Standards and Interpretive Issues in Lung Function Testing

Robert O Crapo MD and Robert L Jensen PhD

Pulmonary function tests are most useful when performed with good technique and with an accurate system. Using standard techniques in performing the tests minimizes diagnostic and therapeutic errors. This report discusses the rationale and limits of standardization and offers practical suggestions for using available standards to increase confidence in test results. *Key words: reference standards, pulmonary function tests, diffusing capacity, spirometry.* [Respir Care 2003;48(8):764–772. © 2003 Daedalus Enterprises]

Standards for Lung Function Testing: What Do They Do?

Standardization reduces the noise in lung function measurements, thus improving the chances of correctly identifying a signal of interest. In research settings it is important to precisely and accurately identify the signal of interest and reduce the noise. What is signal and what is noise depends on the question being asked. For example, if one is trying to measure the effect of smoking on lung function, other sources of variability (eg, age, height, weight, gender) are sources of noise (Fig. 1). If the question concerns the effect of gender on lung function, the signal is gender and smoking becomes noise.

In clinical settings standards help minimize diagnostic and therapeutic errors. When a referring physician gets a pulmonary function report, he/she presumes the measured values, the predicted values, and the lower limits of the healthy-subject ("normal") range are correct. However, we know that measurement biases and errors are common and that a patient's pulmonary function test (PFT) results are not constant.[1] Table 1 (derived from work by Becklake) summarizes the sources of technical and biologic variation in spirometry, and Figure 2 illustrates the magnitude of that variability.

The sources of variability in the test for the lung's carbon monoxide diffusing capacity (D_{LCO}) include all those that operate on spirometry plus additional physiologic variables and more complicated procedural and measurement processes (Figure 3), including the patient's hemoglobin concentration; carboxyhemoglobin concentration; and the inspired oxygen pressure, which changes with altitude and the concentration of oxygen in the test gas. Table 2 illustrates the effects on D_{LCO} of hemoglobin, carboxyhemoglobin, and inspired oxygen pressure.

Standardizing requirements for instrument performance and protocols for testing is one way to reduce variability. Setting standards is an acceptable way to deal with variability, but the standards cannot be static in a changing world. Setting them involves committee choices about a

Robert O Crapo MD and Robert L Jensen PhD are affiliated with the Pulmonary Division, LDS Hospital, and the University of Utah, Salt Lake City, Utah.

Dr Crapo presented a version of this report at the 18th Annual New Horizons Symposium, Pulmonary Function Testing in 2002: Updates and Answers, October 6, 2002, at the American Association for Respiratory Care's 48th International Respiratory Congress in Tampa, Florida.

Robert O Crapo MD oversees the research and development of instruments he recommends. Robert O Crapo MD and Robert L Jensen PhD developed the D_{LCO} simulator used to test D_{LCO} systems, and they receive royalties from sales of that simulator.

Correspondence: Robert O Crapo MD, Pulmonary Division, LDS Hospital, 8th Avenue and C Street, Salt Lake City UT 84143. E-mail: ldrcrapo@ihc.com.

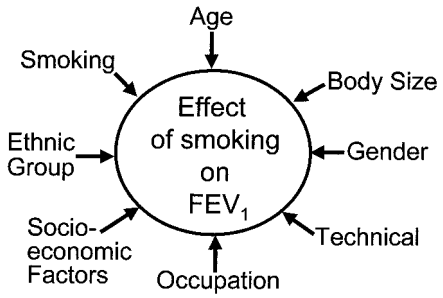

Fig. 1. Sources of noise in spirometry. This figure illustrates the difference between signal and noise. In this case, the signal sought is the effect of smoking on forced expiratory volume in the first second (FEV_1). All of the other sources of variation for FEV_1 are noise that mask the signal. If, for example, the signal of interest were the effect of age on FEV_1, the effect of smoking would then be noise.

standard method, even in the absence of compelling evidence that the method selected is better than other possible methods. In the absence of evidence the choices reflect committee members' knowledge and experience and may involve compromises and controversy. If the controversy is strong enough, it may nudge someone into doing a study to provide evidence that was lacking when the choice was made, and that research may change subsequent standards. The fact that standards evolve is a reminder not to treat them as "immutable truth."

One of the more striking examples of standards changing the environment is seen in the history of spirometry instruments. Early American Thoracic Society (ATS) recommendations for spirometer accuracy were that instruments measure forced vital capacity (FVC) and FEV_1 within the larger of ± 3% or 0.050 L and measure the forced expiratory flow in the middle half of the forced vital capacity ($FEF_{25–75\%}$) within ± 5% or ± 0.2 L/s.[2]

After the standards were set we needed some way for manufacturers to see if they were meeting them. Twenty-four standard volume-time waveforms and waveform simulators were developed to accurately deliver those waveforms.[3,4] Figure 4 shows such a simulator. In an initial study Nelson et al evaluated 62 commercially available spirometers for accuracy in measuring the standard waveforms.[5] Allowing an additional error of 0.5% for the imprecision of the waveform generator, 56.5% of the spirometers performed in the acceptable range, 14.5% were marginal, and 29% had unacceptable performance; some percent errors exceeded 50%. Figure 5 compares Nelson's initial testing results with results from all the instruments we tested in 2002.[5,6] Instrument performance has improved relatively dramatically in response to clear measurable standards and a means of testing instrument performance. Now most (but not all) instruments meet ATS standards for spirometry.

