Tips for teachers of evidence-based medicine: 5. The effect of spectrum of disease on the performance of diagnostic tests

Victor M. Montori, Peter Wyer, Thomas B. Newman, Sheri Keitz, Gordon Guyatt, for the Evidence-based Medicine Teaching Tips Working Group

For clinicians to use a diagnostic test in clinical practice, they need to know how well the test distinguishes between those who have the suspected disease or condition and those who do not. If investigators choose the wrong disease-positive and disease-negative populations for their study of a diagnostic test and thereby introduce what is sometimes called spectrum bias, the results may seriously mislead clinicians.

In this article we present an approach to helping clinicians understand the effect of differences in spectrum of disease, in contrast to the effect of differences in prevalence of disease, on the performance of diagnostic tests. As with other articles in this series, clinical educators experienced in teaching evidence-based medicine developed these tips and have used them extensively. A full description of the development of the tips in this series, as well as pertinent background information, has been presented elsewhere.[1]

For each of the 2 tips in this article, we have provided guidance on when to use the tip, the teaching script for the tip, a "bottom line" section and a summary card. For each tip we have identified the appropriate level of learner experience and provided estimates of the time required for the exercise.

This article addresses 2 stumbling blocks to understanding how the choice of patients for the study of a diagnostic test may affect the estimate of how the test performs. The first tip helps learners understand how the spectrum of disease in disease-positive patients and the spectrum of competing conditions in disease-negative patients can affect the test's apparent diagnostic power. The second tip assumes that the first tip has already been presented. It helps learners understand how, despite the potentially powerful impact of spectrum of disease and competing conditions, prevalence of disease will not alter the test characteristics. Both tips assume a basic familiarity with measures of test performance, including sensitivity, specificity and likelihood ratios.

Generally, clinicians understand that a test may perform better when it is used to evaluate patients with more severe disease and that diagnostic tests may not be needed when the disease is so manifest. The first tip helps them overcome the stumbling block associated with reconciling this notion with the instruction that test characteristics are properties of the test themselves and should not vary with the patient's characteristics.

After overcoming this conceptual hurdle, clinicians need to understand that, although post-test probabilities vary with disease prevalence, sensitivities, specificities and likelihood ratios do not. They may experience this as apparently contradicting their discoveries from tip 1. They may particularly stumble because of the apparent interconnection between disease prevalence and severity in many clinical settings.

For instance, rheumatoid arthritis seen in a family physician's office will be relatively uncommon, and most patients will have a relatively mild case. In contrast, rheumatoid arthritis will be common in a rheumatologist's office, and patients will tend to have relatively severe disease. Tests to diagnose rheumatoid arthritis in the rheumatologist's waiting area (e.g., hand inspection for joint deformity) are likely to be relatively more sensitive not because of the increased prevalence but because of the spectrum of disease present (e.g., degree and extent of joint deformity) in this setting. The second teaching tip aims to overcome this source of confusion by working through some simple illustrative calculations.

Teaching tip 1: The "ideal" spectrum of disease

When to use this tip

This tip is suitable for all clinicians and clinical trainees who do not yet have a clear understanding of the concepts of spectrum of disease and test characteristics. The objective is to foster this understanding. The tip takes 15 to 20 minutes to complete. By the end of this tip learners should be able to:

- Understand the importance of spectrum of disease in the evaluation of diagnostic test characteristics.

Other available resources

- A companion version of this article directed to learners of evidence-based medicine has been published in *CMAJ* and is available online through *eCMAJ* (www.cmaj.ca/cgi/content/full/173/4/385/DC1).
- An interactive version of this article, as well as other tools and resources, is available at www.ebmtips.net/ci001.asp

This tip is useful for learners who are critically appraising an article about a diagnostic test and are trying to assess whether the patients included in the study correspond to the appropriate spectrum of patients to whom the test will be applied.[2]

The script

Begin the tip by ensuring that you and the learners are using the same language. You establish the concepts of disease-positive and disease-negative patients and distinguish between the test that you are evaluating and the criterion or reference standard (or "gold standard") used to unequivocally establish whether patients have or do not have the target disease. You then ask, "How should we choose our disease-positive and disease-negative populations?" You offer 1 alternative: disease-positive patients are those with unequivocal disease (typically, far advanced) and disease-negative patients are normal people who are unequivocally disease-free (often laboratory technicians or medical students). Investigators often choose these populations to avoid misclassification.

