# Comparative Modelling of Learning in a Decision Making Task

**Richard Cooper (R.Cooper@psyc.bbk.ac.uk)**
Department of Psychology, Birkbeck College, University of London, Malet St., London, WC1E 7HX

**Peter Yule (P.Yule@psyc.bbk.ac.uk)**
Department of Psychology, Birkbeck College, University of London, Malet St., London, WC1E 7HX

**Abstract**

In this paper we compare the behaviour of three competing accounts of decision making under uncertainty (a Bayesian account, an associationist account, and a hypothesis testing account) with subject performance in a medical diagnosis task. The task requires that subjects first learn a set of symptom/disease associations. Later, subjects are required to form diagnoses based on limited symptom information. The competing theoretical accounts are embodied in three computational models, each with a single parameter governing the learning rate. Subjects' diagnostic accuracy was used to calibrate the learning rates of the models. The resulting parameter-free models were then used to predict subjects' symptom querying behaviour in a subsequent task. The fit between the Associationist model's predictions and subject behaviour was poor. The fit was slightly better in the case of the Bayesian model, but the hypothesis testing account proved to provide the most adequate account of the data.

## Introduction

Many decisions in real life are made under conditions of uncertain or incomplete information. Bayesian probability theory provides an optimal approach to such decisions when the uncertainty in the evidence can be quantified in the form of probabilities. In many cases, however, such quantitative information is not available, and even when it is, people frequently fail to make correct use of it, as in the well-known cases of base rate neglect (Kahneman & Tversky, 1973).

Although the Bayesian approach to decision making under uncertainty may be optimal, a number of other, sub-optimal, approaches yield plausible decision making behaviour. Gluck & Bower (1988), for example, have demonstrated that an associative network employing a Rescorla-Wagner (1972) learning rule can learn a disease categorisation task in which symptoms are probabilistically associated with diseases. The task is effectively a decision making task in which symptoms are unreliable indicators of possible diseases. Gluck & Bower compared the behaviour of their model with that predicted by a Bayesian account, and with that of human subjects. They report a correlation of 0.94 between subject performance and the model's performance, suggesting that non-Bayesian accounts are indeed capable of producing human-like performance.

Further concerns about the relevance of Bayesian approaches to human decision making under uncertainty are raised by Gigerenzer & Goldstein (1996), who suggest that the cognitive plausibility of such accounts is undermined by human performance limitations. They suggest that in real-life people rely on "fast and frugal" heuristics which approximate optimal behaviour. In support of this position Gigerenzer & Goldstein (1996) compared the decision making behaviour of four algorithms employing different forms of bounded computation with an optimal regression model. Several of the bounded algorithms achieved levels of performance equivalent to the mathematically optimal model. This work demonstrates that, at least on certain tasks, fast and frugal approaches are capable of producing near optimal decision making performance, and hence that viable alternatives to Bayesian decision algorithms do exist.

It would be wrong to suggest, however, that the results of Gluck & Bower (1988) and Gigerenzer & Goldstein (1996) provide unequivocal evidence against Bayesian processes in human decision making. Thus, although the correlation obtained by Gluck & Bower (1988) between subject performance and the performance of their associative network was high (0.94), it was not as high as the correlation between subject performance and that of a Bayesian approach (0.99). More critically, the root mean square error between disease probabilities as predicted by the associative network and the subjects was more than twice that between disease probabilities as predicted by the Bayesian account and the subjects. The correlation of 0.94 (obtained by choosing an appropriate value for the learning rate, a free parameter in the associative network) is less convincing when considered in this light.

The import of the results of Gigerenzer & Goldstein (1996) is similarly open to question. The difficulty here lies in the fact that their evaluation of non-Bayesian approaches did not involve comparison with human data. Fast and frugal algorithms were found to perform as well as a mathematically optimal account — one based on multiple regression, and, in fact, formally equivalent to an Associationist model — but human data was not collected on the task which they investigated.

