# Statistical methods for comparing limits of detection in molecular diagnostics a practical guide

Jesse A Canchola

MedCrave
Gateway into the World of Research

# Statistical methods for comparing limits of detection in molecular diagnostics: a practical guide


## Authors Information

**Jesse A Canchola**

MS, PStat®, Roche Molecular Systems, Inc., Pleasanton, California, USA

## *Correspondance:

Jesse A Canchola, Roche Molecular Systems, Inc., 4300 Hacienda Drive, Pleasanton, California USA, Email jesse.canchola@roche.com



## Published By:

**MedCrave Group LLC**

February 27, 2026

# Contents

## Abstract

The limit of detection (LoD) is a critical performance characteristic in molecular diagnostics, defining the lowest analyte concentration at which an assay achieves reliable detection.[1,2] While numerous approaches exist for estimating LoD in both quantitative and qualitative assays, confusion often arises regarding appropriate methods for statistical comparison between a reference formulation and a new test condition.[3,4] This paper provides a comprehensive overview of theoretical frameworks, statistical models, and practical considerations for LoD comparison, with additional attention to multi-target assay scenarios common in contemporary molecular diagnostics.

We contrast quantitative assays, where LoD is a numeric estimate derived from continuous data, with qualitative assays, where detection probabilities are assessed across dilution levels. Worked examples illustrate correct applications of the delta method,[5] Fieller's theorem,[6] bootstrap,[7] two one-sided test (TOST) procedures[8] and non-inferiority procedures when Just Enough Testing (JET) is the objective. The discussion emphasizes common pitfalls, particularly the inappropriate reliance on overlapping confidence intervals as evidence of equivalence,[9,10] and offers regulatory-compliant recommendations for both single-target and multi-target comparison studies.[11,12] Guidance on sample size determination, multiplicity considerations, and interpretation frameworks ensures that practitioners can design and execute statistically sound LoD comparison studies that meet regulatory expectations or, when using the JET method, for delivery to assay development for optimization.

**Keywords:** non-inferiority, detection, molecular diagnostics, analytical sensitivity, equivalence, multi-target scenarios, statistical guidance, binary outcome

**Abbreviations:** LoD, limit of detection; JET, just enough testing; TOST, two one-sided test; NHST, null hypothesis significance testing; Ct, cycle threshold; MOE, margin of error; TOST, two one-sided tests

# Introduction

## The central role of analytical sensitivity

Analytical sensitivity is a cornerstone of assay validation in molecular diagnostics, representing the ability of a test to detect low concentrations of target analyte.[13] The Clinical and Laboratory Standards Institute (CLSI) defines the limit of detection (LoD) as the lowest analyte concentration that can be consistently detected with a stated probability, typically 95%.[1,2] This seemingly straightforward definition belies considerable complexity in practice, particularly when comparing LoDs between different assay conditions.

In the modern diagnostic landscape, manufacturers frequently need to compare LoDs when introducing formulation changes, new reagent lots, alternative instruments, or modified protocols.[14,15] Regulatory agencies, including the U.S. Food and Drug Administration (FDA) and international bodies, require rigorous evaluation demonstrating that such changes do not adversely affect analytical sensitivity.[11,16] The stakes are high: an undetected degradation in LoD could result in false-negative results with serious clinical consequences,[17,18] while overly conservative statistical approaches may unnecessarily delay beneficial improvements to diagnostic tests and potentially delay patient therapy.

## The equivalence challenge

The fundamental question in LoD comparison studies is not whether two conditions are *identical*, they rarely are, but whether they are *equivalent* within clinically acceptable bounds.[8,19] This distinction is critical but often misunderstood. Traditional null hypothesis significance testing (NHST), which asks "are these different?", is poorly suited to equivalence questions.[20] A non-significant p-value does not prove similarity; it merely indicates insufficient evidence of difference, potentially reflecting inadequate statistical power rather than genuine equivalence.[21]

Despite clear guidance from regulatory authorities and statistical literature,[11,22] misapplications remain common. Perhaps the most persistent error is treating overlapping confidence intervals between two LoD estimates as evidence of equivalence.[9,10,23] This approach is statistically invalid because it conflates individual parameter uncertainty with the uncertainty of their difference. Two LoDs with overlapping confidence intervals may still differ by a clinically meaningful amount, while two with non-overlapping intervals may actually be equivalent when their difference is properly assessed.[24]

## Scope and structure

This paper delineates theoretical and applied approaches for comparing LoDs across assay types, with the goal of harmonizing practice across laboratories and regulatory submissions. We address both quantitative assays (those producing continuous measurements such as PCR quantitation or semi-quantitative cycle threshold (Ct) values) and qualitative assays (those producing binary detected/not-detected outcomes). Additional attention is paid to multi-target scenarios, increasingly relevant as multiplex molecular assays become standard in infectious disease diagnostics,[25,26] pharmacogenomics,[27] and oncology testing.[28]

The methods presented here reflect current best practices as outlined in CLSI guidelines,[1,2] FDA statistical guidance documents,[11] and peer-reviewed statistical literature.[8,19,29] Our aim is to provide both theoretical grounding and practical, implementable approaches that working diagnostics professionals can apply with confidence.

