Cue Selection and Category Learning: A Systematic Comparison of Three Theories[1]

Richard P. Cooper
Birkbeck, University of London, United Kingdom[2]

Peter Yule
Birkbeck, University of London, United Kingdom

John Fox
Cancer Research United Kingdom

We evaluate three approaches to tasks involving categorisation with probabilistic cues (the Bayesian approach, the associationist approach, and the hypothesis testing approach) by comparing the behaviour of three classes of cognitive model with that of human participants on a simulated medical diagnosis task. The task yields dependent measures relating to both categorisation accuracy and cue selection. A systematic exploration of the effects of processing biases within the models reveals that all three approaches are able to account for the effects in the human data, provided that appropriate performance factors and processing biases are incorporated. The discussion focuses on the methodology used to evaluate the approaches and on the role of performance factors and processing biases within the various models.

Keywords: Category learning; Cue selection; Bayesian reasoning; Hypothesis testing; Comparative modelling; Indistinguishability.

Introduction

Many cognitive tasks, ranging from object recognition to medical diagnosis, involve elements of categorisation. Frequently such tasks also involve

[1] This work was supported by grant GR/M89621 from the UK's Engineering and Physical Sciences Research Council

[2] Address for correspondence: Dr. R. Cooper, School of Psychology, Birkbeck, University of London, Malet Street, London WC1E 7HX, United Kingdom. E-mail: r.cooper@bbk.ac.uk (Fax: +44-20-7631-6312). Home page: http://www.psyc.bbk.ac.uk/staff/rc.html

unreliable cues. Thus, in medical diagnosis a cue (i.e., symptom) may be highly suggestive of a particular category (i.e., disease), but the category may still occur in the absence of the cue. A normative, mathematically optimal, approach to such tasks is provided by probability theory and Bayes' theorem. Within this approach, conditional probabilities of cues given categories are used to calculate the probability that a set of cues corresponds to a given category. The conditional probabilities may be estimated on the basis of experience and categorisation may proceed by selecting the most likely category for a given set of cues.

Several studies have compared human performance on categorisation tasks with that predicted both by the Bayesian approach and by alternative non-normative approaches. For example, Fox (1980) demonstrated that a process-oriented approach based on generating and testing hypotheses (e.g., "the current case is an instance of category *x*") yielded a good approximation to human behaviour on a sequential categorisation task (described in detail below). This approximation was no worse than that of the Bayesian approach. In a similar vein, Gluck & Bower (1988) compared the Bayesian approach with an associationist approach motivated by animal learning theory. This approach based categorisation on weighted associations between cues and categories. Gluck & Bower found that both approaches provided good approximations to human performance.

The Bayesian, associationist, and hypothesis testing accounts of categorisation represent three distinct schools of thought. The Bayesian approach is logically optimal and grounded in sound, well-understood, mathematics. On the empirical side, many studies (including Fox, 1980, Gluck & Bower, 1988) have reported reasonable fits between human behaviour and that predicted by Bayesian theory, and further robust empirical findings (e.g., base-rate neglect: Medin & Edelson, 1988; Kruschke, 1996) may be accounted for in terms of systematic departures from the Bayesian norm. In addition, the strategy of accounting for non-normative behaviour in terms of systematic heuristics or biases affecting a normative system has been successfully employed in the related fields of judgement and decision making (Kahneman, Slovic, & Tversky, 1982). However, the approach provides limited insight into the underlying processing mechanisms — do people actually manipulate conditional probabilities, albeit imperfectly, in their heads? — and on precisely why or when departures from the norm may arise. This undermines the predictive power of the approach: it is not always clear which combination of biases or heuristics should apply in any particular situation, or indeed if a new bias or heuristic should be posited (cf. Gigerenzer & Todd, 1999).

The associationist approach also has a sound mathematical grounding in the delta rule of learning (which can be shown under certain conditions to converge to an optimal solution: see, e.g., Hertz, Krogh, & Palmer, 1991). In addition, the mechanisms employed are continuous with those postulated in the animal learning literature (Rescorla & Wagner, 1972). While empirical studies (e.g., Gluck & Bower, 1988; Estes, Campbell, Hatsopoulos, & Hurwitz, 1989) have found support for the associationist approach, it does not provide an account of how verbal information (such as information concerning the probability of relevant events) may affect behaviour. Further, if the task requires a representational mapping that lacks certain formal properties (specifically, that the input vectors are linearly separable with respect to each output feature) then learning algorithms that are generally argued to be biologically implausible are required (cf. Crick, 1989).

The hypothesis testing approach treats categorisation as involving the explicit generation and testing of competing hypotheses coupled with strategies aimed at, for example, verifying one hypothesis or discriminating between alternative hypotheses. The general approach may be traced to the work of Bruner, Goodnow, & Austin (1956), and is instantiated in the "fast and frugal" heuristics of Gigerenzer and colleagues (Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 1999). The approach lacks the mathematical foundation of the other two approaches, and as such may appear to be *ad hoc*. It makes up for this by providing an intuitive account of intermediate processes in decision making. Furthermore, because the approach is process-oriented it is able to address effects of processing limitations such as decay of information from working memory, and is most amenable to implementation within a cognitive architecture such as ACT-R (Anderson & Lebiere, 1998).

It should be noted that none of the approaches as sketched here is sufficiently detailed to yield precise predictions about human behaviour in realistic tasks. The approaches are effectively frameworks within which more detailed theories may be developed. Nevertheless, the approaches may be evaluated by considering a range of detailed theories grounded within each. This is the strategy adopted here: we evaluate the three approaches by comparing the behaviour of three classes of computational model with that of human participants on a simulated medical diagnosis task. The task requires flexible use of acquired knowledge and yields multiple dependent measures relating to categorisation accuracy, cue selection and learning. We find that the principal effects within the data may be accounted for by models based on each of the three approaches, provided that the models incorporate performance factors and processing biases such as a memory decay or a bias towards confirmation or the use of positive cue/category associations.

A second issue explored in the current work concerns cue selection during categorisation. In their standard forms, the various approaches assume that complete cue information is available to the categorisation mechanism, and most empirical and computational studies have focused on categorisation accuracy rather than cue selection. (Exceptions include Bruner *et al.* (1956), Fox (1980) and, to a lesser extent, Berretty *et al.*, (1999).)

However, the assumption of complete cue availability is not always valid. In naturalistic medical diagnosis, for example, patients typically present with one specific cue, and the medical practitioner must seek additional information to complete the categorisation task (e.g., by querying the presence of other symptoms or by ordering appropriate laboratory tests). Efficient categorisation under these conditions requires that symptoms are queried in order of validity or informativeness (and hence that some mechanism for determining informativeness is available), and that a categorisation judgement may be made in the absence of complete cue information. We therefore consider extensions to each class of model that allow cue selection and categorisation with incomplete information.

Background: the medical diagnosis task

Categorisation has frequently been investigated through tasks based on medical diagnosis (e.g., Fox, 1980, Medin *et al.*, 1982; Gluck & Bower, 1988; Medin & Edelson, 1988; Estes *et al.*, 1989; Shanks, 1991; Kruschke, 1996; Ross, 1997). The basic medical diagnosis task involves participants associating symptom configurations with likely diseases (Ledley & Lusted, 1959). The version of medical diagnosis investigated here most closely follows that investigated by Fox (1980). This task was chosen because it provides quantitative data on 1) diagnostic or categorisation accuracy; 2) the information or cues employed in making those categorisations; and 3) the relative priority of that information. Though similar in spirit to Fox's original work, this paper represents a significant advance by considering both a wider range of models of diagnosis and the effects of learning, both in participants and in models grounded in all three theoretical approaches.

The structure of the task is as follows: Participants take the role of a doctor attempting to diagnose a series of patients. They are presented with an initial symptom (e.g., the patient is vomiting) and allowed to query the presence of a limited set of other symptoms before giving a diagnosis. Feedback (in the form of the actual diagnosis) is then given, allowing participants to learn the task. Dependent measures included diagnostic accuracy, the number of symptoms queried in making a diagnosis, and the ordering of those symptom queries.

Although considerably simpler than real-world medical diagnosis, this task has a number of desirable features. Like other rule induction tasks (such as Kendler & D'Amato's (1955) concept learning experiments, Wason's (1960) 2-4-6 task, or Gluck & Bower's (1988) categorisation version of the medical diagnosis task) it provides data on human learning. The difficulty of the task may also be adjusted (by modifying the number of symptoms and diseases involved, or by varying the conditional probabilities that relate symptoms to diseases). The task also offers a number of complexities beyond those found in other learning and rule induction tasks, because participants are initially only provided with information relating to one symptom. Participants must actively seek more information if they require it. Many participants realise that they do not always need to know about all symptoms to make a diagnosis, and reliable questioning strategies emerge. It is thus possible to examine both the information used by participants in making their diagnoses and the order in which that information is sought.

One disadvantage of the task, however, results from the possibility of a trade-off between diagnostic accuracy and information sought. Querying more symptoms often provides more information and allows a more accurate diagnosis. Participants may therefore attempt to maximise their diagnostic accuracy by querying most or all symptoms, or minimise their symptom querying at the expense of compromising diagnostic accuracy. Group data may blur individual differences in this trade-off.

Models of the diagnosis task

Evaluation of the three approaches presents significant methodological difficulties. The strategy adopted here is to develop computational models representative of each framework. Since the frameworks are not fully specified, each approach is embodied within a model having several parameters. Many of those parameters are continuous valued. The result is a high-dimensional space of models. This section presents the models and their parameters. Later sections employ an evaluation methodology in which canonical models are specified within each framework, the behaviour of the canonical models is compared to human data, and the effects of model parameters on this behaviour explored. This approach allows the strengths and weaknesses of the various approaches to be determined. It also clarifies the role of each of the parameters and how they may relate to human behaviour.

Common aspects of the models

All models were developed within COGENT (Cooper & Fox, 1998; Cooper, 2002), a modelling environment that allows models to be specified by constructing and then fleshing out box and arrow diagrams. The environment was used both for the development of the models and for building the computational infrastructure necessary to simulate the experimental conditions under which human participants were tested (including randomisation of stimulus materials, control of stimulus presentation, responding to symptom queries made by the models, and collection and analysis of data). This infrastructure was fixed across all models to ensure that they were functioning under identical experimental conditions, and thus to facilitate comparison of their behaviour.