You ask the learners what they think of this particular choice of study participants. Learners unfamiliar with concepts of critical appraisal generally conclude that this choice is satisfactory. In groups that are more advanced in critical appraisal skills, someone may supply a compelling critique of the choice. In either situation, ask the learners to consider the following issues in deciding on the optimal choice of population for a diagnostic test study.

Distributions of test results among disease-positive and disease-negative patients

Draw *x* and *y* axes and explain that the horizontal axis represents the possible range of results of a hypothetical blood test from very low, or normal, to very high, or abnormal values, and that the vertical axis represents the proportion of patients at each test result (Fig. 1). Ask the learners what they would expect the distribution of test results to be in a population of patients who unequivocally have the target condition and in whom the condition is thus far advanced. They will suggest that such patients will have test results on the right side of the scale. You draw the disease-positive distribution of Fig. 1.

You then ask the learners what the distribution of test results is likely to be in a group of completely normal patients. When they suggest it will be on the left side of the distribution, you draw the distribution of negative results for the target condition of Fig. 1.

Your next question is whether, given these results, the test appears to be a good or bad one, and why. The learners will suggest that the test appears to be good and (sometimes with a little prompting) that the reason is that the curves demonstrate very little overlap. You agree, pointing out that for test results above point B, all patients have the condition, and for all test results below point A, no patients have the condition (Fig. 1).

You now ask the learners what they think the distribution of test results will be in a population of patients with less severe disease — those in whom you might be uncertain whether they were disease-positive or disease-negative. They will suggest a curve to the left of the first disease-positive curve, which you draw as the disease-positive curve in Fig. 2. To allow for visual comparison, draw the curves on Fig. 2 on

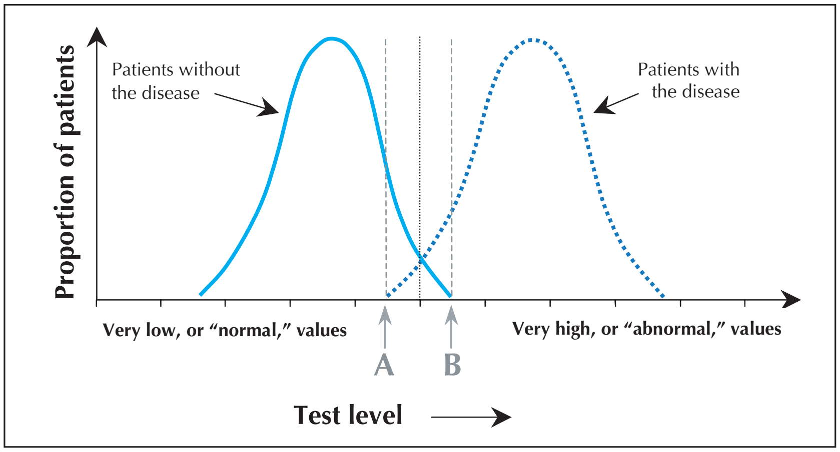

Fig. 1: Spectrum bias. Test performance when differentiating normal volunteers (disease-negative) from severely ill patients (disease-positive). For test results above point B, all patients have the condition, and for all test results below point A, no patients have the condition. The distance between A and B shows the extent of the overlap of test results between the 2 groups.