# Methods

## Quantitative assays: modeling continuous detection probability

### Theoretical framework

For continuous assays, such as real-time PCR where cycle threshold (Ct) values are measured and converted to quantitative values via a reference standard,[30,31] or fluorescence-based assays producing intensity measurements, the probability of detection is modeled as a function of analyte concentration. The standard approach uses logistic or probit regression to characterize the relationship between log-transformed concentration and the binary outcome of detection (above or below a predefined analytical threshold).[4,32,33]

The logistic model takes the form:

$$P(detect|c) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot log_{10}(c))}} \qquad (1)$$

where c is the analyte concentration, $\beta_0$ is the intercept, and $\beta_1$ is the slope parameter reflecting how steeply detection probability rises with concentration. The $LoD_{95}$ is defined as the concentration at which $P(detect|c) = 0.95$, which can be solved analytically:[4]

$$log_{10}(LoD_{95}) = \frac{logit(0.95) - \beta_0}{\beta_1} \qquad (2)$$

where $logit(0.95) = \ln\left(\frac{0.95}{1-0.95}\right) \cong 2.944$. Note that, in general, the $logit_b\pi = log_b\left(\frac{\pi}{1-\pi}\right)$, where b is any positive base except 1. It is of note that the $logit_b\pi$ must use the same base as whatever the link function was used in the logistic regression to estimate $\beta_0$ and $\beta_1$ (typical software uses the natural logarithm) and does not have to match a base used with any logarithmic transformations of the LoD.

### Variance estimation and confidence intervals

The standard error of the $\log_{10}$(LoD) estimate can be obtained through the delta method,[5,34] which approximates the variance of a function of random variables using a first-order Taylor expansion. Alternatively, profile likelihood methods[35] or bootstrap resampling[7,36] can provide more robust confidence intervals, particularly when sample sizes are modest or detection curves are not ideally sigmoidal.

For comparing two conditions (reference, r, and test, t), we estimate $\log_{10}$(LoDs), viz., $\theta_r$ and $\theta_t$, and their respective variances Var($\theta_r$) and Var($\theta_t$). Assuming independence (typically valid when different sample aliquots are used), the difference, $\Delta_{quant}$, is then:

$$\Delta_{quant} = \theta_t - \theta_r \qquad (3)$$

with standard error:

$$SE\left(\Delta_{quant}\right)=\sqrt{Var\left(\theta_r\right)+Var\left(\theta_t\right)} \quad (4)$$

A 95% Wald confidence interval for the "true" log-difference, $\Delta$, is constructed as:

$$\Delta_{quant}=D_{quant}\pm 1.96\Delta SE\left(D_{quant}\right) \quad (5)$$

where

$$D_{quant}=\hat{\theta}_t-\hat{\theta}_r \quad (6)$$

is the observed difference estimator for $\Delta_{quant}$ and where $\hat{\theta}_t$ is an estimator for $\theta_t$ and $\hat{\theta}_r$ is an estimator for $\theta_r$ such that

$$SE\left(D_{quant}\right)=\sqrt{Var\left(\hat{\theta}_r\right)+Var\left(\hat{\theta}_t\right)} \quad (7)$$

is the standard error of the observed difference and where both are estimated empirically from the experimental data.

### The ratio metric

Because the LoD is inherently a multiplicative quantity (i.e., concentrations span orders of magnitude), the ratio of LoDs, $CI_{quant\ ratio}$, is often more interpretable than their difference.[4,37] Back-transforming the confidence interval on the log-difference yields a confidence interval for the ratio using Fieller's theorem or similar approaches:[6,38]

$$CI_{quant\ ratio}=\left(10^{CI_{low}},10^{CI_{high}}\right) \quad (8)$$

This ratio directly answers questions like "Is the test LoD within 25% of the reference LoD?" Equivalence is declared if the entire ratio confidence interval lies within pre-specified bounds, commonly (0.80, 1.25) (say, within ±25%, representing the margin of error or MOE) or (0.85, 1.18) (within ±15%), for example, depending on regulatory requirements and clinical risk assessment.[11,19]

### Study design considerations

Effective quantitative LoD comparison studies require careful attention to experimental design:[1,2,39]

I. **Dilution series**: Typically 5 to 8 concentration levels spanning approximately 0.5x to 3x, covering the anticipated LoD from both ends

II. **Replicates per level**: 20 to 24 replicates provide minimal reasonable precision;[40] more are needed for steep detection curves or if using probit models

III. **Concentration spacing**: In exponential distributions as in PCR assays, log-uniform spacing ensures adequate information across the detection probability range[4]

IV. **Randomization**: Run order should be randomized to avoid systematic biases from time trends or instrument drift[41]

V. **Quality controls**: Include positive and negative controls to validate assay performance throughout the study[1,2]

## Qualitative assays: binary detection outcomes

### The fundamental difference

Qualitative assays produce only binary outcomes: detected or not detected. Unlike quantitative assays, there is no continuous measurement to model. This fundamentally changes both the study design and the statistical analysis.[1,42] The goal shifts from estimating a numeric LoD to demonstrating equivalent detection *probability* at predefined concentration levels.