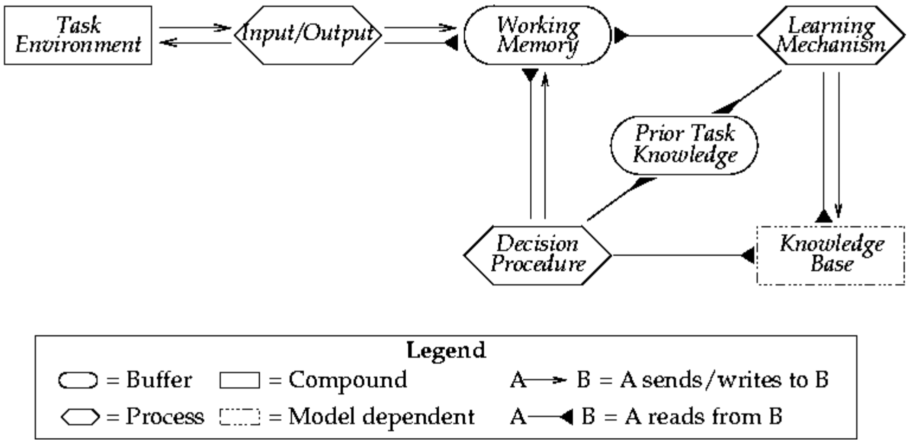

Figure 1: The functional components of all models. The form of *Knowledge Base* (and communication with it) differs between classes of model. In Bayesian and hypothesis testing models it is a buffer, storing quantitative or categorical information about event co-occurrences. In the associationist models it is an associative network, mapping known symptoms to possible diseases.

The "topology" of all models (in terms of the functional sub-components and their interactions) was also fixed to further aid comparability (see Figure 1). The boxes and their functions are as follows:

- *Task Environment* is a compound box containing the infrastructure described above. It is not part of the cognitive model.
- *Input/Output* models the interface between central cognition and the environment. Messages from *Task Environment* trigger additions to *Working Memory* (e.g., adding information about the presence of a symptom), and the existence of appropriate elements in *Working Memory* triggers generation of queries (e.g., Is symptom *x* present?) and diagnoses.
- *Working Memory* is a data store in which information about the current case is stored and manipulated. The representation of that information (quantitative probabilities or symbolic propositions) is dependent upon the type of model.
- *Prior Task Knowledge* contains the complete set of possible symptoms and diseases. This constrains the symptoms that may be queried and the diagnoses that may be generated.
- *Knowledge Base* contains acquired knowledge of symptom/disease associations. In the Bayesian models this knowledge is represented in a frequentistic form, in the associative models it is stored within an

associative network, and in the hypothesis testing models it is represented via symbolic propositions.

- *Decision Procedure* contains production-style rules that combine information from various sources in order to select either a symptom to query or a disease to proffer as a diagnosis. The form of the rules varies with the class of model, but in all cases they consult *Working Memory* (for knowledge of the current trial), *Prior Task Knowledge* (for the set of possible symptoms and diseases) and *Knowledge Base* (for acquired symptom/disease associations).
- *Learning Mechanism* acquires and updates information in *Knowledge Base*. In all cases this process is triggered by the presence of a confirmed diagnosis in *Working Memory*, but the details of the mechanism depend on the representational format of the knowledge being acquired, and hence differ significantly between the classes of model.

Bayesian models

In the Bayesian models *Decision Procedure* implements a Bayesian probability-revision system with a symptom selection procedure based on symptom informativeness (Shannon & Weaver, 1949, Wiener, 1948), as used in related work on categorisation (e.g., Lindley, 1956, Fox 1980, Oaksford & Chater, 1994). This requires that *Knowledge Base* contains information that can be used to derive conditional probabilities of symptoms given diseases. The models assume conditional independence of symptoms given each disease (i.e., they are naïve Bayesian models) — a fact that is true of the experimental task employed below — and employ a frequentistic representation (Gigerenzer & Hoffrage, 1995), in which the frequencies of relevant events are maintained. *Learning Mechanism* updates frequency information upon receipt of feedback at the end of each trial. Such models have been used in practical implementations of, for example, full-scale expert systems (see, e.g., Castillo, Gutiérrez, & Hadi, 1997).

The decision procedure

The Bayesian version of *Decision Procedure* uses the standard Bayesian equation for posterior probability revision:

$$p(D_i \mid S_j) = \frac{p(S_j \mid D_i).p(D_i)}{\sum_i p(S_j \mid D_i).p(D_i)} \qquad [1]$$

where *D* ranges over diseases and *S* ranges over symptoms.

All diseases are initially assumed to be equally likely. Whenever a symptom is discovered to be present or absent Equation 1 is used to calculate the posterior probabilities of all candidate diseases. The conditional probabilities of symptoms given diseases are estimated from the frequency information stored in *Knowledge Base*.

To illustrate, suppose the presenting symptom is vomiting. After applying Equation 1 for each disease *Working Memory* may include the following (where the left column indicates the processing cycle on which the element appeared in *Working Memory* and diseases with a probability of zero are ignored):

```
4:  probability(malengitis, 0.250).
4:  probability(ritengitis, 0.250).
4:  probability(tepittitis, 0.500).
2:  told(vomiting, present).
```

If the posterior probability of any disease exceeds the diagnostic threshold (a parameter of the Bayesian models) then that disease is selected as a diagnosis. Otherwise, the posterior probabilities are taken as new priors, the expected information gain of each of the remaining symptoms is calculated, and the symptom with the greatest expected information gain is queried.

The expected information gain of each symptom is determined from the standard Shannon-Wiener information theoretic formula:

$$I(S) = \sum_x \left[\sum_i \left\{ p(D_i \mid S^x).\log p(D_i \mid S^x) \right\} p(S^x) \right] - \sum_i p(D_i).\log p(D_i) \qquad [2]$$

where x = `present` or `absent`, and

$$p(S^x) = \sum_i p(S^x \mid D_i).p(D_i) \qquad [3]$$

In the example, this results in several items being added to *Working Memory*:

```
6:  informativeness(headache, 0.488).
6:  informativeness(earache, 0.488).
6:  informativeness(stiffness, 0.562).
6:  informativeness(pyrexia, 0.562).
```

In this case the greatest expected information gain is shared by two symptoms, stiffness and pyrexia. One of these is chosen at random and queried. *Task Environment* then replies to the query by referring to the current patient case. This process repeats until a diagnosis is made. If all symptoms are queried but no disease's probability reaches the diagnostic threshold then the disease with highest probability is selected as the diagnosis.

Learning

Learning Mechanism is triggered at the end of each trial by receipt of feedback about the correct diagnosis. For Bayesian models learning involves updating frequency counts: for each symptom/disease pair, *Learning Mechanism* records the number of times the symptom is observed to be present with the disease, and the number of times it is observed to be absent with the disease. Thus, for each disease and symptom *Knowledge Base* contains a proposition of the form:

```
frequency(Disease, Symptom, NumPresent, NumAbsent).
```

where `NumPresent` and `NumAbsent` are respectively the number of present and absent occurrences of the symptom given the disease.

Decision Procedure uses frequency information to estimate conditional probabilities according to Equation 4:

$$p(\texttt{Symptom}|\texttt{Disease}) = \frac{\texttt{NumPresent}}{\texttt{NumPresent} + \texttt{NumAbsent}} \qquad [4]$$

This approximates the true conditional probability, and, assuming unbiased exposure to cases, grows closer to it with experience.

Variants within the Bayesian class

The Bayesian approach to the diagnosis task is mathematically optimal in that it can be shown to maximise diagnostic accuracy while minimising the number of symptoms that are queried. There is a tension between the requirements of maximising accuracy while minimising queries, and the diagnostic threshold allows these requirements to be traded off against each other. A high threshold will maximise accuracy. A low threshold will reduce the number of symptoms queried. The diagnostic threshold therefore represents one parameter that may be varied to yield different behaviours within the class of Bayesian models.

Human behaviour on categorisation tasks is rarely optimal. Arguably sub-optimal performance is the result of factors such as memory limitations and/or processing biases. It is therefore appropriate to consider how such factors might be incorporated into the Bayesian model. Three obvious ways relate to learning: the model may begin with preconceptions about the relations between cues and categories, it may be inefficient at learning, and acquired information may decay over time.

The model's preconceptions about cue/category relations may be manipulated by manipulating the initial values of `NumPresent` and `NumAbsent`. (Note that Equation 4 is undefined if `NumPresent` and `NumAbsent` are zero.) The assumption that symptoms are common may be incorporated by initialising `NumPresent` to a higher value than `NumAbsent`. The reverse assumption, that symptoms are rare, may be incorporated by initialising `NumPresent` to a lower value than `NumAbsent`.

In fact, `NumPresent` and `NumAbsent` may also be used to manipulate the rate of learning. If one compares two Bayesian models, one in which `NumPresent` and `NumAbsent` are doubled relative to the other, then although both will make the same initial assumptions about symptom rarity the one with higher values will learn more slowly because each additional case will have a smaller impact on the ratio in Equation 4. Thus, learning rate within the Bayesian models may be manipulated via the initial values of `NumPresent` and `NumAbsent`.[3]

Decay within the Bayesian model may be incorporated by assuming that frequency counts revert to zero in the absence of increments through learning. Equation 5 specifies the decay function adopted here:

$$\begin{aligned}\texttt{NumPresent} \text{ on trial } t{+}1 &= (1 - \text{Decay}) \times \texttt{NumPresent} \text{ on trial } t\\ \texttt{NumAbsent} \text{ on trial } t{+}1 &= (1 - \text{Decay}) \times \texttt{NumAbsent} \text{ on trial } t \qquad [5]\end{aligned}$$

A final factor of interest relates to the substantial body of empirical evidence that suggests that participants focus more on positive instances or properties, such as the presence of a symptom, than on negative instances or properties, such as the absence of a symptom (e.g., Wason & Johnson-Laird, 1972; Hunt & Rouse, 1981; Hearst, 1991). The effect of such a focus within the Bayesian model may be examined by calculating the informativeness of a symptom (Equation 2) under the assumption that the symptom is present (i.e., with x restricted to `present` in Equation 2). For completeness, one may also consider a model that calculates informativeness under the alternative assumption, that the symptom is absent.

Associationist models

In the associationist class of models, *Knowledge Base* comprises a single layer feed-forward network that learns symptom/disease associations through standard associationist techniques. The network is used by *Decision Procedure* to guide the querying of symptoms and to determine diagnoses.

The decision procedure

Knowledge Base contains one input node for each symptom (five for the task considered below) and one output node for each disease (four for the task considered below). When an element of the form `told(Symptom, Value)` appears in *Working Memory*, *Decision Procedure* constructs an input vector from the currently known symptom information (with symptoms known to be present coding as +1, symptoms known to be absent as –1 and the rest as 0) and sends it to *Knowledge Base*. The resulting output vector contains an activation value for each disease. If any disease node's activation exceeds the model's diagnostic threshold then that disease is selected as a diagnosis.

[3] In standard Bayesian terms, `NumPresent` + `NumAbsent` is the assumed *equivalent sample size*. Increasing the equivalent sample size effectively reduces the learning rate.

As in the other models, if no diagnosis is possible based on the current information then the model should select and then query a symptom. Associationist principles do not generalise well to this aspect of the task. One might attempt to develop an account based on recurrent connectionist techniques, but learning presents serious difficulties within such an approach. The alternative adopted here is as follows. For each remaining symptom, *Decision Procedure* calculates the diagnostic certainty that would result from discovering it to be present and the diagnostic certainty that would result from discovering it to be absent. It then queries the symptom with greatest diagnostic certainty on either of these measures. If no disease's activation exceeds the threshold and all symptoms have been queried, then the disease whose activation is greatest is selected as the diagnosis.

