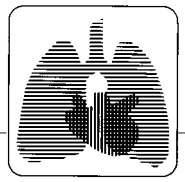

reviews

Measuring Agreement Between Diagnostic Devices*

W. Ward Flemons, MD; and Michael R. Littner, MD, FCCP

There is growing interest in using portable monitoring for investigating patients with suspected sleep apnea. Research studies typically report portable monitoring results in comparison with the results of sleep laboratory-based polysomnography. A systematic review of this research has recently been completed by a joint working group of the American College of Chest Physicians, the American Thoracic Society, and the American Academy of Sleep Medicine. The methods for comparing the results of portable monitors and polysomnography include product-moment correlation, intraclass correlation, mean differences/limits of agreement, sensitivity, specificity, and likelihood ratios. Each approach has advantages and limitations, which are highlighted in this review. ***(CHEST 2003; 124:1535–1542)***

Key words: diagnosis; likelihood ratios; methods; polysomnography; research design; sensitivity and specificity; sleep apnea syndromes

Abbreviations: AHI = apnea-hypopnea index; LR = likelihood ratio; RDI = respiratory disturbance index; ROC = receiver operating characteristic

The recommended method for diagnosing sleep apnea is polysomnography.[1] As physicians and the general population have gained awareness of sleep apnea, there has been a steadily increasing demand for the investigation of patients who are suspected of having this disorder, which in many sleep laboratories has resulted in unacceptably long waiting lists. This problem has prompted increasing interest in other possible diagnostic approaches, the most common of which is some type of portable monitoring that does not require the patient to be studied in a sleep laboratory. Typically, these devices have utilized various combinations of signals that are commonly used during polysomnography, such as oximetry alone, airflow measured by thermistor or nasal pressure, heart rate variation, snoring, or rib-cage/abdominal movement. Some monitors have continued to use EEG and electromyogram recording that allows for sleep staging and the calculation of an apnea-hypopnea index (AHI) [*ie*, the total number of apneas and hypopneas per hour of sleep time], but the majority do not. Instead, they quantitate "respiratory disturbances" (not distinguishing between apneas and hypopneas) and use total monitoring time as the quotient to determine the respiratory disturbance index (RDI). The validity of portable monitors for investigating patients with suspected sleep apnea generally has been studied by comparing their results with those of the accepted reference standard, sleep-laboratory based polysomnography. A systematic review of the research evidence on portable monitoring for investigating patients with suspected sleep apnea has been conducted by a joint working group of the American College of Chest Physicians, the American Thoracic Society, and the American Academy of Sleep Medicine using the principles outlined in this methodology review (see page 1543).

There are several commonly used approaches for assessing how well two different methods, which have been designed to measure a common variable such as breathing disturbances during sleep, agree with each other. Each method has its strengths and

*From the Faculty of Medicine (Dr. Flemons), University of Calgary, Calgary, AB, Canada; and the Veterans Affairs Greater Los Angeles Healthcare System (Dr. Littner), Sepulveda, CA.
Manuscript received May 1, 2003; revision accepted May 2, 2003.

Correspondence to: W. Ward Flemons, MD, Faculty of Medicine, University of Calgary, 1403 Twenty-Ninth St NW, Calgary, AB, Canada, T2N 2T9; e-mail: flemons@ucalgary.ca

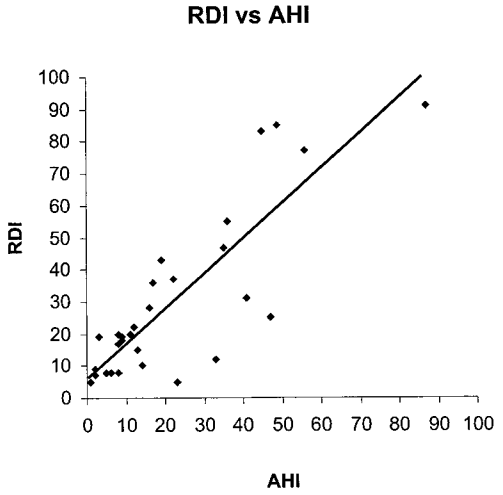

FIGURE 1. Hypothetical data on measurements of breathing disturbances during sleep by RDI recorded by a portable monitor and AHI recorded by polysomnography ($r = 0.83$).

weaknesses. Pearson product-moment correlation coefficients have as an advantage over other methods that they are in common use and that the scale is easily understood. However, they can be misleading and therefore are not recommended.[2] Intraclass correlation coefficients compare total variability among patients, measurement variability, and measurement error.[3] Statistically, they are superior to product-moment correlation coefficients. However, this approach is not intuitive to clinicians and is not commonly used. Calculating the mean differences between two methods of measurement is useful and is preferable to correlation. However, the limits of agreement, the key descriptor that relates how well the measures agree, can be misleading if not calculated properly.

