Tips for teachers of evidence-based medicine: 3. Understanding and calculating kappa

Thomas McGinn, Peter C. Wyer, Thomas B. Newman, Sheri Keitz, Rosanne Leipzig, Gordon Guyatt, for the Evidence-Based Medicine Teaching Tips Working Group

For many studies in the medical literature, clinician-readers will be interested in the extent of agreement among multiple observers. For example, do the investigators in a clinical study agree on the presence or absence of physical, radiographic or laboratory findings? Do investigators involved in a systematic overview agree on the validity of an article or on whether the article should be included in the analysis? In perusing these types of studies, where investigators are interested in quantifying agreement, clinicians will often come across the kappa statistic, and it is likely to mystify them.

Whenever 2 people judge the presence or absence of an attribute, they will agree some of the time by chance. When investigators report the degree of agreement at "face value" (i.e., the percentage of assessments in which the observers agreed), the contribution of chance to this tally, if large, may result in a misleading impression. Statisticians have developed solutions to this problem. When categorical data are involved (test result positive or negative or a number of ordered categories such as high, intermediate and low probability), the most popular is "chance-corrected agreement," quantified as kappa or weighted kappa.

In this article we present 3 approaches to helping clinicians use the concepts of kappa when applying diagnostic tests in practice. As with other articles in this series, clinical educators experienced in teaching evidence-based medicine developed the tips and have used them extensively.[1] We have attempted to capture the interactive process that the educators who developed the approaches characteristically observe when using them. We have also emphasized the stumbling blocks that these educators have most commonly encountered among their learners. A full description of the development of the tips presented in this series, as well as pertinent background information, has been presented elsewhere.[1]

For each of the 3 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. In trying to teach concepts related to observer agreement, it is important to appreciate that several layers are involved, not all of which are necessary for every learner to know. First, there is the simple understanding of how to interpret a kappa value and its implications for clinical practice (tips 1 and 2); then, there are the more difficult concepts related to estimating chance agreement (tip 3).

These exercises are meant to be flexible. Instructors can use the tips in whole or in part, in various settings, with different learners and different goals in mind. Completing all 3 tips takes some time and is best reserved for a group that is very keen; however, the first 2 tips can be accomplished quickly.

Teaching tip 1: Defining the importance of kappa

When to use this tip

This tip is suitable for beginners and intermediate-level learners. The exercise takes 5 to 10 minutes. The general objective is to make clear how calculating kappa differs from simply measuring the percent agreement, with the following specific objectives:

- Understand the difference between measuring agreement and measuring agreement beyond chance.
- Understand the implications of the different values of kappa.

A common stumbling block for learners is grasping the basic concept of agreement beyond chance and, in turn, the importance of correcting for chance agreement. The example below is intended to help them overcome this hurdle.

In the course of a short group discussion, such as might take place during rounds, a group member may ask a ques-

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/171/11/1369/DCI)
- An interactive version of this article, as well as other tools and resources, is available at www.ebmtips.net/ci001.asp

tion such as, "What is the meaning of the statement in the results that 'the kappa for agreement on the presence of Murphy's sign was 0.6?'" Your first answer might be as follows: "People making a decision on the presence or absence of an element of the physical examination, such as Murphy's sign, will sometimes agree simply by chance. The kappa statistic corrects for the chance agreement and tells you how much of the possible agreement over and above chance the clinicians have achieved."

Typically, the faces of the participants will, at this point, reveal some degree of puzzlement or discomfort. If you perceive this, you can say, "If the group is interested in gaining a deeper understanding of kappa and chance-corrected agreement, we can go through an exercise. But perhaps you are satisfied with simply knowing that kappa corrects for chance agreement. How do people feel?" Some groups elect to move on without further explanation, and others will seek a deeper understanding.

The script

A simple example, such as the following, usually helps to clarify the importance of chance agreement. Two radiologists independently read the same 100 mammograms. Reader 1 is having a bad day and reads all the films as negative without looking at them in great detail. Reader 2 reads the films more carefully and identifies 4 of the 100 mammograms as positive (suspicious for malignant disease). The percent agreement between the 2 readers is 96%, even though one of them has arbitrarily decided to call all of the results negative. The learners will see that measuring the simple percent agreement overestimates the degree of clinically important agreement in a fashion that is misleading. The role of kappa is to assess how much the 2 observers agree beyond the agreement that is expected by chance. (The calculation of chance agreement is discussed in teaching tip 3 in this article.) Once learners grasp the importance of kappa and understand that the simple percent agreement may be very misleading, it is worth listing the implications of different kappa values, as outlined in Table 1.