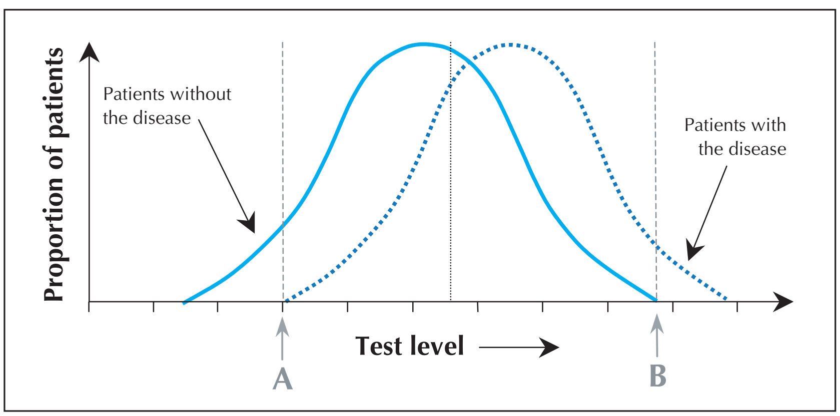

Fig. 2: Spectrum bias. Test performance when differentiating patients who have the disease from those who don't when both groups appear to have the target disease. The distance between A and B is now much wider.

top of those in Fig. 1 or as a separate figure below or above Fig. 1. Finally, you ask them to characterize the distribution of test results of disease-negative patients who present with symptoms or diseases that mimic those of the target condition. They will suggest that this distribution is also closer to the middle of the axis, and you draw it as the disease-negative curve in Fig. 2.

How does the test now perform in distinguishing between patients who have the disease and those who don't? Not very well, as the group will quickly point out as they note the extent of the overlapping distributions.

At this point you ask learners which groups they are interested in differentiating: normal patients from patients with severe disease, or patients who appear as if they might have the target condition and do from those who appear they might have the target condition and don't. The learners will choose the latter, and you agree. Varying ways of expressing the "right" population for a diagnostic test study include:

- Those in whom we're uncertain of the diagnosis;
- Those in whom we'll use the test in clinical practice to resolve our uncertainty; and
- Disease-positive patients with a wide spectrum of severity, and disease-negative patients with a sampling of symptoms commonly associated with the target disease.

The bottom line

- Test performance will vary with the spectrum of disease within a study population.
- The ability of a test to differentiate normal volunteers from severely ill patients may be misleading when the test is applied in clinical practice.
- Clinicians need diagnostic tests when patients with and without the target condition cannot be distinguished without a test. Learners should be interested in the performance of the test in this situation.

See Appendix 1 for the summary card for this tip.

Extensions for advanced learners

1. Sometimes, learners take the perspective of investigators. When doing so, they rightly point out that the initial investigation of a test may appropriately enroll normal people and severely diseased patients. If the test fails to distinguish the 2 groups, the investigators go back to the drawing board. If it succeeds, a subsequent investigation (of considerably more interest to clinicians) will examine test performance in the truly relevant populations.
2. You can help learners gain a deeper understanding of likelihood ratios by pointing out that the likelihood ratio for any given test result is represented by the respective height of the curves for each result represented on the *x* axis (Fig. 3). The point on the *x* axis below the intersection of the 2 curves is the test result with a likelihood ratio of 1. Fig. 3 also identifies test values corresponding to likelihood ratios of 0.25 and 4. Comparing Fig. 1 with Fig. 2, you can point out that the relative heights of the 2 curves, and hence the likelihood ratios, corresponding to a given test value will change as the curves move closer together and the area of overlap increases.
3. Finally, you may want to point out the empirical support for attending to the selection of disease-positive and disease-negative patients. Lijmer and collaborators found that if investigators evaluated the test in a group of patients already known to have the disease and in a separate group of normal people (referred to as a case–control study), their results suggested a much more powerful test than when both disease-positive and disease-negative patients were drawn from the same population with suspected disease (relative diagnostic odds ratio 3.0, 95% confidence interval 2.0–4.5).[3]

Teaching tip 2: Prevalence, spectrum and test characteristics

When to use this tip

This tip is suitable for learners possessing an intermediate level of critical appraisal skills, and works best when presented after tip 1. The numerical exercise works best with learners who are familiar with the computation of test characteristics (e.g., likelihood ratios, sensitivity, specificity and predictive value). The objective is to help learners understand the impact of disease prevalence on test character-