Ultimately, a clinician could accept that the measurement of breathing events using a portable monitor does not agree completely with polysomnography as long as it classifies patients accurately as those with and those without sleep apnea. For these purposes, using sensitivity, specificity, and likelihood ratios (LRs) is more appealing. However, this approach dictates that a patient be classified as having or not having the disorder based on an arbitrary cutoff for the AHI that is variable across studies. There is a wide spectrum of severity of breathing events at night, and the AHI captures only a single dimension. Since a large number of patients have index values around the usual cutoff point, it is possible that a patient's classification could change due to expected variability in the measurement. By dichotomizing results into simply positive or negative, a substantial proportion of information is lost, in particular, information that could better classify a patient as having mild, moderate, or severe disease.

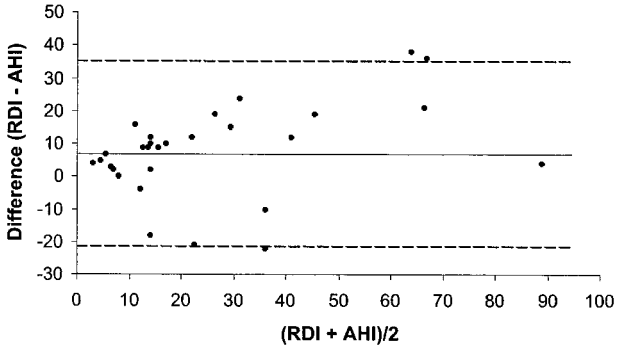

FIGURE 2. Differences (RDI − AHI) plotted against the mean of RDI and AHI. The solid line is the mean difference, and the dotted lines are the limits of agreement.

CORRELATION ANALYSIS

Measuring agreement between two different clinical measurements using a product-moment correlation coefficient can be misleading, since it is only a measure of the strength of a relation. Two methods

Table 1—*Calculating Sensitivity, Specificity, Positive and Negative Predictive Values, and the Effect of Prevalence

DT	RS		Total
	Positive	Negative	
10% prevalence†			
Positive	90‡	100§	190
Negative	10‖	800¶	810
Total	100	900	1,000
50% prevalence#			
Positive	450	55	505
Negative	50	445	495
Total	500	500	1,000

*DT = diagnostic test; RS = reference standard. TP = true-positive; FP = false-positive; FN = false-negative; TN = true-negative. Sensitivity = TP * 100/TP + FN; specificity = TN * 100/TN + FP; positive predictive value = TP * 100/TP + FP; negative predictive value = TN * 100/TN + FN.

†In this hypothetical example, 90 of 100 patients who have a disease (prevalence, 10%) defined by a positive RS test have a positive diagnostic test (sensitivity, 90.0%) and 800 of 900 of those who do not have the disease have a negative test result (specificity, 88.9%). The positive predictive value is 90/190 or 47.3%. The negative predictive value is 800/810 or 98.8%.

‡True-positive result.

§False-positive result.

‖False-negative result.

¶True-negative result.

#In this example, prevalence has increased to 50% with no change in sensitivity or specificity. However, the positive predictive value has increased substantially to 89.1%, and the negative predictive value has dropped to 89.9%.

