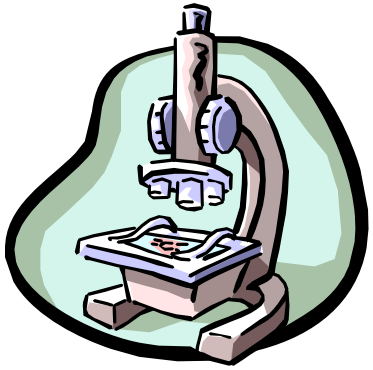
Measuring reliability and agreement



Madhukar Pai, MD, PhD
Assistant Professor of Epidemiology, McGill University
Montreal, Canada
Professor Extraordinary, Stellenbosch University, S Africa

Email: madhukar.pai@mcgill.ca

What is reliability

- Repeatability under similar conditions, either by the same reader or different readers
- When one measure is compared against a “reference standard” [‘truth’] then this process is often called “calibration”

Why is reliability important?

- How can we trust a test that does not give consistent results?
- Good example: in-house PCR for TB produces highly inconsistent results
  - Cannot be used for clinical diagnosis [unless you have validated it in your own setting]
- IFN-g assays for serial testing of healthcare workers
  - How stable are IFN-g values over time and how do we decide who has a IGRA “conversion”?

BMC Microbiology

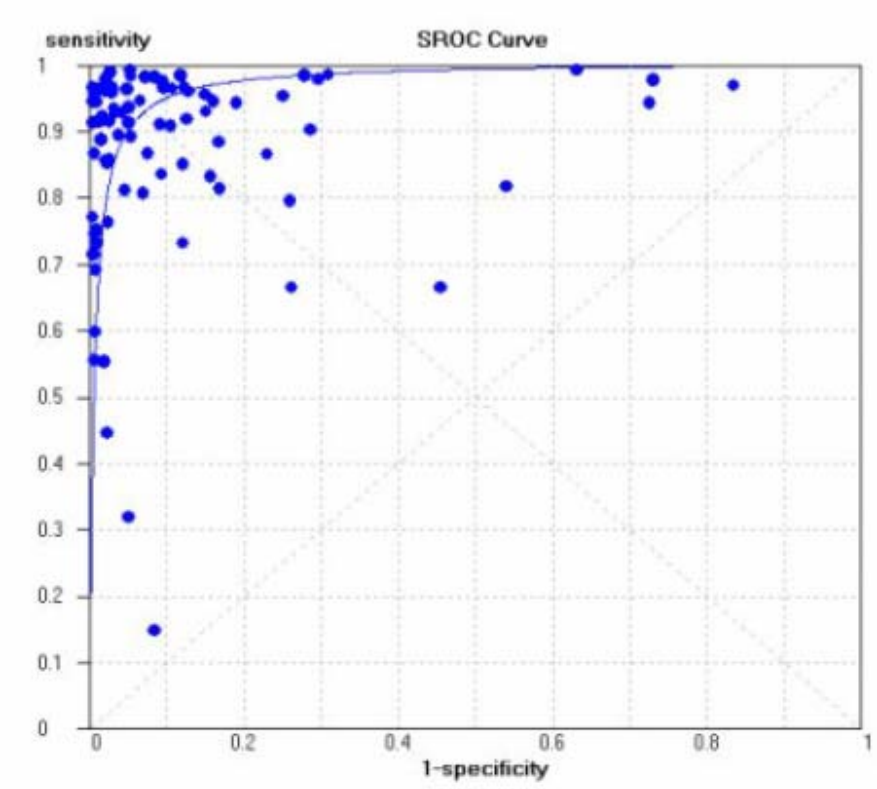
BioMed Central

Research article

Open Access

In-house nucleic acid amplification tests for the detection of *Mycobacterium tuberculosis* in sputum specimens: meta-analysis and meta-regression

Laura L Flores[1,2,3], Madhukar Pai[1,3], John M Colford Jr[1] and Lee W Riley*[1]



JOURNAL OF CLINICAL MICROBIOLOGY, Oct. 1996, p. 2522–2525 Vol. 34, No. 10
0095-1137/96/$04.00+0

Reliability of Nucleic Acid Amplification for Detection of *Mycobacterium tuberculosis*: an International Collaborative Quality Control Study among 30 Laboratories

GERDA T. NOORDHOEK,[1]* JAN D. A. van EMBDEN,[2] AND AREND H. J. KOLK[3]

Public Health Laboratory, Leeuwarden,[1] National Institute for Public Health and Environmental Protection, Research Laboratory for Infectious Diseases, Bilthoven,[2] and Department of Biomedical Research, Royal Tropical Institute, Amsterdam,[3] The Netherlands

Received 28 May 1996/Returned for modification 29 June 1996/Accepted 24 July 1996

Nucleic acid amplification to detect *Mycobacterium tuberculosis* in clinical specimens is increasingly used as a laboratory tool for the diagnosis of tuberculosis. However, the specificity and sensitivity of these tests may be questioned, and no standardized reagents for quality control assessment are available. To estimate the performance of amplification tests for routine diagnosis, we initiated an interlaboratory study involving 30 laboratories in 18 countries. We prepared blinded panels of 20 sputum samples containing no, 100, or 1,000 mycobacterial cells. Each laboratory was asked to detect *M. tuberculosis* by their routine method of nucleic acid amplification. Only five laboratories correctly identified the presence or absence of mycobacterial DNA in all 20 samples. Seven laboratories detected mycobacterial DNA in all positive samples, and 13 laboratories correctly reported the absence of DNA in the negative samples. Lack of specificity was more of a problem than lack of sensitivity. Reliability was not found to be associated with the use of any particular method. Reliable detection of *M. tuberculosis* in clinical samples by nucleic acid amplification techniques is possible, but many laboratories do not use adequate quality controls. This study underlines the need for good laboratory practice and reference reagents to monitor the performance of the whole assay, including pretreatment of clinical samples.

Serial Testing of Health Care Workers for Tuberculosis Using Interferon-γ Assay

Madhukar Pai, Rajnish Joshi, Sandeep Dogra, Deepak K. Mendiratta, Pratibha Narang, Shriprakash Kalantri, Arthur L. Reingold, John M. Colford, Jr., Lee W. Riley, and Dick Menzies

Divisions of Epidemiology and Infectious Diseases, School of Public Health, University of California, Berkeley; Division of Pulmonary and Critical Care Medicine, San Francisco General Hospital, University of California, San Francisco, California; Departments of Medicine and Microbiology, Mahatma Gandhi Institute of Medical Sciences, Sevagram, India; and the Montreal Chest Institute, McGill University, Montreal, Quebec, Canada

TABLE 1. INCIDENCE OF TUBERCULIN SKIN TEST AND QUANTIFERON-TB GOLD IN-TUBE ASSAY CONVERSIONS OVER A 18-MONTH PERIOD AMONG PARTICIPANTS WHO WERE CONCORDANTLY NEGATIVE BY BOTH TESTS AT BASELINE*

Definition of Conversion	No. Serially Tested	No. of Conversions	% Incidence of Conversions (*95% CI*)
TST			
1. Baseline induration of < 10 mm and follow-up TST of ≥ 10 mm, with increment of ≥ 6 mm	147	14	9.5 (5.3–15.5)
2. Baseline induration of < 10 mm and follow-up TST of ≥ 10 mm, with increment of ≥ 10 mm	147	6	4.1 (1.5–8.7)
QFT			
3. Baseline IFN-γ < 0.35 IU/ml and follow-up IFN-γ ≥ 0.35 IU/ml	147	17	11.6 (6.9–17.9)
4. Baseline IFN-γ < 0.35 IU/ml and follow-up IFN-γ ≥ 0.70 IU/ml	147	11	7.5 (3.8–13.0)
Combinations of TST and QFT			
1 or 3	147	22	14.9 (9.6–21.8)
2 or 4	147	11	7.5 (3.8–13.0)
1 and 3	147	9	6.1 (2.8–11.3)
2 and 4	147	6	4.1 (1.5–8.7)

Definition of abbreviations: CI = confidence interval; QFT = QuantiFERON-TB Gold In-Tube assay; TST = tuberculin skin test.
* n = 147.