In previous work (e.g., Fox, 1980; Cooper & Fox, 1997; Yule, Cooper, & Fox, 1998) we have compared the normative, quantitative, Bayesian account of decision making under uncertainty, and a qualitative, hypothesis testing account, with human performance on a medical diagnosis task. Subjects

Table 1: Conditional probabilities of symptom given disease

| | | Symptoms | | | | |
|---|---|---|---|---|---|---|
| | | *sym0* | *sym1* | *sym2* | *sym3* | *sym4* |
| Diseases | *dis0* | 1.00 | 0.00 | 0.25 | 0.00 | 0.25 |
| | *dis1* | 0.00 | 0.50 | 1.00 | 0.50 | 1.00 |
| | *dis2* | 0.50 | 1.00 | 0.00 | 0.00 | 0.25 |
| | *dis3* | 0.00 | 0.00 | 0.25 | 1.00 | 0.00 |

rarely achieved the levels of performance suggested by either computational account, but the general pattern of subject performance more closely resembled the hypothesis testing account. In this paper we extend that work by 1) adding reasonable performance limitations to the Bayesian and hypothesis testing accounts (thus bringing baseline performance into line with subjects); 2) extending the comparison by including an associationist (two-layer feed-forward network) model in the set of competing computational accounts; 3) modifying the experimental task such that subject performance on one dependent measure can be used to calibrate the various models, which may in turn be used to predict subject performance on a second dependent measure; and 4) examining the effect on human and predicted behaviour of different training histories.

We begin by describing the diagnosis task in detail and outlining the mechanisms behind the three models. We then report an experiment in which human subjects learned to perform the diagnosis task. This performance is then used firstly to fix the learning rate in each model, and then to evaluate the resulting parameter-free models.

## The Diagnosis Task

The task of diagnosis is essentially one of categorising a set of features or symptoms as corresponding to one of a set of known diseases. There are numerous variations on the task. Fox (1980) employed five symptoms and five diseases. All diseases were equally likely and subjects were required to query symptoms in sequence before offering a diagnosis. Gluck & Bower (1986) employed four symptoms and two diseases, but the occurrence of one disease was rare in comparison to the other. Here, subjects were allowed complete symptom information when making their diagnoses. The version of the task employed in the current work is derived from that of Fox (1980). Four diseases (hypothetical strains of 'flu) and five symptoms were employed. All diseases were equally likely, and symptoms were unreliable indicators of diseases. Table 1 shows the probability of each symptom occurring for each disease. Thus, one in four patients suffering from *dis0* would have *sym2*.

The diagnosis task can be presented in two forms. In the simplest form, the subject makes a diagnosis based on full symptom information. That is, the presence/absence of each symptom is known, and the subject must select which of the four diseases is most likely. With the symptom/disease associations employed here, it is always possible to discriminate between diseases based on full symptom information.

A more challenging form of the diagnosis task involves presenting subjects initially with one symptom (the presenting symptom), and requiring them to query just those symptoms necessary to make a diagnosis (and then to make a diagnosis when appropriate). This version of the task is naturalistic in that it corresponds closely to the task of a General Practitioner. It also yields rich data in the form of symptom querying strategies. However, the data are difficult to interpret because different subjects appear to employ different "diagnostic thresholds" — some subjects are willing to offer a diagnosis on the basis of few symptoms, whereas others query most or all symptoms before offering a diagnosis.

The freedom allowed to subjects in this more challenging version of the task, and its manifestation in the form of a diagnostic threshold, also poses methodological problems for evaluating computational accounts of subject behaviour. The diagnostic threshold effectively introduces a free parameter into the model of a subject, allowing the modeller an additional degree of freedom with which to account for subject behaviour.

The experiment reported below yields data on symptom querying strategies whilst overcoming the difficulty of diagnostic thresholds by presenting subjects with one symptom, and then requiring them to query exactly one further symptom before offering the most likely diagnosis. There is no scope for a diagnostic threshold in this form of the task. Whilst individual differences may still exist, such differences must be attributed to other factors (such as learning rate, or strategic elements).