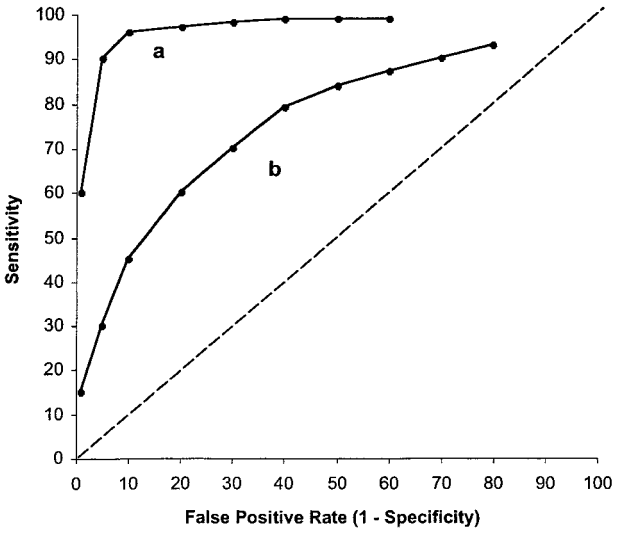

FIGURE 3. ROC curve. Different thresholds for positive diagnostic test produce different combinations of sensitivity and specificity. The dashed line indicates a test that does not alter the pretest probability. The operating characteristics of test a are superior to that of test b, since there are several thresholds that produce a high sensitivity and specificity, and there is greater area under the curve.

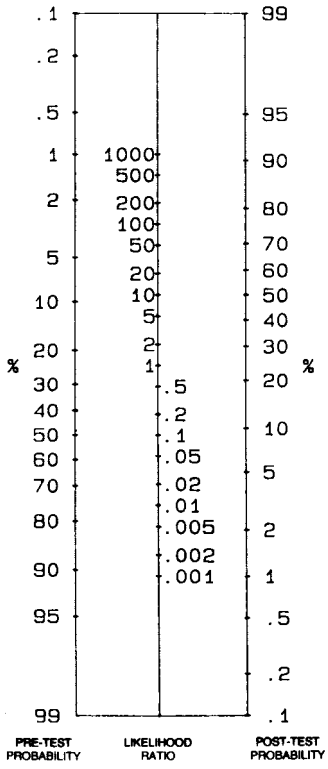

FIGURE 4. A nomogram for converting pretest to posttest probability (probabilities are listed as percentages) using LRs. To use the nomogram, anchor a straight edge at the pretest probability and direct it through the appropriate LR. The intersection of the straight edge with the third (right) line produces the probability result. In this example, a pretest probability of 50% combined with an LR of 20 increases the probability of disease to 95%.

may correlate perfectly but have different scales of measurement, in which case they do not agree. Furthermore, this type of correlation depends on the range of values that are being compared (the wider the range, the stronger the correlation), yet this does not necessarily reflect greater agreement between two methods. For these reasons product-moment correlation coefficients are not recommended as a statistic to describe how well two methods of measurement agree.[2] A hypothetical example of the comparison between the RDI measured by a portable monitor and the AHI determined from polysomnography is shown in Figure 1. The product-moment correlation coefficient is high ($r = 0.83$), highly statistically significant, and would suggest that the two methods have excellent agreement. However, this correlation coefficient is able to indicate only that the two measurements are related.

MEAN DIFFERENCE AND LIMITS OF AGREEMENT (BLAND-ALTMAN)

A widely accepted method of measuring agreement is the approach proposed by Bland and Altman[2] in which the difference between the two measurements for each subject is determined. The mean difference provides an estimate of whether the two methods, on average, return a similar result. A mean difference other than 0 suggests a systematic bias in the way that one method is measuring the clinical result of interest (*eg*, the RDI). The measure of agreement is based on calculating the SDs of the mean difference. The limits of agreement have been defined as ± 2 SDs. How far apart these limits of agreement should be for a measurement method is a question of judgment. Some authors calculate 95% confidence intervals of the estimate of the mean difference and report this rather than the limits of agreement. The two values are related but are not the same. The limits of agreement will be larger. Figure 2 plots the same data that are presented in Figure 1. Each data point represents the difference between the RDI and the AHI for an individual patient. This is plotted against the average value (*ie*, RDI + AHI/2) for each patient. With this presentation of the data, it is easy to observe that the mean difference is approximately 7.6 (indicating a systematic bias) and that the limits of agreement are wide (*ie*, discrepancies of up to 28 events per hour), indicating that there is a substantial lack of agreement.

