Statistical Considerations in Clinical Studies Supporting Blood Donor Screening Assays

Zhen Jiang[1]
[1]Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD 20993

Abstract
21 CFR § 610.40 requires using donor screening assays to control transfusion-transmitted infectious diseases. The implementation of donor screening assays along with donor selection has dramatically reduced disease transmissions risk from blood transfusion. In this paper, we will discuss the unique features of donor screening assays (e.g. intended population, technical characteristics) comparing to diagnostic assays. Some of these features lead to the particular study designs we see and several issues in the clinical sensitivity and specificity studies, which we will illustrate and discuss in detail.

Key Words: Donor screening assay, sensitivity, specificity

Disclaimer: This paper reflects the views of the author and should not be construed to represent FDA's views or policies.

1. Background

Since the beginning of modern blood banking, controlling transfusion-transmitted diseases has been a constant challenge [1]. Besides donor selection including the use of questionnaires, donor screening assays were implemented to address this challenge [2]. Title 21, Section 610.40(a) of the Code of Federal Regulations (21 CFR § 610.40(a)) requires establishments that collect blood or blood components to test each donation of human blood or blood component for evidence of infection due to HIV-1, HIV-2, HBV, HCV, HTLV-I, and HTLV-II. In addition, 21 CFR § 610.40(b) requires using screening assays approved by the U.S. Food and Drug Administration (FDA).

Both regulatory agency and industry have made enormous efforts to develop donor screening assays for transfusion-transmitted infectious agents. For example, since the first-generation screening test for HIV antibodies implemented in 1985, more than 50 versions of HIV serologic assays have been developed [2]. Together with donor selection, the risk of HIV transmission through blood transfusion reduced from 1 in 2500 units prior to HIV testing to a current estimated residual risk of 1 in 1.47 million [3, 4]. Also, no transmissions of HIV, HBV, or HCV have been documented through U.S.-licensed plasma derived products in the past two decades [5].

Table 1 lists select approved donor screening assays for HIV, HBV and HCV. In recent years, several multiplex assays, which simultaneously detect or even differentiate multiple infectious agents, have been developed. A complete list of approved screening assays for infectious agents (i.e., HBV, HCV, HIV, HTLV, T. pallidum, T. cruzi, WNV, and CMV) can be found on FDA's website [6].

Table 1: Select Approved Donor Screening Assays for HIV, HBV and HCV

Infectious Agent	Trade Name	Format	Approval Date
HBV	ABBOTT PRISM HBsAg	ChLIA	7/18/2006
	ABBOTT PRISM HBcore	ChLIA	10/13/2005
	UltraQual™ HBV PCR Assay	PCR	9/1/2011
	COBAS AmpliScreen HBV Test	PCR	4/21/2005
HCV	Abbott PRISM HCV	ChLIA	7/11/2007
	Hepatitis C Virus RT PCR Assay	PCR	2/9/2007
HIV-1	COBAS AmpliScreen HIV-1 Test, version 1.5	Qualitative PCR	12/20/2002
	Human Immunodeficiency Virus, Type 1 RT PCR Assay	Qualitative PCR	1/31/2007
HIV-1 HIV-2	ABBOTT PRISM HIV O Plus assay	ChLIA	9/18/2009
	Genetic Systems HIV-1/HIV-2 Plus O EIA	EIA	8/5/2003
Multiplex: HBV, HCV, HIV-1, HIV-2	COBAS TaqScreen MPX Test	PCR	12/30/2008
	COBAS TaqScreen MPX Test version 2.0	PCR	12/19/2014
Multiplex: HBV, HCV, HIV-1	Procleix Ultrio Assay	TMA	10/3/2006
	Procleix Ultrio Plus Assay	TMA	5/25/2012

Donor screening assays are very different from regular diagnostic assays in many aspects. First, unlike diagnostic assays, the intended population for screening assays is blood donors with very low prevalence rates for infectious diseases of interest. Second, donors tested positive by screening assays will be followed up for additional testing [7, 8]. The thresholds for positive results in screening are set to err on the side of caution, because false negative results causing safety concerns are more serious than false positive results which could probably be ruled out by additional testing anyways. Third, screening assays are designed to test with high throughput using highly automated technology and provide non-subjective readouts. Finally, while clinical labs may develop their own diagnostic assays, all screening assays must be approved by FDA. More specifically, donor screening assays are regulated under the BLA process by the Center for Biologics Evaluation and Research (CBER) in FDA.

Some of these unique features of screening assays lead to the particular designs and analysis strategy of the clinical performance studies. What's more, nowadays, many manufacturers are developing their second or third generation assays after the approval of their first generation assays. This actually raises some new issues in results interpretation. In the following sections, we will illustrate and discuss these issues and the clinical study designs.

