Tips for learners of evidence-based medicine: 3. Measures of observer variability (kappa statistic)

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

Imagine that you're a busy family physician and that you've found a rare free moment to scan the recent literature. Reviewing your preferred digest of abstracts, you notice a study comparing emergency physicians' interpretation of chest radiographs with radiologists' interpretations.[1] The article catches your eye because you have frequently found that your own reading of a radiograph differs from both the official radiologist reading and an unofficial reading by a different radiologist, and you've wondered about the extent of this disagreement and its implications.

Looking at the abstract, you find that the authors have reported the extent of agreement using the κ statistic. You recall that κ stands for "kappa" and that you have encountered this measure of agreement before, but your grasp of its meaning remains tentative. You therefore choose to take a quick glance at the authors' conclusions as reported in the abstract and to defer downloading and reviewing the full text of the article.

Practitioners, such as the family physician just described, may benefit from understanding measures of observer variability. 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.

In this article we present tips aimed at helping clinical learners to use the concepts of kappa when applying diagnostic tests in practice. The tips presented here have been adapted from approaches developed by educators experienced in teaching evidence-based medicine skills to clinicians.[2] A related article, intended for people who teach these concepts to clinicians, is available online at www.cmaj.ca/cgi/content/full/171/11/1369/DC1.

Clinician learners' objectives

Defining the importance of kappa

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

Calculating kappa

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

Calculating chance agreement

- 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.

Tip 1: Defining the importance of kappa

A common stumbling block for clinicians is the basic concept of agreement beyond chance and, in turn, the importance of correcting for chance agreement. People making a decision on the basis of 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 this chance agreement and tells us how much of the possible agreement over and above chance the reviewers have achieved.

A simple example should help to clarify the importance of correcting for 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

Teachers of evidence-based medicine:

See the "Tips for teachers" version of this article online at www.cmaj.ca/cgi/content/full/171/11/1369/DC1. It contains the exercises found in this article in fill-in-the-blank format, commentaries from the authors on the challenges they encounter when teaching these concepts to clinician learners and links to useful online resources.

DOI:10.1503/cmaj.1031981

films more carefully and identifies 4 of the 100 mammograms as positive (suspicious for malignancy). How would you characterize the level of agreement between these 2 radiologists?

The percent agreement between them is 96%, even though one of the readers has, on cursory review, decided to call all of the results negative. Hence, measuring the simple percent agreement overestimates the degree of clinically important agreement in a fashion that is misleading. The role of kappa is to indicate how much the 2 observers agree beyond the level of agreement that could be expected by chance. Table 1 presents a rating system that is commonly used as a guideline for evaluating kappa scores. Purely to illustrate the range of kappa scores that readers can expect to encounter, Table 2 gives some examples of commonly reported assessments and the kappa scores that resulted when investigators studied their reproducibility.

The bottom line

If clinicians neglect the possibility of chance agreement, they will come to misleading conclusions about the reproducibility of clinical tests. The kappa statistic allows us to measure agreement above and beyond that expected by chance alone. Examples of kappa scores for frequently ordered tests sometimes show surprisingly poor levels of agreement beyond chance.

Table 1: Qualitative classification of kappa values as degree of agreement beyond chance[3]

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

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

Assessment	Kappa value
Interpretation of T wave changes on an exercise stress test[4]	0.25
Presence of jugular venous distension[5]	0.56
Detection of alcohol dependence using CAGE questionnaire[6]	0.75
Presence of goitre[7]	0.82–0.95
Bone marrow interpretation by hematologist[8]	0.84
Straight leg raising test[9]	0.82
Diagnosis of pulmonary embolus by helical CT[10]	0.82
Diagnosis of lower extremity arterial disease by arteriography[11]	0.39–0.64

Tip 2: Calculating kappa

What is the maximum potential for agreement between 2 observers doing a clinical assessment, such as presence or absence of Murphy's sign in patients with abdominal pain? In Fig. 1, the upper horizontal bar represents 100% agreement between 2 observers. For the hypothetical situation represented in the figure, the estimated chance agreement between the 2 observers is 50%. This would occur if, for example, each of the 2 observers randomly called half of the assessments positive. Given this information, what is the possible agreement beyond chance?

The vertical line in Fig. 1 intersects the horizontal bars at the 50% point that we identified as the expected agreement by chance. All agreement to the right of this line corresponds to agreement beyond chance. Hence the maximum agreement beyond chance is 50% (100% – 50%).

The other number you need to calculate the kappa score is the degree of agreement beyond chance. The observed agreement, as shown by the lower horizontal bar in Fig. 1, is 75%, so the degree of agreement beyond chance is 25% (75% – 50%).

