Reporting Standards for Diagnostic Testing:

Guidance for authors from editors of Respiratory, Sleep, and Critical Care Journals

Abstract:

Diagnostic testing is fundamental to medicine. However, studies of diagnostic testing in respiratory medicine vary significantly in terms of their methodology, definitions, and reporting of results. This has led to often conflicting or ambiguous results. To address this issue, a group of 20 respiratory journal editors worked to develop reporting standards for studies of diagnostic testing based on rigorous methodology to guide authors, peer reviewers, and researchers when conducting studies of diagnostic testing in respiratory medicine. Four key areas are covered, including defining the reference standard of truth, measures of dichotomous test performance when used for dichotomous outcomes, measures of multichotomous test performance for dichotomous outcomes, and what constitutes a useful definition of diagnostic yield. The importance of using contingency tables for reporting of results is addressed with examples from the literature. A practical checklist is provided as well for reporting studies of diagnostic testing.

The 21st century has brought advances in genomics, proteomics, imaging, robotics, and artificial intelligence that promise to be transformative to the practice of medicine. In parallel there has been a call for improvements in the methodologic rigor used to assess how well these technologies perform. Diagnostic testing is foundational to medicine. However, much of the guidance provided by journals relates to reporting of therapeutic trials and modelling, with the reporting of research assessing the accuracy or utility of diagnostic testing being given relatively minimal coverage.[1-5] This may contribute to conflicting results in the literature, the consequence of ambiguous definitions, opaque and incomplete reporting of results, and use of inadequate methodology. One example of this is field of diagnostic bronchoscopy. While the last three decades have seen tremendous technological progress in bronchoscopy, the results reported in the literature vary widely. Definitions are often not completely specified, the reference standard of 'truth' is sometimes not clearly defined, follow-up periods greatly vary and are often unclear, the results reported are often not sufficiently granular or complete, and the interpretation of the findings often exceeds the limitations of the methods used.[1-3]

Building on the experience of prior guidance provided by respiratory journal editors,[4,5] we convened an *ad hoc* group of 20 respiratory editors in the respiratory, sleep, and critical care domains in order to offer guidance to authors, peer reviewers, and researchers on the design and reporting of studies of diagnostic testing. We anticipate that best practices will continue to evolve, requiring this document to be updated periodically. We believe these changes will increase the rigor, validity, and value of the work published in our journals and will facilitate clinical research and innovations in the discipline. It will also facilitate collaborations between industry and academia by providing a clear framework for designing clinical research studies involving new diagnostic methods.

The goals of this document therefore are to 1) Clarify definitions and standardize certain terminology, 2) Apply best practice to facilitate accurate and transparent reporting as well as valid comparisons of similar technologies across studies, 3) Provide useful templates for data reporting, and 4) Provide a practical checklist to be used when designing studies and reporting results. Throughout the document, we will seek to highlight important concepts by providing illustrative examples. The examples relate primarily to bronchoscopy, because this has been a recent area of controversy, but the underlying concepts apply broadly to all studies of diagnostic testing in respiratory medicine. The document is divided into four parts, with each part focusing on a fundamental concept:

- Reference standards of truth
- Measures of dichotomous test performance when used for dichotomous outcomes
- Measures of multichotomous test performance for dichotomous outcomes
- Diagnostic yield

The four concepts are integrated in a final section that provides practical guidance for constructing appropriate contingency tables and a checklist for carrying out studies of diagnostic testing.

Reference Standards of Truth

Diagnostic accuracy can be defined as how well a test discriminates between two conditions. The two conditions can be health and disease, two different diseases, or even two stages of the same disease. A reference standard is used to identify which of the two conditions is truly present. While frequently the term “gold-standard” is used as a synonym for the reference standard, the term reference standard is preferable, because almost every reference standard can have errors. Ideally, the reference standard should be the best available method for establishing the presence or absence of the target condition.[6] While the reference standard may be a single test (e.g., surgical biopsy), it may

alternatively be composed of multiple other tests and/or clinical follow-up over time. When reporting study methods, it is critical that the reference standard of truth be clearly specified in the report with sufficient detail.

Although the reference standard may be composed of multiple other tests or long-term follow-up, ideally none of the elements that comprise the reference standard should be part of the diagnostic test being evaluated. Occasionally this can be problematic. For example, in studies of advanced diagnostic bronchoscopy for identifying lung cancer in peripheral nodules, the presence of malignant cells on biopsy is generally considered as definitive evidence of cancer (i.e., all positive biopsy results are considered "true positive" for malignancy). Ideally all positive bronchoscopy biopsy results would be confirmed by surgical biopsy, but this would require all patients to be surgical candidates. However, this is not always the case, and if surgical proof was required, then non-surgical patients would not be represented, and this in turn would limit generalizability. Conversely, if all bronchoscopy results showing cancer are considered as true-positive, then specificity will be 100%. Fortunately, data from older studies using surgical biopsy for confirmation suggest that specificity is indeed close to 100% for bronchoscopic biopsy.[7] So based on these older data, it is reasonable, depending on the study design, to consider all biopsies demonstrating malignancy as true-positives since the biopsy material is essentially the same; it is just the instrument that is obtaining the biopsy that has changed. This represents one of the rare exceptions to the rule that none of the elements that comprise the reference standard should be part of the test being evaluated.

An example of how including elements from the reference standard can lead to problems can be seen in studies of bronchoscopy that evaluate diagnostic yield. In prior studies, bronchoscopy *test results* were classified as "diagnostic" based on the finding of non-specific biopsy results (e.g. inflammation) and whether the nodule resolved by 12 months on follow-up chest computed tomography (CT). Note that this is the *test result* definition (as opposed to the reference standard

definition). If the nodule was stable or resolved at 12 months the test was classified as diagnostic, while if the nodule grew the test was considered non-diagnostic. Two patients with identical biopsy histology could have different *test results* based on their 12-month chest CT findings. This definition of "12-month diagnostic yield" has been used in a variety of studies,[3,34] but it is problematic and has led to widely varying and often conflicting results.[1] To know the "test result" you would have to wait 12 months from the bronchoscopy and repeat a chest CT. A test which has a 12-month processing time is not useful. The chest CT result is in the future and is different than the bronchoscopy test result – a test from the future cannot be used to determine test results in the present. However, a test from the future can be used as part of the reference standard for a test being done in the present. Indeed, the underlying reference standard of truth in those studies used stability on follow-up radiographic imaging as one of the criteria for determining if a lung nodule was malignant. Because both the test result and the reference standard shared this radiographic criterion in common, bronchoscopy performance was overestimated, because the test and the reference standard were essentially auto correlated.

For example, it would not be correct to supplement the biopsy results with information from the future (i.e. whether the lung nodule grew, stayed the same, or resolved by 12 months on chest CT) to define a bronchoscopy test result. The CT scan result is different than the bronchoscopy test result and it is in the future – it cannot be part of the test itself. This also violates the rule that test results be separate from the reference standard of truth. Consequently the definition of "12 month diagnostic yield" used previously,[3,34] should be avoided for bronchoscopy.[1]

A more complex problem may arise when a new test has higher sensitivity than the reference standard. For example, quantitative polymerase chain reaction (qPCR) assays along with blood culture often have superior sensitivity than blood culture when used to test for pneumonia in children. However false positive results may arise. Study design in such cases can be especially challenging but are beyond

the scope of this discussion. However, the concept of detailed contingency tables, as described below, can be applied effectively to such cases, provided they are sufficiently granular.