Learning

Learning Mechanism is triggered by the appearance of disease feedback in *Working Memory*. This results in the network being trained using the delta rule (cf. Hertz *et al.*, 1991) with the known symptom pattern as the input vector and the known disease pattern as the target output vector.

Variants within the associationist class

As in the case of the Bayesian class of models, several parameters govern the behaviour of specific instances of associationist models. The diagnostic threshold and learning rate are continuous-valued parameters ranging from 0.00 to 1.00 that govern the degree of diagnostic certainty required for a diagnosis and the degree to which new symptom/disease patterns are integrated into the network. The diagnostic threshold is directly analogous to the corresponding parameter in the Bayesian class of models. Manipulation of the associationist learning rate is equivalent to scaling the initial values of `NumPresent` and `NumAbsent` in the Bayesian class of models.

The other parameters of the Bayesian models also have analogues within the associationist class. *Knowledge Base* decay may be examined by adding noise to weights in the associative network, such that weights tend to "blur" in the absence of learning. Initial assumptions concerning the distribution of symptom/disease associations may be incorporated by selecting appropriate initial weights. Initial weights are assumed to be randomly distributed. If the distribution is centred on 0.00 the model begins in an unbiased state. Adopting a distribution centred on –1.00 implements the assumption that symptoms are rare. The alternative assumption, that symptoms are common, may be incorporated by adopting a distribution centred on +1.00.

A bias towards positive or negative information in symptom querying may also be incorporated by varying the way in which the change in diagnostic certainty is calculated. Thus, unbiased models take into account the possibility of both positive and negative responses (i.e., symptom present and symptom absent responses) when determining which symptom to query. Positivist models only take account of positive responses, and negative models only take account of negative responses.

Associationist models might vary along several other dimensions (e.g., degree of connectivity, learning rule, activation function, activation range). The space of such models is vast. In the interests of tractability, all associationist models reported here employ 100% connectivity, delta rule learning and a sigmoidal activation function with a range of –1 to +1.

Hypothesis testing models

The hypothesis testing models are based on the explicit representation and manipulation of symbolic hypotheses concerning possible diagnoses. They represent information in a propositional form and develop their hypotheses in response to the presence or absence of symptoms. Hypotheses are then tested, using one or more of a variety of strategies, such as verification, elimination, or discrimination (each of which is described below).

The decision procedure

The hypothesis testing models' *Decision Procedure* is a set of inference rules that modify *Working Memory*, implementing one or more hypothesis testing strategies. Processing is initiated when *Input/Output* receives a presenting symptom from *Task Environment*. The corresponding proposition (of the form `told(Symptom, present)`) is added directly to *Working Memory*. This triggers a rule in *Decision Procedure* that augments *Working Memory* with the names of all diseases that are suggested by the presenting symptom. These diseases constitute the model's initial hypotheses about the hypothetical patient's condition. The hypothesis generation rule is independent of strategy and may be expressed as:

IF: `told(Symptom, Value)` is in *Working Memory* and
`association(Disease, Symptom, Value)` is in *Knowledge Base*

THEN: add `hypothesis(Disease, suspected)` to *Working Memory*

Terms within the rule that begin with an upper-case letter are variables, which may be bound to specific symptoms or diseases as appropriate. Thus, the rules are generic with respect to symptoms and diseases, but may be instantiated for the particular task by the participant's beliefs about symptom/disease associations. These beliefs are stored in *Knowledge Base*.

The presence in *Working Memory* of hypotheses concerning suspected diseases prompts the generation of symptom expectations through a second *Decision Procedure* rule:

IF: `hypothesis(Disease, suspected)` is in *Working Memory* and
`association(Disease, Symptom, Value)` is in *Knowledge Base* and
`told(Symptom, AnyValue)` is not in *Working Memory*

THEN: add `expectation(Disease, Symptom, Value)` to *Working Memory*

At this stage *Working Memory* might contain the following elements:

```
5:  expectation(tepittitis, pyrexia, present).
5:  expectation(tepittitis, vomiting, present).
5:  expectation(ritengitis, vomiting, present).
4:  hypothesis(tepittitis, suspected).
4:  hypothesis(ritengitis, suspected).
3:  told(vomiting, present).
```

Propositions of the form `expectation(Disease, Symptom, Value)` are used in symptom query selection. Three possible strategies are:

- *Verification:* Query any symptom expected to be present given any suspected diseases. If the queried symptom is present it will help verify one or more suspected diseases.
- *Elimination:* Query any symptom expected to be absent given any suspected diseases. If the queried symptom is present it will allow one or more suspected diseases to be eliminated.
- *Discrimination:* Query any symptom that is expected to be present given one suspected disease but absent given another. If the symptom is present it will allow one or more suspected diseases to be eliminated and provide support for one or more other suspected diseases.

Verification is a simple confirmatory strategy based upon positive associations between symptoms and diseases. It is prone to error in that, once a hypothesis has been generated, supporting evidence is sought in preference to falsifying evidence. There is a substantial body of evidence that suggests that human judgement is subject to such a confirmation bias (e.g., Wason, 1960; Klayman & Ha, 1987). Elimination seeks evidence that will allow possible hypotheses to be eliminated. It is related to the Categorisation by Elimination heuristic of Berretty *et al.* (1999) which in turn is related to Tversky's theory of Elimination by Aspects (Tversky, 1972). Discrimination is a hybrid strategy that seeks evidence that will, where possible, simultaneously support one hypothesis while refuting another.

Each strategy specifies how querying should proceed in certain circumstances. They do not specify how querying should proceed if those circumstances do not hold. However, if one strategy fails one may fall back on another. For example, one may improve upon pure discrimination by adopting verification when the preconditions for discrimination are not met. We refer to this strategy as *discrim/verify*.

To pursue just one strategy in more depth, verification is reflected in the following query selection rule:

IF: `expectation(Disease, Symptom, present)` is in *Working Memory*
and told(`Symptom, AnyValue)` is not in *Working Memory*

THEN: add `query(Symptom)` to *Working Memory*

This rule differs from the previous ones in that, when its conditions are satisfied, it should fire for just one instantiation of its variables. (The previous rule may fire for all instantiations simultaneously.) The rule augments *Working Memory* with a prompt for *Input/Output* to query a specific symptom:

```
6:  query(pyrexia).
```

The response to this query (generated by *Task Environment* with reference to a simulated patient case) appears in *Working Memory* several cycles later:

```
9:  told(pyrexia, absent).
```

In the current example this counts against `tepittitis`, because `pyrexia` is believed to be present in cases of `tepittitis`. A further rule is therefore required to delete hypotheses (and expectations) from *Working Memory* when they conflict with known symptom information.

Symptom querying continues with the above rules until all symptoms have been queried or a single diagnosis remains. *Learning Mechanism* maintains a set of symptom patterns corresponding to each disease. If the current symptom pattern corresponds to one and only one disease, then the diagnosis rule (not shown) will fire, proffering that disease as the diagnosis.

The ordering of hypothesis generation and symptom querying depends upon both strategy and the recall characteristics of *Knowledge Base* and *Working Memory*. Access to *Working Memory* is determined by primacy (i.e., First-In/First-Out), but access to *Knowledge Base* is determined by recency (i.e., Last-In/First-Out). The net result is that recently acquired information is used in generating hypotheses and possible symptom queries (through access to *Knowledge Base*), but behaviour is ultimately controlled by the first such hypothesis or symptom query to enter *Working Memory*. These access characteristics are based on previous work (Fox, 1980; Fox & Cooper, 1997) that found them to yield plausible symptom query biases in a related diagnosis task.

Learning

The hypothesis testing model makes a clear distinction between task knowledge, which is represented in propositional form in *Knowledge Base*, and strategic knowledge, which is embodied in the rules within *Decision Procedure*. It is assumed that strategic knowledge remains fixed throughout the task. Any improvement in participant performance across blocks is attributed entirely to the accumulation and/or modification of task knowledge. The separation therefore makes clear what the participant must learn during the task.

Knowledge Base contains two types of information, simple associations between diseases and symptoms, represented via `association` terms:

```
association(Disease, Symptom, Value)
```

and more complex pattern information relating configurations of symptoms to diseases, represented via `pattern` terms:

```
pattern(Disease, SymValList)
```

The `association` terms are used to determine which diseases are likely given the exhibited symptoms, and which symptoms to expect when considering the possibility that the patient has a given disease. *Knowledge Base* is initialised (for each simulated participant) with a random set of these terms, and *Learning Mechanism* contains rules that specify how they (and `pattern` terms) are modified and/or acquired throughout the task. These rules are triggered by the appearance in *Working Memory* of diagnostic feedback (generated by *Task Environment*). When such diagnostic feedback appears, the contents of *Knowledge Base* are modified as follows:

- An item of the form `association(Disease, Symptom, Value)` is added to *Knowledge Base* for each item of the form `told(Symptom, Value)` in *Working Memory*.
- Any previous associations between the symptoms and the actual disease are deleted (regardless of whether those symptoms are known to be present or absent).
- A `pattern(Disease, SymValList)` term is added to *Knowledge Base*, with `SymValList` bound to a list of symptom/value pairs corresponding to the symptoms known to be present or absent.
- Any pairs of `pattern` terms involving the same disease and differing only in the presence/absence of one symptom are merged, on the assumption that the disease is independent of the symptom.

These rules are naïve (and logically unsound) in that they make no attempt to merge `association` information from the current case with prior relevant knowledge. For example, if a disease is observed to occur both with and without a symptom, only the most recent observation is recorded. (The mechanism thus implements a recency bias.) Nevertheless, and as discussed below, the mechanism yields a good account of human performance.

Variants within the hypothesis testing class

Unlike the Bayesian and associationist models, the hypothesis testing models do not have an obvious way of trading off diagnostic accuracy against the number of symptoms queried. However, other variations parallel to those in the other models are possible. Thus, in the hypothesis testing models learning rate is the probability that information from the current trial is learnt according to the rules above. Decay may be imposed upon the contents of *Working Memory*, such that elements may spontaneously disappear, with the probability of disappearance on any cycle determined by a decay rate constant. The initial assumptions about symptom/disease associations are embodied in the initial associations contained in *Knowledge Base*. These may be varied by changing the ratio of `present` to `absent` associations. Finally, a bias towards information valence may be incorporated by selecting an appropriate strategy. Verification is biased towards positive associations, elimination is biased towards negative associations, and discrimination is unbiased. (For completeness a fourth strategy, discrim/verify, is also explored below.).

Overview of the models

The models represent three distinct theoretical approaches to the diagnosis task. The Bayesian models rely on quantitative information acquired through monitoring the frequencies of relevant events (e.g., the co-occurrence of a symptom and a disease). The associationist models similarly maintain quantitative information, but in these models that information is represented in the form of associative weights between symptoms and diseases. The hypothesis testing models, by contrast, maintain an ordered set of hypotheses concerning possible diseases, and use general purpose strategies to select from those hypotheses. Only categorical information (of the form symptom *x* is associated with disease *y*) is used in the process.

