# Simulation-Based Evaluation of P-Value Quality in Phase 3 Clinical Trials

**Jihao Zhou[1], Brandon Wales[2], Ray Zhu[3]**

1 Allergan Pharmaceuticals, Irvine, CA 92612, USA
2 University of California, Riverside, CA 92507, USA
3 Allergan Pharmaceuticals, Irvine, CA 92612, USA

## Abstract:

A p-value is a most widely used measure of evidence against a null hypothesis in statistical testing of hypothesis. In Phase 3 clinical trials, a threshold of 0.05 alpha level is usually used to judge against a calculated p-value to conclude whether there is appropriate statistical evidence to support drug regulatory approval decision. The p-value, being derived from a statistical sample via test statistic, has inherent variability, which is generally ignored or not assessed and thus leads to a lack of understanding of its quality when evaluating study outcome from phase 3 clinical trials. In this paper, we use parametric bootstrap approach to assess the p-value variability from published Phase 3 clinical trials.

**Key Words:** p-value, measure of evidence, reproducibility, prediction interval, clinical trial

## 1. Introduction

### 1.1 Motivation

P-value has been widely used statistics in research in almost all scientific disciplines. There are many criticisms about the value of using p-values for research. For example, in drug development, usually two phase 3 studies that demonstrate p-value $\leq 0.05$ are required for FDA approval. However, the quality of the p-value is usually uncertain as only one observed p-value is reported per study. In this research, we utilized publicly available data from a recent FDA approved drug Xiidra for dry eye disease in July 2016 to assess the quality of the p-values from the four clinical studies.

### 1.2 P-value History

Based on Wikipedia (https://en.wikipedia.org/wiki/P-value), records of p-values dated back to 1,770 when a p-value was calculated by Pierre-Simon Laplace when looking at the proportion of births for an excess of boys compared to girls. It was formally introduced by Karl Pearson for Chi-Square test in early 1900s. Then Ronald Fisher popularized the concept in his book, *Statistical Methods for Research Workers* (1925). Fisher also established the 0.05 level for statistical significance.

### 1.3 P-value Definition

From Wikipedia: "The p-value is defined as the probability, under the assumption of hypothesis H, of obtaining a result equal to or more extreme than what was actually observed."

For this research, we define a simple null hypothesis $H_0$, a test statistic $T$, and the observed value of the test statistic $T(x)$; p-value can be defined as:

$$P(T \geq T(x)|H_0)$$

### 1.4 P-value Controversy

A p-value is linked to the idea of statistical significance / evidence adopted in many research applications today. Heated debates in scientific communities including top-notch journals such as *Nature* and *Science. Basic and Applied Social Psychology* (BASP) banned p-values in early 2015. American Statistical Association (ASA) releases statement regarding p-values (2016).

Due to this controversy, we also consider in this assessment the reproducibility probability.

## 2. Notations and the Two Sample Testing Setting

### 2.1 Notations

For the Xiidra analysis we used the large two sample T-test setting comparing a treatment group and placebo (vehicle) group. This drug considered the end points for Eye dryness score (EDS) and inferior fluorescein corneal staining score (ICSS); more details about the observations are discussed in section 5.2. Let the first sample be change of baseline of EDS or ICSS for the treatment group and these observations are notated as $Y_1, Y_2, \dots, Y_{m_1}$ of size $m_1$ and let the second sample be the change of baseline EDS or ICSS defined as $X_1, X_2, \dots, X_{m_2}$ for placebo group of size $m_2$.

Let $Y_1, Y_2, \dots, Y_{m_1}$ have mean $\mu_1$ and variance $\sigma_1^2$ and $X_1, X_2, \dots, X_{m_2}$ have mean $\mu_2$ and variance $\sigma_2^2$. We may notate the collection of sample observations as $\underline{X}$ and $\underline{Y}$The estimates for the $j^{th}$ sample $\mu_j$ and $\sigma_j^2$ are calculated to be the sample average $\hat{\mu}_1 = \frac{1}{m_1}\sum_{i=1}^{m_1} Y_i$ or $\hat{\mu}_2 = \frac{1}{m_2}\sum_{i=1}^{m_2} X$ with sample variances calculated to be $\hat{\sigma}_1^2 = \frac{1}{m_1-1}\sum_{i=1}^{m_1}(Y_i - \hat{\mu}_1)^2$ or $\hat{\sigma}_2^2 = \frac{1}{m_2-1}\sum_{i=1}^{m_2}(X_i - \hat{\mu}_2)^2$.

Let $\lambda$ be the proportion of observations that belong in sample 2, then $\lambda = \frac{m_2}{m_1+m_2}$. Half the harmonic mean $m$ can be calculated as $m = m_1\lambda = \frac{m_1 m_2}{m_1+m_2}$. We can then derive the quantities $\sigma = \sqrt{\lambda\hat{\sigma}_1^2 + (1-\lambda)\hat{\sigma}_2^2}$ and $\hat{\delta} = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sigma}$ which will be used for large sample T-test statistic $T(\underline{Y}, \underline{X})$**.** For a given null hypothesis $H_0$, the p-

value is then calculated to be $P\big(T \geq T(\underline{Y}, \underline{X})|H_0\big)$. We use p-value as an observed value from hypothesis testing and as a random variable in some cases, we use them interchangeably.

### 2.2 Large-sample Two Sample T-test

Since the mean baseline data for treatment group and placebo group in each of the study appear to be identical we can test the hypothesis if the treatment is effective,

$H_0: \mu_1 - \mu_2 = 0$ vs $H_a: \mu_1 - \mu_2 > 0$

A large sample T-test statistic used between treatment and vehicle group would be calculated as,

$$T_{df} = \frac{(\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{m_1} + \frac{\hat{\sigma}_2^2}{m_2}}} = \sqrt{m}\hat{\delta}$$

T is asymptotically normal with mean $\sqrt{m}\delta$ and variance 1. Since we are working with large samples, the null hypothesis $H_0: \mu_1 = \mu_2$ , we could calculate the p-value to be:

$$P\big(T \geq T(\underline{Y}, \underline{X})|H_0\big) \approx P\big(Z \geq T(\underline{Y}, \underline{X})|H_0\big) = P\big(Z \geq \sqrt{m}\delta\big)$$

## 3. Theoretical Distribution of P-values

### 3.1 The Distribution of the P-value Under the Null Hypothesis

The distribution under the null hypothesis of p-value is stochastically greater than or equal to uniform(0,1) distribution. For continuous data, it is equal to uniform(0,1). In other words, if we had data that followed the assumptions of the null hypothesis, then the mean of the p-value distribution E(p)=1/2=0.5, the variance is Var(p)=1/12 ~= 0.0833. These are interesting characteristics of the p-value distribution if in fact the null hypothesis is true.

