Setting Bias Specifications Based on Qualitative Assays With a Quantitative Cutoff Using COVID-19 as a Disease Model

Chun Yee Lim, PhD,1 Wei Zhi Chang,1 Corey Markus,2 Andrea Rita Horvath, MD,3 and Tze Ping Loh, MBBChBAO4

From the 1Engineering Cluster, Singapore Institute of Technology, Singapore; 2Flinders University International Centre for Point-of-Care Testing, Bedford Park, Australia; 3Department of Chemical Pathology, New South Wales Health Pathology, Prince of Wales Hospital, Sydney, Australia; and 4Department of Laboratory Medicine, National University Hospital, Singapore.

ABSTRACT

Objectives: Automated qualitative serology assays often measure quantitative signals that are compared against a manufacturer-defined cutoff for qualitative (positive/negative) interpretation. The current general practice of assessing serology assay performance by overall concordance in a qualitative manner may not detect the presence of analytical shift/drift that could affect disease classifications.

Methods: We describe an approach to defining bias specifications for qualitative serology assays that considers minimum positive predictive values (PPVs) and negative predictive values (NPVs). Desirable minimum PPVs and NPVs for a given disease prevalence are projected as equi-PPV and equi-NPV lines into the receiver operator characteristic curve space of coronavirus disease 2019 serology assays, and the boundaries define the allowable area of performance (AAP).

Results: More stringent predictive values produce smaller AAPs. When higher NPVs are required, there is lower tolerance for negative biases. Conversely, when higher PPVs are required, there is less tolerance for positive biases. As prevalence increases, so too does the allowable positive bias, although the allowable negative bias decreases. The bias specification may be asymmetric for positive and negative direction and should be method specific.

Conclusions: The described approach allows setting bias specifications in a way that considers clinical requirements for qualitative assays that measure signal intensity (eg, serology and polymerase chain reaction).

INTRODUCTION

Qualitative serology assays detect the presence of human antibodies against specific antigens, including antigens from a pathogen in a clinical sample. They often employ principals of immunoassay when applied on stand-alone, automated platforms. In such assays, the presence of the target antibody will generate signals that the instrument detects and measures.1 The observed signal intensity is compared against a manufacturer-defined cutoff for qualitative interpretation (commonly expressed as a signal-to-cutoff [S/C] index or...
cutoff index (COI), and then reported in binary terms of positive and negative detection, but the report may include an equivocal or gray zone.

The discriminating power of a qualitative serological test can be described by its clinical performance in terms of clinical sensitivity and clinical specificity when assessed against a reference criterion—for example, when severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) serological assays are compared against the results of viral detection by molecular techniques, such as polymerase chain reaction (PCR) assays. The receiver operating characteristic (ROC) curve is a graphical representation of clinical performance in binary classification settings, when the discriminating threshold (S/C or COI) varies throughout the measurement range. By varying the discriminating threshold, relative measures of sensitivity and specificity can be obtained and plotted. The derived area under the ROC curve can be considered the probability that an assay will correctly classify a new sample; it is independent of disease prevalence.²

In internal quality control systems and external quality assurance (EQA) programs, serology assays are generally assessed for overall concordance in a qualitative fashion with peer laboratories or the results of molecular diagnostics, such as PCR assays. These qualitative approaches may not detect the presence of analytical shifts or drift—that is, the systematic error characterized by constant or proportional bias—that could affect disease classification.³ Use of the signal output or S/C ratio from automated analytical instruments provides an opportunity to apply the concepts of analytical performance metrics to qualitative serology assays.

Coronavirus disease 2019 (COVID-19), which is caused by the SARS-CoV-2 virus, has focused the attention of clinical practitioners and the public on the clinical performance of laboratory assays. The importance of the clinical sensitivity and specificity of qualitative serology assays for disease detection and public health surveillance measures was recently discussed by Loh et al.² Particularly in a global pandemic, it is important that the assay maintain minimum clinical performance to ensure that it achieves the desired posttest predictive value for the clinical scenario to which the assay is applied.

Using COVID-19 as a model disease, we attempted to define bias specifications for qualitative serology assays by projecting minimum clinical performance metrics onto a ROC space, simultaneously considering predictive values and disease prevalence.

