Use of Gray Zone in IVD Tests

Tie-Hua Ng*, Ph.D. (tiehua.ng@fda.hhs.gov) and
Paul Hshieh*, Ph.D. (paul.hshieh@fda.hhs.gov)
Center for Biologics Evaluation and Research
US Food and Drug Administration
10903 New Hampshire Ave., Silver Spring, MD 20993

Abstract

Although the estimations of the sensitivity and specificity of an IVD (in-vitro diagnostic) test are based on a simple 2 × 2 table, difference aspects of testing algorithm (e.g., retest, gray zone, pool testing, multiplex) could lead to complexity in the study design and analysis. Complexity due to retest and gray zone (GZ) are discussed. In general, the gray zone is defined as Index value or signal over the cutoff value being in a specified range. Although gray zone is often used in the study, the final interpretation of the assay may or may not involve a GZ. In this paper, we will discuss two examples where a GZ is used for two different purposes. In the first example, the GZ is used so that a retest may be done in the study with the purpose to refine the cutoff value. In the second example, the GZ is simply the retest zone. Although refining the cutoff value based on the current data is against the basic principle, there should be a balance between practicality and basic principle. Therefore, further research is needed to assess the bias and to see how such bias may be adjusted.

Key Words: IVD, Gray zone, Assay cutoff, retest

1. Introduction

In vitro diagnostic (IVD) devices are defined as reagents, instruments, and systems intended for use in the diagnosis of disease or other conditions (Russek-Cohen, et al, 2011). This paper focuses on IVDs which have a continuous outcome such as the Signal over the cutoff value (S/CO) or the Index value (IDX). Usually, if S/CO or IDX < 1, it is interpreted Non-Reactive (NR). If it is ≥ 1, it is interpreted as Reactive (R). Very often, we see a third category called Gray Zone (GZ). In this paper, without loss of generality, the GZ is defined as 0.90 – 0.99. Technically, it means ≥ 0.9 and < 1. Other ranges such as 0.8 to 1.2 may also be used.

2. Performance Characteristics

The performance of an IVD is generally assessed by reproducibility/precision, sensitivity and specificity. This paper focuses on the sensitivity and specificity. The sensitivity is the probability that the device will have a Reactive (or positive) test result given that the subject has the disease and the specificity is the probability that the device will have a Non-Reactive (or negative) test result given that the subject does not have the disease. The sensitivity is estimated in the clinical study as the proportion of subjects that are positive by the device among those with the disease and the specificity is estimated as the

[*] The views expressed in this paper are those of the authors and do not necessarily reflect the perspective of the U.S. Food and Drug Administration.

proportion of subjects that are negative by the device among those without the disease. If the results are summarized in a 2 × 2 table as shown in Table 1a, then the sensitivity is estimated by a/(a+c) and specificity is estimated by d/(b+d).

Table 1a. Summary Results by Disease Status

Test result	Disease	
	Yes	No
Reactive	A	b
Non-Reactive	C	d

Sensitivity = a/(a +c); Specificity = d/(b+d)

Table 1b. Summary Results by Reference/Gold Standard

Test result	Reference/Gold Standard	
	Reactive	Non-Reactive
Reactive	TP	FP
Non-Reactive	FN	TN

TP: True Positive; FP: False Positive; TN: True Negative; FN: False Negative
Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP)

In reality, the disease statuses of subjects enrolled in the study are not known. So, we have to rely on a reference or gold standard test. In that case, the results can be summarized in a 2 × 2 table as shown in Table 1b where a, b, c and d in Table 1a are replaced by TP, FP, FN and TN, respectively, where TP, FP, FN and TN denote True Positive, False Positive, False Negative and True Negative, respectively. Therefore, the sensitivity is estimated by TP/(TP + FN) and the specificity is estimated by TN/(TN + FP). In this paper, TP, FP, FN and TN are used interchangeably as the numbers or the acronyms. For example, TP may denote the number of true positives or stands for “true positive”.

3. Comparison of CBER and CDRH

The Center for Biologics Evaluation and Research (CBER) within the Food and Drug Administration (FDA) regulates IVDs for blood donor screening indications detecting viruses such as the human immunodeficiency virus (HIV), the hepatitis B virus (HBV), the hepatitis C virus (HCV), west nile virus (WNV), human T-cell lymphotropic virus (HTLV), etc. CBER also regulates IVDs with diagnostics indications for retroviruses such as HIV and HTLV. In contrast, the Center for Devices and Radiological Health (CDRH) within the FDA regulates most IVDs with diagnostic indications. In CBER, most assays require a retest in duplicate when it is initially reactive with few exceptions, while in CDRH, there is no retest when it is initially reactive. See Russek-Cohen, et al (2011) for the regulatory pathways for IVDs.

