



ORIGINAL ARTICLE

Bridging Strategies for *In Vitro* Diagnostic Clinical Trials in a New Region

Robert Magari, PhD*, Mohamad Hasan, PhD and Karen Lo

Beckman Coulter, Inc, Miami, Florida, USA

*Corresponding author: Robert Magari, PhD, Beckman Coulter, Inc, 11800 SW 147 Ave., Miami, Florida 33196-2500, USA, Tel: 954-329-7214



Abstract

The need for multiregional clinical trials for *in vitro* diagnostic devices and extrapolating results in a new region is increasing in recent years. In this paper we are providing strategies for consideration of bridging studies for the two most common types of clinical trials for *in vitro* diagnostic, method comparison and diagnostic clinical performance trials. Bridging studies should be considered when the expected agreement in the new region is inferior to the agreement in the original trial. In method comparison, a bridging study in the new region is not needed, if bias in this region is equal or smaller than the estimated bias of the original trial. Similarly, no bridging study is necessary for diagnostic clinical performance trials when sensitivity in the new region is equal to or greater than the sensitivity of the original trial. The reproducibility of the original trial has a significant effect on the sample size of the bridging study. Sample size decreases with the increase of reproducibility. We are using the ratio of reproducibility and the ratio of differences from the specification of the new region to the original trial to decide about the clinical trial in the new region.

Introduction

Multiregional clinical trials have been used in the last decade or so to evaluate therapeutic products as part of developing global medicines. Approaches applied to these trials are summarized in the International Conference on Harmonization guideline ICH E5 [1], while several papers addressing bridging studies for therapeutic endpoints have laid the foundation for statistical procedures [2-4]. Recently, market requirements and regulatory landscape are changing for *In Vitro* Diagnostic (IVD) devices as well. Topics of multiregional clinical trials and bridging strategies were vastly discussed during the International Medical Device Regulatory Forum of China

last year [5]. This forum organized by Chinese NMPA and aiming to address policies and regulations on medical devices, evidenced the increasing demand for multiregional clinical trials and bridging strategies to evaluate diagnostic products in new regions.

Accuracy performance is a way to evaluate IVD devices in clinical testing. ISO 5721-1 defines accuracy as the closeness of agreement between a test result and an accepted reference value [6]. The choice of reference depends on the diagnostic test and the type of device being tested. A newly discovered test is commonly evaluated for its accuracy to the clinical diagnosis. Clinical trials where diagnosis is used as a reference method are called 'diagnostic clinical performance' trials and accuracy is commonly evaluated in terms of sensitivity and specificity [7,8]. For other IVD tests, a marketed device can serve as a predicate for the candidate device under test and agreement with the comparative method is evaluated based on the closeness of the agreement of the measurements. These types of clinical trials are referred to as 'method comparison'. Bias between devices is an evaluator of agreement that can be estimated statistically by using a regression approach [9].

Shao & Chow [10] and Chow, et al. [3] introduced the concept of reproducibility to assess the ability of the results from a clinical trial to be extrapolated to a new region. Quan, et al. [11] and Ikeda & Bretz [12] provided approaches for sample size calculation for bridging study in a new region using the concept of consistency defined earlier by Shih [2]. Lan, et al. [13] used weighted z-statistics to design and size bridging studies while Liu, et al. [14] proposed a Bayesian approach using the prior



Citation: Magari R, Hasan M, Lo K (2020) Bridging Strategies for *In Vitro* Diagnostic Clinical Trials in a New Region. Int J Clin Biostat Biom 6:028. doi.org/10.23937/2469-5831/1510028

Accepted: July 08, 2020; **Published:** July 10, 2020

Copyright: © 2020 Magari R, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

information to assess the similarity between the original clinical trial and bridging studies in the new region. All these methodologies were developed for therapeutic endpoints, i.e., comparing test results to a placebo, survival endpoints, etc. We are not aware of any publication addressing bridging methodologies to support the introduction of IVD products into new geographies or region.

Consequently, the purpose of this paper is to provide a strategy for bridging the results of an original clinical trial into a new region for IVD devices. We will consider the two most common types of clinical trials for IVD, method comparison and diagnostic clinical performance trials.

Method Comparison Trials

This section is only for method comparison studies for quantitative measurement procedures. Let X be the measurement with the predicate device and Y the measurement with the test device. Both devices are tested in N subjects, with X_i and Y_i being the measurement of both predicate and test devices, respectively, on the same i subject.

The linear function (1) fits the relationship between devices.

$$Y = b_0 + b_1 X + \varepsilon \quad (1)$$

Where b_0 and b_1 are intercept and slope, and ε is the error.

Since both test and predicate devices measure with a certain degree of imprecision, Linnet [15] proposed an orthogonal least square approach to estimate parameters of (1). This approach is usually referred to in the IVD literature as Deming approach [9]. Non-parametrical estimation of the parameters is also commonly used in the IVD industry [9,16].