### Standard three-level design

The CLSI-recommended design for qualitative LoD comparison includes three concentration levels:[1,2]

I. **0.5x level**: Approximately 40% to 60% detection probability as an empirical design target but can be as high as 78% under the Poisson distribution*; verifies that both assays detect at low frequency when analyte is scarce

II. **1x level**: Approximately 95% detection probability; the primary comparison point where equivalence must be demonstrated

III. **3x level**: Approximately 100% detection probability; verifies that both assays reliably detect when analyte is abundant

*Under the Poisson distribution, at 0.5x, the mean is

$$\lambda=\frac{\lambda_{LoD}}{2}\cong 1.498\Rightarrow P\left(detect\ at\ 0.5x\right)=1-e^{-1.498}\cong 1-0.2236\cong 0.78.$$

The 1x level is the study LoD, chosen based on preliminary testing to identify a concentration yielding roughly 95% detection. This is distinct from the true $LoD_{95}$, which remains unknown and is not estimated in qualitative studies, it is simply assumed to be near the 1x level.[1]

### Sample size determination

The number of replicates required depends critically on the equivalence margin.[43,44] For the primary comparison at 1x ( "hit rate" or p ≅ 0.95, α = 0.05, power = 0.80 to detect equivalence when the true proportions are equal):

I. **±30% equivalence margin:** ~12 replicates per arm/target

II. **±25% equivalence margin:** ~16 replicates per arm/target

III. **±20% equivalence margin:** ~25 replicates per arm/target

IV. **±15% equivalence margin:** ~40 replicates per arm/target

V. **±10% equivalence margin**: ~60 replicates per arm/target

VI. **±7.5% equivalence margin**: ~100 replicates per arm/target

VII. **±5% equivalence margin**: ~200 replicates per arm/target

For pattern verification at 0.5x and 3x, 20 replicates per arm typically suffice, as these are descriptive assessments rather than formal hypothesis tests.[1] Pattern verification in the context of qualitative LoD studies refers to checking that the detection rates at the 0.5x and 3x concentration levels follow expected patterns, while not strictly required, provides confidence that your 1x level (where you're formally testing equivalence) is appropriately calibrated.

### The concept

For qualitative assays, you test at three concentration levels:

I. **0.5x level**: Should show ~40% to 60% detection (both assays detecting roughly half the time), though under the Poisson distribution, can be as high as 78%.

II. **1x level**: Should show ~95% detection (this is where you formally test equivalence)

III. **3x level**: Should show ≥95% to 100% detection (both assays reliably detecting)

The sample size requirement can be calculated using standard formulas for two-proportion equivalence tests, accounting for the TOST procedure's Type I error allocation.[8,43,45]

**Statistical analysis: TOST procedure**

At the 1x level, we define detection counts $X_r$ out of $N_r$ replicates for reference and $X_t$ out of $N_t$ replicates for the test as the population proportions:

$$\pi_r = \frac{X_r}{N_r} \quad (9)$$

$$\pi_t = \frac{X_t}{N_t} \quad (10)$$

Estimated by the observed

$$p_r = \frac{x_r}{n_r} \quad (11)$$

$$p_t = \frac{x_t}{n_t} \quad , \quad (12)$$

respectively.

The population difference is:

$$\Delta_{qual} = \pi_t - \pi_r \quad (13)$$

estimated by

$$D_{qual} = p_t - p_r \quad (14)$$

The population standard error, using the unpooled variance (appropriate for equivalence testing),[46] is:

$$SE\left(\Delta_{qual}\right) = \sqrt{\frac{\pi_r\left(1-\pi_r\right)}{N_r} + \frac{\pi_t\left(1-\pi_t\right)}{N_t}} \quad (15)$$

A 95% confidence interval for $\Delta_{qual}$ is constructed[3,47] as follows:

$$\Delta_{qual} = D_{qual} \pm 1.96.SE\left(D_{qual}\right) \quad (16)$$

Using the Two One-Sided Tests (TOST) procedure,[8,48] equivalence is declared if the entire confidence interval is contained within the equivalence bounds $\left(-\Delta_{qual\ margin}, +\Delta_{qual\ margin}\right)$. This is mathematically equivalent to conducting two one-sided statistical hypothesis tests:

$$H_{01}: \Delta_{qual} \leq -\Delta_{qual\ margin} \ vs.\ H_{11}: \Delta_{qual} > -\Delta_{qual\ margin} \quad (17)$$

$$H_{02}: \Delta_{qual} \geq +\Delta_{qual\ margin} \ vs.\ H_{12}: \Delta_{qual} < +\Delta_{qual\ margin} \quad (18)$$

Equivalence requires rejecting both null hypotheses at a stated significance level, say at $\alpha = 0.05$.[8,19]

**Choice of equivalence margin**

The equivalence margin should be based on clinical considerations, not statistical convenience.[49,50] Common choices include:

I. **±30%**: **Exploratory stringency**. Best suited for early feasibility/JET hand-offs or internal screening to rule out gross sensitivity loss, with small sample sizes (~12 per arm). Not generally appropriate for regulatory equivalence claims; use to confirm expected 0.5×/1×/3× patterns before a tighter study;

II. **±20%**: **Development-phase stringency**. Acceptable when clinical risk is low and the change is remote from target-specific chemistry (e.g., minor buffer/process tweaks). Enables moderate sample sizes (~25 per arm) while still guarding against meaningful degradation; often used for internal bridging/verification;

III. **±10%**: **Standard regulatory stringency.** For many diagnostics submissions when preserving analytical sensitivity is important but not lifesaving. Typical design uses ~60 replicates per arm at 1x to achieve ≅80% power when the true hit rate p ≅0.95, with 0.5x/3x pattern checks. Appropriate for formulation changes that could plausibly affect sensitivity and for target-by-target claims in multiplex assays;

IV. **±7.5%**: **Heightened stringency.** For assays where modest sensitivity loss could impact patient management or public health decisions. Plan ~100 replicates per arm at 1x (plus pattern verification at 0.5x and 3x) to maintain adequate power. Recommended when clinical risk tolerance is low, when prior data suggest marginal performance, or when seeking strong equivalence evidence across multiple targets;

V. **±5%**: **High-consequence stringency.** Reserved for critical use cases (e.g., high-risk pathogens, triage/therapy decisions) or when prior evidence must be confirmed with minimal residual uncertainty. Requires large studies (~200 replicates per arm at 1x) and tight operational control; consider independent replication days and enhanced QC to manage variance and avoid ceiling effects at 1x.

The margin should be **pre-specified** in the study protocol and justified based on assay research and development optimization status or clinical risk assessment if submitting to a regulatory agency.[11,22] Retrospectively widening margins to achieve equivalence is scientifically and ethically inappropriate.[51,52]

**Pattern verification**

The 0.5x and 3x levels serve as pattern checks, ensuring the detection curves have appropriate shape[1,2] and serve as a sanity check for your study design. Expected patterns:

I. **0.5x**: Both assays should show 40-60% detection (or 78% under the Poisson distribution); if one shows <30% or >80%, Your 1x concentration is probably too low or too high, respectively, and most likely you are not actually testing near the true LoD. Investigate whether the 1x level was mis-calibrated;

II. **3x**: Both assays should show ≥95% detection (ideally, 100% at this level); lower rates (<95%) suggest systematic assay problems such that the assay is not reliably detecting.

III. **If patterns are asymmetric**: One assay shows expected patterns but the other doesn't, suggesting they have fundamentally different detection curves - even if they appear equivalent at 1x, they may not be truly comparable.

Gross discordance in patterns (e.g., reference at 55% but test at 30% at the 0.5x level) indicates the assays may not have comparable analytical sensitivity even if the 1x comparison shows equivalence, potentially due to different detection curve slopes.[1]

## Multi-target considerations

### The contemporary reality

Modern molecular diagnostic assays frequently target multiple analytes simultaneously. Sexually transmitted infection (STI) panels may detect 4 to 7 organisms;[25,53] respiratory pathogen panels may target 15 to 20 viruses and bacteria;[26,54] pharmacogenomic assays may interrogate dozens of genetic variants.[27,55] When comparing formulations for such multiplex assays, the question arises: "How should multiple LoD comparisons be handled?"

### Strategic approaches

**Approach 1: Target-by-target analysis (Recommended for regulatory submissions or when handing off from research to assay development)**

The most straightforward and defensible approach is to treat each target independently:[56,57]

I. Design and execute separate LoD comparison studies for each target
II. Apply the same equivalence criteria to each target
III. Report results target-by-target with no pooling or adjustment
IV. Require all targets to demonstrate equivalence for overall formulation approval

This approach is conservative but provides maximum transparency and is readily understood by regulators.[11] It also accommodates targets with genuinely different sensitivities or where formulation changes might affect targets differentially (e.g., due to primer design changes in PCR assays).

**Approach 2: Pooled analysis with target as a stratification factor**

If there is strong reason to believe the formulation change affects all targets similarly, a pooled analysis may be considered:[58,59]

I. Use logistic regression (for qualitative data) or mixed-effects models (for quantitative data) with target as a fixed or random effect
II. Estimate an overall formulation effect while accounting for target-specific baselines
III. Test whether the formulation by target interaction is significant; if not, the pooled estimate may be reported
IV. Still report target-specific results as secondary analyses

This approach gains statistical power by borrowing strength across targets[60] but requires careful justification and should be supplemented with target-specific evaluations to ensure no individual target is masked by pooling.