Measures of Dichotomous Test Performance for Dichotomous Outcomes

Provided the reference standard and all possible test results are clearly specified, we can then measure test performance. There are a variety of methods available to quantify test performance, some of which measure discriminatory power while others measure predictive ability. Frequently used measures of test performance include sensitivity and specificity, diagnostic accuracy, positive and negative predictive value, likelihood ratios (LRs), and area under the receiver operating characteristics curve (ROC).

Sensitivity and Specificity

When there are only two possible test results (e.g., positive or negative) for a dichotomous outcome, sensitivity and specificity can be used to measure the discriminatory power of a diagnostic test. Sensitivity is the proportion of patients who truly have the disease that test positive. This can be annotated in the following manner:

$$Sensitivity = p(T+ \mid D+) \quad (1)$$

This can be read as sensitivity is the conditional probability (p) that the test will be positive (T+), *given (|) that the disease is present (D+)*.

Using the same annotation, specificity is the proportion of patients who truly do not have the disease that test negative:

$$Specificity = p(T- \mid D-) \quad (2)$$

This can be read as specificity is the conditional probability (p) of having a negative test (T−), *given* (|) that the disease is absent (D−).

Calculation of sensitivity and specificity can be done using a contingency table, as shown in Table 1. Importantly, measures of diagnostic accuracy are not fixed indicators of a given test's performance in all situations, but rather are measures of test performance in a given context. There is increasing recognition that mix and severity of disease within a given study will impact estimates of a test accuracy.[8]

In the case of sensitivity, if the burden of disease is low in patients that have the disease, the sensitivity of a test will be lower than in patients with greater disease burden. If the selection criteria for a cohort selects preferentially for patients with lower disease burden, the reported sensitivity will be lower than in a different cohort where disease burden amongst patients with the disease is greater. The term disease burden is sometimes ambiguous. In this context disease burden means the amount of disease present in a patient with disease. This is not to be confused with the term disease burden *in a population*, which is synonymous with prevalence. For example, in patients that have positron emission tomography (PET) N0 disease (cN0), a meta-analysis of nine studies with 1,146 patients demonstrated endobronchial ultrasound guided transbronchial needle aspiration (EBUS-TBNA) had a sensitivity of 49% (95% confidence interval [CI] 41-57%).[9] In contrast, in cohorts that selected for patients that have large (> 1 cm), PET positive mediastinal nodes (cN2), EBUS-TBNA sensitivity is greater.[9-12] In eight studies of patients with PET N2 or N3 disease, EBUS-TBNA sensitivity ranged from 73% to 95%.[13-20] The difference in EBUS-TBNA sensitivity between cN0 and cN2/N3 studies reflects significantly lower disease burden among the cN0 patients that truly did have nodal disease (and hence were counted when calculating sensitivity) as compared to their cN2/N3 counterparts who also truly had nodal disease. Note that the lower sensitivity of EBUS-TBNA in the cN0 studies is not caused by the lower prevalence of disease in those studies (only patients with lung cancer in the lymph nodes were used to calculate sensitivity), but

rather is a function of lower disease burden (i.e., fewer malignant cells in the lymph nodes) among the patients that really did have cancer in their lymph nodes.

Specificity for a given test may also vary based on contextual factors. Specificity by definition only measures test performance amongst those patients ultimately proven not to have the disease of interest. Therefore, it is the mix of conditions *other than the target disease* of interest that may impact specificity. For example, PET specificity for malignancy will vary based on the prevalence of other conditions in the population that can cause a PET to be positive, such as histoplasmosis.[21,22]

Positive and Negative Predictive Value

The probability of disease given a positive or negative test result is of importance to clinicians, and this is reflected by the predictive value of the test. The positive predictive value (PPV) of a test is the conditional probability (p) of having the disease (D+), given (|) that the test is positive (T+), and can be written as p(D+|T+). Similarly, the negative predictive value (NPV) of a test is the conditional probability (p) of not having the disease (D−), given (|) that the test is negative (T−). Table 1 provides more detailed formulas for calculating PPV and NPV.

PPV and NPV are a function of the sensitivity and specificity of a test, as well as the prevalence of the condition, as dictated by Bayes' theorem (Figure 1). As disease prevalence increases, PPV increases and NPV declines for any given test. It is therefore important when reporting PPV and NPV results that disease prevalence be reported.

Diagnostic Accuracy

Diagnostic accuracy is often reported to provide a quantitative assessment of test performance. Diagnostic accuracy is the extent of agreement between the outcome of the new test and the reference standard.[23] In essence, diagnostic accuracy is the proportion of all test results, whether positive or

negative, that might be considered "correct".[24] For a dichotomous test for a dichotomous outcome, accuracy is defined as (TP + TN) / (TP + FN + TN + FP). Diagnostic accuracy, like sensitivity and specificity, is also not a fixed indicator, and varies between studies based on the same contextual factors noted for sensitivity and specificity above. In addition, diagnostic accuracy, unlike sensitivity, is impacted by disease prevalence in the population. While frequently reported, it has limited usefulness unless it is reported along with sensitivity, specificity, and prevalence, since it is a single aggregate measure that is influenced by multiple other factors.[8,24] A report of high diagnostic accuracy alone does not necessarily mean the test is useful. For example, if we do a study of bronchoscopy for all screen detected sub-centimeter pulmonary nodules, diagnostic accuracy will be more than 95%, because bronchoscopy has essentially 100% specificity, the prevalence of cancer in the population is very low, and hence there will be many true negatives resulting in high diagnostic accuracy even if sensitivity is zero. But the test is not useful and will not yield any new or valuable information for the patient.

Reporting Considerations for Sensitivity, Specificity, and Predictive Values

When reporting results of studies of diagnostic tests, it is important to clearly and completely specify the clinical context and selection criteria for the population so that disease burden can be estimated. This facilitates appropriate extrapolation of the results to other centers, provided the populations are similar. Relevant details of the study population should also be reported, since this can impact estimates of disease burden. Note that regarding specificity, other concurrent diseases that do not necessarily impact disease burden should also be reported (e.g., in a study of PET imaging of lung nodules, reporting the specific causes of the non-malignant nodules is important because it impacts specificity). This will facilitate proper cautious comparisons between studies, since sensitivity and specificity can vary between studies purely due to differences in the population of patients rather than the tests themselves. When reporting PPV and NPV, it is important to report disease prevalence, since

predictive probabilities are a function of both test characteristics (i.e., sensitivity and specificity) as well as disease prevalence (Figure 1). Appropriate 95% confidence intervals should be reported for each measure.

Measures of Test Performance for Multichotomous Tests for Dichotomous Outcomes

While sensitivity and specificity together are useful to describe the discriminatory ability of a test, they cannot be directly applied to tests with more than two possible test results. For purposes of this discussion, we will define tests that have three or more possible results as multichotomous tests. Multichotomous tests are common in medicine, and their interpretation can be challenging. Examples include ventilation-perfusion scans, EBUS cytology results, and certain molecular assays.

One solution for multichotomous tests is to consolidate groups together. If a multichotomous test has three possible results: low, intermediate, and high, then low and intermediate could be consolidated into a negative group while high is considered positive. However, this results in loss of information and the discriminatory power of the test will decrease. It is also not necessarily true that intermediate should be lumped in with low rather than being included with high. This is even more problematic if test results are not ordinal, which is frequently the case with biopsy results.