The procedural aspects of testing are at least as important—though not as easily standardized—as the instrumentation. The difficulty is that PFTs are effort-dependent and it is difficult to standardize patient effort. The ATS recommendations for spirometry require at least 3 acceptable tests and define "acceptable" as a maximum inhalation, a good start of test (with no hesitation, pauses, or substantial coughs in the first second), and a reasonable duration (defined as a plateau and at least 6 seconds of exhalation).[2] The reproducibility requirement is that the

Table 1. Sources of Technical and Biologic Variability in Pulmonary Function Tests

Technical Variability	Instrument	Instrument accuracy and imprecision (intra-instrument and inter-instrument)
	Procedure	Number of trials Choice of results to be reported
	Observer	Test administration (eg, coaching) Evaluation of results
	Subject	Comprehension Cooperation Illness
	Interactions	Between subject, observer, and instrument
	Other	Temperature Altitude
Biologic Variability	Intra-subject variation	Measurement variability and error Effects of diurnal, circadian, and seasonal changes
	Inter-subject variability	All of the above plus: 1. Personal characteristics such as body size, age, gender, degree of physical activity, muscularity, ethnicity, past and present health, and other factors. 2. Environmental factors such as smoking, occupation, residence (urban or rural), and air pollution (urban, home)
	Between-population variability	All of the above plus study inclusion and exclusion factors.

(Adapted from Reference 1.)

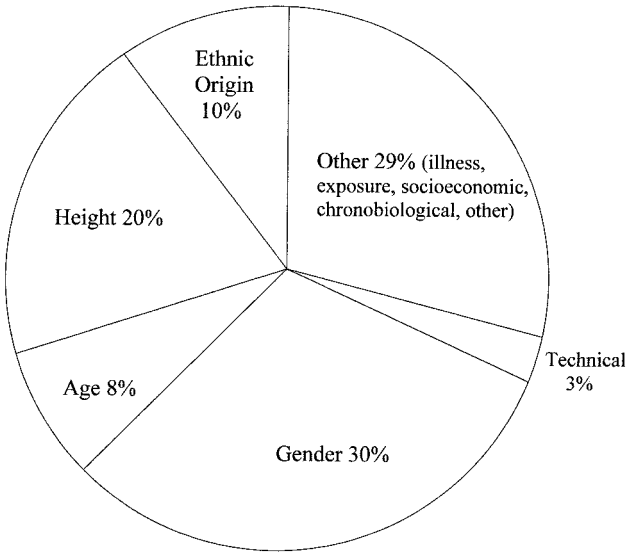

Fig. 2. Magnitude of the sources of inter-individual variability for lung function testing. Approximately 29% of the inter-individual variability for lung function testing is unexplained but is thought to relate to illness, exposure, socioeconomic factors, biologic factors, weight, and other factors. (Adapted from Reference 1.)

largest 2 values of FVC and FEV_1 be within 200 mL of each other. The requirement for 3 tests and reproducible results assures that the tests are representative of the patient. Reports from research settings show the acceptability and reproducibility criteria can be met about 90% of the time.[7–9] In a study of patients with spinal cord injuries acceptability criteria were met 77–90% of the time and reproducibility criteria were met 64–74% of the time.[10] In contrast, in 30 primary care practices in New Zealand, Eaton et al found that the full acceptability and reproducibility criteria were met only 13.5% of the time when practitioners had a relatively brief training session.[11] For those who started out with new spirometers and manuals but no formal training the full criteria were met only 3.4% of the time. Minimal criteria (at least 2 acceptable blows) were met 33.1% and 12.5% of the time in the 2 groups.

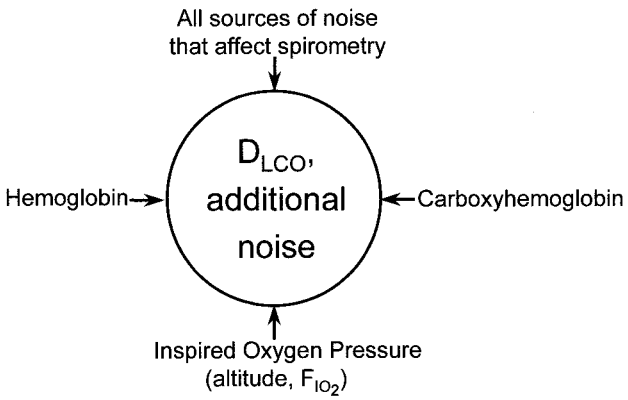

Fig. 3. Additional sources of variability in the measurement of the lung's diffusing capacity for carbon monoxide (D_{LCO}), which is subject to all of the sources of variation that affect spirometry. In addition, the technical aspects of measuring D_{LCO} are more demanding than spirometry. D_{LCO} is affected by hemoglobin and carboxyhemoglobin concentrations and by the inspired oxygen pressure (P_{IO_2}), which varies with altitude and the concentration of oxygen in the test gas (F_{IO_2}).

