Available online at www.sciencedirect.com
ScienceDirect
Annals of Diagnostic Pathology 14 (2010) 88–93
Annals of DIAGNOSTIC PATHOLOGY

Interobserver diagnostic variability at "moderate" agreement levels could significantly change the prognostic estimates of clinicopathologic studies: evaluation of the problem using evidence from patients with diffuse lung disease

Alberto M. Marchevsky, MD*, Ruta Gupta, MD

Department of Pathology and Laboratory Medicine, Cedars-Sinai Medical Center Los Angeles, CA 90048, USA

Abstract Does interobserver diagnostic variability (IODV) influence the accuracy of prognostic estimates of clinicopathologic studies? "Best evidence" from usual interstitial pneumonia (UIP) and nonspecific interstitial pneumonia (NSIP) patients was used to investigate the effects of IODV. Systematic literature review identified studies of UIP and NSIP providing "best evidence." Survival proportions from studies were compared using χ^2 and meta-analysis. Interobserver diagnostic variability was simulated in the data arbitrarily at 5% to 30% intervals. The various "diagnoses" were evaluated with κ, and χ^2 statistics were used to evaluate the interobserver agreement and compare survival proportions. The survival proportions for UIP and NSIP patients in 7 retrospective level III studies ranged from 11% to 58% and 39% to 100%, respectively. Analysis of simulation results with κ and χ^2 statistics showed that IODV greater than 10% resulted in significantly different survival proportion estimations. Interobserver diagnostic variability at moderate agreement levels significantly influences prognostic estimates. Evaluation and minimization of IODV in future clinicopathologic studies are indicated.

Keywords: Evidence based; Interobserver variability; Interstitial lung disease

1. Introduction

Clinicopathologic entities are usually described when a significant statistical association is established between a set of diagnostic features and survival, recurrence rates, response to a particular treatment, or other independent prognostic variables [1]. For example, nonspecific interstitial pneumonia (NSIP) was identified by Katzenstein and Fiorelli [2] in 1994 as a new variant of chronic interstitial pneumonia distinct from usual interstitial pneumonia (UIP) after noticing that patients that exhibited certain histopathologic features on open lung biopsies had better prognosis and response rates to steroid therapy than those with the UIP pattern of interstitial fibrosis. The introduction of a new clinicopathologic entity, such as NSIP, is often followed by the identification of new practical problems in research and clinical practice as additional cases that exhibit pathologic or clinical features that overlap with those seen in "older" entities such as UIP are encountered. These problems result in diagnostic variability among different pathologists. For example, Park et al [3] has shown only moderate agreement between different investigators diagnosing UIP and NSIP, with $\kappa = 0.590$ and $\kappa = 0.420$, respectively. In addition, as additional patients with NSIP have been investigated by other research groups, often very different prognostic estimates have been reported, raising questions as to the validity of the prognostic data that were used to validate the difference between 2 distinct clinicopathologic entities such as UIP and NSIP. For example, Nicholson et al [4] have reported 11% survival rates for patients with UIP, whereas Riha et al [5] have reported 58% survival rates for individuals with the same diagnosis. Nicholson et al [4]

* Corresponding author. Department of Pathology, Cedars Sinai Medical Center, Los Angeles, CA 90048, USA.
E-mail address: alberto.marchevsky@cshs.org (A.M. Marchevsky).

1092-9134/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.anndiagpath.2009.12.002

have reported 39% survival rates for patients diagnosed with NSIP, whereas Travis et al [6] have described 100% survival rates for individuals with the same diagnosis. This considerable prognostic variability could result from a variety of factors that can influence the conclusions of retrospective observational studies such as demographics, sample size, treatment effect, length of follow-up, and other features; but it also raises questions as to whether different investigators are diagnosing lung biopsies as UIP or NSIP using similar interpretations of the diagnostic criteria.

Although the problem of interobserver variability has been documented in multiple studies of diffuse lung disease, various neoplasms, and other entities, there have been no previous attempts at evaluating whether it can significantly influence the results of clinicopathologic studies [7-9]. We used a simple simulation tool and statistics to investigate the potential influence of interobserver diagnostic variability on prognostic estimates using as an example recently published data from UIP and NSIP patients in an effort to understand at what level of interobserver agreement, as measured with κ statistics, the problem of interobserver variability would significantly change the results of clinicopathologic studies.