Bias (B) between two methods at a certain medically important level X_c can be calculated as:

$$B = b_0 + (b_1 - 1) X_c \quad (2)$$

The standard error of B can be calculated based on ordinary least square variances and covariance of the slope and intercept. Linnet [17] argued that for a relatively large analytical measuring interval (AMI), slope close to 1 and equal precision of both devices, orthogonal least square estimate of $\text{var}(B)$ can be approximated by its ordinary least square estimate. Considering variances-covariance of the regression estimates, $\text{se}(B)$ can be approximated as:

$$\text{se}(B) \approx 2 \sqrt{\frac{\sigma^2}{N} \left(1 + \frac{(X_c - \bar{X})^2}{\text{var}(X)}\right)} \quad (3)$$

Where σ is the precision and \bar{X} is the average of the predicate device. $\text{se}(B)$ is proportional to precision, AMI, as well as the distance of X_c from the mean. The function under the square root is multiplied by 2 since both

devices are measured with error. Knowledge about the precision of the device, AMI of the analyte, and a pre-specified clinical tolerable bias (B_s) can be used to prospectively calculate power and size the clinical trial.

Hypotheses for equivalence between the predicate and test device in method comparison are:

$$H_0: B \leq -B_s \text{ or } B \geq B_s$$

$$H_1: -B_s < B < B_s$$

Prespecified bias, B_s is usually symmetrical around zero. The absolute value of the bias represents the clinical relevance for an analyte while the sign indicates the direction of the bias. However, in specific comparisons, bias might be asymmetrical.

Let P be the prospective power to size the original clinical trial. Let \hat{B} and $\text{se}(\hat{B})$ be the estimated bias and standard error from the original clinical trial, and t_c a statistic that follows a t-distribution with $(N-2)$ degree of freedom.

$$t_c = \frac{\hat{B} - B_s}{\text{se}(\hat{B})}$$

Let \hat{P} be the post experimental power of the original trial. We can use Owen's function as [18]:

$$Q = \frac{\sqrt{2\pi}}{\Gamma\left(\frac{N-2}{2}\right) 2^{\frac{n-4}{2}}} \int_0^u \Phi\left(\frac{tx}{\sqrt{N-2}} - t_c\right) x^{n-3} \varphi(x) dx$$

Where t is from t-distribution with $1-\alpha$ confidence and $n-2$ degrees of freedom, $\Phi(\cdot)$ is the cumulative distribution function and $\varphi(\cdot)$ is the density function of the standard normal distribution, while,

$$u = \frac{(N-2)^{0.5} 2B_s}{2\text{se}(\hat{B})t}$$

Power can be calculated as the difference:

$$\hat{P} = Q\left(-t, \frac{\hat{B} - B_s}{\text{se}(\hat{B})}, 0, u\right) - Q\left(t, \frac{\hat{B} + B_s}{\text{se}(\hat{B})}, 0, u\right) \quad (4)$$

Shao and Chow [10] defined post experimental power as reproducibility probability of the original clinical trial (\hat{P}).

When devices are tested on a different set of subjects in the new region, the estimate of bias might be different from the original trial due to population differences. Let \hat{B}_{new} denote the estimate of bias in the new region and $\text{se}(\hat{B}_{new})$ denotes the estimate of standard error. The values of these estimates can be based on knowledge, if available, or several different values can be used in the calculation to perform a sensitivity analysis. Difference between the estimates in the original clinical trial and the new region clinical trial is defined by Chow, et al. [3] as the similarity index (Δ). In case of method comparison, $\text{se}(\hat{B}) = \text{se}(\hat{B}_{new})$ since precision and AMI are characteristic of the device they do not de-

pend on the population. Consequently, Δ is only a function of the difference in bias.

Let $\widehat{B}_{new} = \widehat{B} + \Delta$ and $t_{c(new)} = \frac{\widehat{B}_{new} - B_s}{se(\widehat{B})}$. Equation (4) can be used to calculate the reproducibility probability of the trial in the new region, \widehat{P}_{new} . Results of the original region can be extrapolated to the new region when $\frac{\widehat{P}_{new}}{\widehat{P}} \leq 0.9$. When $\frac{\widehat{P}_{new}}{\widehat{P}} > 0.9$, a bridging study in the new region is not required [3].

Let $D = B_s - \widehat{B}$ and $D_{new} = B_s - \widehat{B}_{new}$. Consistency between the results of the original trial and the expected results in the new region can be expressed as [12]:

$$p\left(\frac{D_{new}}{D} > \rho\right) = \gamma \tag{5}$$

A consistency probability of $\gamma \geq 0.8$ and the values of $\rho \geq 0.5$ are generally acceptable. When D and D_{new} are normally distributed, Ikeda & Bretz [12] showed that equation (5) can be approximated by a normal cumulative density function and the proportion of subjects in the new region can be calculated as:

$$p = \frac{z_{1-\gamma}^2}{(1-\rho)^2(z_{1-\alpha} + z_{1-\beta})^2 - z_{1-\gamma}^2\rho(\rho-2)} \tag{6}$$

where $\beta = 1 - \widehat{P}$, and $\alpha = 0.025$ for 95% confidence. Since the sensitivity of using different values of \widehat{B}_{new} is evaluated, both \widehat{P}_{new} and \mathbf{p} represent vectors of values. The number of subjects in the new region is:

$$N_{new} = N \min\{p \mid \frac{\widehat{P}_{new}}{\widehat{P}} \leq 0.9 \text{ and } \rho \geq 0.5\}$$

Table 1: Bias at two different medically important levels.