In pulmonary function testing, standard-setting started with the simplest, most widely used test (spirometry) and progressed to D_{LCO}. The need for such standards was manifest in an experience I had as a pulmonary fellow. My D_{LCO} was measured in 4 university-affiliated hospitals on the same day, with results ranging from 30 to 63 mL CO/mm Hg/min (coefficient of variation was about 18%).[12] I was tested again after all the labs had been brought into compliance with the then-current Intermountain Thoracic Society standards, and the coefficient of variation fell to 6%. My experience was not unique. High inter-laboratory variability for D_{LCO} has been confirmed by other investigators.[13]

D_{LCO} instrument variability can also be reduced with clear standards[14] and a system to measure how well an instrument meets the standards. We have been testing a variety of D_{LCO} systems with a D_{LCO} simulator we developed (see Fig. 4). As with that early study on spirometers, simulator testing has identified a number of problems with D_{LCO} systems, including incorrect measurement of inspired volume, carbon dioxide interference with the carbon monoxide measurements, the carbon monoxide gas analyzers, computer interference with data acquisition, errors from oscillations induced by demand valves, broken valves, plumbing problems, software errors in analyzing sensor drift, and body-temperature-and-pressure-saturated (BTPS) calculation errors. When these problems were brought to the attention of the manufacturers, they were fixed or at least did not show up on subsequent testing. We take that as evidence that they were problems of the D_{LCO} systems and not related to the simulator.

The D_{LCO} testing procedure is also important.[14] Though studies of accuracy and reproducibility for D_{LCO} testing are not as widely available as for spirometry, it is not unreasonable to expect that poor procedural processes can lead to large errors, even with excellent testing equipment.

What Can You Do?

As a respiratory therapist or pulmonary function technologist you can take the following steps to assure test quality in your laboratory.

1. Make sure your instruments meet ATS recommendations for accuracy and precision. Elements that influence purchase include ease of use, personal preferences about the displays and reports, and, of course, cost. I suggest you do not even think about those until you are sure the instrument can deliver quality data. The first question to ask a manufacturer's representative is what evidence he/she

Table 2. Effects of Hemoglobin, Carboxyhemoglobin, and Inspired Oxygen Pressure on the Lung's Diffusing Capacity for Carbon Monoxide

	Direction of Change in D_{LCO}	Magnitude of Effect on D_{LCO}
Hemoglobin	As hemoglobin concentration increases, D_{LCO} increases As hemoglobin concentration decreases, D_{LCO} decreases	3–4% per g/dL change in hemoglobin
Carboxyhemoglobin	As carboxyhemoglobin concentration increases, D_{LCO} decreases (and vice versa)	1% per 1% change in carboxyhemoglobin
P_{IO_2}	As P_{IO_2} increases, D_{LCO} decreases (and vice versa)	0.35% per mm Hg change in P_{IO_2}

D_{LCO} = diffusing capacity of the lung for carbon monoxide
P_{IO_2} = inspired oxygen pressure

has that the instrument meets performance standards. We are scientists; we need numbers and hard evidence that the system has been tested and meets ATS recommendations for accuracy and precision.[14] Many manufacturers have pulmonary function simulators, and some test each instrument before it is shipped.

2. Once the instrument has been purchased make sure it is maintained and calibrated on schedule. Set up a schedule for maintenance, calibration, and a monitoring system to make sure your instrument continues to perform within specifications.

a. The manufacturer will provide maintenance recommendations. It would be good to assure they meet the minimum recommendations set by the ATS.[14]

b. It is important to have a high-quality 3-L calibration syringe and to assure that the syringe continues to perform within its manufacturer's standards. You will want: (1) to leak test the syringe regularly, using different starting volumes, (2) to monitor the position of the collar that stops the movement of the piston (shifts in the collar position mean the syringe is out of calibration and thus will not accurately deliver the 3 L you need), (3) to assume that a calibration syringe that has been dropped has been damaged and is out of calibration until you can assure it is functioning properly.

c. Your quality monitoring system should include biologic controls. Your biologic controls will be healthy members of the laboratory staff who are regularly available for lung function measurements. A good starting program would be weekly measurements of the biologic controls, with additional measurements any time there is a question about instrument performance. Graphing the biologic control results will allow you to easily identify aberrant values and trends that suggest a problem. Once your monitoring program is established, the frequency of biologic control testing can be adjusted based on your data and on how frequently you test patients.

d. Pay attention. Other reasons to suspect an instrument problem include patient data that appear to be well out of a usual range (eg, frequent measurements in which FVC is more than 120% of predicted or in which biologic control measurements show a sudden step change, or there are step changes in a patient's PFT values that are not associated with any clinical evidence that there has been a real change). Check instrument performance and your procedures when you suspect problems.

3. Work with your medical director to assure that all staff performing tests receive feedback about the quality of the

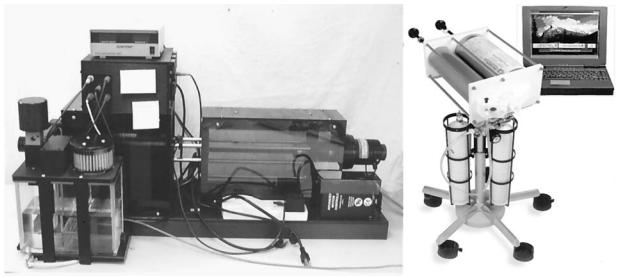

Fig. 4. Simulators used to test instruments that measure pulmonary function and the lung's diffusing capacity for carbon monoxide (D_{LCO}). Left: The latest version of the spirometry waveform simulator. Right: The recently developed D_{LCO} simulator. Simulators allow more definitive evaluation of instruments. In addition to defining whether performance is adequate, they can often identify problems. In our experience, instruments that performed poorly on first testing performed better on later testing, in part because problems had been identified in testing.

