Logic in Medicine

Diagnostic systems as an aid to clinical decision making

R P KNILL-JONES

My contribution to this series on logic in medicine aims to carry the theme of "diagnostic logic" discussed by Macartney further in the direction of clinical decision making. The example of dyspepsia has been chosen, partly because it has long been one of my special interests; the underlying principles exemplified are, however, applicable to many other clinical problems.

Dyspepsia diagnosis

However defined, dyspepsia is a common condition that accounts for a considerable proportion of consultations in general practice, referrals to hospital, and working days lost to illness. Diagnosis is often difficult and negative results of investigations common, which leads to many patients having the symptoms but no evidence of an organic cause for them. An accurate history is therefore essential, yet one may not be taken because of lack of time, the easy availability of investigations such as endoscopy (which seems sometimes to be used as an alternative to a properly taken history), or a simple failure to appreciate the discriminating value of standardised questions. Several solutions to this problem have been explored, including the use of standard forms that can be entered on a computer, automated history taking, which can be used as a decision aid, and simple weighting systems for symptoms.

All these solutions aim to reduce the diagnostic uncertainty of the doctor so that more rational therapeutic, investigative, and referral procedures develop. Diagnostic uncertainty about patients who have dyspepsia is high in general practice and is probably equally high among junior hospital staff; despite such uncertainty, there is little teaching about which are the best questions to ask—that is, which are the Popperian searchlights alluded to by Macartney.

What is required is a simple screening system that minimises unnecessary referrals and negative results of investigations. The development of such a system has many similarities to problems in other clinical specialties. For example, in the assessment of risk for ischaemic heart disease the use of weights (or scores) for each factor is becoming widespread. In this instance the approach is a practical one for assessing risk, and it can help patients to reduce their own risk factors. In dyspepsia a similar approach should lead to better use of existing resources and in particular improve the overall management of patients when care is shared between the general practitioner and the hospital based specialist.

This paper shows first how a simple weighting system for symptoms can be devised, given some carefully collected data. The method is clearly related to that described by Macartney but is expressed in a different way. For application in clinical practice this process has been partly automated: a computer asks the relevant questions of the patient and prints the results as a summary; this is effectively a screening process indicating which patients require referral and why, and, for those who do not indicating appropriate management to be carried out in general practice.

Weighting of symptoms

There are several steps in the process of developing weights for symptoms.

(1) DEFINING THE POPULATION OF PATIENTS

The first step is to define the population of interest. In this example I have used a wide definition of dyspepsia which includes many patients with non-acute gastrointestinal symptoms. This broad definition reflects the poor localisation of symptoms within the gastrointestinal tract and is as follows: "Episodic, persistent, or recurrent abdominal pain or discomfort or any other symptoms referable to the alimentary tract except for rectal bleeding and jaundice as the main symptom."

(2) DEFINING (RELEVANT) SYMPTOMS

The second step is to define the symptoms relevant to dyspepsia. This is much more important than is commonly realised, the underlying assumption being that we all know what is meant by simple terms such as "flatulence" and "nausea." If colleagues are asked to write down definitions, however, there are big differences, not only in length but also in meaning. For example, "flatulence" to 19 of 40 gastroenterologists meant passing wind upwards, to seven passing wind through the rectum, and to 11 passing wind in either direction. Thus the evidence that we can discard explicit definitions in favour of our implicit ones is not good. We have to agree on some kind of definition of a symptom and try to stick to it when collecting data. Clinicians are not entirely comfortable with this concept, but epidemiologists are, having learnt to live with it.

In our studies on dyspepsia definitions have been given for every symptom on the form used for collecting data; though this does not completely remove inconsistencies among observers, it does seem to reduce them. For example, the written definition we have used for flatulence is: "The patient brings up what he regards as an excessive amount of wind," while for nausea the patient feels "sick without actually being sick." By contrast, the form widely used for assessing acute pain in the abdomen does not contain definitions of each term used; definitions, however, are now available on separate
Some 1200 patients have now been seen in the study of dyspepsia at the Southern General Hospital in Glasgow, and, of these, nearly half were aged under 40 at the time of presentation, just over half having had symptoms for less than two years. The commonest presenting symptom was abdominal pain or discomfort (64%), the next commonest being vomiting and diarrhoea (7% each) and heartburn (5%). On direct questioning 87% of patients agreed that they had had some abdominal pain or discomfort, even if in a quarter it had not been their presenting symptom. One fifth of all our patients had had at least one episode of pain sufficiently severe to call in a doctor as an emergency or to attend an accident and emergency department. Few had a final diagnosis normally recognised as causing an "acute abdomen." Many must therefore have been patients who when presenting with acute pain were discharged with a diagnosis of "non-specific abdominal pain," a term used for 34% of presentations in a recent multicentre study. There are many definable and treatable diagnoses in this heterogeneous group, which can be somewhat misleadingly used as a diagnostic dustbin.