None of the models is particularly novel. The Bayesian class of models is a direct implementation of well established normative principles (Shannon & Weaver, 1949; Wiener, 1948; Lindley, 1956). The diagnostic processes and learning mechanism of the associationist models are basically the same as those proposed by Shanks (1991), with extensions to cope with the requirements of querying symptoms. The hypothesis testing class of models is based on that of Fox (1980). It includes insights from the work of Bruner *et al.* (1956), and shares some similarities with the RULEX model (rule-plus-exception) of Nosofsky, Palmeri, & McKinley (1994) in that it effectively develops categorical rules for mapping symptom patterns to diseases. These rules may either be general (involving relatively few symptoms) or relate to exceptional cases (involving specific values for most or all symptoms). However, the learning mechanism in RULEX is somewhat different, and RULEX is not designed to handle the kinds of indeterminate symptom/disease associations considered here.

All classes of model include parameters to control learning rate, decay of acquired information, assumptions about initial symptom rarity, and symptom selection bias. The Bayesian and associationist models also include a diagnostic threshold parameter. All models also include a bias towards recent events. In the Bayesian models this results from *Knowledge Base* decay. In the associationist models a similar effect is obtained by the use of delta rule learning. The effect is compounded by weight blurring. Finally, in the hypothesis testing models recency emerges through the interaction of the access characteristics of *Knowledge Base* and *Working Memory*.

In summary, although the models are based on different theoretical approaches to the diagnosis task, they are parameterised such that they share important characteristics. This allows evaluation of the relative importance of the underlying approaches (Bayesian, associationist or hypothesis testing) and the performance limitations and processing biases embodied in the shared characteristics (memory decay, recency biases, etc.) .

Experiment: symptom selection in medical diagnosis

Rationale

Previous empirical work with the medical diagnosis task has found that participants are able to achieve high levels of diagnostic accuracy. Fox (1980), who used medical students trained in diagnostic procedures as participants, reported accuracy of 81% on a five disease/five symptom version of the task. Models from each of the classes described above are all, with judicious selection of parameters, able to perform the diagnosis task at this level. As such, diagnostic accuracy is a poor discriminator of models. Further dependent measures derived from the task offer the opportunity to evaluate model performance at a finer grain. An experiment was therefore designed to provide appropriate dependent measures. Of primary interest were the learning profiles of human participants (diagnostic accuracy and numbers of symptoms queried, across a series of blocks), and their symptom querying behaviour at the end of the learning schedule.

Two major differences between the theoretical approaches concern their treatment of cues that are unreliable indicators of a category and the way that they implement positive biases. The experiment was designed to provide data relevant to these differences. Concerning the first, symptoms were associated with diseases in a probabilistic (rather than deterministic) fashion. For example, in one condition headache was associated with the (fictional) disease bonanoma on 75% of trials. In this condition headache may be suggestive of bonanoma, but absence of headache does not rule out bonanoma. Concerning the second, two conditions were investigated, with the average number of symptoms present and absent varying between the conditions. In the first condition the relationship between diseases and symptoms was "dense", in that relatively many symptoms were associated with each disease. In the second condition the relationship was "sparse" — on average fewer symptoms were associated with each disease.

Table 1 lists the conditional probabilities for each symptom given each disease for both conditions. For example, from the table the probability of a hypothetical patient having headache given that he/she has mesiopathy is 0.75 in the dense condition and 0.25 in the sparse condition. The dense and

Table 1: Conditional probabilities of symptoms given diseases in each condition

Condition 1: Dense matrix

	Disease			
Symptom	Mesiopathy	Ritengitis	Katalgia	Bonanoma
Diarrhoea	1.00	0.50	1.00	0.00
Fever	0.00	1.00	1.00	1.00
Headache	0.75	0.00	1.00	0.75
Paralysis	0.75	0.00	0.75	1.00
Vomiting	1.00	0.50	0.00	1.00

Condition 2: Sparse matrix

	Disease			
Symptom	Mesiopathy	Ritengitis	Katalgia	Bonanoma
Diarrhoea	0.00	0.50	0.00	1.00
Fever	1.00	0.00	0.00	0.00
Headache	0.25	1.00	0.00	0.25
Paralysis	0.25	1.00	0.25	0.00
Vomiting	0.00	0.50	1.00	0.00

sparse conditions are symmetrical in that the conditional probability of a symptom given a disease in one condition and the corresponding conditional probability in the other condition always sum to 1. The two conditions are therefore mathematically equivalent: the probability of a patient having disease *D* with symptoms *V*, *W*, and *X*, but not *Y* and *Z* in the dense condition is the same as the probability of a patient having disease *D* without symptoms *V*, *W*, and *X*, but with *Y* and *Z* in the sparse condition. Any bias towards treating positive and negative information differently should be seen in different behavioural patterns between these two mathematically equivalent conditions.

Method

Participants

40 second year psychology students from Birkbeck, University of London, took part, with 20 in each of the dense and sparse matrix conditions. Birkbeck teaches mature students, and ages ranged from 20 to 50, with a mean of 32. The experiment was conducted between 6pm and 8pm during an evening laboratory class.

Design

Each participant was randomly assigned to either the dense or sparse matrix condition. Participants performed four blocks of 20 trials. Each block comprised 5 trials with each of the four diseases listed in Table 1, in pseudo-random sequence, and with the symptom pattern of each trial generated randomly in accordance with Table 1. Mean diagnostic accuracy and number of symptoms queried were recorded for each block. In addition, the order of symptom queries was recorded for each trial of the final block.

Software

The task was computer-based, mouse driven and administered by a client-server system on the departmental intranet using a network of PCs. The client portion, written in JavaScript for Netscape Navigator 4, randomised trials within blocks, presented stimuli and collected responses. The server, which has since been developed into a general web-based experiment presentation system (Yule & Cooper, 2001), assigned participants to matrix conditions and recorded and collated the data.

The client system was launched by clicking on a button at the foot of a web page of instructions. This opened a new window. On each trial the program displayed a series of rectangles labelled with symptom names, running across the top half of the window, and a series of rectangles labelled with disease names running across the bottom. In order to avoid artefacts in question sequencing behaviour, the left-to-right order of symptom names and disease names was randomised from trial to trial.

The symptom rectangles could be clicked by the participant. They would then change to reveal whether the selected symptom was present or absent in the current hypothetical patient. At the beginning of each trial, one of the symptom rectangles was already in this changed state, giving the participant information about the patient's presenting symptom. Participants made their diagnosis by clicking on a disease rectangle in the lower half of the screen. When a disease rectangle was clicked, a new rectangle appeared in the centre of the screen stating whether the diagnosis was correct or not. If the diagnosis was incorrect, this rectangle also gave the name of the correct disease, allowing the participant to learn. The next trial was initiated by clicking on this rectangle.

At the end of each block, the program presented a score (the number of correct diagnoses out of 20), saved the accumulated data to the server, and gave the participant the chance to pause before proceeding to the next block.

Instructions

The launch page of the experimental client system described the screen layout and the block structure of the experiment. Participants were also verbally instructed to attempt to diagnose efficiently, that is, to minimise the number of symptom queries they made, provided that this did not compromise their diagnostic accuracy.

Table 2: Mean (standard deviation) diagnostic accuracy (Human data; N = 20 in each condition)

	Block				
Condition	1	2	3	4	Mean (s.d.)
Dense	34.3 (14.0)	46.5 (20.8)	48.8 (23.2)	49.0 (21.9)	44.6 (20.9)
Sparse	53.0 (17.1)	62.3 (20.6)	67.0 (23.4)	75.3 (19.8)	64.4 (21.5)
Mean (s.d.).	43.6 (18.1)	54.4 (22.0)	57.9 (24.8)	62.1 (24.5)	54.5 (23.2)

Table 3: Mean (standard deviation) number of symptoms queried (Human data; N = 20 in each condition)

	Block				
Condition	1	2	3	4	Mean (s.d.)
Dense	3.26 (1.09)	3.37 (1.07)	3.50 (0.85)	3.39 (1.10)	3.38 (1.02)
Sparse	2.83 (1.23)	2.54 (1.42)	2.37 (1.52)	2.36 (1.50)	2.52 (1.41)
Mean (s.d.).	3.04 (1.17)	2.95 (1.31)	2.93 (1.34)	2.88 (1.40)	2.95 (1.30)

Results

Table 2 shows mean percentage diagnostic accuracy of participants over all blocks and both conditions. (The data are shown in graphical form in the top left panels of Figures 2, 3, and 4.) There are highly significant effects of both block ($F(3,114) = 18.86$, $p < 0.001$) and matrix ($F(1,38) = 12.43$, $p < 0.01$). Participants in the sparse condition performed significantly better than those in the dense condition, and learning is shown by an increase in diagnostic accuracy over the four blocks. Analysis of simple effects shows that the effect of block is highly significant in each condition (dense: $F(3,57) = 6.261$, $p < 0.001$; sparse $F(3,57) = 15.894$, $p < 0.001$). Although the effect of block on diagnostic accuracy appears to be stronger in the sparse condition, there is no significant interaction ($F(3,114) = 1.54$).

Table 3 shows the number of symptoms queried by participants in each block and each matrix condition. (The data are shown in graphical form in the top right panels of Figures 2, 3, and 4.) There is a modestly significant effect of matrix ($F(1, 38) = 5.20$, $p < 0.05$), such that more symptoms are queried in the dense condition than in the sparse condition, but no overall effect of block ($F(3, 114) = 1.12$). However, there is a significant interaction ($F(3, 114) = 5.51$, $p < 0.01$). Analysis of simple effects reveals that the reduction in symptoms queried across blocks in the sparse condition is significant ($F(3, 57) = 3.60$, $p < 0.02$), but the apparent increase in the dense condition is not ($F(3, 57) = 2.39$). There was no significant correlation between diagnostic accuracy and number of symptoms queried in either condition (dense condition: Pearson's r = 0.17, $t(18)$ = 0.72; sparse condition: Pearson's r = –0.11, $t(18)$ = –0.46). There is therefore no evidence of a trade-off between diagnostic accuracy and symptom querying.

Strategies or biases that may have developed in symptom querying can be investigated by tabulating the first query made by participants against the presenting symptom. Table 4 shows this "one-ply" tabulation for performance on the final block. The table shows, for example, that in the dense condition diarrhoea was given as the presenting symptom on 73 trials.[4] Participants immediately made a diagnosis of bonanoma on just 1.4% of these trials, on 23.2% of trials they first queried fever, on a further 23.2% of trials they first queried headache, and so on.

The one-ply data are not amenable to standard statistical techniques as the individual cell entries are not independent. Nevertheless, a number of qualitative comments may be made. The strongest tendencies in the dense condition are 1) a general tendency to query further symptoms rather than immediately offering a diagnosis, and 2) a tendency to query diarrhoea when the presenting symptom is headache. The tendency away from immediate diagnosis is absent in the sparse condition. This is consistent with the symptom querying data in Table 3. The sparse condition does, however, lead to other tendencies. For example, when the presenting symptom is diarrhoea, the dominant response is to diagnose bonanoma, but when paralysis is presented the preference is to either query vomiting or diagnose ritengitis, and there is an aversion to querying the presence of fever.