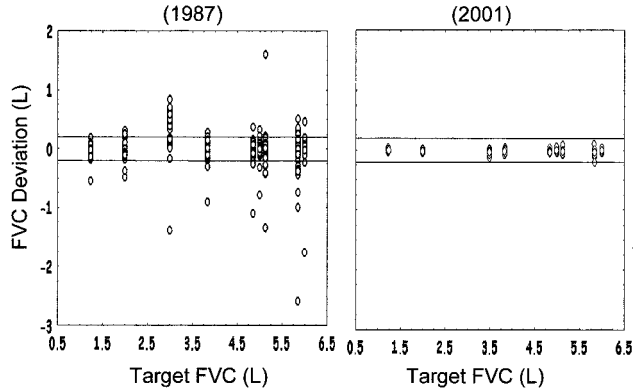

Fig. 5. Change in spirometer performance from 1987 to 2001. The 1987 data are from the instruments studied by Nelson et al.[5] The 2001 data are from the spirometers tested in our laboratory.[6] The forced vital capacity (FVC) deviation from target is in liters; a negative value indicates the values were lower than the target; the horizontal bars indicate the boundaries of acceptable performance, and positive values indicate the measured values were higher than the target. (Adapted from Reference 6.)

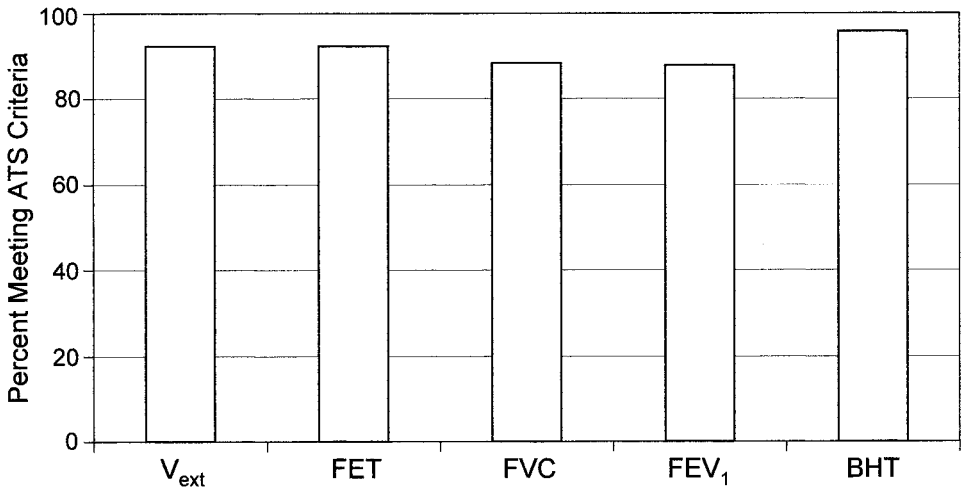

Fig. 6. Performance of the LDS Hospital pulmonary laboratory in meeting American Thoracic Society (ATS) standards. V_{ext} = extrapolated volume. FET = forced expiratory time > 6 s. Forced vital capacity (FVC) and forced expiratory volume in the first second (FEV_1) measure the frequency with which reproducibility criteria (± 200 mL) are met. BHT = breath-hold time in the test of the lung's diffusing capacity for carbon monoxide (D_{LCO}) (standard is 9–11 s). The vertical axis indicates the percent of tests that met ATS standards. Such data provide laboratories an evaluation of their current performance and a starting point for tracking performance and targeting quality control programs. (Adapted from Reference 15.)

tests they submit. Ideally feedback would occur with every test. It should at least be done regularly so technicians know when the test quality is good and when it is not.

4. Set up a monitoring system so you'll know if quality is beginning to drift down. Pick quality elements and track how frequently they are met. For example, monitor how frequently the FVC and FEV_1 meet ATS reproducibility standards or how often the inhaled volume in the D_{LCO} test exceeds 90% of the largest previously-measured FVC. Remember that not all patients will be able to do the test, so you should not expect 100% compliance with the standards. In our laboratory (in which we frequently see very sick patients) a review of 505 tests established a baseline for our lab's performance (Fig. 6).[15] Establishing a similar performance level for your laboratory will define quality control starting points and target areas for improvement. Ask the instrument manufacturer's representatives if they provide software to help you track test quality. This process will be even more effective and rewarding if there are frequent and direct interactions between the technical staff and the physicians who are reading the tests.

Lung Function Testing: Interpretive Issues for Doctors and Respiratory Therapists

The rationale for ATS's interpretive standards can be summarized by this statement from the document: "The clinical value of lung function tests is maximized when good quality tests are interpreted with appropriate reference values and appropriate interpretive schemes."[16] Choosing appropriate reference values should be a matter of careful consideration by laboratory directors and should not be defaulted to manufacturers of automatic equipment.[16] This recommendation may seem obvious, but in a survey of 139 adult pulmonary training programs, 34% of the respondents sent in only report sheets from their machines and 3 institutions stated they did not know the source of their reference equations.[17]

Clinical information does not exist in a vacuum; without comparisons, observations mean nothing. For example, knowing that a person's FEV_1 is 3.60 L has no meaning by itself. It becomes meaningful only in comparison to reference values, and those can take a variety of forms.