The limits of agreement may be misleading in the case in which the difference between measurements varies in a systematic way over the range of measure-

Table 2—*Impact of Pretest Probability and LRs on Posttest Probability*

Pretest Probability, %	LR	Posttest Odds	Posttest Probability, %
10	0.05	0.01	0.6
10	0.1	0.01	1.1
10	0.2	0.02	2.2
10	0.4	0.04	4.3
10	0.7	0.08	7.2
10	2	0.22	18.2
10	5	0.56	35.7
10	10	1.11	52.6
10	20	2.22	69.0
20	0.05	0.01	1.2
20	0.1	0.03	2.4
20	0.2	0.05	4.8
20	0.4	0.10	9.1
20	0.7	0.18	14.9
20	2	0.50	33.3
20	5	1.25	55.6
20	10	2.50	71.4
20	20	5.00	83.3
30	0.05	0.02	2.1
30	0.1	0.04	4.1
30	0.2	0.09	7.9
30	0.4	0.17	14.6
30	0.7	0.30	23.1
30	2	0.86	46.2
30	5	2.14	68.2
30	10	4.29	81.1
30	20	8.57	89.6
40	0.05	0.03	3.2
40	0.1	0.07	6.3
40	0.2	0.13	11.8
40	0.4	0.27	21.1
40	0.7	0.47	31.8
40	2	1.33	57.1
40	5	3.33	76.9
40	10	6.67	87.0
40	20	13.33	93.0
50	0.05	0.05	4.8
50	0.1	0.10	9.1
50	0.2	0.20	16.7
50	0.4	0.40	28.6
50	0.7	0.70	41.2
50	2	2.00	66.7
50	5	5.00	83.3
50	10	10.00	90.9
50	20	20.00	95.2
60	0.05	0.08	7.0
60	0.1	0.15	13.0
60	0.2	0.30	23.1
60	0.4	0.60	37.5
60	0.7	1.05	51.2
60	2	3.00	75.0
60	5	7.50	88.2
60	10	15.00	93.8
60	20	30.00	96.8
70	0.05	0.12	10.4
70	0.1	0.23	18.9
70	0.2	0.47	31.8
70	0.4	0.93	48.3
70	0.7	1.63	62.0
70	2	4.67	82.4
70	5	11.67	92.1
70	10	23.33	95.9
70	20	46.67	97.9
80	0.05	0.20	16.7
80	0.1	0.40	28.6
80	0.2	0.80	44.4
80	0.4	1.60	61.5
80	0.7	2.80	73.7
80	2	8.00	88.9
80	5	20.00	95.2
80	10	40.00	97.6
80	20	80.00	98.8
90	0.05	0.45	31.0
90	0.1	0.90	47.4
90	0.2	1.80	64.3
90	0.4	3.60	78.3
90	0.7	6.30	86.3
90	2	18.00	94.7
90	5	45.00	97.8
90	10	90.00	98.9
90	20	180.00	99.4

ments.[3] A comparison of measurements of breathing disturbances often shows larger differences as the AHI increases. This results in limits of agreement that are too wide for small values of the AHI and are not wide enough for higher values. Therefore, the use of a logarithmic transformation of differences between polysomnography and portable monitors is recommended, but this is rarely done in practice. When it is not done, the quoted limits of agreement can be misleading and should be interpreted with caution.

Sensitivity/Specificity

The operating characteristics of tests are often summarized as sensitivity (*ie*, the proportion of patients with disease who have a positive test result, or the *true-positive* rate) and specificity (*ie*, the proportion of patients without disease who have a negative result, or the *true-negative* rate). Using sensitivity and specificity to describe the utility of a diagnostic test has some limitations, since they indicate the probability that the test result will be positive if the patient has the disease, and the probability that the test will be negative if the patient does not have the disease. Clinicians cannot apply these numbers directly, because they do not know whether or not the patient has the disease. What the physician wants to know is conditional, the probability that the patient has the disease if the test is positive or negative (*ie*, the positive and negative predictive values of the test, respectively). Sensitivity and specificity can be determined by analyzing the

columns in a 2 × 2 table (Table 1), while the positive and negative predictive values are obtained by analyzing the rows. By convention, the reference standard is at the top and the new diagnostic test that is being compared to it is on the side. For sleep apnea, the reference standard is a definition of sleep apnea based on the AHI (the most common cutoffs used are 10 or 15), and the diagnostic test values are the result of the portable monitor. Sensitivity, specificity, prevalence (or pretest probability), and predictive values provide valuable information about a diagnostic test. However, it can be a challenge to interpret several different numbers that all describe the operating characteristics and outcomes of a test.