2. Materials and methods

A previously described "Evidence-based pathology" systematic process was used [10,11]. It included formulating specific questions relevant to the problem of interobserver variability, reviewing current "best evidence" available in the literature, assessing evidence levels using the schema proposed by Sackett et al [12], and analyzing the data with statistics and meta-analysis [12]. The specific questions included the following: (1) What is the current "best evidence" available regarding the survival proportions of UIP and NSIP patients? (2) Are the survival proportions of UIP patients reported in different studies significantly different? (3) Are the survival proportions of NSIP patients reported in different studies significantly different? (4) Does meta-analysis confirm that prognosis of UIP and NSIP patients is significantly different? (5) Does meta-analysis show data heterogeneity among different studies? (6) Is it possible that if the studies providing current "best evidence" had been performed by other pathologists, their estimates of the proportions of patients surviving their disease would have been significantly different? (7) What is the level of interobserver agreement needed to minimize the effect of interobserver diagnostic variability on the prognostic estimates for UIP and NSIP patients?

A systematic review of the English literature was performed for the period 2000-2007 using the National Library of Medicine Database to identify studies that have evaluated the survival proportions of both UIP and NSIP patients. The following search terms were used: *usual interstitial pneumonia*, *UIP*, *nonspecific interstitial pneumonia*, *NSIP*, *prognosis*, and *survival*. Only studies that evaluated more than 50 cases of interstitial lung disease and reported survival information for both UIP and NSIP patients were selected as providing "best evidence." The following data were extracted from each study: number of patients diagnosed as UIP and NSIP, number of patients with UIP and NSIP that survived their interstitial lung disease, and the survival proportions of UIP and NSIP patients. The 95% confidence intervals (CIs) of all proportions were calculated using Medcalc software (Medcalc, Mariakerke Belgium) and compared with χ^2 statistics. Data were also analyzed with Comprehensive Meta-analysis version 2.0 software (Biostat, Inc, Englewood, NJ) to determine whether the survival differences between UIP and NSIP patients were significant and whether the data showed significant heterogeneity.

Five data sets were generated to simulate in a controlled manner the effects of different levels of interobserver variability on the diagnosis and prognostic estimates of UIP and NSIP patients. In each data set, the UIP or NSIP data reported in each of the studies providing "best evidence" were labeled arbitrarily as "A." These data included the total number of UIP and NSIP patients, number of patients with each condition that were reported to survive their disease, and survival percentage. Four different levels of interobserver variability were simulated by reducing the number of UIP cases by 5%, 10%, 20%, and 30% and proportionally increasing the number of NSIP cases for each study so that the total number of cases reported in each study remained constant. The data from the 4 simulated levels of interobserver variability were arbitrarily designated as "B to E." For example, the study by Parra et al [13] reported 55 UIP patients, labeled as "A." The following numbers of UIP cases were arbitrarily generated for "B to E": 52, 50, 45, and 40. The same study reported 22 NSIP patients. The following numbers of NSIP cases were arbitrarily generated for "B to E": 25, 27, 32, and 37. In sets "A-E," the total number of 77 cases remained constant. The numbers of UIP and NSIP cases from "A-E," by study, were analyzed with κ statistics using Medcalc software. The analysis was performed by comparing the data from "B to E" with the actual data (A) one set at a time in a sequential manner. κ values were interpreted using the following scale: 0.01 to 0.20, slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 0.99, almost perfect agreement [14].

Survival percentages for the "A-E" data were calculated, by study, using the number of patients that survived their disease in the actual data (A) as a constant numerator. For example, the study by Parra et al [13] reported that 20 of 55 of their UIP patients survived their disease. The actual data (A) survival percentage in this study is therefore (20/55)*100 = 36.4%. The following survival percentages were calculated for "B to E": (20/52)*100 = 38.5%, (20/50)*100 = 40.0%, (20/45)*100 = 44.4%, and (20/40)*100 = 50.0%, respectively [13]. A similar method was used to calculate the survival proportions for NSIP cases. Thus, the survival proportions for NSIP data based on