Table 4 also shows the "query rate" for each query in each matrix condition. This is a weighted average of the query column totals, and is calculated as the number of times a symptom was queried divided by the number of opportunities for querying that symptom. This statistic was calculated to investigate whether participants showed any non-specific tendency towards a particular symptom. Such a tendency would appear as a peak in the query rate distribution that was independent of condition. There is no evidence of such a peak.

Discussion

The results confirm that participants are able to learn the task. In both conditions diagnostic accuracy is greater than chance (25%), even in the first block. In particular, mean diagnostic accuracy of 53% in the first block of the sparse condition indicates that participants are can learn some associations rapidly and apply that learning before the first block is complete.

[4] The data in each table are derived from 400 trials — 20 trials for each of 20 participants in each condition. The frequencies of presenting symptoms are unequal because, from Table 1, not all symptoms are equally likely.

Table 4: One-ply table for final block performance (Human data). Entries represent the percentage of responses in each category.

Condition 1: Dense matrix

Presenting		Diagnosis				Query				
Symptom	N	Bonanoma	Katalgia	Mesiopathy	Ritengitis	Diarrhoea	Fever	Headache	Paralysis	Vomiting
Diarrhoea	73	1.4	0	0	0	–	23.2	23.2	20.6	31.5
Fever	138	0	4.3	0	1.4	15.9	–	23.9	26.8	27.5
Headache	55	1.8	1.8	0	0	45.5	21.8	–	14.6	14.6
Paralysis	63	0	7.9	4.8	0	25.4	11.1	14.3	–	36.5
Vomiting	71	0	1.4	1.4	0	29.6	18.3	35.2	14.1	–
Query rate						25.7	18.7	24.3	20.1	28.0

Condition 1: Sparse matrix

Presenting		Diagnosis				Query				
Symptom	N	Bonanoma	Katalgia	Mesiopathy	Ritengitis	Diarrhoea	Fever	Headache	Paralysis	Vomiting
Diarrhoea	101	32.7	0	0	2.0	–	14.9	19.8	15.8	14.9
Fever	89	0	1.1	34.8	3.4	5.6	–	28.1	14.6	12.4
Headache	56	1.8	3.6	1.8	7.1	30.4	25.0	–	8.9	21.4
Paralysis	46	4.3	0	2.2	23.9	19.6	4.3	15.2	–	30.4
Vomiting	108	1.9	20.4	0.9	2.8	14.8	19.4	20.4	19.4	–
Query rate						15.7	16.7	21.5	15.6	17.8

Learning is also seen in the development of symptom querying biases, and, at least in the sparse condition, by a decrease in the number of symptoms queried before arriving at a diagnosis. In the dense condition there is some evidence that diagnostic accuracy reaches a ceiling. Diagnostic accuracy levels off at just below 50% in this condition after block two, while in the sparse condition it continues to rise over all four blocks, ending at over 75%. However, comparing diagnostic accuracy and symptoms queried in blocks three and four reveals an improvement in "efficiency" in both conditions: in the sparse condition accuracy increases without compromising the number of symptoms queried; in the dense condition the number of symptoms queried decreases (albeit marginally) without compromising accuracy.

Differences in performance over the dense and sparse conditions are of particular interest. Although the conditional probabilities of symptoms given diseases in the two conditions are complementary (and so, as described above, the conditions are mathematically equivalent), the first symptom given to a participant — the presenting symptom — is always present/positive rather than absent/negative. A positive symptom narrows the search space more in the sparse condition, in which diseases are associated with few positive symptoms, than in the dense condition, in which diseases are associated with many positive symptoms.

One consequence of this asymmetry is that a greater number of symptom queries may be expected in the dense condition than in the sparse condition. This was indeed found to be the case. The asymmetry does not, however, imply any difference between conditions in diagnostic accuracy, or that there should be an interaction between matrix condition and the number of symptoms queried across blocks. These are therefore a curious aspects of the data that pose a challenge for the computational models.

Some of the querying tendencies identified in the previous section can be understood in terms of the associations between symptoms and diseases given in Table 1. The tendency not to offer an immediate diagnosis in the dense condition, for example, may arise from the fact that, given any presenting symptom, three diseases always remain possible. In contrast, occurrence of fever in the sparse condition guarantees that the hypothetical patient is suffering from mesiopathy. Many participants appear to realise this. The tendency not to diagnose and not to query paralysis when headache is presented in the sparse condition can similarly be understood. The roots of other tendencies are less clear. When, in the dense condition, headache is the presenting symptom, three diagnoses are possible. Querying the presence of diarrhoea will reduce the number of possible diagnoses to two (if diarrhoea is present) or one (if it is absent). Querying the presence of vomiting would appear to be equally useful, but in this situation diarrhoea was queried over three times more frequently than vomiting. Again, the data provide a challenge for the computational models.

Modelling results

Methodology

It would be neither practical nor instructive to examine and report simulation results covering all regions of all parameter spaces for all models. Such an approach might yield one or more models that fit the empirical data well, but it would be little more than a data fitting exercise. Instead, the methodological strategy adopted here is to present for each approach one "canonical model", to describe the strengths and weakness of that model's behaviour, and then to explore the effects of varying each of the model's parameters independently in an attempt to characterise how each parameter impacts upon behaviour. Clear interactions between parameters are also reported. Finally, for each approach we present a "preferred model". Preferred models result from a judicious choice of parameters based on the documented main effects and interactions. They yield better fits to the data than canonical models, but do not necessarily represent optimal fits. This methodology is designed to clarify the role of each parameter in influencing the models' behaviour. Strengths and weaknesses of the methodology are discussed in the General Discussion.

In all cases graphs are presented showing diagnostic accuracy and number of symptoms queried across the four blocks of the task, for the human data and both the canonical and preferred models. Querying behaviour generated by the models is also compared with human querying behaviour for each condition via two statistics: the root mean square (RMS) differences between the 40 non-empty entries of Table 4 and the equivalent table generated by the model, and the Pearson's product moment correlation coefficient calculated over the same 40 pairs of table entries. Neither of these statistics is entirely appropriate because one-ply table entries are not independent. Thus, and in order to establish the reliability of these statistics, each model was run for 40 virtual subjects on ten separate occasions (i.e., for ten replications of the complete experiment), and the resulting sample used to generate means (and standard deviations) for the statistics.

Bayesian models

The canonical Bayesian model

Parameter settings for the canonical Bayesian model are motivated by normative considerations: decay is 0.00, informativeness calculation is unbiased, and the number of present and absent occurrences of each symptom/disease combination is assumed to be 1.00. In addition, the diagnostic threshold is set to 0.90 to ensure high diagnostic accuracy.

The middle panels of Figure 2 show the diagnostic accuracy and number of symptoms queried by the canonical Bayesian model (cf. the top panels of

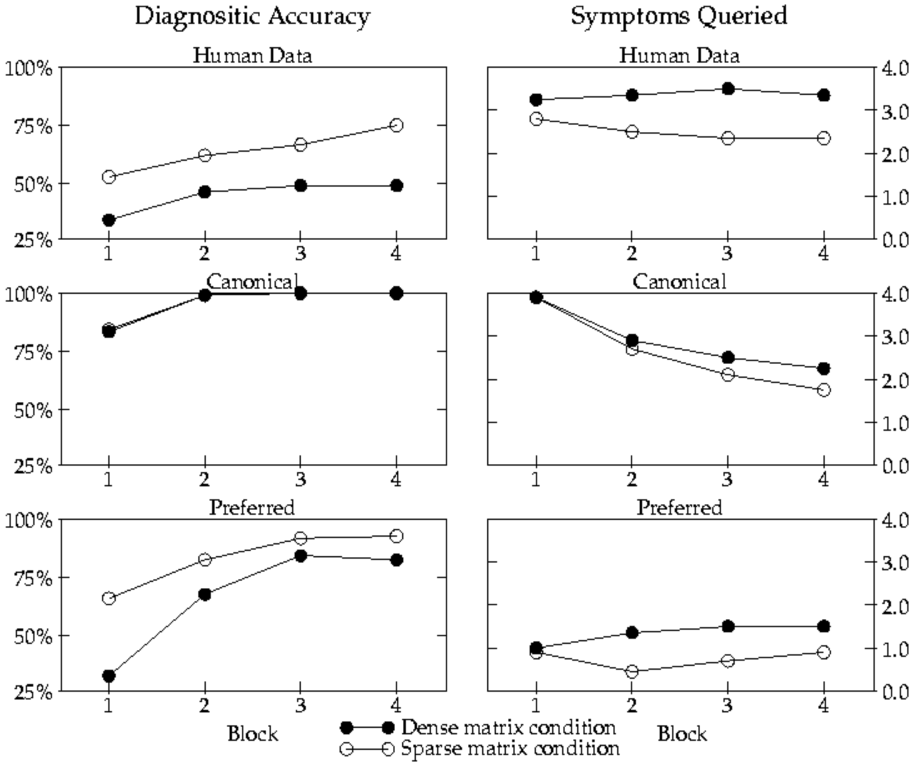

Figure 2: Mean diagnostic accuracy and number of symptoms queried for both matrix conditions for human participants and Bayesian models

the same figure, which show the human data). Perfect diagnostic performance is obtained in both conditions after just one block. However, there is continued evidence of learning as the number of symptoms queried continues to fall across all 4 blocks. This reflects increasing diagnostic efficiency across blocks — as the model learns better approximations to the conditional probabilities of symptoms given diseases it is able to generate more appropriate symptom queries (or extract more information from queried symptoms), and hence diagnose more efficiently. More symptoms are queried in the dense condition than in the sparse, presumably because the presenting symptom is more informative in the sparse condition than in the dense condition.

Comparison of the graphs in Figure 2 suggests that the canonical Bayesian model is a poor model of human performance. This is reflected in the one-ply statistics (see the top row of Table 5) which reveal high RMS and low to modest correlations between the model's querying behaviour and that of the human participants.

Table 5: Means (and standard deviations) for one-ply statistics generated from the Bayesian models

	Dense		Sparse	
Model	RMS	*r*	RMS	*r*
Canonical	24.432 (0.558)	0.356 (0.014)	26.502 (0.219)	0.142 (0.010)
Preferred	10.727 (1.445)	0.782 (0.037)	19.932 (0.948)	0.455 (0.019)

Simple effects of parameters

Reducing the diagnostic threshold leads to a reduction in the number of symptoms queried and a reduction in accuracy. The effect is similar in both matrix conditions. These results are consistent with the diagnostic threshold acting to trade off diagnostic accuracy against the number of symptoms queried. However, even with a low threshold (e.g., 0.60) accuracy remains high (around 95% in the final block of both conditions). Reduction of the diagnostic threshold also leads to a slightly better fit of querying behaviour in the sparse condition, with *r* increasing to over 0.35. This appears to be due to the model's increased tendency to diagnose in some cases on the basis of just the presenting symptom — an effect also seen in participant behaviour.

