Spirometry and Dyspnea in Patients With COPD*

When Small Differences Mean Little

Donald A. Redelmeier, MD, MSc; Roger S. Goldstein, MD, FCCP; Simon T. Min, MD; and Robert H. Hyland, MD, FCCP

**Objective:** To determine when a difference in FEV₁ is sufficiently large to be associated with a noticeable difference in dyspnea symptoms for patients with chronic lung disease.

**Design:** Cross-sectional analysis of 15 groups (n=112 patients, 832 contrasts).

**Setting:** Respiratory rehabilitation program.

**Patients:** Patients with COPD (mean FEV₁=35% predicted).

**Measures:** Patients' perspectives assessed through subjective comparison ratings of dyspnea and of overall health. Relation between the FEV₁ and patients' perspectives determined the smallest difference in spirometry that was associated with a noticeable difference in patients' symptoms.

**Results:** The FEV₁ was moderately correlated with patients' ratings of dyspnea (r=0.29; 95% confidence interval [CI], 0.22 to 0.35). In contrast, the FEV₁ was minimally correlated with patients' ratings of overall health (r=0.10; 95% CI, 0.03 to 0.17). The FEV₁ needed to differ by 4% predicted for the average patient to stop rating his or her dyspnea as "about the same" and start rating his or her dyspnea as either "a little bit better" or "a little bit worse" relative to other patients (95% CI, 1.5 to 6.5). This was equivalent to the average patient's FEV₁ increasing by 112 mL (starting from 975 mL and ending at 1,087 mL).

**Conclusions:** Some statistically significant differences in the FEV₁ are so small that they may not represent important differences in symptoms for the average patient with severe COPD; an awareness of the smallest difference in FEV₁ that is noticeable to patients can help clinicians interpret the effectiveness of symptomatic treatments.

*(CHEST 1996; 109:1163-68)*

**Key words:** COPD; dyspnea; spirometry

No treatment for COPD eliminates all symptoms for all patients. Some treatments can help, yet an individual's response cannot usually be predicted until therapy is tried. Also, an individual's response is usually partial, not complete. Determining whether a particular therapy is effective for a specific patient requires that clinicians followup individuals over time and monitor changes in the level of subjective and objective clinical findings. To complement the history and physical examination, spirometry is a popular method for assessing the effectiveness of treatments in patients with COPD because it provides objective data on some of the physiologic characteristics of the disease. Indeed, spirometry measures are often obtained as part of research trials to establish the effectiveness of new treatments for COPD. Yet the results of spirometry studies are often difficult to interpret.

Small differences in pulmonary physiology do not necessarily indicate important symptomatic benefits. For example, a difference in the FEV₁ of 500 mL would probably be associated with a substantial difference in dyspnea symptoms for patients with COPD, whereas a difference of 5 mL would probably not be noticed. Between these two extremes is a threshold: the point where a difference in the FEV₁ is sufficiently large to correspond to a noticeable difference in dyspnea for the average patient. When considering symptomatic treatments for COPD, clinicians need an explicit estimate of this threshold as a guide to judging whether a small difference in the FEV₁ is worthwhile. Such judgments are especially important when interpreting medical literature and identifying find-
ings that are both statistically significant and clinically significant.2

How much does the FEV₁ need to improve to signify a noticeable difference in patient symptoms? The literature provides no rigorous estimate of this threshold because of difficulties in developing a patient-centered approach to determining clinically important effect sizes.3 In particular, changes in pulmonary physiology often occur so slowly that patients cannot accurately remember their past health and judge if their dyspnea is still “about the same” or has become “a little bit better.”4 Calculating a threshold by evaluating changes in the FEV₁ relative to patients’ judgments on whether they feel the same or different may be further biased if some individuals are unavailable for follow-up or suffer acute changes. In this study, we applied a method that avoided these difficulties and we obtained an estimate of the threshold where a difference in the FEV₁ was sufficiently large to correspond to a noticeable difference in dyspnea for the average patient.