In the first example in Table 1, if the diagnostic test was being used to exclude sleep apnea, 81% of the tests would be negative and only 10 of 810 of those negative test results (or 1%) would be false-negative. If the test was being used to confirm a diagnosis of sleep apnea, only 19% of the tests would be positive, and of those testing positive, more than half (53%) would be false-positive. If the same test was used in the second example in Table 1, in which the prevalence (*ie*, pretest probability) is much higher, the number of negative results would be much lower (49.5%), and the percentage of false-positive results would rise to 10%. However, in the second example in Table 1 the test would have more usefulness to rule in the disorder (patients testing positive, 50.5%; false-positive results, 11% [of those testing positive]).

When two methods of measurement do not completely agree, the potential user of the test should understand the interaction of sensitivity, specificity, and pretest probability that will dictate the number of tests that will, on average, come back positive or negative and the percentage of times that a positive result will be false-positive, and a negative result will be false-negative. The thresholds of sensitivity and specificity that dictate the ability of a test to exclude or confirm a diagnosis in a substantial percentage of cases and the acceptable rate of false results will be affected by several factors, such as the potential risk to a patient of having test results being labeled false-negative or false-positive. In the former case, it could potentially deny the patient a trial of beneficial therapy. However, this risk could be reduced if symptomatic patients were offered a second test or a polysomnogram. In the latter case, the patient may be offered a trial of therapy in circumstances in which it may not otherwise be indicated. The risk associated with this in sleep apnea is likely to be small, but unless the trial was conducted with polysomnography, it would be difficult to determine whether the use was warranted.

Changing the threshold of what constitutes a normal or abnormal diagnostic test result will change the sensitivity and specificity. Lowering the threshold will increase sensitivity but lower specificity, resulting in more true-positive results (and therefore fewer false-negative results) but also in more false-positive results (and fewer false-negative results). The converse (*ie*, increasing the threshold) will have the opposite effect (*ie*, lower sensitivity and increased specificity). A threshold that results in a low false-negative rate (*ie*, high sensitivity) is useful to exclude disease, but very few patients may actually have a negative result, so, practically, the impact of

Table 3—*Combinations of Sensitivity and Specificity with Corresponding Positive and Negative LRs**

		LR	
Sensitivity	Specificity	Pos	Neg
98	98	49.0	0.02
98	95	19.6	0.02
98	90	9.8	0.02
98	85	6.5	0.02
98	80	4.9	0.03
98	75	3.9	0.03
98	70	3.3	0.03
95	98	47.5	0.05
95	95	19.0	0.05
95	90	9.5	0.06
95	85	6.3	0.06
95	80	4.8	0.06
95	75	3.8	0.07
95	70	3.2	0.07
90	98	45.0	0.10
90	95	18.0	0.11
90	90	9.0	0.11
90	85	6.0	0.12
90	80	4.5	0.13
90	75	3.6	0.13
90	70	3.0	0.14
85	98	42.5	0.15
85	95	17.0	0.16
85	90	8.5	0.17
85	85	5.7	0.18
85	80	4.3	0.19
85	75	3.4	0.20
85	70	2.8	0.21
80	98	40.0	0.20
80	95	16.0	0.21
80	90	8.0	0.22
80	85	5.3	0.24
80	80	4.0	0.25
80	75	3.2	0.27
80	70	2.7	0.29
75	98	37.5	0.26
75	95	15.0	0.26
75	90	7.5	0.28
75	85	5.0	0.29
75	80	3.8	0.31
75	75	3.0	0.33
75	70	2.5	0.36

*Pos = positive; Neg = negative.

Table 4—*Hypothetical Results on 1,000 Patients Who Have Undergone a Polysomnogram (AHI) and a Portable Monitor Test (RDI)**

	AHI	
RDI	≥ 10†	< 10‡
≥ 40	178	8
20–39	129	31
10–19	53	33
5–9	27	189
< 5	13	339
	400	600

*Values given as No. of patients, unless otherwise indicated.
†Positive for sleep apnea.
‡Negative for sleep apnea.

performing the test on a population of patients may be small. The converse is true for setting the threshold of a test quite high in order to improve specificity. Very few patients actually may receive a positive result. A receiver operating characteristic (ROC) curve often is used to illustrate the effect of changing thresholds for a positive diagnostic test result (Fig 3).