### 3.2 Exact Distribution of the P-value Under the Alternative Hypothesis

For the two sample T-test the exact distribution of the p-value can be derived by Hung, Biometrics (1997). The density of p-value is as follows,

$$g_\delta(p) = \frac{\phi\big(\Phi^{-1}(1-p) - \sqrt{m}\delta\big)}{\phi\big(\Phi^{-1}(1-p)\big)}$$

Where $\phi(x)$ is the standard normal density evaluated at $x$ and $\Phi^{-1}(1-p) = Z_p$ percentile of the standard normal distribution.

### 3.3 Asymptotic Distribution Under the Alternative Hypothesis

Asymptotic normality of –log(p-value) has been shown by Lambert and Hall (1982). For a hypothesis test where $n$ is the sample size, $\theta$ is a population parameter of interest, $c(\theta)$ and $\tau^2(\theta)$ are the asymptotic mean and variance for the asymptotic distribution,

$$\sqrt{n}\left(\frac{-\log(p)}{n} - c(\theta)\right) \xrightarrow{d} N\left(0, \tau^2(\theta)\right) \text{ as } n \to \infty$$

In our two-sample T test scenario, we can define $\theta$ to be the true population mean difference $\theta = \mu_1 - \mu_2$; we then have $c(\theta) = \frac{m\lambda^2\delta^2}{2}$ and $\tau^2(\theta) = m\lambda^2\delta^2$

## 4. Statistical Simulation and Computation Methods

For each study and endpoint, we only have one observed p-value based on the data. If we could replicate each study multiple times, we would have multiple p-values and would be able to understand some characteristics of the p-value distribution such as mean and variance. Since we don't have each study replicated, we will use parametric bootstrap prediction intervals to generate the empirical p-value distribution to assess the mean and variance of p-values for each study. Using this, we can assess the reproducibility probabilities for each study.

### 4.1 Bias-Corrected (BC) Bootstrap Prediction Intervals

If we have a random sample $Y_1, Y_2, \ldots, Y_n$ and an independent replicate $X_1, X_2, \ldots, X_m$, then we can calculate two p-values based off these replicates to get $P_{Y,n}$ and $P_{X,m}$. Define $P_{Y,n}^{(1)}$ be derived from the first resample. Repeating the process independently B times results in $P_{Y,n}^{(1)}, P_{Y,n}^{(2)} \ldots, P_{Y,n}^{(B)}$. A $1-\alpha$ bias-corrected (BC) bootstrap prediction interval for $P_{X,m}$ has the prediction limits to be in the form $\{\hat{\eta}_B(\alpha_1)), \hat{\eta}_B(1-\alpha_2)\}$ where $\hat{\eta}_B(\alpha)$ is the $\alpha$th sample quantile. We would calculate $\alpha_1 = \Phi\left(z_{\frac{\alpha}{2}}\left(1+\frac{m}{n}\right)^{\frac{1}{2}} + \hat{z}_0\left(\frac{m}{n}\right)^{\frac{1}{2}}\right)$ and $\alpha_2 = \Phi\left(z_{\frac{\alpha}{2}}\left(1+\frac{m}{n}\right)^{\frac{1}{2}} - \hat{z}_0\left(\frac{m}{n}\right)^{\frac{1}{2}}\right)$. $z_{\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ percentile of the standard normal distribution or $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$. We would also need to calculate $\hat{z}_0 = \Phi^{-1}\left(\widehat{K}_B\left(P_{Y,n}\right)\right)$, where $\widehat{K}_B(\cdot)$ is the empirical CDF for $P_{Y,n}^{(i)}$.

### 4.2 Reproducibility Probability

Following Shao & Chow (2002), define RP to be the reproducibility population parameter:

$$RP = P(p_{new} \leq \alpha)$$

For the two-sample test scenario:

$$\widehat{RP} = 1 - \Phi(-T(\boldsymbol{x}) + z_\alpha) = \Phi(-\Phi^{-1}(p_{obs}) - z_\alpha)$$

## 5. Clinical Trial Data

The study data for this research comes from a recent FDA approved drug label which is publically available. The drug name is called Xiidra™ (lifitegrast ophthalmic solution) 5%. This data was used here for p-value quality research purposes only.

### 5.1 The Drug

From Xiidra label, Lifitegrast is a chemical drug. The molecular formula of lifitegrast is C29H24Cl2N2O7S and its molecular weight is 615.5 $g \cdot mol^{-1}$. Xiidra (lifitegrast ophthalmic solution) 5% is a lymphocyte function-associated antigen-1 (LFA-1) antagonist supplied as a sterile, clear, colorless to slightly brownish-yellow colored, isotonic solution with a pH of 7.0–8.0 and an osmolality range of 200–330 mOsmol/kg. *In vitro* studies demonstrated that lifitegrast may inhibit secretion of inflammatory cytokines in human peripheral blood mononuclear cells. The exact mechanism of action of lifitegrast in dry eye disease is not known.

### 5.2 Four Clinical Studies and the Endpoints

The safety and efficacy of lifitegrast for the treatment of dry eye disease (DED) were assessed in a total of four 12-week, randomized, multi-center, double-masked, vehicle-controlled studies. Patients were randomized to Xiidra or vehicle (placebo) in a 1:1 ratio and dosed twice a day. Use of artificial tears was not allowed during the studies. The mean age was 59 years (range, 19–97 years). Most patients were female (76%). Enrollment criteria included, minimal signs (i.e., Corneal Fluorescein Staining (CFS) and non-anesthetized Schirmer Tear Test (STT)) and symptoms (i.e., Eye Dryness Score (EDS) and Ocular Discomfort Score (ODS)) severity scores at baseline.