4. Complexity

The assessment of the performance discussed in Section 2 is nothing but a simple 2 × 2 table such as Table 1b. However, the study design or testing algorithm can be very complex and confusing due to retest, gray zone (GZ), multiplex, pool testing and other factors or dimensions. Multiplex means the assay can detect two or more analytes. For example, Procleix Ultrio Plus Assay is a multiplex which can simultaneously detect HIV-1, HBV and HCV. A nucleic acid testing (NAT) for blood donor screening is often done on pooled samples due to the cost and labor intensiveness. In Section 5, two examples are given to show how complexity arises due to retest and GZ.

5. Examples

In Example 1, if the initial result is NR, there is no retest. The final interpretation is NR. However, if the initial result is R, we say it is initially reactive (IR) and it is required to retest in duplicate using the same samples in the sense that the samples are coming from the same drawn. If both retest results is < 1, the final interpretation is NR. If at least one retest results ≥ 1, the final interpretation is R. In that case, we say that it is Repeatedly Reactive (RR). Note that the sensitivity and specificity are calculated based on the final interpretation instead of the initial interpretation. We see that it gets a little complicated but we simply use the final interpretation in the performance calculations.

Table 2a. Example 1: Retest in Duplicate

Initial Result	Initial Interpretation	Retest Procedure	Retest Result	Final Interpretation
< 1.00	Non-reactive			Non-reactive
≥ 1.00	Reactive (IR[1])	Retest in duplicate	Both < 1.00	Non-reactive
			One ≥ 1.00	Reactive (RR[2])

[1] IR: Initially Reactive

[2] RR: Repeatedly Reactive

In Example 2, we have a gray zone (GZ). In addition, we also have a retest in duplicate when the initial result falls in the GZ. If both retest results are < 1, the final interpretation is NR. If at least one retest results are ≥ 1, the final interpretation is R. If the initial result is outside the GZ, i.e. < 0.9 or ≥ 1, no retest is needed and the final interpretation is the same as the initial interpretation. With the GZ plus retest, it gets a little more complicated and interesting.

Table 2b. Example 2: Gray Zone plus Retest

Initial Result	Initial Interpretation	Retest Procedure	Retest Result	Final Interpretation
< 0.90	Non-reactive			Non-reactive
0.90 – 0.99	Gray Zone	Retest in duplicate	Both < 1	Non-reactive
			One ≥ 1.00	Reactive
≥ 1.00	Reactive			Reactive

In both examples, we are talking about how the device is going to be used in practice. This testing algorithm is described in ether Package Insert (PI), Labelling, User Manual or Instructions for Use (IFU). In addition, the sensitivity and specificity are calculated based on the final interpretation. Although the second example has a GZ, the final interpretation is either NR or R. The GZ in this example is simply a retest zone.

6. Testing Algorithm in the Clinical Study

Testing Algorithm in the study is more complicated for two reasons. First, we have to do other tests to determine the disease status. Second, there could be additional testing using the investigational device. This paper focuses on the second one.

Why do we do additional testing using Investigational device? The reason is to refine the cutoff value in Example 1 and to expand the Gray Zone (e.g., ≥ 0.9 and < 1.05) in Example 2. In Example 1, we have data to adjust the cutoff upward in the sense to interpret the results but we don't have the data to adjust cutoff downward. For example if you want to adjust the cutoff from 1 to 0.95, we don't have the retest results if the initial result is, say, 0.96. So, we need additional testing in the study. To do so, we create a GZ defined by 0.90 and 0.99. In the study, we need to retest in duplicate if the initial result is in GZ (see Table 3a).

Table 3a. Example 1: Retest in Duplicate in the study to Refine the Assay Cutoff

Initial Result	Initial Interpretation	Retest Procedure
< 0.90	Non-reactive	
0.90 – 0.99	Gray Zone	*Retest in duplicate*
1.00	Reactive	Retest in duplicate

In Example 2, we have the GZ originally and we don't have the data to expand the GZ upward in the sense to interpret the results. So, we need additional testing. So, in the study, we need to retest in duplicate if the initial result is R (see Table 3b).

Table 3b. Example 2: Retest in Duplicate in the study to Expand the Gray Zone

Initial Result	Initial Interpretation	Retest Procedure
< 0.90	Non-reactive	
0.90 – 0.99	Gray Zone	Retest in duplicate
1.00	Reactive	*Retest in duplicate*

Although the testing procedures in the package insert are different in the two examples, the same testing algorithm is used in the study for both examples, i.e., we retest in duplicate if the initial result is GZ or R. The GZ is used in Example 1 in the study so that the sample will be retested if the initial result is in GZ. The purpose is to refine the cutoff value. On the other hand, in Example 2, the GZ is simply the retest zone when the device is used in practice and the purpose of additional testing in the study is to expand the GZ/retest zone.

Therefore, the evaluation of the sensitivity and specificity of an IVD is not as simple as it appears. Furthermore, to assess the study design (or testing algorithm), one must know how the device is going to be used in practice.