**MATERIALS AND METHODS**

**Setting**

We recruited individuals with severe COPD who were participating in a supervised respiratory rehabilitation program. Participants admitted to the rehabilitation program came from the community, were receiving maximal medical therapy, and had no other active medical conditions (including congestive heart failure, significant anemia, or substantial obesity).3 Most individuals in the rehabilitation program had been referred because of reduced exercise tolerance, inability to perform activities of daily living, and dyspnea at rest or during exertion. Individuals were excluded from the rehabilitation program if they showed poor motivation, unrealistic expectations, or inadequate comprehension. Most had smoked in the past, although all had quit prior to starting the program. Most had received oral steroids at some time in the past, although none were taking oral steroids during the program.

**Patients**

Consecutive study patients were identified by clinicians associated with the rehabilitation program and recruited by a research assistant. Individuals were excluded if they were either experiencing an acute worsening of their lung disease, expected to be discharged from the rehabilitation program within less than 5 days, or refused to provide informed consent. After being familiarized with the goals of the study, individuals were organized into one large group and observed each other during various activities for the next several days. Groups were initiated about every 2 months (as new members entered the rehabilitation program) and included consecutive patients between 1992 and 1994. All patients underwent spirometry testing prior to starting the study using a rolling seal spirometer (Sensormedics 2450 Pulmonary Function Laboratory) with results corrected for individual characteristics and reported as “percent predicted” for each patient.5

**Patients’ Judgments of Themselves**

We examined how patients judged themselves relative to others, rather than relative to their memories of the times. Patients first observed each other continually for 5 consecutive days during exercise sessions, physiotherapy classes, informal walks, social activities, and leisure time. Individuals also interacted with each other on a one-to-one basis, discussed their health, and talked about whatever issues they thought important. Subsequently, patients privately rated themselves relative to every other patient on two dimensions: breathing ability and overall health. The breathing question was: “Compared to this person, my breathing is . . . ?” The overall health question was: “Compared to this person, my overall health is . . . ?” The response categories for each question were: “Much better,” “Somewhat better,” “A little bit better,” “About the same,” “A little bit worse,” “Somewhat worse,” and “Much worse.” We refer to these judgments as subjective comparison ratings.

**Analysis of Single Contrasts**

To assess the relationship between objective and subjective measures of breathing, we contrasted the FEV₁ with the subjective comparison ratings for every possible pair of patients. First, we calculated the difference between the two patients’ spirometry measures. For example, if patient A had an FEV₁ of 40% predicted and patient B had an FEV₁ of 30% predicted, the difference was 10% predicted. Second, we identified how each patient in the pair considered his or her breathing relative to his or her partner. For example, did patient A rate his or her breathing ability as better than patient B? We evaluated all pairings in each group; for example, a group of eight patients provided seven pairings for each patient and, therefore, a total of 56 different pairings for the group (7 x 8). These pairings, in turn, would provide 56 different contrasts of breathing ability and 56 different contrasts of overall health.

**Estimating the Threshold of Clinical Importance**

To calculate the smallest difference in the FEV₁ that was noticeable to patients (the threshold of clinical importance), we compared the objective and subjective data from all contrasts. For example, we evaluated whether patients rated their breathing as about the same if the FEV₁ values were quite similar, and rated their breathing as better or worse if the FEV₁ values were quite different. A difference that was noticeable to the average patient defined the threshold of clinical importance. Specifically, we calculated how much the FEV₁ needed to differ, on average, for patients to stop rating their breathing as about the same and start rating their breathing as either a little bit better or a little bit worse. For example, if the mean difference when patients said about the same was 5% predicted, and the mean difference when patients said a little bit better was 20% predicted, the threshold from these data would be 15% predicted (20-5).