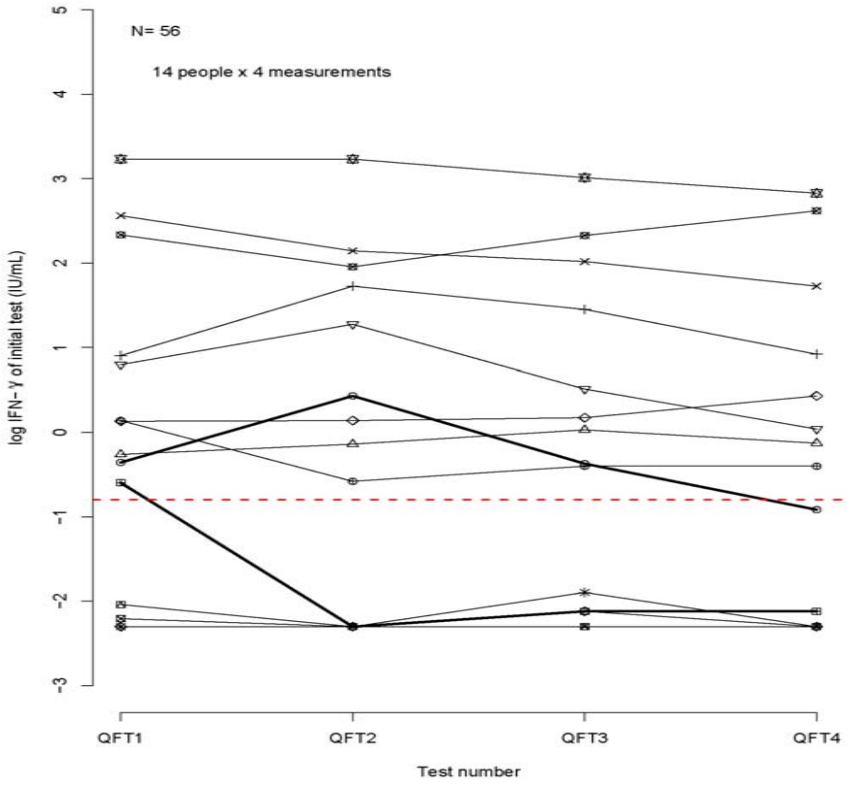
OPEN ACCESS Freely available online

PLoS one

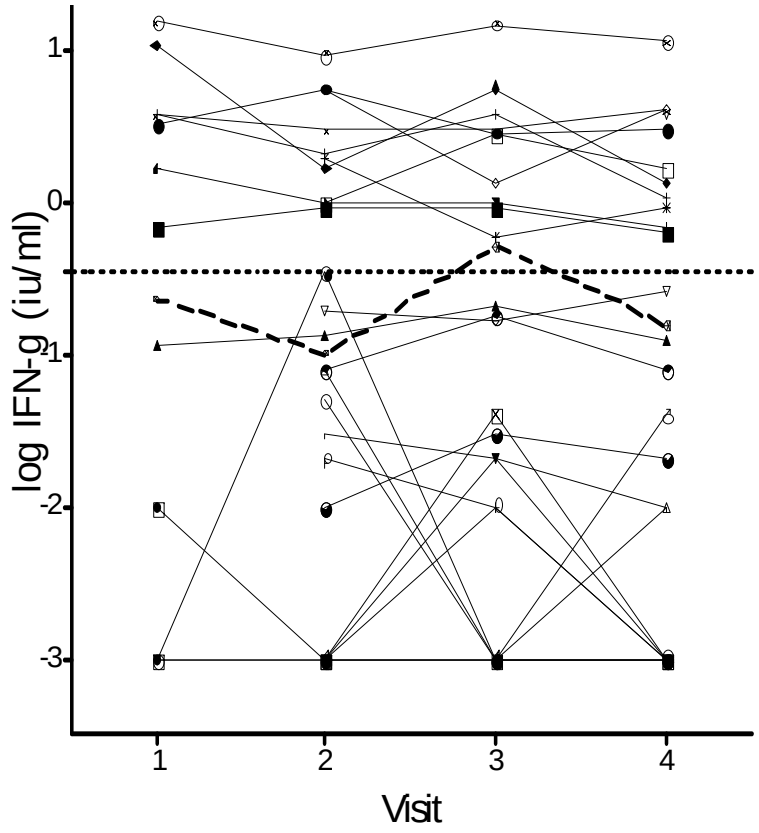
T-Cell Assays for Tuberculosis Infection: Deriving Cut-Offs for Conversions Using Reproducibility Data

Anandharaman Veerapathran[1,2], Rajnish Joshi[1,2,3], Kalyan Goswami[1,2], Sandeep Dogra[4], Erica E. M. Moodie[5], M. V. R. Reddy[1,2], Shriprakash Kalantri[1,2], Kevin Schwartzman[5,6], Marcel A. Behr[5,7], Dick Menzies[5,6], Madhukar Pai[5,6]*

1 Department of Biochemistry, Mahatma Gandhi Institute of Medical Sciences, Sevagram, Maharashtra, India, **2** Department of Medicine, Mahatma Gandhi Institute of Medical Sciences, Sevagram, Maharashtra, India, **3** Division of Epidemiology, School of Public Health, University of California, Berkeley, California, United States of America, **4** Acharya Shri Chander College of Medical Sciences and Hospital, Jammu, India, **5** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, **6** Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, McGill University, Montreal, Canada, **7** Division of Infectious Diseases and Medical Microbiology, McGill University Health Centre, Montreal, Canada



Source: Veerapathran et al. *PLoS ONE* 3(3): e1850



van Zyl Smit R, et al. *Am J Resp Crit Care Med* 2009

Types of measurements

- Continuous
  - E.g. blood pressure, cholesterol, CD4+ counts
- Ordinal
  - Scales and scores (e.g. Glasgow Coma; Apgar score)
  - Protein energy malnutrition
  - Pain
- Dichotomous
  - Tests with pos/neg results

Measures of reliability with continuous test results

Measures of reliability

- Within subject standard deviation
  - When measures are repeated on the same subject
  - Same as within subject standard deviation
- Correlation coefficient
- Coefficient of variation
- 95% limits of agreement

Example: duplicate glucometer values

Calculation of Within-Subject Standard Deviation on Duplicate Glucose Measurements

	Glucose Measurement (mg/dL)			
Specimen	1	2	Difference	Variance = $(M_1 - M_2)^2/2$
1	80	92	−12	72
2	89	92	−3	4.5
3	93	109	−16	128
4	97	106	−9	40.5
5	103	87	16	128
6	107	104	3	4.5
7	100	105	−5	12.5
8	112	104	8	32
9	123	110	13	84.5
10	127	120	7	24.5

