Agreement between two *M. tuberculosis* Screening Methods

Time to Complete Exercise: ~30 minutes

LEARNING OBJECTIVES

- Know how to construct 2-by-2 tables for reliability/agreement.
- Assess the strengths and limitation of using the kappa statistic in assessing reliability.
- Distinguish between reliability and validity.

ASPH EPIDEMIOLOGY COMPETENCIES ADDRESSED

Identify the principles and limitations of public health screening programs

AKNOWLEDGEMENT

This material was developed by George Khalil, MPH, Marian R. Passannane, Ph.D., and the staff at the Global Tuberculosis Institute (GTBI), one of four Regional Training and Medical Consultation Centers funded by the Centers for Disease Control and Prevention. The original source is:

> New Jersey Medical School Global Tuberculosis Institute. /Incorporating Tuberculosis into Public Health Core Curriculum./ Epidemiology Case Study 2: Reliability, Validity, and Tests of Agreement in M. Tuberculosis Screening. Student Version 1.0.

INTRODUCTION

In the United States, no vaccines are given to prevent the transmission of tuberculosis (TB) due to their current lack of efficacy (MMWR Recomm Rep 1998:47(RR-13):1-6). Because scientists are still working to create a more efficient vaccine, the generally accepted approach to TB control relies on screening, surveillance, and contact investigations (*MMWR Recomm Rep* 1996;45(RR-4):1-18).

Identifying and treating persons with latent TB infection (LTBI) at high risk for developing TB is part of the current TB elimination strategy in the United States. Screening is essential in the identification phase of this strategy (*MMWR Recomm Rep* 1996;45(RR-4):1-18). The most common method of screening is with purified protein derivative (PPD), a type of tuberculin skin test (TST). Infection with *Mycobacterium tuberculosis* produces a delayed-type hypersensitivity reaction to certain antigenic components of the organism that are contained in extracts of culture filtrates called tuberculin. The skin test is injected in the forearm and read by a trained clinician after 48 to 72 hours. The size of the reaction is measured in millimeters and interpreted according to its size, using cutoff points corresponding to the degree of induration (Ayub et al., 2008). Another type of screening test is called interferon gamma assay (IGA), which measures the production of the cellular interferon gamma by T-cells after sensitization with *M. tuberculosis* antigens. Although researchers believe that interferon tests are preferable to the TSTs, they are much more expensive (Madariaga et al., *J Am Board Fam Med* 2007;20:540-47). This exercise is based on the study:

Mazurek, G. H., LoBue, P. A., Daley, C. L., Bernardo, J., Lardizabal, A. A., Bishai, W. R., . . . Rothel, J. S. (2001). Comparison of a whole-blood interferon gamma assay with tuberculin skin testing for detecting latent Mycobacterium tuberculosis infection. *JAMA*, 286(14), 1740-1747. Sections of this article have been reprinted according to fair use principals. Here's the abstract:

Context Identifying persons with latent tuberculosis infection (LTBI) is crucial to the goal of TB elimination. A whole-blood interferon (IFN-γ) assay, the Quanti-FERONTB test, is a promising in vitro diagnostic test for LTBI that has potential advantages over the tuberculin skin test (TST).

Objectives To compare the IFN-γ assay with the TST and to identify factors associated with discordance between the tests.

Design and Setting Prospective comparison study conducted at 5 university affiliated sites in the United States between March 1, 1998 and June 30, 1999.

Participants A total of 1226 adults (mean age, 39 years) with varying risks of Mycobacterium tuberculosis infection or documented or suspected active TB, all of whom underwent both the IFN-γ assay and the TST.

Main Outcome Measure Level of agreement between the IFN-γ assay and the TST.

METHODS

The study was conducted at 5 sites: Boston University School of Medicine, Mass; Johns Hopkins School of Hygiene and Public Health, Baltimore, MD; University of California at San Francisco; New Jersey Medical School, Newark; and University of California at San Diego, using a common protocol. These sites were randomly coded as A-E in the analysis. Ethical approval for the study was obtained from the institutional review boards at the Centers for Disease Control and Prevention (CDC), which supported the study and the 5 study sites prior to enrolling any subjects. All participants provided written informed consent.

Persons recruited for the study were 18 years or older and included persons requesting a pre-employment or preschool enrollment TST (low-risk group) and persons being screened with a TST because they were considered to be at high risk for Latent TB Infection (high-risk group). Note that the actually study included two additional groups (persons in whom TB was clinically suspected and persons who had active TB confirmed by a positive culture). However, for the purposes of this exercise, we will examine the data from Group 1 (low-risk group for TB infection) and Group 2 (high-risk group) only.

Table 1 in the article compares the responses to both TST and IFN-γ tests for Groups 1 (low risk) and 2 (high risk).

Table 1. Response to TST and IFN-γ tests in high- and low-risk groups

	Group 1 (n = 98)	Group 2 (n = 947)
Positive TST and positive IFN-γ assay	1	146
Negative TST and negative IFN-γ assay	89	649
Negative TST and positive IFN-γ assay	7	73
Positive TST and negative IFN-γ assay	1	79

Question 1.

A. For Group 1, create a 2X2 table to assess the agreement between TST and IFN-γ readings.

	IFN-γ Positive	IFN-γ Negative	Total
TST Positive			
TST Negative			
Total			

B. For Group 2, create a 2X2 table to assess the agreement between TST and IFN-γ readings.

	IFN-γ Positive	IFN-γ Negative	Total
TST Positive			
TST Negative			
Total			

Question 2.

A. For Group 1, calculate an overall percent agreement by TST and IFN-γ assay and interpret this result (Formula 10.1).

B. Do the same for Group 2.

Question 3.

A. Calculate an expected proportion of agreement due to chance using Formula 10.2 for Group1.

B. Do the same for Group 2.

Question 4.

A. What is the kappa statistic for the test for Group 1 (low-risk TB)?

B. What is the kappa statistic for Group 2 (high-risk)?

Question 4. Use the benchmarks from Table 10.4 in the text to describe the reliability of the results in the two groups.

Question 5. Note that prior-probability ("prevalence") has a large impact on the kappa value. Read about "The kappa paradox" on pp. 227 – 228 in your text. To overcome this limitation, one can calculate one of the measures of agreement that are resistant to the kappa paradox (e.g., the Brennan-Prediger kappa, Gwet's AC1). Alternatively, on can accompany kappa with p_{pos} and p_{neg}. Calculate p_{pos} and p_{neg} (Formula 10.4 and Formula 10.5) in Group 1 and Group 2. Interpret these results.

Optional. If you are able to run WinPEPI, calculate the Brennan-Prediger kappa and Gwet's AC1. These are kappa alternatives that are resistant to the kappa paradox.

Comment: A useful screening test is both reliable and valid. These concepts are analogous to that of precision (reliability) and lack of bias (validity) addressed in Chapter 9 of *Epidemiology Kept Simple* (3e). Unfortunately, a test for validity would require a "gold standard" and in TB screening there is no such standard. Therefore, this is an exercise only looking at reliability.