New approaches to enhance the accuracy of the diagnosis of reflux disease

P Moayyedi, J Duffy, B Delaney

The accuracy of symptoms in diagnosing gastro-oesophageal reflux disease (GORD) is complicated by the lack of a gold standard test. Statistical techniques such as latent class and Bayesian analyses can estimate accuracy of symptoms without a gold standard. Both techniques require three independent diagnostic tests. Latent class analysis makes no assumptions about the performance of the tests. Bayesian analysis is useful when the accuracy of the other tests is known. These statistical techniques should be used in the future to validate GORD symptom questionnaires comparing them with endoscopy, oesophageal pH monitoring, and response to proton pump inhibitor therapy. Studies that evaluate GORD symptoms are usually done in secondary care. The prevalence of GORD in primary care will be lower and this reduces the positive predictive value of symptoms. There will be some bias in the type of patient referred for diagnosis and this usually decreases the specificity of symptom diagnosis.

SUMMARY
Symptoms such as predominant heartburn and regurgitation are believed to be important in the diagnosis of gastro-oesophageal reflux disease (GORD). These symptoms can be regarded as a diagnostic test for the presence of GORD and have been the basis for recruitment of patients into some trials and also for epidemiological surveys of the prevalence of reflux disease. The diagnostic accuracy of reflux symptoms is important for the interpretation of these studies. This article will discuss issues relevant to the assessment of the efficacy of symptoms in diagnosing GORD.

Expressions of diagnostic accuracy
The accuracy of a diagnostic test can be expressed as either sensitivity and specificity, or as positive and negative predictive values. The methods for calculating these indices are given in table 1. Sensitivity is asking: “if a patient has the disease, what is the probability that the test will be positive?” whereas specificity is asking: “if the patient does not have the disease, what is the probability of the test being negative”? What the clinician wants to know however is: “if the test is positive, what is the probability of the patient having the disease?” or “if the test is negative, what is the probability of the patient not having the disease”? These questions are answered by the positive and negative predictive values respectively. These diagnostic indices are unified by the use of positive and negative likelihood ratios that are derived from the sensitivity and specificity of the test (table 1) and can be used to calculate positive and negative predictive values if the prevalence of disease is known.2

DIAGNOSTIC ACCURACY IN DIFFERENT POPULATIONS
Influence of prevalence of disease on diagnostic accuracy
Prevalence of disease varies widely in different settings. Usually a disease is rare in the general population and becomes more common as patients move from primary to secondary and tertiary care settings. Positive and negative predictive values are the most clinically relevant diagnostic indices, but these vary with the prevalence of disease. Positive predictive value decreases as the prevalence of disease falls and negative predictive value decreases as the prevalence of disease rises in the population. In the hypothetical exam of reflux symptoms having a sensitivity and specificity of 80%, the positive predictive value varies from 94% in populations with an 80% prevalence of GORD to 50% in populations with a 20% prevalence of GORD (fig 1). Reflux symptoms might, therefore, be expected to have higher positive predictive values in secondary care than primary care, and, conversely, absence of these symptoms will have a higher negative predictive value in the general population compared with the hospital setting.

It is important to note that the sensitivity and specificity of reflux symptoms will not change whatever the prevalence of disease, provided there is no bias in referral and evaluation of GORD patients. In theory, therefore, the sensitivity and specificity of reflux symptoms derived from an unbiased secondary care study can be applied to primary care or the general population. The positive and negative predictive values will vary in these populations, but these can be calculated using likelihood ratios if the prevalence of GORD is known in these settings.

Influence of spectrum and selection bias
The sensitivity and specificity of reflux symptoms can be applied to different settings whatever the prevalence of GORD, provided there is no bias in recruiting patients to the study. Bias is, however, a problem with many epidemiological and diagnostic accuracy studies. There is a long list of biases that can influence the results of a

Abbreviations: GORD, gastro-oesophageal reflux disease; LCA, latent class analysis

See end of article for authors’ affiliations

Correspondence to:
Professor P Moayyedi,
Gastroenterology Division,
McMaster University-HSC
4W8, 1200 Main Street
West, Hamilton, Ontario,
Canada L8N 3Z5;
evansl@mcmaster.ca


www.gutjnl.com

on September 21, 2023 by guest. Protected by copyright.
diagnostic accuracy study, but the most important of these are spectrum and selection bias.