Average Variance = 53.1

$S_w = 7.3$

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

Example: duplicate glucometer values

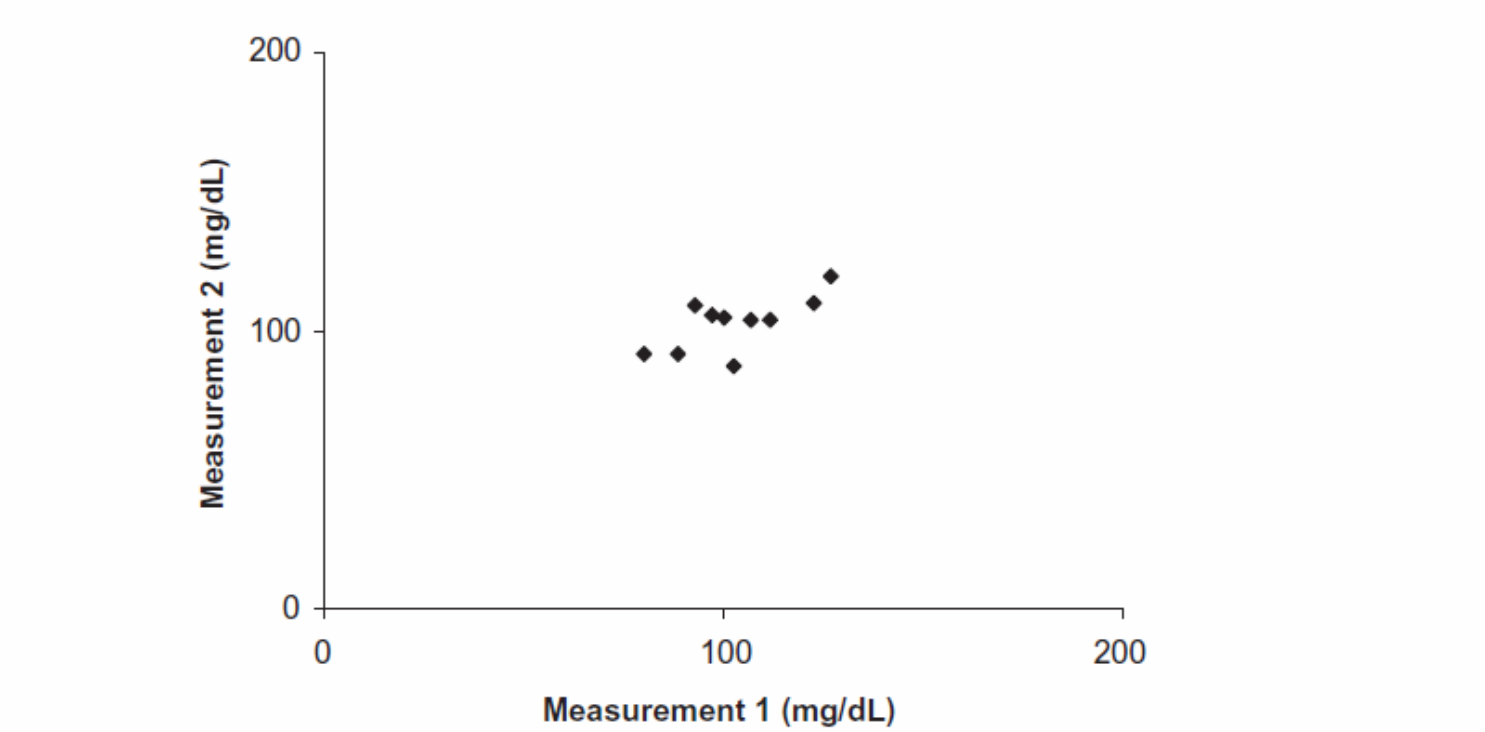


Figure 2.2 Scatterplot of the glucometer readings in Example 2.11. Correlation coefficient = 0.67.

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

Example: duplicate glucometer values

Duplicate Glucose Measurements from Example 2.11 (except for the last observation)

	Glucose Measurement (mg/dL)			
Specimen	**1**	**2**	**Difference**	**Variance**
1	80	92	−12	72.0
2	89	92	−3	4.5
3	93	109	−16	128.0
4	97	106	−9	40.5
5	103	87	16	128.0
6	107	104	3	4.5
7	100	105	−5	12.5
8	112	104	8	32.0
9	123	110	13	84.5
10	***300***	***600***	***−300***	**45000.0**

Average Variance = 4550.7

$S_w = 67.5$

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

Example: duplicate glucometer values

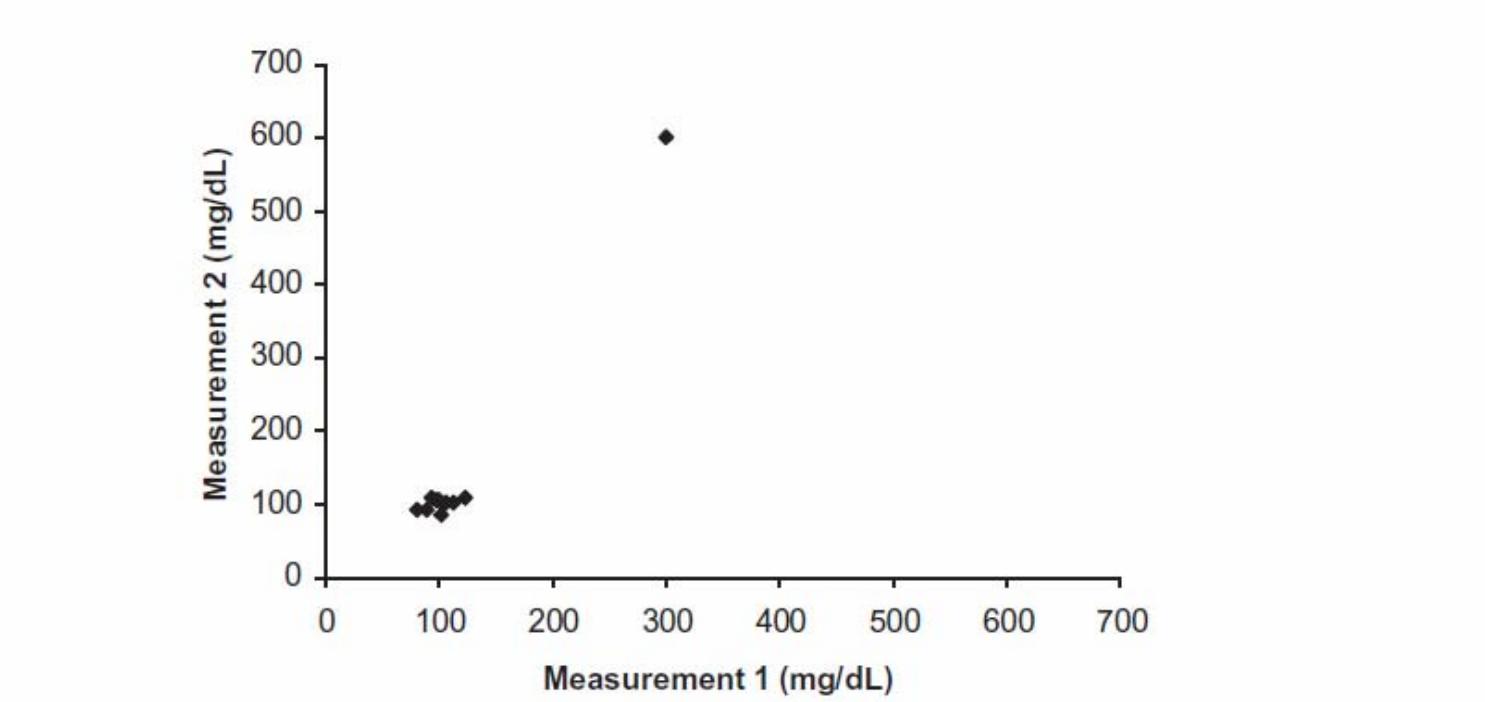


Figure 2.3 Scatterplot of the glucometer readings in Example 2.12. Correlation coefficient = 0.99.

Be cautious in using simple correlation coefficients!

