Where Do We Go Wrong in Assessing Risk Factors, Diagnostic and Prognostic Tests? The Problems of Two-by-two Association

Helena Chmura Kraemer, PhD; and Robert D. Gibbons, PhD

Psychiatric Annals, Volume 39, Issue 7, July 2009

CME EDUCATIONAL OBJECTIVES

1. Review two-by-two contingency tables.
2. Express methods for assessing the strength of association between two binary variables.
3. Explain design, sampling, and estimation issues in assessing risk, prognostic, and diagnostic factors.

ABOUT THE AUTHOR

Helena Chmura Kraemer, PhD, is Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California. Robert D. Gibbons, PhD, is with the Center for Health Statistics, University of Illinois at Chicago.

Address correspondence to: Helena Chmura Kraemer, PhD, Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, 1116 Forest Avenue, Palo Alto, CA 94301; or e-mail hckhome@pacbell.net.

Dr. Kraemer has disclosed no relevant financial relationships. Dr. Gibbons disclosed the following relevant financial relationships: National Institutes of Mental Health: research grant recipient (R56-MH078580 and R01-MH8012201).

doi: 10.3928/00485713-20090625-05

PARTICIPANT ATTESTATION

___ I certify that I have read the article(s) on which this activity is based, and claim credit commensurate with the extent of my participation.

COMMERCIAL BIAS EVALUATION

Please rate the degree to which the content presented in this activity was free from commercial bias. No bias Significant bias
5 4 3 2 1

Comments regarding commercial bias: ______________________________
____

INSTRUCTIONS

1. Review the stated learning objectives of the CME articles and determine if these objectives match your individual learning needs.

2. Read the articles carefully. Do not neglect the tables and other illustrative materials, as they have been selected to enhance your knowledge and understanding.

3. The following quiz questions have been designed to provide a useful link between the CME articles in the issue and your everyday practice. Read each question, choose the correct answer, and record your answer on the CME REGISTRATION FORM at the end of the quiz. Retain a copy of your answers so that they can be compared with the correct answers should you choose to request them.

4. Type your full name and address and your date of birth in the space provided on the CME REGISTRATION FORM.

5. Complete the evaluation portion of the CME REGISTRATION FORM. Forms and quizzes cannot be processed if the evaluation portion is incomplete. The evaluation portion of the CME REGISTRATION FORM will be separated from the quiz upon receipt at PSYCHIATRIC ANNALS. Your evaluation of this activity will in no way affect the scoring of your quiz.

6. Your answers will be graded, and you will be advised whether you have passed or failed. Unanswered questions will be considered incorrect. A score of at least 80% is required to pass. Your certificate will be mailed to you at the mailing address provided. Upon receiving your grade, you may request quiz answers. Contact our customer service department at (856) 994-9400.

7. Be sure to complete the CME REGISTRATION FORM on or before July 31, 2010. After that date, the quiz will close. Any CME REGISTRATION FORM received after the date listed will not be processed.

8. This activity is to be completed and submitted online only.

Indicate the total time spent on the activity (reading article and completing quiz). Forms and quizzes cannot be processed if this section is incomplete. All participants are required by the accreditation agency to attest to the time spent completing the activity.

CME ACCREDITATION

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. There are no specific background requirements for participants taking this activity. Learning objectives are found at the beginning of each CME article.

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and PSYCHIATRIC ANNALS. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits™*. Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the relevant financial relationships of the planners, teachers, and authors involved in the development of CME content. An individual has a **relevant financial relationship** if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

HOW TO OBTAIN CME CREDITS BY READING THIS ISSUE

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. Physicians can receive *AMA PRA Category 1 Credits*™ by reading the CME articles in PSYCHIATRIC ANNALS and successfully completing the quiz at the end of the articles. Complete instructions are given subsequently. Educational objectives are found at the beginning of each CME article.

CME ACCREDITATION

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and PSYCHIATRIC ANNALS. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits*™. Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the **relevant financial relationships** of the planners, teachers, and authors involved in the development of CME content. An individual has a relevant financial relationship if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

Financial Disclosures: Stanley Caroff, MD, has disclosed no relevant financial relationships. John M. Davis, MD, has disclosed no relevant financial relationships. Jan Fawcett, MD, has disclosed the following relevant financial relationships: Merck: Member of Editorial Advisory Board, Merck Manual. Paula Hensley, MD, has disclosed no relevant financial relationships. Andrew A. Nierenberg, MD, has disclosed the following relevant financial relationships: Abbott, Brain Cells Inc., Bristol Myers Squibb (BMS), Eli Lilly, GlaxoSmithKline (GSK), Innapharma, PGx, Ortho-McNiel Janssen, Novartis, Pfizer, Sepracor, Shire, and Somerset: Member of Advisory Board/Consultant: Massachusetts General Hospital (MGH) Structured Interview Guide to the Montgomery-Åsberg Rating Scale (MADRS), and the Clinical Positive Affect Scale, licensed exclusively to the MGH Clinical Trials Network and Institute. Copyright Holder: BMS, Cederroth, Cyberonics, Eli Lilly, Forest, GSK, Ortho-McNeil Janssen; Lichtwer, NARSAD, National Institute of Mental Health; Pfizer, Stanley Foundation, Wyeth: Research Grant Recipient; and BMS, Cyberonics, Eli Lilly, Forest, GSK, Wyeth: Member of Speakers' Bureau. The staff of *Psychiatric Annals* have disclosed no relevant financial relationships.

INSTRUCTIONS

1. Review the stated learning objectives of the CME articles and determine if these objectives match your individual learning needs.

2. Read the articles carefully. Do not neglect the tables and other illustrative materials, as they have been selected to enhance your knowledge and understanding.