In fact, previous research has shown large between-subject differences on diagnosis tasks (e.g. Yule, Cooper & Fox, 1998), even with little apparent variation in diagnostic threshold. The task can be taxing, and motivational factors are likely to play a role. However, between-subject differences may also be attributable to differences in training history. That is, if during training subjects are exposed to randomly generated sequences of cases of diseases, then their diagnostic behaviour may be influenced by idiosyncratic features of their training materials. This is especially likely to be true if training is limited. Training history is therefore another factor considered in the experiment reported below.

## Theoretical Accounts of Diagnosis

### The Bayesian Approach

The Bayesian approach to the categorisation element of diagnosis is straightforward and well documented (see, for example, Fox, 1980). The probability of each disease can be calculated from symptom information provided that disease base rates and the conditional probabilities of each symptom given each disease are known (assuming independence of symptoms). Approximations to each of these probabilities can be computed from frequency information, acquired as part of the learning of symptom/disease associations.

Symptom selection, when incomplete symptom informa-

tion is provided, may be determined through calculation of the informativeness of each indeterminate symptom. Oaksford & Chater (1994) suggest that cue selection on the basis of informativeness (in conjunction with assumptions about the distribution of cues) provides a good account of subject behaviour in related tasks, and point to appropriate information theoretic measures of informativeness (Shannon & Weaver, 1949; Wiener, 1948).

Previous research has shown that this Bayesian approach significantly outperforms subjects when learning the diagnosis task (Yule, Cooper & Fox, 1998). In order to reduce performance to human levels we suggest imperfect recording of frequency information relating to both base rates of diseases and disease/symptom co-occurrences. In particular, the model that generated the data reported below includes a parameter, the learning rate, which specifies the probability that frequencies will be updated on any given trial. Thus, when this parameter is set to 0.10 (a value that yields behaviour at levels comparable to human subjects), frequency information will be recorded (and hence employed in determining the informativeness of symptoms and the probabilities of possible diagnoses) on 1 in 10 trials on average.

### The Hypothesis Testing Approach

Fox (1980) described an approach to the diagnosis task which generated propositional hypotheses about possible diseases and reasoned over those hypotheses when determining appropriate symptoms to query or diagnoses to offer. Cooper & Fox (1997) extended that model to account for learning during the diagnosis task, and Yule, Cooper & Fox (1998) compared the performance of that model and a Bayesian model with subject performance on a version of the diagnosis task.

We do not describe the model again here, except to say that the presence or absence of symptoms triggers the model into forming hypotheses about possible diseases. These hypotheses lead the model to expect further symptoms, which form the basis of the model's querying strategy: the model will ask about symptoms if it expects them to be present given any of the hypothesised diseases, in an order determined by recency in its knowledge base. When provided with diagnostic feedback, the model adjusts its beliefs about the relations between symptoms and diseases, and about the symptom patterns associated with diseases.

In the model employed in the current work this updating is not performed on all trials — diagnostic feedback is sporadically ignored. The probability of updating on a given trial is determined by a learning rate parameter analogous to that in the Bayesian model. As in the Bayesian account, this probabilistic learning allows the model's diagnostic performance to be brought approximately into line with that of subjects.

### The Associationist Approach

The diagnosis task may also be performed by an associative network employing a Rescorla-Wagner (1972) learning rule. As noted above, Gluck & Bower (1988) argue that this learning rule, which is formally equivalent to the Delta rule, accounts well for human performance in their version of the diagnosis task.

Our implementation of the associationist approach follows traditional lines: a two-layered network maps five input nodes (corresponding to symptoms, and set equal to $+1.00$ when a symptom is present, $-1.00$ when a symptom is absent and 0.00 when a symptom's status is unknown) to four output nodes (corresponding to the four diseases). On testing cycles the symptom vector is fed to the network and the most active disease node is offered as the diagnosis. On training cycles the network's weights are adjusted according to the standard Delta rule. The learning rate constitutes a parameter of the model (again allowing the model's diagnostic accuracy to be calibrated with that of human subjects) that scales weight adjustments within the network, with weights changing by an amount equal to the learning rate times the difference between the network's output and the training signal.

