Value Based Interpretation of Pulmonary Function Tests*

Philip Harber, M.D., F.C.C.P.†

Two contrasting errors in spirometric interpretation are the inappropriate conclusion of “normality” (type I) and of “abnormality” (type II). A survey of 67 health professionals showed major interpersonal and intersituational differences in opinion about the optimal relative proportions of type I and type II errors. This suggests the need for caution in the commonly employed practice of interpretation based on a 5 percent false positive rate.

Pulmonary function tests, particularly spirometry, are generally performed to affect decisions about individuals rather than to describe populations. Alone or in combination with other factors, such tests often lead to dichotomous decisions (eg, should prednisone be prescribed for a patient with sarcoidosis, is lung function so abnormal that a worker be considered medically fit for a job?). Furthermore, significant errors may occur on either side of a choice—eg, risk of inappropriate steroid therapy vs risk of permanent fibrosis if untreated; unnecessary denial of a job versus risk of subsequent injury. However, the commonly employed conceptual basis for separating “normal” from “abnormal” is based on a statistical technique for population description and does not explicitly consider the significance of both types of errors. Optimally, an appropriate balance between both errors should be achieved, reflecting the gravity of each error and the values of the patient and physician.

Clinical decision analysis techniques are increasingly used to guide the use of clinical data. Therefore, a limited study was conducted to determine whether clinicians implicitly use a system different from the explicit statistical model commonly described and to assess the extent of interphysician differences in importance associated with the alternate errors. Use of spirometry in several occupational medicine situations was examined to assess the assumptions underlying clinical pulmonary function test interpretation and to describe the range of values held by clinicians.

METHODS

Several simple situations involving dichotomous choices based on spirometry data were presented in written form to participants in several seminars at the national meetings of the American Thoracic Society and the American Academy of Occupational Medicine in 1984. Participants could voluntarily and confidentially submit their answers. The questions are shown in the Appendix. The first (Q1) deals with preemployment testing, the second (Q2) is about transfer to a “lighter” job, and the third (Q3) concerns payment of workers’ compensation disability benefits. Each question involved possible errors in reaching conclusions from “borderline” spirometry results. A type I error was in concluding that a worker was “fit” when he actually was not. A type II error was the conclusion that a worker was “unfit” when he actually was “fit.” Participants were asked to indicate how they would prefer that their errors be distributed for each situation (Q1, Q2, Q3) by describing how 100 errors should be split between type I and type II.

The results were tabulated and analyzed with an HP-87 microcomputer (Hewlett-Packard). The frequency distribution of answers was determined for each question and described with descriptive statistics. Paired t tests and Wilcoxon tests assessed the “statistical significance” of differences in individuals’ opinions for the different situations.

RESULTS

The significance attached to the alternate possible errors is summarized in Table 1 and Figures 1 and 2. The percentage of errors that the participants would prefer to be type I errors is used to express their values. Table 1 describes the parameters of the distribution of choices. In general, participants considered both types of error to be significant. There was considerable interpersonal variation in expressed values. For each question, the answers ranged from 0 to 100. Table 1 shows large SDs as an indication of the wide variability. The large interquartile ranges (difference between 25th and 75th percentiles) confirm that the variability is not simply the result of a small number of “outliers” with very unusual views, but rather reflects large differences in opinion in the entire group.

In addition to the large interpersonal differences, individuals showed differences in their preferred error distributions for the different situations. There was a tendency to be more willing to accept type II errors when protection of a worker’s health is at stake (Question 2) than when considering compensation (Question 3) or job placement (Question 1). The last two columns of Table 1 show the mean difference and statistical significance of differences between individuals’ choices of error distributions for the three situations. Errors of inappropriately considering test results significantly abnormal were preferred more often in the

*From the Occupational Medicine Branch, Pulmonary Division, Department of Medicine, University of California, Los Angeles, and University of California Southern Occupational Health Center, Los Angeles.
†Assistant Professor of Medicine.
Supported in part by a grant from the UCLA Academic Senate.
Manuscript received April 26; revision accepted July 8.
second than the other situations. Figures 1 and 2 compare individuals’ values for Questions 1 and 2 and for Questions 2 and 3, respectively. Deviations from the diagonal indicates persons whose choice of optimal means of interpretation differ depending on the situation. The figures show that such situation-dependent value choices are used by a significant fraction of the participants.