**Individual Patients Relative to the Average Patient**

Are patients similar in how they judge a difference in pulmonary function or do individual patients differ substantially from the average? To examine whether the threshold of clinical importance accurately predicted individual subjective comparison ratings, we examined the proportion of individual judgments of a little bit better, somewhat better, or much better at increasing differences in FEV₁. If patients are consistent in their judgments, the threshold of clinical importance would be sharp and most differences larger than the threshold would be associated with judgments of at least a little bit better. If patients are inconsistent in their judgments, the threshold of clinical importance would be fuzzy and many differences larger than the threshold would not be associated with judgments of at least a little bit better. Similar to evaluating dose-response relationships in pharmacology, this analysis provided an assessment of how patients appeared to respond to larger and larger differences in the FEV₁.
RESULTS

We recruited 112 patients in groups of between 5 and 13 individuals (15 groups). The typical participant was 67 years old with symptomatic COPD for 10 years. Half were men and half were women. The distributions of age, gender, and spirometry were similar to the general characteristics of patients enrolled in the respiratory rehabilitation program. Patients varied in their severity of disease, with FEV₁ values ranging from 12% predicted to 86% predicted. The mean FEV₁ was 35% predicted (SD=16), which reflected a predicted FEV₁ of 2.79 L and a measured FEV₁ of 975 mL for the average patient. The mean FVC was 68% predicted (SD=21), the mean forced expiratory flow at 75% of vital capacity (FEF75) was 10% predicted (SD=4), and the mean relaxed vital capacity (VC) was 86% predicted (SD=22).

In total, 832 contrasts were completed out of a possible 832 contrasts (100% completion rate). No data were missing. The distribution of differences in FEV₁ was bell shaped (Fig 1). The distribution of subjective comparison ratings of breathing ability was skewed (Fig 2). As in other studies, patients tended to judge themselves slightly more positively than others. Differences in patients’ FEV₁ and subjective comparison ratings of breathing ability were moderately correlated (Fig 3). Differences in the FEF75, in contrast, were not significantly correlated with subjective comparison ratings of breathing ability (Table 1). Overall, all spirometry measures were more strongly correlated with subjective comparison ratings of breathing ability than with subjective comparison ratings of overall health (Table 1).

To estimate the threshold for the FEV₁, we examined contrasts where patients judged their breathing as about the same, a little bit better, or a little bit worse. The mean FEV₁ difference for ratings of about the same was -1.7% predicted, for ratings of a little bit better it was +1.3% predicted, and for ratings of a little bit worse it was -6.6% predicted. The FEV₁ needed to differ by 3% predicted for the average patient to stop rating his or her breathing as about the same and start rating his or her breathing as a little bit better (1.3 to -1.7), and needed to differ by 4.9% predicted to stop

![Figure 1](http://journal.publications.chestnet.org/pdfaccess.ashx?url=/data/journals/chest/21731/)

**Figure 1.** Distribution of FEV₁ differences as measured in percent predicted for individual patients. The histogram summarizes 832 contrasts by the difference in the FEV₁ for the two patients. For example, a patient having an FEV₁ of 40% predicted who compared himself or herself with a patient having an FEV₁ of 30% predicted would contribute a single contrast characterized by a difference in the FEV₁ of 10% predicted.

![Figure 2](http://journal.publications.chestnet.org/pdfaccess.ashx?url=/data/journals/chest/21731/)

**Figure 2.** Distribution of subjective comparison ratings of breathing ability as judged by individual patients. The histogram summarizes 832 contrasts by the patient’s judgments of himself or herself relative to his or her partner. For example, a patient who judged his or her breathing as “somewhat better” than another patient would contribute a single contrast in ratings of breathing.

![Figure 3](http://journal.publications.chestnet.org/pdfaccess.ashx?url=/data/journals/chest/21731/)

**Figure 3.** Relationship between objective and subjective measures of breathing. The scatter plot summarizes 832 contrasts by the mean and 95% confidence interval of the mean. For example, a patient whose FEV₁ was 10% predicted better than another patient’s and who judged his or her breathing as “somewhat better” than the other patient would contribute a single contrast of breathing ability with an X-axis coordinate of +10 and a Y-axis coordinate of “somewhat better.”
rating his or her breathing as about the same and start rating his or her breathing as a little bit worse (-6.6 to -1.7). The average of the two estimates provided the threshold (4% predicted; 95% confidence interval, 1.5 to 6.5). Given that the normal FEV₁ was 2.79 L for the average patient, this estimate was equivalent to a difference of 112 mL (4.0% × 2.79).