An alternative approach to the problem of multichotomous tests is to use LRs.[25] LRs are well described but have not been used as widely as sensitivity and specificity. However, they are well suited to the problem of multichotomous tests and can provide clinicians and researchers with a more nuanced understanding of diagnostic testing. The LR of a given test result (T) is the probability of seeing that test result in patients with the disease divided by the probability of seeing that same test result in patients without the disease:

$$Likelihood\ ratio\ for\ a\ test\ result\ (T) = LR(T) = \frac{p(T|D+)}{p(T|D-)} \quad (3)$$

Knowing the pretest probability of disease and the LR for a test, it is possible using Bayes' theorem to calculate the posttest probability of disease (Figure 2).

A test with a high LR increases the odds of disease, a low LR decreases the odds of disease, while a LR of one indicates that a test result is non-informative – meaning the posttest odds are unchanged from the pretest odds of disease. Note that each specific test result has its own LR, so if a test has only two possible results, positive (+) and negative (−), then there will be two LRs, one for each result (LR+ and LR−). Unlike sensitivity and specificity, LRs can be used for multichotomous tests as well. So, if there are three possible test results (R1, R2, R3), then there will be three corresponding LRs: LR(R1), LR(R2), and LR(R3) respectively. A qualitative way to interpret LR is shown in Table 2.[26,27]

As an example, when doing EBUS-TBNA for isolated mediastinal lymphadenopathy in a patient with suspected lymphoma, the biopsy could show lymphoma, granulomas, another specific disease other than lymphoma, or be non-diagnostic with either adequate or inadequate lymphocytes in the specimen.[28] With four possible test results, one approach is to consider a finding of lymphoma as positive, and lump all other test results together as "negative". But in this case not all "negative" results are truly equal. The probability of having lymphoma is different, depending on which negative result was obtained. Using LR highlights this and captures the differences between groups (Table 3). In patients with suspected lymphoma, the probability of having lymphoma decreases to a much greater degree if granulomas or a specific alternative diagnosis is found on EBUS than if merely adequate lymphocytes are found. Using LRs for reporting bronchoscopy results is useful because it highlights how different "negative" findings have different implications. These subtle but important differences are lost when we lump all findings into either positive or negative bins.[28] Notice that in table 3 a biopsy finding of adequate lymphocytes or inadequate lymphocytes are lumped into a single category. Based on clinical grounds, the finding of inadequate lymphocytes was judged to be clearly very different from the finding

of granulomas or lymphoma and it clearly does not constitute a specific diagnosis other than lymphoma. But are inadequate lymphocytes really similar to a finding of adequate lymphocyte, such that they should be aggregated into a single category? Perhaps, but possibly not. In this study, the number of patients with inadequate lymphocytes was very small, so aggregation with adequate lymphocytes was chosen when reporting the primary result. However, in addition to reporting the LR for the combined category (adequate or inadequate lymphocytes) which was 0.31 (95% CI 0.018-0.55), the investigators also reported the LR for adequate lymphocytes (LR 0.25, 95% CI 0.14-0.49) and the LR for inadequate lymphocytes (LR 1.06, 95% CI 0.24-4.59).

LRs are also useful when tests results are indeterminate. Even a dichotomous test may fail if the sample is insufficient or for technical processing reasons. Ignoring indeterminate test results (i.e., excluding them) can produce biased estimates of accuracy if the indeterminate results do not occur at random. The STARD guidelines[6] recommend reporting how indeterminate tests results are handled, and one option is to capture these indeterminate test results as their own test category. Using LR facilitates this, since the indeterminate, positive, and negative test results are each assigned their own LR.

Finally, LRs are also useful for diagnostic tests that provide a continuous test result (e.g., size of a lymph node on CT). For tests with a continuous result, the area under the receiver operating characteristics curve (ROC) provides a useful overall measure of the discriminatory function of a diagnostic test. ROC curve methodology is well established and described.[29] The ROC curve plots sensitivity on the y-axis and the false positive rate (i.e., 1 – specificity) on the x-axis. Any continuous test result variable can be partitioned into two categories (negative and positive) by using a single cutoff (Figure 3A). The points on the ROC curve merely represent the sensitivity and the corresponding false positive rate that are obtained for all possible cutoff values of the diagnostic test. The shape of the curve represents the trade-off occurring between sensitivity and the false positive rate as the cut-point is varied. The greater the area under the ROC curve (AUC), the better the test. Conceptually, the AUC is

equal to the probability that the test will correctly rank a randomly chosen patient with disease higher than a randomly chosen patient without disease. A completely non-informative test would still be correct half the time. Therefore, the AUC varies from a minimum of 0.5 (completely non-informative test without discriminatory power) up to a maximum of 1.0, which represents a test with perfect discriminatory power.

However, physicians may find it hard to apply a test with an AUC of 0.8 at the bedside. So, it is often useful for physicians to partition a continuous variable into two or more intervals. If partitioning results in only two intervals (e.g., negative, and positive), then sensitivity and specificity can be used to describe and quantify test performance. LRs can be used when there are only two categories, and in such instances the LR+ and the LR− can be expressed as a function of sensitivity and specificity as follows:

$$LR+= \frac{p(T+|D+)}{p(T+|D-)} = \frac{sensitivity}{1-specificity} \quad (4)$$

$$LR- = \frac{p(T-|D+)}{p(T-|D-)} = \frac{1-sensitivity}{specificity} \quad (5)$$

However, if partitioning results in three or more intervals (e.g., low, medium, high), sensitivity and specificity are problematic, since sensitivity and specificity are predicated on having only two possible test results. In such instances, calculating sensitivity and specificity for the three or more intervals is meaningless and instead LRs should be used. Note that when there are three or more intervals, equations (2) and (3) do not apply to those intervals directly.[25]

Calculation of LRs for intervals for a continuous variable is more complex but can be linked directly to ROC curves. Figure 3 illustrates the key concepts and the relationship between LRs and ROC curves for continuous tests. For any point on the ROC curve, X, if that point is used as the criterion to partition the continuous test into positive and negative results, then the slope between the origin and X is equal to the LR+ (Figure 3A).[29] This is readily apparent when we calculate the slope and then compare

that to equation (2), which is identical. Similarly, the slope from X to the coordinates (1,1) represents the LR−.[25]

If n cut-points are used to partition a continuous test, this results in (n+1) intervals. The slope of the line connecting two consecutive cut-points on the curve corresponds to the LR for a test result falling in the interval bounded by those two points. If n=2, there are two cut-points (X and Y), and there must be 3 possible categories of results corresponding to the three intervals created (Figure 3B). For test results that fall in the interval bounded by X and Y, the slope of the line connecting X to Y is equal to the LR for this category. Equations (2) and (3) do not apply directly to the interval (X, Y) or the n+1 intervals when n is 2 or higher. However, the LR for a given interval (X, Y) can be calculated if we know the sensitivity and specificity for both X and Y when each is used as a binary cut-point:

$$Likelihood\ ratio\ for\ test\ result\ in\ inerval\ (X,Y) = LR(X,Y) = \frac{Sensitivity\ (X) - Sensitivity\ (Y)}{Specificity\ (Y) - Specificity\ (X)}$$

We can imagine an infinite number of cut-points, such that the intervals between consecutive cut-points becomes infinitely small. Hence, the tangent to a point on the ROC curve corresponds to the LR for a single test value represented by that point (Figure 3C).[29] For mathematical proof and more in depth analysis, we direct readers to methodological papers that more fully describe the relationship between the ROC curve and LRs.[29-31]

Reporting Considerations for Multichotomous Tests

When there are three or more possible categorical test results with substantively different clinical interpretations, such as is often the case with bronchoscopy or studies using biopsy results (e.g., pleural biopsy could show tuberculosis, malignancy, non-specific pleuritis, or normal pleura), consider using LRs. For multichotomous test with categorical results, such as biopsies, some aggregation of rare categories will be required. The rationale for aggregating categories should be clearly articulated and based on clinical judgment. In addition, consider using data supplements to report data in a sufficiently

granular form such that LRs for smaller subcategories can be determined. This allows editors, reviewers, and readers to see the individual elements that were subsequently aggregated for creation of the LRs and judge for themselves whether this was warranted.