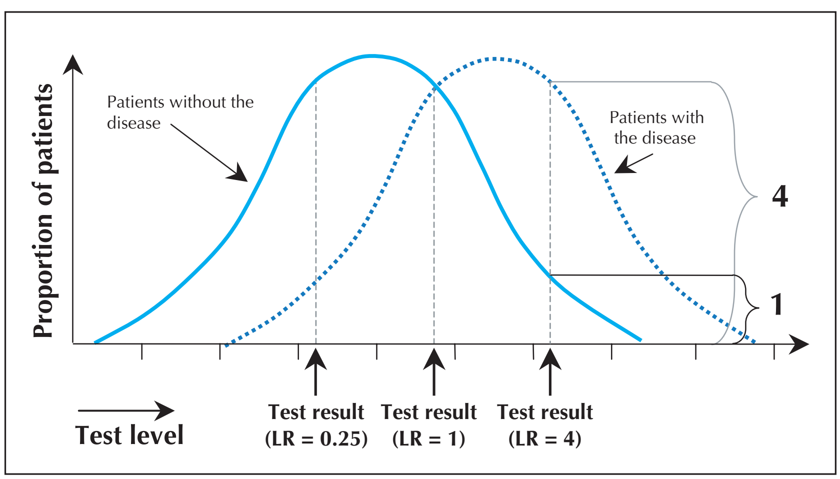

Fig. 3: Likelihood ratios and spectrum of disease. The likelihood ratio of a test result represented by a point on the horizontal axis is the height of the curve corresponding to disease-positive patients divided by the height of the curve corresponding to disease-negative patients.

istics. This tip takes 15 to 20 minutes to complete. By the end of this tip learners should be able to:

- Understand the lack of impact of disease prevalence on sensitivity, specificity and likelihood ratios.
- Understand the impact of disease prevalence on the post-test probability of disease (predictive value of the test).

After experiencing tip 1, learners will frequently ask about the effect of disease prevalence on diagnostic test characteristics. This tip attempts to disentangle the common coincidence of greater disease severity and increased prevalence and to help learners to understand that, although the predictive values of a test change with prevalence, the likelihood ratios (and sensitivity and specificity) remain constant.

The script

This tip uses a graphical and a numerical exercise. The graphical exercise involves using Fig. 2 after presenting tip 1. For this tip, explain that researchers were taking samples of subjects from different populations with similar disease spectrums (and therefore appropriate candidates to receive the test) but different prevalence of the target disease.

Referring to Fig. 2 on the board, ask the learners what happens to the performance of the test if researchers take a different sample, involving a different proportion of disease-positive and disease-negative patients from that depicted in Fig. 2. Although the researchers draw their sample from the same groups of disease-positive and disease-negative patients, on this occasion they find that the number of patients with the target disease is twice as large as in the sample used to generate Fig. 2, wherease the number of disease-negative patients is unchanged. The learners may initially mistake the height of the curves with the number of patients in each group. Remind them that the height of the curves represents the proportion of patients with a particular test result. They will then realize that the proportions will not change and that the performance of the tests (in this case measured using the likelihood ratios) is unaffected.

Ask what happens to the test if the researchers take another sample with a similar spectrum of disease but where the number of patients with the target disease is half as large as in the sample used to generate Fig. 2. They will quickly point out that nothing happens to the test characteristics. Conclude that the relative number of patients with and without the target disease, or prevalence of the target disease, is irrelevant to the estimation of test characteristics.

The numerical exercise starts by drawing a 2 × 2 table where the proportion of patients with disease is 50% and the proportion of patients without the disease is 50% (prevalence of disease of 50%) (Fig. 4A). Ask the learners what will happen if the researchers sample from a population with a

A

	Disease	No disease	Total
Test +	A	B	
Test –	C	D	
Total	50	50	100

A 2 × 2 table corresponding to a study of a diagnostic test in which the prevalence of disease is 50%.

B

Test +	A × 2	B	
Test –	C × 2	D	
Total	100	50	150

In this study the prevalence of disease is 67%.

C

Test +	A	B × 2	
Test –	C	D × 2	
Total	50	100	150

Here the prevalence of disease is 33%.