These first two steps allow us to define an area of interest and collect data in a reasonably reproducible way. An analysis of these data allows a population of patients to be described simply—giving for example, the prevalence of symptoms in a hospital clinic, as was done briefly in the preceding paragraph.

DEFINING THE DIAGNOSIS

The third step is to define the diagnosis. To find out which symptoms are diagnostic rather than descriptive it is clearly essential to obtain a "final diagnosis" for every patient. This is not particularly easy for dyspepsia. Although we could not provide any diagnosis at all in only 2% of our patients (usually because of failure of follow up), we found it difficult to be absolutely certain about a final diagnosis on many occasions. Furthermore, there is always some residual uncertainty about whether a patient's symptoms are in fact due to the particular diagnosis with which he has been labelled, given the poor specificity of symptoms arising from the gastrointestinal tract. This also makes an algorithmic approach inherently implausible, at least for most gastrointestinal disease.

As a result of this uncertainty, early in the study it proved to be necessary, firstly, to allow the possibility of there being more than one gastrointestinal diagnosis in a patient, and, secondly, to indicate for each final diagnosis given whether it was certain, probable, or possible. This reflected the completeness of follow up and the adequacy of investigation. For instance, a patient who had a duodenal ulcer that was seen clearly at endoscopy and noted then as being present without any doubt and was confirmed after a barium meal would have a "certain" final diagnosis of a duodenal ulcer. If in addition the patient had several symptoms of irritable bowel syndrome, sufficient to fit a pattern about which there was considerable agreement, then he would have a second diagnosis of irritable bowel syndrome. The order in which the diagnoses were given would reflect the clinician's opinion on which was the more important condition for that patient. About one third of our patients had a second diagnosis.

Recording the final diagnosis in this way is important for developing weights for symptoms. In an analysis of the symptoms of patients who have a duodenal ulcer the patients can be selected according to whether they have one or more than one final diagnosis. If the weights of the symptoms seem to be similar then both groups could be combined; if not then care would be necessary in selecting both patients and symptoms.

The same procedure can be applied to the certainty of final diagnosis. If patients who have a certain final diagnosis of duodenal ulcer are symptomatically similar to those who have a probable diagnosis both groups can be included when calculating symptom weights.

Table I shows the broad groupings of the final diagnoses in the first 1200 patients in our study. No final diagnosis could be made in 2%, who were excluded. About one quarter of all the patients had duodenal ulcer disease, and a further 74 (6%) had gastric ulcer. Another quarter had non-organic dyspepsia or functional dyspepsia, and to these some might wish to add the 177 patients (15%) who had irritable bowel syndrome. The final diagnosis was certain in about 60% of the patients who had duodenal ulcer, nearly 81% of the patients who had gastric ulcer, about 40% of those who had non-organic dyspepsia, and half of the patients who had a primary diagnosis of irritable bowel syndrome.

EVALUATING THE WEIGHTS OF SYMPTOMS NUMERICALLY

There have been many reported examples of the use of Bayes's theorem to produce a diagnostic aid of some kind—for example, in jaundice and in the very different problem of diagnosis after head injury. In some of these problems the presentation of the basic method has been altered so that clinicians are given a series of weights of symptoms that can be printed on a simple card and added up according to which symptoms the patient has. Thus the fourth step, the data on patients having been obtained and their problems finally diagnosed, is to analyse the data in such a way as to obtain the weights of symptoms. The most successful example of this has been for jaundice, where the COMIK group in Copenhagen have described and tested a pocket chart based on a scoring system derived from Bayes's theorem. This work developed from earlier work on jaundice by way of a careful comparison, carried out in 1980, of variation among observers in assessing symptoms. Similar scoring systems have not been developed for the acutely painful abdomen, although several groups interested in the problem will probably produce one soon.

The assumption is, of course, that the explicit provision of weights for symptoms is better understood by clinicians than a little black box that calculates the probability of disease in the way that the classical applications of Bayes's theorem have done. Although there is no definite evidence to confirm this assumption, it does seem likely to be right.