After you have listed these numbers, it is also helpful to list examples of each, so that the students can understand their clinical relevance (Table 2).

The examples should ideally be a mix of invasive procedures (bone marrow analysis, arteriography) and noninvasive tests (history and physical examination). In the list shown in Table 2, the kappa values for the CAGE questionnaire[5] and for detection of goitre[6] are higher than for the more invasive arteriography for lower extremity arterial disease[10] and the more elaborate exercise stress test.[3]

At this point, depending on the setting, the learners may wish to move on to other pressing issues, such as writing discharge orders or performing a lumbar puncture. However, if time and interest allow, you can move on to teaching tip 2, a more detailed explanation of calculating kappa values. Many learners are curious as to how the value judgements of slight, fair, moderate and substantial (Table 1) were assigned to the various kappa values. This is a good starting point for the more detailed exploration that will allow them to assign their own value judgements to these numbers.

Table 1: Qualitative classification of kappa values as degree of agreement beyond chance[2]

Kappa value	Degree of agreement beyond chance
0	None
0–0.2	Slight
0.2–0.4	Fair
0.4–0.6	Moderate
0.6–0.8	Substantial
0.8–1.0	Almost perfect

The bottom line

- Stumbling block: Understanding the concept of agreement beyond chance and why it is important to correct for chance agreement.
- This tip leads learners to understand that kappa allows us to measure agreement above and beyond that expected by chance alone.
- Concrete examples of kappa scores for frequently ordered tests help to anchor this statistic in clinical practice and stimulate the learners to move on to more complex concepts related to kappa.

See Appendix 1 for the summary card for this tip.

Teaching tip 2: Calculating kappa

When to use this tip

Once the learners understand the meaning of kappa, as outlined in tip 1, they may wish to know how the value is calculated. The second teaching tip explains the calculation and is suitable for beginners and intermediate-level learn-

Table 2: Representative kappa values for common tests and clinical assessments

Assessment	Kappa value
Interpretation of T wave changes on an exercise stress test[3]	0.25
Presence of jugular venous distension[4]	0.56
Detection of alcohol dependence using CAGE questionnaire[5]	0.75
Presence of goitre[6]	0.82–0.95
Bone marrow interpretation by hematologist[7]	0.84
Straight leg raising test[8]	0.82
Diagnosis of pulmonary embolus by helical CT[9]	0.82
Diagnosis of lower extremity arterial disease by arteriography[10]	0.39–0.64

ers. The exercise takes 5 to 10 minutes. This tip has the following specific objectives for learners:

- Understand the basics of how the kappa score is calculated.
- Understand the importance of "chance agreement" in estimating kappa.

A common stumbling block for learners is understanding how to derive kappa values. Once learners understand this calculation, they can better appreciate the clinical importance and shortcomings of the kappa score. A simple bar diagram, presented on the board with the assistance of the group, can help explain the calculation.

The script

Start by asking the group to state the potential for total agreement between 2 observers. The group should have no difficulty providing the answer: 100%. Draw a horizontal bar to represent 100% agreement (Fig. 1). For the sake of the exercise, tell the group that for a hypothetical situation, the estimated chance agreement between the 2 observers is 50% and the observed agreement is 75%. Draw 2 additional bars illustrating chance and observed agreement (Fig. 1). Ask the group to identify the possible agreement beyond chance. Most learners give the correct answer quickly after viewing the diagram: 50%.