**Discussion**

Spirometry and other pulmonary function tests are frequently used for clinical decision making. Their use may be implicit (for example, spirometry is one factor affecting the decision about use of a drug). Explicit decision-making based on spirometry is also common. For example, written policies on “physical standards” for placement are used by many employers; the Cotton Dust standard explicitly states how spirometry will be used to choose workers for entry into medical surveillance programs; and the Social Security Administration has published tables for disability benefit decisions based on spirometry.

A physician’s personal values, perception of typical values of a patient population, or understanding of a specific patient’s values might affect the choice of optimal balance between type I and type II errors. This study shows that the preferred values of persons who may utilize spirometry tests differ considerably between people and between situations. Furthermore, they consider both of the possible errors associated with a dichotomous choice to be significant. Because spirometry is regularly used to make clinical decisions about individuals, an appropriate basis for interpretation should be used. Unfortunately, interpretation of spirometric tests commonly entails a rigid system for the separation of “normal” from “abnormal” using a statistical approach that may be inconsistent with these expressed values and goals for several reasons. First, it explicitly attaches importance only to one of the two possible errors. Second, it arbitrarily chooses the frequency of the error. Third, it may be answering a different question from the one

---

**Table 1—Preferred Distribution of Errors, Expressed as Percentage of Type I**

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>25th</th>
<th>75th</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>67</td>
<td>37.3</td>
<td>25.0</td>
<td>33.8</td>
<td>5.0</td>
<td>70.0</td>
<td>10.3†</td>
<td>-6.8NS</td>
</tr>
<tr>
<td>Q2</td>
<td>67</td>
<td>27.0</td>
<td>20.0</td>
<td>28.9</td>
<td>5.0</td>
<td>45.0</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Q3</td>
<td>67</td>
<td>44.0</td>
<td>50.0</td>
<td>35.5</td>
<td>5.0</td>
<td>80.0</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

*For each question (Q1, Q2, Q3), the subjects’ preferred distribution of errors is shown as the percentage of errors which should be type I errors. N = number of respondents. The differences in choice represent Row-Column differences. Significance levels were same by paired t and Wilcoxon tests.

†p<.05.
‡p<.01.

---

**Figure 1.** Each individual’s preferred distribution of errors, expressed as preferred percentage of type I (declare “fit” when subject is not). Answer for Question 2 is on X-axis and for Question 1 is on Y-axis. Slight offsets were used where two subjects’ responses coincided.

**Figure 2.** Using the same format as Figure 1, preferred choices for Question 2 (X-axis) and Question 3 (Y-axis) are shown.
actually posed by the test requestor.

Typically, a spirometric test result such as forced vital capacity (FVC) is interpreted by comparison to a population average value (the "predicted"); generally, a "percent predicted" FVC (FVC percent) is calculated. Then, the FVC percent is compared to a single arbitrary reference point in the distribution of FVC percent results from a somewhat arbitrary reference population (that used for derivation of the prediction equation). The reference point chosen is that FVC percent value which theoretically separates the upper 95 percent from the lowest 5 percent of subjects in the reference population. If the clinical subject's FVC percent is lower than the reference point FVC percent, then he/she is considered "abnormal." If the FVC test is frequently performed, this implies that 5 percent of healthy people will be considered abnormal ("false positive" as defined by Vedal and Crapo). Thus, this method fixes the false positive rate (analogous to the type II error rate), but does not consider the frequency of the alternate error (calling a "sick" person "normal," equivalent to a type I error). It thus ignores false negative errors and is inflexible, assuming that the same (5 percent) false positive rate is appropriate in all situations.