X_c	\widehat{B}	$se(\widehat{B})$	Lower CI	Upper CI	B_s
150	-1.86	2.20	-6.20	2.48	15
450	4.14	5.01	-5.74	14.02	45

Table 2: Decision for a method comparison trial in the new region ($X_c = 150$ cells/ $\mu\text{L} \times 10^3$).

Δ	$\frac{\widehat{P}_{new}}{\widehat{P}}$	ρ	Decision New Region	Proportion of Subjects for Bridging Study	Sample Size for Bridging Study
1	1.000	0.92	No Clinical Trial	.	.
2	0.999	0.85	No Clinical Trial	.	.
3	0.996	0.77	No Clinical Trial	.	.
4	0.984	0.70	No Clinical Trial	.	.
5	0.955	0.62	No Clinical Trial	.	.
6	0.894	0.54	Bridging Study	0.09	18
7	0.783	0.47	New Clinical Trial	.	.
8	0.637	0.39	New Clinical Trial	.	.
9	0.460	0.32	New Clinical Trial	.	.
10	0.292	0.24	New Clinical Trial	.	.
11	0.159	0.16	New Clinical Trial	.	.
12	0.074	0.09	New Clinical Trial	.	.

Example

A clinical trial to comply with the pre-market regulatory requirement for testing Beckman Coulter DxH 520 hematology analyzer was conducted at different sites in the United States. DxH 520 was the test device while UniCel DxH 800 Coulter Cellular Analysis System was the predicate device. Hematology analyzers measure several hematology parameters at the same time. In the context of this example, we will focus on only one parameter, platelet (PLT) which is a cell type related to blood coagulation. PLT expressed as cell counts/ $\mu\text{L} \times 10^3$ was measured by both devices on $N = 196$ subjects. Because of a large number of counts for each measurement, both X_i and Y_i are assumed to be normally distributed, with means μ_x and μ_y and the same precision, σ .

Method comparison was used to evaluate the agreement and establish equivalence between devices. PLT counts in humans exhibit relatively wide range depending on clinical conditions. This can affect the variability of measurements throughout the AMI. In addition, the relative variability (CV%) of precision in the low range is much larger in comparison to the rest of AMI. Consequently, there is a pattern of dependency of variability within the AMI that required weighted adjustments for the estimation of regression parameters.

Weighted Deming approach [15] was used to estimate regression parameters. The estimated slope was 1.02 (95% CI; 0.97, 1.06) and intercept was -4.86 (95% CI; -15.21, 5.49). There are two medically important levels for PLT; $X_c = 150$ cell/ $\mu\text{L} \times 10^3$ and $X_c = 450$ cell/ $\mu\text{L} \times 10^3$. Biases at these levels were calculated using equation (2) while a jackknife resampling was used to estimate $se(\widehat{B})$. The two-sided confidence intervals of \widehat{B} were calculated as ($\alpha = 0.05$ and $N = 196$):

$$\widehat{B} \pm t_{\frac{\alpha}{2}, N-2} se(\widehat{B})$$

A 10% bias is usually tolerable for PLT. Consequently, $B_s = 15 \text{ cells}/\mu\text{L} \times 10^3$ at $X_c = 150 \text{ cells}/\mu\text{L} \times 10^3$ and $B_s = 45 \text{ cells}/\mu\text{L} \times 10^3$ at $X_c = 450 \text{ cells}/\mu\text{L} \times 10^3$. Results are provided in Table 1.

Using this information and equation (4), $\hat{P} = 0.999$. The value of Δ in the new region is unknown. We calculated \hat{P}_{new} for different values of Δ in the interval $\hat{B} \leq \Delta \leq B_s$. The respective values of $\rho = \frac{D_{new}}{D}$ were also calculated for the same interval. The summary of the results for $X_c = 150$ are shown in Table 2. No clinical trial is necessary in the new region when $\frac{\hat{P}_{new}}{\hat{P}} > 0.9$, bridging study is recommended when $\frac{\hat{P}_{new}}{\hat{P}} \leq 0.9$ and $\rho \geq 0.5$, while a fully powered new clinical trial is needed when $\rho < 0.5$. A different set of Δ can be used if there are no solutions for one of the above conditions. Based on PLT results, a bridging study is recommended only when $\Delta = 6 \text{ cells}/\mu\text{L} \times 10^3$ that correspond to a sample size of 18 (Table 2).

Decision for a bridging study corresponds to $\Delta = 6$ for $X_c = 150$ (Table 2) and the fraction of the difference from the prespecified clinical tolerable bias is $\frac{(B_s - \Delta)}{B_s} = 0.6$. This fraction corresponds to $\Delta = 18$ for $X_c = 450$. Calculation for \hat{P} and ρ and decision are shown in Table 3 starting from $\Delta = 17$. All decision results for a reasonable interval indicate that no clinical trial is need in the new region.

Based on the results from the two medically important levels, is it evident that bias at $X_c = 150$ level represents the worst-case scenario and drives the decision on the clinical trial. The relative value of this bias is 1.24% compare to 0.92% at $X_c = 450$.