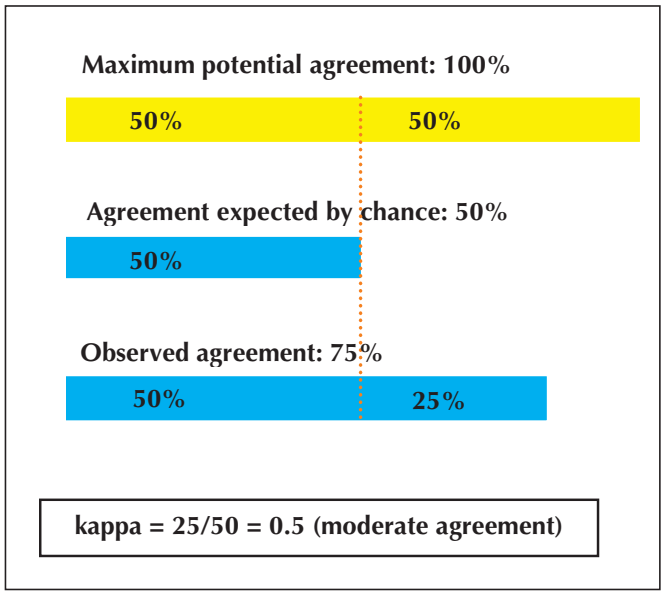

Fig. 1: Two observers independently assess the presence or absence of a finding or outcome. Their assessments agree in 75% of the cases. The yellow bar represents potential agreement (100%), the upper blue bar represents chance agreement (50%), and the lower blue bar represents actual agreement (75%). The portion of each coloured bar that lies to the left of the dotted vertical line represents the agreement expected by chance. The observed agreement above chance is half of the possible agreement above chance. The ratio of these 2 numbers is the kappa score.

Formula for calculating kappa

(Observed agreement – agreement expected by chance) ÷ (100% – agreement expected by chance)

Another way of expressing this formula:
(Observed agreement beyond chance) ÷ (maximum possible agreement beyond chance)

Then, draw a vertical line through the 3 bars at a point corresponding to 50% agreement. The portion of each bar lying to the right of this line corresponds to agreement beyond chance. You may at this point tell the group that the kappa value in this instance equals 0.50 (25 ÷ 50). Then, ask the group to figure out how this score was calculated. Allow them to struggle a bit with the calculations. Some will determine the calculation quickly, others will take longer. You may choose to write the formula for calculating kappa on the board (see box). Once all of the learners understand how you arrived at the value of 0.50, ask them if they agree that a kappa value of 0.4 to 0.6 is "moderate" agreement beyond chance, as suggested in the previous tip (see Table 1).

The bottom line

- Stumbling block: Understanding how kappa scores are derived.
- A simple graphic can be used to demonstrate the derivation of kappa.
- The graphic allows learners to begin to see how kappa is calculated. The visual image helps learners to overcome the common stumbling block more easily than would be the case if they merely worked through the numbers.

See Appendix 1 for the summary card for this tip.

Teaching tip 3: Calculating chance agreement

When to use this tip

This tip is suitable for intermediate and advanced learners. The exercise takes 15 to 20 minutes. This tip goes beyond a qualitative approach to understanding kappa and has the following specific objectives for learners:

- Understand how to calculate the kappa score given different distributions of positive and negative results.
- Understand that the more extreme the distributions of positive and negative results, the greater the agreement that will occur by chance alone.
- Understand how to calculate chance agreement, agreement beyond chance and kappa for any set of assessments by 2 observers.

Two questions naturally arise from tip 2: How was chance agreement calculated? and Does chance agreement equal 50% for all situations? These questions highlight a common stumbling block for learners: understanding chance agreement when it is not equal to 50%. In appropriate settings (in a classroom or during a workshop), with motivated learners, the following example can be used to walk learners through the concepts of chance agreement.

The script

Start by introducing the following arbitrary scenario. "Let us assume that 2 'hopeless' clinicians are assessing the presence of Murphy's sign. However, they have no idea what they are doing, and their evaluations are no better than blind guesses. Let us say they are each guessing in a 50:50 ratio: half the time they guess that Murphy's sign is present and the other half that it is absent. If you were completing a 2 × 2 table, with these 2 clinicians evaluating the same 100 patients, how would the cells, on average, get filled in?"

