Graphical presentation of diagnostic information

Penny F Whiting*, Jonathan AC Sterne1, Marie E Westwood2, Lucas M Bachmann3, Roger Harbord1, Matthias Egger4 and Jonathan J Deeks5

Address: 1Department of Social Medicine, Canynge Hall, Whiteladies Road, Bristol, BS8 2PR, UK, 2Centre for Reviews and Dissemination, University of York, YO10 5DD, UK, 3Horten Centre, Zürich, Switzerland, 4Department of Social and Preventive Medicine, University of Bern, Switzerland and 5Medical Statistics Group/Diagnostic Research Group, Department of Public Health and Epidemiology, University of Birmingham, B15 2TT, UK

Email: Penny F Whiting* - penny.whiting@bristol.ac.uk; Jonathan AC Sterne - jonathan.sterne@bristol.ac.uk; Marie E Westwood - mew3@york.ac.uk; Lucas M Bachmann - lucas.bachmann@usz.ch; Roger Harbord - roger.harbord@bristol.ac.uk; Matthias Egger - egger@ispm.unibe.ch; Jonathan J Deeks - J.Deeks@bham.ac.uk

* Corresponding author

Abstract

Background: Graphical displays of results allow researchers to summarise and communicate the key findings of their study. Diagnostic information should be presented in an easily interpretable way, which conveys both test characteristics (diagnostic accuracy) and the potential for use in clinical practice (predictive value).

Methods: We discuss the types of graphical display commonly encountered in primary diagnostic accuracy studies and systematic reviews of such studies, and systematically review the use of graphical displays in recent diagnostic primary studies and systematic reviews.

Results: We identified 57 primary studies and 49 systematic reviews. Fifty-six percent of primary studies and 53% of systematic reviews used graphical displays to present results. Dot-plot or box-and-whisker plots were the most commonly used graph in primary studies and were included in 22 (39%) studies. ROC plots were the most common type of plot included in systematic reviews and were included in 22 (45%) reviews. One primary study and five systematic reviews included a probability-modifying plot.

Conclusion: Graphical displays are currently underused in primary diagnostic accuracy studies and systematic reviews of such studies. Diagnostic accuracy studies need to include multiple types of graphic in order to provide both a detailed overview of the results (diagnostic accuracy) and to communicate information that can be used to inform clinical practice (predictive value). Work is required to improve graphical displays, to better communicate the utility of a test in clinical practice and the implications of test results for individual patients.

Background

Readers of a research report evaluating a diagnostic test may wish to assess the test's characteristics (diagnostic accuracy) or evaluate the impact that its use has on diagnostic decisions (predictive value) for individual patients. Graphical displays of results of test accuracy studies allow researchers to summarise and communicate the key findings of their study. We discuss the types of graphical dis-
play commonly encountered in primary diagnostic accuracy studies and systematic reviews of such studies, and systematically review the use of graphical displays in recent diagnostic systematic reviews and primary studies. Table 1 defines the various measures of diagnostic accuracy used.

**Types of graphical display**

**Primary studies**

Figure 1 illustrates four types of graphical display commonly used to present data on diagnostic accuracy for primary diagnostic accuracy studies. We used data from a study of the biochemical tumour marker CA-19-9 antigen to diagnose pancreatic cancer to construct these graphs [1].

**Dot plots (Figure 1a) and Box-and-whisker plots (Figure 1b)**

Dot plots are used for test results that take many values, and display the distribution of results in patients with and without the target condition. Box and whisker plots summarise these distributions: the central box covers the interquartile range with the median indicated by the line within the box. The whiskers extend either to the minimum and maximum values or to the most extreme values within 1.5 interquartile ranges of the quartiles, in which case more extreme values are plotted individually [2]. Sometimes an indication of the threshold used to define a positive test result is included, for example by adding a horizontal line or shading at the relevant point. Such plots can be used to clearly summarise a large volume of data, but are only able to display differences in the distribution of test values between patients with and without the target condition; they do not directly display the diagnostic performance of the test.