Spectrum bias occurs when a diagnostic test is validated in one population and then applied to another with a different clinical spectrum of disease. For example, the accuracy of reflux symptom assessment in secondary care may be greater than that seen in primary care or in the general population, as patients that are referred to hospital tend to have more severe symptoms. It is usually easier to discriminate between “disease” and “no disease” if patients have more extreme manifestations of the disorder. This is distinct from variations in prevalence of disease. The absolute proportion of patients with GORD is not the issue—it is the distribution of mild versus severe GORD in patients with the disease. If patients with more severe symptoms are referred to secondary care and the diagnostic accuracy of symptom assessment evaluated in this setting, this will increase the sensitivity and reduce the specificity of reflux symptoms in the diagnosis of GORD. The positive and negative likelihood ratios will also both be reduced. The impact of spectrum bias on sensitivity and specificity is not usually as notable as the effect of prevalence on positive and negative predictive values, except in extreme circumstances. For example, if the diagnostic utility of heartburn is evaluated in patients with only typical and clearly definable symptoms, this might have a hypothetical sensitivity of 74% and specificity of 90%. If heartburn was evaluated in a group of patients with less clearly defined symptoms, it is estimated that the sensitivity would increase to 85% and the specificity fall to 83%.

Selection bias occurs when there is an association between the test result and the probability of being included in the study that is validating the test. For example, if primary care doctors were only interested in diagnosing GORD in patients presenting with upper abdominal/retrosternal symptoms and referred patients on the basis of Rome II criteria, then most of the study population would have predominant heartburn. This is subtly different from spectrum bias, as there is a conscious decision to refer patients with a specific symptom rather than the phenomenon that the spectrum of disease is usually more severe in secondary compared with primary care. If there is marked selection of patients on the basis of predominant heartburn, this will slightly increase the sensitivity but will dramatically reduce the specificity of this symptom in diagnosing GORD. For example, the specificity of right lower quadrant tenderness in the diagnosis of acute appendicitis fell from 89% in primary care to 16% in tertiary care. Thus, there is the danger that the diagnostic value of a reflux symptom may be “used up” along the referral pathway.

It is important to be aware of the influence of spectrum and selection bias on the accuracy of reflux symptoms in GORD, but this must be kept in perspective. It is perfectly acceptable to evaluate reflux symptoms in selected populations to assess their diagnostic accuracy. Patients willing to undergo endoscopy and pH monitoring are likely to be different from the general population but, pragmatically, there is little that can be done to correct this. It is better to carefully assess the sensitivity and specificity of reflux symptoms in the diagnosis of GORD in a selected group than not to do this. The results can be applied to a more general population, although the sensitivity may increase and specificity may decrease in this setting.

Further studies of the utility of reflux symptom assessment in terms of patient outcomes (for example, response to proton pump inhibitor treatment) can then be conducted. Finally, it is necessary to be aware that even if symptoms successfully identify GORD, some patients will still be referred to secondary care. If the referral process is heavily based on reflux symptoms, the specificity will fall dramatically. This is not necessarily a bad thing if the diagnostic value of reflux symptoms has been maximised. If hospital clinicians are only referred patients at the point at which symptoms are not useful predictors of disease, this is an appropriate filtration because it channels patients who have need for more invasive tests only available at referral centres, such as endoscopy and 24 hour pH monitoring.

**THE “GOLD STANDARD” PROBLEM**

So far we have assumed that we can compare reflux symptoms with a “gold standard” test that is 100% sensitive and specific. This type of test rarely exists in clinical practice, but often there is a single test that is sufficiently accurate to serve as a reference standard. Even this is not available for the diagnosis of GORD, as endoscopy and pH monitoring are not sufficiently sensitive and specific for use as a reference standard. The lack of at least a reference test has bedevilled the assessment of the accuracy of reflux symptoms to diagnose GORD, yet this is not a unique problem. Psychiatrists make diagnoses without the benefit of laboratory, x ray, or pathology reports. The lack of any reference standard in psychiatry has been overcome by using techniques that avoid the need for comparison with a single accurate test. These techniques can be broadly divided into latent class analysis (LCA) and Bayesian analysis. Both have been used widely in psychiatry as well as other disciplines, but, as yet, have not been applied to the evaluation of the accuracy of symptom assessment in GORD.