To show how simple weightings for symptoms can be developed I shall show how such a system differentiates between two groups of patients presenting with dyspepsia, one with duodenal ulcer as the cause and the other with gastric ulcer as the cause. At this point you might like to consider carefully which symptoms might conceivably differentiate between these two conditions.

**Table I**—Broad disease classification (primary diagnosis) of cases of dyspepsia seen at Southern General Hospital, Glasgow.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>No. of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncomplicated oesophageal disease</td>
<td>45 (4)</td>
</tr>
<tr>
<td>Severe oesophageal disease</td>
<td>63 (5)</td>
</tr>
<tr>
<td>Gastric ulcer</td>
<td>74 (6)</td>
</tr>
<tr>
<td>Carcinoma of stomach</td>
<td>32 (3)</td>
</tr>
<tr>
<td>Duodenal ulcer disease</td>
<td>380 (27)</td>
</tr>
<tr>
<td>Cholelithiasis</td>
<td>50 (4)</td>
</tr>
<tr>
<td>Irritable bowel syndrome</td>
<td>177 (15)</td>
</tr>
<tr>
<td>Non-organic dyspepsia</td>
<td>294 (25)</td>
</tr>
<tr>
<td>Alcohol related dyspepsia</td>
<td>48 (4)</td>
</tr>
<tr>
<td>Organic bowel disease, undiagnosed</td>
<td>87 (7)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1200 (100)</strong></td>
</tr>
</tbody>
</table>

Ratios, relative risk, and screening tests

As encouragement one immediate suggestion is that it is well known for duodenal ulcers to be commoner in men and for gastric ulcers to be commoner in women. Perhaps you would agree that men make up some 70% of the patients who have duodenal ulcer...
compared with only 35% of the patients who have gastric ulcer? A fairly obvious way of expressing the weight of this particular "symptom"—that is, the sex of the patient—is to divide 70 by 35 to give a ratio of 2. This means that a patient who has a duodenal ulcer is twice as likely to be a man than a patient who has a gastric ulcer. The ratio is also equivalent to the epidemiological idea of relative risk, and this natural ratio is also closely related to the well known concepts of sensitivity and specificity, which are normally applied to screening tests; the ratio is simply: sensitivity/(1—specificity). This ratio is, of course, exactly the same as the likelihood ratio mentioned by Macartney.1

One way of thinking about the clinical consultation is to regard every question or investigation as a kind of screening test. Thus the apparent dissimilarity between the clinical consultation on the one hand and the assessment of screening tests on the other is not perhaps quite as great as it might at first seem.

The example taken below includes about 75% of our patients who had duodenal ulcer, excluding those who had multiple diagnoses or an uncertain diagnosis, and about 60% of our patients who had gastric ulcer, selected in the same way.

The actual figures from the study (table II) show that there was a preponderance of men among the cases of duodenal ulcer—namely, 169 of 248 (68%)—compared with 17 of 43 cases of gastric ulcer (40%). If the dyspepsia sufferer is a man duodenal ulcer is therefore more likely, and this would naturally be expected to give a positive weight towards the diagnosis of duodenal ulcer. Using the example of the simple ratio given above we find that the actual ratio is 68/40, or 1.7.

Turning ratios into scores and weights

To turn the ratio into a simple score that we can add up we now have to take the logarithm of this ratio (technically a likelihood ratio), because logarithms can be added rather than needing to be multiplied, thus simplifying the process. In this case the natural logarithm of the ratio is 0.531. Again to make it simple to use, this value is multiplied by 10 and rounded to allow us to quote some whole numbers as weights; in this case the weight to be applied in favour of duodenal ulcers for patients who are men is 10×0.531=5.

If, however, the patient in our clinic is a woman then a different weight has to be calculated. In this case the ratio is 32%/60%=0.53, for which the logarithm is negative (−0.629). Multiplying by 10 gives us a score of −6.

The weights seem to be reasonably natural, in the sense that if the patient is a man we add some five points towards a diagnosis of duodenal ulcer and if the patient is a woman then we subtract six from the score for duodenal ulcer.

There are two negative scores (−6 for being a woman and −4 for food aggravation) that total −10 and three positive scores (+10 for age, +3 for no daily pain, and +17 for prior probability) that add up to +30, giving a net score of +20. Given that she has some kind of ulcer, this converts to a probability of this woman having a duodenal ulcer of 88% compared with her chances of having a gastric ulcer of 12%.