Although the CA-19-9 antigen test to diagnose pancreatic cancer (used to construct Figure 1) is an example of continuous data, it is also possible to construct similar graphs for categorical test results providing that the number of categories is reasonably large. Alternatively, for smaller numbers of categories, similar information can be conveyed using paired bar charts/histograms. Paired histograms show the distribution of test results in patients with the target condition above the x-axis and the distribution in patients without the target condition below the x-axis. These types of graphical display are less commonly used.

**Table 1: Definitions of measures of diagnostic accuracy**

<table>
<thead>
<tr>
<th>Test result</th>
<th>Target condition</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Likelihood ratio (LR)</th>
<th>Diagnostic odds ratio (DOR)</th>
<th>Predictive value</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>Present</td>
<td>a/(a + c)</td>
<td>d/(b + d)</td>
<td>LR +ve = sensitivity/(1 - specificity)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>Absent</td>
<td>b/(b + d)</td>
<td>(1 - sensitivity)/specificity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Present</td>
<td>c/(a + c)</td>
<td>ad/bc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>Absent</td>
<td>d/(c + d)</td>
<td>DOR = [a/(a + c)]/[(d/(b + d)]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sensitivity**

- Proportion of true positives that are correctly identified by the test [31]

**Specificity**

- Proportion of true negatives that are correctly identified by the test

**Likelihood ratio (LR)**

- Describes how many times a person with disease is more likely to receive a particular test result than a person without disease [32]

**Diagnostic odds ratio (DOR)**

- Used as an overall (single indicator) measure of the diagnostic accuracy of a diagnostic test. It is calculated as the odds of positivity among diseased persons, divided by the odds of positivity among non-diseased. When a test provides no diagnostic evidence then the DOR is 1.0. [33] This measure has a number of limitations: by combining sensitivity and specificity into a single indicator the relative values of the two are lost i.e. the DOR can be the same for a very high sensitivity and low specificity as for very high specificity and low sensitivity [33] Further, tests that are effective for classifying persons as having or not having the target condition have DORs that whose magnitude is much greater (e.g. 100) than usually considered as indicating strong associations in epidemiological studies. [34]

**Predictive value**

- Positive predictive value (PPV) = a/(a + b)
- Negative predictive value (NPV) = d/(c + d)

Positive predictive values depend on disease prevalence, the more common a disease is, the more likely it is that a positive test result is right and a negative result is wrong. [35]
It is not possible to construct any of these graphs for truly dichotomous test results. However, truly dichotomous tests rarely occur in practice. Examples of dichotomous tests include dipstick tests that change colour if the target condition is said to be present (although these are based on an underlying implicit threshold) or the presence/absence of certain clinical symptoms.

**Example graphical displays for primary study data.**

- **a.** Dot plot.  
- **b.** Box-and-whisker plot.  
- **c.** ROC Plot.  
- **d.** Flow diagram.

**Receiver operating characteristic (ROC) plot (Figure 1c)**

ROC plots show values of sensitivity and specificity at all of the possible thresholds that could be used to define a positive test result [3]. Typically, sensitivity (true positive rate) is plotted against 1-specificity (false positive rate): each point represents a different threshold in the same group of patients. Stepped lines are used for continuous
test results while sloping lines are used for ordered categories. ROC curves may be derived directly from the observed test sensitivity and specificity corresponding to different test thresholds, or by fitting curves based on parametric [4], semi-parametric [5,6], or non-parametric methods [7]. The area under the ROC curve (AUC) is a summary of diagnostic performance, and takes values between 0.5 and 1. The more accurate the test, the more closely the curve approaches the top left hand corner of the graph (AUC = 1). A test that provides no diagnostic information (AUC = 0.5) will produce a straight line from the bottom left to the top right. ROC curves may be restricted to a range of sensitivities or specificities of clinical interest.