Comparisons can be made intuitively (based on clinical experience), can be based on knowledge of anatomy or physiology, or can emerge from a formal algorithm. They can be based on data from healthy subjects, subjects with disease, or from "ideal comparisons" that take into account information about risks.[18] The differences between a population-based reference value and an "ideal comparison" can be illustrated by considering a cholesterol level of 220 g/dL. For a 55-year-old American man eating a typical American diet, 220 mg/dL would probably be within the "normal" range of asymptomatic American males of the same age. If, on the other hand, the 220 g/dL were compared with data based on risk of cardiovascular events, that 55-year-old man would be advised to change his diet, exercise more, and possibly be given a medication.

The typical reference value comparisons used in pulmonary function interpretation include a comparison with data based on healthy subjects selected from a population, comparisons with a subject's values from previous tests,

and comparisons with disease patterns (eg, airway obstruction, chest restriction).[18] This brief summary emphasizes that reference data are not just "normal" data, and comparisons performed in the laboratory extend beyond just those with numbers derived from healthy subjects. The term "normal" is in quotes here as a reminder that using it can be misleading. PFT values labeled "normal" imply health and those labeled "abnormal" imply disease—implications that are not always correct.[18]

With that caveat, the traditional starting point for analyzing lung function tests is comparing a measured value against a representative sample of healthy, usually ambulatory, subjects drawn from a comparable population. Comparisons rest on at least these assumptions:

1. That the measurements have been made accurately and that the measured value therefore accurately represents the patient being tested at the time of the testing
2. That the reference values selected for the comparison are appropriate for the patient being tested
3. That the lower limit of normal has been properly selected
4. That an appropriate interpretive scheme is being used

The validity of the comparison can break down at each of the following points:

Accuracy and precision of the measurement. The principle is that the patients' measured values and reference values should be comparable in terms of analytical imprecision and biologic variation.[16,18] Figure 7 illustrates one way of assuring technical comparability. If the reference value study was performed with instruments and procedures that have been verified to meet ATS standards for accuracy and precision and if the patient's values are measured to the same standards, technical comparability between the reference values and the patient's values will be reasonably well established. In the laboratory, establishing that connection involves proper maintenance and daily checks to assure that instrument performance has not changed for the worse.

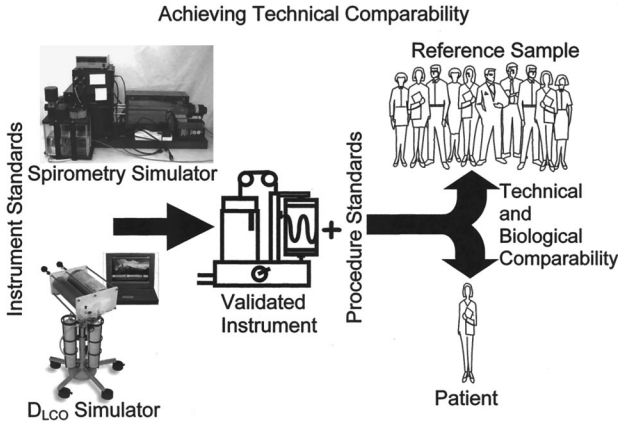

Fig. 7. Establishing technical and biologic comparability. Technical comparability can be established by ensuring that the reference values and patient values are collected with instruments and procedures that meet current standards. Establishing biologic comparability requires that the laboratory select reference values that that are biologically appropriate for the patients the laboratory serves.

Selecting biologically appropriate reference values. The patient and reference values should also be comparable with regard to biologic sources of variation.[16,18] The sources of biologic variability typically used for lung function tests are height, age, and gender; others such as ethnic group, weight, body mass index, or arm span can also be used.

Ethnic origin is an established source of biologic variability in FVC and FEV_1, though it is not clear exactly how to deal with racial, ethnic, and cultural differences. Their boundaries are vague and the differences are compounded by covariates such as socioeconomic status, educational achievement, and even ethnocentricity (the tendency to judge a culture using one's own culture as the standard).[19,20] It can affect every element of a study from design to interpretation. The best available study of ethnic differences is the National Health and Nutrition Examination Survey (NHANES III),[21] which is based on a random sample of the United States population ages 8 to 80 years; African Americans and Mexican Americans were oversampled (ie, there were more in the study than in the population at large). Subjects categorized themselves into the ethnic groups. The study sample size was large (7,429) and technical quality control was excellent. The study confirms significant differences in spirometry values for white Americans, African Americans, and Mexican Americans. Specific prediction equations were developed for each ethnic group, and the data also confirm that ethnic differences in lung function cannot be controlled by applying a single correction factor to white-based reference values (Fig. 8). NHANES III has the best available spirometry reference equations for United States citizens for the 3 ethnic groups included in the study (whites, African Americans, and Mexican Americans),[21] and it should be the starting point for spirometry reference values for laboratories in the United States. To appropriately select the ethnic group equation for your patients use the method used in the NHANES III study: that is, ask the subject to categorize himself or herself.