Multichotomous tests should also be considered when there are a significant number of indeterminate results, since excluding indeterminate tests can result in bias.[6] For tests with continuous readouts that will be converted into multichotomous tests (e.g., low, middle, high), LRs when calculated correctly can be valuable. However, care must be taken to use the correct equations, since misapplication of the wrong equation can lead to misleading results and inaccurate estimates of disease probablity.[25,29,32]

Diagnostic Yield

Diagnostic yield is not used frequently in epidemiology texts or literature, but it has been used frequently in the bronchoscopy literature. While some authors treat it as a synonym for diagnostic accuracy, this is not the case. As described above, diagnostic accuracy measures the concordance rate of the test with the reference standard of truth for a specific disease. Diagnostic yield is the likelihood that a test or procedure will provide the information needed to establish a diagnosis of *any disease that is present*. Diagnostic yield is thus not disease specific.

To illustrate this, it is first useful to consider what is in the numerator and denominator of diagnostic yield for a dichotomous test and then to consider the case of a multichotomous test that can make many different diagnoses. For a dichotomous test for a dichotomous outcome, diagnostic yield is similar to diagnostic accuracy in that the denominator for diagnostic yield is all patients undergoing the test. However, the numerator for diagnostic yield consists of all cases in which the test was sufficient to "establish a diagnosis". If we consider a contingency matrix (table 1), TP results seem like they would be

in the numerator. Would false positives be included as well? The answer is yes, since there is no way for the test user to distinguish between TP and FP. The test user cannot know the underlying truth at the time of testing and interprets both TP and FP as establishing a diagnosis. A negative test, even if it is a TN, does not establish a diagnosis, so in most contexts TN and FN would not be in the numerator. Hence diagnostic yield can be written as:

$$Diagnostic\ yield = \frac{TP + FP}{(\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP})} = p(T+)$$

Therefore, high diagnostic yield is not necessarily a desirable attribute, unless the number of FP results is zero or extremely low and of little consequence. However, in the special case when a test has 100% specificity, such that all positives are TP, then diagnostic yield is meaningful. In this special case, diagnostic yield is equal to the prevalence of disease × sensitivity. But in many other circumstances, diagnostic yield is not that informative for dichotomous tests for dichotomous outcomes.

In other contexts, quantifying diagnostic yield may be useful if the test is a multichotomous test that has the potential to establish many different diagnoses. Measures like sensitivity, specificity, and diagnostic accuracy are disease specific. These measures of test performance work well if we are evaluating a test that can diagnose only one disease (e.g., HIV blood test). However, bronchoscopy can establish many different diagnoses. In the situation where a single test can diagnose multiple different diseases, the interpretation of disease specific diagnostic accuracy becomes subtle and not always intuitive.

To illustrate this, we use data from the AQuIRE registry, which is shown in table 4.[33] In a subset of the registry, four centers collected data on 334 subjects undergoing peripheral bronchoscopy. Peripheral sampling by bronchoscopy established a lung cancer diagnosis in 144 patients, which because of the high specificity of bronchoscopy were considered as TP. There were 51 FN results in which peripheral sampling was negative, but lung cancer was eventually proven by other means. Note there were 44 subjects lost to follow-up. In this example we assume all subjects lost to follow-up truly did not

have lung cancer and calculate the maximum sensitivity of bronchoscopy *for lung cancer* in the periphery as 144/195 = 74%, specificity being 100% (Table 5). Maximum diagnostic accuracy *for lung cancer* in this case is (144 + 139)/334 = 85%. But bronchoscopy did not establish a diagnosis 85% of the time. The diagnostic yield *for lung cancer* would be 144/334 = 43%. We may be interested in asking how often bronchoscopy was able to establish *any* diagnosis. We see from table 4 that bronchoscopy was potentially informative in another 41 patients. Some of these 41 patients had specific diagnoses made (e.g., metastatic solid tumor to the lung or tuberculosis) by bronchoscopy that were clearly sufficient to establish a diagnosis, while others might require adjudication (e.g., granulomatous inflammation). If we count all 41 patients as diagnostic, then they would go in the numerator when calculating total aggregate diagnostic yield, so (144 + 41) / 334 = 55% of the time bronchoscopy established a diagnosis.

A key distinction is that diagnostic accuracy, like sensitivity and specificity, is disease specific while diagnostic yield is not necessarily limited to just one disease. Aggregate diagnostic yield for bronchoscopy can be informative, but only when certain assumptions are met. For diagnostic yield to be meaningful, it must have a direction (i.e., higher is better). But if there is the potential for FP results in the numerator, then diagnostic yield is less meaningful since a higher diagnostic yield might just be due to more FP results. However, if the criteria for "diagnostic" is narrowly defined, such that each possible result is nearly 100% specific for a given disease (e.g., bronchoscopy showing tuberculosis), then diagnostic yield is more interpretable. In this special case, it reflects how often bronchoscopy was informative for a range of diseases in that clinical context. Diagnostic accuracy, because it is disease specific and includes TN in the numerator, does not reflect the range of other diseases that bronchoscopy is useful for and does not necessarily reflect how often bronchoscopy was able to establish a diagnosis.

However, aggregate diagnostic yield does not necessarily inform us as to how well bronchoscopy performed for a particular disease - it is an aggregate measure which in the special case of

100% specificity is really a function of multiple other factors. The first is the prevalence of each of the separate diseases in the cohort that bronchoscopy could potentially diagnose. The second is the sensitivity of bronchoscopy for each of those diseases. Given the baseline assumption that we are restricted to only diseases for which bronchoscopy is 100% specific, then aggregate diagnostic yield for a population that has *n* different diseases potentially diagnosable by bronchoscopy is:

$$Aggregate\ Diagnostic\ Yield = \sum_{i=1}^{n} prevalence_i \times sensitivity_i$$

Where *i* is the index for a particular disease that is present in the population at a given *prevalence,* for which bronchoscopy has a given sensitivity$_i$ and with a specificity of 100%. Therefore, it is not necessarily informative to compare alternative types of tests for the same disease based on their diagnostic yield in different studies to see which test is "better". A higher diagnostic yield may just be due to more FP results or to differences in the prevalence of disease(s) in the populations being studied.

Reporting Considerations for Diagnostic Yield

When reporting diagnostic yield, it is important to completely specify in the methods section which test results are included in the numerator and the rationale for this classification. For tests such as bronchoscopy that can establish more than one diagnosis, this is particularly important since "positive" results that are not 100% specific may introduce a significant number of FP results into the numerator depending on the definition used. Precision in language is important, so that diagnostic accuracy and diagnostic yield are not confused.

