Effect of Improving the Quality of Radiographic Interpretation on the Ability to Predict Pulmonary Tuberculosis Relapse

Jason E. Stout, MD, MHS, Andrzej S. Kosinski, PhD, Carol Dukes Hamilton, MD, MHS, Philip C. Goodman, MD, Ann Mosher, MPH, FNP, Dick Menzies, MD, Neil Schluger, MD, Awal Khan, PhD, John L. Johnson, MD and the Tuberculosis Trials Consortium

Rationale and Objectives: Chest radiographic findings are important for diagnosis and management of tuberculosis. The reliability of these findings is therefore of interest. We sought to describe interobserver reliability of chest radiographic findings in pulmonary tuberculosis, and to understand how the reliability of these findings might affect the utility of radiographic findings in predicting tuberculosis relapse.

Materials and Methods: Three blinded readers independently reviewed chest radiographs from a randomly selected group of 10% of HIV-seronegative subjects participating in a tuberculosis treatment trial. The three readers then arrived at a fourth, consensus radiographic interpretation.

Results: A total of 241 films obtained from 99 patients were reviewed. Agreement among the independent readers was very good for the findings of bilateral disease (κ = 0.71–0.86 among readers) and cavitation (κ = 0.66–0.73). The original interpretation was reasonably sensitive and specific (compared to the consensus interpretation) for bilateral disease, but the sensitivity for cavity decreased from 81% for the 2-month film to 47% at end of treatment (P = 0.013). Substituting the consensus interpretation for the original interpretation increased the odds ratio for the association between cavitation on early chest radiograph and subsequent tuberculosis relapse from 4.97 to 8.97.

Conclusion: Radiographic findings were reasonably reliable between independent reviewers and the original interpretations. The original investigators, who knew the patient's clinical course, were less likely to identify cavitation on the end of treatment chest radiograph. Improving the reliability of these findings could improve the utility of chest radiographs for predicting tuberculosis relapse.

Key Words: Tuberculosis; radiography, thoracic; reliability and validity.

Acad Radiol 2010; 17:157–162

From the Division of Infectious Diseases & International Health, Department of Medicine (J.E.S., C.D.H., A.M.), Department of Biostatistics and Bioinformatics (A.S.K.), and Division of Thoracic Imaging, Department of Radiology (P.C.G.), 3306-Duke University Medical Center, Durham, NC 27710; Montreal Chest Institute, McGill University, Montreal, Quebec, Canada (D.M.); Division of Pulmonary Allergy & Critical Care Medicine, Columbia University College of Physicians & Surgeons, New York, NY (N.S.); Centers for Disease Control and Prevention, Atlanta, GA (A.K.); Division of Infectious Diseases, Case Western Reserve University, Cleveland, OH (J.L.J.); the Tuberculosis Trials Consortium. USPHS/TB Trials Consortium Study 22 was sponsored by the United States Centers for Disease Control and Prevention and was funded in part through a Memorandum of Understanding between the CDC and the Washington, D.C. Veterans Affairs Medical Center. Hoechst Marion Roussel, the manufacturer of rifapentine, provided rifapentine and contributed to the cost of 3 investigator meetings, but did not participate in original or secondary analysis study design, data collection, data analysis, data interpretation, or writing of the report. Dr. Stout received salary support from the National Institutes of Health (K23AI051409). Dr. Hamilton also received salary support from the National Institutes of Health (K24AI001833). Received July 31, 2009; accepted August 10, 2009. **Address correspondence to:** J.E.S. e-mail: stout002@mc.duke.edu

doi:10.1016/j.acra.2009.08.013

Chest radiographic findings have traditionally been important for diagnosis and management of tuberculosis (TB). The current American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America guidelines for TB treatment recommend extending treatment from 6 to 9 months for patients with both a cavity on early (taken after 2 months or less of treatment) chest radiograph (CXR) and a positive sputum culture after 2 months of antituberculous therapy (1). This recommendation was based to a large extent on the findings of the Tuberculosis Trials Consortium Study 22. Study 22 was a multicenter, randomized, controlled trial that compared a standard twice-weekly TB treatment regimen during the continuation phase of therapy with a once-weekly regimen substituting a new drug, rifapentine, for the standard agent, rifampin. The final study cohort of 1004 subjects was restricted to HIV-seronegative individuals (13). Five characteristics were found to be independently associated with increased risk for disease relapse: positive sputum cultures for *Mycobacterium tuberculosis* after 2 months of antituberculous therapy; the presence of a cavity on a CXR taken

during the first 2 months of treatment, being underweight, having bilateral pulmonary involvement, and being a non-Hispanic white person. The risk for TB relapse was 23% in the group with both culture-positive sputum after 2 months of therapy plus a cavity on an early CXR, and less than 2% in the group with neither a cavity on chest film nor positive sputum cultures after 2 months of antituberculous therapy (2).