LRs

Although sensitivity and specificity are more likely to be used to infer the utility of a diagnostic test to exclude or confirm a disease, either of these statistics, when considered in isolation, can be misleading. This is because positive and negative predictive values depend on the combination of sensitivity and specificity. The utility of a test for excluding or confirming a disorder can be captured in a single number, the LR. The LR for a positive test result is the ratio of the proportion of patients with disease who have a positive test result (*ie*, the true-positive rate or sensitivity) to the proportion of people without disease who have a positive test result (*ie*, the false-positive rate).[4] Similarly, the LR for a negative test result is the ratio of the proportion of patients with disease who have a negative test result (*ie*, the false-negative rate) to the proportion of people without disease who have a negative test result (*ie*, the true-negative rate or specificity). Using the example of the 2 × 2 table in the first example in Table 1, the LR for a positive result is calculated as 0.90/0.11 = 8.1. The LR for a negative result is calculated as 0.10/0.89 = 0.11. Mathematically, when using LRs to convert pretest to posttest probabilities, the pretest probability estimate (*ie*, the estimated prevalence) is first converted to an odds expression (*ie*, pretest odds = pretest probability/1 − pretest probability), then is multiplied by the LR to obtain the posttest odds, which then are converted back to a probability statement (*ie*, posttest probability = posttest odds/posttest odds + 1). This process can be greatly simplified with the use of a nomogram (Fig 4).[5] The nomogram and Table 2 also highlight the interaction between pretest probability and LR on posttest probability. Of note, the posttest probabilities are equal to the positive predictive value and 100 minus the negative predictive value (expressed as the percentages).

The LR in a single number captures the utility of a test to change the probability of disease, and

Table 5—*The Effect on Sensitivity, Specificity, Positive and Negative Predictive Values and False-Negative and False-Positive Results of Changing the RDI Threshold for a Positive Result*

	AHI			
RDI	≥ 10*	< 10†	Total	LR
≥ 40‡§	178	8	186	44.5/1.3 = 34.2
< 40‖	222	592	814	54.5/98.7 = 0.55
Total	400	600	1,000	
≥ 20‡¶	307	39	346	76.8/6.5 = 11.8
< 20‖	93	561	654	24.2/93.5 = 0.26
Total	400	600	1,000	
≥ 10‡#	360	72	432	90.0/12.0 = 7.5
< 10‖	40	528	568	10.0/88.0 = 0.11
Total	400	600	1,000	
≥ 5‡**	387	261	648	96.8/43.5 = 2.2
< 5‖	13	339	352	3.2/56.5 = 0.06
Total	400	600	1,000	

*Positive for sleep apnea.
†Negative for sleep apnea.
‡Positive result.
§For RDI cutoff of 40: sensitivity, 178/400 (44.5%); specificity, 592/600 (98.7%); positive predictive value, 178/186 (95.7%); negative predictive value, 592/81 (72.7%); patients testing positive, 186/1,000 (18.6%); patients testing positive with a false-positive result, 8/186 (4.3%); patients testing negative, 814/1,000 (81.4%); patients testing negative with a false-negative result, 222/814 (27.3%).
‖Negative result.
¶For RDI cutoff of 20: sensitivity, 307/400 (76.8%); specificity, 561/600 (93.5%); positive predictive value, 307/346 (88.7%); negative predictive value, 561/654 (85.8%); patients testing positive, 346/1000 (34.6%); patients testing positive with a false-positive result, 39/346 (11.3%); patients testing negative, 654/1,000 (65.4%); patients testing negative with a false-negative result, 93/654 (14.2%).
#For RDI cutoff of 10: sensitivity, 360/400 (90.0%); specificity, 528/600 (88.0%); positive predictive value, 360/432 (83.3%); negative predictive value, 528/568 (93.0%); patients testing positive, 432/1,000 (43.2%); patients testing positive with a false-positive result, 72/432 (16.7%); patients testing negative, 568/1,000 (56.8%); patients testing negative with a false-negative result, 40/568 (7.0%).
**For RDI cutoff of 5: sensitivity, 387/400 (96.8%); specificity, 339/600 (56.5%); positive predictive value, 387/648 (59.7%); negative predictive value, 339/352 (96.3%); patients testing positive, 648/1,000 (64.8%); patients testing positive with a false-positive result, 261/648 (40.2%); patients testing negative, 352/1,000 (35.2%); patients testing negative with a false-negative result, 13/352 (3.7%).