Only categories of test results should be in the numerator of diagnostic yield, without any information from the underlying reference standard of truth or any additional information from the future that would not be part of the test itself. Data reporting should also be sufficiently granular so that it is clear what underlying data were used to calculate sensitivity, specificity, and LRs. This makes

interpretation of diagnostic yield more meaningful because many possible diseases may be contributing to aggregate diagnostic yield. Each different test result that was counted as a diagnosis established by the test should be clearly enumerated and the frequency of that disease in the cohort should also be reported. Otherwise, two studies could report the same diagnostic yield of peripheral bronchoscopy as 55%, but in one cohort this might be due to 55% of the cases being diagnosed as lung cancer and in the other cohort it could be 30% lung cancer and 25% tuberculosis. While the aggregate diagnostic yield is the same the two are not equivalent in terms of clinical interpretation. A further benefit of this reporting approach is that it allows calculation of sensitivity of the test for each disease present. For example, in table 4 each specific non-cancer diagnosis is listed as well, and we can determine the sensitivity of peripheral bronchoscopy for each disease in this cohort. Thus, the need for granularity in reporting applies to both the test results as well as the underlying reference standard of truth.

Practical Guidance: Contingency Tables and Checklists

We suggest authors review other applicable guidelines on reporting of diagnostic testing, such as the STARD guidelines, and use them as appropriate based on the strength of the methodology described and their potential applicability to the question being studied.[6] However, existing guidelines do not provide sufficient discipline specific details or rationale to always be useful for physicians. Understanding the reasons behind the recommendations is essential since guidelines cannot cover all possible situations. In addition, deeper understating improves the data abstraction process by illustrating the importance of capturing and reporting data in a highly granular and organized fashion. Whether the test is dichotomous or multichotomous, careful, complete, and granular reporting of selection criteria, reference standards of truth, test results, and other diseases present in the population are important. However, it can be difficult to do this in a sufficiently granular and transparent manner

while keeping the data manageable. When a test can establish more than one diagnosis, this makes matters even more complex.

One potential approach is to construct a more detailed contingency table to serve as a foundation for analysis. To illustrate how to construct a more detailed contingency table, we again use the AQuIRE data (Table 6).[33] Unlike the classic 2 x 2 contingency matrix (table 5) that simplifies tests and results into positive and negative, table 6 is more detailed with one row for each possible test result and one column for each of the multiple diseases that were present in the cohort. Using table 6, it is easy to construct a 2 x 2 contingency table for any disease by specifying which columns are aggregated together to constitute the disease of interest and which rows will be aggregated together to constitute negative and positive results. For example, in table 6, if primary lung cancer of any type is the disease of interest, then the first three columns will constitute the diseased group (n=179 + 8 + 8 = 195); rows can similarly be aggregated, with the result being a 2 x 2 contingency matrix.

Providing this table in a data supplement provides the necessary granularity in terms of the other diseases present that may impact specificity as noted above. In addition, if instead of viewing bronchoscopy as a dichotomous test for lung cancer, we wish to view it as a multichotomous test for lung cancer, this can be done by specifying which rows are aggregated together for each test result in a multichotomous test. This facilitates transparent calculation of LRs and makes explicit the underlying classification scheme used by the investigators. Readers can judge for themselves the validity of the underlying classification scheme, and recalculate LRs if necessary, based on their own classification.

Finally, this more detailed contingency table allows transparent reporting of diagnostic yield. All that is required is to specify which rows are being aggregated together and considered as diagnostic, and which rows are considered as non-diagnostic. The key rule for diagnostic yield is that a given row is indivisible. The entire row is considered as either establishing a diagnosis or not. This is different than accuracy, where some cells in a row would be in the numerator (TP or TN) and other cells in the same

row might not be in the numerator (FN or FP). This facilitates transparent reporting of diagnostic yield, and if reviewers or readers disagree with the definition of what constitutes diagnostic yield, they can recalculate for themselves, so long as each row is indivisible. Applying this to the "12-month diagnostic yield" approach, which uses CT results from the future to categorize test results as diagnostic, makes the problem clear. Not only does such an approach use information from a future CT scan, but it also allows patients with the same biopsy result to be classified into different test categories (i.e., it requires splitting a row), which is invalid.

Table 7 provides a checklist of concepts covered in this document and can serve as a supplement to the STARD checklist that we recommend.[6]

Final Comments

The above approach and this document in general are intended to provide thoughtful guidance for studies of diagnostic accuracy as well their rationale, rather than absolute rules. A summary checklist of considerations is provided in Table 7. The guidance provided is not authority-based and should always be open to questioning and academic discourse. Their applicability or lack thereof comes from the quality of the reasoning and the methods behind them. As such the considerations listed are presented as a framework to facilitate efficient discussion between researchers, reviewers, and editors. Researchers should be free to modify the approach described so long as they provide additional details in their methods explaining the rationale for the choices made, which may differ from published guideline recommendations. In this way their methods and results will be presented in a sufficiently granular and transparent fashion that the community of scientists can more efficiently evaluate the evidence and arguments being made. The goal is to raise the rigor and quality of academic discourse in

our journals and to facilitate transparent communication of research findings. This, in turn, will enhance the validity of our science and create value for patients and providers alike.

Table 1. Contingency matrix for calculating sensitivity and specificity

		Disease status based on reference standard		Total	Predictive Value
		Positive	Negative		
Test result	Positive	**A** True positive (TP)	**B** False positive (FP)	Total test positive results (A + B)	Positive Predictive Value (PPV) $\frac{True\ positives}{All\ positive\ results} = \frac{TP}{TP+FP} = \frac{A}{A+B}$
	Negative	**C** False Negative (FN)	**D** True Negative (TN)	Total test negative results (C + D)	Negative Predictive Value (NPV) $\frac{True\ negatives}{All\ negative\ results} = \frac{TN}{TN+FN} = \frac{D}{C+D}$
Total		All disease patients (A + C)	All non-diseased patients (B + D)		
		Sensitivity $\frac{TP}{TP+FN} = \frac{A}{A+C}$	Specificity $\frac{TN}{TN+FP} = \frac{D}{B+D}$		

TP: true positive; FP: false positive; TN: true negative; FN: false negative

Table 2. Likelihood ratio interpretation*

Likelihood ratio	Qualitative interpretation
0.1	Large and often conclusive decrease in the likelihood of disease
0.2	Moderated decrease in the likelihood of disease
0.5	Small decrease in the likelihood of disease
1	No change in the likelihood of disease
2	Small increase in the likelihood of disease
5	Moderate increase in the likelihood of disease
10	Large and often conclusive increase in likelihood of disease

* These are general guidelines that must be correlated with the clinical scenario.