EDS was used to measure the symptoms of DED and was rated by patients using a visual analogue scale (VAS) (0 = no discomfort, 100 = maximal discomfort) at each study visit (Days 0, 14, 42, 84). While ICSS was used to measure the signs of DED and was rated a score (0 = no staining, 1 = few/rare punctate lesions, 2 = discrete and countable lesions, 3 = lesions too numerous to count but not coalescent, 4 = coalescent) which was recorded at each study visit (Days 0, 14, 42, 84). For our research purposes, we have taken data to be the change of baseline at day 84 for both EDS and ICSS.

### 5.3 Safety Data

The most common adverse reactions reported in 5-25 % of patients: instillation site irritation, dysgeusia and reduced visual acuity. Other adverse reactions reported in 1% to 5% of the patients: blurred vision, conjunctival hyperemia, eye irritation, headache, increased lacrimation, eye discharge, eye discomfort, eye pruritus and sinusitis. The drug seems to be safe, for this research we will not assess safety data and we will focus on efficacy only.

### 5.4 Efficacy Data

Efficacy data including symptoms and signs measures for EDS and ICSS. In this study we focus on study endpoint day 84 (12 weeks). The data was extracted from the drug

label. See table 1 for EDS change from baseline at Day 84 and table 2 for ICSS change from baseline at Day 84.

**Table 1.** EDS - Change from baseline at day 84.

| | $m_i$ | | $\hat{\mu}_i$ | | $\hat{\sigma}_i$ | |
|---|---|---|---|---|---|---|
| **Study** | Ctl | Xiidra | Ctl | Xiidra | Ctl | Xiidra |
| **1** | 58 | 58 | -7.2 | -14.4 | 25.29 | 25.26 |
| **2** | 295 | 293 | -11.2 | -15.2 | 28.78 | 31.48 |
| **3** | 360 | 358 | -22.8 | -35.3 | 28.60 | 28.40 |
| **4** | 356 | 355 | -30.5 | -37.7 | 28.03 | 28.91 |

**Table 2.** ICSS - Change from baseline at Day 84.

| $m_i$ | | $\hat{\mu}_i$ | | $\hat{\sigma}_i$ | |
|---|---|---|---|---|---|
| Ctl | Xiidra | Ctl | Xiidra | Ctl | Xiidra |
| 58 | 58 | 0.38 | 0.04 | 0.79 | 0.75 |
| 295 | 293 | 0.17 | -0.07 | 0.82 | 0.87 |
| 360 | 358 | -0.71 | -0.73 | 0.94 | 0.93 |
| 356 | 355 | -0.63 | -0.80 | 0.91 | 0.94 |

## 6. P-value Quality Assessment Results

### 6.1 P-value Estimates from the Two-sample T-test Results

To roughly assess the statistical significance, we performed large sample based two sample two-sided T-test comparing the drug and vehicle as described in section 2. The comparison results for EDS is presented in table 3 and ICSS is presented in table 4. From the analysis, we can see that in study 3 and 4 from EDS endpoints were statistically significant at $\alpha = 0.05$ and study 1, 2, and 4 were statistically significant at $\alpha = 0.05$.

**Table 3.** Two sample T-test comparing drug and vehicle for change of baseline of EDS at day 84.

| study | diff | Est. sigma | Est. effect size | T-test statistic | p-value |
|---|---|---|---|---|---|
| 1 | -7.20 | 25.28 | -0.28 | -1.53 | 0.0625 |
| 2 | -4.00 | 30.16 | -0.13 | -1.61 | 0.0539 |
| 3 | -12.50 | 28.50 | -0.44 | -5.89 | 0.0000 |
| 4 | -7.20 | 28.47 | -0.25 | -3.37 | 0.0004 |

**Table 4.** Two-sample T-test comparing drug and vehicle for change of baseline of ICSS at day 84.

| Study | Diff | Est. sigma | Est. effect size | T-test statistic | p-value |
|---|---|---|---|---|---|
| 1 | -0.34 | 0.77 | -0.44 | -2.39 | 0.0084 |
| 2 | -0.24 | 0.84 | -0.28 | -3.45 | 0.0003 |
| 3 | -0.02 | 0.93 | -0.02 | -0.29 | 0.3872 |
| 4 | -0.17 | 0.93 | -0.18 | -2.45 | 0.0071 |

### 6.2 Empirical Distribution

To understand the p-value data, we first used parametric bootstrap method to obtain the empirical distribution for the p-values comparing the drug and vehicle for symptom measure EDS and sign measure ICSS.

#### 6.2.1 Empirical Distribution for P-value

A few percentiles of empirical p-value distribution are presented such as $5^{th}$, $10^{th}$, $25^{th}$, $50^{th}$ (median), $75^{th}$, $90^{th}$, and $95^{th}$ percentile in table 5 (EDS) and table 6 (ICSS).

**Table 5.** Percentiles of empirical P-value distribution comparing drug and vehicle for change of baseline of EDS at day 84.

| | | | | Percentile | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Study | Mean | Median | SD | 5th | 10th | 25th | 75th | $90^{th}$ | 95th |
| 1 | 0.1383 | 0.0613 | 0.1799 | 0.0007 | 0.0024 | 0.0133 | 0.1928 | 0.3991 | 0.5463 |
| 2 | 0.1276 | 0.0545 | 0.1707 | 0.0006 | 0.0019 | 0.0109 | 0.1767 | 0.3671 | 0.5167 |
| 3 | 0.0000 | 0.0000 | 0.0003 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 4 | 0.0081 | 0.0004 | 0.0288 | 0.0000 | 0.0000 | 0.0000 | 0.0036 | 0.0175 | 0.0388 |

**Table 6.** Percentiles of empirical P-value distribution comparing drug and vehicle for change of baseline of ICSS at day 84.