therefore is recommended over sensitivity and specificity for this purpose. The relationship of LRs to different combinations of sensitivity and specificity is shown in Table 3. A guide to the interpretation of LRs is given as follows: < 0.05, very large reduction; 0.05 to 0.1, large reduction; 0.1 to 0.2, modest reduction; 0.21 to 5.0, little change; 5.1 to 10.0, modest increase; 10.1 to 20.0, large increase; > 20.0, very large increase.

Evaluating Multiple Threshold Values for Best Sensitivity and Best Specificity

When trying to address the issue of whether a portable monitor can reduce the probability that a patient has sleep apnea, the focus is on sensitivity. A high sensitivity will result in a low number of false-negative results and a low LR. Conversely, when addressing the issue of whether a portable monitor can increase the probability of sleep apnea, the focus is on specificity. A high specificity will result in a high LR and a low number of false-positive results.

When the best sensitivity and the best specificity are at different RDI thresholds (*ie*, different points on the ROC curve), then some patients will meet one or the other criteria, but some will meet neither and therefore will have indeterminate results. For example, if authors reported their best sensitivity at an RDI cutoff of 5 and their best specificity at an RDI cutoff of 15, those patients whose RDI fell between these two thresholds would have a result that did not substantially reduce or increase the probability that they had sleep apnea. If the results are such that a majority of patients fall into this "gray zone," then the test may not be as useful as the best sensitivity and best specificity might suggest. In this regard, it is useful to examine the percentage of patients who meet the criteria for a negative result (*ie*, best sensitivity) and the percentage that meet the criteria for a positive result (*ie*, best specificity). Once it is clear what percentage of patients meets the criteria for a negative result or a positive result, the final important questions is what percentage of those patients meeting the criteria actually had a false result (*ie*, were misclassified by the diagnostic test). This result will be affected by the prevalence as well as the operating characteristics of the test (*ie*, sensitivity, specificity, and LRs). The best test obviously will be the one with the largest majority of patients who meet the criteria and have a low misclassification rate. The ideal test is one in which there is a single cutoff that has both a high sensitivity and high specificity so that patients have either a negative or positive result, and there is no gray zone.

Using different thresholds for positive and negative results generates combinations of sensitivity and specificity, and different LRs that can be confusing. The example below illustrates this.

Example

Consider 1,000 patients who are suspected of having sleep apnea, 40% of whom ultimately have a positive polysomnogram finding (*ie*, AHI ≥ 10). All patients have an RDI measured from a portable monitor, and the breakdown of the results is shown in Table 4. If the threshold for a positive result on the portable monitor is changed between the upper group (*ie*, AHI < 40 or ≥ 40) and the lowest group (*ie*, < 5 or ≥ 5), a series of 2 × 2 tables can be constructed to evaluate the effect on several different parameters (*eg*, sensitivity, specificity, predictive values, and percentage of patients with a positive or negative result). This is illustrated in Table 5.

When several thresholds are being evaluated for a diagnostic test, it is possible to calculate pairs of LRs for each level of the portable monitor result. However, it is more appropriate to calculate LRs for each intermediate threshold, as shown in Table 6, rather than calculating them for each pair (Table 5). Note

Table 6—*LRs for Each Level of RDI Result*

	AHI				
	≥ 10*		< 10†		
RDI	Patients, No.	Proportion	Patients, No.	Proportion	LR
≥ 40	178	178/400 = 0.445	8	8/600 = 0.013	0.445/0.013 = 34.2
20–39	129	129/400 = 0.323	31	31/600 = 0.052	0.323/0.052 = 6.21
10–19	53	53/400 = 0.133	33	33/600 = 0.055	0.133/0.055 = 2.42
5–9	27	27/400 = 0.068	189	189/600 = 0.315	0.068/0.315 = 0.21
< 5	13	13/400 = 0.033	339	339/600 = 0.565	0.033/0.565 = 0.06
Total	400		600		

*Positive for sleep apnea.
†Negative for sleep apnea.

that the only LRs that correspond between these two approaches are the LR for a positive result at the highest threshold and the LR for a negative result at the lowest threshold.