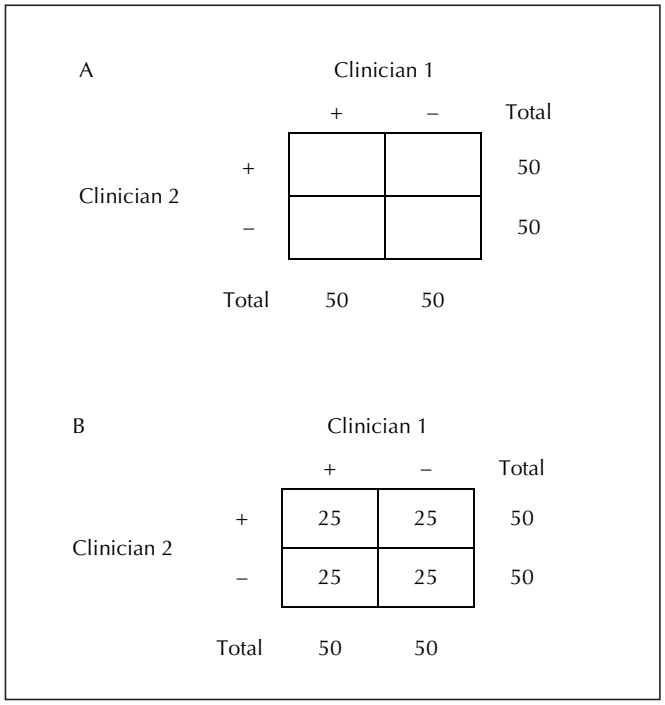

Fig. 2: The educator uses a 2 × 2 table as a worksheet to lead learners through the process of calculating chance agreement. (A) Blank 2 × 2 table. The educator may include or omit the total number of patients assessed as "positive" and "negative" by the 2 clinicians. When included, as here, these totals are commonly called "marginals." The examples shown here and in subsequent figures take advantage of the fact that, when the total number of patients is 100, each marginal equals the total of each row and each column and also the percentage of patients corresponding to that total. (B) Agreement table for 2 clinicians who randomly assess the presence or absence of a clinical finding, such as Murphy's sign, in the same 100 patients. Each assesses half of the patients as positive for the finding. The numbers in each box reflect the number of patients in each agreement category. The marginals are the totals of the patients in the respective rows and columns.

Then draw a 2 × 2 table on the board, with or without the marginals (Fig. 2A).

When a participant makes a suggestion, fill in the 4 cells accordingly, turn to the group, and ask them if they agree. Usually, it won't be long before someone provides the correct values (see Fig. 2B).

Then ask the group the proportion of patients for which the 2 hopeless clinicians have agreed. The first or second suggestion is usually the correct one: 50%. Record the first row of a simple table on the board, in an area where what is written can be retained and added to (Table 3).

Draw another empty 2 × 2 table and ask the participants to fill in the numbers for 2 clinicians who are once again "hopeless" (in that their guesses are no better than chance), but both guessing in a ratio of 60 positive to 40 negative.

At this point the participants begin to struggle. A frequent initial suggestion is that shown in Fig. 3A. Enter the suggested numbers and ask the group if they are correct. Even the person who made the suggestion is usually uncomfortable and not very confident. However, someone will come up with the correct values (Fig. 3B) in a matter of 3 minutes or so.

Make these entries in the table and ask the group if they are correct. A few people will begin to nod their heads in the affirmative, but others will look baffled. Ask the person who offered the correct answer how she arrived at that answer. Usually, the explanation will be reasonable, but neither articulate nor easy to understand. Restate the explanation as follows, adding the marginals to the table as you go along (Fig. 3C).

"Of the 60 patients whom clinician 1 guesses are positive [write in number for marginal *g*), clinician 2 will guess that 60% [write in number for marginal *e*] are positive. Sixty percent of 60 is 36 [point to cell *a*]. Of the same 60 patients whom clinician 1 guesses are positive [point to marginal *g*], clinician 2 will guess that 40% [write in number for marginal *f*] are negative. Forty percent of 60 is 24 [point to cell *c*]. Of the 40 patients whom clinician 1 guesses are negative [write in number for marginal *h*], clinician 2 will guess that 60% are positive [point to marginal *e*]. Sixty percent of 40

Table 3: Chance agreement when 2 observers randomly assign positive and negative results, for successively higher rates of a positive call

Proportion positive (%)	Agreement by chance (%)
50	50
60	52
70	58
80	68
90	82

is 24 [point to cell *b*]. Of the same 40 patients whom clinician 1 guesses are negative [point to marginal *b*], clinician 2 will guess that 40% are negative [point to marginal *f*]. Forty percent of 40 is 16 [point to cell *d*]."

Then ask the group what the chance agreement is in this situation, where both observers are guessing in a ratio of 60 to 40. When the group provides the answer, write this number in the second row of the evolving Table 3.