Newman T, Kohn MA. Evidence-based diagnosis. 2009, Cambridge Univ Press

CV%

- Specifying the standard deviation is not helpful without the additional specification of the mean value
- It makes a big difference if $s = 5$ with a mean of $= 100$, with a mean of $= 3$.
- Relating the standard deviation to the mean resolves this problem. In other words, we need a normalized measure of dispersion
- The coefficient of variation is therefore equal to the within-subject standard deviation divided by the mean

$$coefficient\ of\ variation = \frac{standard\ deviation}{mean}$$

Other approaches

Ultrasound Obstet Gynecol 2003; 22: 85–93
Published online 9 May 2003 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/uog.122

Applying the right statistics: analyses of measurement studies

J. M. BLAND* and D. G. ALTMAN†

**St George's Hospital Medical School, London and †Cancer Research UK Medical Statistics Group, Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK*

Limits of agreement

- We start with the differences between measurements by two methods
- We calculate the mean and SD of these differences.
- Then we calculate the mean difference $\pm$ 2 SDs.
- We would expect 95% of differences between measurements by two methods to lie between these limits.

Bland & Altman 2003

Limits of agreement

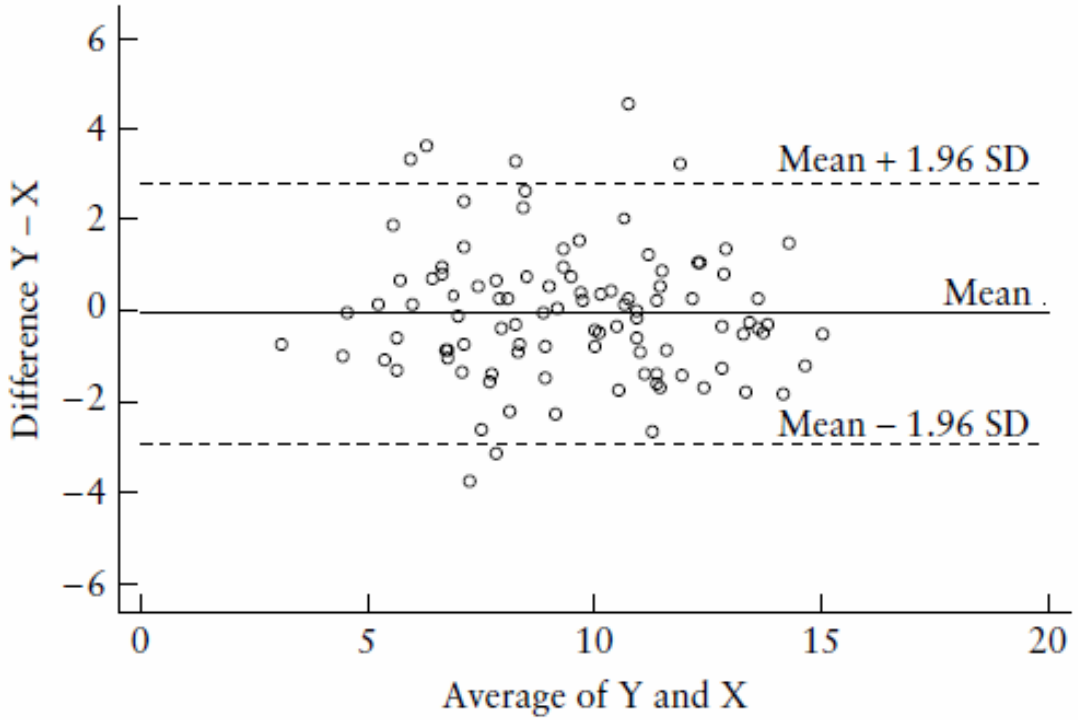


Figure 12 Plot of difference against mean for the artificial data X and Y (as in Figure 11a), with mean difference and 95% limits of agreement indicated.

Bland & Altman 2003

Measures of reliability with dichotomous or ordinal test results

Clinicians often disagree

- Clinicians often disagree in their assessment of patients.
  - When 2 clinicians reach different conclusions regarding the presence of a particular physical sign, either different approaches to the examination or different interpretation of the findings may be responsible for the disagreement.
  - Similarly, disagreement between repeated applications of a diagnostic test may result from different application of the test or different interpretation of the results.
- Researchers may also face difficulties in agreeing on issues such as whether patients in a trial have experienced the outcome of interest (eg, they may disagree about whether a patient has had a transient ischemic attack or a stroke or about whether a death should be classified as a cardiovascular death), or whether a study meets the eligibility criteria for a systematic review

Chance Will Always Be Responsible for Some of the Apparent Agreement Between Observers

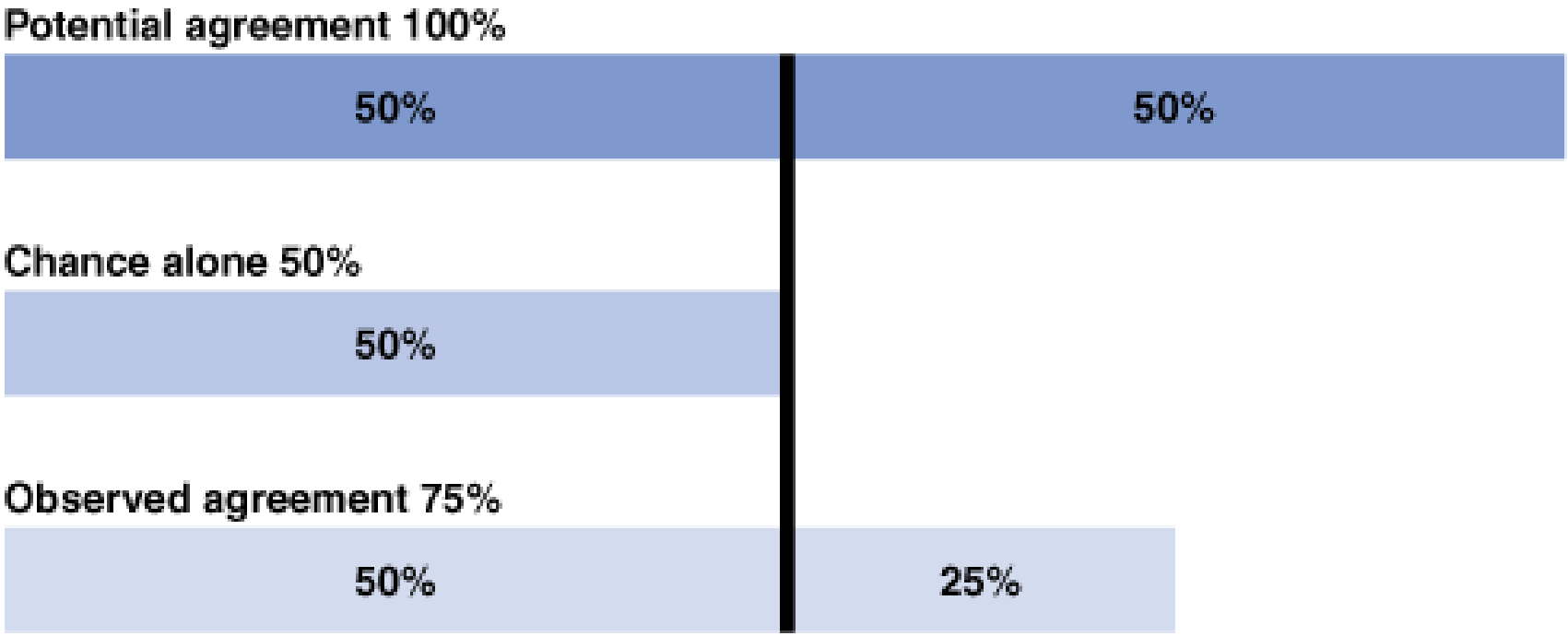
- Any 2 people judging the presence or absence of an attribute will agree some of the time simply by chance.
- Similarly, even inexperienced and uninformed clinicians may agree on a physical finding on occasion purely as a result of chance.
- This chance agreement is more likely to occur when the prevalence of a target finding (a physical finding, a disease, an eligibility criterion) is high.
- When investigators present agreement as raw agreement (or crude agreement)—that is, by simply counting the number of times agreement has occurred—this chance agreement gives a misleading impression.