Diagnostic Clinical Performance Trials

Let C be a binary variable denoting the true clinical status for a disease. $C = 1$ when the disease is present and $C = 0$ when the disease is absent. Let X be the results of an IVD test, where $X = 1$ indicates the presence of the disease while $X = 0$ indicates the absence of the disease. X can have a binary outcome or can be measured on a continuous scale. When measured on a continuous scale, X can be transformed into binary by using a cut-off point that separates the values of X in two categories, disease present or absent.

Table 3: Decision for a method comparison trial in the new region ($X_c = 450 \text{ cells}/\mu\text{L} \times 10^3$).

$\frac{(B_s - \Delta)}{B_s}$	Δ	$\frac{\hat{P}_{new}}{\hat{P}}$	ρ	Decision New Region
0.62	17	0.997	0.58	No Clinical Trial
0.6	18	0.995	0.56	No Clinical Trial
0.58	19	0.992	0.54	No Clinical Trial
0.56	20	0.986	0.51	No Clinical Trial
0.53	21	0.977	0.49	No Clinical Trial

Comparison of a test result to clinical truth can have four outcomes; True positive = ($C = 1, X = 1$), False negative = ($C = 1, X = 0$), False positive = ($C = 0, X = 1$) and True negative = ($C = 0, X = 0$). Let N be the total number of subjects in the trial.

$$N = N_{TP} + N_{FN} + N_{FP} + N_{TN}$$

where N_{TP} , N_{FN} , N_{FP} , and N_{TN} are the numbers of true positive, false negative, false positive and true negative respectively. Based on these outcomes, several conditional probabilities can be calculated to evaluate the diagnostic ability of the test [19]. The most common are sensitivity, specificity, positive/negative predicted values, positive/negative likelihood ratios, etc. In this section, we will focus on sensitivity, which is the proportion of subjects with disease that are tested positive. The same approach can be used for specificity and predicted values.

$$S = p(X = 1 | C = 1)$$

Let S be a target value for sensitivity and S_s a lower limit of sensitivity that is clinically acceptable. Prospective power (P) can be directly calculated from the cumulative normal distribution function as:

$$P = \phi\left(Z_\alpha - \sqrt{N_{Disease}} \frac{S - S_s}{\sqrt{S(1-S)}}\right)$$

While the required number of diseased subjects for the original clinical trial is:

$$N_{Disease} = S(1 - S) \left(\frac{z_p + z_{1-\alpha}}{S - S_s}\right)^2 \quad (8)$$

Let \hat{S} be the estimated sensitivity from the original clinical trial.

$$\hat{S} = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

$se(\hat{S})$ is the standard error of a binomial proportion, while normal approximation can be used to calculate confidence intervals. Post experimental power is:

$$\hat{P} = \phi\left(z_\alpha - \sqrt{N_{TP} + N_{FN}} \frac{\hat{S} - S_s}{\sqrt{\hat{S}(1-\hat{S})}}\right) \quad (9)$$

The similarity index is only a function of differences in sensitivity between the original clinical trial and the expected estimates in the new region, $\Delta = S - S_{New}$, while \hat{P}_{new} can be calculated based on the power function (9), S and Δ . When $D = S - S_s$ and $D_{new} = S_{new} - S_s$, consistency between the results of the original trial and new region and proportion of subjects in the new region can be calculated based on equations (5) and (6).

Example

Data from the sepsis clinical trial presented in Crouser, et al. [20] were used for demonstration purposes. Clinical trial conducted at three different sites in the US, enrolled $N_{Disease} = 385$ sepsis positive subjects out of a

Table 4: Decision for a diagnostic clinical performance trial in the new region (sensitivity).

Δ	$\frac{\hat{P}_{new}}{\hat{P}}$	ρ	Decision New Region	Proportion of Subjects for Bridging Study	Sample Size for Bridging Study
0	1.00	1.00	No Clinical Trial	.	.
0.01	0.95	0.89	No Clinical Trial	.	.
0.02	0.86	0.78	Bridging Study	0.50	78
0.03	0.72	0.67	Bridging Study	0.31	
0.04	0.56	0.56	Bridging Study	0.20	
0.05	0.38	0.44	New Clinical Trial	.	.
0.06	0.24	0.33	New Clinical Trial	.	.
0.07	0.13	0.22	New Clinical Trial	.	.
0.08	0.06	0.11	New Clinical Trial	.	.

Table 5: Decision for a diagnostic clinical performance trial in the new region (specificity).

Δ	$\frac{\hat{P}_{new}}{\hat{P}}$	ρ	Decision New Region	Proportion of Subjects for bridging study	Sample size for bridging study
0.031	0.92	0.55	No Clinical Trial	.	.
0.032	0.91	0.54	No Clinical Trial	.	.
0.033	0.89	0.52	Bridging Study	0.073	122
0.034	0.88	0.51	Bridging Study	0.069	
0.035	0.86	0.49	New Clinical Trial	.	.
0.036	0.84	0.48	New Clinical Trial	.	.

total of $N = 2158$ subjects. Monocyte distribution width (MDW) was measured and evaluated for the ability to diagnose sepsis. The MDW was measured on a continuous scale and the use of a cut-off ($X_c = 20$) separated the values into $X = 1$ when $MDW > 20$ and $X = 0$ when $MDW \leq 20$ [20].