The main effect of variation of decay (cf. Equation 5) is similar to that of reduction in the diagnostic threshold: non-zero decay reduces both the diagnostic accuracy and the number of symptoms queried. The effect on accuracy is quite mild, however. Even with decay of 0.50, accuracy begins at 75% in the first block and rises to 90% by the fourth block. The number of symptoms queried is more strongly affected, falling to approximately 2.0 on the first block and approaching 1.0 by the final block. As noted previously, the addition of decay imposes a recency bias on the model, so it would seem that reasonable diagnostic accuracy can be obtained by basing responses primarily on recent cases. More significantly, however, high rates of decay lead to querying behaviour that resembles that of human participants, especially in the dense condition, which with decay of 0.75 yields a correlation of more than 0.70 and an RMS error of less than 11 — both substantially better than the corresponding figures for the canonical model.

Variation of bias in the calculation of informativeness has little effect on diagnostic accuracy. It does affect the number of symptoms queried, however. With a positive bias, the difference between the number of symptoms queried in each condition is reduced slightly, whereas with a negative bias it is slightly increased. More interestingly, positive bias greatly improves the correlation between the model's query behaviour and the human data (from 0.36 to 0.69 in the dense condition and from 0.14 to 0.25 in the sparse condition). A much smaller improvement in correlation is observed with the negative bias. The effect is restricted to the correlation statistic; the RMS error is only marginally reduced in each condition.

The main effects of varying the initial distribution of symptom/disease associations are on diagnostic accuracy and the number of symptoms queried. The initial distribution has little effect on the correlation or RMS statistics. In detail, the effects are as follows. If `NumPresent` and `NumAbsent` are equal and much greater than 1.00, the model tends to query all four symptoms in all trials (regardless of condition or block). Diagnostic accuracy in this case is perfect after the initial block. If `NumPresent` and `NumAbsent` are equal, greater than zero, but less than one, then the model queries fewer than four symptoms in the first block. The rate of querying remains constant after the second block, with more queries in the dense condition than in the sparse condition, and high but imperfect accuracy in both conditions. If `NumPresent` is greater than `NumAbsent` (e.g., 10 compared with 0.1), then diagnostic accuracy in the dense condition is greater than in the sparse condition, and more queries are made in the sparse condition than in the dense condition. These behaviours are contrary to that of participants. However, if `NumPresent` is less than `NumAbsent` (e.g., 0.1 compared with 10), the effects are reversed, and the effects of condition on the model are similar in quality (though not in quantity) to effects on the human participants.

Interactions between parameters

There are few cases where parameters of the Bayesian model interact in their effects on dependent measures. For example, simultaneously lowering the diagnostic threshold and increasing the decay rate results in near additive effects on both diagnostic accuracy (which falls marginally due to both manipulations) and the number of symptoms queried (which also falls for similar reasons), but diagnostic accuracy remains insensitive to condition, and the number of symptoms queried remains higher in the dense condition than in the sparse condition, as in the canonical model.

One case where an interaction is apparent — between the `NumPresent` and `NumAbsent` parameters — has already been described. These parameters also interact with decay. Increasing decay reduces the effect of scaling `NumPresent` and `NumAbsent`. This is easily explained, as a high decay rate results in recency, rather than initial values, dominating behaviour as the task progresses.

The preferred Bayesian model

The relative independence of the parameters of the Bayesian model simplifies the task of crafting a set of parameters that yields a reasonable fit to the human data. First, `NumAbsent` must be greater than `NumPresent` to ensure that accuracy is greater in the sparse condition and that more symptoms are queried in the dense condition. Second, a positive bias must be employed in the calculation of informativeness, as this yields better fits to the query bias data without compromising other dependent measures. Provided that the difference between `NumPresent` and `NumAbsent` is sufficiently large, these two manipulations yield a model that produces a reasonable fit to the data, with the exception that accuracy is much higher than that obtained by our participants. The fit to the query bias data can be improved by increasing decay and scaling up `NumPresent` and `NumAbsent` to counteract its effect on the other dependent measures. Finally, decreasing the threshold will reduce diagnostic accuracy, allowing accuracy levels to be brought more in line with the human data. This also reduces the number of symptoms queried, however, so `NumPresent` and `NumAbsent` must be scaled up further to counteract this effect.

The data shown as the preferred Bayesian model in Figure 2 and Table 5 were obtained with a diagnostic threshold of 0.7, decay of 0.3, the positive bias in the calculation of informativeness, and initial values of `NumPresent` and `NumAbsent` of 0.02 and 200 respectively. The preferred model fits the data somewhat better than the canonical model. Of particular interest is the correlation between the human and model query behaviour, which is 0.78 in the dense condition and 0.46 in the sparse condition. However, there are several substantive differences between the human and model data. Perhaps most seriously, the human data suggest that participants in the dense condition reach a ceiling of 50% in their diagnostic accuracy by the second block, while accuracy in the sparse condition continues to increase across all four blocks. None of the parameter variations lead to a Bayesian model that replicates this pattern. In all cases accuracy of the Bayesian model appears to asymptote to the same value in both conditions, with that value being approached more rapidly in the sparse condition than the dense condition.

Associationist models

The canonical associationist model

Within the canonical associationist model learning rate is 1.00, there is no weight blurring, query strategy is unbiased, the initial weight distribution is normal with a mean of 0.00 (and a standard deviation of 0.30) and the diagnostic threshold is 0.80 (equivalent to that of the canonical Bayesian model given that activations range from –1.0 to +1.0).

The middle panels of Figure 3 show the diagnostic accuracy and number of symptoms queried by the canonical associationist model (cf. the top panels of the same figure, which show the human data). The fit of the canonical associationist model is substantially better than the fit of the canonical Bayesian model. As in the human data, accuracy in the sparse condition is greater than in the dense condition, while more symptoms are queried in the dense condition than in the sparse. The top row of Table 6 also reveals that the one-ply querying behaviour of the canonical associationist model is more realistic than that of the canonical Bayesian

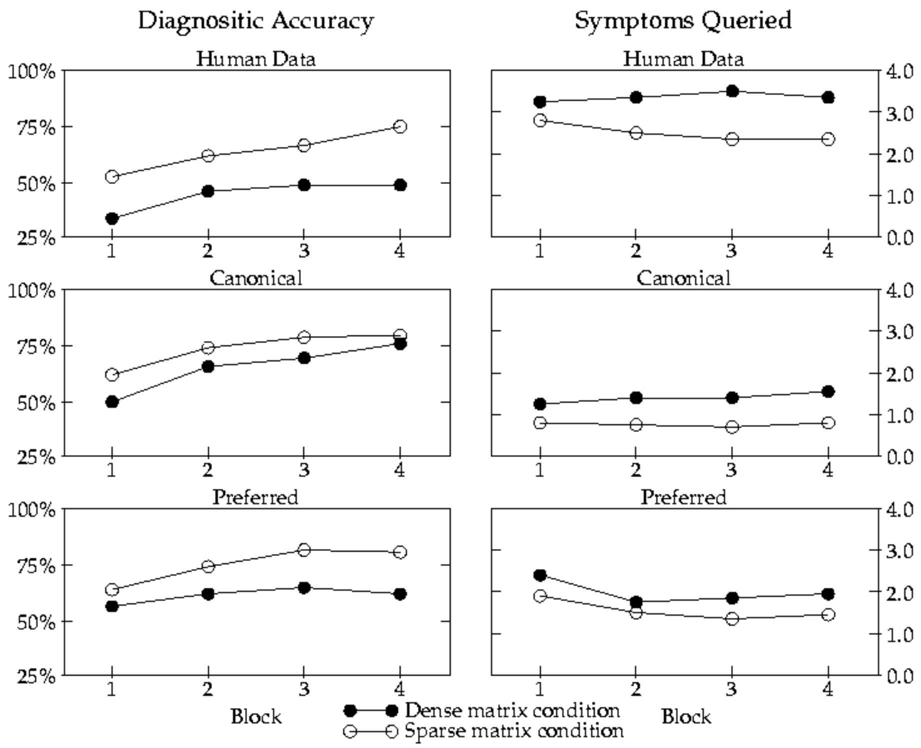

Figure 3: Mean diagnostic accuracy and number of symptoms queried for both matrix conditions for human participants and associationist models

model. The fit in the sparse condition even exceeds the fit obtained with the preferred Bayesian model.

Simple effects of parameters

As might be anticipated, reducing the diagnostic threshold in the canonical associationist model reduces the number of symptoms queried. It also increases the difference between conditions in diagnostic accuracy. Its effect on the fit of the one-ply data is, however, detrimental, with the correlation between human and model behaviour being negative in the dense condition. Increasing the threshold results in an increase in the number of symptoms queried and a reduction in the difference in diagnostic accuracy between conditions. It does not further improve the fit of the one-ply data.

The associationist version of *Knowledge Base* decay is weight blurring. Small amounts of weight blurring (with standard deviation of 0.01) yield an increase in the difference in diagnostic accuracy between conditions, no effect on the number of symptoms queried, and a slight improvement in the one-ply fit under the sparse condition. The one-ply fit in the dense condition

Table 6: Means (and standard deviations) for one-ply statistics generated from the associationist models

	Dense		Sparse	
Model	RMS	*r*	RMS	*r*
Canonical	13.484 (0.950)	0.483 (0.048)	14.499 (1.200)	0.580 (0.039)
Preferred	9.089 (1.759)	0.731 (0.093)	13.140 (1.440)	0.589 (0.042)

is adversely affected, however. These effects are magnified with larger degrees of weight blurring. In all cases weight blurring maintains the differences in conditions for diagnostic accuracy and the number of symptoms queried produced by the canonical model.

The effects of variation in bias are generally small. A positive bias leads to an improved one-ply fit in the dense condition but the fit in the sparse condition is impaired. It has little effect on the number of symptoms queried in either condition, but results in a slightly larger difference in diagnostic accuracy between conditions. A negative bias reduces the difference in accuracy between conditions but has little effect on the one-ply fits or the number of symptoms queried in either condition.

Varying the centre of the initial weight distribution (and thereby varying the initial assumptions about symptom/disease associations) has little effect on the number of symptoms queried, though it compromises the one-ply fit of the model to the human data. The effect on diagnostic accuracy is for positive weight centres to increase the difference between conditions and negative weight centres to decrease the difference between conditions.

Decreasing the learning rate from its default of 1.00 improves the model's diagnostic accuracy, reducing the difference in accuracy between conditions (with accuracy in both conditions approaching ceiling). Lower learning rates also lead to more querying of symptoms but poorer fits between human and model one-ply query behaviour.