Table 3. Likelihood ratios (LR) for EBUS results in patients with suspected lymphoma[28]

Test result	Lymphoma absent	Lymphoma present	LR for test result
Lymphoma	0	63	∞
Specific non-lymphoma diagnosis	12	0	0
Granulomatous inflammation	41	0	0
Adequate or inadequate lymphocytes	53	12	0.32*
Total	106	75	

***** The value of 0.31 reported in the original text uses a correction factor to allow for calculation of confidence intervals, since there are cells with zero counts. As a result, the LR reported is 0.31 in the original manuscript. For didactic purposes, we calculate this manually which comes out to 0.32. LR: Likelihood ratio; ∞: infinity

Table 4. Peripheral bronchoscopy results from the AQuIRE registry[33]

Final Diagnosis Used as a Reference Standard	**Diagnosed by bronchoscopy of peripheral lesion**	**Diagnosed by EBUS-TBNA of mediastinal nodes but peripheral lesion was non-diagnostic**	**Diagnosed by subsequent procedure or serial imaging and clinical follow-up**
Non-small cell lung cancer - adenocarcinoma	74	3	19
Non-small cell lung cancer - squamous	50	4	17
Non-small cell lung cancer - non specified (undifferentiated)	8	1	3
Small cell lung cancer	8	0	0
Other primary lung cancer	0	0	3
Carcinoid tumor of the lung	3	0	1
Non-small cell lung cancer-large cell	1	0	0
Metastatic to the lung from a hematological origin*	7	0	3
Metastatic to the lung originating from a solid tumor	6	0	8
Hamartoma	1	0	1
Infection- bacterial	8	1	1
Prototheca wickerhamii	1	0	0
Infection- fungal, aspergillus	2	1	1
Infection- fungal, histoplasmosis	1	0	3
Infection- fungal, other	4	1	1
Infection- tuberculosis (TB)	1	0	2
Infection- *Mycobacterium avium-intracellulare* (MAI)	0	0	1
Viral infection- other	1	0	0
Infection – pathogen not identified	0	0	1
Sarcoidosis†	1	2	2
Granulomatous inflammation†	3	0	0
Bronchiolitis Obliterans Organizing Pneumonia (BOOP)	2	0	1
Usual interstitial pneumonia	0	0	1
ILD - Other	1	0	0
Diffuse alveolar hemorrhage	0	0	1

Atypical myxoid spindle cell tumor	1	0	0
Pseudotumor	0	0	1
Seroma	0	0	1
Benign hyalinized scar	0	0	1
Lipoid pneumonia	1	0	0
Pulmonary nodule or mass resolved by 1 year under surveillance, presumed benign	0	0	16
Stable pulmonary nodule or mass for at least 1 year or greater, under surveillance, no growth to date	0	0	5
Lost to follow up without a final diagnosis made	0	0	44

***** Hematologic origin refers to any lymphomas, leukemia, or myeloma. † Sarcoidosis was a combined clinical-pathologic diagnosis. If the physician found granulomatous inflammation without evidence of infection but was not certain this was sarcoidosis this was termed granulomatous inflammation.

Table 5. Contingency matrix for diagnosing lung cancer in the AQuIRE peripheral bronchoscopy registry

		Disease status based on reference standard		Total	Predictive Value
		Positive	Negative		
Bronchoscopy result	Positive	**144** True positive (TP)	**0** False positive (FP)	Total test positive results 144	Positive Predictive Value (PPV) $\frac{True\ positives}{All\ positive\ results} = \frac{TP}{TP+FP} = \frac{144}{144}$
	Negative	**51** False Negative (FN)	**139** True Negative (TN)	Total test negative results 190	Negative Predictive Value (NPV) $\frac{True\ negatives}{All\ negative\ results} = \frac{TN}{TN+FN} = \frac{139}{51+139}$
Total		All disease patients 195	All non-diseased patients 139		
		Sensitivity $\frac{TP}{TP+FN}$ $= \frac{144}{144+51}$ $= 0.74$	Specificity $\frac{TN}{TN+FP}$ $= \frac{139}{0+139}$ $= 1.0$		

Table 6. Hypothetical higher dimensional contingency table for a peripheral diagnostic bronchoscopy study*

	Reference Standard of Truth Result														
Bronchoscopy result	Non-small cell lung cancer	Small cell lung cancer	Other primary lung cancer	Metastatic cancer to the lung	Aspergillus	Other fungal infection	TB	Other mycobacteria	Bacterial, viral or other infectious pathogen	Sarcoid	ILD or BOOP	Other†	No change over 2 years	Disappeared during observation	Total
Non-small cell lung cancer	132	0	0	0	0	0	0	0	0	0	0	0	0	0	132
Small cell lung cancer	0	8	0	0	0	0	0	0	0	0	0	0	0	0	8
Other primary lung cancer	0	0	4	0	0	0	0	0	0	0	0	0	0	0	4
Metastatic cancer to the lung	0	0	0	13	0	0	0	0	0	0	0	0	0	0	13
Aspergillus	0	0	0	0	2	0	0	0	0	0	0	0	0	0	2
Other fungal infection	0	0	0	0	0	5	0	0	0	0	0	0	0	0	5
Tuberculosis	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Other mycobacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bacterial or viral or other identified infection	0	0	0	0	0	0	0	0	10	0	0	0	0	0	10
Granulomas without growth of organism	0	0	0	0	0	0	0	0	0	1	0	0	3	0	4

ILD or BOOP	0	0	0	0	0	0	0	0	0	0	3	0	0	0	3
Other†	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2
Acute inflammation without granulomas, no infectious agent grown	3	0	1	1	0	1	1	0	1	1	1	0	5	2	17
Non-specific normal lung tissue	44	0	3	10	2	4	1	1	2	3	1	4	44	14	133
Total	**179**	**8**	**8**	**24**	**4**	**10**	**3**	**1**	**13**	**5**	**5**	**6**	**52**	**16**	334

TB: Tuberculosis; ILD: Interstitial lung disease; BOOP: Bronchiolitis obliterans organizing pneumonia. * The data is hypothetical but matches the AQuIRE data shown in table 4. The hypothetical aspect is that it presumes complete 2-year follow-up for all patients and that all patients lost-to-follow up also had no growth. †Aggregation of the underlying disease (reference standard) lumps atypical myxoid spindle cell tumor, pseudotumor, seroma, benign hyalinized scar, and lipoid pneumonia together as one column labelled other. The row listed as other, lumps atypical myxoid spindle cell tumor and lipoid pneumonia together. The distinction between acute inflammation without granuloma, no infectious agent grown, and non-specific normal lung tissue is hypothetical. It is purely didactic and used only to illustrate the point that different histology findings (inflammation vs. normal), while not definitive, might have different likelihood ratios (LR). For the diagnosis of any lung cancer (first three columns), the LR of acute inflammation in this hypothetical example is 0.22, while the LR of non-specific normal lung tissue is 0.39. In the actual AQuIRE data there are sub-groups – the acute inflammation without granuloma and the non-specific normal tissue are just one group; with a LR of 0.26 assuming all lost-to-follow-up was truly negative.

Table 7. Checklist of Considerations for Diagnostic Studies

Concept	Comments
Study Design	
Consider whether a test result is truly dichotomous or multichotomous	Multichotomous is often the better choice for biopsy results, since not all "negative" results may be equal in terms of clinical implications.
Cohort selection	
Detailed eligibility criteria?	This needs to be sufficiently granular to replicate.
How were potential patients identified?	Consecutive, random, convenience samples.
Detailed description of the setting / clinical context.	This impacts disease burden, which influences sensitivity. Concurrent conditions may influence specificity.
Reference Standard	
Is the reference standard clearly specified?	This should be provided with sufficient detail for replication, and the classification scheme should be mutually exclusive and completely exhaustive. Each patient must map to one and only one category.
Is the reference standard separate from the test being evaluated?	Ideally the reference standard should not include any part of the test being evaluated, with rare exceptions. If there is overlap, provide a rationale.
Does the test result only use information from the test, and not information from the future?	Using information from other tests and from tests in the future (e.g., outcome after 12 months of clinical follow-up) is suitable when defining the reference standard, but information from the future cannot be used as part of a test-result.