3. The following quiz questions have been designed to provide a useful link between the CME articles in the issue and your everyday practice. Read each question, choose the correct answer, and record your answer on the CME REGISTRATION FORM at the end of the quiz. Retain a copy of your answers so that they can be compared with the correct answers should you choose to request them.

4. Type or print your full name and address and your date of birth in the space provided on the CME REGISTRATION FORM.

5. Complete the evaluation portion of the CME REGISTRATION FORM. Forms and quizzes cannot be processed if the evaluation portion is incomplete. The evaluation portion of the CME REGISTRATION FORM will be separated from the quiz upon receipt at PSYCHIATRIC ANNALS. Your evaluation of this activity will in no way affect the scoring of your quiz.

6. Send the completed form, with your $25 payment (check, money order, or credit card information) to: VINDICO MEDICAL EDUCATION, PO Box 36, Thorofare NJ 08086. Payment should be made in US dollars drawn on a US bank.

7. Your answers will be graded, and you will be advised whether you have passed or failed. Unanswered questions will be considered incorrect. A score of at least 80% is required to pass. Upon receiving your grade, you may request quiz answers. Contact our customer service department at (856) 994-9400.

8. Be sure to mail the CME REGISTRATION FORM on or before the deadline listed. After that date, the quiz will close. Any CME REGISTRATION FORM received after the date listed will not be processed.

Indicate the total time spent on the activity (reading article and completing quiz). Forms and quizzes cannot be processed if this section is incomplete. All participants are required by the accreditation agency to attest to the time spent completing the activity.

EDUCATIONAL OBJECTIVES OVERVIEW

If any readers of *Psychiatric Annals* read any psychiatric articles in any journal so that they can keep up to date and use the best-available evidence in their clinical practice, then the statistical papers in this issue are essential not only to read, but also to re-read, study, and master. What do these excellent statistical papers have to do with clinical psychiatric practice? They contain the tools that will allow you to critically appraise and interpret the psychiatric literature. They demystify the statistics used to generate the evidence and help you to become statistically literate. They give you the tools to turn data into information and knowledge.

TABLE OF CONTENTS

RESPONSIBILITY FOR STATEMENTS

All opinions expressed by authors and quoted sources are their own and do not necessarily reflect the opinions of the editors, publishers, or editorial boards of *Psychiatric Annals* or its employees, Vindico Medical Education or its employees, or the University of New Mexico. The acceptance of advertising in no way implies endorsement by the editors, publishers, or editorial boards of *Psychiatric Annals* .

The material presented at or in any *Psychiatric Annals* or Vindico Medical Education continuing education activity does not necessarily reflect the views and opinions of Vindico Medical Education or *Psychiatric Annals*. Neither *Psychiatric Annals*, Vindico Medical Education, nor the faculty endorse or recommend any techniques, commercial products, or manufacturers. The faculty/authors may discuss the use of materials and/or products that have not yet been approved by the U.S. Food and Drug Administration. Articles are intended for informational purposes only and should not be used as the basis of patient treatment. All readers and continuing education participants should verify all information before treating patients or utilizing any product.

CME

Where Do We Go Wrong in Assessing Risk Factors, Diagnostic and Prognostic Tests?
The Problems of Two-by-two Association

Helena Chmura Kraemer, PhD; and Robert D. Gibbons, PhD

CME **EDUCATIONAL OBJECTIVES**

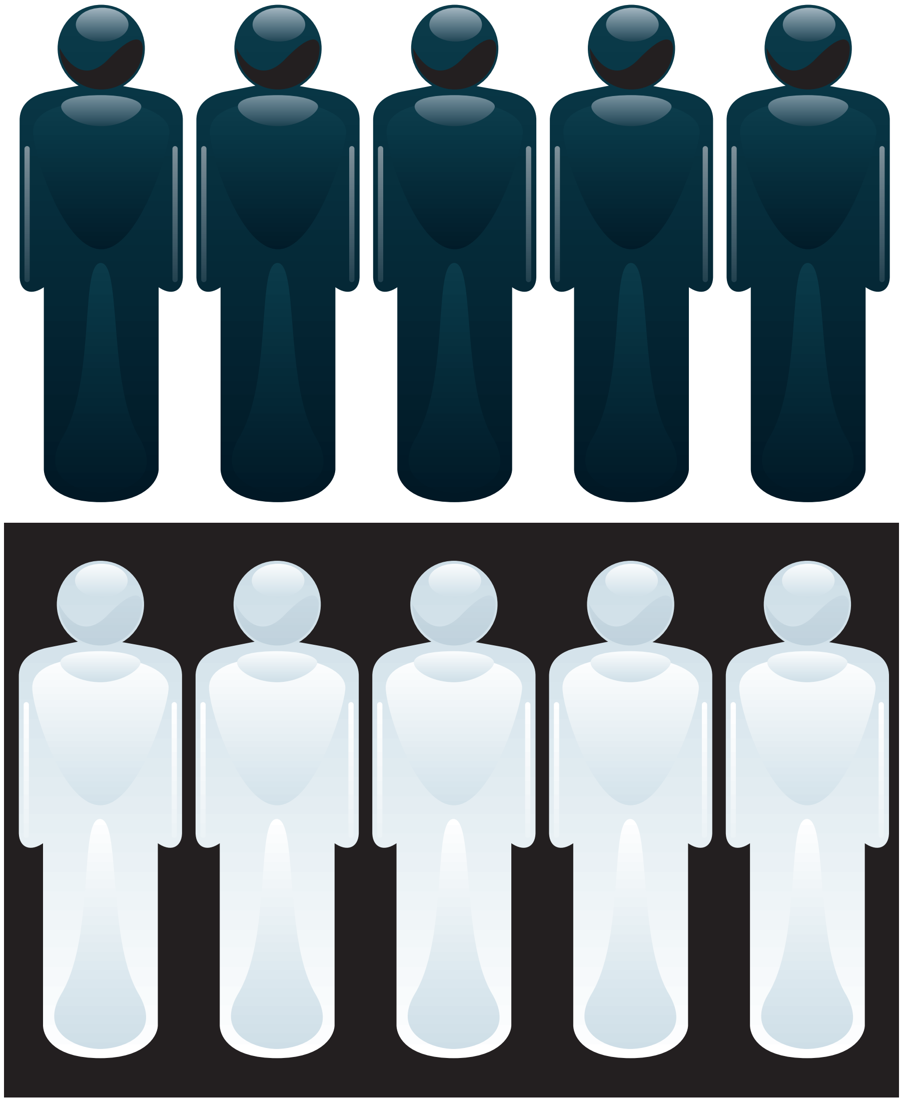
1. Review two-by-two contingency tables.
2. Express methods for assessing the strength of association between two binary variables.
3. Explain design, sampling, and estimation issues in assessing risk, prognostic, and diagnostic factors.