Interactions between parameters

The analysis of parameter interactions within the associationist models is hindered by the generally small effects of parameters within the models. There are no clear interactions, for example, between weight centre (positive, zero, negative) and query bias (positive, unbiased, negative) on diagnostic accuracy or number of symptoms queried. However, there is an interaction on the fit between human and model query behaviour, with a positive bias and a negative weight centre yielding a better fit in the sparse condition. Independent manipulation of the parameters impairs this measure.

Learning rate interacts in complex ways with most other parameters. For example, manipulation of learning rate and weight centre independently impairs the one-ply fit in the dense condition, but this impairment is reduced when both parameters are manipulated together. This interaction is not present in the sparse condition. In a similar vein, a low learning rate eliminates the difference in diagnostic accuracy between matrix conditions in the unbiased and negatively-biased models, but not in the positively-biased model, and a positive bias and slow learning rate increase the one-ply fit in the dense condition more than additively, whereas a negative bias and slow learning rate increase the fit less than additively. In contrast, the adverse effect of a decreased learning rate on the one-ply fits is compounded when combined with a very high threshold (e.g., 0.90). Further interactions also occur (e.g., between learning rate and weight blur, and between weight blur and diagnostic threshold). The effects of parameters on the behaviour of the associationist models are therefore far from transparent.

The preferred associationist model

Many of the associationist models yield good fits to the observed data on diagnostic accuracy and number of symptoms queried. They differ primarily in their fit with the one-ply data. However the picture of parameter interactions is confusing as most interactions have differential effects in the two matrix conditions. While this makes finding an optimal parameter fit particularly difficult, it is possible to improve upon the performance of the canonical model. Nevertheless, the analysis of this class of models remains less satisfactory than the corresponding analysis of the Bayesian class.

The preferred associationist model is obtained by extrapolating the assumptions of the preferred Bayesian model to the parameters of the associationist model class. Specifically, the preferred associationist model has a positive query bias procedure, an initial distribution centred on –2.00 (reflecting few positive initial symptom/disease associations), a learning rate of 1.00, mild weight blurring of 0.01 (implementing a recency bias), and a diagnostic threshold of 0.99. The high diagnostic threshold is the only one of these parameters obtained through direct parameter fitting.

The behaviour of the preferred associationist model is illustrated in the lower panels of Figure 3. The basic effects of condition are present in both graphs. The fits between the graphs are better than for the optimal Bayesian models, but are still less than perfect, and arguably no better than the fits of the canonical associationist model. However, the one-ply data are more convincing, particularly in the dense condition (see Table 6).

Hypothesis testing models

The canonical hypothesis testing model

The parameter settings for the canonical hypothesis testing model are, as for the other classes of model, aimed at producing a maximally unbiased model. To this end, decay is set at 0.00, the strategy is set to discrimination (which uses both positive and negative cue information equally), the initial

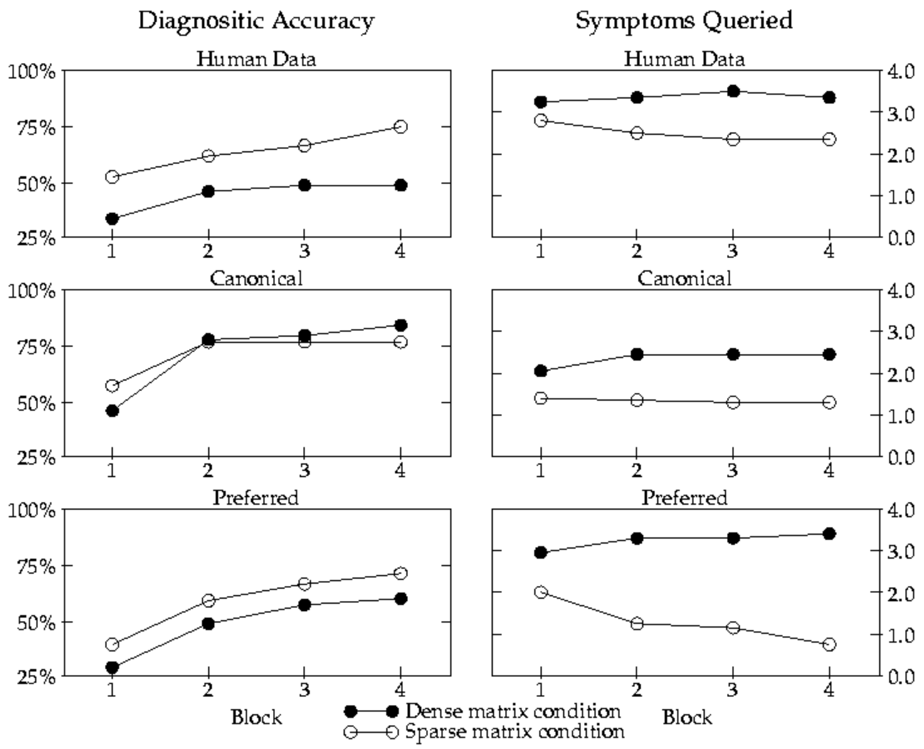

Figure 4: Mean diagnostic accuracy and number of symptoms queried for both matrix conditions for human participants and hypothesis testing models

distribution of symptom/disease associations consists of equal numbers of present and absent symptoms, and the learning rate is 1.00.

The middle panels of Figure 4 show the diagnostic accuracy and number of symptoms queried by the canonical hypothesis testing model (cf. the top panels of the same figure, which show the human data). The canonical model fails to show an effect of condition on diagnostic accuracy. However it replicates the qualitative effect of condition on number of symptoms queried, with more queries in the dense condition than in the sparse condition. The top row of Table 7 reveals that the one-ply query behaviour produced by the canonical hypothesis testing model is plausible. In both conditions the fits are better than for any other model discussed so far. The fit in the dense condition is remarkably good — with a correlation of over 0.84 and an RMS error of just over 7.

Simple effects of parameters

The behaviour of the hypothesis testing model is highly dependent upon strategy. When the verify strategy is adopted, the number of symptoms

Table 7: Means (and standard deviations) for one-ply statistics generated from the hypothesis testing models

	Dense		Sparse	
Model	RMS	*r*	RMS	*r*
Canonical	7.128 (0.915)	0.847 (0.039)	11.690 (1.575)	0.659 (0.049
Preferred	7.141 (0.649)	0.837 (0.031)	13.378 (0.806)	0.638 (0.032

queried in the dense condition increases to just below 4.0 (i.e., ceiling), while in the sparse condition it remains near 1.5. Diagnostic accuracy also increases, particularly in the dense condition, where it reaches 97%. There is a slight reduction in the fit with the one-ply data in both conditions, though the fits remain better than in most other models. The elimination strategy results in more symptom queries in the sparse condition than in the dense condition. Accuracy in the sparse condition is correspondingly higher than in the dense condition (though both are high: 95% and 90% respectively). The one-ply data fit drops in both conditions, with the sparse condition particularly affected. The combined discrim/verify strategy leads to diagnostic accuracy and number of query curves that are very similar to those of the verify strategy, with greater accuracy and more queries in the dense condition than the sparse condition. The one-ply fit is better though, particular in the dense condition, where the correlation approaches 0.90. This fit is better than that obtained with any other model variant.

Variation of the initial distribution of symptom/disease associations also has a marked affect on the model's behaviour. Biasing the distribution in either direction (i.e., towards most initial associations being either positive or negative) impairs diagnostic accuracy in the dense condition but not the sparse condition. The result is greater accuracy in the sparse condition than the dense condition (as observed in the participant data). In both cases the number of symptoms queried is also reduced, though this effect is larger when the initial bias is towards negative associations than positive ones. An initial negative bias also results in a substantially poorer one-ply fit, though an initial positive bias has a slightly detrimental effect on the fit in the dense condition but a positive effect on the fit in the sparse condition.

Mild decay of *Working Memory* elements reduces accuracy, particularly in the dense condition. This captures the basic effect of condition on diagnostic accuracy present in the human data, but also yields fewer symptom queries and poorer one-ply fits in both conditions. The learning rate is less critical to the model's behaviour. Halving the parameter's value from 1.00 to 0.50 causes a small reduction in diagnostic accuracy in both conditions, but has no discernible effect upon the number of symptoms queried. The effect on the one-ply fit is also small (though unfavourable). More substantial reduction impairs diagnostic accuracy in both conditions.

To summarise, the hypothesis testing models generally produce good one-ply fits. They also generally yield more symptom queries in the dense condition than in the sparse condition (with the exception of the elimination strategy). However, the canonical model fails to account for the greater accuracy of participants in the sparse condition. This effect can be captured by biasing the initial distribution of symptom–disease associations (either positively or negatively) or by adding mild decay to W*orking Memory*.

Interactions between parameters

There is substantial independence of parameters within the hypothesis testing models. In fact, the effects of learning rate, decay and initial distribution of symptom/disease associations all appear to be additive. There is only one interaction of particular interest. The one-ply fit of the discrimination strategy is more affected by working memory decay than for the other strategies. Working memory decay reduces the fit of the discrimination strategy by one-third in both conditions, but has little effect on the fit of the other strategies. Thus, while the discrimination strategy of the canonical model might appear to produce a good fit to the one-ply behaviour of participants, it does less well on accounting for the effect of condition on diagnostic accuracy, and while introducing working memory decay improves the latter, it has a serious detrimental effect on the former.

The preferred hypothesis testing model

The preferred hypothesis testing model is based on the observation that *Working Memory* decay does not adversely affect the one-ply fit in the verificationist model. The preferred model is just the canonical model with a verificationist strategy and *Working Memory* decay. This model retains the canonical model's good one-ply fit in both conditions (see the lower row of Table 7, which differs non-significantly from the upper row), but also yields the appropriate main effects of condition on both diagnostic accuracy and number of symptoms queried (see the bottom panels of Figure 4). The main differences between the model's behaviour and that of the human participants concerns the sparse condition, where human participants queried more symptoms and achieved greater accuracy. The fit between model and participant behaviour is not improved by variation of either the initial distribution of symptom/disease associations or learning rate.

General discussion

Participants showed a number of robust effects on a sequential symptom selection version of a medical diagnosis task: greater accuracy with less symptom querying when few rather than many symptoms were associated with each disease, and non-random querying of symptoms. The data present a challenge for the three theoretical approaches, but the modelling results demonstrate that judicious choice of parameter values can lead to a model based on any of the three approaches that produces a reasonable account of the data. We begin this discussion by reviewing the preferred models and the roles of the parameters and biases in those models.

A further issue raised by the above is that of methodology. Human behaviour has been compared with behaviour generated by three models based on the competing theoretical approaches. This involved exploring the effects of parameter modification within each model in an attempt to fit aspects of the data. This form of *comparative modelling* has not been widely used within the cognitive sciences. We therefore evaluate the methodology.