There are several reasons why it is important to better understand the reliability of CXR interpretations. First, there are now clinical ramifications for recognizing or missing the presence of a cavity on an early (taken either at initiation of treatment or at 2 months after initiation of treatment) CXR. Second, in response to the findings of Study 22, clinical trials examining new therapeutic options for treatment of TB are basing enrollment strategies and analyses on subjects' radiographic pattern. Unfortunately, several studies have described only moderate agreement among independent observers in evaluating radiographs for manifestations of TB (3–7). Furthermore, agreement may vary depending on the observer's specialty and level of experience (3,8). When the individual interpreting the chest radiograph is familiar with the patient's clinical condition, the interpretation of the chest radiograph may also be biased by the clinical data. Bias and interreader variability may both affect the utility of radiographic data in predicting TB outcomes. We therefore conducted a substudy to validate the radiographic findings of Study 22 and to assess how interreader reliability and bias might have affected the association between radiographic findings and TB relapse.

MATERIALS AND METHODS

We conducted a validation substudy of United States Public Health Service Study 22, a randomized, controlled trial comparing standard twice-weekly isoniazid and rifampin to once-weekly isoniazid and rifapentine during the continuation phase of TB treatment in HIV-seronegative subjects (2). Data from up to three standard posteroanterior chest radiographs were obtained from each subject. Chest radiographs were required within 2 weeks of randomization, which occurred after 8 weeks of standard therapy, and at the end of TB treatment. A CXR at the time of initial diagnosis was not required for the study, but if available, this radiograph was assessed in the same way as the study-required films and included in the study database. Central reading of CXRs was not performed during Study 22. The reported findings were based only on standard posteroanterior and lateral chest radiographs. Although site investigators were either chest or infectious disease physicians with considerable experience in the diagnosis and management of TB, site personnel did not receive specific study-related training on CXR interpretation or study reporting before or during Study 22. Specifically, site investigators were asked to report presence or absence of cavity based on their judgment and experience, as well as interpretations from local radiologists, but the term "cavity" was not otherwise defined. For each chest radiograph, site investigators reported three binary variables: 1) normal/abnormal; 2) presence/absence of bilateral TB disease; and 3) presence/absence of one or more pulmonary cavities. For the second and third radiographs, site investigators also reported whether there was extension of TB lesions present on a previous radiograph, and whether new TB lesions were present compared to a previous radiograph.

A simple, random sample of 20% of the original 1004 Study 22 subjects was selected for the validation substudy. The original Study 22 radiographs were mailed to Duke University Medical Center, where subject, radiograph date, and site identifiers were obscured by study personnel not involved in interpreting the radiographs. Three readers (one chest radiologist, PG, and two infectious diseases/TB clinicians, CDH and JES) independently interpreted the radiographs using the original five Study 22 variables (normal, bilateral disease, cavity, extension of old lesions, and new lesions). Readers were presented the films in random order and blinded to the original order of the films as well as to all clinical and microbiological data. All radiographs for each subject (up to three) were read at the same time by each reader, and readers were asked to make pairwise comparisons of all films to assess extension of old lesions and appearance of new lesions. After all three readers had independently interpreted the radiographs, a consensus meeting was held to resolve any differences of opinion regarding radiograph interpretation. The consensus interpretation, which was held by at least a majority of readers, was held to be the "gold standard" for purposes of this analysis.

Statistical Methods

We used the kappa statistic to describe agreement between radiographic interpretations and McNemar's chi-square test to test whether systematic differences (biases) were present between the blinded and original site radiographic interpretations (9). We compared the original sites' interpretations to the blinded consensus interpretations and summarized the findings by sensitivity and specificity, using the blinded consensus readings as the referent or gold standard. We then used a latent class model, with an assumption of independence of the blinded readings conditional on TB relapse status, to assess the validity of the consensus reading in the sampled chest radiographs (10,11). The relationship between cavity status (present/absent) and TB relapse status was summarized by odds ratios (OR) for the original site readings and for available blinded consensus readings. To estimate the relationship between the "gold standard" presence of cavity and relapse we also utilized a maximum likelihood approach (12), which simultaneously provides OR estimates for associations between TB relapse and the true cavity status (as determined by blinded consensus), as well as estimates of sensitivity and specificity of the original site readings with the consensus cavity status as the gold standard. This approach uses in one model all the original Study 22 readings and available consensus readings for those in the validation substudy. We considered a model with nondifferential misclassification of