Repeat this exercise a few more times, having the 2 hopeless clinicians guess in ratios of 70 to 30, 80 to 20, and 90 to 10, until the group has no further difficulty with the calculation. As the group completes each example, fill in Table 3 accordingly. A common question that arises during this process is, "Why does chance agreement vary?" The simple answer is that the estimated prevalence varies. If both clinicians expect a low prevalence of disease, based on their experience, they will anticipate a high rate of normal results, which leads to a high rate of agreement due to chance alone.

Finally, summarize what the group has learned to this point: first, even if the clinicians have no idea what they are doing, there will be substantial agreement by chance alone, and second, the magnitude of the agreement by chance increases as the imbalance between positive and negative increases.

A

Clinician 2 \ Clinician 1	+	–
+	30	20
–	30	20

B

Clinician 2 \ Clinician 1	+	–
+	36	24
–	24	16

C

Clinician 2 \ Clinician 1	+	–	Total
+	*a* 36	*b* 24	*e* 60
–	*c* 24	*d* 16	*f* 40
Total	*g* 60	*h* 40	

Fig. 3: The same 2 clinicians randomly assess the presence or absence of the clinical finding in another 100 patients. This time, each assesses 60% of the patients as positive for the finding and 40% as negative. (A) An example of the entries that learners might guess for this situation; here, the entries are not in fact correct. (B) The correct agreement table for 2 clinicians guessing as to the presence or absence of a clinical finding, and assessing 60% of the patients as positive and 40% as negative). (C) The agreement table, with the marginals included, to help learners understand why the values are correct. See text for explanation of the letter designations.

Extension for advanced learners

At this point, you may ask the group if they are satisfied or if they wish to go one step further and calculate kappa. If they wish to go further, present them with another 2 × 2 table. Depending on the sophistication of the group, you can choose either an arithmetically simple table such as Fig. 4A or something a little more challenging such as Fig. 4B.

In either case, ask the group to determine the extent to which the 2 clinicians would have agreed by chance alone. Conceptually, this involves another step: realizing that agreement by chance is obtained by using the marginal numbers to determine the numbers of patients assessed as positive and negative by clinician 1 and then multiplying those by the percentages of corresponding assessments by clinician 2. This step may seem trivial, given the exercises that the group has just completed, but most group members usually find it challenging. If one group member figures out the solution ahead of the others (as often happens), you can ask that person to summarize her approach; you can then repeat, in clear fashion, the logic of the calculations. If the group seems to hit a brick wall at this point, proceed as follows.

"Well, how did we decide on the agreement by chance when the 2 observers were just guessing and the proportion in which they were guessing was 70:30? To find the agreement by chance in cell *a*, we multiplied 70 by 70%. And to find the agreement by chance in cell *d*, we multiplied 30 by 30%." While making these statements, you can refer to a table with the appropriate marginals and cell contents. Having completed this step, you then ask, "What would be the analogous process here?" Turn back to the table at hand and, with the group's help, add the appropriate numbers (Fig. 4C).

"What proportion has clinician 1 decided are positive?" A group member answers "50%," and you write that in the position for marginal *g*. "What proportion has clinician 1 decided are negative?" A group member answers "50%," and you write that in the position for marginal *h*. "What proportion has clinician 2 decided are positive?" A group member answers "50%," and you write that in the position for marginal *e*. "What proportion has clinician 2 decided are negative?" A group member answers "50%," and you write that in the position for marginal *f*.

"So, if these clinicians had simply been guessing, what number of observations would have ended up in cell *a*?" A

group member answers "25," and you write that number in parentheses in cell *a* (see Fig. 4C). "And again, if the clinicians were just guessing, what number of observations would have ended up in cell *d*?" A group member answers "25," and you write that number in parentheses in cell *d* (see Fig. 4C). "Now, what would the agreement by chance be in this instance?" A group member answers "50%." "What agreement did the clinicians actually achieve?" A group member answers "80%."

You can then explain how kappa is calculated: (Observed agreement [80%] – agreement expected by chance [50%]) ÷ (total possible agreement [always 100%] – agreement expected by chance [50%]). In this case, the final calculation of kappa is 30% ÷ 50% = 0.6. (See also the box on page 3.)