Alternatives for Dealing With the Problem of Agreement by Chance

- When we are dealing with categorical data (i.e., placing patients in discrete categories such as mild, moderate, or severe or stage 1, 2, 3, or 4), the most popular approach to dealing with chance agreement is with chance-corrected agreement.
- Chance-corrected agreement is quantitated as kappa, or weighted kappa.

Chance-Corrected Agreement, or kappa

- kappa removes most of the agreement by chance and informs clinicians of the extent of the possible agreement over and above chance.
- The total possible agreement on any judgment is always 100%.
- Figure depicts a situation in which agreement by chance is 50%, leaving possible agreement above and beyond chance of 50%.
- As depicted in the figure, the raters have achieved an agreement of 75%. Of this 75%, 50% was achieved by chance alone. Of the remaining possible 50% agreement, the raters have achieved half, resulting in a value of 0.25/0.50, or 0.50.

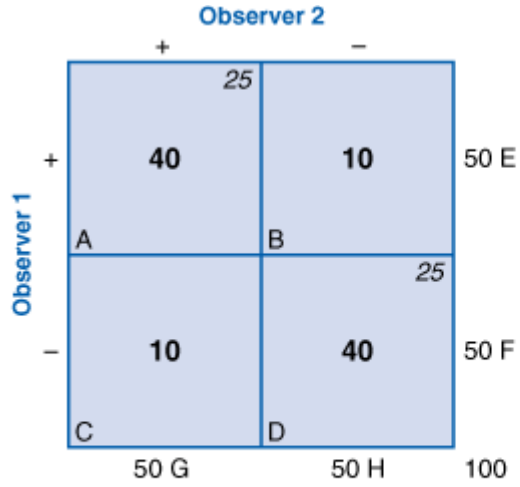


$\kappa = 0.25/0.50 = 0.50$ (good agreement)

Source: Guyatt G, Rennie D, Meade MO, Cook DJ: *Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice*, 2nd Edition: http://www.jamaevidence.com

How is kappa calculated?

- Assume that 2 observers are assessing the presence of Murphy sign, which may help clinicians detect an inflamed gallbladder.
- First, we calculate the agreement observed:
- In 40 patients, the 2 observers agreed that Murphy sign was positive (cell A) and they further agreed that in another 40 patients, it was negative (cell D).
- Thus, the total agreement is 40 + 40, or 80%.

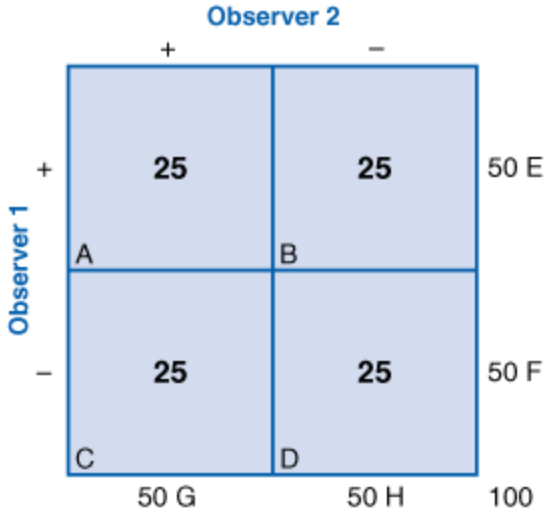


Source: Guyatt G, Rennie D, Meade MO, Cook DJ: *Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice*, 2nd Edition: http://www.jamaevidence.com

Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd Edition

How is kappa calculated?

- Now assume they have no skill at detecting the presence or absence of Murphy sign, and their evaluations are no better than blind guesses.
- Let us say they are both guessing in a ratio of 50:50; they guess that Murphy sign is present half of the time and that it is absent half of the time.
- On average, if both raters were evaluating the same 100 patients, they would achieve the results presented in Figure.
- Referring to that figure, you observe that these results demonstrate that the 2 cells that tally the raw agreement, A and D, include 50% of the observations. Thus, simply by guessing (and thus by chance), the raters have achieved 50% agreement.



Source: Guyatt G, Rennie D, Meade MO, Cook DJ: *Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice*, 2nd Edition: http://www.jamaevidence.com

Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd Edition

How is kappa calculated?

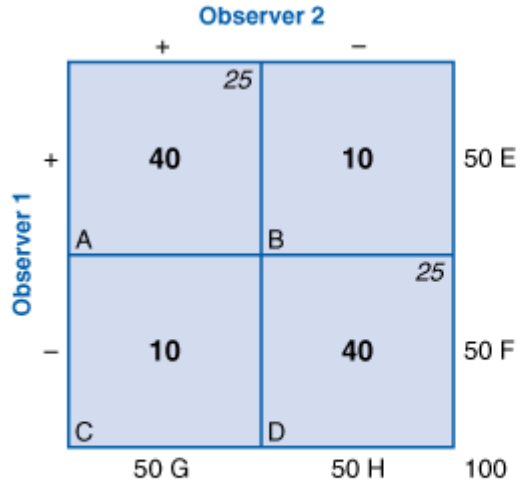
- The total agreement by chance is 0.25 + 0.25, or 0.50, 50%.
- Observed agreement is 80%

We can then calculate κ using the principle illustrated

$$\frac{(\text{agreement observed} - \text{agreement by chance})}{(\text{agreement possible} - \text{agreement by chance})}$$

or in this case:

$$\frac{80 - 50}{100 - 50} = \frac{30}{50} = 0.6$$



Source: Guyatt G, Rennie D, Meade MO, Cook DJ: *Users' Guides to the Medical Literature: A Manual for Evidence-Based Practice*, 2nd Edition: http://www.jamaevidence.com

95% CI = 0.44 to 0.76

Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd Edition

What is a good kappa value?

- There are a number of approaches to valuing the k levels raters achieve. One option is the following:
  - 0 = poor agreement;
  - 0 to 0.2 = slight agreement;
  - 0.21 to 0.4 = fair agreement;
  - 0.41 to 0.6 = moderate agreement;
  - 0.61 to 0.8 = substantial agreement; and
  - 0.81 to 1.0 = almost perfect agreement

Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd Edition

Kappa with 3 or More Raters, or 3 or More Categories

- Using similar principles, one can calculate chance-corrected agreement when there are more than 2 raters
- Furthermore, one can calculate when raters place patients into more than 2 categories (eg, patients with heart failure may be rated as New York Heart Association class I, II, III, or IV).
- In these situations, one may give partial credit for intermediate levels of agreement (for instance, one observer may classify a patient as class II, whereas another may observe the same patient as class III) by adopting a so-called weighted kappa statistic.
- Weighting refers to calculations that give full credit to full agreement and partial credit to partial agreement (according to distance from the diagonal on an agreement table)

Limitation of kappa

- Despite its intuitive appeal and widespread use, the k statistic has one important disadvantage:
  - As a result of the high level of chance agreement when distributions become more extreme, the possible agreement above chance agreement becomes small, and even moderate values of k are difficult to achieve.
  - Thus, using the same raters in a variety of settings, as the proportion of positive ratings becomes extreme, k will decrease even if the raters' skill at interpretation does not

Kappa examples

Accuracy and reliability of palpation and percussion for detecting hepatomegaly: a rural hospital-based study

Rajnish Joshi, Amandeep Singh, Namita Jajoo, Madhukar Pai,* S P Kalantri

Department of Medicine, Mahatma Gandhi Institute of Medical Sciences, Sevagram 442 102, Maharashtra; and *Division of Epidemiology, University of California at Berkeley, Berkeley, CA 94720, USA

Table 3: Inter-physician agreement in determination of palpable liver and percussion liver span ≥10 cm

	Palpation		Percussion	
	Percent agreement	Kappa (95% CI)	Percent agreement	Kappa (95% CI)
Physician 1 vs. physician 2	82	0.44 (0.29, 0.60)	68	0.33 (0.19, 0.47)
Physician 2 vs. physician 3	84	0.49 (0.33, 0.65)	64	0.31 (0.17, 0.44)
Physician 1 vs. physician 3	84	0.53 (0.38, 0.68)	57	0.17 (0.04, 0.30)

Ind J Gastro 2004

Kappa examples

Respiratory Medicine (2007) **101**, 431–438

ELSEVIER

respiratoryMEDICINE

Accuracy and reliability of physical signs in the diagnosis of pleural effusion

Shriprakash Kalantri[a], Rajnish Joshi[a,c], Trunal Lokhande[a], Amandeep Singh[a], Maureen Morgan[b], John M. Colford Jr[c], Madhukar Pai[d,*]

Table 2 Reproducibility of physical signs for pleural effusion.