Helena Chmura Kraemer, PhD, is Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California. Robert D. Gibbons, PhD, is with the Center for Health Statistics, University of Illinois at Chicago.

Address correspondence to: Helena Chmura Kraemer, PhD, Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, 1116 Forest Avenue, Palo Alto, CA 94301; or e-mail hckhome@pacbell.net.

Dr. Kraemer has disclosed no relevant financial relationships. Dr. Gibbons disclosed the following relevant financial relationships: National Institutes of Mental Health: research grant recipient (R56-MH078580 and R01-MH8012201).

doi: 10.3928/00485713-20090625-05



© iStockphoto.com

A vital issue in medical research is that of assessing the relationship between two binary variables: examining how well the presence/absence of a certain factor (eg, smoking, obesity) predicts a future event (prognostic test); how well a test result (eg, a blood, urine, skin, or imaging test) discriminates between those who do and do not have a disorder (a diagnostic test); and how well a decision rule (based on genes, gender, age, ethnicity, etc.) discriminates between those who will or will not succeed, etc.

The core of such problems is that there are two binary variables, T and D, where T is intended to be positively related to D in some population, and, where observing T may lead to certain decisions appropriate to D, that, if wrong, may do harm. The "test" T is evaluated by considering what happens when the test and "diagnosis" D are applied ("blinded" to each other) to the subjects in a population of interest and the results are compiled into a two-by-two table (see Table, page 713).

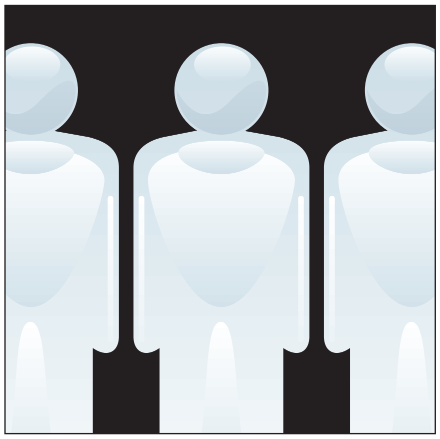
In what follows, we will define the language medical researchers commonly use to describe the association between T and D in the population, and consider and compare common measures of two-by-two association, and how they are estimated and interpreted for positive association. The purpose of doing so is to review such methods, but more so, to point out many common errors made in this context that may cost the health, well-being, and perhaps even lives of patients.

THE BASIC TERMS

In the two-by-two table, there are four outcomes: TP represents the probability of a "true-positive result," FN that of a "false-negative result," FP that of a "false-positive result," and TN that of a "true-negative result."

The probability that D = 1 is the "baserate" P, which, depending on how the outcome is defined, may be a prevalence or an incidence in the population of interest. The probability that T = 1, is called the "level" Q. The level of a test is determined by how loosely or stringently the criteria for a positive test are set, and often can be varied by the decision of the researchers.

When the results in the two-by-two table are consistent with random decision making, TP = P•Q, FN = P•Q', FP = P'•Q, FN = P'•Q'. Rational and high-quality decision making requires that one do a lot better. Ideally, of course, one would want to make no wrong decisions, ie, that TP = P = Q and TN = P' = Q', and FN = FP = 0, but this can seldom happen in medical applications because the criterion D is seldom measured with perfect reliability.



Rational and high-quality decision making requires that one do a lot better.

Perhaps the sole issue concerning two-by-two tables on which everyone tends to agree is how to test the null hypothesis of random association between T and D: some variation of the two-by-two chi-square test. However, our focus is not on "proving" non-randomness. By the time there is scientific rationale and justification for proposing such a test, the null hypothesis of randomness will seldom be absolutely true.[1,2] Any non-random association (no matter how weak) is likely to be "statistically significant" at whatever level specified, provided only that the sample size is large enough. The question at issue here is one of assessing clinical significance. That's where the disagreements begin, largely for reasons having to do with the now well-recognized problems with the use of statistical significance tests.[2-11]

SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUES

Sensitivity (Se), specificity (Sp), and the predictive values of a positive and negative test (PVP and PVN) are commonly used measures to describe the association between T and D and are defined in the Table (see page 713). It still occasionally happens in the medical literature that these terms are interchanged, or that different names are used. For example, the "true positive rate" might refer to TP, or to sensitivity, or to PVP. The definitions presented here are standard definitions, but readers of the medical literature must still check the definitions given in each text.