Sensitivity; $\hat{S} = 0.74$ (95% CI; 0.69, 0.78) based on $N_{TP} = 285$ and $N_{FN} = 100$. The lowest clinically acceptable value for sensitivity was $S_s = 0.65$. Based on this information reproducibility of the original trial was $\hat{P} = 0.97$. Since the value of Δ in the new region is unknown we choose a series of values in the interval $S_s \leq \Delta \leq \hat{S}$. \hat{P}_{new} and ρ are also calculated for the same interval of Δ values. The results are shown in Table 4. At least 78 sepsis positive samples are needed in the new region.

Specificity; $\hat{S} = 0.719$ (95% CI; 0.69, 0.74) based on $N_{TN} = 1275$ and $N_{FP} = 498$. The lowest clinically acceptable value was $S_s = 0.65$. Results are shown in Table 5. At least 122 negative samples are needed in the new region.

Reproducibility of the Original Trial

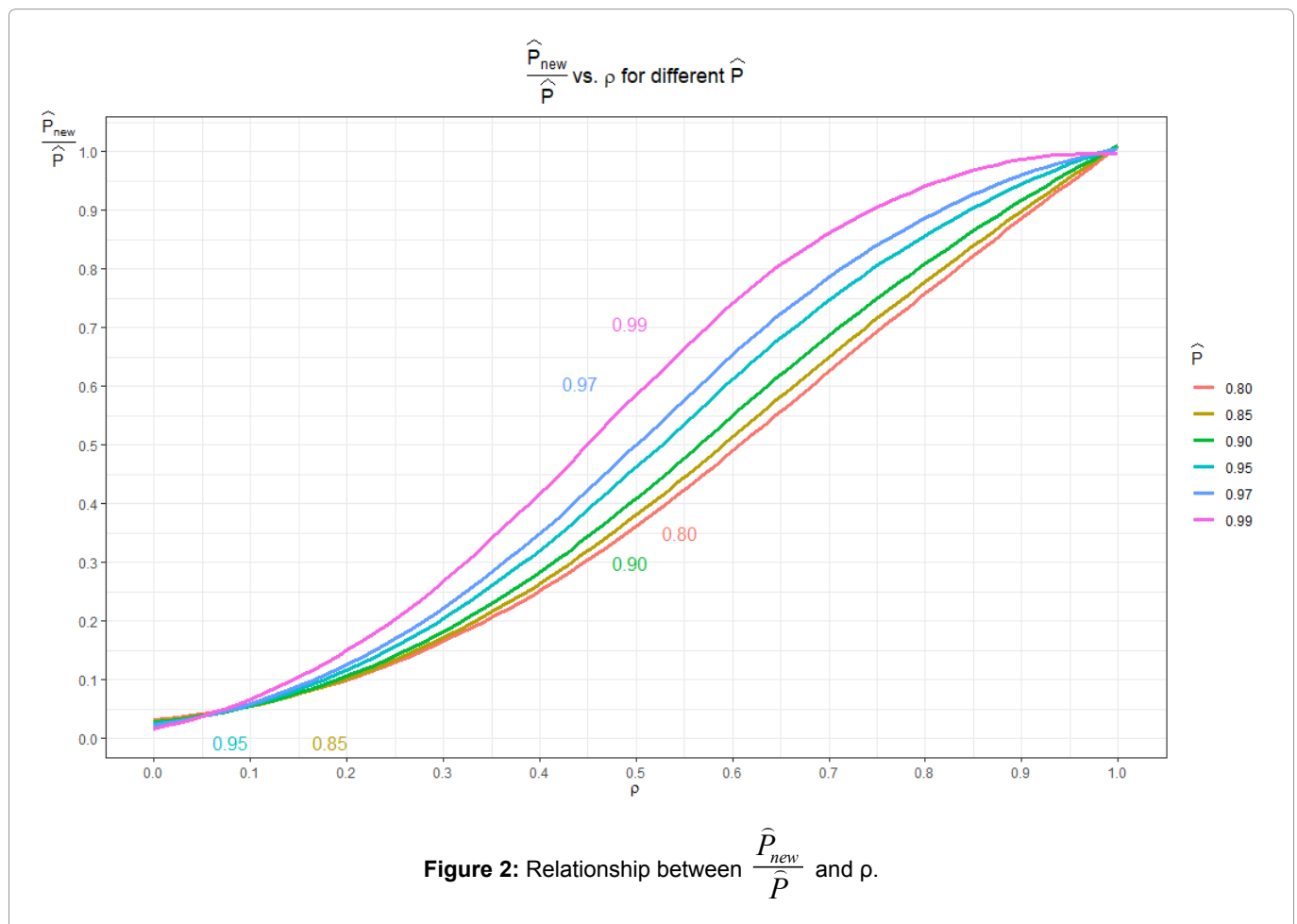
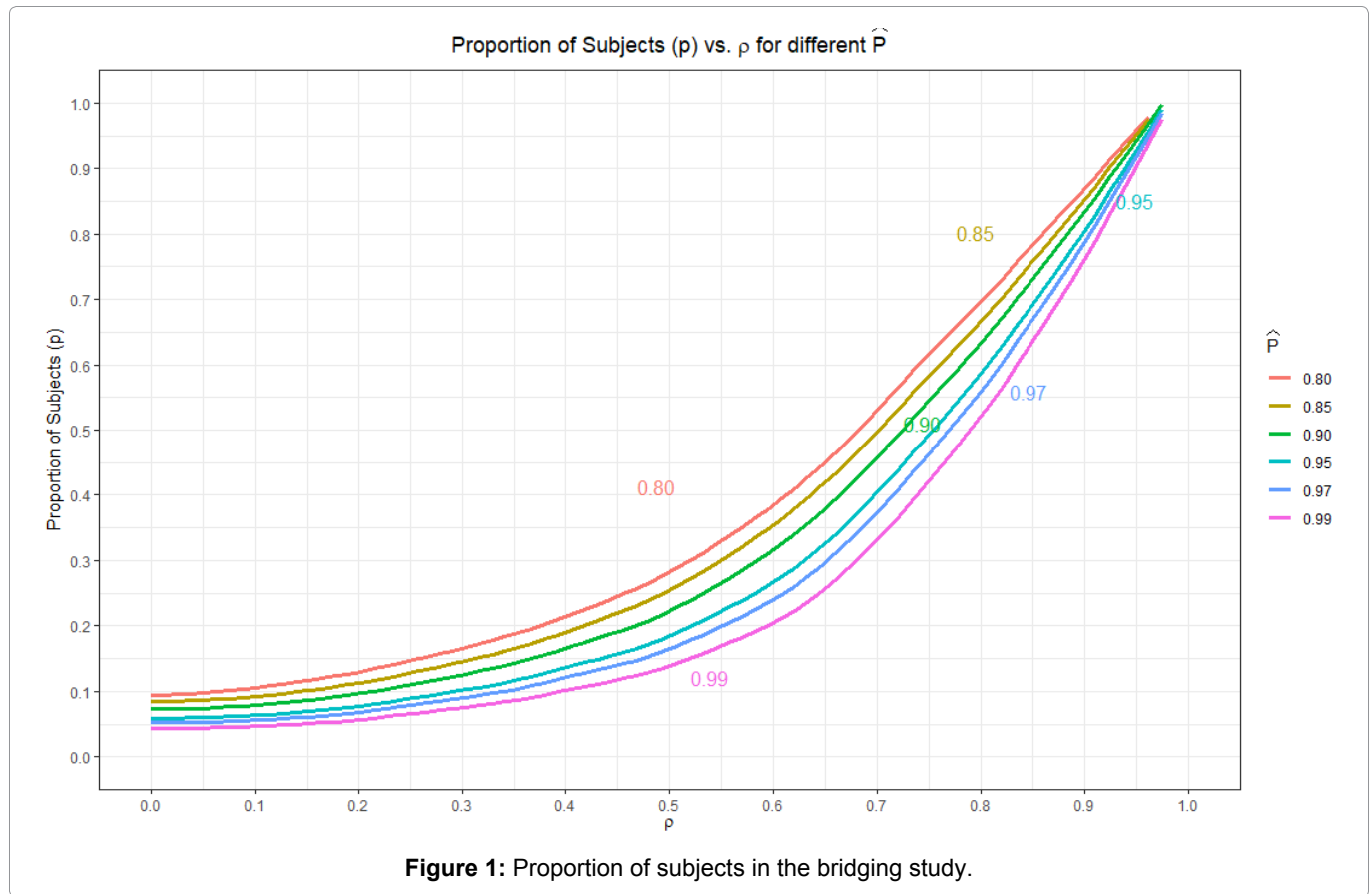
The performance in the original trial expressed in terms of reproducibility has a significant effect on sample size of bridging study. To demonstrate this, we simulated data calculating sample size for bridging study by varying \hat{P} and ρ . The proportion of the sample size was calculated for the significance level of $\alpha = 0.05$ and con-

