The value of basic science in clinical diagnosis: creating coherence among signs and symptoms

Nicole N Woods, Lee R Brooks & Geoffrey R Norman

BACKGROUND We investigated whether learning basic science mechanisms may have mnemonic value in helping students remember signs and symptoms, in comparison with learning the relation between symptoms and diagnoses directly.

PURPOSE To compare 2 approaches to learning diagnosis: learning how features of various conditions relate to underlying pathophysiological mechanisms and learning the conditional probabilities of features and diseases.

METHODS Undergraduate students \((n=36)\) were taught 4 disorders (upper motor neuron lesion, lower motor neuron lesion, neuromuscular junction disease and muscular disease), either using basic science explanations or \((\text{symptom} \times \text{disease})\) probabilities. They were tested with diagnostic cases immediately after learning and 1 week later.

RESULTS On the immediate test, there was no difference in the results. One week later, the accuracy of the mechanism group remained at 0.52, but the performance of the probability group had dropped to 0.43.

CONCLUSIONS Knowledge of basic science may have value in clinical diagnosis by helping students recall or reconstruct the relationships between features and diagnoses.

KEYWORDS education, medical, undergraduate/standards; professional competence/standards; diagnostic techniques and procedures/standards; teaching/methods; science/education.

INTRODUCTION

Medical students spend a minimum of 2 years studying basic science. Paradoxically, although educators may believe that this is a necessary foundation for clinical medicine, studies of clinician reasoning have shown little evidence that clinicians use basic science in routine diagnosis.\(^1\) It would seem that an understanding of mechanisms has little direct heuristic value to the diagnostic task. The results of studies by Wolf\(^2\) et al. and Elieson and Papa\(^3\) suggest that students benefit most in performing diagnosis from learning mathematical probabilities. Participants in the study by Elieson and Papa\(^3\) were divided into 6 groups. Three of these groups were of particular interest to the present study; 1 group received a matrix of conditional probabilities; another received the same probability information in narrative form, and another received the probabilities converted to phrases like ‘usually’ or ‘infrequently.’ Mean post-test scores for the 2 groups that received probability information were 70\% and 69\%, respectively, while that for the group that received only the verbal probability descriptors was 58\%. Elieson and Papa\(^3\) concluded that:

‘... the use of soft text descriptors should be avoided ... and that medical students would benefit from having either conditional probabilities presented in matrix form or prototypical descriptions of the diseases which they could more easily use to develop their own mental associations and/or prototypes.’ (p 83)

But this may not tell the whole story. By placing the focus on probability matrices and prototypes, the necessity for students to learn, retain and integrate... 
information has been neglected. While these probability data may have greater initial utility, they are likely to be much more difficult to remember and recall than richer narrative information. This may also be the area in which the students’ knowledge of basic science becomes critical. Basic science may be useful for creating a coherent mental representation of diagnostic categories and their signs and symptoms. Students may well use an understanding of mechanisms to help them remember or reconstruct the features of a disease. Conversely, the likelihood that students can retain probabilistic information over time has not been addressed in these studies.

There is certainly a large body of evidence from psychology that meaningful material is retained and recalled better than non-meaningful material. Perhaps the earliest evidence arose from studies of chess masters,4 where the master showed vast superiority to novices in recall of legitimate chess positions, but performed at a novice level when positions were placed at random on the board. Norman et al.5 showed similar findings in nephrology: expert nephrologists were much better able than medical students to recall electrolyte data when the data represented legitimate configurations, but no better with random data.

Arguably, lists of signs and symptoms or probability matrices contain very little meaningful information. It may be that basic science, by providing causal explanations linking diseases with signs and symptoms, may provide a framework of meaning and hence aid retention and recall. In the present study, we directly assessed the value of basic science explanations against direct probability estimates in assisting students in making clinical diagnoses. Critically, performance was assessed immediately after initial learning and again 1 week later in order to test retention.

Research hypotheses

1. The provision of a basic science explanation linking signs and symptoms to diseases will result in better memory of features and diagnostic performance after a time delay.
2. There will be a correlation between diagnostic ability and mastery of (a) causal basic science knowledge, or (b) conditional probabilities.

METHOD

Participants

Participants were recruited from an introductory psychology subject pool. We elected not to use medical students in order to ensure that all subjects were complete novices with virtually no knowledge of neurological disorders. A total of 36 students participated.

Stimuli

Four neurological disease categories were selected: muscle disorders, neuromuscular junction disorders, upper motor neuron lesions and lower motor neuron lesions. The principal complaint for each disorder was muscle weakness. The symptoms and disorders are identical to those used by Elieson and Papa.3

Training material

A set of written material describing the 4 disease categories was created for each of the learning conditions. Each training booklet contained essentially the same diagnostic information; only the specific format of presentation varied. In the probability condition (PB), the training material consisted of a table listing the 4 disorders (in columns) and 18 features (in rows). Numerical values indicated the
probability of each feature, given a particular disorder. In each instance, the probability estimate was determined by the consensus of independent ratings by 3 neurologists. While no formal assessment of agreement among the 3 neurologists was conducted, it was our impression that the agreement was remarkably high.