Only if P = Q is Se = PVP and Sp = PVN, and only if P = Q = 1/2 are all four of these probabilities the same. Generally, Se, Sp, PVP, and PNV are different numbers and can vary widely from each other.

All the probabilities (TP, FN, FP, TN) within the two-by-two table can be expressed in terms of (P, Se, Sp) or in terms of (Q, PVP, PVN) (see Table, page 713). Thus, Se and Sp can be expressed in terms of Q, PVP and PVN, and PVP and PVN can be expressed in terms of P, Se, Sp (Bayes' theorem). For example, applying the definition of PVP but using the cell probabilities expressed in terms of P, Se and Sp:

$$\mathrm{PVP} = \frac{\mathrm{PSe}}{\mathrm{PSe} + \mathrm{P'Sp'}}$$

It is an easy matter to get either sensitivity or specificity near 1. Give everyone a positive test (T = 1) to get a sensitiv-

TABLE.

Population Definitions of the Basic Probabilities Used in the Evaluation of a Test (T) against a Criterion (D) and Definitions of Common Effect Sizes with their Relationship to the Weighted Kappa, k(w)

	T = 1	T = 0	Total
D = 1	TP (true positives) = P Se = Q PVP	FN (false negatives) = P Se’ = Q’ PVN’	P = baserate
D = 0	FP (false positives) = P’ Sp’ = Q PVP’	TN (true negatives) = P’ Sp = Q’ PVN	P’ = 1-P
Total	Q = level	Q’ = 1-Q	1

Weighted Kappa:

$$k(w) = \frac{TP*TN - FP*FN}{PQ'w+P'Qw'} = PQ'wk(1)+P'Qw'k(0)$$

$O \leq wM \leq$, $w' = 1-w$

Measure	Relationship to Weighted Kappa
Sensitivity = Se = TP/P	k(1) = (Se-P)/P’
Specificity = Sp = TN/P’	k(0)=(Sp-P’)/P
Predictive value of a positive test = PVP = TP/Q	k(0) = (PVP-P)/P’
Predictive value of a negative test = PVN = TN/Q’	k(1) = (PVN-P’)/P
Risk Differences	
RD1 = Se + Sp-1	k(P’) = RD1
RD2 = PVP + PVN-1	k(Q) = RD2
Risk Ratios	
RR1 = Se/(1-Sp)	k(0) = (RR1-1)/(RR1 + P’/P)
RR2 = Sp/(1-Se)	k(1) = (RR2-1)/(RR2 + P/P’)
RR3 = PVP/(1-PVN)	k(1) = (RR3-1)/(RR3 + Q’/Q)
RR4 = PVN/(1-PVP)	k(0) = (RR4-1)/(RR4 + Q/Q’)
Phi = (TP TN-FN FP)/(PP’QQ’)$^{1/2}$	k(w*) = Phi = (k(0)k(1))$^{1/2}$; w*= [(PP’QQ’)$^{1/2}$-P’Q]/(PQ’-P’Q)

Odds Ratio = (TP TN)/(FN FP)

ity of 1.0; give everyone a negative test result (T = 0) for a specificity of 1.0. A high sensitivity or specificity or PVP or PVN alone does not guarantee the accuracy of the test. The trick is to get a test with both high enough sensitivity and high enough specificity, or both predictive values high enough, and that is not easy, particularly since “high enough” is so vaguely defined.

With random decision making, Se = Q, Sp = Q’, PVP=P, PVN=P’. Thus all four of these indices (Se, Sp, PVP, PVN) are uncalibrated measures [ie, their random value depends on either the baserate (P) or the level (Q)]. Yet it still happens in the literature that only specificity, or only sensitivity is reported, which is somewhat like reporting a temperature, but forgetting to report whether the scale was Fahrenheit, Celsius, Kelvin, or on some other temperature scale.

We can standardize these four measures, exactly as we would recalibrate Fahrenheit or Kelvin readings to the standard Celsius scale, treating random as the “freezing point,” perfect decision making as the “boiling point.” Recalibrated sensitivity is: (Se-Q)/Q’; recalibrated specificity is (Sp-Q’)/Q; recalibrated PVP is (PVP-P)/P’; recalibrated PVN is (PVN-P’)/P. Then, it turns out that rescaled sensitivity and PVN are exactly the same number, as are rescaled specificity and PVP. Those who choose to report sensitivity/specificity and those who choose to report predictive values are in essence reporting the same information, measured on scales calibrated differently. Nevertheless, we still have at least two different numbers that might convey different messages about the association between T and D. Which of these two numbers best reflects the quality of a test for a diagnosis? The answer to that question lies in the consideration of a crucial issue often ignored in evaluation of medical decision-making, the relative clinical importance of the two types of errors, false positives, and false negatives.

WEIGHTING THE CLINICAL IMPORTANCE OF THE TWO TYPES OF ERRORS: WEIGHTED KAPPA

In 1968,[12] Cohen noted that in the two-by-two situation, the costs or risks associated with two types of errors in any population are often very different. He introduced an index that incorporates consideration of such relative costs into the assessment of the association: the weighted kappa k(w).

The weight in this kappa, w, is some number between 0 and 1 that reflects the relative importance of avoiding false negatives to avoiding false positives.[13,14] How much difference does it make to those who will have the outcome whether or not they have a positive test (avoiding a false negative)? How much difference does it make to those who will not have the outcome whether or not they have a negative test (avoiding a false positive)? Of these two, how much more important is one than the other? This is a clinical judgment based on assessment of what the sequelae will be for a positive and for a negative test in the population of interest.