**MATERIALS AND METHODS**

**Proposed Approach for Setting Analytical Performance Specifications**

Clinical sensitivity refers to a test’s ability to correctly identify patients with a disease or condition, whereas clinical specificity refers to a test’s ability to correctly identify individuals without the disease or condition. Mathematically, sensitivity and specificity are represented by the following equation:

\[
\text{Clinical sensitivity} = \frac{TP}{TP + FN}
\]

where TP is a positive laboratory result in a patient with the disease or condition, TN is a negative laboratory result in a patient without the disease, FP is a positive test result in the absence of disease, and FN is a negative test result in the presence of disease.

The ROC is a statistical tool that describes the relationship between clinical sensitivity and clinical specificity at different thresholds used to define a positive result. The ROC curve is plotted with sensitivity (true-positive rate) on the y-axis and 1 − specificity (false-positive rate) on the x-axis. Thomas and colleagues previously described an approach that expresses posttest predictive values in a ROC space for assessing the introduction of new prognostic tests. We extended this concept to qualitative serology assays in this study to define maximum analytical performance specifications that produce minimum clinical performance specifications.

Under this approach, the ROC curve for a laboratory assay is first constructed using the raw data from diseased and nondiseased populations. Each point on the ROC curve is obtained by calculating the sensitivity and false-positive rate when the S/C value is varied. The analytical cutoff value recommended by the manufacturer is also indicated on the ROC for reference. Following this, the desired minimum positive predictive value (PPV) and negative predictive value (NPV), representing the minimum clinical performance measure, for a particular clinical use case are

\[
\text{Clinical specificity} = \frac{TN}{TN + FP}
\]
defined. Incorporating the prevalence of the disease, post-test predictive values are then plotted on the ROC.

Of note, the posttest predictive values are dependent on the prevalence of the disease or condition in the population where the assay is applied. The required minimum PPV and NPV are represented on the ROC curve as equi-PPV and equi-NPV lines (ie, the values of equi-PPV and equi-NPV along the respective lines are constant for a particular value of disease prevalence).

The equation for the equi-PPV line can be expressed as follows:

$$Sensitivity = \frac{(1 - prevalence)}{prevalence} \cdot \frac{PPV}{(1 - PPV)} \cdot (1 - specificity)$$

The equation for the equi-NPV line can be derived as follows:

$$Sensitivity = \frac{(1 - prevalence)}{prevalence} \cdot \left(\frac{1}{NPV} - 1\right) \cdot (1 - specificity)$$

$$+ \left[1 - \frac{(1 - prevalence)}{prevalence} \cdot \frac{1}{NPV} - 1\right]$$

The equations are derived to show the slope and intercept for the 2 lines on the ROC curve, where the x-axis and y-axis are a false-positive rate equivalent to (1 – specificity) and sensitivity, respectively (see Supplemental Material; all supplemental materials can be found at American Journal of Clinical Pathology online). Hence, the slope of each line is the coefficient for the (1 – specificity) term in the equation. The intercept of the equi-PPV line is 0, while the intercept for the equi-NPV line is the last term in the equation (in square brackets).

The 2 equi-predictive value lines divide the ROC into 4 quadrants. The upper left quadrant (the shaded region in Figure 1) represents the allowable area of performance (AAP) within which the assay meets the a priori defined minimum clinical performance measures. The presence of an analytical bias or drift can influence the S/C ratio and result in movement of the ROC curve. If a positive bias is present, the S/C ratio will be higher; hence, the sensitivity and false-positive rate of the assay will be increased. This approach has the same effect as adopting an analytical cutoff value (the denominator in the S/C ratio) lower than the manufacturer’s recommended cutoff.

The maximum analytical bias whereby the ROC values remain within the AAP is considered the performance specification for the assay because it maintains the a priori defined posttest predictive values. The maximum allowable negative bias is determined as the S/C ratio that corresponds to the intersection between the minimum NPV line and the ROC curve. Similarly, the maximum allowable positive bias is determined as the corresponding S/C ratio obtained at the intersection between the minimum PPV line and the ROC curve. Using this AAP approach for assay performance while maintaining minimum clinical performance, we applied these calculations to 4 SARS-CoV-2 serology assays.