2. Clinical Specificity Study

The clinical specificity study for donor screening assay is usually conducted with low-risk donor population. The study usually requires a minimal of 10,000 individuals and/or 10,000 pools [7], depending on whether the assay tests pools or not. Each sample (individual or pool) is tested by the investigational assay and a comparator assay in

parallel. To resolve discordant results between investigational and comparator assays, additional tests are required.

As we mentioned before, many manufacturers are now developing their second or third generation assays. In this case, they usually use previously approved version as the comparator though it is sometimes augmented with additional testing. Table 2 shows the usual study design and test algorithm. When a sample is tested negative or positive by both assays (i.e., "double negative" or "double positive"), this sample will be considered as true negative or true positive. True positive samples will be excluded from specificity calculation. Additional testing will be conducted to determine the true sample status only when discordant results are observed. In contrast with a diagnostic performance study, only a handful of subjects out of 10,000 typically tested will be positive.

Table 2: Test Algorithm in Clinical Specificity Study

		Comparator (Previous Version)	
		Nonreactive	**Reactive**
Investigational Assay	**Nonreactive**	True Negative	Additional Testing
	Reactive	Additional Testing	True Positive*

* *Exclude from specificity calculation*

One issue in above test strategy is that the true sample status is determined with concordant results from the investigational assay and its previous version. However, these "double positive" or "double negative" results may be false positive or false negative by both assays due to their similar technical characteristics. For example, these two assays may have similar primer and probe designs which are sensitive to similar interference substances or conditions. In other words, there may be systematic false positive and/or false negative results by both assays. These systematic biases will not occur if an independently developed assay is used as the comparator.

Treating "double positive" samples as true positive may lead to bias in specificity estimation. However, due to the nature of this study, it is acceptable to consider "double negative" samples as true negatives. By definition, we know specificity equals one minus false positive rate (FPR) as indicated in Eq (1). Since the study population in clinical specificity study is low-risk donor population, one should observe very low number of positive readings let alone false positives. Therefore, the numerator in FPR is very small relative to the denominator. Any false negative results will only affect the denominator and should have minimal impact on the FPR and specificity calculation. On the other hand, any false positive results will affect the numerator of FPR and have a much bigger impact on the specificity estimate.

$$\text{Specificity} = \frac{\text{True Negative}}{\text{Total Number of Negative}} = 1 - \frac{\text{False Positive}}{\text{Total Number of Negative}} \qquad \text{Eq (1)}$$

Table 3 illustrates the impact of these two types of misclassifications on FPR and specificity calculation. Let us assume, based on the test algorithm in Table 2, we observe a FPR of 0.10% which corresponds to a specificity of 99.90% for the investigational assay. Let us also assume we actually have 2, 5, 10, or 20 false negative results by treating "double negative" samples as true negatives. As the second and third columns of Table 3 show, there is a minimal impact on both FPR and specificity estimates, because false negatives only affect the denominator which is typically a very big number.

However, if we miss same number of false positives, as the last two columns of Table 3 show, there is a much bigger impact on both FPR and specificity estimates, because false positives affect the numerator.

Table 3: Impact of Misclassifications on Specificity and False Positive Rate (FPR)

	Missed False Negatives		**Missed False Positives**	
	FPR	**Specificity**	**FPR**	**Specificity**
N=0	10/10000 **0.10%** (0.05%, 0.18%)	9990/10000 **99.90%** (0.05%, 0.18%)	10/10000 **0.10%** (0.05%, 0.18%)	9990/10000 **99.90%** (99.82%, 99.95%)
N=2	10/9998 **0.10%** (0.05%, 0.18%)	9988/9998 **99.90%** (0.05%, 0.18%)	12/10002 **0.12%** (0.06%, 0.21%)	9990/10002 **99.88%** (99.79%, 99.94%)
N=5	10/9995 **0.10%** (0.05%, 0.18%)	9985/9995 **99.90%** (0.05%, 0.18%)	15/10005 **0.15%** (0.08%, 0.25%)	9990/10005 **99.85%** (99.75%, 99.92%)
N=10	10/9990 **0.10%** (0.05%, 0.18%)	9980/9990 **99.90%** (0.05%, 0.18%)	20/10010 **0.20%** (0.12%, 0.31%)	9990/10010 **99.80%** (99.69%, 99.98%)
N=20	10/9980 **0.10%** (0.05%, 0.18%)	9970/9980 **99.90%** (0.05%, 0.18%)	NA*	NA*

* It is highly unlikely to observe 20 false positives results from 10,000 donor population.

As demonstrated in Table 3, it is important to further test "double positive" samples so that all potential false positive results can be further examined. These additional testing will not be too burdensome to the manufacturers, since there should be only a minimal number of "double positive" samples in donor population. Of course, another alternative is to use an independently developed assay as the comparator, so that the possibility of systematic false positive by both the investigational and comparator assays is minimal. However, this may not be feasible in practice since the investigational assay and its previous version are more readily available to their manufacture.