If separate thresholds are used to define a group of patients with sleep apnea and another group of patients without sleep apnea, some patients will fulfill neither criterion. The following tables give the approach used to calculate the number of patients without a positive or negative result. Table 7 combines best specificity (*ie*, RDI ≥ 40) and the best sensitivity (*ie*, RDI < 5) [see Table 5 for corresponding calculations]. The outcome is that 462 patients do not have a positive or negative result (*ie*, they have an RDI between 5 and 40), and among 538 patients with a positive or negative result there are 13 false-negative results and 8 false-positive results.

Table 8 demonstrates the effect of using a lower RDI threshold to define patients with sleep apnea (*ie*, RDI ≥ 20) while maintaining the threshold for defining patients without disease (*ie*, RDI < 5). The choice of which set of RDIs to use depends on many factors such as the repercussions to the patient of false-positive or false-negative test results. In Table 8, there are more patients who have a positive or negative result (69.8%). There are still 13 false-negative results, but now the number of false-positive results has risen to 39. The difference is that 159 more patients now have a result, but among these 159 patients are 31 more false-positive results.

Conclusions

Although there are several approaches to measuring agreement between two methods of measurement, such as portable monitoring and polysomnography, each one has limitations. The two recommended approaches are the following: (1) the Bland-Altman calculation of mean differences and limits of agreement; and (2) sensitivity, specificity, and LRs. The latter approach is in most common use. To correctly interpret the meaning of the combination of sensitivity and specificity or a LR, it is important to understand how test operating characteristics interact with pretest probability (*ie*, prevalence) to generate predictive values (posttest probabilities), which is the information that a clinician requires to make an informed choice about a diagnostic test. A test result above a threshold level that is associated with a high LR indicates an increased probability of disease. The higher the LR, the higher the probability. A test result below a threshold that is associated with a low LR indicates a decreased probability of disease. The lower the LR, the lower the probability. It is important to remember that when different thresholds are used to generate high and low LRs (or high specificities or sensitivities, respectively), a certain and possibly substantial proportion of patients may have a nondiagnostic result. This percentage should be considered when evaluating the utility of the test in a particular patient population.

Table 7—*Effect of Using as RDI Threshold of 40 to Define Patients With Sleep Apnea and an RDI Threshold of 5 to Define Patients Without Sleep Apnea*

	AHI		
RDI	≥ 10*	< 10†	Total
≥ 40	178	8	186
< 5	13	339	352
Total‡	191/400	347/600	538§/1,000

*Positive for sleep apnea.
†Negative for sleep apnea.
‡Values given as patients with a positive or negative result/total No. of patients.
§Of these, 462 had neither a positive nor a negative result.

Table 8—*Effect of Using an RDI Threshold of 20 to Define Patients With Sleep Apnea and an RDI Threshold of 5 to Define Patients Without Sleep Apnea*

	AHI		
RDI	≥ 10*	< 10†	Total
≥ 20 (+ve)	307	39	346
< 5 (+ve)	13	339	352
Total‡	320/400	378/600	698§/1,000

*Positive for sleep apnea.
†Negative for sleep apnea.
‡Values given as patients with a positive or negative result/ total No. of patients.
§Of these, 302 had neither a positive nor a negative result.

References

1 American Sleep Disorders Association. Practice parameters for the indications for polysomnography and related procedures: Polysomnography Task Force, American Sleep Disorders Association Standards of Practice Committee. Sleep 1997; 20:406–422
2 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307–310
3 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development. 2nd ed. New York, NY: Oxford University Press, 1995
4 Sackett DL, Strauss SE, Richardson WS, et al. Evidence-based medicine: how to practice and teach EBM. 2nd ed. Edinburgh, Scotland, UK: Churchill Livingstone, 2000
5 Fagan TJ. Nomogram for Bayes theorem [letter]. N Engl J Med 1975; 293:257