Commercial SARS-CoV-2 Assay Data

Four qualitative SARS-CoV-2 serology assays performed on major commercial stand-alone instruments by Siemens, Roche, DiaSorin, and Abbott were comprehensively evaluated by the National SARS-CoV-2 Serology Assay Evaluation Group. In that study, the instrument signal from healthy participants with no history of past SARS-CoV-2 infection and participants with PCR-confirmed SARS-CoV-2 infection were used to determine the clinical sensitivity and specificity of the assays. The publicly available raw data were obtained from the journal website and used for the simulations outlined below.

The ROC curves for each of the 4 SARS-CoV-2 serology assays is plotted using the raw data from the previous study, along with the manufacturer-recommended threshold value. In this study, to determine the AAP, the desired PPV and NPV were arbitrarily set as 80% and 96%, respectively, for the purpose of diagnosing past SARS-CoV-2 infection in a population with 5%, 10%, and 20% prevalence. For comparison, a second, stricter set of specifications with minimum PPV and NPV set to 90% and 99%, respectively, were also examined under the same prevalence settings. A minimum NPV value of 96% was selected because the y-intercept (sensitivity) for the equi-NPV line will include negative values if the NPV is set to 95% or less for a prevalence of 0.05 (see Supplementary Figure 1).

Ethics Approval

This study involved only simulations that used publicly available data and was exempted from local ethics board review.

RESULTS

The ROC of the 4 commercial assays, together with the predefined posttest predictive values and AAP, are shown in Figures 2-5. In general, more stringent clinical specifications produce smaller AAP (ie, the area within which the assay meets the a priori defined minimum clinical performance measures).

When higher NPVs are required, there is lower tolerance for negative biases. Conversely, when higher PPVs are required, there is less tolerance for positive biases. As prevalence increases, so too does the allowable positive bias; however, the allowable negative bias decreases. The bias specifications for the 4 SARS-CoV-2 assays are summarized in Table 1.

DISCUSSION

A positive analytical shift can increase the false-positive rate, which can lead to overdiagnosis of a condition. Overdiagnosis may lead to unnecessary downstream clinical investigation, treatment, patient anxiety, and an overall increase in health care costs. Some of these clinical interventions may be invasive or even harmful. In the case of COVID-19, a false-positive result may lead to unnecessary quarantine of a person with other patients who has the infection (eg, at a centralized health care facility) and thus may expose a healthy person to the disease. In contrast, a negative analytical shift can increase the false-negative rate, leading to underdiagnosis of a condition. Underdiagnosis can cause delayed investigation and treatment that may lead to adverse outcomes. In the case of COVID-19, a
false-negative result may lead to further community spread by the index patient and suboptimal clinical management.\textsuperscript{5}

It is therefore important to consider the clinical implications of the false-negative and false-positive results when selecting the appropriate NPV and PPV. When a high NPV is clinically required, there is a reduced tolerance for negative biases because they will (artificially) reduce observed assay signal intensity and erroneously turn positive results into negative. This reduced tolerance can be seen by the horizontal upward shift in the equi-NPV line, which reduces the lower boundary of the quadrant [FIGURES 2-5]. Conversely,
A stringent PPV requirement carries a reduced tolerance for positive bias because the bias will increase the false-positive rate, as represented by the leftward vertical shift in the equi-PPV line (again reducing the boundary of the quadrant).

When the disease prevalence is high and to maintain the predefined minimum NPV, there is a further reduction in the tolerance for negative biases. This phenomenon is reversed when the prevalence is low.
Analytical performance specifications are important laboratory criteria that help maintain assay performance, ensuring that they continue to meet required clinical purposes. The proposed AAP approach overcomes the challenge of setting bias specification for qualitative serology tests, but there are several important considerations when setting bias specification using the proposed approach.

