Communicating test accuracy terms to practicing physicians – a controlled study

Johann Steurer, Joachim E. Fischer, Lucas M. Bachmann, Michael Koller, Gerben ter Riet

*BMJ. 2002 Apr 6;324(7341):824*
Abstract

Objective: To assess the extent to which different forms of summarising diagnostic test information influence practicing physicians' ability to estimate disease probabilities.

Design: Controlled non-clinical study

Setting: Three conferences in continuous medical education

Participants: 263 practicing physicians

Measurements: Physicians completed three multiple-choice questions assessing their knowledge and application of terms of test accuracy. Participants were then asked to estimate the disease probability in a 65-year-old woman presenting with uterine bleeding. Physicians had to revise their estimate in the light of a positive transvaginal ultrasound scan. They received this test result in one of three versions, which were allocated in a concealed fashion: the test result only; the test result and the information that sensitivity and specificity were 80% and 60%, respectively; or the test result and information on the corresponding positive likelihood ratio of 2. The latter was presented in common sense wording.

Results: 76% of the physicians chose the correct definition for sensitivity, 61% for predictive value, but only 22% chose the correct answer for the post-test probability of a positive screening test. In the clinical example, physicians not provided with information on the accuracy of the scan overestimated its diagnostic value (median attributed likelihood ratio (aLR) = 9.0, against 2.54 reported in the literature). Providing the scan's sensitivity and specificity reduced the overestimation (median aLR = 6.0), but to a lesser extent than common sense wording of the likelihood ratio (median aLR = 3.0).

Conclusion: Most general practitioners recognise the correct definitions for sensitivity and positive predictive value, but do not apply them correctly. Compared to the reporting of sensitivity and specificity, conveying the same information in terms of a common-sense verbal expression of the likelihood ratio appears to improve GPs ability to estimate disease probabilities accurately.
Introduction

Physicians are expected to be proficient in integrating diagnostic information from history, physical examination and other diagnostic procedures. Effective therapeutic action rests on the correct interpretation of such data.

Usually, the accuracy of tests is reported in terms of their sensitivity, specificity and predictive values. In settings with low disease prevalence, most physicians grossly overestimate the probability of disease in patients with a positive result from a screening test. Apparently, they confuse the sensitivity of a screening test with its positive predictive value. Less is known about practicing physicians’ aptitude to apply test accuracy data to clinical settings with a higher prevalence of disease. We therefore presented a structured questionnaire with a vignette of a clinical problem to general practitioners (GP). The primary aim of the study was to assess the extent to which different forms of presenting test accuracy information affected the physicians’ post-test probability estimates.

Methods

GPs attending one of three continuous medical education conferences in Switzerland were eligible. On average, participating physicians had more than ten years of professional experience. Although GPs do not formally act as gatekeepers, they are usually the first health care provider, who is contacted if new medical problems arise.

The questionnaire (see BMJ website), which was developed and piloted in a different group of 45 physicians, consisted of two parts. The first part – using a 4-item multiple-choice format - assessed knowledge and interpretation of the terms sensitivity and predictive value. The first two questions asked for the definition of the terms ‘sensitivity’, and ‘positive predictive value’, respectively. The third question asked to estimate the post-test probability when a screening test with a sensitivity and specificity of 95% returns a positive result in a population with a disease prevalence of 1%. Choices were “< 25%”, “approximately 50%”, “nearly 100%”, and “don’t know”.