Standard associative network approaches to the diagnosis task do not obviously generalise to the version of the task in which symptom selection is required. In our implementation, a second associative network is trained on instances of the identity map between symptom vectors corresponding to those symptom patterns actually presented to the main symptom/disease network. The rationale for this is that, after training, incomplete symptom information will be mapped by this network to a symptom vector resembling a previously seen pattern. The most active symptom in this vector whose presence is indeterminate is the symptom that is queried. This approach is far from optimal, but it is closer to associative principles than the most obvious alternative: to test the symptom/disease network on all possible extensions of the current symptom information and then select the symptom corresponding to the extension yielding the strongest diagnosis.

## Experiment: Rationale and Method

52 second year psychology students from Birkbeck College took part in the experiment. There were four conditions, with 13 subjects in each condition. The conditions differed only in the sequences of cases presented, as described below.

Subjects in all conditions completed 5 blocks of trials. Blocks 1 and 3 were training trials in which full symptom and disease information was presented to all subjects. Subjects were required to step through trials in these blocks at their own pace, whilst attempting to learn the symptom/disease associations. These blocks comprised 12 trials each. The degree of learning in the training blocks was assessed in blocks 2 and 4. Here, subjects were required to make diagnoses based on full symptom information. Feedback (in the form of the actual underlying disease) was given on these trials, allowing subjects to further improve their diagnostic accuracy. These testing blocks also consisted of 12 trials. In the fifth and final block subjects were presented with a single presenting symptom. They were required to query the presence of exactly one further symptom and then make a "best guess" diagnosis. Diagnostic feedback (i.e., the actual underlying disease) was given in the case of error. This block consisted

Table 2: Mean diagnostic accuracy (%) in each block and training history condition (Human data).

| *Training* | | *Block 2* | | *Block 4* | | *Block 5* | |
|---|---|---|---|---|---|---|---|
| *History* | N | Mean | SD | Mean | SD | Mean | SD |
| 1 | 13 | 50.6 | 25.1 | 55.1 | 24.7 | 49.6 | 21.0 |
| 2 | 13 | 50.6 | 19.7 | 62.8 | 20.3 | 48.5 | 20.2 |
| 3 | 13 | 60.9 | 22.7 | 59.6 | 22.5 | 53.1 | 21.9 |
| 4 | 13 | 50.0 | 18.0 | 59.6 | 14.8 | 49.2 | 11.5 |
| Total | 52 | 53.0 | 21.4 | 59.3 | 20.5 | 50.1 | 18.6 |

of 20 trials. Subjects were instructed of the block structure before commencing the task, and knew that they would be required to identify diseases based on minimal information in the final block. They were instructed that in this final block they should query the symptom that would be "most helpful to them in making their diagnosis".

Symptoms and diseases were related according to the probabilities given in Table 1, with four strains of 'flu (Austrian 'flu, Belgian 'flu, Greek 'flu, and Danish 'flu) being mapped onto the disease names and five common 'flu symptoms (headache, shivering, sore throat, coughing, and sneezing) being mapped onto the symptom names. This mapping was randomised across subjects. Three cases of each disease occurred in each of the first four blocks, with five cases of each disease in the final block. In order to examine the effect of training history the sequence of cases presented to each subject was generated from one of four random seeds. Subjects were allocated at random to a training history group. (Training history was thus a between-subject independent variable.)

The experiment was administered by networked software running over the department's intranet. The software, written in JavaScript for use with web browsers, randomly assigned subjects to each of the four training conditions and collected subject responses at the end of each block.[1]

## Results

**Diagnostic accuracy** Table 2 shows mean diagnostic accuracy for each training history condition, in each of the blocks in which a diagnosis was required, namely 2, 4 and 5. A two-factor, mixed-model ANOVA shows a significant effect of block ($F(2,96) = 7.70$, $p < 0.0008$), but no effect of training history ($F(3,48) = 0.28$) and no interaction ($F(6,96) = 0.91$). The increase in diagnostic accuracy between blocks 2 and 4 ($F(1,48) = 5.32$, $p < 0.0254$), and the decrease between blocks 4 and 5 ($F(1,48) = 20.89$, $p < 0.0001$), are both significant. So there is evidence of learning during the training phase of the experiment, and of disruption of performance by the different task in the final block, but no evidence of any effects of training history on diagnostic accuracy.