sistency of $\gamma = 0.80$. Results in Figure 1 and Table 6 show the proportion of subjects (ρ) required for the different threshold of $\rho = \frac{D_{new}}{D}$ when \hat{P} in the original region varies. Each curve in Figure 1 corresponds to a specific reproducibility value. Values of $\hat{P} \geq 0.8$ are selected for these simulations because multicenter trials for IVD are usually prospectively powered at $P \geq 0.8$ level. The proportion of the subjects increases with the increase of ρ , however, decreases with the increase of the reproducibility of the original trial.

The relationship between $\frac{\hat{P}_{new}}{\hat{P}}$ and ρ is shown in Figure 2. Values of $\rho < 0.5$ are not possible when $\frac{\hat{P}_{new}}{\hat{P}} > 0.9$ substantiating the decision of no need for bridging study when only $\frac{\hat{P}_{new}}{\hat{P}} > 0.9$. The value of ρ that corresponds to $\frac{\hat{P}_{new}}{\hat{P}} = 0.9$ is approximately 0.75 and bridging is necessary when $0.5 \leq \rho \leq 0.75$. From Figure 1, we also see the most separation of the sample sizes for different \hat{P} occur in the $0.5 \leq \rho \leq 0.75$ interval. Consequently, any bridging study decision and subsequent sample size calculations should be in the region $\frac{\hat{P}_{new}}{\hat{P}} \leq 0.9$ and $\rho \geq 0.5$.