The second part evaluated the participants’ ability to apply these terms to a clinical vignette. For simplicity of further calculations, the vignette stated that the prevalence of endometrial cancer in all women with abnormal uterine bleeding is 10%. First, physicians were requested to estimate the probability of endometrial cancer given the additional information that the patient is 65 years old. The second piece of information on the result of a transvaginal ultrasound scan was provided in three different versions: (1) “Transvaginal ultrasound showed a pathological result compatible with cancer”; (2) “Transvaginal ultrasound showed a pathological result compatible with cancer. The sensitivity of this test is 80%, its specificity is 60%”; (3) “Transvaginal ultrasound showed a pathological result compatible with cancer. A positive result is obtained twice as frequently in women with an endometrial cancer than in women without this disease”. Version three aimed to present the positive likelihood ratio of 2 using non-technical wording.
Data collection

Participants received the questionnaires presenting the test result in one of three versions, which were allocated in a concealed fashion. Questionnaires were handed out prior to the start of a lecture on evidence-based medicine. Participants were given oral instructions and 10 minutes time to complete the questionnaires. Thereafter, the questionnaires were collected. If GPs attended more than one conference, only the first completed questionnaire was included in the analysis.

Data analysis

For the 3 questions testing formal knowledge we calculated the proportions of GPs who chose the correct answer and their 95% confidence intervals (CI). For the second part of the questionnaire, we derived the implicitly attributed likelihood ratios (aLR) by comparing the given pre-test probability of 10% to the probability estimate after adding information on age and ultrasound scan. We used the equation $aLR = \frac{\text{post-test odds}}{\text{pre-test odds}}$, where odds = probability / (1- probability). Likewise, the likelihood ratio attributed to the positive ultrasound result was calculated (probability estimate after providing ultrasound information compared to probability estimate after providing information on age). Eight post-test probability estimates of 100% were recoded into 99.999% to avoid needless missing values. An overall comparison between the three ways of presenting information on the positive scan result were made using the Kruskal-Wallis test using the SAS statistical software package (version 8.1, SAS Inc., Cary, NC, USA). Other differences were tested using the Mann-Whitney ranksum test.

To obtain an empirical likelihood ratio corresponding to the test result “this woman is 65 years old” we used data from 248 consecutive outpatients presenting with abnormal uterine bleeding at the Birmingham (UK) Women’s Hospital rapid access ambulatory diagnostic clinic (RAAD) between November 1996 and December 1997. This database contains information on patient age and uses the endometrial biopsy as a gold standard. In this database, a woman aged 60 to 70 presenting with abnormal bleeding are 3.1 times more likely to have endometrial cancer than women outside that age range. The provided sensitivities and specificities for transvaginal ultrasound were chosen to approximate the median from data provided in the literature with rounding for simplicity of calculation.

Results

Of the 263 eligible GPs, between 251 and 261 answered the 3 questions, which tested their knowledge. Seventy-six percent (95% CI from 70 to 81%) chose the correct definition of the term “sensitivity” and 61% (95% CI from 54 to 67%) of the term “positive predictive value”. Only 22% (95% CI from 17 to 27%) chose the correct multiple-choice option “less than 25% probability” in the given example of a positive result from a screening test (sensitivity and specificity 95%, disease prevalence 1%), while 56% (95% CI from 49 to 62%) selected a post-test probability of “close to 100%”. 

90
In the clinical vignette, providing the information that the woman was in fact 65, led 48.4% of GPs to change their probability estimates. 51.6% of GPs did not change their probability estimate in the light of the additional information on age. The figure presents the key findings on the effect of (1) omitting numerical information, and (2) the two ways of presenting the same test accuracy information. Physicians, who were not given any information on test accuracy (n=92), regarded a positive transvaginal ultrasound as an accurate test to prove the presence of endometrial cancer. The median attributed likelihood ratio was 9 (IQR 3.25-68.5). Physicians provided with the literature-based sensitivity and specificity of the scan (n=92) appeared to apply lower estimates of the likelihood ratio (aLR = 6, IQR 2.3 – 22.1, p = 0.019). Presentation of the test accuracy in common sense wording (n = 79) brought the attributed likelihood ratio still closer to the literature-based ones (aLR = 3.0, IQR 2.25-9, overall p-value (testing whether all groups are equal) = 0.0013; the p-value for the comparison between group 1 versus groups 2 and 3 combined is 0.0006, indicating gross overestimation when numerical information is omitted; the p-value for the difference between groups 2 and 3 is 0.228, representing somewhat weaker evidence that the presentation of diagnostic information in the form of common sense wording of the likelihood ratio leads to estimates closer to the true one compared to its presentation as sensitivity and specificity.