Is the reference standard a suitable standard?	The rationale for choosing this as the reference standard should be provided if it is not standard. This is particularly true if alternative options exist. The best available method for establishing the truth will vary based on clinical context, so providing the context may be critical.
Were test result assessors blinded to the reference standard and vice versa?	This is particularly relevant for retrospective studies with tests that require judgment and clinical interpretation (e.g., diagnostic yield and use of a multidisciplinary team conference that knows the results of the biopsy, so the test influences the reference standard of truth). This also impacts interpretation of diagnostic yield.
Test classification scheme	
Specify if the test is dichotomous, multichotomous, or continuous	Provide sufficient granularity in the contingency table so that there is no ambiguity of classification.
When using diagnostic yield as a metric, clearly specify and provide a rationale as to which results are considered “diagnostic”. Provide details as to whether the classification scheme was pre-specified or whether it was established after data review	Diagnostic results should be close to 100% specific for a disease. Diagnostic results should be linked to specific rows in the contingency matrix, should be based solely on the test and should not use information from the future (e.g., results of 12-month follow up are not a *test result* although they can be used as part of the reference standard of truth).
Results	
Detailed description of the cohort	The cohort’s characteristics, which are influenced by the selection criteria, will influence disease burden. Descriptive measures should be sufficiently granular to capture this.

Sufficiently detailed contingency matrix of the test results and the reference standard so there is no ambiguity	The contingency matrix should be sufficiently granular, with each possible test result being clearly specified as well as each reference standard of truth. Test results should be mutually exclusive and completely exhaustive. The same applies to reference standards. Footnotes can be used to detail which rows and columns were subsequently aggregated.
Report estimates of precision of the measures (e.g., 95% confidence intervals)	

Figure 1. Prevalence, Sensitivity, Specificity, and Predictive value

$$Positive\ predictive\ value\ (PPV) = \frac{TP}{TP + FP}$$

$$= \frac{(Prevalence\ \times Sensitivity)}{(Prevalence\ \times Sensitivity) + (1 - Prevalence)(1 - specificity)}$$

$$Negative\ predictive\ value = \frac{TN}{TN + FN}$$

$$= \frac{(1 - Prevalence)(Specificity)}{(1 - Prevalence)(Specificity) + (Prevalence)(1 - Sensitivity)}$$

Figure 2. Bayes' theorem to calculate post-test probability of disease

Convert the pretest probability of disease to pretest odds of disease, using the following the equation:

$$Pretest\ Disease\ Odds = \frac{pretest\ probability\ of\ disease}{1 - pretest\ probability\ of\ disease} = \frac{p(D+)}{1-p(D+)}$$

Use Bayes' theorem to calculate the posttest disease odds as follows:

$$Posttest\ odds = pretest\ disease\ odds \times LR(Test\ result)$$

Then convert posttest odds back to posttest probability:

$$Posttest\ Probability\ of\ Diseaes = \frac{Posttest\ Odds}{Posttest\ Odds+1}$$

Figure 3. Relationship between ROC curve and likelihood ratios for a continuous test value

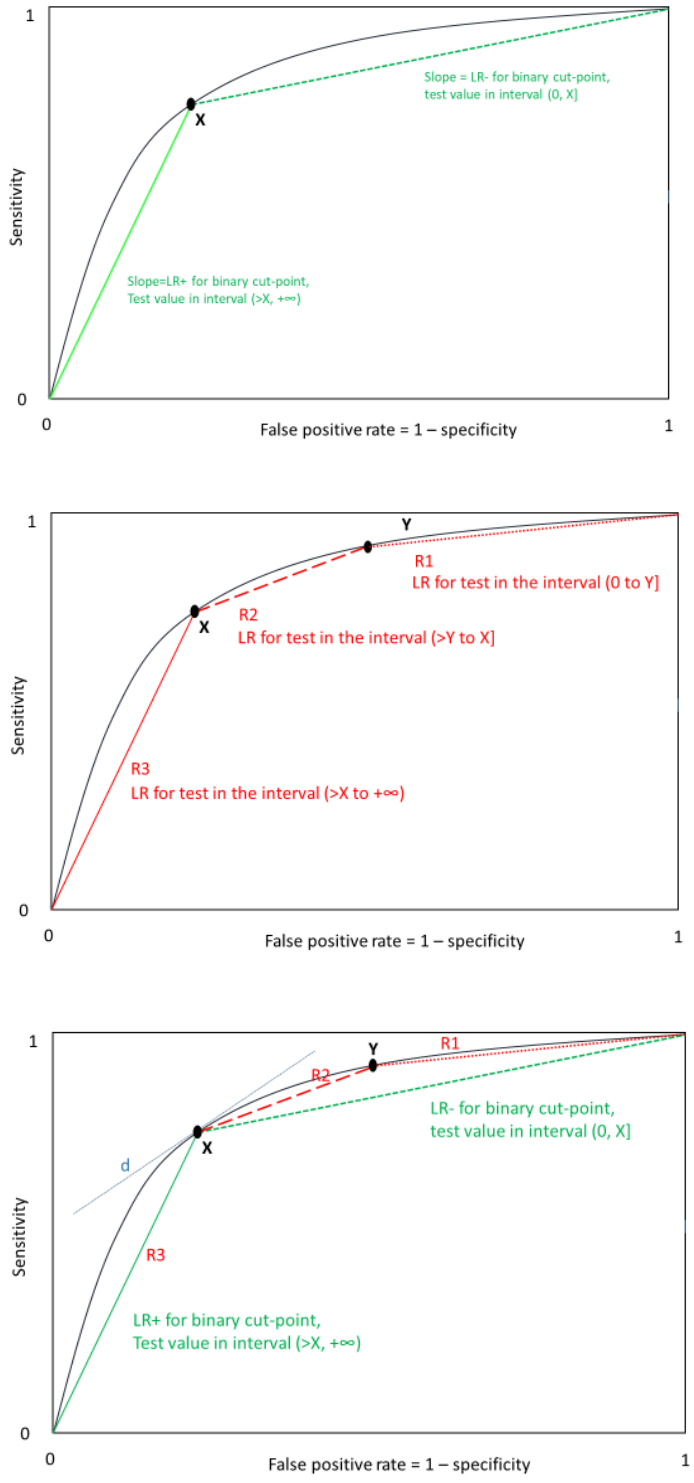

Figure 3. Receiver operating characteristic (ROC) curve relationship to likelihood ratios (LR) for a hypothetical test. The test has a continuous read-out, which can range from 0 to infinity, and disease is associated with higher test values. (A) Dichotomous classifier: A binary cut-point X is chosen, so tests with values ≤ X are considered negative, while tests with a value > X are positive. For any value of X, the slope of the solid green line to X is equal to the LR+; the slope of the dashed green line from X to (1, 1) is equal to the LR−. (B) For a multichotomous test, with 2 cut-points (X and Y), there are three intervals corresponding to three possible test results. R1 is the interval for a test from (0 to Y], R2 is the interval for a test from (>Y to X], and R3 is the interval for a test result > X. The LR for interval R1 is equal to the slope of the dotted red line from Y to (1,1), the LR for interval R2 is equal to the slope of the red long dashed line from X to Y, and the LR for interval R3 is equal to the slope of the solid red line from (0, 0) to X. (C) The LR for the a continuous test with value X is given by the tangent (X) = slope d (blue line). Dichotomous formulation with a single cut-point (green), and interval formulations (red) are shown for comparison. If a continuous test (black ROC curve) is converted into a dichotomous test (green lines), the AUC of the dichotomous test is less than the continuous version of the test. This reflects the impact of simplification and the resulting loss of information. As the number of intervals increases (red lines in B), the area approaches that of the continuous test (i.e., less information is loss).