To take another example, if we had a woman aged 65 who had daily pain and symptoms suggesting that food made the pain worse then she would have a total score of −32, which, when combined with the prior probability score of +17, gives a total of −15, in turn converting to a probability of a duodenal ulcer of 18% and a much higher probability of a gastric ulcer of 82%.

Turning scores back to probability of disease

The conversion of a patient’s cumulative score to probability is straightforward. Remember to divide the score by 10 to reverse the previous step of multiplying the logarithm by 10. To take the first example above:

Cumulative score (T) =30+(−10)=20.
Probability of duodenal ulcer=1−0.88=0.12.

This can be done more quickly by using a simple graph.29
(5) LOGISTIC DISCRIMINATION TO ADJUST WEIGHTS FOR DEPENDENCE BETWEEN SYMPTOMS

Effect of adjustment

The fifth step required by this method reflects a well known statistical problem with Bayes's theorem: the assumption, in its basic form, of independence of symptoms within a class of diseases. For example, in a patient with jaundice it is clear that in biliary obstruction the symptom of passing dark urine is highly correlated with (or dependent on) another symptom—namely, passing pale stools. Thus these two symptoms are clearly not independent, though on occasions when only one of the two occurs without the other, some diagnostic information might be obtained. The uncritical use of simple Bayes's theorem, which some authors have been known to describe as "idiot's Bayes," particularly in its blunderbuss form, leads to considerable problems in overestimating the probabilities of disease. This happens if many symptoms that are highly dependent on each other are included.

It is therefore necessary to find statistical ways of making an adjustment, and this can be done by several methods, the most appropriate of which seems to be logistic discrimination. (An extensive example of its application to dyspepsia is available.)

What this does is to reduce the weights for the symptoms for those indicators that are dependent on, or correlated with, other indicators, thereby avoiding the resulting overestimation of probabilities. The effect of this is also shown in table II, which compares the crude weightings obtained earlier with the slightly reduced weightings obtained after adjustment by logistic discrimination. For this particular example the effects on the probabilities of disease for the two patients given in an earlier section are not very great. The first example had a combined score of 20 when crude weights were used, and this becomes 18 when adjusted weights are used, changing the probability of duodenal ulcer from 88% to 86%. In the second example the probability of duodenal ulcer changes from 18% to 21%.

You should not go away with the idea that adjustments to crude weights inevitably lead to only small changes in the calculated probability of disease. The effect can be very much greater for particular diagnostic problems, and if there is no adjustment completely misleading probabilities can be produced. Standard errors can also be calculated for both crude and adjusted weightings and are shown for the adjusted weights in table II.

Which is the "best" symptom?

An inspection of the crude weights given in table II also allows us to compare the symptoms. For example, the range of crude weights for sex is from -6 to +5, a range of 11. On the other hand, the range of weights for age, split in the way that it has been done, is from -10 to +18, a range of 28. The bigger the range of an indicant, the greater its potential value as a diagnostic item. The symptom conveying the least amount of evidence is the one derived by asking about the effect of food. The range there is only 8, considerably less than for the other symptoms shown.

Further analysis along these lines allows us to identify the most important symptoms for particular clinical problems, the selection of symptoms being something we should consider when teaching students. Many years ago Fletcher asked us to stop teaching and using unreliable signs, his advice coming from a rigorous assessment of variation among observers. Perhaps others should now be left out—namely, those that convey little information when assessed in the way suggested above.

You might, however, see the limitations in regarding the importance of symptoms simply in terms of the range of weights that they provide. Clearly, it is possible to have a very powerful symptom that is so rarely present that it would not normally be ascertained in a routine clinical consultation. An example is given by Willfrid Card, who once had a patient with steatorrhoea who presented because the patient had had to call the plumber to unblock the drain (as a result of the massive steatorrhoea). Clearly, if a patient gives this "symptom"—calling the plumber—then steatorrhoea is highly likely. Nobody, however, would consider it reasonable to ask patients in a general gastrointestinal clinic a direct question about whether they have had to call the plumber to unblock their lavatories.

What has been provided above are the weights to be applied to the relevant responses to a particular question.

Which is the best question?

The final step is then to consider the relative value of asking each question with respect to discovering whether a patient has the disease of interest. This is clearly related to two things—firstly, the weights of the various responses to that question as derived above, and, secondly, some measure of how common that symptom is in the disease of interest. (This gets round the problem posed by Card's symptom.) It is a slightly more complex subject than the derivation of crude weights, and you may wish to skip the next two paragraphs.