The default value for w is 1/2. In a situation in which one is primarily concerned with avoiding false negatives, w approaches 1; in a situation in which one is primarily concerned with avoiding false positives, w approaches 0. When both types of errors are of about equal concern, w = 1/2, the most common situation.

The definition of k(w) 15 appears in the Table (see page 713). Note that the numerator of k(w) does not actually depend on the weight w; only the reference value in the denominator does. Moreover, k(w) is a weighted average of k(1) (rescaled sensitivity or PVN) and k(0) (rescaled specificity or PVP), with weights determined by P, Q, and w. When P = Q, all values of k(w) are equal, regardless of what w is. Finally when k(w) = 0 for one value of w, k(w) equals 0 for all values of w. In any case, the magnitude of k(w) indicates how far between random and perfect a particular test is, when the errors are weighted as specified by w. There are no fixed standards for judging the magnitude of k(w), and generally one test is either compared with another test or against the quality of D regarded as a "gold standard."

INVARIANCE ISSUES

The values of all these probabilities and k(w) vary from one clinical population to another. That P varies is, of course, no surprise. Men are more likely to develop schizophrenia; women to develop depression, etc. That Q varies roughly paralleling P, is welcome: one would hope to have more positive tests in populations where there are more positive diagnoses. It is well understood that PVP and PVN vary from one population to another, typically with PVP increasing roughly as P increases, and with PVN decreasing.

What is more likely to be surprising, although it shouldn't be, is that the sensitivity and specificity of a test evaluated against the same diagnosis vary from one clinical population to another.[15-18] If the test is reasonably accurate for D, sensitivity tends to increase as P (and Q) increase, and specificity to decrease, although there is no exact functional relationship between P and either sensitivity or specificity across different clinical populations. It is quite possible that two populations with very similar base-rates (P) will have very different sensitivities or specificities, while two populations with very different base-rates may have quite similar values. It is one of the longest lasting and most tenacious myths in medical science that the sensitivities and specificities for a particular test evaluated against a particular diagnosis are test constants, the same in any clinical population.

So pervasive is this myth that medical students are sometimes required to demonstrate that they believe this myth on the medical board exam. Their question goes something like this: Suppose that it has been shown in one population that the sensitivity of a test is 90% and its specificity is 80%. Now it is proposed to use that test in a different population in which the prevalence of the disorder the test is meant to detect is 2%. In this latter population, what percentage of those with positive tests is likely to have the disorder?

What the question is meant to elicit is demonstration of the understanding of the definitions of prevalence, sensitivity, specificity, predictive values, and Bayes' theorem. Thus, the answer that will be counted as correct is that since the sensitivity was 90% and the specificity was 80% in the population with 2% prevalence, PVP = 8.4%. However, the correct answer is that there is insufficient information to answer their question. The fact that the sensitivity of that test is 90% and the specificity is 80% in another population does not tell anything about the sensitivity and specificity of the test in this population with prevalence 2%.

RISK RATIOS, RISK DIFFERENCES, NNT

To make matters worse, there is also a proliferation of other measures (and more being suggested each year) to describe the association between test and diagnosis, measures that do not necessarily agree with each other.

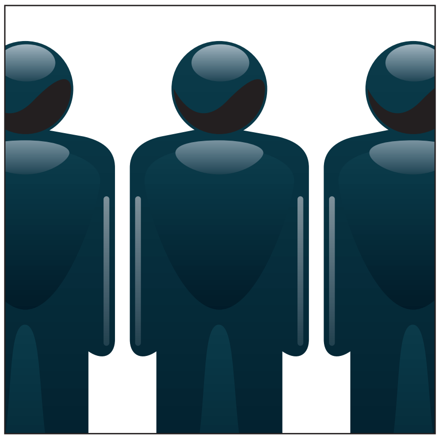
There are, for example, two risk differences (see the Table, page 713), which are usually different from each other, and from k(0), k(1/2) or k(1). However, it is easy to show that mathematically RD1 = k(P'), and RD2 = k(Q).

When comparing two or more tests against the same diagnosis in the same population (*P* fixed), RD1 often seems a reasonable choice. It is less reasonable to use RD2, because the weight RD2 places on false negatives versus false positives reflects the researchers' often arbitrary choice of test criterion (the level Q). This changes from one test to another when evaluated against the same diagnosis in the same population, making it impossible to compare tests.

Number Needed to Take (NNT) equals 1/RD1, and is the number needed to take with D = 1 to find one more T = 1 than if the same number were taken with D = 0.[19] Thus, random association approaches infinity; perfect association has NNT = 1. This measure is useful to convey clinical meaning to clinicians and medical consumers, because the units are patients rather than probability points, but in computations, is difficult to use. Its mathematical equivalent, RD1, is preferable.

In addition, there are also four risk ratios (see the Table, page 713). Risk ratios are on a completely different scale from the k(w) or the risk differences, where a value of 0 indicates random association, +1 perfect positive association and -1 perfect negative association. For the risk ratios, the value indicating random association is 1, and perfect-positive association would be indicated by a risk ratio approaching infinity, perfect negative association by a risk ratio approaching zero. The infinite scale in one direction, and the asymmetry of positive (1 to infinity) and negative (0 to 1) association, makes judging the magnitude of any risk ratio very difficult.

However, that problem is easily resolved, for a non-linear recalibration of the risk ratios that controls for how rapidly the various risk ratios increase from 1 to infinity (or decrease from 1 to 0) aligns them exactly with the k(w) (see Table, page 713). In brief, once each risk ratio is recalibrated, it is equivalent to either k(0) or k(1). Thus all four risk ratios from any two-by-two table, disparate though they may be from each other, are meaningful, but first require recalibration using the appropriate row or column probability, which reduces the four to two, and, after that, which of the two is relevant (if either) depends on the choice of the weight w.