| | | | | Percentile | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Study | Mean | Median | SD | $5^{th}$ | 10th | 25th | 75th | 90th | 95th |
| 1 | 0.0465 | 0.0093 | 0.0938 | 0.0000 | 0.0002 | 0.0012 | 0.0447 | 0.1342 | 0.2275 |
| 2 | 0.0075 | 0.0003 | 0.0293 | 0.0000 | 0.0000 | 0.0000 | 0.0029 | 0.0152 | 0.0349 |
| 3 | 0.4223 | 0.3889 | 0.2842 | 0.0283 | 0.0616 | 0.1726 | 0.6557 | 0.8440 | 0.9185 |
| 4 | 0.0416 | 0.0068 | 0.0899 | 0.0000 | 0.0001 | 0.0009 | 0.0363 | 0.1182 | 0.2129 |

**Figure 1.** Histogram of P-values for EDS change of baseline day 84 with theoretical P-value distribution imposed.

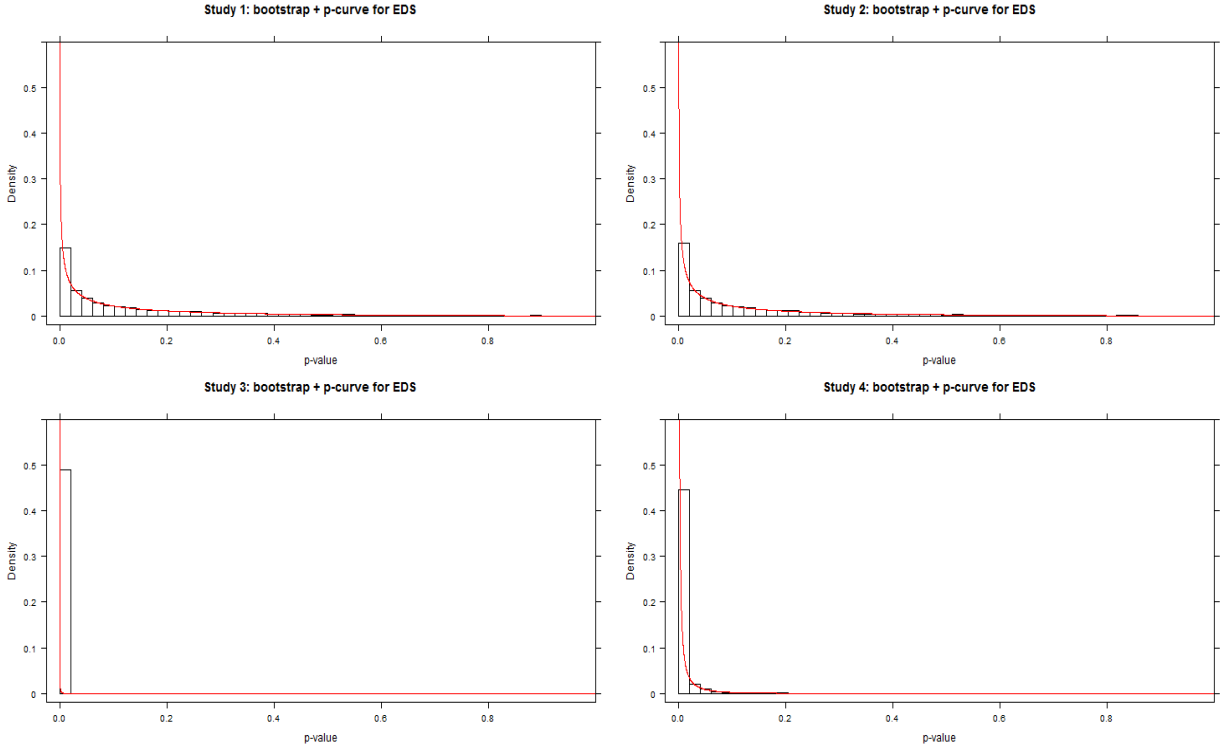
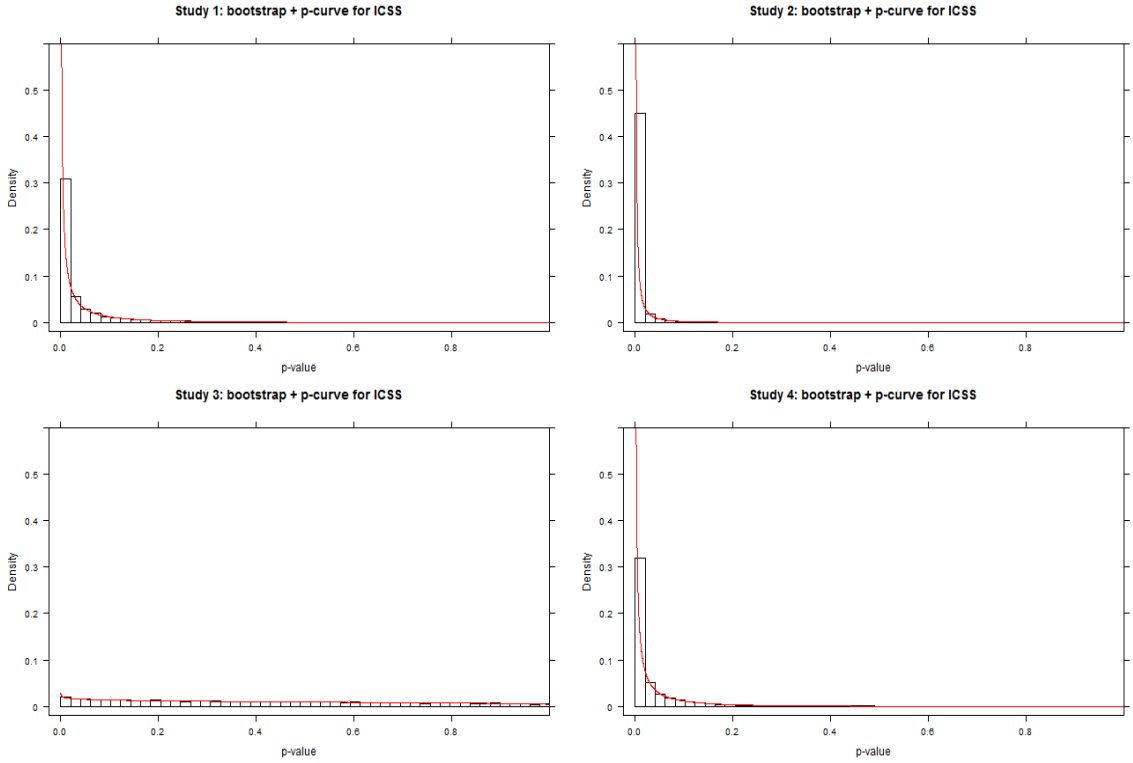