ROC plots show how estimated sensitivity and specificity vary according to the threshold chosen, and can be used to identify suitable thresholds for clinical practice if the points on the curve are labelled with the corresponding threshold as in Figure 1c, which shows for example that the sensitivity and specificity corresponding to a threshold of 39.3 are 74% and 90%, respectively. Confidence intervals can be added to indicate the uncertainty in estimates of test performance at each point. ROC plots also allow comparison of the performance of several tests independently of choice of threshold, by plotting data sets for multiple tests in the same ROC space. However, they are thought to be difficult to interpret as they describe the characteristics of the test in a way which does not relate directly to its usefulness in clinical practice; research has shown that ROC plots are generally poorly understood by clinicians [8].

Flow charts (Figure 1d)
These depict the flow of patients through the study: for example how many patients were eligible, how many entered the study, how many of these had the target condition, and the numbers testing positive and negative. Such charts require categorisation of test results, for example as "positive" and "negative". Although flow charts do not directly present diagnostic accuracy data, addition of percentages to the test result boxes (as in Figure 1d) can be used to report test sensitivity (68/90 = 76%) and specificity (46/51 = 90%). Charts that first separate individuals according to test result before classification by disease status may similarly be used to depict positive and negative predictive values. The STARD (standards for reporting of diagnostic accuracy) statement, an initiative to improve the reporting of diagnostic test accuracy studies similar to the CONSORT statement for clinical trials, recommends the inclusion of a flow diagram in all reports of primary diagnostic accuracy studies [9]. This should illustrate the design of the study and provide information on the numbers of participants at each stage of the study as well as the results of the study. The example flow chart in Figure 1d is not a full STARD flow diagram as we do not have data on numbers of withdrawals or uninterpretable results from this study. It does, however, show the design (diagnostic case-control) and results of the study.

Systematic reviews
Figure 2 illustrates two graphical displays commonly used to present data on diagnostic accuracy in diagnostic systematic reviews. Data from a systematic review of dipstick tests for urinary nitrite and leukocyte esterase to diagnose urinary tract infections were used to construct these graphs [10].

Forest plots (Figure 2a)
Forest plots are commonly used to display results of meta-analysis. They display results from the individual studies together with, optionally, a summary (pooled) estimate. Point estimates are shown as dots or squares (sometimes sized according to precision or sample size) and confidence intervals as horizontal lines [11]. The pooled estimate is displayed as a diamond whose centre represents the estimate and tips the confidence interval.

For diagnostic accuracy studies, measures of test performance (sensitivity, specificity, predictive values, likelihood ratios or diagnostic odds ratio) are plotted on the horizontal axis. Diagnostic test performance is often described by pairs of summary statistics (e.g. sensitivity and specificity; positive and negative likelihood ratios), and these are depicted side-by-side. Between-study heterogeneity can readily be assessed by visual examination. Results may be sorted by one of a pair of test performance measures, usually that which is most important to the clinical application of the test. A disadvantage of paired forest plots is that they do not directly display the inverse association between the two measures that commonly results from variations in threshold between studies.

ROC plots and summary ROC (SROC) curves (Figure 2b)
ROC plots can be used to present the results of diagnostic systematic reviews, but differ from those used in primary studies as each point typically represents a separate study or data set within a study (individual studies may contribute more than one point). A summary ROC (SROC) curve can be estimated using one of several methods [12-15] and quantifies test accuracy and the association between sensitivity and specificity based on differences between studies. As with forest plots, ROC plots provide an overview of the results of all included studies. However, unless there are very few studies, it is not feasible to display confidence intervals as the plot would become cluttered. Results for several tests can be displayed on the same plot, facilitating test comparisons. It is also possible to display pooled estimates of sensitivity and specificity together with associated confidence intervals or prediction regions. ROC plots may also be used to investigate possible expla-
nations for differences in estimates of accuracy between studies, for example those arising from differences in study quality. Figure 3 shows results for a recent review that we conducted on the accuracy of magnetic resonance imaging (MRI) for the diagnosis of multiple sclerosis (MS) [16]. By using different symbols to illustrate studies that did (diagnostic cohort studies) and did not (other study designs) include an appropriate patient spectrum we were able to show that studies that included an inappropriate patient spectrum grossly overestimated both sensitivity and specificity.