Figure: Distribution of the likelihood ratios (aLR) attributed by GPs to a positive transvaginal ultrasound in women presenting with abnormal uterine bleeding. Group 1 received no information on test accuracy, Group 2 was provided with a sensitivity of 80% and a specificity of 60%, Group 3 received a common sense wording of the corresponding a Likelihood ratio of 2.
Discussion

In this study we evaluated practising physicians’ knowledge on terms commonly used to describe a test’s accuracy. Participants were asked to calculate the post-test probability of disease after a positive screening test and to provide probability estimates for a clinical vignette, which described an unspecified woman presenting with abnormal uterine bleeding. In the latter, physicians had to use two pieces of information (age and a positive ultrasound result) to update their probability estimate for the presence of endometrial cancer.

Four findings emerged from this study: (1) while the majority of physicians were able to recognise the correct multiple choice answer for sensitivity and positive predictive value, only a minority correctly estimated the (low) probability of disease when provided with a 1% prevalence, 95% sensitivity and 95% specificity. (2) Compared to the empirical data found in women referred to an outpatient clinic, physicians underestimated the diagnostic value of the information on patient age. (3) Participants, who were not provided with data on test accuracy, grossly overestimated the diagnostic accuracy of a positive transvaginal ultrasound result – both if compared to a recently published systematic review as well as compared to the data provided in the Swiss guidelines on the workup of women with post-menopausal bleeding. (4) Presentation of test accuracy by expressing it as the positive likelihood ratio using common-sense wording seemed to be more effective for eliciting correct post-test probabilities than presenting it as sensitivity and specificity.

If – as seems likely – in our sample GPs with an interest in evidence-based medicine were overrepresented our findings may even overestimate to some extent the average GP’s achievements on the knowledge questions we measured. One may speculate whether the findings in the comparative part of this study would be affected by prior knowledge on test accuracy measures and their presentation. One may argue that, in general, not all mis-estimations of disease probability estimates are of equal importance. In particular, two numerically different estimates may not be clinically different if they lead to the same clinical decision. However it is difficult to be specific about these thresholds for action as they may depend on many subjective factors unless formal decision analyses exist.

Despite a long tradition of reporting test accuracy in terms of sensitivity and specificity, only a minority of physicians could correctly apply this information to a setting with low prevalence or to a clinical judgement. This inability to quickly perform the required calculations probably explains their little use in general practice. Rather than blaming physicians for this lack of aptitude, medical editors and authors of diagnostic research studies alike should reconsider the communication of research data. We showed, that simple presentation of a positive likelihood ratio of 2 in common sense wording enabled physicians to almost correctly estimate the diagnostic information associated with a positive ultrasound scan. In future research, other potentially useful ways to present diagnostic data, e.g. disease probability estimates, prediction rules or decision trees could also be explored.

Our study raises the question, to what extent the observed overestimation of the diagnostic value of the technical procedure contributes to the steadily increasing utilisation of laboratory and imaging testing.
Another potential reason to underestimate the diagnostic value of information from the patient history may be the lack of well-designed studies tackling that issue. This study gives no insight into what consequences physicians would draw from a positive ultrasound result in real practice. However, if other reasoning approaches do not correct for the observed overestimation of a positive ultrasound result, it is conceivable that this favouritism of technical data may adversely affect both doctor-patient communication and further action.

In summary, we observed a considerable overestimation of the diagnostic value of a positive transvaginal ultrasound in elderly women with abnormal uterine bleeding. Presenting test accuracy data in common sense wording markedly reduced this overestimation.

Acknowledgments
We thank Wim Verstappen, MD (Dept. General Practice, Maastricht University) for sharing with us his questionnaire for testing GPs knowledge of test accuracy issues and Drs K.S. Khan and J.K Gupta for providing the outpatient database.

References