The differences for Asian Americans are less clear. Korotzer et al compared 40 Asian Americans with 40 European Americans and found spirometric and lung volumes to be about 7% lower in the Asian Americans for the same height.[22]

Previous values measured with a subject provide another important comparison. In fact, the best predicted value for any individual is a value measured when he/she

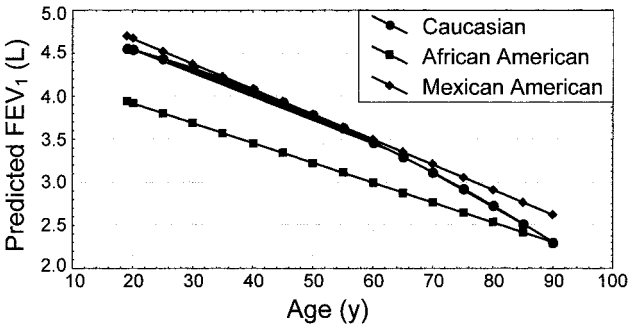

Fig. 8. Predicted forced expiratory volume in the first second (FEV_1) from the National Health and Nutrition Examination Survey (NHANES III) equations (from male subjects 175 cm tall). The figure shows that for average-size men the differences between the 3 ethnic groups are not constant with age. Similar findings would be present if change in FEV_1 with height was graphed for a fixed age. The figure illustrates the problem of adjusting white-based values with a fixed factor (eg, multiplying by 0.88). The full equations for different ethnic groups should be used. (Adapted from Reference 21.)

is a healthy adult (ie, when lung function is not increasing). Figure 9 illustrates the advantage of self-comparison. For a person of average height and age the upper and lower confidence intervals for FVC and FEV_1 are approximately ± 20% from the mean value (the inter-individual variability). For a person performing the test well, the intra-individual variability would be about ± 5%. The graph illustrates that a person whose FEV_1 started at the upper end of the distribution would have to traverse a large portion of the population variability before he/she would be classified as abnormal. This could mean a person at the upper threshold of the reference distribution could lose 40% of his/her lung function before the test results would

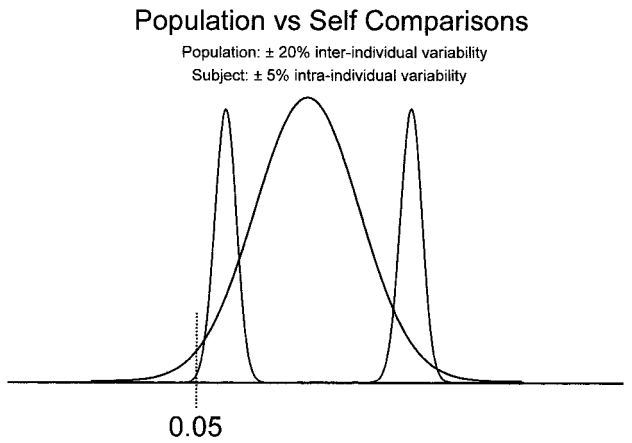

Fig. 9. There are advantages to comparing a patient's new measurements to their previous measurements. Comparisons to self (intra-individual variability) involve about one quarter of the variability of comparisons to population-based reference values (inter-individual variability).

be classified as below "normal." Eliminating inter-individual sources of variability vastly improves the ratio of signal to noise.

Despite the advantages for comparison it is not practical to obtain baseline measurements on every healthy subject. However, baseline studies can improve the ability to detect disease early in subjects with exposures known to be associated with lung injury.

The best comparison values are the subject's own PFT results from a time when the subject is healthy. Second best would be a healthy identical twin's PFT results. Third best would be PFT results from a comparable, randomly selected group of healthy subjects. Fourth best would be PFT results from a poorly selected healthy subject reference set. In the real world the best most of us can hope for is a comparison of a good measurement with a carefully selected reference set. An understanding of these limits should be reflected in increased caution in interpretation.

Lower limits of the reference range. Properly setting the limits of the reference range is just as important as selecting the reference subjects, and selecting an improper lower limit can lead to important errors in interpretation. Ideally, such boundaries should take into consideration the distributions of lung function values in both healthy subjects and patients with the relevant diseases.[18] These distributions need to be independent of each other (eg, one couldn't use spirometry to define a chronic obstructive pulmonary disease population for the purpose of setting the boundaries). Consideration of both distributions would allow boundaries to be set based on the consequences of the errors (false positive and false negative results). Unfortunately, we do not have independent distributions for the obstructive lung disease populations, since there is no clear way to define them without using lung function tests. We are, therefore, left with the less than optimal method of using a statistical estimate to infer that the patient's values are unlikely to fall within a distribution of healthy-subject values. The most common method is to define the lower or upper 95% confidence intervals or percentiles.

For spirometry the limits are usually just lower limits, but upper and lower limits may be relevant for D_{LCO} and lung volumes. The statistical lower limits of the healthy subject range are typically defined using confidence limits or percentiles. Confidence intervals are appropriate when the distributions are at least roughly gaussian (so-called normal or bell-shaped distribution). FVC and FEV_1, for example, have gaussian distributions in healthy subjects. If the distribution of healthy subjects is not gaussian (as is true for midflows such as $FEF_{25–75\%}$ and the instantaneous flows),[22] percentiles are a more appropriate way to define the lower limits of the healthy subject range.

Statistical limits of normal are currently the best available way to define subjects who are outside the range for

healthy subjects. Schemes that appear to be simpler are actually more problematic because their meaning is less clear. For example, using ± 20% of the predicted value as a "normal" range works reasonably well for patients with average heights and ages but tends to fail with subjects outside the average, and the failures worsen as distance from the average increases. A fixed ratio, such an FEV_1/FVC of 0.7, to define the presence of airway obstruction, may also lead to interpretive errors for anyone who is over 40–50 years of age. The statistically defined lower limit of normal for FEV_1/FVC drops below 0.70 in healthy women as they reach their early 50s (Fig. 10) and in healthy men in their early 40s.[23] False positive categorizations will increase with age after those ages. We do not know what happens to the false negative and true positive categorizations, because we do not have reliable methods of identifying obstructive lung diseases that are independent of lung function tests.