Discussion

There are three available options when determining whether to conduct a clinical trial in the new region;



extrapolate the results of the original trial to the new region (no need for a new clinical study), perform a bridging study, and conduct a new fully powered clinical trial. The decision about these choices is contingent on

Table 6: Simulated proportion of subjects.

$\hat{P} = 0.80$			$\hat{P} = 0.90$			$\hat{P} = 0.99$		
$\frac{\hat{P}_{new}}{\hat{P}}$	ρ	Proportion of Subjects (\hat{p})	$\frac{\hat{P}_{new}}{\hat{P}}$	ρ	Proportion of Subjects (\hat{p})	$\frac{\hat{P}_{new}}{\hat{P}}$	ρ	Proportion of Subjects (\hat{p})
0.03	0.00	0.09	0.03	0.00	0.07	0.03	0.00	0.04
0.04	0.05	0.10	0.04	0.05	0.07	0.04	0.05	0.05
0.06	0.10	0.11	0.06	0.10	0.08	0.06	0.10	0.05
0.08	0.15	0.12	0.08	0.15	0.09	0.09	0.15	0.06
0.10	0.20	0.13	0.11	0.20	0.10	0.13	0.20	0.06
0.13	0.25	0.15	0.14	0.25	0.11	0.18	0.25	0.07
0.16	0.30	0.17	0.18	0.30	0.13	0.24	0.30	0.08
0.21	0.35	0.19	0.23	0.35	0.15	0.31	0.35	0.09
0.25	0.40	0.21	0.28	0.40	0.17	0.39	0.40	0.11
0.30	0.45	0.25	0.34	0.45	0.19	0.47	0.45	0.12
0.36	0.50	0.28	0.41	0.50	0.23	0.55	0.50	0.15
0.42	0.55	0.33	0.48	0.55	0.26	0.63	0.55	0.18
0.49	0.60	0.38	0.55	0.60	0.31	0.71	0.60	0.21
0.56	0.65	0.45	0.62	0.65	0.37	0.78	0.65	0.26
0.63	0.70	0.52	0.69	0.70	0.45	0.84	0.70	0.32
0.70	0.75	0.61	0.76	0.75	0.54	0.89	0.75	0.41
0.76	0.80	0.71	0.82	0.80	0.65	0.93	0.80	0.52
0.83	0.85	0.81	0.87	0.85	0.76	0.95	0.85	0.66
0.89	0.90	0.91	0.92	0.90	0.88	0.98	0.90	0.81
0.95	0.95	0.98	0.96	0.95	0.97	0.99	0.95	0.95
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

the difference in agreement between the original trial and the expected agreement in the new region. We also showed that the reproducibility of the original trial significantly affects sample size of the bridging study.

The reproducibility of the original trial is related to the prospectively calculated sample size and estimated agreement performance in the trial. Sample size should be based on at least 80% power (90% is preferred for method comparisons) while estimated performance depends on the product being tested. Prospectively powering a method comparison trial separately for slope and intercept has been discussed by Linnet [17] and Passing & Bablok [21]. We are showing an approach to approximate standard error of bias based on variance-covariance structure of slope and intercept and use the knowledge about bias acceptance limits, precision, and measuring interval to calculate sample size for a predetermined power and confidence level. Other approaches for evaluating agreement in method comparison clinical trials are not discussed in this paper [22,23]. Powering of diagnostic clinical performance trials is based on normal approximation of a binomial proportion. Approaches like the exact binomial calculations [24], powering both sensitivity and specificity Pepe [19],

and others are not considered here but might be subject to future publications.

Bridging studies should be considered when the expected agreement in the new region is inferior to the agreement in the original trial. In method comparison, a bridging study in the new region is not needed, if bias in this region is equal or smaller than the estimated bias of the original trial. Similarly, no bridging study is necessary for diagnostic clinical performance trials when sensitivity in the new region is equal to or greater than the sensitivity of the original trial.

We are using the ratio of reproducibility and the ratio of differences from the specification (ρ) of the new region to the original trial to decide about the clinical trial in the new region. Shao and Chow [10] recommended a 90% level of reproducibility ratio as a cutoff level for considering a bridging study in a new region. Values of $\rho \geq 0.5$ are based on the Japanese Ministry of Health, Labor, and Welfare guidance and are used in several publications [11,12,25]. Both these values are related to therapeutic drug testing and authors recommend modifying them according to the specific product being tested.

There is no historical evidence or literature that we know about using these values to design bridging studies for IVD. We are recommending the following steps:

Estimate reproducibility of the original trial (\hat{P})

Simulate a range of expected differences in the new region (Δ)

Estimate reproducibility for each of these differences in the new region (\hat{P}_{new})

Calculate the ratio of reproducibility in the new region versus the original trial ($\frac{\hat{P}_{new}}{\hat{P}}$)

Calculate $\rho = \frac{D_{new}}{D}$

Guide to inform decision:

No need for bridging study when $\frac{\hat{P}_{new}}{\hat{P}} > 0.9$

Consider bridging study when $\frac{\hat{P}_{new}}{\hat{P}} \leq 0.9$ and $\rho \geq 0.5$

New clinical trial when $\rho < 0.5$

Sometimes there might be no solutions for the combination of values above. In these cases, a new set of expected differences in step 2 will need to be simulated. We also caution that the values of $\rho > 0.6$ could require the bridging study to be too large and not practical. Ko, et al. [25] also argued that larger values of ρ would indicate that the overall results are dominated by the new region and consistency of the results between the original trial and the new region might not be valid. For these reasons, we recommend that the sample size for bridging study be based on the minimum proportion in the interval of, $\frac{\hat{P}_{new}}{\hat{P}} \leq 0.9$ and $\rho \geq 0.5$.