Covariate	Agreement between observers 1 and 2	
	%	κ (95% CI)
Asymmetric chest expansion	95.04	0.85 (0.73, 0.98)
Reduced vocal fremitus	94.66	0.86 (0.74, 0.97)
Dull percussion	93.1	0.84 (0.71, 0.94)
Decreased or absent breath sounds	95.80	0.89 (0.77, 1.00)
Reduced vocal resonance	91.60	0.78 (0.66, 0.89)
No Crackles	87.73	0.67 (0.56, 0.79)
Pleural rub	94.27	−0.02 (−0. 57–0.78)
Auscultatory percussion	92.37	0.76 (0.64, 0.84)

Abbreviation: κ, kappa statistic.

Resp Med 2007

Kappa examples

Mycobacterium tuberculosis Infection in Health Care Workers in Rural India

Comparison of a Whole-Blood Interferon γ Assay With Tuberculin Skin Testing

Madhukar Pai, MD, PhD
Kaustubh Gokhale, MD
Rajnish Joshi, MD
Sandeep Dogra, MD
Shriprakash Kalantri, MD, MPH
Deepak K. Mendiratta, MD
Pratibha Narang, MD
Charles L. Daley, MD
Reuben M. Granich, MD, MPH
Gerald H. Mazurek, MD
Arthur L. Reingold, MD
Lee W. Riley, MD
John M. Colford, Jr, MD, PhD

Context *Mycobacterium tuberculosis* infection in health care workers has not been adequately studied in developing countries using newer diagnostic tests.

Objectives To estimate latent tuberculosis infection prevalence in health care workers using the tuberculin skin test (TST) and a whole-blood interferon γ (IFN-γ) assay; to determine agreement between the tests; and to compare their correlation with risk factors.

Design, Setting, and Participants A cross-sectional comparison study of 726 health care workers aged 18 to 61 years (median age, 22 years) with no history of active tuberculosis conducted from January to May 2004, at a rural medical school in India. A total of 493 (68%) of the health care workers had direct contact with patients with tuberculosis and 514 (71%) had BCG vaccine scars.

Interventions Tuberculin skin testing was performed using 1-TU dose of purified protein derivative RT23, and the IFN-γ assay was performed by measuring IFN-γ response to early secreted antigenic target 6, culture filtrate protein 10, and a portion of tuberculosis antigen TB7.7.

Main Outcome Measures Agreement between TST and the IFN-γ assay, and comparison of the tests with respect to their association with risk factors.

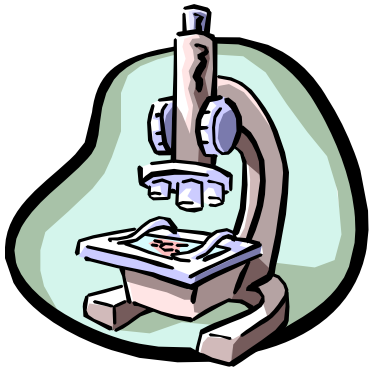
Table 3. Agreement Between TST and IFN-γ Assay Results (n = 719)

Results*	TST Cutpoint, mm ≥5	≥10	≥15
Positive TST/positive IFN-γ assay	259	226	148
Negative TST/negative IFN-γ assay	254	359	412
Positive TST/negative IFN-γ assay	177	72	19
Negative TST/positive IFN-γ assay	29	62	140
Agreement, %	71.4	81.4	77.9
κ (95% CI)	0.45 (0.39-0.51)	0.61 (0.56-0.67)	0.51 (0.44-0.57)

Abbreviations: CI, confidence interval; IFN-γ, interferon γ; TST, tuberculin skin test.
*IFN-γ assay cutpoint was at least 0.35 IU/mL.

JAMA 2005

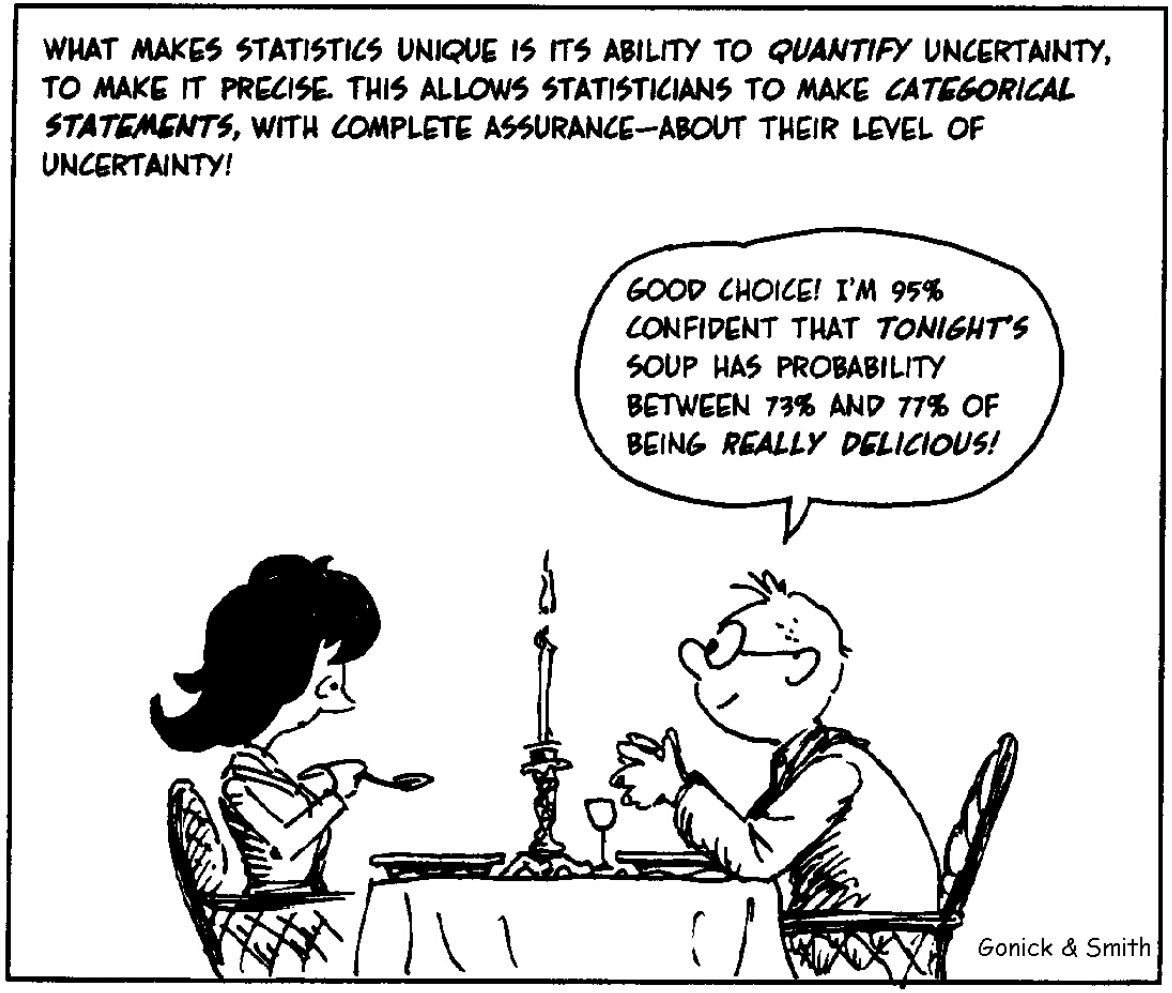
Sample size, precision and power



Madhukar Pai, MD, PhD
Assistant Professor of Epidemiology, McGill University
Montreal, Canada
Professor Extraordinary, Stellenbosch University, S Africa