7. Refining the Cutoff Value in Example 1

We consider 4 cutoff values, namely, 0.90, 0.95, 1.00 and 1.05. For example, if the cutoff is 0.95, then the result is interpreted as NR, if it is < 0.95 and R, otherwise (see Table 4a).

We use 5 hypothetical samples with initial results as well as retest results as given in Table 4b. The result of the Gold Standard (GS) for each sample can be either R or NR. So, there are 32 (= 2^5) possible scenarios. We consider one scenario where Samples 1, 2 and 3 are R and Samples 4 and 5 are NR by the GS. These are highlighted in yellow in Table 4b.

For each combination of sample and cutoff value, we can determine whether the final interpretation of the sample is NR or R (same as RR). Following the definitions in Table 1b (replacing "initial results" by "final interpretation"), (i) a NR sample is a TN, if the GS is NR and a FN, if the GS is R, and (ii) a RR sample is a TP, if the GS is R and a FP, if the GS is NR (see Table 4b). For example, for Sample 2, if we keep the cutoff value as 1, then the initial result is NR and so is the final interpretation; therefore, Sample 2 is a TN, if the GS is NR, and a FN, if the GS is R. With a cutoff value of 0.95, the initial result for Sample 2 is R and the final interpretation is RR; therefore, Sample 2 is a TP, if the GS is R, and a FP, if the GS is NR.

Table 4a. Interpretations of Four Cutoff Values

Cutoff values	Results	Interpretations
0.90	< 0.90	Non-Reactive
	≥ 0.90	Reactive
0.95	< 0.95	Non-Reactive
	≥ 0.95	Reactive
1.00	< 1.00	Non-Reactive
	≥ 1.00	Reactive
1.05	< 1.05	Non-Reactive
	≥ 1.05	Reactive

Table 4b. Five Hypothetical Samples and the Interpretations

Sample	Initial results	Retest results		Reference/Gold Standard	Cutoff Values			
					0.90	0.95	1.00	1.05
1	0.94	0.97	0.89	Reactive	TP	FN	FN	FN
				Non-Reactive	FP	TN	TN	TN
2	0.97	0.92	0.96	Reactive	TP	TP	FN	FN
				Non-Reactive	FP	FP	TN	TN
3	0.98	1.09	1.12	Reactive	TP	TP	FN	FN
				Non-Reactive	FP	FP	TN	TN
4	1.02	0.94	1.07	Reactive	TP	TP	TP	FN
				Non-Reactive	FP	FP	FP	TN
5	1.06	0.98	1.08	Reactive	TP	TP	TP	TP
				Non-Reactive	FP	FP	FP	FP

Table 4c. Summary Results

Interpretations	Cutoff Values			
	0.90	0.95	1.00	1.05
TP	3	2		
FN		1	3	3
TN				1
FP	2	2	2	1

The numbers of TPs, FNs, TNs and FPs for each cutoff value are summarized in Table 4c. There are 3 FNs and 2 FPs if we keep the cutoff value of 1.0. If the cutoff value is increased from 1.00 to 1.05, the number of FNs remains the same while the number of FPs is decreased by 1. If the cutoff value is decreased from 1.00 to 0.95 and 0.90, the number of FPs remains the same while the number of FNs is decreased by 2 and 3, respective. Therefore, the cutoff value of 0.90 appears to be "optimal". See Section 8 for a discussion of the issues due to refining the cutoff value.

In general, lowering the cutoff (e.g., from 1 to 0.9) will increase the sensitivity at the expense of lower specificity. On the other hand, raising the cutoff (e.g., from 1 to 1.05) with increase the specificity at the expense of lower sensitivity.

8. Discussion and Further Research

This paper discusses the complexity which arises due to retest and gray zone. Complexity may arise due to many other factors such as multiplex, pooling, types of assay (e.g., serology, NAT) and sample types (e.g., plasma, serum, and finger stick) which are not discussed in this paper. Determination of the disease status is another aspect of the clinical study which could lead to complexity in the study design but is not discussed in this paper.

In principle, the assay cutoff value should be determined before the clinical study is conducted with the objective to validate the assay cutoff value. Refining the cutoff based on the current data is against this basic principle. Using the refined cutoff value will overestimate the sensitivity and specificity because the data is used twice --- one in redefining the cutoff value and one the estimation. However, it is not practical to ask for a second study to validate the new cutoff. There should be a balance between practicality and the basic principle.

A big question is: How do we adjust the bias? Further research is needed to assess the bias and to see how such bias may be adjusted.

Reference

Russek-Cohen E, Feldblyum T, Whitaker KB, and Hojvat S (2011). FDA Perspectives on Diagnostic Device Clinical Studies for Respiratory Infections. *Clinical Infectious Diseases* 52(S4):S305-S311.