Bias in method comparison studies might be calculated at different medically important levels. We recommend identifying the highest relative bias to the prespecified clinical tolerable bias at a medical level and make decision for bridging studies for that bias. In addition, some IVD devices provide measurements for multiple analytes. In these cases, we recommend that calculations be performed for each analyte separately and decision be made based on the worst performing analyte. Similarly, both sensitivity and specificity need to be considered in diagnostic clinical performance trials.

Sample size for bridging study represents the minimum sample size to satisfy the pre-determined requirements and conditions. However, if necessary, more subjects can be tested to uniformly cover the AMI.

Like other authors in the therapeutic domain, we recommend that the choice of $\frac{\hat{P}_{new}}{\hat{P}}$ and ρ to be based on the clinical relevance of IVD under test as well as regulatory requirements in the new region. In addition, recommendations for bridging studies in this paper are

based on statistical considerations. Other factors may influence the final decision.

References

1. International Conference on Harmonization (1998) Tripartite Guidance E5, Ethnic factors in the acceptability of foreign data. Federal Register 83: 31790-31796.
2. Shih WJ (2001) Clinical trials for drug registration in Asian-Pacific countries: Proposal for a new paradigm from a statistical perspective. *Control Clin Trials* 22: 357-366.
3. Chow SC, Shao J, Hu OYP (2002) Assessing sensitivity and similarity in bridging studies. *J Biopharm Stat* 12: 385-400.
4. Chow SC, Hsiao CF (2010) Bridging diversity: Extrapolating foreign data to a new region. *Pharmaceutical Medicine* 24: 349-362.
5. CIMDR (2019) 10th China International Medical Device Regulatory Forum. Suzhou, China.
6. ISO (1994) Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions. International Organization for Standardization (ISO) 5725-1, Geneva, Switzerland.
7. Altman DG, Bland JM (1994) Diagnostic tests. 1: Sensitivity and specificity. *BMJ* 308: 1552.
8. Garrett PE, Lasky FD, Meier KL (2008) User protocol for evaluation of qualitative test performance. (2nd edn), Clinical and Laboratory Standards Institute (CLSI), Wayne, PA, USA.
9. Budd JR, Durham AP, Gwise TE, Hawkins DM, Holland M, et al. (2018) Measurement procedure comparison and bias estimation using patient samples. (3rd edn), Clinical and Laboratory Standards Institute (CLSI), Wayne, PA, USA.
10. Shao J, Chow SC (2002) Reproducibility probability in clinical trials. *Statistics in Medicine* 21: 1727-1742.
11. Quan H, Zhao PL, Zhang J, Rosessner M, Aizawa K (2010) Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. *Pharmaceutical Statistics* 9: 100-112.
12. Ikeda K, Bretz F (2010) Sample size and proportion of Japanese patients in multi-regional trials. *Pharm Stat* 9: 207-216.
13. Lan GKK, Soo Y, Siu C, Wang M (2005) The use of weighted Z-tests in medical research. *J Biopharm Stat* 15: 625-639.
14. Liu JP, Hsiao CF, Hsueh HM (2002) Bayesian approach to evaluation of bridging studies. *J Biopharm Stat* 12: 401-408.
15. Linnet K (1990) Estimation of the linear relationship between the measurements of two methods with proportional errors. *Stat Med* 9: 1463-1473.
16. Passing H, Bablok (1983) A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Chemistry, Part I. *J Clin Chem Clin Biochem* 21: 709-720.
17. Linnet K (1999) Necessary sample size for method comparison studies based on regression analysis. *Clin Chem* 45: 882-894.
18. Owen DB (1965) A special case of a bivariate non-central t-distribution. *Biometrika* 52: 437-446.

19. Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York, NY.
20. Crouser ED, Parrillo JE, Seymour CW, Angus DC, Bicking K, et al. (2019) Monocyte distribution width: A novel indicator of Sepsis-2 and Sepsis-3 in high-risk emergency department patients. *Crit Care Med* 47: 1018-1025.
21. Passing H, Bablok W (1984) Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in *Clinical Chemistry, Part II. J Clin Chem Clin Biochem* 22: 431-445.
22. Magari RT (2004) Bias estimation in method comparison studies. *J Biopharm Stat* 14: 881-892.
23. Lu MJ, Zhong WH, Liu YX, Miao HZ, Li YC, et al. (2016) Sample size for assessing agreement between two methods of measurement by Bland-Altman method. *Int J Biostat* 12.
24. Johnson NL, Kotz S, Kemp AW (1992) *Univariate discrete distributions*. (2nd edn), John Wiley & Sons, 565.
25. Ko FS, Tsou HH, Liu JP, Hsiao CF (2010) Sample size determination for a specific region in a multiregional trial. *Journal of Biopharmaceutical Statistics* 20: 870-885.