In the general science (GS) condition, the study booklet included a brief overview of neuroanatomy, describing communication between the motor neurons, the neuromuscular junction and the muscle. The specific symptoms were described as resulting from various disruptions to the system. The conditional probability for each symptom was translated into what Elieson and Papa referred to as ‘soft’ descriptors. Features that had an associated probability of 0.60 or greater were referred to as ‘usually occurring’. Features with a probability between 0.20 and 0.50 were described as ‘sometimes occurring’. Those features with a probability less than 0.20 were omitted from the category.

**Supporting tests**

In an attempt to encourage subjects to learn all available material, the experiment included a test intended to measure learning of the supporting information from each condition – information that could aid diagnosis but would not necessarily be part of a feature list or rule. For the general science condition, the supporting test included questions on the pathways and causation associated with symptoms. The supporting test for the probability condition asked subjects to complete blank cells of the probability matrix. For counterbalancing purposes, 2 forms of each test were created, matched for difficulty. The test was administered on both testing occasions (see below).

**Diagnostic tests**

The testing booklets included 15 cases. Each case consisted of the name, age and at least 4 presenting symptoms for a fictional patient. The specific cases and correct diagnosis were selected from a database of over 100 cases provided to us by Dr F Papa. Each case had been reviewed and diagnosed by a doctor. The correct answer, therefore, was based on the diagnosis at the time of treatment.

The symptoms for each case were also submitted to Bayesian calculations, producing the most probable disorder. This allowed us to verify that it was in fact possible to arrive at the correct diagnosis using the probabilities provided by our doctors. For counterbalancing purposes, 2 forms of the diagnostic test booklet were created. Both forms consisted of 15 cases, matched for difficulty. During the immediate test, half of the participants in each condition received form A while the remaining half were tested on form B. The forms were then reversed for the delayed test.

**Procedures**

Participants undertook the study in cohorts of up to 6 people. In both conditions, the students were given 25 minutes of study with the instructions that they were to learn to diagnose each of the 4 disease categories. They were also informed that they would be tested on all of the material in the study booklets, including exact probabilities and diagrams. Participants were aware that they would be re-tested 1 week later and that they would not have the opportunity to study again.

Immediately after training, participants turned in their study booklets and completed supporting test 1. No time limits were imposed. Upon completion of the test, participants looked at the answer key and compared their answers. The experimenter discussed each answer and encouraged participants to ask questions about any unclear responses. This feedback ensured that all participants had the correct supporting information before writing the diagnostic test. Each participant then completed the diagnostic test. Again, no time limits were imposed.

The delayed testing phase took place 1 week after initial training. There was no opportunity to review the test material. In the test, participants were given an additional 15 cases from the alternative testing form. After completing the diagnostic test, participants completed a novel test on supporting information (supporting test 2). Once their score had been tabulated, each participant received a cash reward, depending on their performance in supporting test 2. The reward had been explained to participants before the beginning of the experiment and was intended to improve motivation.

**RESULTS**

The number of correct responses (maximum = 15) on both immediate and delayed diagnostic tests was recorded for each participant. A $2 \times 2$ repeated measures ANOVA was conducted using the experimental condition (probability/general science) as the
between-subject variable and time (immediate/delay) as the within-subject variable. The scores on the immediate diagnostic test were almost identical for both groups. On the delayed test, however, participants in the PB group showed a reduction in accuracy but those in the GS group maintained their initial performance (Table 1). The results were confirmed by the ANOVA, which showed a significant interaction between time and condition, $F(1,34) = 4.65$, $P < 0.05$. The difference amounted to an effect size of $0.09 / 0.19 = 0.47$, which is in the range of a moderate effect size.

These results were mirrored in the supporting tests. Overall, students in the general science group performed better (although, of course, the tests were not directly comparable). More critically, they showed that they had forgotten less after 1 week than the probability group (Table 2).

To examine the relation between understanding of the concepts and performance on the diagnostic test, we examined the correlation between performance on the relevant supporting test and the diagnostic test for immediate and delayed testing in both groups (Table 3). The pattern of correlations suggests that, on immediate testing, performance was less strongly related to the understanding assessed in the supporting test than it was on delayed testing. After 1 week, the correlations were in the range of 0.63–0.73. A possible explanation is discussed below.

**DISCUSSION**

Performance on the immediate diagnostic test shows that participants in both conditions were equally successful in acquiring knowledge useful for diagnosing the cases. For the group provided with the accurate probabilities, it was entirely possible to use Bayes’ theorem (or some simpler heuristic applied to the conditional probabilities) to identify the correct diagnosis. For the other group, the basic science descriptions provided were equally useful in isolating those features relevant to diagnosis.

However, the real benefit of acquiring an understanding of basic science mechanisms lies in retention and retrieval, which became evident only after the passage of time. For participants in the probability condition, scores on the diagnostic test fell off after 1 week, as did their memory of the probabilities. Conversely, students who learned the basic science forgot little over a week, and their diagnostic scores stayed constant.