When comparing two or more tests against the same diagnosis in the same population, RD1 often seems a reasonable choice.

WHAT OF THE PHI COEFFICIENT?

In behavioral research, less so in medical research, a common index used to describe the strength of association in this situation is the phi coefficient, the product moment correlation coefficient between T and D. The phi coefficient is the geometric mean of the two extreme kappa coefficients: phi = $(k(1)k(0))^{1/2}$. If P = Q, then k(w) are all equal to each other, and all equal the phi coefficient. Otherwise, phi = k(w*) where w* (see Table, page 713) depends on how close P and Q are to each other. Thus, as was the case with RD2, the weight w will vary with the choice of Q, and like RD2, phi is not recommended.

One final comment on phi: If a representative sample from the population is used for estimation, then the chi-square statistic $X = N \bullet phi^2$, where phi is the sample phi estimate, can be used to test the null hypothesis that the population phi is zero. Because the power to detect deviations from randomness depends in part on the relative sizes of P and Q, phi, rather than any of the other measures of two-by-two association, is most closely related to testing the null hypothesis of randomness. However, if the sample is stratified, the relationship between the chi-square statistic and population phi no longer holds, for the chi-square statistic will then reflect the researchers' choices of sample sizes in each stratum, but the population phi coefficient, of course, does not.

WHAT OF THE ODDS RATIO?

Finally there is the Odds Ratio (OR),[20] perhaps the most commonly used indication of strength of association in medical research. Below are 5 equivalent formulas for the Odds Ratio, all equivalent to that in the Table (see page 713):

$$OR = \frac{TP \bullet TN}{FP \bullet FN} = \frac{Se \bullet Sp}{(1 - Se) \bullet (1 - Sp)} =$$

$$\frac{PVP \bullet PVN}{(1 - PVP) \bullet (1 - PVN)} = RRI \bullet RR2 = RR3 \bullet RR4$$

The Odds Ratio has the same scale as do the Risk Ratios. However, because OR is the product of two Risk Ratios, for positive association OR is always bigger than the biggest risk ratio. Thus, Odds Ratio often conveys the impression of much stronger association than any of the other known measures of two-by-two association. This may be one of the reasons for its continued popularity in the face of considerable evidence suggesting how problematic its behavior is for assessing association in two-by-two tables.

Moreover, the relationship between Odds Ratio and the Risk Ratios has led to a widely held misperception, that the Odds Ratio is approximately the same as the largest Risk Ratio. Researchers often even interchange the labels Odds Ratio and Risk Ratio. It is true that when P is near zero or near 1, the Odds Ratio approximates the largest Risk Ratio, but this happens only because another Risk Ratio is near its null value of 1. If RR3 = 10 and RR4 = 1.01, for example, then clearly Odds Ratio = 10.1 is approximately equal to RR3. However, which of the two risk ratios here is the one to believe? Since rescaled RR3 equals k(0) and rescaled RR1 equals k(1), the answer to that question depends crucially on which of the two errors is considered clinically more important, an issue completely ignored by the Odds Ratio.

In the very special case when P = Q = 1/2, not only are all the k(w)'s the same for all values of w, but they all equal Yule's Index:

$$Y = \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1}$$

To correct the asymmetry in the OR scale, and the problems associated with a measure with infinite range, as well as to facilitate comparison with other measures of association, Y seems the best choice of rescaling the Odds Ratio to the standard (0,1) scale.

What then can be said about Y? 1) When P = Q = 1/2, Y equals k(w) for all values of w; 2) With random association Y is zero, as are all the k(w). 3) Y and all the k(w) always have the same sign. 4) Y is always greater than k(P').[21] However, Y may approach 1 when association is far from perfect, indeed often when most other measures of association indicate near random association.

One way of re-expressing k(w) is:

$$k(w) = \frac{TP \bullet TN - FP \bullet FN}{PQ'w + P'Qw'}$$

while one way of re-expressing Y is:

$$Y = \frac{TP \bullet TN - FP \bullet FN}{(\sqrt{TP \bullet TN} + \sqrt{FP \bullet FN})^2}$$

(ie, same numerator, different denominators). Thus, the basic difference between k(w) and Y is that the reference value in the denominator for k(w) depends on the baserate (P), how loosely or stringently the criteria for T are set (Q) and the relative clinical importance of false negatives to false positives (w), while the reference value for Y depends heavily on how closely one or the other of the error rates (FP and FN) comes to zero. Any situation in which Q<<P, or Q>>P will force one of the error rates toward zero, Y may approach 1 (Odds Ratio approaching infinity), even if that error rate corresponds to the error of least clinical importance.

The Odds Ratio appears to have been introduced originally as the likelihood ratio test statistic to test the null hypothesis of randomness comparing two Binomial distributions. The magnitude of an Odds Ratio unequal to 1, however, is uninterpretable as an effect size. Nevertheless, there are still many useful applications of the Odds Ratio, for example, it is an excellent indicator of non-randomness (eg, in Logistic Regression Analyses).

Moreover, unlike every other measure of two-by-two association, population Odds Ratio can be estimated (even with the same computation) from a representative sample, a prospective stratified sample or a case-control sample, provided all three are unbiased samples with unbiased measurements from the same population.[22] This is a computational convenience, that has led to a serious misunderstanding.