**Latent class analysis**

Traditional regression techniques describe relations between observed variables. For example, a logistic regression model may suggest a relation between smoking and lung cancer that is independent of other variables in the model, such as alcohol intake, sex, or social class. Any variation of the data in this model that is not explained by these observed variables is assumed to occur at random. LCA postulates the existence of an unobserved categorical variable that divides the population of interest into classes (hence the term “latent class”). Members of the population with a set of observed

---

**Table 1** Calculation of different diagnostic indices

<table>
<thead>
<tr>
<th>Test result</th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>A</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Absent</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
</tbody>
</table>

Sensitivity = a/(a+c); Specificity = d/(b+d); Positive predictive value = a/(a+b); Negative predictive value = d/(c+d); Positive likelihood ratio = sensitivity/(1−specificity); Negative likelihood ratio = (1−sensitivity)/specificity.

**Figure 1** Variation in positive predictive value (PPV) and negative predictive value (NPV) with prevalence of GORD for reflux symptoms with a hypothetical sensitivity and specificity of 80%.
variables will respond differently depending on the latent class to which they belong. This technique can be applied to the problems related to diagnostic testing, with the unobserved categorical variable being “disease present” or “disease absent”. The observed variables might typically be the results of diagnostic tests, none of them being a gold standard. LCA could then be applied in an attempt to divide the population into “true” positives and negatives. This approach has been shown to require at least three different types of diagnostic test. LCA can then be applied to derive the population into “true” positives and negatives. This technique can be applied to the diagnosis of reflux disease.

LCA assumes that association between the results of diagnostic tests arise solely from the disease status of the individuals. It is, therefore, important that there should be no other form of dependence between variables entered into an LCA model. An example of such dependence would be when one test relates to the presence or absence of a particular symptom, and another to its severity or frequency. These tests would then be unsuitable to be entered together into LCA. This type of analysis has been applied to such diverse problems as the diagnosis of depression, visceral leishmaniasis, Chlamydia trachomatis, and Behcet’s disease, but we are only aware of one application related to gastroenterology. This paper applied LCA to the diagnosis of Helicobacter pylori infection and reported the sensitivity and specificity of histology to be approximately 95%. This result subsequently agreed with results from a gold standard of four other tests.

Bayesian analysis
Standard statistical tests assume that there are no prior expectations of the study results, in adherence with the principle that science is objective. Statisticians analysing data in this way are known as frequentists. This approach has been questioned because, in most scientific experiments, there is an expectation of what the outcome will be from prior knowledge, which should be incorporated in the analysis. An epidemiological study that showed no association between smoking and lung cancer, for example, would not shake the belief of the scientific community that the two are strongly linked. Thomas Bayes proposed a theory over 200 years ago to try and overcome this problem. Bayes’ theorem is a formula that describes how our existing beliefs (described as probability distributions) are changed by new study data. Distributions of beliefs before new information becomes available are known as priors, those after the assimilation of new information are posteriors. Prior distributions can be obtained from existing research evidence, expert opinion, or be set to be “uninformative” (a flat distribution that does not influence the analysis). Posterior distributions are described as means (or proportions) and credible intervals. Credible intervals are the Bayesian equivalent of confidence intervals.

The vast majority of statisticians are frequentists rather than Bayesians as, until recently, the latter required extremely complex calculations. The advent of powerful computers has given Bayesians a new lease of life and consequently software, such as WinBUGS (Bayesian analysis Using Gibbs Sampling, MRC Biostatistics Unit, Cambridge, UK; http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), has been developed to perform this type of analysis. This software uses an iterative “resampling” technique to construct “posterior” distributions. This cycle of sampling is repeated many thousands of times until the values “converge”—that is, stabilise to new estimates based on the new data.

The advantage of Bayesian analysis for diagnostic test evaluation is that no reference standard is required. A relation is formed between the prior and posterior distributions allowing for the imperfect reference standards in the same manner as described previously for LCA. Again, a set of observed data consisting of at least three different, independent, tests is required from the same subjects. The advantage of Bayesian analysis over LCA is that previous knowledge of the accuracy of diagnostic tests for GORD can be Incorporated. A very high specificity (with a relatively low sensitivity) can, therefore, be assigned to endoscopy and also a sensitivity and specificity of 80–90% for pH monitoring. The disadvantage of Bayesian statistics is that the calculations are complex. If little is known about the accuracy of other diagnostic methods, LCA will give similar answers and is more straightforward.

CONCLUSION
There is a great deal of confusion surrounding the accuracy of reflux symptoms in diagnosing GORD. Future studies should apply LCA or Bayesian analysis to overcome the problem of not having a gold standard. This will give a much more realistic estimate of the accuracy of reflux symptoms for the diagnosis of GORD.

REFERENCES