Calculating the relative value of asking each question is in fact a calculation of "the expected weight of evidence" from that question. This gives a single value for each question and can be expressed by multiplying the probability that a particular response to a question will occur by the corresponding weight and adding in the same for each of the other possible responses to the same question. For example, the symptom of daily pain occurs in 21/235 cases of duodenal ulcer (table II). From adjusted weights (table II) the expected weight of evidence:

\[
\text{expected weight of evidence} = \left( \frac{21}{235} \times -11 \right) + \left( \frac{214}{235} \times 2 \right)
\]

\[
= -0.96 + 1.82
\]

\[
= 0.84 \text{ (always a positive value).}
\]

The sex of the patient gives a value of 0·1, and for the patient's age the expected weight of evidence is 4·0, much greater in fact than the other symptoms in the table. As it happens, age is a more useful piece of information to ask about than the patient's sex—which is obvious anyway.

In principle therefore the questions with the highest expected weight of evidence are clearly those that should be taught to medical students in the hope that they might be able to remember more easily those that are discriminating. Using these as signposts would help to guide them through the uncertainty that besets us all at some time or another when faced with a plethora of information, most of which has somehow to be discarded. These are the searchlights alluded to earlier, which, if taught to students, can help us all.

Conclusion

It is probably unwise to speculate too much about future developments. Provided that the use of such systems is seen as a diagnostic aid, not a decision maker, and as a screening test or guide, not an automatic initiator of treatment or of further investigation, then these ideas will probably find a place in providing more efficient medical care for certain problems.

I acknowledge the financial support of the Scottish Home and Health Department and more recently of Smith, Kline and French. I am deeply indebted to my colleagues G P Crean, R J Holden, and D J Spiegelhalter for their contributions but most of all to the late W I Card, who pioneered the work described here.

References

How To Do It

Make a video tape

PETER CULL

Video recordings have become a popular and accepted tool of record, analysis, and communication in medicine and they vary from the relatively simple, unedited demonstration of the clinical examination of a patient—which to make may need just one camera and single operator—to a full scale teaching production involving many staff, studio and location work, multiple cameras, complex editing with insertion of graphics, film and animation sequences, overdubbed narration, and music. At each point in this spectrum, the technical and creative requirements are very different and so in this short article it is possible only to consider some general points. For those who, having read it and pondered the implications, wish to proceed further there are some excellent reference works available and these are listed below.

Improvements in video technology have taken place at a remarkable pace in recent years, and while quality and versatility have risen costs have come down. Furthermore, cameras and recorders are now so simple to operate that no amateurs need be deterred from using this excellent medium, providing, that is, that they recognise their own creative limitations and that of their equipment, and are prepared to invest the necessary time and effort.

The technical, production, and creative elements of video production are inseparable so I shall discuss them together.

Video formats

There are several video recording formats currently in use; the two most popular in the amateur/professional range are VHS and U-Matic, though others are fast developing. Both of these are cassette tape systems; VHS, using 1/2" recording tape, is the cheaper whereas U-Matic, which uses 3/4" tape, is mainly used by professionals. Generally speaking the wider the tape the higher the quality of recording and this is particularly important when tapes need to be edited or copied as each rerecording results in a loss of quality.

From How To Do It 2, a new collection of useful advice on topics that doctors need to know about but won't find in the medical textbooks. Just published, price £6.95, this is a companion volume to the popular How To Do It: 1, also published by the BMJ.

It is perhaps worth mentioning here that in addition to various formats there are also three different television broadcast standards in the world. In Britain and most European countries PAL standard is used but both France and the USA among others use different standards. As a result a video tape made in Britain cannot be replayed in a country operating on a different standard unless the relevant equipment is available or the tape is copied in the appropriate standard.

Recorders

Irrespective of the format employed, the video recorder is the key element in the equipment, and it is worth spending a bit more to gain reliability and obtain the best results. It is advisable to choose a professional or industrial recorder with an "Edit" facility which ensures that recordings begin with a clean cut picture rather than a broken image that takes time to settle down (the use of editing in compiling a programme will be discussed later). Battery driven portable recorders with built in playback facilities are available in both formats and some recorders are now actually incorporated in the camera body itself—the so called Cam-corders. Portable equipment is particularly useful for a one man operation and in locations remote from a power source, or where trailing power, camera, and sound cables are a nuisance or dangerous. It is worth bearing in mind that when making recordings in sensitive situations—for example, clinical interviews and examinations, performance evaluation, role play, etc—the less obtrusive the equipment the less alarming or distracting it is for the participant, and a better, less selfconscious performance will be obtained.