References

1. Vachani A, Maldonado F, Laxmanan B, Kalsekar I, Murgu S. The Impact of Alternative Approaches to Diagnostic Yield Calculation in Studies of Bronchoscopy. *Chest.* 2022;161(5):1426-1428.
2. Thiboutot J, Yarmus LB, Lee HJ, Rivera MP, Ost DE, Feller-Kopman D. Real-World Application of the NAVIGATE Trial. *J Thorac Oncol.* 2019;14(7):e146-e147.
3. Folch EE, Pritchett MA, Nead MA, et al. Electromagnetic Navigation Bronchoscopy for Peripheral Pulmonary Lesions: One-Year Results of the Prospective, Multicenter NAVIGATE Study. *J Thorac Oncol.* 2019;14(3):445-458.
4. Lederer DJ, Bell SC, Branson RD, et al. Control of Confounding and Reporting of Results in Causal Inference Studies. Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. *Annals of the American Thoracic Society.* 2019;16(1):22-28.
5. Leisman DE, Harhay MO, Lederer DJ, et al. Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. *Crit Care Med.* 2020;48(5):623-633.
6. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open.* 2016;6(11):e012799.
7. Rivera MP, Mehta AC, Wahidi MM. Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest.* 2013;143(5 Suppl):e142S-e165S.
8. Bossuyt P, Davenport C, Deeks J, Hyde C, Leeflant M, Scholten R. Chapter 11: Interpreting results and drawing conclusions. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Accuracy Version 0.9*: The Cochrane Collaboration; 2013.
9. Leong TL, Loveland PM, Gorelik A, Irving L, Steinfort DP. Preoperative Staging by EBUS in cN0/N1 Lung Cancer: Systematic Review and Meta-Analysis. *J Bronchology Interv Pulmonol.* 2019;26(3):155-165.
10. Silvestri GA, Gonzalez AV, Jantz MA, et al. Methods for staging non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest.* 2013;143(5 Suppl):e211S-e250S.
11. Guinde J, Bourdages-Pageau E, Collin-Castonguay MM, et al. A Prediction Model to Optimize Invasive Mediastinal Staging Procedures for Non-small Cell Lung Cancer in Patients With a Radiologically Normal Mediastinum: The Quebec Prediction Model. *Chest.* 2021;160(6):2283-2292.
12. Vakil E, Jackson N, Sainz-Zunega PV, et al. Optimizing Diagnostic and Staging Pathways for Suspected Lung Cancer: A Decision Analysis. *Chest.* 2021;160(6):2304-2323.
13. Yasufuku K, Chiyo M, Koh E, et al. Endobronchial ultrasound guided transbronchial needle aspiration for staging of lung cancer. *Lung Cancer.* 2005;50(3):347-354.
14. Lee HS, Lee GK, Lee HS, et al. Real-time endobronchial ultrasound-guided transbronchial needle aspiration in mediastinal staging of non-small cell lung cancer: how many aspirations per target lymph node station? *Chest.* 2008;134(2):368-374.
15. Gilbert S, Wilson DO, Christie NA, et al. Endobronchial ultrasound as a diagnostic tool in patients with mediastinal lymphadenopathy. *Ann Thorac Surg.* 2009;88(3):896-900; discussion 901-892.

16. Hwangbo B, Kim SK, Lee HS, et al. Application of endobronchial ultrasound-guided transbronchial needle aspiration following integrated PET/CT in mediastinal staging of potentially operable non-small cell lung cancer. *Chest.* 2009;135(5):1280-1287.
17. Cetinkaya E, Seyhan EC, Ozgul A, et al. Efficacy of convex probe endobronchial ultrasound (CP-EBUS) assisted transbronchial needle aspiration for mediastinal staging in non-small cell lung cancer cases with mediastinal lymphadenopathy. *Ann Thorac Cardiovasc Surg.* 2011;17(3):236-242.
18. Lee KJ, Suh GY, Chung MP, et al. Combined endobronchial and transesophageal approach of an ultrasound bronchoscope for mediastinal staging of lung cancer. *PloS one.* 2014;9(3):e91893.
19. Dziedzic D, Peryt A, Szolkowska M, Langfort R, Orlowski T. Endobronchial ultrasound-guided transbronchial needle aspiration in the staging of lung cancer patients. *SAGE Open Med.* 2015;3:2050312115610128.
20. Fréchet B, Kazakov J, Thiffault V, Ferraro P, Liberman M. Diagnostic Accuracy of Mediastinal Lymph Node Staging Techniques in the Preoperative Assessment of Nonsmall Cell Lung Cancer Patients. *J Bronchology Interv Pulmonol.* 2018;25(1):17-24.
21. Croft DR, Trapp J, Kernstine K, et al. FDG-PET imaging and the diagnosis of non-small cell lung cancer in a region of high histoplasmosis prevalence. *Lung cancer (Amsterdam, Netherlands).* 2002;36(3):297-301.
22. Deppen S, Putnam JB, Jr., Andrade G, et al. Accuracy of FDG-PET to diagnose lung cancer in a region of endemic granulomatous disease. *Ann Thorac Surg.* 2011;92(2):428-432; discussion 433.
23. Services USDoHaH, Administration FaD, Health CfDaR, Branch DD, Biostatistics Do, Biometrics OoSa. Guidance for Industry and FDA Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests. 2007.
24. Fletcher RH, Flectcher SW, S. FG. Chapter 8: Diagnosis. *Clinical Epidemiology: The Essentials, Fifth Edition*. Baltimore, MD, USA: Lippincott Williams & Wilkins; 2014.
25. Ost DE. Interpretation and Application of the Likelihood Ratio to Clinical Practice in Thoracic Oncology. *J Bronchology Interv Pulmonol.* 2022;29(1):62-70.
26. McGee S. Simplifying likelihood ratios. *Journal of general internal medicine.* 2002;17(8):646-649.
27. Rubinstein ML, Kraft CS, Parrott JS. Determining qualitative effect size ratings using a likelihood ratio scatter matrix in diagnostic test accuracy systematic reviews. *Diagnosis (Berl).* 2018;5(4):205-214.
28. Grosu HB, Iliesiu M, Caraway NP, et al. Endobronchial Ultrasound-Guided Transbronchial Needle Aspiration for the Diagnosis and Subtyping of Lymphoma. *Annals of the American Thoracic Society.* 2015;12(9):1336-1344.
29. Choi BC. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American journal of epidemiology.* 1998;148(11):1127-1132.
30. Ost DE. Interpretation and Application of the Likelihood Ratio to Clinical Practice in Thoracic Oncology. *J Bronchology Interv Pulmonol.* 2021.
31. Black WC, Armstrong P. Communicating the significance of radiologic test results: the likelihood ratio. *AJR Am J Roentgenol.* 1986;147(6):1313-1318.
32. Glazer GM, Orringer MB, Gross BH, Quint LE. The mediastinum in non-small cell lung cancer: CT-surgical correlation. *AJR Am J Roentgenol.* 1984;142(6):1101-1105.
33. Ost DE, Ernst A, Lei X, et al. Diagnostic Yield and Complications of Bronchoscopy for Peripheral Lung Lesions. Results of the AQuIRE Registry. *Am J Respir Crit Care Med.* 2016;193(1):68-77.
34. Kalchiem-Dekel O, Connolly JG, Lin IH, et al. Shape-Sensing Robotic-Assisted Bronchoscopy in the Diagnosis of Pulmonary Parenchymal Lesions. *Chest.* 2022;161(2):572-582.