**Approach 3: Hierarchical decision rules**

A middle ground involves pre-specifying decision rules.[61] For example:

I. **Primary rule**: All targets must show equivalence at the ±10% margin
II. **Secondary rule**: If 1-2 targets fail the ±10% margin but meet ±12.5%, conduct clinical risk assessment
III. **Stopping rule**: If >2 targets fail or any target shows >15% degradation, formulation is rejected

This provides flexibility while maintaining rigor. The key is that rules must be defined *a priori* and scientifically justified, not crafted *a posteriori* to accommodate unexpected results.[51]

### Multiplicity and Type I error

A vigorous debate exists in statistical literature about whether multiplicity adjustments (e.g., Bonferroni correction) are appropriate in equivalence testing contexts.[62,63] Arguments against adjustment include:

I. Equivalence testing is inherently conservative, i.e., erroneously concluding equivalence is less likely than erroneously concluding difference with Null Hypothesis Significance Testing (NHST).[8,19]
II. The goal is protecting against falsely declaring inequivalence, not false equivalence[64]
III. Multiplicity adjustments can make equivalence nearly impossible to demonstrate[65]

Arguments for adjustment include:

I. With many targets, the probability that at least one shows spurious equivalence increases[66]
II. Regulatory consistency favors conservative approaches[11]
III. Family-wise error rate control provides a unified framework[67]

**Our recommendation**: For regulatory submissions, do not apply formal multiplicity adjustments, but require all targets to individually meet the equivalence criterion. For exploratory or internal decision-making, document any adjustments clearly and justify the approach.[56,62]

### Reporting multi-target results

Clear, tabular reporting is essential. We recommend the format following the Table 1 example.

**Table 1** Recommended tabular reporting format

| Target | Reference Pr | Test Pt | Difference Pt - Pr | 95% CI for % Difference | Equivalence Assessment 95% CI within (-10%, +10%)? |
|---|---|---|---|---|---|
| CT | 57/60 (95.0%) | 58/60 (96.7%) | 1.70% | (−6.0, +9.4%) | Yes |
| NG | 56/60 (93.3%) | 52/60 (86.7%) | −6.7% | (−18.2, +4.9%) | No |
| TV | 58/60 (96.7%) | 59/60 (98.3%) | 1.70% | (−5.2, +8.5%) | Yes |
| MG | 55/60 (91.7%) | 57/60 (95.0%) | 3.30% | (−7.1, +13.8%) | No |
| HPV | 59/60 (98.3%) | 54/60 (90.0%) | −8.3% | (−16.9, +0.2%) | No |

CT=Chlamydia trachomatis, NG=Neisseria gonorrhoeae, TV=Trichomonas vaginalis, MG=Mycoplasma genitalium,

HPV=Human Papillomavirus

**Overall conclusion**: Test formulation demonstrates equivalence to reference for 2/5 targets at the ±10% margin. Targets NG, MG, and HPV warrant clinical risk assessment due to point estimates suggesting lower test sensitivity.

## A single level design for a qualitative LoD experiment

### Non-inferiority testing (aka just enough testing – JET): Newcombe hybrid score interval as an alternative to TOST

#### *Conceptual framework: when non-inferiority is appropriate*

While the Two One-Sided Tests (TOST) procedure addresses the fundamental question "Are these assays equivalent?", non-inferiority testing addresses a related but distinct question: "Is the new assay no worse than the reference by more than a pre-specified margin?" This distinction is critical in regulatory and commercial contexts.[1,2] Non-inferiority designs may be utilized for delivering a "good enough" assay with its updated constituents to assay development ready for optimization. This is called Just Enough Testing or JET.

The fundamental difference lies in the hypothesis being tested:

**Equivalence (TOST):** $H_0: |\pi_{test} - \pi_{reference}| \geq \delta$. We reject this null hypothesis if BOTH the lower bound > -δ AND the upper bound < +δ.

**Non-inferiority:** $H_0: \pi_{test} - \pi_{reference} \leq -\delta$ (i.e., the test is not worse than δ). We reject this null hypothesis if the lower confidence bound exceeds -δ.

Non-inferiority testing is appropriate when:

I. **Regulatory pathway permits it:** FDA guidance recognizes non-inferiority designs for certain modifications.[2,3] For LoD studies, non-inferiority may be acceptable for minor reagent changes, alternative instruments, or protocol modifications where full equivalence is not mandated.[4]

II. **Clinical context justifies it:** When clinical and practical considerations support acceptance of a slightly less sensitive assay (within predefined bounds), non-inferiority suffices. For example, if a new buffer formulation is less expensive or more stable, accepting 10-15% lower sensitivity may be reasonable if all detections occur well above the actual LoD.

III. **Efficiency is important:** Non-inferiority testing typically requires smaller sample sizes than equivalence testing, because only one bound (the lower bound) requires comparison to the margin, not both bounds. This advantage grows as margins become stricter.[5]

IV. **Facilitating Efficient R&D Handoffs:** Non-inferiority testing is ideal when transitioning an assay from research to development. If a component is modified (e.g., a new formulation), a full equivalence study is often unnecessary. This is a perfect application for "Just Enough Testing (JET)": using a smaller non-inferiority study to efficiently demonstrate that the modified assay is "good enough" with JET to be handed over to the development team for further optimization.