Email: madhukar.pai@mcgill.ca

Key issue to understand: all measures are "estimates" [subject to error]



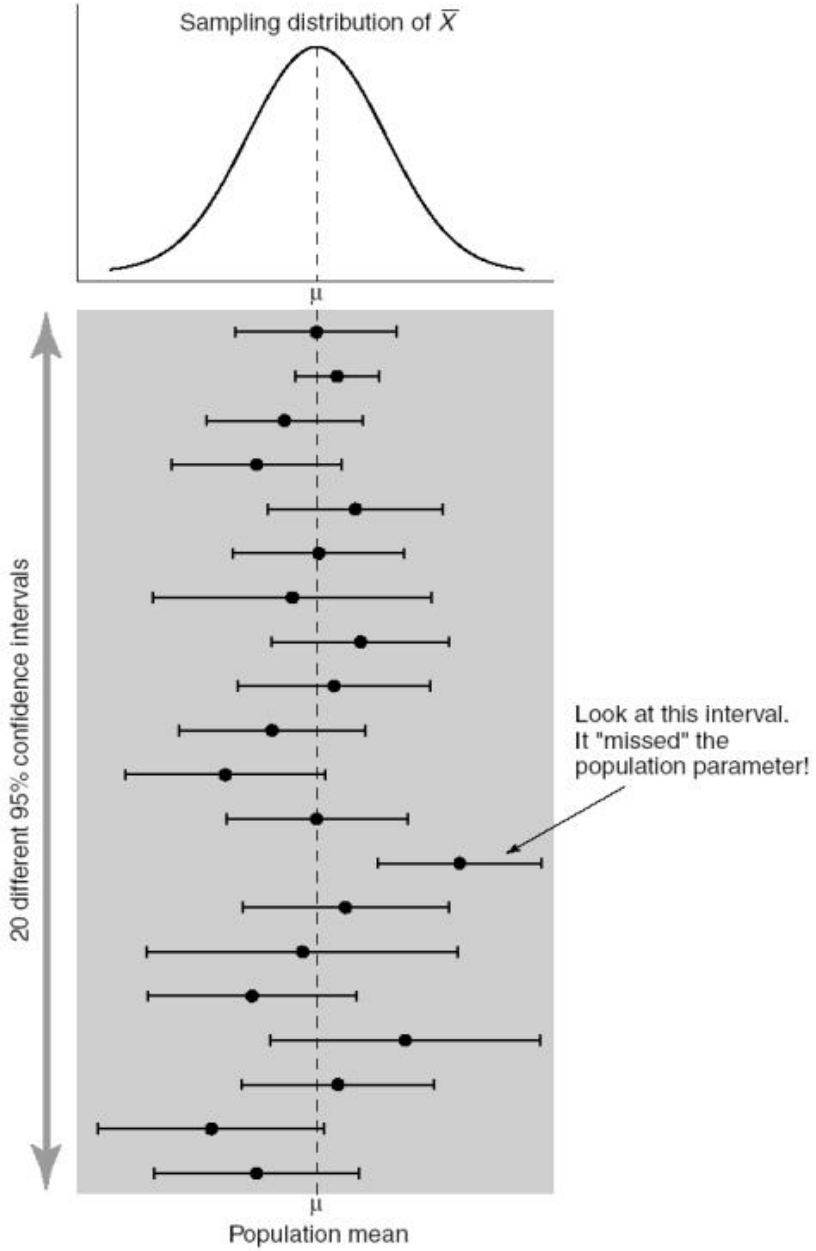


FIGURE 16.2 A sampling distribution of the mean (based on all possible samples of size 100) and an illustration of the 95 percent confidence intervals for twenty possible samples. The width of the intervals will be slightly different because they are estimated from different random samples. In the long run, 95 percent of confidence intervals will capture the population mean.

http://www.southalabama.edu/coe/bset/johnson/dr_johnson/index.htm

Therefore,
all estimates
must
be reported
with a
confidence
Intervals

CI is a measure
of “precision”

What are 95% confidence intervals?

- The interval computed from the sample data which, were the study repeated multiple times, would contain the true effect 95% of the time
- Incorrect Interpretation: "There is a 95% probability that the true effect is located within the confidence interval."
  - This is wrong because the true effect (i.e. the population parameter) is a constant, not a random variable. Its either in the confidence interval or it's not. There is no probability involved (in other words, truth does not vary, only the confidence interval varies around the truth).

Confidence Intervals for diagnostic accuracy

- Since many of the measures (sens, spec, NPV, PPV) are simple proportions, 95% CI is easy to compute (even by hand)
- For proportions:
  - General formula:
    - Proportion +/- 1.96 standard error
  - Standard error for a proportion (p):

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Does not work well for large proportions!

Need to use exact methods

Example: Serological test for TB

		Culture (gold standard)		
		Yes	No	
Serological Test	Positive	14	3	17
	Negative	54	28	82
		68	31	99

Sensitivity = 21% (95% CI 12, 31)
Specificity = 90% (95% 76, 98)

Clin Vacc Immunol 2006;13:702-03

Sample size estimation

- Depends on study design
  - If objective is sensitivity and specificity, then its simple
    - See next slides (can easily do with OpenEpi)
  - If objective is multivariable (added value of a new test), then sample size is more complicated
    - For logistic regression models, the rule of thumb is 10 disease events for every covariate in the model

Statistical Methods in Medical Research 1998; **7**: 371–392

Sample size calculations in studies of test accuracy

Nancy A Obuchowski Department of Biostatistics and Epidemiology, The Cleveland Clinic Foundation, Cleveland, Ohio, USA

Methods for determining sample size for studies of the accuracy of diagnostic tests are reviewed. Several accuracy indices are considered, including sensitivity and specificity, the full and partial area under the receiver operating characteristic curve, the sensitivity at a fixed false positive rate, and the likelihood ratio. Sample size formulae are presented for studies evaluating a single test and studies comparing the accuracy of tests. Four real examples illustrate the concepts involved in sample size determination. Lastly, various study design issues are discussed, such as sampling methods, choices in format for the test results, and the issue of replicated readings.