Another issue we want to discuss is that the confirmed positive donor samples in the clinical specificity study are most often overlooked in the assay sensitivity assessment. As we will discuss in the next section, the clinical sensitivity study for donor screening assays is usually conducted with specimens from subjects known to be infected and then separately a cohort from a high risk population, which are not the real intended population. It is very likely that the specimen from a previously healthy donor who may have just been infected and possibly still in the window period shortly after infection will react with the assay differently than the specimen from a diagnosed patient who may be symptomatic already. The test results among those confirmed positive donor samples may be the only direct measurement of assay sensitivity in its intended population. Of course, we fully understand that there will not be enough positive samples in donor population alone to provide adequate sensitivity estimate with reasonable level of uncertainty. However, we believe those samples should not be overlooked in sensitivity assessment. More specifically, if any false negative results are observed in donor population, one should at least not be able to reject the null hypothesis that the assay sensitivity is at certain acceptable or claimed level (e.g., H_o: sensitivity $\geq$ 99.5% vs. H_a: sensitivity < 99.5%). For example, if an investigational assay gives 1 false negative result

out of 6 confirmed positive donor samples. We can reject the null hypothesis that the assay sensitivity is larger or equal to 99.5% with a p value of 0.03 (i.e., $P(x \geq 1, n = 6 \mid \text{sensitivity} = 99.5\%) = 0.03$). In other words, when the above result is observed in the intended population, we need to be very cautious about a sensitivity claim higher than 99.5% even if the clinical sensitivity study produces such estimate, since the sensitivity is estimated based on results from known positives or high risk population rather than the intended donor population. Pepe [9] refers this potential bias as spectrum bias. In this case, further in-depth investigation is needed. However, it may be impractical because prevalence of the screened diseases is so low in donor population.

3. Clinical Sensitivity Study

The clinical sensitivity study for screening assay is conducted with known infected samples and high risk population, because the intended population has very low prevalence rates for the infectious diseases of interest. Although the sensitivity study is not blinded (i.e., operator may know certain samples are from known infected individuals), the usual observer bias is ignorable here, because donor screening assays are highly automated and the operators play minimal roles in results interpretation.

The clinical sensitivity study usually requires about 1000 positive samples for HIV-1, 500 for HBV or HCV, and 200 for HIV-2 [7]. Each sample is tested by the investigational assay and a comparator assay in parallel. The test algorithm and analysis is very straight forward for known infected samples. For high risk population, the test algorithm is usually as shown in Table 4 when a previously approved version is used as comparator. It is similar to clinical specificity study that "double negative" samples are considered as true negative and discordant results will be followed up by additional testing. The difference is that all samples with positive results from either investigational or comparator assay will be further tested to determine their true status.

Table 4: Test Algorithm for High Risk Population

		Comparator (Previous Version)	
		Nonreactive	**Reactive**
Investigational Assay	**Nonreactive**	True Negative*	Additional Testing
	Reactive	Additional Testing	Additional Testing

** Exclude from sensitivity calculation*

As we explained in section 2, treating "double negative" samples as true negative in clinical specificity study is acceptable. However, it is actually not the case in clinical sensitivity study. By definition, we know sensitivity equals one minus false negative rate (FNR) as indicated in Eq (2). Following similar logic, any false negative results will affect the numerator of FNR and have a bigger impact on the FNR and sensitivity calculation comparing to any false positive results which only affect the denominator.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{Total Number of Positive}} = 1 - \frac{\text{False Negative}}{\text{Total Number of Positive}} \qquad \text{Eq (2)}$$

Table 5 illustrates the impact of these two types of misclassifications on FNR and sensitivity calculation. Assuming a FPR of 0.50% and a sensitivity of 99.50% for the investigational assay, the impact of 2, 5, or 10 false negative results on the sensitivity

estimate is much bigger than the same number of false positive results. Therefore, it is important to further test at least a percentage of the "double negative" samples in clinical sensitivity study, so that all potential false negative results can be further examined [10].

Table 5: Impact of Misclassifications on Sensitivity and False Negative Rate (FNR)

	Missed False Positives		**Missed False Negatives**	
	FNR	**Sensitivity**	**FNR**	**Sensitivity**
N=0	5/1000 **0.50%** (0.16%, 1.16%)	995/1000 **99.5%** (99.84%, 99.84%)	5/1000 **0.50%** (0.16%, 1.16%)	995/1000 **99.5%** (99.84%, 99.84%)
N=2	5/998 **0.50%** (0.16%, 1.17%)	993/998 **99.5%** (99.83%, 99.84%)	7/1002 **0.70%** (0.28%, 1.43%)	995/1002 **99.30%** (98.57%, 99.72%)
N=5	5/995 **0.50%** (0.16%, 1.17%)	990/995 **99.5%** (99.83%, 99.84%)	10/1005 **1.00%** (0.48%, 1.82%)	995/1005 **99.00%** (98.18%, 99.53%)
N=10	5/990 **0.51%** (0.16%, 1.17%)	985/990 **99.5%** (99.83%, 99.84%)	15/1010 **1.49%** (0.83%, 2.44%)	995/1010 **98.51%** (97.56%, 99.17%)