Are some patients more sensitive to a difference in FEV₁ than others? We found that our estimate of the threshold of clinical importance was not a perfect predictor of how individual patients judged their breathing ability (Fig 4). For example, when the difference in FEV₁ was at least 4% predicted, only a small majority of the subjective comparison ratings were at least a little bit better. When the difference in FEV₁ was at least 20% predicted, about two thirds of the ratings were at least a little bit better. And when the difference in FEV₁ was at least 40% predicted, about three quarters of the ratings were at least a little bit better. A similar lack of consistency in judgments was found when we examined large differences (such as the threshold between about the same and much better) and when we examined negative differences (such as the threshold between about the same and a little bit worse).

Multivariable analysis confirmed that no simple combination of age, gender, or duration of symptoms accurately identified patients who would have a particularly high or low threshold of clinical significance for the FEV₁. The only patient characteristic that seemed potentially important was the individual's current severity of disease. The threshold for individuals with an FEV₁ below 1,000 mL tended to be smaller than the threshold for individuals with an FEV₁ above 1,000 mL, although the difference was not statistically significant (2.2 vs 5.3% predicted; p=0.09). In other words, a small difference in FEV₁ may be more noticeable for individuals with particularly poor lung function than for individuals with relatively good lung function.

**DISCUSSION**

Dyspnea in patients with chronic lung disease is a complex subjective phenomenon accompanied by complex neurophysiology. In this study, we estimated the threshold at which a difference in FEV₁ tended to be associated with a small but noticeable difference in breathing for the average patient with severe COPD. By comparing objective differences in FEV₁ to subjective comparison ratings of breathing ability, we found that the threshold for the FEV₁ was about 112 mL for patients who had an average FEV₁ of about 1 L. A difference in FEV₁ greater than 112 mL was associated with a small noticeable difference in dyspnea for the average patient, whereas a difference less than 112 mL was not associated with noticeable difference in dyspnea. Individual patients varied considerably in their judgments, however, so that differences larger than the threshold were not unanimously associated with noticeable differences in dyspnea.

How do results from our patient-centered approach compare with other estimates of the clinically important difference of the FEV₁ in patients with COPD? The American Thoracic Society has suggested that a prompt relative difference of 13% indicates a clinically important improvement. For the patients in our study who had an average FEV₁ of 975 mL, this difference is equivalent to 127 mL and is larger than our estimate. In contrast, statisticians have proposed that a difference of one fifth of an SD indicates a small clinical improvement. For the patients in our study who had an FEV₁ SD of 525 mL, this difference is equivalent to 105 mL and is smaller than our estimate. All of these

<table>
<thead>
<tr>
<th>FEV₁</th>
<th>0.29</th>
<th>0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVC</td>
<td>0.26</td>
<td>0.13</td>
</tr>
<tr>
<td>Vital capacity</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td>FEV₁/FVC</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td>FEF₇₅</td>
<td>0.01</td>
<td>0.06</td>
</tr>
</tbody>
</table>

*Values are Pearson correlation coefficients. The SE of each estimate is ±0.03. Values that exceed 0.07 are statistically significantly different from zero. Values that differ by 0.09 are statistically significantly different from each other.

**Table 1—Correlations Between Selected Spirometry Measures and Subjective Comparison Ratings**

**Figure 4.** Probability of rating "a little bit better" if the FEV₁ difference is at least a given amount. The numbers at the bottom represent the number of contrasts from which each probability estimate was obtained. For example, there were 258 contrasts in which the difference in FEV₁ values was at least 10% predicted. Of these, 148 (57%) of the subjective comparison ratings were either a little bit better, somewhat better, or much better. These data yield a point with an X-axis coordinate of 10 and a Y-axis coordinate of 57.
estimates are within the 95% confidence interval obtained from our patient-centered approach. Such agreement is encouraging, yet more research is needed to prospectively validate patient-centered methods by assessing changes in symptoms and spirometry observed when following up patients longitudinally over time.