[1] A demonstration of the client system is available at http://redback.psyc.bbk.ac.uk/expts/jdm4/demo/

Table 3: Overall mean diagnostic accuracy (%) in each block for each model ($N = 52$ each).

| | | *Block 2* | | *Block 4* | | *Block 5* | |
|---|---|---|---|---|---|---|---|
| *Model* | L.R. | Mean | SD | Mean | SD | Mean | SD |
| Bayes. | 0.10 | 39.3 | 16.3 | 60.3 | 17.4 | 74.8 | 15.1 |
| Hypot. | 0.25 | 41.0 | 15.6 | 57.8 | 13.9 | 56.3 | 16.1 |
| Assoc. | 0.015 | 36.3 | 17.2 | 58.0 | 20.3 | 57.0 | 14.6 |

**Model calibration** Each model contains one free parameter, a learning rate, that determines the speed and accuracy of learning. Human diagnostic accuracy data for block 4 was used to calibrate learning in all models as follows. For each model, simulations (comprising 52 virtual subjects, 13 in each training history condition) were conducted for a range of learning rates. Learning rates in each model were then fixed at values leading to diagnostic accuracy on block 4 which was most commensurate with the human data. This approach resulted in a learning rate of 0.10 for the Bayesian model, 0.015 for the Associationist model, and 0.25 for the Hypothesis Testing model. (These rates are not comparable because of the different learning mechanisms within each model.) Once calibrated, all models make parameter-free predictions for symptom query patterns and diagnostic accuracy on the final block. The experiment was deliberately designed to allow this approach to model testing.

Table 3 shows mean diagnostic accuracy for each calibrated model. It is clear that all the models show larger differences between blocks 2 and 4 than do human subjects. Also, the Bayesian model's diagnostic performance improves markedly on the final block, whereas the other models' performances do not, but none of the models show the significant decrease in performance observed in the human data.

**Human symptom queries** Table 4 shows frequencies of each possible symptom query for the final instance of each presenting symptom in the final block. For each row, there is a $\chi^2$ test for nonrandom distribution of queries (d.f.=3). From the table, there are significant departures from random distribution for queries following *sym1* presenting and *sym4* presenting. By inspection of the peaks in each row, we can see that given *sym1* presenting, human Ss tend to query *sym0*, and given *sym4* presenting, Ss tend to query *sym2*.

**Bayesian Model** Table 5 shows the final symptom query frequency table for the Bayesian model. There are only two significant query biases, for *sym0* and *sym3* presenting. These do not correspond to either of the significant human biases. However, given *sym0* presenting, the Bayesian model tends to query *sym1*, and although the human effect does not reach significance, its maximum is also *sym1*. But the Bayesian tendency to query *sym4* given *sym3* is not reflected in the Human data.

Moreover, the significant human effects are not paralleled

Table 4: Final symptom query frequencies for each presenting symptom (Human data, row $N = 52$).

| Pres. Sym. | Query *sym0* | *sym1* | *sym2* | *sym3* | *sym4* | $\chi^2(3)$ | p |
|---|---|---|---|---|---|---|---|
| *sym0* | | 19 | 9 | 14 | 10 | 4.77 | |
| *sym1* | 19 | | 9 | 14 | 10 | 9.69 | .021 |
| *sym2* | 13 | 8 | | 14 | 17 | 3.23 | |
| *sym3* | 11 | 17 | 13 | | 11 | 1.85 | |
| *sym4* | 14 | 6 | 21 | 11 | | 9.08 | .028 |

Table 5: Final symptom query frequencies for each presenting symptom (Bayesian model, row $N = 52$).