Other plots
Various other graphical methods have been developed to display the results of systematic reviews and meta-analyses [17,18]. Although not generally developed specifically for diagnostic test reviews these can be adapted to display the results of such reviews. Funnel plots [19] and Galbraith plots [20] are often used to assess evidence for publication bias or small study effects in systematic reviews of the effects of medical interventions assessed in randomized controlled trials. However, their application to systematic reviews of diagnostic test accuracy studies is
problematic [20]. Diagnostic odds ratios are typically far from 1, and it has been shown that, for data of this type, sampling variation can lead to artefactual associations between log odds ratios and their standard errors [21]. It is therefore recommended that the effective sample size funnel plot be used in reviews of test accuracy studies [20].

Predictive value

A number of graphical displays aim to put results of diagnostic test evaluations into clinical context, based either on primary studies or systematic reviews. Two graphical displays commonly used for this purpose are the likelihood ratio nomogram (Figure 4a) and the probability-modifying plot (Figure 4b). Each allows the reader to estimate the post-test probability of the target condition in an individual patient, based on a selected pre-test probability. To use the likelihood ratio nomogram, the reader needs an estimate of the likelihood ratios for the test. He then draws a line through the appropriate likelihood ratio on the central axis, intersecting the selected pre-test probability, to derive the post-test probability of disease. The probability-modifying plot depicts separate curves for positive and negative test results. The reader draws a vertical line from the selected pre-test probability to the appropriate likelihood ratio line and then reads the post-test probability off the vertical scale. Both graph types are based on a single estimate of test accuracy (likelihood ratio), although it is possible to plot separate curves on the probability-modifying plot or lines on the nomogram to depict confidence intervals around the estimated likelihood ratios. Each assumes constant likelihood ratios across the range of pre-test probabilities. However, this assumption may be violated in practice [22], because populations in which the test is used may have different spectrums of disease to those in which estimates of test accuracy were derived.

Figure 3
Sensitivity plotted against specificity, separately for cohort studies and for studies of other designs for MRI for diagnosis of multiple sclerosis.

Figure 4
Example graphs for interpreting diagnostic study result. a. Likelihood ratio nomogram. b. Probability modifying plot.
Use of graphical displays in the literature

Methods

We systematically reviewed how graphical displays are currently incorporated in studies of test performance. We included primary diagnostic accuracy studies published in 2004, identified by hand searching 12 journals (Table 2), and diagnostic systematic reviews published in 2003, identified from DARE (Database of Abstracts of Reviews of Effects) [23]. Searches were conducted in 2005 and so these years were the most complete available years for searching (there is a delay in adding studies to DARE). Diagnostic accuracy studies were studies that provided data on the sensitivity and specificity of a diagnostic test and that focused on diagnostic (whether the patient had the condition of interest) rather than prognostic (disease severity/risk prediction) questions. Journals were selected to provide a mixture of the major general medical and specialty journals. We particularly aimed to select journals that clinicians read. We extracted data on the different graphical displays used to summarise information about test performance, defined as any graphical method of summarising data on diagnostic accuracy or the predictive value of a test (Table 1).

We located 56 primary studies and 49 systematic reviews (Web Appendix). Fifty-seven percent of primary studies and 53% of systematic reviews used graphical displays to present results. In publications using graphics, the number of graphs per publication ranged from 1 to 51 (median 2, IQR 1 to 3 for primary studies and median 4, IQR 2 to 7 for systematic reviews). Table 3 summarises the categories of tests evaluated in the primary studies and systematic reviews. None of the tests evaluated in any of the primary studies were truly dichotomous: they all gave continuous or categorical results. Three of the eight systematic reviews that assessed clinical examination looked at whether a variety of signs or symptoms were present or absent: these can be considered as truly dichotomous tests. All other reviews evaluated continuous or categorical tests.