**Figure 2.** Histogram of P-values for ICSS change of baseline day 84 with theoretical P-value distribution imposed.



### 6.2.2 Empirical Distribution for -log(P-value)

As described in section 3.3, -log(p-value) is asymptotically normal. Here we present the -log transformed p-values results of a few percentiles for the empirical p-value distribution are presented such as $5^{th}$, $10^{th}$, $25^{th}$, $50^{th}$ (median), $75^{th}$, $90^{th}$, and $95^{th}$ percentile in table 7 (EDS) and table 8 (ICSS).

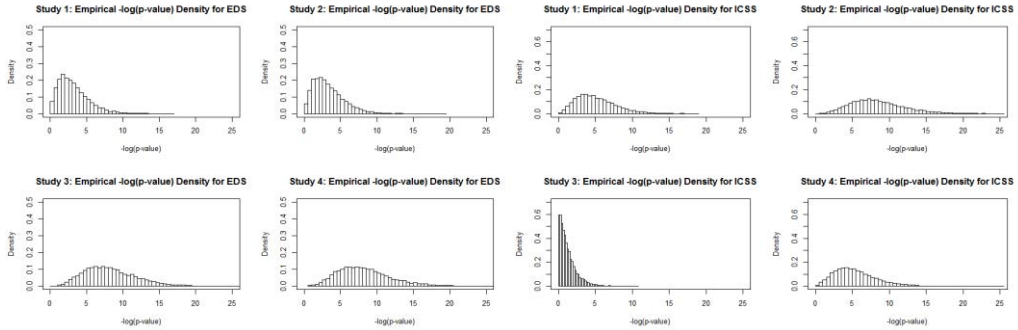
**Table 7.** Percentiles of empirical -log(P-value) distribution comparing drug and vehicle for change of baseline of EDS at day 84.

| | | | | Percentile | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Study | Mean | Median | SD | 5th | 10th | 25th | 75th | 90th | 95th |
| 1 | 3.2117 | 2.7924 | 2.1184 | 0.6045 | 0.9184 | 1.6458 | 4.3181 | 6.0292 | 7.2651 |
| 2 | 3.3353 | 2.9101 | 2.1312 | 0.6603 | 1.0021 | 1.7330 | 4.5129 | 6.2624 | 7.3618 |
| 3 | 20.0832 | 19.4471 | 5.8625 | 11.4445 | 13.1075 | 15.9350 | 23.9280 | 27.9327 | 30.2392 |
| 4 | 8.2934 | 7.8133 | 3.6472 | 3.2498 | 4.0444 | 5.6313 | 10.3868 | 13.2856 | 15.1840 |

**Table 8**. Percentiles of empirical -log(P-value) distribution comparing drug and vehicle for change of baseline of ICSS at day 84.

| | | | | Percentile | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Study | Mean | Median | SD | 5th | 10th | 25th | 75th | 90th | 95th |
| 1 | 5.1070 | 4.6757 | 2.7055 | 1.4808 | 2.0084 | 3.1072 | 6.6872 | 8.7576 | 10.1832 |
| 2 | 8.4701 | 8.0333 | 3.6351 | 3.3558 | 4.1880 | 5.8507 | 10.6178 | 13.3367 | 15.1207 |
| 3 | 1.2694 | 0.9445 | 1.1652 | 0.0850 | 0.1697 | 0.4220 | 1.7567 | 2.7870 | 3.5664 |
| 4 | 5.3976 | 4.9839 | 2.8423 | 1.5467 | 2.1356 | 3.3154 | 7.0093 | 9.2014 | 10.7346 |

**Figure 3**. Histogram of -log(P-values) for EDS (left) and ICSS (right) change of baseline day 84 with theoretical P-value distribution imposed.



### 6.3 Bootstrap Prediction Intervals

Since p-values is a statistic, which is a random variable, to quantify its variability we have used bootstrap prediction intervals for EDS and ICSS change from baseline. The bootstrap method used is described in section 4.

#### 6.3.1: Bootstrap Prediction Intervals for P-value

**Table 9**. Bias corrected bootstrap prediction intervals for P-value for EDS change from baseline data.

| study | Point Estimate | Standard Error | 95% Interval Lower | 95% Interval Upper |
|---|---|---|---|---|
| 1 | 0.15 | 0.006 | 0.136 | 0.159 |
| 2 | 0.13 | 0.006 | 0.121 | 0.143 |
| 3 | 0.00 | 0.000 | 0.000 | 0.000 |
| 4 | 0.01 | 0.001 | 0.007 | 0.011 |

**Table 10.** Bias corrected bootstrap prediction intervals for P-value for ICSS change from baseline data.

| study | Point Estimate | Standard Error | 95% Interval Lower | 95% Interval Upper |
|---|---|---|---|---|
| 1 | 0.0364 | 0.0049 | 0.0365 | 0.0461 |
| 2 | 0.0087 | 0.0010 | 0.0066 | 0.0107 |
| 3 | 0.4220 | 0.0089 | 0.4045 | 0.4395 |
| 4 | 0.0405 | 0.0027 | 0.0352 | 0.0457 |

#### 6.3.2: Bootstrap Prediction Intervals for -log(P-value)

**Table 11.** Bias corrected bootstrap prediction intervals for -log(P-value) for EDS change from baseline data.

| study | Point Estimate | Standard Error | 95% Interval Lower | 95% Interval Upper |
|---|---|---|---|---|
| 1 | 3.13 | 0.063 | 3.011 | 3.258 |
| 2 | 3.32 | 0.069 | 3.187 | 3.456 |
| 3 | 19.78 | 0.184 | 19.420 | 20.140 |
| 4 | 8.21 | 0.107 | 8.003 | 8.423 |