#### *Why small sample sizes require enhanced methods*

The TOST procedure as typically presented relies on normal approximation to construct confidence intervals for the difference in proportions:

$$\pi_1 - \pi_2 = p_1 - p_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

This standard Wald approach works well when np and n(1-p) are both ≥ 5 for each group.[6] However, in LoD studies:

I. Typical sample sizes are 20 to 60 per arm

II. Qualitative assays near the LoD often show detection probabilities of 60% to 90%, not extreme values

III. The Wald method can undercover (i.e., the actual confidence interval is narrower than the nominal 95%), leading to type I errors[7]

**Example**: For n=21 and p=0.86, we have np(1-p) = 21·(0.86)·(0.14) = 2.53, well below the rule-of-thumb minimum. The normal approximation becomes questionable.

Therefore, two methods to consider for LoD qualitative analysis are based on the Wilson score and Clopper-Pearson exact methodologies. We focus on the Wilson score methodology and leave the Clopper-Pearson exact implementation to Appendix A, B.

The Wilson score interval[47] and its extension to differences, the Newcombe hybrid score interval,[3] address this limitation by:

1. Not assuming a specific distribution shape (binomial, not normal)
2. Using an exact principle (inverting hypothesis tests) rather than approximation
3. Maintaining proper coverage probability even for extreme proportions and small samples[3]

#### *Mathematical formulation*

**Wilson score interval for a single proportion**

For a single proportion p estimated from x successes in n trials, the Wilson Score interval is:

$$L_p = \frac{2x + z^2 - z\sqrt{z^2 + 4x\left(1 - \frac{x}{n}\right)}}{2\left(n + z^2\right)}$$

$$U_p = \frac{2x + z^2 + z\sqrt{z^2 + 4x\left(1 - \frac{x}{n}\right)}}{2\left(n + z^2\right)}$$

where Z is the critical value from the standard normal distribution. For a one-sided non-inferiority test at α=0.05, Z = 1.645 (not 1.96, which would be used for a two-sided test).

**Newcombe hybrid score interval for difference of proportions**

Newcombe's method[3] combines the individual Wilson intervals using a Pythagorean distance approach. The lower confidence limit for the difference ($\pi_t$ - $\pi_r$) is:

$$LCL_{difference} = (p_t - p_r) - \sqrt{(L_t - p_t)^2 + (U_r - p_r)^2}$$

where $L_t$ is the lower Wilson bound for $p_t$ and $U_r$ is the upper Wilson bound for $p_r$.

**Non-inferiority decision rule**

Declare non-inferiority if:

$$LCL_{difference} > -\delta$$

where δ is the pre-specified non-inferiority margin. This compares only the lower bound; there is no requirement for an upper bound.

***Implementation: Step-by-Step***

**Step 1: Define the margin and significance level**

Pre-specify the non-inferiority margin δ before the study. Common choices:

I. δ = 0.10 (10%): Standard choice for many diagnostics; permits test assay detection probability down to (reference - 0.10)

II. δ = 0.15 (15%): More generous; appropriate when other assay properties (cost, speed, stability) provide offsetting benefits

III. δ = 0.30 (30%): Used when LoD comparison is not the primary regulatory concern, or when only pattern equivalence is required, or when implementing the JET methodology

Also pre-specify the significance level, α, say α = 0.05 for a one-sided test (z = 1.645).

**Step 2: Design the experiment**

Follow the same three-level design as for equivalence testing:

I. 0.5x level: ~40-60% expected detection but can be as high as 78% (Poisson)

II. 1x level: ~95% expected detection (primary comparison point)

III. 3x level: ≥95% expected detection

For the 1x level where non-inferiority is formally tested, sample size depends on the margin (α=0.05, 80% power, $p_t = p_r$) as shown in Table 2.

**Table 2** Sample sizes required per arm for non-inferiority testing margins

| Margin | N per arm ($\alpha$=0.05, 80% power, pt = pr) |
|---|---|
| ±10% | ~60 |
| ±15% | ~40 |
| ±20% | ~25 |
| ±30% | ~12 |

Though not strictly required, for the 0.5x and 3x levels, n=20 to 30 per arm typically suffices for pattern verification.

**Step 3: Conduct the assays**

Run replicates of both reference and test assays at the 1x concentration. Record detection/non-detection for each replicate.

**Step 4: Calculate proportions**

$$p_t = \frac{\text{detections}_t}{n_t}, \quad p_r = \frac{\text{detections}_r}{n_r}$$

**Step 5: Calculate Wilson intervals for each assay**

Using the formulas above, calculate $L_t$, $U_t$ for the test assay and $L_r$, $U_r$ for the reference assay.

**Step 6: Calculate LCL for the difference**

Apply the Newcombe formula.

**Step 7: Make a decision**

I. If LCL > -δ: Declare non-inferiority of the test assay

II. If LCL ≤ -δ: Cannot declare non-inferiority; further investigation needed

***Worked example: Qualitative PCR assay comparison***

**Study design:** A clinical laboratory is introducing a new reagent buffer (test) for a qualitative PCR assay. The current buffer (reference) is well-established. The laboratory wants to verify that the new buffer is non-inferior in terms of analytical sensitivity.