| Pres. Sym. | Query *sym0* | *sym1* | *sym2* | *sym3* | *sym4* | $\chi^2(3)$ | p |
|---|---|---|---|---|---|---|---|
| *sym0* | | 26 | 7 | 9 | 10 | 17.69 | .001 |
| *sym1* | 15 | | 18 | 8 | 11 | 4.46 | |
| *sym2* | 13 | 14 | | 10 | 15 | 1.08 | |
| *sym3* | 6 | 10 | 12 | | 24 | 13.85 | .003 |
| *sym4* | 14 | 15 | 12 | 11 | | 0.77 | |

Table 6: Final symptom query frequencies for each presenting symptom (Hyp. Testing model, row $N = 52$).

| Pres. Sym. | Query *sym0* | *sym1* | *sym2* | *sym3* | *sym4* | $\chi^2(3)$ | p |
|---|---|---|---|---|---|---|---|
| *sym0* | | 25 | 12 | 6 | 9 | 16.15 | .001 |
| *sym1* | 27 | | 7 | 3 | 15 | 25.85 | .001 |
| *sym2* | 5 | 2 | | 5 | 40 | 75.23 | .001 |
| *sym3* | 4 | 7 | 18 | | 23 | 18.61 | .001 |
| *sym4* | 1 | 9 | 33 | 9 | | 44.30 | .001 |

Table 7: Final symptom query frequencies for each presenting symptom (Associationist model, row $N = 120$).

| Pres. Sym. | Query *sym0* | *sym1* | *sym2* | *sym3* | *sym4* | $\chi^2(3)$ | p |
|---|---|---|---|---|---|---|---|
| *sym0* | | 33 | 35 | 22 | 30 | 3.27 | |
| *sym1* | 25 | | 25 | 42 | 28 | 6.60 | (.086) |
| *sym2* | 35 | 24 | | 31 | 30 | 2.07 | |
| *sym3* | 28 | 25 | 23 | | 44 | 9.13 | .028 |
| *sym4* | 32 | 29 | 29 | 30 | | 0.20 | |

by the corresponding non-significant Bayesian effects; given *sym1* the Bayesian model tends to query *sym2*, unlike the human preference for *sym0*, and given *sym4* the Bayesian model queries *sym1*, not *sym2*.

**Hypothesis Testing Model** Table 6 shows the final symptom query frequency table for the Hypothesis Testing model. This exhibits strong, highly significant biases for each presenting symptom. With regard to the presenting symptoms which give significant querying biases in the human data, *sym1* and *sym4*, the Hypothesis Testing model generates maxima in the same places as do human subjects; it tends to query *sym0* given *sym1* (cf. Bayesian model), and *sym2* given *sym4*. But also, even where the human bias is non-significant, the model still predicts the most frequent query correctly in two of three cases, with *sym0* and *sym2* presenting, and fails to predict the human bias only with *sym3* presenting.

**Associationist Model** Unfortunately the Associationist model produced no significant symptom querying biases at all when calibrated to the human level of diagnostic accuracy and subject numbers. Consequently these data are not presented; instead, the model was rerun with a larger number of virtual subjects (120), at the same learning rate, resulting in Table 7.

As Table 7 shows, there is only one significant symptom query bias, for *sym3* presenting, when the model tends to query *sym4*. (Curiously, all the models predict the same bias for *sym3*, but the human bias is non-significant and in a different direction.) The Associationist model also shows an almost-significant bias to query *sym3* given *sym1*, unlike the other models and unlike the significant human tendency to query *sym0*.

## Discussion of results

The relative absence of significant biases in the Bayesian model symptom query data is attributable to the large amount of random variance in the model's behaviour, a consequence of its low learning rate. With higher learning rates, or with larger numbers of virtual subjects at the same learning rate, the model's predictions are quite clear for all presenting symptoms. But as things stand, such predictions as there are from the Bayesian model are not very well borne out in the human data.

Unfortunately, even with large numbers of virtual subjects the predictions of the Associationist model are minimal, and do not correspond to human query biases. So we can conclude that of the three, the Associationist model gives the poorest account of the human data.