Statistically defined lower limits of normal for spirometric indices and D_{LCO} are our "best shot" until we have better information.

Interpretive schemes. A robust interpretive scheme is important to maximize the accuracy of interpretations. The basic elements in such a scheme are:[16]

1. Limit the number of tests analyzed
2. Use appropriate reference values and appropriate lower limits of the "normal" range
3. Use caution near the lower limits
4. Use the FEV_1/FVC ratio as the primary means of defining airway obstruction
5. Resist using $FEF_{25–75\%}$ as a measure of small airways disease in individuals
6. Use caution about inferring restriction based on spirometry alone

The ATS recommends limiting the number of tests used. For spirometry the recommendation is to focus on FVC, FEV_1, and FEV_1/FVC.[16] The number of false positive tests

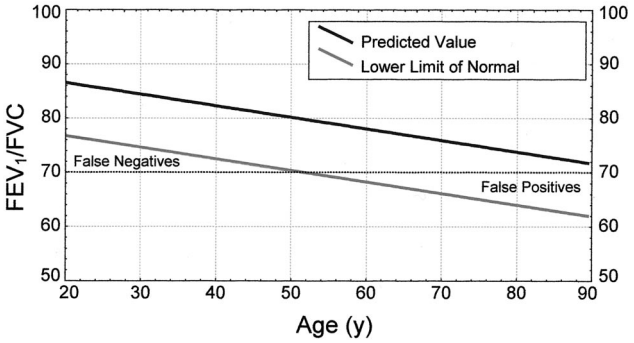

Fig. 10. The ratio of forced expiratory volume in the first second to forced vital capacity (FEV_1/FVC) in healthy white women falls below 0.70 at about age 52. This would occur in men in their early 40s. (Adapted from Reference 21.)

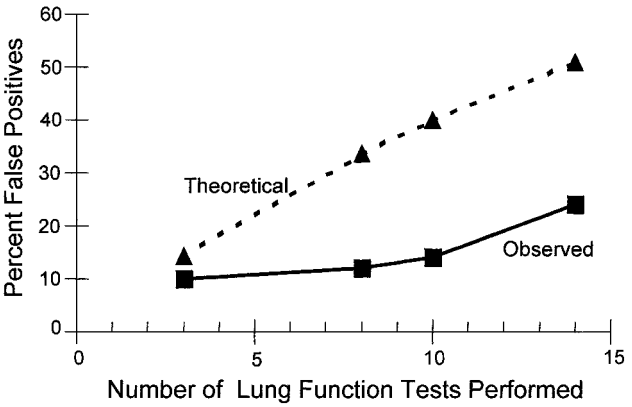

Fig. 11. In a reference value study with healthy subjects Vedal and Crapo[24] found that false positive categorizations increased significantly as the number of tests increased. The theoretical curve is what would be predicted if the lung function variables were completely independent of each other. The observed curve represents what actually happened, and it is lower than the theoretical curve because PFTs are highly intercorrelated.

increases with the number of tests performed (Fig. 11).[24] Using only FVC, FEV_1, and FEV_1/FVC, approximately 10% of healthy subjects will be classified as "abnormal" based on one of those measurements. With a full battery of lung function tests, including spirometry, D_{LCO}, and lung volumes, more than 20% of healthy subjects will be falsely classified as "abnormal."[24]

Because of uncertainties in defining both predicted values and lower limits of normal, values near the limits of normal should be interpreted with caution (Fig. 12). When a measured value lies near a threshold, the interpretation should include some estimate of the known probability of abnormality.[16] For example, a value just *below* the lower limit of normal in a healthy, asymptomatic person with no risk factors, who is being tested for insurance or employment reasons, would be more properly classified as within

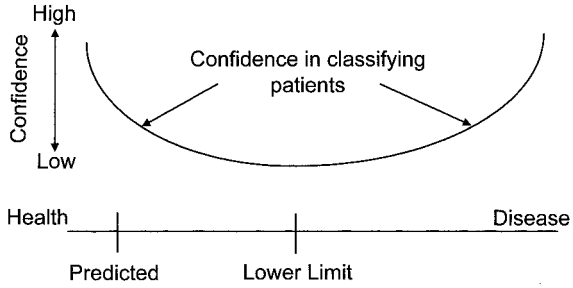

Fig. 12. A schematic of the degree of confidence one should have in interpreting pulmonary function test results, as a function of how close a measured value is to the lower limit of the distribution of the reference values. The figure illustrates the flaw in making arbitrary, univariate categorizations based on whether a data point lies above or below the "normal" threshold. Choose reference limits carefully and interpret borderline values with caution.

the normal range. On the other hand, a person with a strong smoking history and symptoms of cough and dyspnea and values just *above* the lower limit would be more accurately classified as having airway obstruction. In both instances the known probability of disease increases the chance that categorization based only on the position of the borderline value would be wrong. This becomes difficult, if not impossible, when the clinician interpreting tests knows nothing about the patient or the reason for the test. Any interpretation made in that circumstance should include a statement of caution.