Primary studies

Dot-plots or box-and-whisker plots were the most commonly used graphic and were included in 22 (39%) studies. Generally the plots showed individual test results separately for patients with and without the target condition, with four including an indication of the threshold used to define a positive test result. Three studies included both a dot plot and a box-and-whisker plot on the same figure. Other variations included separate plots for different patient subgroups, different symbols to indicate different stages of disease, or separate plots for different tests. The majority of studies using these types of plots were of laboratory tests. An ROC curve was displayed in 15 (26%) studies. All of these plotted full ROC curves; only two provided any indication of the thresholds corresponding to one or more of the points. Thirteen studies included separate ROC curves for different tests, either on the same plot (10 studies) or on separate plots (3 studies). Five studies included separate ROC plots for different patient subgroups. Although all the primary studies were published in 2004, after the publication of the STARD guidelines, only one included a STARD flow diagram.

Systematic reviews

ROC plots were included in 22 (45%) reviews. Twenty showed individual study estimates of sensitivity and specificity, 14 fitted SROC curves, and two displayed a summary point. One study, which did not fit an SROC curve, added a box and whisker plot to each axis to show the distributions of sensitivity and specificity. One study plotted only summary estimates of sensitivity and specificity in ROC space, with no SROC curves. Some reviews included separate plots for different tests, for different patient subgroups, or for different thresholds used to define a positive test result.

Table 2: Number of primary studies identified from the journals searched together with the number of studies from each journal that included graphical displays

<table>
<thead>
<tr>
<th>Journal</th>
<th>Number of studies</th>
<th>Number with graphs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Chemistry</td>
<td>25</td>
<td>18 (72)</td>
</tr>
<tr>
<td>American Journal of Obstetrics and Gynecology</td>
<td>1</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Annals of Internal Medicine</td>
<td>6</td>
<td>3 (50)</td>
</tr>
<tr>
<td>BMJ</td>
<td>3</td>
<td>0 (0)</td>
</tr>
<tr>
<td>European Journal of Pediatrics</td>
<td>1</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Gastroenterology</td>
<td>7</td>
<td>4 (57)</td>
</tr>
<tr>
<td>JAMA</td>
<td>5</td>
<td>2 (40)</td>
</tr>
<tr>
<td>British Journal of Radiology</td>
<td>1</td>
<td>1 (100)</td>
</tr>
<tr>
<td>Lancet</td>
<td>3</td>
<td>2 (67)</td>
</tr>
<tr>
<td>New England Journal of medicine</td>
<td>3</td>
<td>2 (67)</td>
</tr>
<tr>
<td>Thorax</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
The inclusion of some form of graphical presentation of test accuracy data has a number of advantages compared to not using such displays. It allows fuller reporting of results, for example (S)ROC plots can display results for multiple thresholds whereas reporting test accuracy results in a text or table generally requires the selection of one or more thresholds. In addition, (S)ROC plots depict the trade-off between sensitivity and specificity at different thresholds. Use of such displays also have the advantage of presenting all of the results of a primary study or systematic review without the need for selected analyses, which may be biased depending on the analyses selected. The inclusion of graphical displays, such as SROC plots or forest plots, in systematic reviews of test accuracy studies allows a visual assessment of heterogeneity between studies by showing the results from each individual study included in the review. There is also a suggestion that graphical displays may be easier to interpret than text or tabular summaries of the same data.

Diagnostic accuracy studies will usually need to include more than one graphic in order both to provide a detailed description of results (diagnostic accuracy) and to communicate appropriate summary measures that can be used to inform clinical practice (predictive value); the more detailed graphic provides context for the interpretation of summary measures. Further work is required to improve on existing graphical displays. The starting point for this should be further evaluation of the types of graphical display most helpful to assessing the utility of a test in clinical practice and the implications of test results for individual patients.