The Hypothesis Testing model successfully predicts the significant symptom querying biases in the human data, as well as the directions of most of the non-significant ones, so it easily fares best of the three models. We are in the process of collecting more human data, in order to determine if more of the human query biases are significant. With more human data, we also expect to be able to investigate possible effects of training history on symptom querying strategies.

The assumption that human levels of performance can be simulated by manipulating models' learning rates was reasonably successful, in that it produced a good fit between human symptom querying patterns and the predictions of one of the models. However, none of the models generated learning curves of the same shape as the human curve, since the cali-

brated models performed more poorly than humans on block 2, and better than humans on block 5. It seems that humans reach peak performance quite early in the experiment, but find it hard to improve much thereafter. Their performance is then significantly disrupted on the final block, where the task is somewhat different.

## General Discussion

The observed superiority of the Hypothesis Testing model over the Bayesian in fitting human symptom querying behaviour replicates previous findings (Yule *et al.*, 1998; Fox, 1980), despite substantial variations in task, materials, experimental interface and methods of analysis. The Hypothesis Testing model owes its success to two factors: it only queries symptoms expected to be present given any of the hypothesised diseases (i.e., it contains a confirmation bias), and queries are ordered according to a recency principle in memory (such that behaviour is determined more by recent events than by those in the more distant past). It is reasonable to ask if incorporation of these biases in the other models would improve their fit with the human data.

With respect to the Bayesian model, symptom queries are selected on the basis of expected information gain, and as in previous studies, while the model can predict human behaviour in a few cases, it does not yield a good overall fit with human questioning patterns. Oaksford & Chater (1994) have argued that Bayesian approaches in similar information seeking tasks can give a good account of human performance when they are supplemented with a "rarity assumption". In the current task, such an assumption would have the effect of restricting the search for evidence to symptoms expected to be present given the most likely diseases. In other words, in the current task such an assumption would amount to a Bayesian implementation of a confirmation bias. Incorporation of the rarity assumption into the Bayesian model is therefore of some importance.

The second factor contributing to the Hypothesis Testing model's superior performance, recency, might also be incorporated into a Bayesian model by weighting recent events in the estimation of event frequencies used to determine the various numerical factors required by Bayes' theorem. Such a weighting is appealing given recency effects, but its incorporation would add an extra parameter to the Bayesian model, thus raising further difficulties in model evaluation.

The Associationist model could also benefit from a re-evaluation of its approach to symptom querying. The difficulty here is that in the first four blocks the model is given full symptom/disease information. There is no obvious way in which an associative network can produce sequential symptom querying behaviour from this static information. Standard associative network approaches to sequencing (recurrent networks) offer little assistance with this problem. Network models employing competitive activation may be appropriate, but such models have little in common with the associative framework from where we started.

A final methodological point is in order. The precise form of the experiment was dictated by the requirements of model testing. We have not simply attempted to fit models to the data. Rather, the experiment was designed to yield two dependent measures: diagnostic accuracy and symptom query strategy. The first measure was used to fix the single free parameter in each model. The result was a set of predictive models. Tables 5, 6, and 7 are model *predictions* — generable (in principle) before subjects begin the final block of the experiment. Few cognitive models are parameter free. For those that are not, we strongly advocate a methodology such as ours where the requirements of model evaluation determine aspects of subsequent empirical work. This methodology, we aver, is far more sound than the more common approach of data fitting via the adjustment of parameter values.

In sum, the comparative evaluation of three very different models of decision making under uncertainty, as applied to the medical diagnosis task, leaves us cautiously optimistic with respect to fast and frugal alternatives to optimal Bayesian accounts. The evidence for purely associative processes, however, appears weak.

## References


Cooper, R., & Fox, J. (1997). Learning to make decisions under uncertainty: The contribution of qualitative reasoning. In Langley, P., & Shafto, M. (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 125–130.

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

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650–669.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General, 117*(3), 227–247.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

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.

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

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

Yule, P., Cooper, R., & Fox, J. (1998). Normative and information processing accounts of medical diagnosis. In Gernsbacher, M. A., & Derry, S. J. (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pp. 1176–1181. Madison, WI.