Fig. 4: Disease prevalence and diagnostic test characteristics.

similar spectrum of disease but select patients in such a fashion that the proportion with and without the target disease is 2:1 (Fig. 4B).

The learners should be able to note that the prevalence of disease has increased to 67% but that the percentage of people with and without the disease having positive and negative test results has not. Those interested in calculating test characteristics using this table will notice that the proportion of people with the disease and a positive test result (or sensitivity) has not changed before and after the increase in disease prevalence (A ÷ 50 v. 2A ÷ 100). They should also notice that the proportion of people with a positive (or negative) test result and the target disease has changed before and after the increase in disease prevalence. For example, for a positive test: $A \div (A + B) \neq 2A \div (2A + B)$.

Finally, ask what will happen if the researchers sample from a population with a similar spectrum of disease but select patients in such a fashion that the proportion with and without the target disease is 1:2 (Fig. 4C). The learners will note that the sensitivity remains the same despite a decrease in the proportion of disease-positive patients from 50% to 33%, and that the specificity remains the same despite an increase in the proportion of disease-negative patients to 67%. Once again, conclude that disease prevalence is irrelevant to the estimation of test characteristics.

The bottom line

- Stumbling block: Confusion about the impact of disease prevalence on test characteristics because disease severity and prevalence are frequently interconnected.
- With this tip, in graphical and numerical form, learners discover that disease prevalence has no effect on test characteristics (e.g., likelihood ratios, sensitivity and specificity).

See Appendix 1 for the summary card for this tip.

Extensions for advanced learners

1. Motivated by tip 2, learners may enquire about a "real world" clinical situation in which patients present with similar disease spectra but different disease prevalence. "Real world" examples of conditions that may present with equal severity in people with different demographic characteristics (age, sex, ethnic origin) but that may be much more prevalent in 1 group than in another include mild osteoarthritis of the knee in young and old patients, and asymptomatic thyroid abnormalities in men and women. In both examples, the spectrum of severity remains constant and diagnostic tests will have the same sensitivity, specificity and likelihood ratios in young and old and in men and women respectively. However, influenced by the higher prevalence of osteoarthritis among elderly patients and thyroid abnormalities among women, the proportion of those with a positive test result who do in fact have the disease will be much higher in these 2 groups.
2. Learners may become worried about applying test results generated in populations and settings different from those of their practice. When transporting a test studied in one population or setting to another, test characteristics may be affected by factors other than differences in disease prevalence and severity. In the new setting, there may be a different level of expertise in conducting and interpreting the test of interest and the reference standard, and different prevalence of competing diagnoses for both disease-positive and disease-negative conditions. This may lead to misclassification and deterioration of the test characteristics in the new setting. We think that it would be safe to apply results from valid diagnostic studies unless there are important differences in the conduct or interpretation of the test or in the prevalence of competing conditions.

Report on field-testing

One of us (S.K.), an experienced teacher of evidence-based medicine, field-tested these tips with medical residents in the United States during one 1-hour teaching session. There were 16 residents in the session: 8 were start-of-year interns (naive learners) with very little experience in evidence-based medicine, 7 were residents who were fairly comfortable with the material, and 1 was a senior resident who was very comfortable with evidence-based medicine. S.K. used 1.5 hours to prepare these scripts. In that time, she produced transparencies and handouts for the learners to generate the figures. The session was divided into 2 parts to correspond to each of the tips, and each part took 25 minutes to complete.

For tip 1, S.K. considered it important to emphasize that the y axis in Fig. 1 corresponds to the proportion of patients rather than to the number of patients and that the x axis has labels of "very normal" and "very abnormal" rather than specific numbers. The extension to tip 1 was used successfully after the group had a chance to review likelihood ratios during a subsequent session. When using tip 2, the learners were very interested in discussing sensitivity, specificity and predictive values before embarking on these tips. This perceived need was successfully dealt with by using the 2×2 table.