Parameters and biases in model behaviour

The thrust of the modelling work was not to produce a single best fitting model. Rather it was to investigate possible performance factors and processing biases in the various approaches and to discover what, if any, role they may have in generating human-like behaviour. From this perspective the modelling has been successful. The preferred Bayesian model incorporated a positive query bias, an assumption of mainly negative initial symptom/disease associations, a moderate threshold (0.7, on a scale of 0.5 to 1.0), mild decay of frequency representations (0.3 on each trial) and a sub-optimal learning rate. The preferred associationist model also incorporated a positive bias in symptom querying, decay of acquired symptom/disease knowledge (in the form of mild weight blurring), and a negatively skewed initial distribution of initial symptom/disease associations. Unlike the Bayesian model, however, it required a high threshold (0.99 on a scale of 0.0 to 1.0) and fast learning (a learning rate of 1.0). Finally, the preferred hypothesis testing model employed the verificationist strategy (another positively biased querying strategy) and working memory decay. Within this approach it was not necessary to introduce any bias in the initial assumptions about symptom/disease associations, or any concept of a diagnostic threshold.

All preferred models share two biases: a positive bias and a recency bias. These biases are well documented (e.g., Wason & Johnson-Laird, 1972; Hunt & Rouse, 1981; Hearst, 1991). The current work re-iterates their importance in the context of a complex realistic task. It also demonstrates that they are independent of theoretical framework. The two biases are not sufficient to generate human-like behaviour, however. Additional biases or performance factors are required in each framework to fully account for the data. Thus, negative skew of symptom/disease associations is needed within the Bayesian models to ensure that accuracy is greater in the sparse condition than in the dense condition. Negative skew is also useful in the associationist model, though here its effect is to improve the one-ply fit. Skew, whether positive or negative, is not helpful for the hypothesis testing model, which performs in its most human-like way when initial associations are equally likely to be positive or negative. A second parameter with differential effects in the models is diagnostic threshold. The hypothesis testing models have no threshold: diagnosis is determined by matching symptom patterns against stored templates. A diagnostic threshold is essential in the other approaches, but the Bayesian approach appears to require a modest threshold while the associationist model requires a high threshold.

That models based on all approaches can account for the experimental findings may be interpreted in two ways. The experiment may be a poor discriminator of theories, and further experimentation might yield data that would discriminate between the theories. We suggest that this is unlikely. Simultaneously accounting for all dependent measures presents a significant challenge. The models were only able to do this through the inclusion of processing biases and performance factors. Alternatively, it may be that at the level of human behaviour the three approaches when combined with processing biases and performance factors are indistinguishable. Similar indistinguishability arguments have been made in other cognitive domains, including serial versus parallel processing (Townsend, 1971), analogical versus propositional representation (Anderson, 1978), and rule versus model based reasoning (Stenning & Yule, 1997).

Accounting for the experimental effects

Model behaviour may also be analysed in terms of specific effects in the human data and the mechanisms required in the models to capture these effects. Diagnostic accuracy is greater in the sparse condition than in the dense condition. This appears to be a natural consequence of the associationist class of models. It requires the assumption of an initial negative skew within the Bayesian model, and the assumption of working memory decay within the hypothesis testing model. The Bayesian account of this effect is not entirely satisfactory, however. If the effect is due to initial assumptions about the space of symptom/disease associations, then the effect of condition should disappear with block (as learning will reduce the impact of initial assumptions). This occurs in the Bayesian model (and the canonical associationist model) where accuracy in the two conditions appears to converge in later blocks. It does not occur in the human data, in which accuracy in the dense condition appears to reach a ceiling well below that of the sparse condition (which appears to continue to increase into the fourth block). The hypothesis testing account, in terms of working memory decay, is more able to produce this divergence between conditions, because in the hypothesis testing models the dense condition involves more hypotheses (and more expected symptoms), and will therefore be affected more by working memory decay than the sparse condition.

The difference between the number of symptoms queried within each condition is simpler to account for. The task structure ensures that the presenting symptom carries more information in the sparse condition (where symptoms are relatively rare) than in the dense condition. All models, with the exception of the elimination version of the hypothesis testing model, displayed this effect. However, it is an effect that develops with learning. The number of symptoms queried in the first block presumably indicates something about the initial state of the system (e.g., its initial assumptions about symptom/disease associations). In the Bayesian and hypothesis testing models learning alters the querying behaviour so that it approaches a level determined by a combination of the structure of the symptom/disease associations and the efficiency of the diagnostic procedure. Learning has little effect upon the number of symptoms queried by the associationist models. This may be because the query selection algorithm is not closely tied to the learning algorithm.

The third set of dependent variables relate to the one-ply data. While the positive query bias strategies of all models go some way towards producing human-like querying behaviour, the hypothesis testing model is particularly good at replicating this data. This is largely due to the access characteristics of *Working Memory* and *Knowledge Base*, which were fixed in the simulations reported here but which have been shown in previous work to affect the model's query bias behaviour (Fox, 1980; Fox & Cooper, 1997). That work found that the best fit with human data on a closely related task using a different matrix of symptom/disease associations was obtained with access to *Knowledge Base* based on recency and access to *Working Memory* based on primacy. These access characteristics were adopted in the current models, and the relatively good fit obtained further supports both the general hypothesis testing family of models and the access characteristics of the buffers involved. It is particularly significant to note that the previous work was based on a model of final block performance (i.e., a model that did not learn) on a task using different symptom/disease associations. The current work demonstrates that the earlier findings hold over related tasks and when the models incorporate a learning mechanism.

Comparative modelling and "parameter fitting"

The approach adopted in this paper — the comparative evaluation of competing accounts of a task through the development of multiple cognitive models — is currently rare within the cognitive sciences. The approach, which we term comparative modelling, augments the methodology of previous work involving model comparisons (e.g., Gluck & Bower, 1988; Estes, *et al.*, 1989). Important features of comparative modelling include the use of a common "benchmark" task (for which empirical data involving multiple dependent measures exists), the testing of all models on that task under identical conditions, and the exploration of the behavioural consequences of systematic variations to all models. The methodology is intended to allow the evaluation of the strengths and weaknesses of competing approaches. In the current work it has been successful in showing how, for example, the effect of condition on diagnostic accuracy may be accounted for in different ways by different theories, and how similar biases may be incorporated into models based on different theories.

The comparative modelling approach requires significant computational infrastructure to support task environment and scoring modules (into which the various models may be plugged) and to support the execution of computational experiments analogous to standard laboratory experiments. In the present case this infrastructure was provided by the COGENT modelling environment (Cooper & Fox, 1988; Cooper, 2002). We know of no other modelling environment that provides this level of modelling support.

The use of parameterised models is not an essential aspect of comparative modelling. Parameters serve two purposes in the current work. First, each of the approaches is under-specified. Parameters allow alternative elaborations of the same basic mechanism to be incorporated in the models. Second, parameters allow the introduction of performance factors and processing biases such as memory decay and skewing of the initial symptom/disease associations. Such factors are essential in each of the theoretical accounts if they are to capture human diagnostic performance. The difficulty with parameterised models is that parameters allow extra degrees of freedom. It has been argued that if a model includes many parameters then selecting appropriate parameter values will allow the model to fit virtually any data set (Roberts & Pashler, 2000). This is only true when parameters have independent effects on behaviour. In any case, the methodology employed here avoids the charge of parameter fitting by 1) employing multiple dependent measures and demanding that a single set of parameters provide reasonable fits to all dependent measures, and 2) adopting a canonical model and investigating the effects of parameter variation within that model. This second point ensures that data fitting is not performed without a thorough understanding of the roles of all parameters.

Conclusion

The present study supports previous work that suggests that recency and positive biases are strong determinants of behaviour in decision making and categorisation tasks. It extends this work by showing that such biases also have a role in cue selection and by showing that the biases may be incorporated in models developed within any of the three main theoretical approaches to diagnosis and categorisation.

The fact that the data presented cannot discriminate between the theoretical accounts is not of major concern. There is a long-standing debate about which of the three approaches to categorisation is "correct". This work suggests that this debate is framed at the wrong level. Specifically, we suggest that behavioural differences between the frameworks are masked by performance factors and processing biases, and, at least with respect to the task examined here, incorporation of these factors and biases into each framework can result in models that mimic each other's behaviour.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*, 249–277.

Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought.* Lawrence Erlbaum Associates, Mahwah, NJ.

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by Elimination: Using few cues to choose. In Gigerenzer, G., & Todd, P. M. (Eds.), *Simple Heuristics That Make Us Smart*, pp. 235–254. Oxford University Press, Oxford, UK.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A Study of Thinking*. John Wiley & Sons, New York, NY.

Castillo, E., Gutiérrez, J. M., & Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models*. Monographs in Computer Science. Springer, New York, NY.

Cooper, R. P. (2002). *Modelling High Level Cognitive Processes*. Lawrence Erlbaum Associates, Mahwah, NJ. Includes contributions by P. Yule, J. Fox and D. W. Glasspool.

Cooper, R. P., & Fox, J. (1998). COGENT: A visual design environment for cognitive modelling. *Behavior Research Methods, Instruments, & Computers*, *30*, 553–564.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 556–576.

Fox, J. (1980). Making decisions under the influence of memory. *Psychological Review*, *87*, 190–211.

Fox, J., & Cooper, R. P. (1997). Cognitive processing and knowledge representation in decision making under uncertainty. In Scholz, R. W., & Zimmer, A. C. (Eds.), *Qualitative Aspects of Decision Making*, pp. 83–106. Pabst Science Publishers, Lengerich, Germany.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Gigerenzer, G., & Todd, P. M. (Eds.). (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press, Oxford, UK.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*, 227–247.

Hearst, E. (1991). Psychology and nothing. *American Scientist*, *79*, 432–443.

Hertz, J. A., Krogh, A. S., & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Santa Fe Institute Studies in the Sciences of Complexity. Addison-Wesley, Reading, MA.

Hunt, R., & Rouse, W. (1981). Problem-solving skills of maintenance trainees in diagnosing faults in simulated power plants. *Human Factors*, *23*, 317–328.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, New York, NY.

Kendler, H. H., & D'Amato, M. F. (1955). A comparison of reversal shifts and nonreversal shifts in human concept formation behavior. *Journal of Experimental Psychology*, *49*, 165–174.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 3–26.

Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis. *Science*, *130*, 9–21.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, *27*, 986–1005.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *8*, 37–50.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68–85.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In Black, A. H., & Prokasy, W. F. (Eds.), *Classical Conditioning II: Current Research and Theory*. Appleton-Century-Crofts, New York, NY.

Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Ross, B. (1997). The use of categories affects categorization. *Journal of Memory and Language*, *37*, 240–267.

Shanks, D. R. (1991). Categorisation by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 433–443.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, *34*, 109–159.

Townsend, J. T. (1971). Some results on the identifiability of parallel and serial processes. *Perception and Psychophysics*, *10*, 161–163.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281–299.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of Reasoning: Structure and Content*. Batsford, London, UK.

Wiener, N. (1948). *Cybernetics*. Wiley, New York, NY.

Yule, P., & Cooper, R. (2001). Towards a technology for computational experimentation. In Altmann, E., Cleeremans, A., Schunn, C. D., & Gray, W. D. (Eds.), *Proceedings of the 4th International Conference on Cognitive Modelling*, pp. 223–228. Fairfax, VA. Lawrence Erlbaum Associates.