**Protocol specifications:**

I. Non-inferiority margin: δ = 0.30 (i.e., new buffer's detection probability can be 30 percentage points lower)

II. Significance level: α = 0.05 (one-sided)

III. Study design: Three-level (0.5x, 1x, 3x)

IV. Sample size at 1x: n = 21 per arm (based on pragmatic resource constraints)

**Experimental results at 1x level:**

I. Reference buffer: 18 detections out of 21 replicates ($p_r$ = 0.857)

II. Test buffer: 16 detections out of 21 replicates ($p_t$ = 0.762)

III. Observed difference: $p_t - p_r$ = -0.095 (9.5 percentage points lower)

**Analysis:**

Table 3 Step-by-step tabular results for non-inferiority testing of buffer example.

| Calculation | Formula/Result |
|---|---|
| z-critical (one-sided) | NORMSINV(1-0.05) = 1.6449 << using Excel formula >> |
| $z^2$ | $(1.6449)^2$ = 2.7055 |
| Wilson $P_r$ Lower | $\frac{36+2.706-1.645.\sqrt{2.706+4.18.\frac{3}{21}}}{2.(21+2.706)}$ =0.6913 |
| Wilson $P_r$ Upper | $\frac{36+2.706+1.645.\sqrt{2.706+4.18.\frac{3}{21}}}{2.(21+2.706)}$ = 0.9414 |
| Wilson $P_t$ Lower | $\frac{36+2.706-1.645.\sqrt{2.706+4.16.\frac{5}{21}}}{2.(21+2.706)}$ = 0.5850 |

| | |
|---|---|
| Wilson $P_t$ Upper | $\frac{36+2.706+1.645.\sqrt{2.706+4.16.\frac{5}{21}}}{2.(21+2.706)} = 0.8790$ |
| **LCL_ difference** | $-0.095-\sqrt{(0.5850-0.762)^2+(0.9414-0.857)^2}$ = -0.095 - 0.196 = -0.291 |
| **Decision** | Is -0.291 > -0.30? **YES → Non-inferiority PASS** |

## Interpretation

"The 95% lower confidence limit for the difference in detection probabilities ($p_t$ - $p_r$) is -0.291. This represents the lowest plausible value for how much worse the test buffer could be. Since -0.291 > -0.30 (our pre-specified margin), we conclude that the test buffer is non-inferior to the reference buffer. In practical terms, the test buffer may be approximately 29 percentage points less sensitive in the worst-case scenario, but our equivalence margin of 30 percentage points comfortably accommodates this."

*Comparison: Non-Inferiority vs. Equivalence (TOST)*

Using the same data (18/21 vs 16/21), let's compare with equivalence testing:

**Equivalence analysis (TOST with ±0.20 margin):**

I. Equivalence requires: Lower bound > -0.20 AND Upper bound < +0.20

II. Our LCL = -0.291 < -0.20 → **FAIL** (lower bound requirement not met)

III. Conclusion: Cannot declare equivalence with ±0.20 margin

**Why the difference:**

I. Equivalence requires BOTH bounds within symmetric limits: impossible when one proportion is notably lower

II. Non-inferiority only requires the lower bound to exceed the negative margin: a weaker requirement

III. For a new assay that is expected to be similar but potentially slightly less sensitive, non-inferiority is realistic; equivalence may be too stringent

**Sample size comparison:**

I. For equivalence with ±0.15 margin and these observed proportions: ~120 per arm needed

II. For non-inferiority with δ=0.30 and these proportions: ~20-25 per arm sufficient

III. Five-fold reduction in study size!

*Advantages and limitations*

**Advantages of newcombe hybrid score interval:**

I. **Small-sample robust:** Maintains proper coverage (actual ≈ nominal 95%) for n as small as 20

II. **Exact-based:** Uses binomial principles rather than normal approximation, valid regardless of sample size

III. **Handles extremes:** Works well for proportions near 0 or 1, common at LoD boundaries

IV. **Valid bounds:** Confidence interval always lies within [0,1]

V. **Conservative:** Wider intervals when sample is small, appropriately protecting against false claims

VI. **Peer-reviewed:** Extensively validated in statistical literature[3]

VII. **Efficiency:** Smaller studies than equivalence when non-inferiority is adequate

**Limitations:**

I. **One-sided vs. two-sided:** Non-inferiority claims are inherently asymmetric; regulators must explicitly accept this framework

II. **Computational complexity:** Requires iterative methods or software; cannot be hand-calculated easily

III. **Less familiar:** Clinical scientists may be unfamiliar with non-inferiority testing compared to simple equivalence

IV. **Regulatory acceptance:** Not all pathways permit non-inferiority; must verify in advance with regulatory body.

**Table 4** When to Choose Each Approach.

| Decision Factor | Use Equivalence (TOST) | Use Non-Inferiority (Newcombe) |
|---|---|---|
| Regulatory requirement | Full equivalence demanded | Non-inferiority pathway available |
| Clinical context | True exchange of assays needed | New assay "at least as good" acceptable |
| Sample size | Resources allow 80-200 per arm | Constrained to <60 per arm |
| Assay type | Well-established comparison | New development vs. reference |
| Symmetry needed | Yes, symmetric bounds | No, only lower bound matters |

**Practical recommendation:** Before study design, consult with the relevant regulatory body (and/or reference guidance documents). Determine whether non-inferiority pathway is acceptable for your specific modification. If so, choose non-inferiority for smaller, more efficient studies (i.e., Research handoffs to Development for optimization). If full equivalence is required, design accordingly with larger sample sizes.