**Table 12.** Bias corrected bootstrap prediction intervals for -log(P-value) for ICSS change from baseline data.

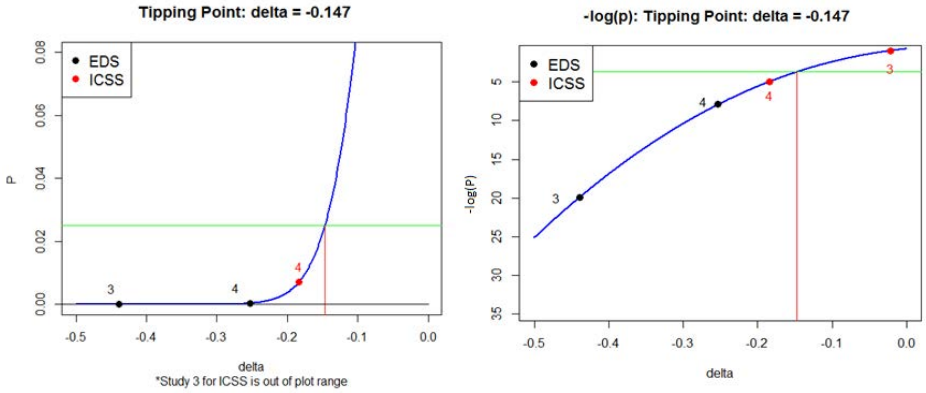
| study | Point Estimate | Standard Error | 95% Interval Lower | 95% Interval Upper |
|---|---|---|---|---|
| 1 | 5.29 | 0.088 | 5.116 | 5.462 |
| 2 | 8.47 | 0.118 | 8.235 | 8.697 |
| 3 | 1.22 | 0.033 | 1.159 | 1.288 |
| 4 | 5.50 | 0.090 | 5.319 | 5.673 |

### 6.4: Tipping Point Analysis

Tipping point analysis is usually used for sensitivity analysis; we had adopted the idea to assess the p-value quality relative to clinical meaningfulness/effect size. When assessing the quality of the p-value, tipping point is the threshold of the effect size value to transition between a non-significant result and a significant result. We brought this idea to assess the quality of the p-value in relation to the effect size which is directly linked to clinical meaningfulness of the results. The variability for both EDS and ICSS across the four clinical studies were stable.

### 6.4.1: Tipping Point Assessment

**Figure 4.** Tipping point analysis for P-value (left) and -log(P-value) (right).

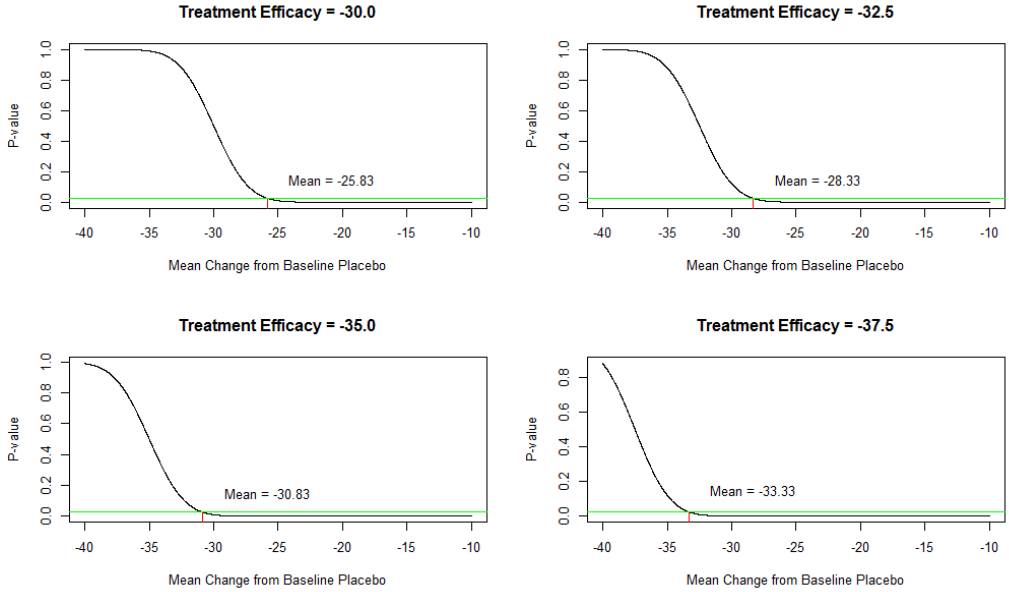



### 6.4.2: Explore the Impact of Placebo Effect

We explore the potential impact of placebo effect on the quality of p-value. After checking the variability of change of baseline in EDS and ICSS at day 84, we have found the variance is similar between 3 and 4 so we hold the variability ($\sigma_{EDS} = 28.5, \sigma_{ICSS} = 0.93$) and sample size constant ($n = 360$ per group) as observed. Then we explore the relationship of p-value and Cohen effect size, we can see from the figure that we have identified the maximum placebo effects before the p-value is no longer significant at $\alpha = 0.025$. For instance, assuming the EDS change from baseline at day 84 in the treatment group efficacy is -30,

then the maximum tolerable placebo effect can be -25.83 before the results are no longer significant at $\alpha = 0.025$.

**Figure 5.** Exploring the impact of the placebo effect.




## 6.5: Reproducibility Probability

Reproducibility probability (RP) is one of the crucial topics for clinical trials. There are different definitions for reproducibility probability. In this research paper, we follow Shao & Chow (2002) and defined RP to be the reproducibility population parameter:

$$RP = P(p_{new} \leq \alpha)$$

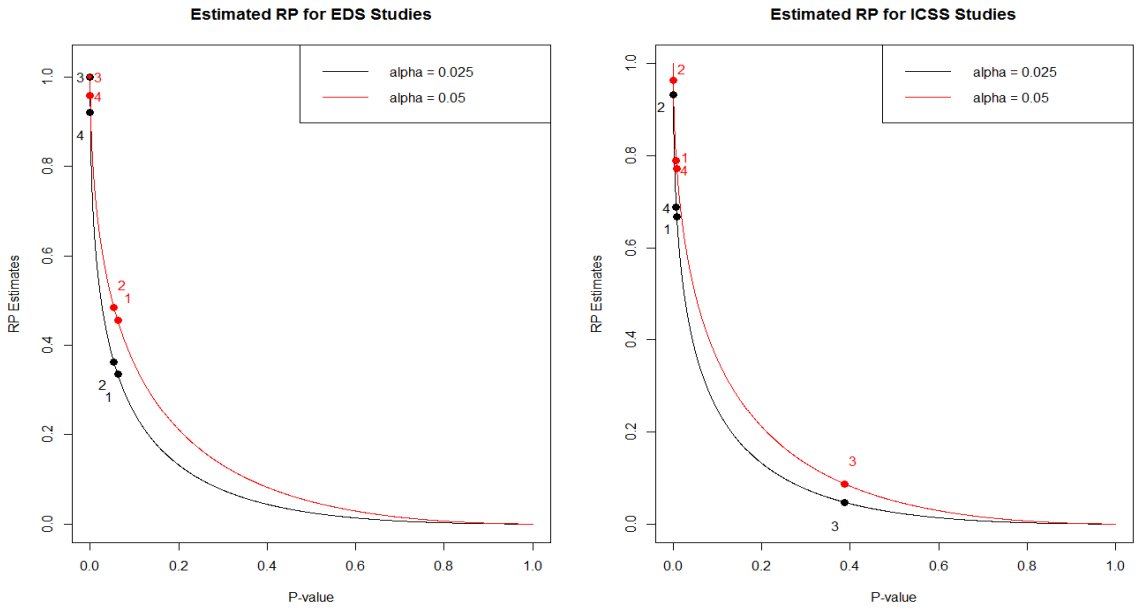
For the two-sample test scenario:

$$\widehat{RP} = 1 - \Phi(-T(x) + z_\alpha) = \Phi(-\Phi^{-1}(p_{obs}) - z_\alpha)$$

### 6.5.1: Reproducibility Probability Point Estimate

From the derivation of RP, we can see a direct one to one relationship between p-value and $\widehat{RP}$. The issue is p-value is a statistic and thus RP which is a mathematical function of p-value is also inevitably a random variable. Therefore, the usefulness of RP is limited.

**Figure 6**. Reproducibility probability for EDS (left) and ICSS (right) for the change of baseline between the drug and vehicle at day 84.




### 6.5.2: Reproducibility Probability Variability

**Table 13.** Bias Corrected Bootstrap Prediction Intervals for -log(P-value) for EDS Change from Baseline Data.

| **EDS:** study | Point Estimate | BC Estimated Standard Error | 95% Interval Lower | 95% Interval Upper |
|---|---|---|---|---|
| 1 | 0.33 | 0.015 | 0.297 | 0.355 |
| 2 | 0.36 | 0.015 | 0.326 | 0.384 |
| 3 | 0.99 | <0.0001 | 0.9999 | 0.9999 |
| 4 | 0.91 | 0.009 | 0.893 | 0.929 |

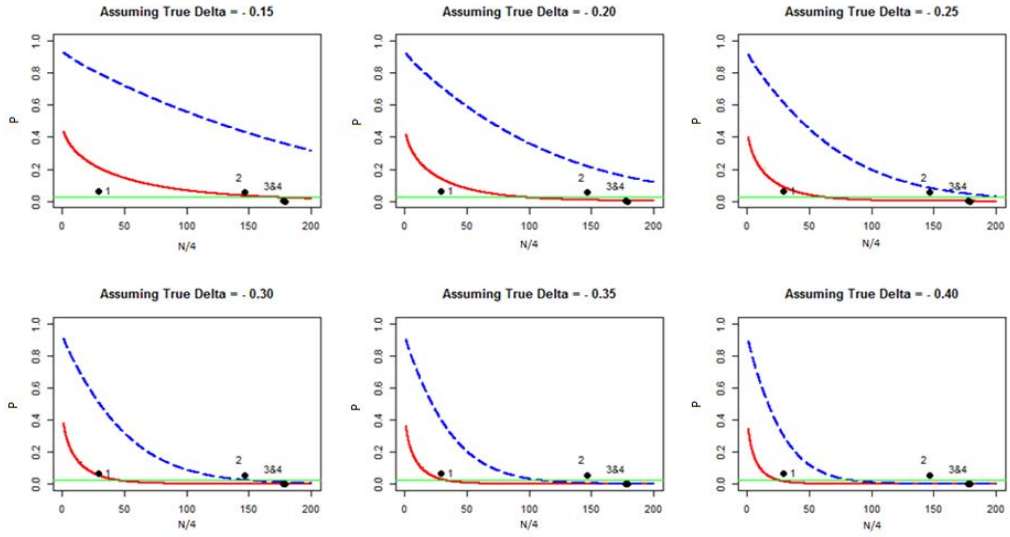
**Table 14.** Bias Corrected Bootstrap Prediction Intervals for -log(P-value) for ICSS Change from Baseline Data.

| **ICSS:** study | Point Estimate | BC Estimated Standard Error | 95% Interval Lower | 95% Interval Upper |
|---|---|---|---|---|
| 1 | 0.67 | 0.015 | 0.646 | 0.703 |
| 2 | 0.93 | 0.008 | 0.913 | 0.945 |
| 3 | 0.05 | 0.007 | 0.036 | 0.062 |
| 4 | 0.67 | 0.015 | 0.645 | 0.704 |