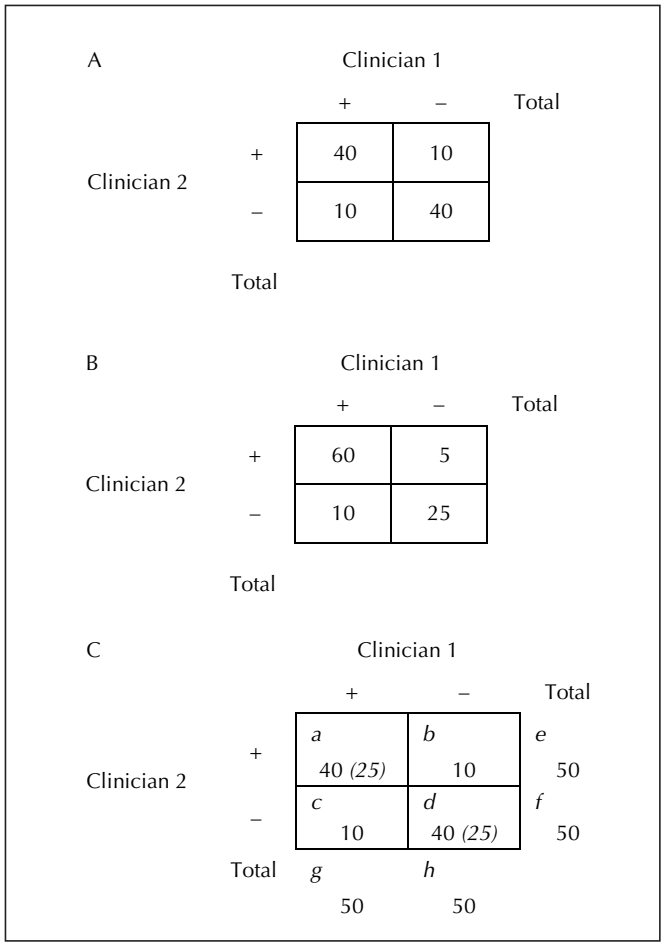

Fig. 4: The educator presents the learners with a situation in which the agreement between the 2 clinicians may be greater than chance. (A) A relatively simple example of agreement that is greater than chance. (B) A more complex example to illustrate a situation where agreement is greater than chance. (C) The correct version of the 2 × 2 table in Fig. 4A when the learners have correctly completed it. The values in parentheses correspond to the results that would be expected had each clinician randomly guessed that half of the patients had a positive result. See text for further explanation.

Thus, kappa expresses the proportion of possible agreement above and beyond chance that the clinicians have achieved. The parallel numbers for the arithmetically more difficult example shown in Fig. 4B are as follows: agreement observed = 60% + 25% = 85%, agreement by chance = 45.5% + 10.5% = 56%, and kappa = (85% – 56%) ÷ (100% – 56%) = 0.66. If necessary, both of these examples can be further clarified using the previously described bar diagram (Fig. 1).

The bottom line

- Stumbling block: Understanding chance agreement when it does not equal 50%.
- Learners can easily estimate the calculation of chance agreement and kappa by using 2 × 2 tables and increasing the percentage of positive findings by each observer.
- Give the learners several opportunities to struggle with calculating chance agreement, and use a running table to demonstrate how the estimate of chance agreement changes with prevalance.
- As the proportion of positive results increases, chance agreement also increases.
- Conclude the session by having the learners calculate chance agreement and then kappa.

See Appendix 1 for the summary card for this tip.

Report on field-testing

One of us (S.K.) field-tested these tips in June 2000 with 14 medical residents during a 1.5-hour teaching session. The leader of the session led the group through all of the tips, including the extension for tip 3, during one session. The learners were 7 new interns (naive learners), 1 intern with advanced exposure to the concept of kappa and 6 upper-level residents with moderate exposure.

The learners seemed to struggle most with the concept and calculation of chance agreement and with understanding how the estimation of chance agreement would change in different circumstances. The concept was clarified by explaining to the learners that chance is affected by the prevalence of the disease being diagnosed.[11]

The learners appeared most excited by the bar diagram (Fig. 1) used in tip 2, which seemed to clarify the concept of kappa most succinctly. They independently returned to using the bar diagram later on, to help perform the calculations needed during tip 3.

All participants rated most components of the exercise very highly. The 14 residents were asked to assign a score between 0 and 10 (hopeless to perfect) for the performance of each of the tips, and means were calculated for these ratings. They assigned a mean rating of 6.0 for the clarity and relevance of tip 1. Their mean ratings for tips 2 and 3 and for the extension to tip 3 ranged between 8.4 and 9.1. They

assigned a mean rating of 9.0 for the overall exercise with respect to promotion of understanding and retention.