Some practitioners consider a low $FEF_{25–75\%}$ in the presence of a normal FEV_1/FVC an indication of small airways disease, and, indeed, a statistical correlation has been shown in studies of groups of subjects. However, the overlap between the disease and nondisease states is large, so it is not useful for classifying individual patients.[16] It is even more problematic if the lower limit of normal is defined as 80% of predicted, because $FEF_{25–75\%}$ does not have a gaussian distribution. Percentile estimates of the lower limit of the "normal" range in middle-aged and older subjects approach 50% of predicted.[23] The ATS standards do not endorse the $FEF_{25–75\%}$ for diagnosing small airways disease in individuals and offer no way to use PFTs to diagnose small airways disease in an individual.[16]

Summary

PFT laboratories can substantially improve the value of their PFTs by paying careful attention to the key elements: (1) select good instruments and maintain them carefully, (2) carefully adhere to ATS procedural standards and monitor compliance with them, (3) select reference equations that are appropriate to the clientele your laboratory serves, (4) select appropriate lower limits of the reference ranges, (5) use appropriate interpretive schemes.

A step-by-step procedure manual for PFT has been developed by the ATS. It covers all of the routine PFTs and is updated regularly to include new and updated standards, and it can be modified to fit the specific needs of individual laboratories.[25]

REFERENCES

1. Becklake MR. Concepts of normality applied to the measurement of lung function. Am J Med 1986;80(6):1158–1164.
2. Standardization of spirometry, 1994 Update. American Thoracic Society. Am J Respir Crit Care Med 1995;152(3):1107–1136.
3. Hankinson JL, Gardner RM. Standard waveforms for spirometer testing. Am Rev Respir Dis 1982;126(2):362–364.
4. Hankinson JL, Reynolds JS, Das MK, Viola JO. Method to produce American Thoracic Society flow-time waveforms using a mechanical pump. Eur Respir J 1997;1997:10(3):690–694.
5. Nelson SB, Gardner RM, Crapo RO, Jensen RL. Performance evaluation of contemporary spirometers. Chest 1990;97(2):288–297.
6. Howell HM, Flint AK, Crapo RO, Jensen RL. Then and now: improvement in spirometer performance (abstract). J Investig Med 2002; 50:81A.
7. Hankinson JL, Bang KM. Acceptability and reproducibility criteria of the American Thoracic Society as observed in a sample of the general population. Am Rev Respir Dis 1991;143(3):516–521.
8. Enright PL, Johnson LR, Connett JE, Voelker H, Buist AS. Spirometry in the Lung Health Study. 1. Methods and quality control. Am Rev Respir Dis 1991;143(6):1215–1223.
9. Stoller JK, Buist AS, Burrows B, Crystal RG, Fallat RJ, McCarthy K, et al. Quality control of spirometry testing in the registry for patients with severe α_1-antitrypsin deficiency. α_1-Antitrypsin Deficiency Registry Study Group. Chest 1997;111(4):899–909.
10. Ashba J, Garshick E, Tun CG, Lieberman SL, Polakoff DF, Blanchard JD, Brown R. Spirometry—acceptability and reproducibility in spinal cord injured subjects. J Am Paraplegia Soc 1993;16(4): 197–203.
11. Eaton T, Withy S, Garrett JE, Mercer J, Whitlock RM, Rea HH. Spirometry in primary care practice: the importance of quality assurance and the impact of spirometry workshops. Chest 1999;116(2): 416–423.
12. Clausen J, Crapo R, Gardner R. Interlaboratory comparisons of pulmonary function testing (abstract). Am Rev Respir Dis 1984;129: A37.
13. Wanger J, Irvin C. Comparability of pulmonary function results from 13 laboratories in a metropolitan area. Respir Care 1991;36(12): 1375–1382.
14. American Thoracic Society. Single-breath carbon monoxide diffusing capacity (transfer factor): recommendations for a standard technique—1995 update. Am J Respir Crit Care Med 1995;152(6 Pt 1):2185–2198.
15. Flint A, Howell H, Jensen RL, Crapo Ro. Compliance with spirometry and DLCO quality criteria in a tertiary referral center (abstract). J Investig Med 2003(51 Suppl 2):S129.
16. American Thoracic Society. Lung function testing: selection of reference values and interpretative strategies. Am Rev Respir Dis 1991; 144(5):1202–1218.
17. Ghio AJ, Crapo RO, Elliott CG. Reference equations used to predict pulmonary function. Chest 1990;97(2):400–403.
18. Solberg HE, Grasbeck R. Reference values. Adv Clin Chem 1989; 27:1–79.
19. McKenzie KJ, Crowcroft NS. Race, ethnicity, culture, and science (editorial). BMJ 1994;309(6950):286–287.
20. Senior PA, Bhopal R. Ethnicity as a variable in epidemiological research. BMJ 1994;309(6950):327–330.
21. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general United States population. Am J Respir Crit Care Med 1999;159(1):179–187.
22. Korotzer B, Ong S, Hansen JE. Ethnic differences in pulmonary function in healthy nonsmoking Asian-Americans and European-Americans. Am J Respir Crit Care Med 2000;161(4 Pt 1):1101–1108.
23. Knudson RJ, Lebowitz MD, Holberg CJ, Burrows B. Changes in the normal maximal expiratory flow-volume curve with growth and aging. Am Rev Respir Dis 1983;127(6):725–734.
24. Vedal S, Crapo RO. False positive rates of multiple pulmonary function tests in healthy subjects. Bull Eur Physiolpathol Respir 1983; 19(3):263–266.
25. Wanger J. Editor-in Chief. Pulmonary function laboratory management and procedure manual. A project of the American Thoracic Society. 1998. Available at http://www.thoracic.org/education/labmanual.asp (accessed 6/11/03).