We hope that this paper will contribute to an increase in the use and quality of graphical displays in diagnostic accuracy studies and systematic reviews of these studies. To achieve this, journal guidelines and the STARD statement need to encourage the use of graphs in reports of test accuracy. Currently, journal guidelines say very little about this issue. A brief review of the instructions for authors from a selection of leading medical journals (Annals of Internal Medicine, BMJ, Clinical Chemistry, JAMA, Lancet, New England Journal of Medicine) found that these only provide formatting guidelines rather than discussing when and what type of graphical displays should be used, although all except the New England Journal of Medicine recommend that the STARD guidelines be followed and

---

**Discussion**

Research in the area of cognitive psychology suggests that sensitivity and specificity are generally poorly understood by doctors [8,24] and are often confused with predictive values [8,25,26]. Doctors tend to overestimate the impact of a positive test result on the probability of disease [27,28] and this overestimation increases with decreasing pre-test probabilities of disease [29]. This research suggests that the most informative measures for doctors may be estimates of the post-test probability of disease (predictive value), which can be presented as a range corresponding to different pre-test probabilities. However, graphical displays that facilitate the derivation of post-test probabilities, such as likelihood ratio nomograms, are usually based on summary estimates of test characteristics (positive and negative likelihood ratios) without allowing for the precision of the estimate, or its applicability to a given population. Use of summary estimates in this way is questionable in the context of reviews of diagnostic accuracy studies, which typically find substantial between-study heterogeneity [30]. It is particularly problematic if the summary estimate is the only information conveyed in a graphic and the graphic is taken as the key message of the paper.

The inclusion of some form of graphical presentation of test accuracy data has a number of advantages compared

---

**Table 3: Number of studies evaluating each category of tests in the primary studies and systematic reviews.**

<table>
<thead>
<tr>
<th>Test category</th>
<th>Number of primary studies</th>
<th>Number of systematic reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical examination</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Imaging</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>Laboratory</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>Questionnaires</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Combination of different categories</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Ten reviews (20%) used forest plots to display individual study results. One study provided a plot of diagnostic odds ratios, while all others displayed paired plots of sensitivity and specificity (8 reviews), positive and negative likelihood ratios (3 reviews), or positive and negative predictive values (1 review). Several studies displayed more than one set of forest plots, including plots for more than one summary measure, for different stages of diagnosis, different test thresholds or for different tests. One study included a forest plot of summary data only, showing how pooled estimates of positive and negative likelihood ratios varied for different patient subgroups.

**Predictive value**

None of the studies included a likelihood ratio nomogram. One primary study and five systematic reviews included a probability-modifying plot.

---

**Note:**

The information provided is a natural representation of the document content as per the instructions. The table and other data elements have been rendered in a readable format. For precise citation, this page number is not intended for citation purposes. For more detailed information, refer to the original document.
include references to the STARD flow diagram. STARD itself does not comment on how graphical displays should be used to convey results of test accuracy studies other than to recommend the inclusion of a flow diagram and to provide an illustration of a dot-plot as a suggestion for how individual study results may be displayed. Guidelines on the type of graphical displays that should be included in reports of test accuracy studies could be considered when STARD is next updated, and should be considered by journals in their instructions for authors.

Conclusion
Our review suggests that graphical displays are currently underused in primary diagnostic accuracy studies and systematic reviews of such studies. Graphical displays of diagnostic accuracy data should provide an easily interpretable and accurate representation of study results, conveying both diagnostic accuracy and predictive value. This is not usually possible in a single graphic: the type of information presented in the most commonly used graphs does not directly allow clinicians to assess the implications of test results for an individual patient.

Competing interests
The author(s) declare that they have no competing interests.

Authors’ contributions
All authors contributed to the design of the study and read and approved the final manuscript. PFW and MEW identified relevant studies and extracted data from included studies. PFW carried out the analysis and drafted the manuscript with help from JD and RH.

Web Appendix: Studies included in the review

a. Primary studies


b. Systematic reviews


46. Trowbridge RL, Rutkowski NK, Shojania KG. Does this patient have acute cholecystitis? JAMA 2003;289:80-86.


Acknowledgements

This work was supported by the MRC Health Services Research Collaboration. Jonathan Deeks is funded by a Senior Research Fellowship in Evidence Synthesis from the Department of Health.

References


23. Centre for Reviews and Dissemination databases 2008 [http://www.crd.york.ac.uk/CRDweb/].


Pre-publication history
The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/8/20/prepub