Kappa is calculated as the observed agreement beyond chance (25%) divided by the maximum agreement beyond chance (50%); here, kappa is 0.50.

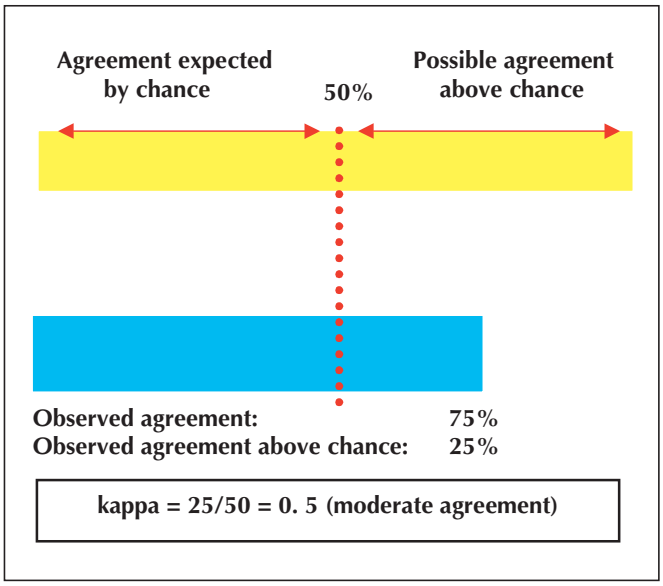

Fig. 1: Two observers independently assess the presence or absence of a finding or outcome. Each observer determines that the finding is present in exactly 50% of the subjects. Their assessments agree in 75% of the cases. The yellow horizontal bar represents potential agreement (100%), and the turquoise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

The bottom line

Kappa allows us to measure agreement above and beyond that expected by chance alone. We calculate kappa by estimating the chance agreement and then comparing the observed agreement beyond chance with the maximum possible agreement beyond chance.

Tip 3: Calculating chance agreement

A conceptual understanding of kappa may still leave the actual calculations a mystery. The following example is intended for those who desire a more complete understanding of the kappa statistic.

Let us assume that 2 hopeless clinicians are assessing the presence of Murphy's sign in a group of patients. They have no idea what they are doing, and their evaluations are no better than blind guesses. Let us say they are each guessing the presence and absence of Murphy's sign 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 represents the completed 2×2 table. Guessing at random, the 2 hopeless clinicians have agreed on the assessments of 50% of the patients. How did we arrive at the numbers shown in the table? According to the laws of chance, each clinician guesses that half of the 50 patients assessed as positive by the other clinician (i.e., 25 patients) have Murphy's sign.

How would this exercise work if the same 2 hopeless clinicians were to randomly guess that 60% of the patients had a positive result for Murphy's sign? Fig. 3 provides the answer in this situation. The clinicians would agree for 52 of the 100 patients (or 52% of the time) and would disagree for 48 of the patients. In a similar way, using 2×2 tables for higher and higher positive proportions (i.e., how often the observer makes the diagnosis), you can figure out how often the observers will, on average, agree by chance alone (as delineated in Table 3).

At this point, we have demonstrated 2 things. First, even if the reviewers have no idea what they are doing, there will be substantial agreement by chance alone. Second, the magnitude of the agreement by chance increases as the proportion of positive (or negative) assessments increases.

But how can we calculate kappa when the clinicians whose assessments are being compared are no longer "hopeless," in other words, when their assessments reflect a level of expertise that one might actually encounter in practice? It's not very hard.

Let's take a simple example, returning to the premise that each of the 2 clinicians assesses Murphy's sign as being present in 50% of the patients. Here, we assume that the 2 clinicians now have some knowledge of Murphy's sign and their assessments are no longer random. Each decides that 50% of the patients have Murphy's sign and 50% do not, but they still don't agree on every patient. Rather, for 40 patients they agree that Murphy's sign is present, and for 40 patients they agree that Murphy's sign is absent. Thus, they agree on the diagnosis for 80% of the patients, and they disagree for 20% of the patients (see Fig. 4A). How do we calculate the kappa score in this situation?

Recall that if each clinician found that 50% of the patients had Murphy's sign but their decision about the presence of the sign in each patient was random, the clinicians would be in agreement 50% of the time, each cell of the 2×2 table would have 25 patients (as shown in Fig. 2), chance agree-

		Clinician 1		
		Sign present	Sign absent	Total
Clinician 2	Sign present	25	25	50
	Sign absent	25	25	50
	Total	50	50	

Fig. 2: Agreement table for 2 hopeless clinicians who randomly guess whether Murphy's sign is present or absent in 100 patients with abdominal pain. Each clinician determines that half of the patients have a positive result. The numbers in each box reflect the number of patients in each agreement category.