### 6.6: Aggregated Assessment of Multiple Studies

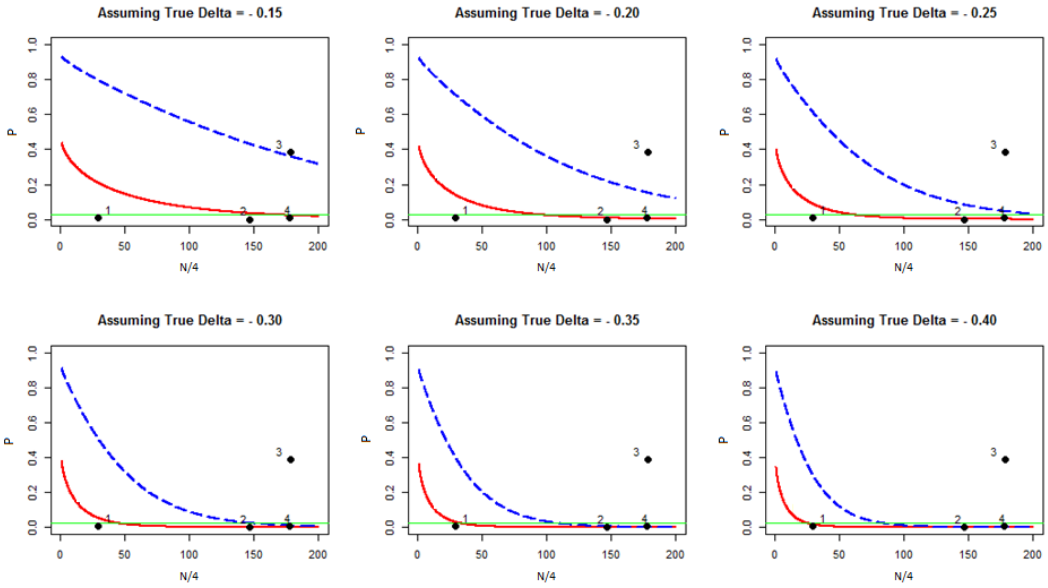
Up to now, we have assessed the p-value based on individual studies. In order to assess the quality of p-value totality, we followed the Biometrics paper by Hung, O'Neill, Bauer, and Kohne (1997) which proposed Phyp plots for meta-analysis purposes. Here we employed the Phyp plots to aid the assessment of p-value quality for multiple studies. The theoretical 95$^{th}$ and 50$^{th}$ percentile of P-value distribution for different Cohen effect sizes. The observed outcomes (p-values and sample size) for the four clinical studies are plotted to assess the totality of the statistical significance for decision making.

**Figure 7.** Phyp Plot for Different Effect Sizes $\delta$ Values for EDS Change of Baseline Comparing the Drug and Vehicle at Day 84.




*Blue line = 95th percentile ; Red line = 50th percentile

**Figure 8.** Phyp Plot for Different Effect Sizes $\delta$ Values for ICSS Change of Baseline Comparing the Drug and Vehicle at Day 84.




## 7. Limitations

This study has limitations. First only summarized data (sufficient statistics) are available for this research. Thus, we had to use parametric bootstrap to generate empirical p-value distributions. Nonparametric bootstrap might be a better choice if individual data are available as it does not put on a parametric distribution constraint. For this research the mean change of baseline in EDS, ICSS, at day 84 follow a parametric Gaussian statistical models. Secondly, the original study used analysis of covariance (ANCOVA) model. Here we used large-sample T test instead due to lack of subject level data. Therefore, we cannot assess the impact of potential confounding factors or missing data on p-value quality.

## 8. Discussions

A p-value is a most widely used statistical measure of evidence against a null hypothesis in statistical hypothesis testing and is generally required in Phase 3 clinical trials for drug registration approval process. A threshold of 0.05 alpha level is usually used to judge against a calculated p-value to conclude whether there is appropriate statistical evidence to support drug regulatory approval.

In pivotal clinical trial, the p-value is a required statistic, however, only one observed p-value (point estimate) is usually reported for a study. The p-value, being derived from a sample (data) via test statistic, has inherent variability, which is usually ignored or not assessed and thus leads to a lack of understanding

of its quality. The quality of this p-value is unknown. In this paper, we assessed the quality of p-value using parametric bootstrap approach to assess the p-value distribution and variability for these four Phase 3 clinical studies from the Xiidra package insert.

First, to describe the distribution profile of the p-value, we have developed better/deeper understanding on the study data of EDS and ICSS. Secondly, prediction interval is used to assess the variability of p-value. For instance, the change of baseline at day 84 in EDS for study 3, the empirical p-value distribution indicates that the 95 percentile is still below 0.0001 and the 95% prediction interval is (0.0000,0.0001) [Note: In the Table 9, we reported up to 3 decimal points.]. On the other hand, change of baseline at day 84 in ICSS for study 3, the empirical p-value distribution indicates that the $10^{th}$ percentile is above 0.06 and the 95% prediction interval is (0.4045,0.4395).

Reproducibility can be misleading if only the p-value without variability is used.

Tipping point analysis can be a critical tool assess clinical meaningfulness in relationship to p-value. Can help with decision making. To explore the placebo effect, we also use a tipping point idea fixing the treatment effect and understanding the magnitude of the placebo's impact on the study outcome.

In conclusion, the descriptive statistics of p-value distribution helps better understanding the study p-values. Characterizing p-value and its prediction intervals leads to better understanding about the variability of the p-value and the study data outcome. We recommend reporting the p-value along with its characteristics such as percentiles and prediction interval in clinical studies. P-value itself may not be very useful as we demonstrated through the reproducibility probability. The assessment of p-value quality should be conducted in linkage with external information such as Cohen's effect size and sample size.

## 9. Acknowledgements

Authors would like to thank Allergan for the support of this research.

## References

Boos DD (2009). The variability of p-values. NC State Statistics Department Tech Report #2626.

Boos DD and Stefanski LA (2011). P-value precision and reproducibility. *The American Statistician*, 65(4): 213-221.

Halsey LG, Curran-Everett D, Vowler SL and Drummond GB (2015). The fickle p value generates irresponsible results. *Nature Methods* 12 (3): 179-185.

Hung HMJ, O'Neill RT, Bauer P, and Kohne K (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*; 53(1): 11-22

Lambert D and Hall WJ (1982). Asymptotic lognormality of p-values. The *Annuals of Statistics*; 10: 44-64

Senn S (2007). Significance tests and p-values. In Senn S. Statistical Issues in Drug Development 2nd ed., John Wiley & Sons Ltd, Chichester, West Sussex PO19 8SQ, England.

Shao J and Chow S-C (2002). Reproducibility probability in clinical trials. Statistics in Medicine, 21: 1727-1742.

Wasserstein RL and Lazar NA (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2): 129-133.

Shire US Inc (2016). Xiidra® US FDA full prescribing information. 07/2016.