Three limitations of our methods merit comment. First, we assessed the threshold by evaluating how patients rated themselves relative to others, not relative to their past. If individuals are more sensitive to changes in themselves than differences with others, our approach may yield an estimate of the threshold that is too large. Yet studies suggest that such overestimation is often small, given that observed thresholds were similar to established standards. 11-14 Second, we used the term “a little bit better” to denote the smallest difference important to patients. If some differences are noticeable but not important, our approach may underestimate the threshold. Yet we selected this term because it is commonly used in clinical practice. Third, we studied individuals who were motivated and cooperative; thus, our results may apply to patients in clinical trials but not all individuals with COPD. Replication in other settings should be encouraged.

We evaluated patients with extensive lung disease and do not know if the results apply to individuals with lesser degrees of severity. Indeed, our data suggest that a small change in FEV₁ may be more noticeable for individuals with particularly poor lung function than for individuals with relatively good lung function. Differences in sensitivity are likely. Differences in sensitivity are also predicted by Weber’s law. 15 Specifically, Weber’s law states that people are primarily sensitive to relative rather than absolute changes; for example, a gift of $100 tends to be more noticeable for people with annual incomes of $10,000 than for people with annual incomes of $100,000. 16 To the extent that Weber’s law also characterizes patients’ perceptions of dyspnea, we caution against applying our threshold to individuals with mild lung disease.

An explicit estimate of the minimal difference in FEV₁ that is sufficiently large to be associated with a noticeable difference in dyspnea can assist in the design of clinical trials of symptomatic treatments. Specifically, calculating an appropriate sample size requires knowing the smallest difference in outcomes that could be potentially important. For patients in this study, our results suggest that this threshold for the FEV₁ is about 112 mL. Given this quantitative estimate, analysts might calculate the number of patients needed for a trial to have sufficient statistical power for a given statistical test. To obtain 80% power using an unpaired t test and a two-tailed p value of 0.05, for example, would require about 136 total patients if the sample SD was 250 mL. 17 The appropriate sample size might be smaller, of course, if each patient’s FEV₁ was measured many times and more sophisticated statistical tests were applied.

An explicit estimate of the threshold of clinical importance also provides a rough guide for interpreting published clinical trials. Our results suggest that the threshold for the FEV₁ is about 112 mL for patients averaging about 1 L. A rigorous study of theophylline for patients with COPD indicated that high doses of this medication, relative to placebo, caused a significant improvement in the FEV₁ within 2 h of administration (970 mL vs 800 mL; p<0.05). 18 Given a threshold of 112 mL, the results suggest that the average patient would notice an improvement in dyspnea with high-dose theophylline therapy. Low-dose theophylline therapy, relative to placebo, also caused a significant improvement in the FEV₁ within 2 h of administration (830 mL vs 800 mL; p<0.05). Given a threshold of 112 mL, the results suggest that the average patient would not notice an improvement in dyspnea with low-dose theophylline therapy (unless theophylline has other effects on dyspnea). 19

Our study emphasizes that the FEV₁ is not a good predictor of an individual patient’s dyspnea. 20-22 First, the FEV₁ is not perfectly reliable and can vary by as much as 5% when tested on the same patient during the same day. 23 Second, dyspnea is not just a function of spirometric flows, but is also influenced by other clinical factors such as lung volumes, airway resistance, and respiratory muscle strength. 24 Third, the symptoms of individuals reflect both their biology and their psychology, which will vary depending on personal attitudes, expectations, and tolerance. 25 Fourth, patients’ symptomatic judgments are not always reliable and unbiased, particularly when elicited through interpersonal comparisons. 26 Our threshold, therefore, is not a substitute for asking individual patients how they feel. Our threshold is a patient-centered guide for interpreting small differences in spirometry observed in large groups of patients.

REFERENCES

3 Naylor CD, Llewellyn-Thomas HA. Can there be a more patient centered approach to determining clinically important effect sizes for randomized treatment trials. J Clin Epidemiol 1994; 47:757-95

Downloaded From: http://journal.publications.chestnet.org/pdfaccess.ashx?url=/data/journals/chester/21731/ on 06/25/2017
13 Redelmeier DA, Guyatt GH, Goldstein RS. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol (in press)
17 Hulley SB, Cummings SR. Designing clinical research. Baltimore: Williams & Wilkins, 1988