4. Discussion

In this paper, we first provided some background information on regulatory requirements and the development of donor screening assays. The evidence has shown that the implementation of donor screening assays along with donor selection is remarkably successful in reducing transmission of infectious diseases transmitted by blood transfusion [1, 2, 3, 4, 5]. We also discussed some unique features of screening assays. For example, its intended donor population has very low prevalence rates for infectious diseases of interest. Hence, assay sensitivity has to be evaluated separately using known infected samples or high risk population. Also, although the clinical sensitivity study is not blinded, the observer bias is ignorable, because donor screening assays are highly automated and provide non-subjective readouts.

The performances of currently approved donor screening assays are very good. It is this author's view that there are still several statistical issues in the design and analysis of clinical performance studies. Addressing these issues will produce even more accurate clinical performance estimates for donor screening assays. As discussed in sections 2 and 3, the confirmed positive donor samples in clinical specificity study should not be overlooked, especially when any false negative results are observed. Also, in recent years, more and more manufacturers are developing their second or third generation assays and using the previously approved version as the comparator in clinical performance studies. In this case, when the true sample status are determined with only the concordant results from investigational assay and its previous version, it is very likely that systematic false positive and/or false negative results may occur due to the similar technical characteristics of these two assays. Tables 3 and 5 show the potential bias of false positive results in specificity estimate and the potential bias of false negative results in sensitivity estimate. This author believes it is helpful to further test "double positive" samples in clinical specificity study and at least a percentage of the "double negative" samples in clinical sensitivity study, while as the bias of false negative results in clinical specificity study is really minimal even without any further testing.

There are many other issues in assessing the performance of donor screening assays which are not covered in this paper. For example, how to handle indeterminate results including gray zone results and invalid results due to procedure failures. Some blood donor screening systems allow for pools from multiple donors to be tested and we have not discussed how to assess complex test algorithm for reactive pool results. How to assess assay's ability to detect novel strains from abroad that do not align with U.S. strains may have to be tested using laboratory studies rather than evaluating this as part of a prospective study. Similar to the issues discussed before, addressing the above issues requires not only relevant statistical knowledge, but also a clear understanding of the unique features of donor screening assays.

References

[1] Stramer, S.L. and Dodd, R.Y. (2013), "Transfusion-transmitted emerging infectious disease: 30 years of challenges and progress", *Transfusion*, 53, 2375-2383.

[2] Perkins, H.A. and Busch, M.P. (2010), "Transfusion-associated infections: 50 years of relentless challenges and remarkable progress", *Transfusion*, 50, 2080-2099.

[3] Ward, J.W., Grindon, A.J., Feorino, P.M., Schable, C., Parvin, M. and Allen, J.R. (1986), "Laboratory and epidemiologic evaluation of an enzyme immunoassay for antibodies to HTLV-III", *Journal of the American Medical Association*, 256, 357-361.

[4] Zou, S., Stramer, S.L. and Dodd, R.Y. (2012), "Donor testing and risk: current prevalence, incidence, and residual risk of transfusion - transmissible agents in US allogeneic donations", *Transfusion Medicine Reviews*, 26, 119-128.

[5] Klamroth, R., Gröner, A. and Simon, T.L. (2014), "Pathogen inactivation and removal methods for plasma-derived clotting factor concentrates", *Transfusion*, 54, 1406-1417.

[6] Complete List of Donor Screening Assays for Infectious Agents and HIV Diagnostic Assays.*http://www.fda.gov/BiologicsBloodVaccines/BloodBloodProducts/ApprovedProducts/LicensedProductsBLAs/BloodDonorScreening/InfectiousDisease/UCM080466*

[7] FDA CBER (1999), "Guidance for Industry: In the Manufacture and Clinical Evaluation of In Vitro Tests to Detect Nucleic Acid Sequences of Human Immunodeficiency Viruses Types 1 and 2".

[8] FDA (1989), "Draft Points to Consider in the Manufacture a Clinical Evaluation of In Vitro Test to Detect Antibodies to Human Immunodeficiency Virus Type 1".

[9] Pepe, M.S. (2003), *The Statistical Evaluation of Medical Tests for Classification Prediction*, Oxford, UK: Oxford University Press.

[10] Kondratovich, M.V. (2008), "Comparing two medical tests when results of reference standard are unavailable for those negative via both tests", *Journal of Biopharmaceutical Statistics*, 18(1), 145-166.