Conclusions

The ability to both understand measures of interobserver agreement and calculate kappa from data presented in clinical trials and systematic reviews is an essential skill for clinicians. We have presented a series of tips previously developed and used by experienced teachers of evidence-based medicine for the purpose of overcoming learners' common pitfalls in acquiring these skills. The results of a field test of these tips by an independent teacher, also skilled in teaching evidence-based medicine to clinical learners but previously unfamiliar with these approaches, suggests that such educators may find this material useful in their own teaching.

This article has been peer reviewed.

From the Department of Medicine, Division of General Internal Medicine (McGinn), and the Department of Geriatrics (Leipzig), Mount Sinai Medical Center, New York, NY; the Columbia University College of Physicians and Surgeons, New York, NY (Wyer); the Departments of Epidemiology and Biostatistics and of Pediatrics, University of California, San Francisco, San Francisco, Calif. (Newman); Durham Veterans Affairs Medical Center and Duke University Medical Center, Durham, NC (Keitz); and the Departments of Medicine and of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont. (Guyatt)

Competing interests: None declared.

Contributors: Thomas McGinn developed the original idea for tips 1 and 2 and, as principal author, oversaw and contributed to the writing of the manuscript. Thomas Newman and Roseanne Leipzig reviewed the manuscript at all phases of development and contributed to the writing as coauthors. Sheri Keitz used all of the tips as part of a live teaching exercise and submitted comments, suggestions and the possible variations that are described 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 tip 3, reviewed the manuscript at all phases of development, contributed to the writing as a coauthor, and, as general editor, reviewed and revised the final draft of the manuscript to achieve accuracy and consistency of content.

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 [editorial]. *CMAJ* 2004;171(4):347-8.
2. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987;126:161-9.
3. Blackburn H. The exercise electrocardiogram: differences in interpretation. Report of a technical group on exercise electrocardiography. *Am J Cardiol* 1968;21:871-80.
4. Cook DJ. Clinical assessment of central venous pressure in the critically ill. *Am J Med* Sci 1990;299:175-8.
5. Aertgeerts B, Buntinx F, Fevery J, Ansoms S. Is there a difference between CAGE interviews and written CAGE questionaires? *Alcohol Clin Exp Res* 2000;24:733-6.
6. Kilpatrick R, Milne JS, Rushbrooke M, Wilson ESB. A survey of thyroid enlargement in two general practices in Great Britain. *BMJ* 1963;1:29-34.
7. Guyatt GH, Patterson C, Ali M, Singer J, Levine M, Turpie I, et al. Diagnosis of iron-deficiency anemia in the elderly. *Am J Med* 1990;88:205-9.
8. McCombe PF, Fairbank JC, Cockersole BC, Pynsent PB. 1989 Volvo Award in clinical sciences. Reproducibility of physical signs in low back pain. *Spine* 1989;14:908-18.
9. Perrier A, Howarth N, Didier D, Loubeyre P, Unger PF, de Moerloose P, et al. Performance of helical computed tomography in unselected outpatients with suspected pulmonary embolism. *Ann Intern Med* 2001;135:88-97.
10. Koelemay MJ, Legemate DA, Reekers JA, Koedam NA, Balm R, Jacobs MJ. Interobserver variation in interpretation of arteriography and management of severe lower leg arterial disease. *Eur J Vasc Endovasc Surg* 2001;21:417-22.
11. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543-9.

Correspondence to: *Dr. Peter C. Wyer, 446 Pelhamdale Ave., Pelham NY 10803, USA; fax 914 738-9368; 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, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for teachers of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. Available: www.cmaj.caicgi/content/full/171/4/353/DCI.

Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Guyatt G, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for teachers of evidence-based medicine: 2. Confidence intervals and p values. Available: www.cmaj.ca/cgi/content/full/171/6/611/DCI.

Appendix 1: Summary cards for 3 teaching tips on the kappa statistic

This appendix has been designed so that it can be printed on two sheets of 8 ½ × 11 inch paper. The individual summary cards can then be cut out, if desired, for use during teaching sessions.

Teaching tip 1: Understanding kappa

Scenario: Two radiologists read the same 100 mammograms. Reader 1 is barely paying attention and reads all 100 films as normal. Reader 2 reads the films more carefuly and identifies 4 of the 100 mammograms as suspicious for malignant disease. The percent agreement is 96%. The learners see that this clearly overestimates the level of clinically important agreement. The role of kappa is to assess the degree to which 2 observers agree beyond that expected by chance. The level of chance agreement is influenced by the estimated prevalence of the condition.