When asked about the importance of the concepts, the learners gave tip 1 an average score of 8.3 and tip 2 an average score of 8.5 out of 10. When asked about the clarity of the presentations, the learners gave tip 1 an average score of 7.3 and tip 2 an average score of 7.1 out of 10. For learners, the most important message involved the impact of prevalence of disease on test characteristics.

Conclusion

Clinicians must understand how the spectrum of disease in disease-positive patients, and the spectrum of competing conditions in disease-negative patients, can affect the apparent diagnostic power of a test power. They also need to understand why, despite the potentially powerful impact of spectrum of disease and competing conditions, prevalence of disease will not have an impact on test performance. We have presented 2 teaching tips, developed and used by experienced clinician-educators, that help overcome the learner difficulties commonly encountered in teaching these concepts.

This article has been peer reviewed.

From the Knowledge and Encounter Research Unit, Department of Medicine, Mayo Clinic College of Medicine, Rochester, Minn. (Montori); the Departments of Epidemiology and Biostatistics and of Pediatrics, University of California, San Francisco (Newman); Durham Veterans Affairs Medical Center and Duke University Medical Center, Durham, NC (Keitz); the Columbia University College of Physicians and Surgeons, New York, NY (Wyer); and the Departments of Medicine and of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont. (Guyatt)

Competing interests: None declared.

Contributors: Victor Montori, as principal author, oversaw and contributed to the writing of the manuscript. Thomas Newman reviewed the manuscript at all phases of development and contributed to the writing as coauthor of tip 2. Sheri Keitz used all of the tips as part of a live teaching exercise and submitted comments, suggestions and the possible variations that are reported in the manuscript. Peter Wyer reviewed and revised the final draft of the manuscript to achieve uniform adherence with format specifications. Gordon Guyatt developed the original idea for tips 1 and 2, reviewed the manuscript at all phases of development, contributed to the writing as coauthor, and reviewed and revised the final draft of the manuscript to achieve accuracy and consistency of content as general editor.

References

1. Wyer PC, Keitz S, Hatala R, Hayward R, Barratt A, Montori V, et al. Tips for learning and teaching evidence-based medicine: introduction to the series. *CMAJ* 2004;171(4):347-8.
2. Jaeschke R, Guyatt G, Lijmer J. Diagnostic tests. In: Guyatt G, Rennie D, eds. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 121-40.
3. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.

Correspondence to: *Dr. Peter C. Wyer, 446 Pelhamdale Ave., Pelham NY 10804; fax 212 305-6792; pwyer@att.net*

Members of the Evidence-Based Medicine Teaching Tips Working Group: *Peter C. Wyer (project director), College of Physicians and Surgeons, Columbia University, New York, NY; Deborah Cook, Gordon Guyatt (general editor), Ted Haines, Roman Jaeschke, McMaster University, Hamilton, Ont.; Rose Hatala (internal review coordinator), University of British Columbia, Vancouver, BC; Robert Hayward (editor, online version), Bruce Fisher, University of Alberta, Edmonton, Alta.; Sheri Keitz (field test coordinator), Durham Veterans Affairs Medical Center and Duke University Medical Center, Durham, NC; Alexandra Barratt, University of Sydney, Sydney, Australia; Pamela Charney, Albert Einstein College of Medicine, Bronx, NY; Antonio L. Dans, University of the Philippines College of Medicine, Manila, The Philippines; Barnet Eskin, Morristown Memorial Hospital, Morristown, NJ; Jennifer Kleinbart, Emory University School of Medicine, Atlanta, Ga.; Hui Lee, formerly Group Health Centre, Sault Ste. Marie, Ont. (deceased); Rosanne Leipzig, Thomas McGinn, Mount Sinai Medical Center, New York, NY; Victor M. Montori, Mayo Clinic College of Medicine, Rochester, Minn.; Virginia Moyer, University of Texas, Houston, Tex.; Thomas B. Newman, University of California, San Francisco, San Francisco, Calif.; Jim Nishikawa, University of Ottawa, Ottawa, Ont.; Kameshwar Prasad, Arabian Gulf University, Manama, Bahrain; W. Scott Richardson, Wright State University, Dayton, Ohio; Mark C. Wilson, University of Iowa, Iowa City, Iowa.*

Articles to date in this series

Barratt A, Wyer PC, Hatala R, McGinn T, Dans AL, Keitz S, et al. Tips for teachers of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. Available: www.cmaj.ca/cgi/content/full/171/4/353/DCI.

Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Moyer V, et al. Tips for teachers of evidence-based medicine: 2. Confidence intervals and *p* values. Available: www.cmaj.ca/cgi/content/full/171/6/611/DCI.

McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, Guyatt G, et al. Tips for teachers of evidence-based medicine: 3. Understanding and calculating kappa. Available: www.cmaj.ca/cgi/content/full/171/11/1369/DC1.

Hatala R, Keitz S, Wyer P, Guyatt G. Tips for teachers of evidence-based medicine: 4. Assessing heterogeneity of primary studies in systematic reviews and whether to combine their results. Available: www.cmaj.ca/cgi/content/full/172/5/661/DCI.

Appendix 1: Summary cards for 2 teaching tips on spectrum of disease and the performance of diagnostic tests

This appendix has been designed so that it can be printed on one 8½ × 11 inch page. The individual summary cards can then be cut out, if desired, for use during teaching sessions.

Teaching tip 1: The ideal spectrum of disease

Scenario: A series of studies of diagnostic test performance involving different approaches to selecting disease-positive (people with the disease of interest) and disease-negative (people without the disease of interest) subjects are represented graphically. Consider the impact of various choices of disease-positive and disease-negative patients on the performance of the test compared with that of the criterion standard.

1. Ask the learners to comment on the choice of subjects with advanced disease for the disease-positive group and of normal patients for the disease-negative group.
2. Draw bell curves with minimal overlap in which the horizontal axis is the test result and the vertical axis the proportion of patients with that test result for these 2 groups and highlight the lack of overlap.
3. Draw extensively overlapping curves for patients with less severe disease for the disease-positive group and with symptoms resembling the disease of interest for the disease-negative group and highlight the poor differentiation between these groups.
4. Ask learners whether they are interested in differentiating between those with far advanced disease from normal patients or between groups of patients in whom we're uncertain of the diagnosis.

Summary points

- Test performance will vary with the spectrum of disease within the study population.
- The ability of a test to differentiate normal volunteers from severely ill patients may be misleading when the test is applied in clinical practice.
- Clinicians need diagnostic tests when patients with and without the target condition cannot be distinguished without a test. Learners should be interested in the performance of the test in this situation.

Teaching tip 2: Prevalence versus spectrum

Scenario: A study of diagnostic test performance is represented graphically with 2 bell curves depicting the distribution of test results within disease-positive (people with the disease of interest) and disease-negative (people without the disease of interest) subjects. The learners have previously seen this figure used to demonstrate the effect of disease spectrum on the performance of the test. It is now used to illustrate that the proportion of subjects in each group with a particular test result is independent of the number of subjects in the group. The same concept is driven home with a series of calculations using 2 × 2 tables.

1. Draw 2 bell curves depicting the distribution of test results in disease-positive and disease-negative patients and point out that the height of the curves represents the proportion of patients with a particular test result. Ask them what would happen to that height if the sample taken had twice the number of disease-positive patients and twice the number of disease-negative patients.
2. Draw a 2 × 2 table with 50 disease-positive and 50 disease-negative subjects in the 2 columns. Define the rows as "test +" and "test –." Use letters to denote the number of subjects in each of the 4 boxes. The learners determine that the prevalence of disease is 50%. Change the number of subjects in the disease-positive column to 100. The learners understand that the sensitivity and specificity of the test remain constant even though the prevalence of disease has increased to 67%. Repeat the exercise with a 1:2 proportion of patients with and without the target condition and calculate the prevalence, sensitivity and specificity.

Summary points

- Stumbling block: Confusion about the impact of disease prevalence on test characteristics because disease severity and prevalence are frequently interconnected.
- Learners discover that disease prevalence has no effect on test characteristics (e.g., sensitivity, specificity and likelihood ratios).