Elieson and Papa argued that students benefit most from clearly formatted, quantitative information, but their experiment only included an immediate test. As we showed, whatever advantages may exist from the learning of probability information (and in this study there were none) may be expected to decay over time. Conversely, students in the general science condition showed no loss of performance over time. There is evidence from psychology to support this observation. Murphy and Medin presented the idea that concepts are organised around personal theories and argued that the coherence of a concept depends on its fit with a theoretical framework. In our experiment, participants in the general science condition learned a neuroanatomy causal frame-

---

**Table 1** Scores (percent correct) on the diagnostic test immediately after instruction and 1 week later

<table>
<thead>
<tr>
<th></th>
<th>Immediate Mean</th>
<th>SD</th>
<th>Delayed Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.54</td>
<td>0.16</td>
<td>0.43</td>
<td>0.17</td>
</tr>
<tr>
<td>(n = 18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General science</td>
<td>0.52</td>
<td>0.16</td>
<td>0.52</td>
<td>0.21</td>
</tr>
<tr>
<td>(n = 18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2** Scores (percent correct) on the supporting test immediately after instruction and 1 week later

<table>
<thead>
<tr>
<th></th>
<th>Immediate Mean</th>
<th>SD</th>
<th>Delayed Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.42</td>
<td>0.22</td>
<td>0.25</td>
<td>0.14</td>
</tr>
<tr>
<td>General science</td>
<td>0.74</td>
<td>0.18</td>
<td>0.64</td>
<td>0.17</td>
</tr>
</tbody>
</table>

**Table 3** Correlation between scores on the supporting and diagnostic tests for immediate and delayed tests

<table>
<thead>
<tr>
<th></th>
<th>Immediate</th>
<th>Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.03</td>
<td>0.73</td>
</tr>
<tr>
<td>General science</td>
<td>0.36</td>
<td>0.63</td>
</tr>
</tbody>
</table>
work that could be used to add coherence to the feature list.

Why do we see that the correlation between the supporting and diagnostic test increased over time? One possibility is this: at the time of initial learning, the presented information, either mechanistic or probabilistic, is synthesised to create a list of diagnostic features for each condition. In completing the immediate diagnostic test, the participant reads the features of the case and starts a memory search for a direct matching of features. The category with the most features present is selected as the appropriate diagnosis. Participants can use the supporting information (either the probability estimates or the causal pathways) to weight the features and resolve any conflicts or ties. As the diagnostic approach amounts to a count of features present in the case, the relation between memory of the specific aspects of the initial information (probabilities or causal pathways) and diagnostic accuracy at immediate test, is relatively low.

After a delay, this process changes. Participants in both conditions may once again begin with a memory search for basic features. It is likely in both conditions that the individual features have become far more confused and the category boundaries less clear. Participants now need some way of reconstructing the features. For the probability group, this reconstruction relies on the relatively poorly remembered probabilities. In contrast, the causal model presented to the general science group provides a framework that can now be used to reorganise and recreate the feature list. As a consequence, there is now a much stronger relationship between performance on the supporting test and diagnostic performance for both groups than was initially the case. However, because the probabilities are less memorable, overall performance drops.

The study has some potential limitations. The use of undergraduate students, while necessary to ensure that they did not bring prior knowledge to the task, might limit generalisability. The use of subjective probability estimates might be seen to limit validity, and it might be considered preferable if published values were used. However, this is both impractical and unnecessary. Firstly, it would be very unlikely to have all 72 conditional probabilities available from a single population. Secondly, for the purposes of this particular experiment, agreement with the published literature was not particularly relevant. For the group learning conditional probabilities, optimal performance would result from making a Bayesian calculation using these probabilities, just as we did in computing the ‘correct’ diagnosis. The group learning basic science would clearly have to use a simpler strategy amounting to counting supporting features, and whether this would handicap their performance is an empirical question. However, the contrast between the groups would be equally valid in terms of the specific research questions, whether they were learning artificial diseases or features labelled as ‘X’, ‘Y’ and ‘Z’.

In conclusion, the study has shown that, in comparison with students who learned conditional probabilities, students provided with a basic science explanation for diagnostic categories were better able to accurately diagnose cases after a delay. A plausible explanation is that the basic science information, because of its conceptual coherence, was itself more memorable, and that it also provided a means to reconstruct the features of individual disease categories after the initial symptom lists had been forgotten.

Contributors: NNW is a graduate student in psychology. She was principally responsible for conceptu alising and conducting the study and writing the manuscript. LRB is a professor in psychology and GRN is a professor in epidemiology and biostatistics. Both contributed to the discussions around the design of the study, the analysis and the writing of the paper. Acknowledgement: we acknowledge the generosity of Dr Frank Papa in contributing some of the materials used in the study. Funding: the study was funded in part by grants to LRB and GRN from the Natural Sciences and Engineering Research Council (NSERC) of Canada, and by a Canada Research Chair. Conflicts of interest: none. Ethical approval: ethical approval for the study was received under the terms of GRN’s NSERC grant.

REFERENCES


Received 18 August 2003; editorial comments to authors 18 November 2003; accepted for publication 1 March 2004