1. Review kappa scores and their implications:

0	No agreement beyond chance
0–0.2	Slight agreement beyond chance
0.2–0.4	Fair agreement beyond chance
0.4–0.6	Moderate agreement beyond chance
0.6–0.8	Substantial agreement beyond chance
0.8–1.0	Almost perfect agreement beyond chance

2. Examples of kappa scores for clinical correlation:

Interpretation of T wave changes on an exercise stress test	0.25
Assessment of jugular venous distension	0.56
Detection of alcohol dependence using CAGE questionnaire	0.75
Presence of goitre	0.82–0.95
Bone marrow interpretation by hematologist	0.84
Straight leg raising test	0.82
Diagnosis of pulmonary embolus by helical CT	0.82
Diagnosis of lower extremity arterial disease by arteriography	0.39–0.64

Summary points

- Stumbling block: Understanding the concept of agreement beyond chance and why it is important to correct for chance agreement.
- This tip leads learners to understand that kappa allows us to measure agreement above and beyond that expected by chance alone.
- Concrete examples of kappa scores for frequently ordered tests help to anchor this statistic in clinical practice and stimulate the learners to move on to more complex concepts related to kappa.

Teaching tip 2: Calculating kappa

Scenario: A simple bar diagram can help in explaining how kappa scores are derived.

1. Ask the learners to determine the total potential for agreement between 2 observers (100%), and draw a horizontal bar to represent this 100% agreement.
2. Tell the group that for a hypothetical situation, the estimated chance agreement between the 2 observers is 50%. Draw a second horizontal bar, representing 50% agreement.
3. Tell the group that the observed agreement is 75%, and draw a third bar representing this 75% agreement.
4. Ask the group to identify the possible agreement beyond chance (25%). Draw a vertical line across the 3 bars at the point of chance agreement (50%), whereby agreement to the right of the vertical line is agreement beyond chance.
5. Tell the group that the kappa score in this instance is 0.5. Then ask the group to figure out how this score was calculated.
6. Ask the group if they agree that a kappa score of 0.4 to 0.6 represents a moderate degree of agreement.

Summary points

- Stumbling block: Understanding how kappa scores are derived.
- A simple graphic can be used to demonstrate the derivation of kappa.
- The graphic allows learners to begin to see how kappa is calculated. The visual image helps learners to overcome the common stumbling block more easily than would be the case if they merely worked through the numbers.

Teaching tip 3: Calculating chance agreement

Scenario: Two observers are assessing the presence of Murphy's sign in a patient, but their assessments are no better than blind guesses. A series of demonstrations with 2×2 tables is performed with successive positive call rates of 50%, 60%, 70%, 80% and 90%. Consider how chance agreement increases with the positive call rate.

1. Draw a 2×2 table, writing "Clinician 1" and "Clinician 2" along the top and left side respectively.
2. Tell the learners that the clinicians guess that Murphy's sign is present half the time.
3. Ask the learners to fill the cells of the 2×2 table. The number 25 should appear in each cell.
4. Ask the learners how often the reviewers agree (cells $a + d = 50\%$). This is the level of chance agreement.
5. Repeat the exercise with a new 2×2 table. Ask the learners to fill in the numbers when the 2 clinicians both guess that 60% of cases are positive and 40% are negative. The cells representing agreement (a and d) will have the numbers 36 and 16 respectively. Cells b and c will both be 24. Cells $a + d = 52\%$. Again, this is the level of chance agreement.
6. Repeat this exercise, increasing the proportion of positive assessments, until 90% of the assessments are positive for both clinicians.

Proportion positive, %	Agreement by chance, %
50	50
60	52
70	58
80	68
90	82

Summary points

- Stumbling block: Understanding chance agreement when it does not equal 50%.
- Learners can easily estimate the calculation of chance agreement and kappa by using 2×2 tables and increasing the percentage of positive findings by each observer.
- Give the learners several opportunities to struggle with calculating chance agreement, and use a running table to demonstrate how the estimate of chance agreement changes with prevalence.
- As the proportion of positive results increases, chance agreement also increases.
- Conclude the session by having the learners calculate chance agreement and then kappa.