenced by activity on the letter level, which in turn influences the activity at the word level. Since these authors were modeling word-superiority results, their direct interest was in reporting at the letter level, but information existed in the model that enabled them to also report at the word level. Of course, connectionist models in general, depending on how they are set up, can have the same properties. However, this interdependence of feature identification and higher unit identification is not a property of many "top-down" or "interactive" models, such as the linear decision unit (perceptron) models. The interaction referred to in these models often means that a decision about the higher unit is influenced by information both about features and about the processing context in which the higher unit is occurring. Decisions about the higher unit are typically not modeled as affecting the process by which the evidence itself is assessed, although there is no difficulty, in principle, in giving the models this property.

Models that have interdependent decisions on feature and higher unit levels capture an aspect of investigations with which most scientists are familiar. Data in a scientific investigation are potentially influenced by many other factors than the process under investigation. At a minimum,

a highly discrepant data point is a good candidate for recalculation or inspection for some other form of error. But, more generally, when a scientist has a convincing or generally useful theory, a nonconfirmatory experiment should be thought through to see whether the experimental arrangements and data analysis provide as appropriate a test as originally seems to be the case. In radiology, borderline findings might be reinspected to see whether a stronger interpretation of them could legitimately be made. However, this process need not be just the quantitative one of increasing one's certainty about the presence of a feature that has originally been taken to be marginal. The nature of the structure being investigated can also be in doubt. Again, to quote Lesgold et al. (1988), "Almost as if they were taking an embedded figures test, [the residents] were unable to see a collapsed lung tissue as occupying that region because they had already assigned it to normal arterial structure" (p. 331). Clearly, both expertise and expectations from valid clinical information can help to guide interpretation of the structure. As in the case of scientific investigations, rationalizing away discrepant findings to make an overall neat picture is a process that needs to be restrained.

It would be of practical importance if Lesgold's (1988) claim that experts are better able to discount normal variations or correct prior misconceptions than novices was apparent in this series. Instead, we found that both groups were equally susceptible to influence from a history. Thus, it would appear that experience acquired in the normal course of practice is insufficient to avoid such effects. Perhaps therein lies a practical implication. From the literature that we reviewed, it is apparent that radiologists do not give adequate attention to the distinction between features and inferential judgments, nor are they apparently aware that features themselves are subject to interpretation. Training might focus less on the combination of features (i.e., "What are the features of aortic aneurism?") and more on the source and interpretation of the features themselves. This study shows that consideration of multiple determinations of data and fidelity to underlying processes can play a role in the investigations conducted by medical practitioners as well as those in science. However, on the presumption that these radiologists are normally receiving valid history—that is, not from psychologists—there is no reason to believe that the influence of history on their feature and diagnostic calls is normatively improper.

REFERENCES

BARROWS, H. S., NORMAN, G. R., NEUFELD, V. R., & FEIGHTNER, J. W. (1982). The clinical reasoning process of randomly selected physicians in general medical practice. *Clinical & Investigative Medicine*, **5**, 49-56.

BERBAUM, K. S., EL-KHOURY, G. Y., FRANKEN, E. A., KATHOL, M., MONTGOMERY, W. J., & HESSON, W. (1988). Impact of clinical history on fracture detection with radiography. *Radiology*, **168**, 507-511.

BERBAUM, K. S., FRANKEN, E. A., DORFMAN, D. D., BARLOON, T., ELL, S. R., LU, C. H., SMITH, W., & ABU-YOUSEF, M. M. (1986). Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. *Investigative Radiology*, **21**, 532-553.

BERBAUM, K. S., FRANKEN, E. A., DORFMAN, D. D., ROOHOLAMINI, S. A., KATHOL, M. H., BARLOON, T. J., BEHIKE, F. M., SATO, Y., LU, C. H., EL-KHOURY, G. Y., FLICKINGER, F. W., & MONTGOMERY, W. J. (1990). Satisfaction of search in diagnostic radiology. *Investigative Radiology*, **25**, 133-140.

BROOKS, L. R., NORMAN, G. R., & ALLEN, S. W. (1991). The role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, **120**, 278-287.

COBLENTZ, C. L., BABCOOK, C. J., ALTON, D., RILEY, B. J., & NORMAN, G. R. (1991). Observer variation in detecting the radiologic features associated with bronchiolitis. *Investigative Radiology*, **26**, 115-118.

COHEN, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.

DOUBILET, P., & HERMAN, P. G. (1981). Interpretation of radiographs: Effect of clinical history. *American Journal of Radiology*, **137**, 1055-1058.

FISCHHOFF, B., & BEYTH-MAROM, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, **90**, 239-260.

GOOD, B. C., COOPERSTEIN, L. A., DEMARINO, G. B, MIKETIC, L. M., GENNARI, R. C., ROCKETTE, H. E., & GUR, D. (1990). Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? *American Journal of Radiology*, **154**, 709-712.

GUPPY, K. H., DETRANO, R., ABBASSI, N., JANOSI, A., SANDHU, S., & FROELICHER, V. (1989). The reliability of probability analysis in the prediction of coronary artery disease in two hospitals. *Medical Decision Making*, **9**, 181-189.

HOLDAWAY, D., ROME, A. C., & GARDNER, P. S. (1967). The diagnosis and management of bronchiolitis. *Pediatrics*, **39**, 924-928.

LESGOLD, A. M., RUBINSON, H., FELTOVICH, P., GLASER, R., KLOPFER, D., & WANG, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M. T. H. Chi, R. Glaser,& M. Farr (Eds.), *The nature of expertise* (pp. 322-351). Hillsdale, NJ: Erlbaum.

LUSTED, L. B. (1968). *Introduction to medical decision making*. Springfield, IL: C. C. Thomas.

MARKUS, J. B., SOMERS, S., O'MALLEY, B. P., & STEPHENSON, G. W. (1989). Double-contrast barium enema studies: Effect of multiple readings on perception error. *Radiology*, **175**, 155-156.

MCCLELLAND, J. L., & RUMELHART, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, **88**, 375-407.

MYLES-WORSLEY, M., JOHNSTON, W., & SIMONS, M. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 553-557.

NORMAN, G. R., BROOKS, L. R., & ALLEN, S. W. (1989). Recall by expert medical practitioners as a record of processing attention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 1166-1174.

SCHREIBER, M. H. (1963). The clinical history as a factor in roentgenographic interpretation. *Journal of the American Medical Association*, **185**, 137-139.

SLOVIC, P., RORER, L. G., & HOFFMAN, P. J. (1971). Analyzing use of diagnostic signs. *Investigative Radiology*, **6**, 18-26.

SWANSON, D. B., FELTOVICH, P. J. & JOHNSON, P. E. (1977). Psychological analysis of physician expertise: Implications for design of decision support systems. In D. B. Shires & H. L. Wolf, (Eds.), *Proceedings of the Second World Conference on Medical Informatics* (pp. 161-164). Amsterdam, North-Holland.

SWENSSON, R. G. (1980). A two-stage detection model applied to skilled visual search by radiologists. *Perception & Psychophysics*, **27**, 11-16.

WIGTON, R. S. (1988). Use of linear models to analyze physicians' decisions. *Medical Decision Making*, **8**, 241-252.

(Manuscript received April 30, 1991;
revision accepted for publication April 3, 1992.)