		Clinician 1		
		Sign present	Sign absent	Total
Clinician 2	Sign present	36	24	60
	Sign absent	24	16	40
	Total	60	40	

Fig. 3: As in Fig. 2, the 2 clinicians again guess at random whether Murphy's sign is present or absent. However, each clinician now guesses that the sign is present in 60 of the 100 patients. Under these circumstances, of the 60 patients for whom clinician 1 guesses that the sign is present, clinician 2 guesses that it is present in 60%; 60% of 60 is 36 patients. Of the 60 patients for whom clinician 1 guesses that the sign is present, clinician 2 guesses that it is absent in 40%; 40% of 60 is 24 patients. Of the 40 patients for whom clinician 1 guesses that the sign is absent, clinician 2 guesses that it is present in 60%; 60% of 40 is 24 patients. Of the 40 patients for whom clinician 1 guesses that the sign is absent, clinician 2 guesses that it is absent in 40%; 40% of 40 is 16 patients.

ment would be 50%, and maximum agreement beyond chance would also be 50%.

The no-longer-hopeless clinicians' agreement on 80% of the patients is therefore 30% above chance. Kappa is a comparison of the observed agreement above chance with the maximum agreement above chance: 30%/50% = 60% of the possible agreement above chance, which gives these clinicians a kappa of 0.6, as shown in Fig. 4B.

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

A

		Clinician 1: Sign present	Clinician 1: Sign absent
Clinician 2	Sign present	40	10
	Sign absent	10	40

B

		Clinician 1: Sign present	Clinician 1: Sign absent	Total
Clinician 2	Sign present	40 (*25*)	10 (*25*)	50
	Sign absent	10 (*25*)	40 (*25*)	50
	Total	50	50	

κ = (observed agreement – agreement expected by chance) ÷ (100 – agreement expected by chance)
= (80% – 50%) ÷ (100% – 50%)
= 30% ÷ 50%
= 0.6

Fig. 4: Two clinicians who have been trained to assess Murphy's sign in patients with abdominal pain do an actual assessment on 100 patients. A: A 2 × 2 table reflecting actual agreement between the 2 clinicians. **B:** A 2 × 2 table illustrating the correct approach to determining the kappa score. The numbers in parentheses correspond to the results that would be expected were each clinician randomly guessing that half of the patients had a positive result (as in Fig. 2).

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)

Hence, to calculate kappa when only 2 alternatives are possible (e.g., presence or absence of a finding), you need just 2 numbers: the percentage of patients that the 2 assessors agreed on and the expected agreement by chance. Both can be determined by constructing a 2 × 2 table exactly as illustrated above.

The bottom line

Chance agreement is not always 50%; rather, it varies from one clinical situation to another. When the prevalence of a disease or outcome is low, 2 observers will guess that most patients are normal and the symptom of the disease is absent. This situation will lead to a high percentage of agreement simply by chance. When the prevalence is high, there will also be high apparent agreement, with most patients judged to exhibit the symptom. Kappa measures the agreement after correcting for this variable degree of chance agreement.

Conclusions

Armed with this understanding of kappa as a measure of agreement between different observers, you are able to return to the study of agreement in chest radiography interpretations between emergency physicians and radiologists[1] in a more informed fashion. You learn from the abstract that the kappa score for overall agreement between the 2 classes of practitioners was 0.40, with a 95% confidence interval ranging from 0.35 to 0.46. This means that the agreement between emergency physicians and radiologists represented 40% of the potentially achievable agreement beyond chance. You understand that this kappa score would be conventionally considered to represent fair to moderate agreement but is inferior to many of the kappa values listed in Table 2. You are now much more confident about going to the full text of the article to review the methods and assess the clinical applicability of the results to your own patients.

The ability to understand measures of variability in data presented in clinical trials and systematic reviews is an important skill for clinicians. We have presented a series of tips developed and used by experienced teachers of evidence-based medicine for the purpose of facilitating such understanding.

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 article. 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. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department: Is the radiologist really necessary? *Postgrad Med J* 2003;79:214-7.
2. 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.
3. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987;126:161-9.
4. Blackburn H. The exercise electrocardiogram: differences in interpretation. Report of a technical group on exercise electrocardiography. *Am J Cardiol* 1968;21:871-80.
5. Cook DJ. Clinical assessment of central venous pressure in the critically ill. *Am J Med Sci* 1990;299:175-8.
6. Aertgeerts B, Buntinx F, Fevery J, Ansoms S. Is there a difference between CAGE interviews and written CAGE questionnaires? *Alcohol Clin Exp Res* 2000;24:733-6.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.

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