**FIGURE 4** Receiver operating characteristic (ROC) curves for the Elecsys Anti-SARS-CoV-2 assay (Roche). The minimum equi-positive predictive value (PPV) and equi-negative predictive value (NPV) lines are plotted for a prevalence of 0.05 (A), 0.10 (B), and 0.20 (C). The star symbols represent the clinical performance the assay achieved with the recommended cutoff value (1.0). The minimum equi-PPV and equi-NPV in the subplot on the left charts are specified as 80% and 96%, respectively, while the minimum equi-PPV and equi-NPV in the subplot on the right charts are set to 90% and 99%, respectively. The shaded upper left quadrants show the region within which the assay can satisfy the predefined posttest predictive values.
Bias Specifications Are Linked to the Clinical Use Case

Using the clinical performance characteristics and a priori defined posttest predictive values as boundaries, the bias specification defined is directly linked to the expected clinical use case. The analytical performance specification can be defined differently for different clinical use scenarios while maintaining the same empirical linkage to a priori defined clinical requirements.
Bias Specification Should Be Tailored to the Specific Assay

The ROC of the 4 commercial assays showed different characteristics, with varying prevalence and performance requirements (TABLE 1), because the assays vary in discriminating power between diseased and nondiseased samples (ie, clinical sensitivity and clinical specificity). Additionally, the greater instrument signal range that the DiaSorin and Roche assays produce are associated with considerably larger bias specifications (ie, greater numerical tolerance for bias). The use of a common analytical performance specification for all 4 assays is likely to result in suboptimal monitoring, being too lenient for some and too stringent (unachievable) for others.

Bias Specifications May Be Asymmetric

Conventionally, analytical performance specifications are set such that the limits are symmetrical around some target value. In reality, the analytical cutoff may be optimized in favor of clinical sensitivity or specificity. As such, an assay may have asymmetric tolerances to analytical bias in either direction.

To apply this approach, it is envisioned that EQA programs distribute samples with antibody concentrations near the manufacturer-defined cutoffs to participating laboratories. The participating laboratories return numerical signal intensity (eg, S/C ratio or COI) to the EQA provider, and a target value can be assigned using a consensus value approach, either peer-mean or peer-median, for a given method. The performance of an individual laboratory is then assessed against the target value (eg, peer-mean or peer-median) and the bias specifications within the same method group and clinical use case scenario (ie, same NPV/PPV requirements). Graphically, this assessment may be presented by plotting the performance of the individual laboratory along with other peer laboratories within the same ROC space bound by the same equi-NPV and equi-PPV. This approach can also be applied to other laboratory assays that generate numerical signal values with qualitative interpretation based on a cutoff, such as real-time PCR assays. Moreover, this approach may be applicable to other areas of laboratory practices that seek to monitor analytical shift or drift, such as internal quality control, EQA/proficiency testing, and lot-to-lot reagent verification. Nevertheless, this approach requires raw instrument data (signal intensity), which may not be readily accessible to all laboratory users.

We have described performance metrics for serology assays based on a continuous measurement scale (S/C ratio) that are interpreted in a qualitative manner. These performance metrics are consistent with the Milan hierarchy level 1b through the impact of analytical performance on patient classification. The AAP derived from the projection of equi-predictive values in a ROC space takes into consideration prevalence and clinical use to ensure that assays maintain an a priori minimum clinical performance.

REFERENCES

1. National SARS-CoV-2 Serology Assay Evaluation Group. Performance characteristics of five immunoassays for SARS-CoV-2: a head-to-head benchmark comparison. Lancet Infect Dis. 2020;20:1390-1400.
2. Thomas G, Kenny LC, Baker FN, et al. A novel method for interrogating receiver operating characteristic curves for assessing prognostic tests. Diagn Progn Res. 2017;1:17.
3. Koerbin G, Liu J, Eigenstetter A, et al. Missed detection of significant positive and negative shifts in gentamicin assay: implications for routine laboratory quality practices. Biochem Med (Zagreb). 2018;28:201705.
4. Rhoads DG. Lab Statistics Fun and Easy: A Practical Approach to Method Validation. Kennett Square, PA: David G. Rhoads Associates; 1996.
5. Loh TP, Lord SJ, Bell K, et al. Setting minimum clinical performance specifications for tests based on disease prevalence and minimum acceptable positive and negative predictive values: practical considerations applied to COVID-19 testing. Clin Biochem. 2021;88:18-22.