The question commonly addressed in the formal interpretation of tests is often different from the one for which the test was ordered. In particular, tests are generally formally interpreted to answer the question, "Is the subject likely to be a member of the population sampled for derivation of the prediction equations; that is, is he/she normal?" (If unlikely to be a member of the "normal" population, by how much does he/she differ; that is, is the abnormality mild, moderate, or severe?) However, the question of interest is more often related to a clinical decision, such as whether to treat with prednisone or whether to support hiring, job transfer, or payment of disability benefits. For example, "Is this patient's FVC similar to those of the population which benefits from prednisone therapy?" or "Is this person likely to be a member of the population which is successfully working (for job placement and disability decisions)?" Clearly, the latter questions should be answered by comparison to the distribution of FVCs from a worker population (from a specific job or more generally) rather than from a population selected for absence of smoking, symptoms, or history of lung disease. Unfortunately, while worker population-specific prediction equations are sometimes developed for epidemiologic research purposes, they are currently rarely available for clinical decisions.

Participants in this study have shown differences in their views of the optimal balance between tendencies for "overinterpretation" (with a high rate of false positives or type II errors) and "overcaution" (with a high rate of type I or false negative errors). The "cautious overinterpretation" advocated by Butler may be appropriate in some situations, but the lack of any useful effect of spirometry and its interpretation may make it a mere "stage prop" in other situations. Therefore, careful thought about the goals of interpretation is necessary. A survey instrument similar to the one employed in this study may be developed to assess opinions about choice of the optimal "operating point" for any given situation along the continuum between high type I and type II error rates.

Because study participants were not selected according to a rigorous sampling plan, the choices of values of different error types summarized here may not be truly representative of those of all physicians and certainly may not represent the relative values of patients themselves. At the very least, this study suggests that there are major situational and interpersonal differences in the preferred goals of spirometric interpretation. It is therefore necessary to understand carefully the limited meaning of the term "normal" as commonly used and to recognize the arbitrary nature of the common assumption of 5 percent "false positives" without regard to the magnitude of other errors.

APPENDIX

Questions Utilized
1. Mr. X is applying for a job which involves walking and carrying. Spirometric tests are performed to see if he is "fit for the job." If the results are "borderline," you are right most of the time. Of 100 errors, how would you prefer that they are distributed?
   You say he is fit, but he is not ______%.
   You say he is not fit, but he is ______%.
   TOTAL 100%

2. Mr. X has COPD and is applying for "medical" transfer to a lighter job. Again, results are borderline. Of your errors, how would you prefer that they be distributed?
   You say he is fit, but he is not ______%.
   You say he is not fit, but he is ______%.
   TOTAL 100%

3. Mr. X is applying for disability benefits under Workers' Compensation, claiming that dust at work injured him. Again, how would you want your errors to be distributed when you are asked if he is truly "disabled?"
   You say he is fit, but he is not ______%.
   You say he is not fit, but he is ______%.
   TOTAL 100%

REFERENCES
POSTGRADUATE COURSE:
HYPERTENSION 1986: PRACTICAL ASPECTS FOR THE PRACTICING PHYSICIAN

Dates: February 6-8, 1986
Location: Sundial of Sanibel
Sanibel Island, Florida
Sponsor: American College of Chest Physicians

Course Co-Directors:
Donald G. Vidt, M.D., FCCP
Sheldon G. Sheps, M.D., FCCP
Alexander Schirger, M.D., FCCP

Course Description:
This educational program has been designed primarily to meet the needs of the practicing physician in the management of hypertension. Appropriate evaluation and treatment decisions for the hypertensive patient in 1986 will be discussed in detail. Non-drug therapy, exercise and new antihypertensive drugs will be reviewed and discussed thoroughly. Particular emphasis will be placed on the practical use of antihypertensive drugs alone and in combination, especially in special clinical situations. Ample time is planned for discussion in the form of dialogue and panel discussions. The course will also feature two evenings of clinical workshops in which registrants can interact with the faculty on selected cases. Upon completion of this course the physician should be able to apply the criteria presented for the diagnosis and management of hypertension.

Time has also been provided for registrants and their families to enjoy the many recreational opportunities available at this fine resort.

For information contact Division of Education, ACCP, 911 Busse Highway, Park Ridge, Il 60068 (312:698-2200).