Table 2 General sample size formulae*

Sample size formula for studies comparing the accuracy of a single test to a null value:

$$n = \frac{\left[z_{\alpha/2}\sqrt{V_0(\hat{\theta}_1)} + z_{\beta}\sqrt{V_A(\hat{\theta}_1)}\right]^2}{(\theta_0 - \theta_1)^2} \quad \text{(T1)}$$

Sample size formula for constructing a CI of length L:

$$n = z^2_{\alpha/2} V(\hat{\theta}_1)/L^2 \quad \text{(T2)}$$

Sample size formula for studies comparing the accuracy of two tests:

$$n = \frac{[z_{\alpha/2}\sqrt{V_0(\hat{\theta}_1 - \hat{\theta}_2)} + z_{\beta}\sqrt{V_A(\hat{\theta}_1 - \hat{\theta}_2)}]^2}{(\theta_1 - \theta_2)^2} \quad \text{(T3)}$$

where

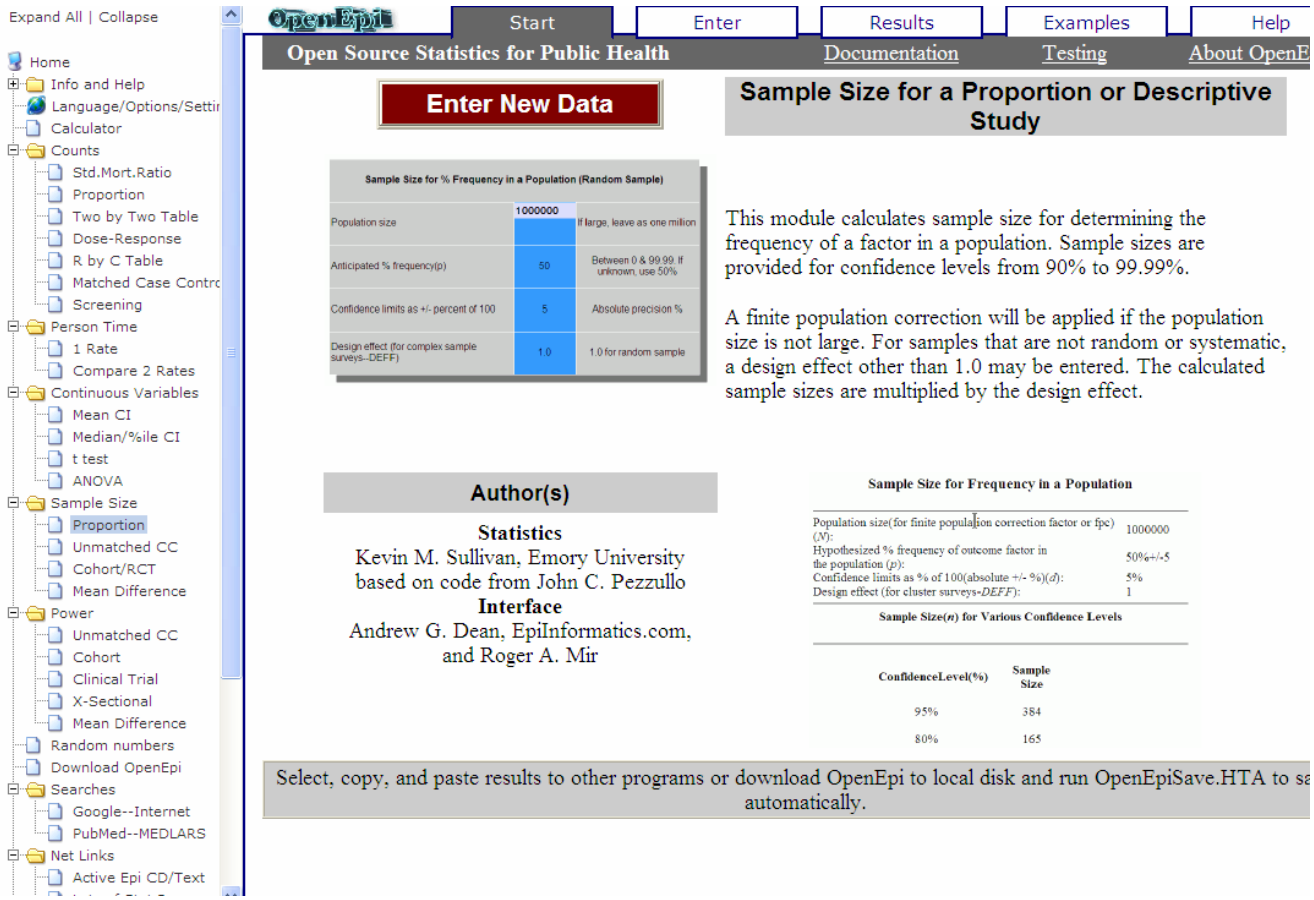
$$V(\hat{\theta}_1 - \hat{\theta}_2) = n\text{Var}(\hat{\theta}_1 + n\text{Var}(\hat{\theta}_2) - 2n\text{Cov}(\hat{\theta}_1, \hat{\theta}_2) \quad \text{(T4)}$$

*See Table 3 for definitions.

Table 3 Common notation

$\hat{\theta}_i$	estimated accuracy of test i
θ_0	prespecified value of accuracy, i.e. null value
$V_0(\hat{\theta}_i)$	$= n\text{Var}_0(\hat{\theta}_i)$, where $\text{Var}_0(\hat{\theta}_i)$ is the variance of $\hat{\theta}_i$ under the null hypothesis
$V_A(\hat{\theta}_i)$	$= n\text{Var}_A(\hat{\theta}_i)$, where $\text{Var}_A(\hat{\theta}_i)$ is the variance of $\hat{\theta}_i$ under the alternative hypothesis
$C(\hat{\theta}_1, \hat{\theta}_2)$	$= n\text{Cov}(\hat{\theta}_1, \hat{\theta}_2)$, where $\text{Cov}(\hat{\theta}_1, \hat{\theta}_2)$ is the covariance between $\hat{\theta}_1, \hat{\theta}_2$
N_D	number of subjects with the condition required for the study, i.e. patients
N_{ND}	number of subjects without the condition required for the study, i.e. controls
N	total number of subjects needed for the study, i.e. $N_D + N_{ND}$
R	ratio of sample sizes of controls to patients, i.e. N_{ND}/N_D
$z_{\alpha/2}$	upper percentile of standard normal distribution, where α = type I error rate
z_{β}	upper percentile of standard normal distribution, where β = type II error rate
$\Phi^{-1}(c)$	inverse of the cumulative normal distribution function at c
L	desired width of one-half of the CI
B	$= \sigma_{ND}/\sigma_D$, where X and Y denote the test results of controls and patients, respectively, such that $X \sim (\mu_{ND}, \sigma^2_{ND})$ and $Y \sim (\mu_D, \sigma^2_D)$
A	$= (\mu_D - \mu_{ND})/\sigma_D$

Sample size for sens/spec



Need to specify expected sens or spec, and desired precision (CI)

Sample size for LRs

J Clin Epidemiol Vol. 44, No. 8, pp. 763–770, 1991
Printed in Great Britain

0895-4356/91 $3.00 + 0.00
Pergamon Press plc

LIKELIHOOD RATIOS WITH CONFIDENCE: SAMPLE SIZE ESTIMATION FOR DIAGNOSTIC TEST STUDIES

DAVID L. SIMEL,* GREGORY P. SAMSA and DAVID B. MATCHAR

Center for Health Services Research in Primary Care, Durham Veterans Administration Medical Center and Division of General Internal Medicine, Duke University Medical Center, Durham, North Carolina, U.S.A.

(Received in revised form 19 December 1990)

Abstract—Confidence intervals are important summary measures that provide useful information from clinical investigations, especially when comparing data from different populations or sites. Studies of a diagnostic test should include both point estimates and confidence intervals for the tests' sensitivity and specificity. Equally important measures of a test's efficiency are likelihood ratios at each test outcome level. We present a method for calculating likelihood ratio confidence intervals for tests that have positive or negative results, tests with non-positive/non-negative results, and tests reported on an ordinal outcome scale. In addition, we demonstrate a sample size estimation procedure for diagnostic test studies based on the desired likelihood ratio confidence interval. The renewed interest in confidence intervals in the medical literature is important, and should be extended to studies analyzing diagnostic tests.

Best resource for confidence intervals