Because the estimation of Odds Ratio in a sample is not affected by the percentage of the sample selected with and without the diagnosis in an unbiased case-control study, it is often suggested that the Odds Ratio in the population does not depend on the baserate P, and thus that Odds Ratio is invariant over populations with different P. This claim is clearly not true for k(w), and k(w) has often been criticized because of its "dependence on baserate." However, Odds Ratio, like k(w), relating a test and diagnosis in different clinical populations, is not a constant.[16-18] This particular misunderstanding may also have arisen in part from the myth that sensitivity and specificity for a particular test evaluated against the diagnosis are constants across different clinical populations, since OR can be computed using only Se and Sp.

All things considered, we strongly recommend against use of the Odds Ratio (or Y) for the evaluation of medical tests,[20,23-26] but this should not limit its use for other previously noted more appropriate applications in medical research.

DESIGN, SAMPLING, AND ESTIMATION ISSUES

Design Issues

When a test is meant to detect a condition that exists at the time of testing (a diagnostic test), T and D are both assessed within a time span short enough that the situation does not change during that time. When a test is meant to predict a condition that may or may not occur in the future (a prognostic test), the test (T) is done at one time point and the subject followed prospectively to a later time point at which D is determined. In either case, to avoid measurement bias, the determinations of T and D should be done "blinded" to each other to avoid pseudo-correlation due to bias in the raters.

Naturalistic Sampling

If a representative sample is drawn from the population of interest and T and D observed for each subject in the sample, each of the probabilities in the Table (see page 713) can be estimated by substituting the proportions of the sample seen in each cell of the two-by-two table for the corresponding probability in the population definition. For accuracy of estimates, sample sizes must be particularly large when the baserate, P, or the level, Q, is either very small or very large.

Prospective Stratified Sampling (Two-stage Sampling)

When P is very small, and/or D can be determined only with great difficulty or cost, prospective stratification is an efficient option. Suppose we proposed to sample a large sample of subjects from the population of interest and assess T (Stage 1). From this sample we can estimate Q, the level of the test. Then (Stage 2), we can randomly sample N1 from among those with T = 1, and randomly sample N0 from among those with T = 0, over-sampling the rarer result for assessment of D. From the second stage sample we can estimate PVP and PVN and thus RD2, RR3, RR4 and Odds Ratio, but not P,Q, sensitivity, specificity, or any of the k(w).

Without P we cannot recalibrate PVP and PVN. Without Q, we do not know what the value of the weight is for RD2=k(Q). However, combining the estimate of Q from Stage 1, and the estimates of PVP and PVN from Stage 2, we can estimate all the cell probabilities, and thus all the probabilities and consequently all the k(w).

Having Stage 1 satisfies two crucial needs: 1) to generate representative samples of those within the population with T = 1 and with T = 0 for Stage 2, and 2) to obtain an unbiased estimate of Q in this population. Without satisfying both of these needs, the assessment of the quality of the decision process can be badly flawed. In short, approaching the evaluation of a test using prospective stratified sampling is quite viable and in many cases advisable, but the sampling must be done with the full two-stage sampling approach and an estimation procedure appropriate to that sampling approach used.

Retrospective Stratified Sampling (case-control)

The situation with respect to retrospective stratified sampling is quite different. A representative sample of those with D = 1 in 2008, for example, is not necessarily a representative sample of those in the population in 1998 who would have gone on to have D = 1 in 2008. Moreover, assessment in 2008 of what T was in 1998 often depends on retrospective recall and faulty records. Most crucially, the retrospective recall or interpretation of records can be influenced by whether or not D = 1 in 2008, thus compromising "blindness."

If there were neither a sampling bias nor a measurement bias, from a case-control study, one could estimate Se, Sp, RD1, RR1, RR2 and Odds Ratio, but not P,Q, the predictive values, or any of the k(w) other than RD1 = k(P'). This is the context in which Odds Ratio was recommended. But why not recommend RD1 = k(P') rather than Odds Ratio? Generally, advocates of the Odds Ratio point out that RD1 in such situations tends to be very small, while Odds Ratio tends to be very large, a questionable logic. However, proposing the use of RD1 in place of Odds Ratio would not solve the problem in any case, for it is difficult to avoid the sampling biases and measurement biases of case-control studies.

A case-control study meant as an exploratory study to investigate the possibility of non-random association between T and D in a population may be very useful as a preliminary to proposing generating a naturalistic sample or designing a two-stage prospective sample in the population in a future study. However, to assess the quality of decision-making, we recommend against a case-control design.

Estimation

For either prospective design, it is important not only to derive a point estimate of the appropriate k(w), but to recognize its estimation error, by presenting a confidence interval estimate. The best current approach is to use a bootstrap method[27-29] appropriate to the sampling method.

CONCLUSION

Clearly, the discussion leads to recommendations about appropriate sampling, measurement, design, analysis, and presentation and evaluation of results. Rather than repeating these, let us instead indulge in a flight of fancy on "How to Lie with Statistics."[30]

Suppose someone really wanted to demonstrate a strong association between some test (T) and an outcome (D) and didn't much care how. What would they do? First of all, they would set the criterion for T = 1 stringently (Q<<P or Q>>P) to force one of the error rates toward zero. Then, they would do a case-control study with a very large sample size, using inclusion/exclusion criteria that assure that the "cases" are sure to have D = 1 and the "controls" sure to have D = 0. Such two samples are more likely to represent the extremes of the population, which will exaggerate any appearance of association, although unlikely to produce an association where there is none. Then, in the informed consent process, researchers would emphasize the scientific and empirical evidence to date that T is associated with D, and then use retrospective recall for T. What this does is to induce "cases" to be much more sensitive to the possibility that T = 1 in their past, which will tend to result in over-reporting in that group. This too serves to exaggerate association, now perhaps even inducing the appearance of association when there is none. Then, they would use the Odds Ratio as the measure of association, because Odds Ratio will indicate association much stronger than any other measure of association, particularly under the above circumstances.

Finally, there is precedent to overstating the interpretation of the magnitude of whatever Odds Ratio is obtained. An Odds Ratio of 1.5 is often regarded as "large," even though it corresponds only to Y = .1 (It is sometimes even described as a 50% increase in risk.) If the Odds Ratio has (by such inflated standards) a "large" magnitude, but is not statistically significant despite the large sample size, one would emphasize the Odds Ratio and ignore the *P* value (perhaps reporting a "trend"). If the Odds Ratio has an admittedly small magnitude (ie, 1.1, Y = .02) but is statistically significant (which is likely with a large sample size), one would emphasize the *P* value and ignore the magnitude of the Odds Ratio.

In neither case, would one present any other descriptive statistics (eg, the sensitivity/specificity or predictive values), or discuss the potential impact on clinical decision making (w), because that might raise doubts about the clinical value of the test, whatever its Odds Ratio.

Unfortunately, this may be remarkably close in some respects to the way many epidemiological studies, genetic studies, and medical test evaluations have been conducted in the past, and may help explain why such studies are so difficult to replicate or confirm.

In contrast, how would we conduct a rigorous evaluation of T against D? One would begin by proposing to obtain a representative sample from the population of interest, with (T,D) evaluated by two different evaluators, done in such a way that neither affects the other's decision. In dealing with a rare event, we might propose two-stage sampling as done above, estimating P from the first stage, and then PVP and PVN from the second stage, combining these to estimate the four probabilities of the two-by-two table. We would then consider the relative importance of false positives versus false negatives, and set w. Then, we would proceed to execute the study, to estimate probabilities in the two-by-two table and to compute the appropriate k(w) from the data. To get 95% two-tailed confidence intervals, we would use a bootstrap procedure. If zero is not contained within that confidence interval we would be assured that association was "statistically significant" at the two-tailed 5% level. To assess the clinical significance of the finding, we might compare the estimated k(w) against the intraclass reliability kappa for D, or against other possible tests for D.

The results would undoubtedly be far less exciting than those from the first study, but would likely be closer to the truth, and more likely to be replicable across future studies.

REFERENCES

1. Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology.* 1978;46:806-834.
2. Jones LV, Tukey JW. A sensible formulation of the significance test. *Psychol Methods.* 2000;5(4):411-414.
3. Borenstein M. Hypothesis testing and effect size estimation in clinical trials. *Ann Allergy Asthma Immunol.* 1997;78(1):5-16.
4. Borenstein M. The shift from significance testing to effect size estimation. In: Hersen M, ed. Research & Methods, Comprehensive Clinical Psychology. vol 3. Burlington, MD: Elsevier Science Publishing; 1998:319-349.
5. Cohen J. The Earth is round (p<.05). *American Psychologist.* 1995;49:997-1003.
6. Dar R, Serlin RC, Omer H. Misuse of statistical tests in three decades of psychotherapy research. *J Consul Clinical Psychol.* 1994;62(1):75-82.
7. Hunter JE. Needed: a ban on the significance test. Psychological Science. 1997;8(1):3-7.
8. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods.* 2000;5(2):241-301.
9. Shrout PE. Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science.* 1997;8(1):1-2.
10. Thompson B. Journal editorial policies regarding statistical significance tests: heat is to fire as p is to importance. *Educational Psychology Review.* 1999;11:157-169.
11. Wilkinson L. Task Force on Statistical Inference. Statistical methods in psychology journals: guidelines and explanations. *American Psychologist.* 1999;54:594-604.
12. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull.* 1968;70:213-229.
13. Bloch DA, Kraemer HC. 2X2 kappa coefficients: measures of agreement or association. *Biometrics.* 1989;45:269-287.
14. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 1960;20:37-46.
15. Kraemer HC. Evaluating Medical Tests: Objective and Quantitative Guidelines. Newbury Park, CA: Sage Publications; 1992.
16. Hlatky MA, Mark DB, Harrell FE, Lee KL, Califf RM, Pryor DB. Rethinking sensitivity and specificity. *Am J Cardiol.* 1987;59(12):1195-1198.
17. Hlatky MA, Pryor DB, Harrell FE, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. *Am J Med.* 1984;77(1):64-71.
18. Fleiss JL. On the asserted invariance of the odds ratio. *Br J Prev Soc Med.* 1970;24(1):45-46.
19. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ.* 1995;310(6977):452-454.
20. Sackett DL. Down with odds ratios! *Evidence-Based Medicine.* 1996;1:164-166.
21. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry.* 2006;59(11):990-996.
22. Cornfield J. A statistical problem arising from retrospective studies. In: Neyman J, ed. Proceedings of the Third Berkeley Symposium. vol IV. Berekely, CA: University of California Press; 1956:135.
23. Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinical or policy significance. *Psychological Methods.* 1999;4(3):257-271.
24. Kraemer HC. Reconsidering the odds ratio as a measure of 2X2 association in a population. *Stat Med.* 2004;23(2):257-270.
25. Kraemer HC. Correlation coefficients in medical research: from product moment correlation to the odds ratio. *Stat Methods Med Res.* 2007;15(6):525-545.
26. Newcombe RG. A deficiency of the odds ratio as a measure of effect size. *Stat Med.* 2006;25(24):4235-4240.
27. Efron B, Tibshirani R. *Computer-Intensive Statistical Methods.* Stanford, CA: Division of Biostatistics, Stanford University Press; 1995:174.
28. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician.* 1983;37:36-48.
29. Efron B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics.* 1979;7:1-26.
30. Geis I. *How to Lie with Statistics.* New York, NY: W.W. Norton & Company; 1954.