***Regulatory references***

For LoD comparison studies, relevant regulatory guidance includes:

I. **CLSI M49** [Clinical and Laboratory Standards Institute standards] - Describes LoD study designs, recommends three-level approach for qualitative assays

II. **FDA Statistical Guidance** [Reference specific FDA documents on non-inferiority] - Outlines when non-inferiority is acceptable, typical margins for diagnostic studies

III. **ICH E10** [International Council for Harmonisation] - Provides statistical framework for non-inferiority trials in clinical contexts; principles apply to analytical validation

Consult these documents along with your regulatory affairs team before finalizing the study protocol.

***Computational tools***

Manual calculation of Newcombe intervals is tedious; statistical software is recommended using R, SAS, Python or Excel:

I. **R:** PropCIs package, wilson2ci() function
II. **SAS:** PROC FREQ with BINOMIAL statement, EXACT option
III. **Python:** statsmodels.stats.proportion module
IV. **Excel:** JET Calculator (available upon request) automates the calculations for qualitative LoD comparisons

The JET Calculator is particularly useful for laboratory scientists without statistical software access; all formulas are transparent and viewable.

## Results via examples

**Example 1: Quantitative assay comparison**

**Clinical context**: A manufacturer has modified the polymerase enzyme in a quantitative HCV viral load assay and needs to demonstrate that analytical sensitivity is preserved.

**Study design**: Eight concentration levels (1, 2, 4, 8, 16, 32, 64, 128 IU/mL) with 24 replicates per level for both reference and test formulations. Detection defined as Ct ≤ 40.

**Analysis**:

I. Reference $LoD_{95}$: 12.0 IU/mL ($log_{10}$ = 1.079, SE = 0.08)
II. Test $LoD_{95}$: 13.0 IU/mL ($log_{10}$ = 1.114, SE = 0.08)
III. Log difference: 0.035 (95% CI: −0.187, 0.256)
IV. Back-transformed ratio CI: (0.65, 1.80)

**Interpretation**: The 95% confidence interval for the LoD ratio extends from 0.65 to 1.80, indicating the test LoD could be anywhere from 35% lower to 80% higher than the reference. This interval extends well beyond the pre-specified equivalence bounds of [0.80, 1.25], therefore equivalence is not demonstrated. The point estimate (1.08) suggests minimal practical difference, but the study lacks statistical power to definitively demonstrate equivalence.

**Recommendation**: Increase sample size (e.g., 40 replicates per level) or reduce the number of concentration levels while maintaining replicates to gain precision in the LoD estimates.

**Example 2: Qualitative assay comparison (Single target)**

**Clinical Context**: A rapid antigen test manufacturer has changed the membrane material and needs to verify that analytical sensitivity for SARS-CoV-2 detection is maintained.

**Study design**: Three concentration levels with study 1x = 2.0 x $10^4$ copies/mL:

I. 0.5x level (1.0 x $10^4$ copies/mL): 20 replicates per arm
II. 1x level (2.0 x $10^4$ copies/mL): 60 replicates per arm
III. 3x level (6.0 x $10^4$ copies/mL): 20 replicates per arm

**Results at 1x level**:

I. Reference: 57/60 detected (95.0%)
II. Test: 58/60 detected (96.7%)
III. Difference: +1.7%
IV. Standard error: 3.9%
V. 95% CI: (−6.0%, +9.4%)

**Equivalence assessment** (pedagogical examples here margins are determined *a priori*):

I. With ±10% margin: CI (−6.0%, +9.4%) is fully contained within [−10%, +10%] → **Equivalent**
II. With ±5% margin: CI upper bound (+9.4%) exceeds +5% → **Not equivalent**

**Pattern verification**:

I. 0.5x level: Reference 11/20 (55%), Test 10/20 (50%) → appropriate pattern
II. 3x level: Reference 20/20 (100%), Test 20/20 (100%) → appropriate pattern

**Interpretation**: The test formulation demonstrates equivalence to the reference at the clinically meaningful ±10% margin. The detection patterns at 0.5x and 3x are appropriate, suggesting comparable analytical sensitivity curves. If a stricter ±5% margin would have been required, additional testing would have been needed to achieve adequate statistical power.

**Example 3: Multi-target qualitative assay**

**Clinical context**: A multiplex STI panel has undergone buffer reformulation and requires validation across five bacterial targets.

**Study design**: For each of CT (Chlamydia trachomatis), NG (Neisseria gonorrhoeae), TV (Trichomonas vaginalis), MG (Mycoplasma genitalium), and HPV (high-risk types), conduct three-level design with 60 replicates at 1x, 20 replicates at 0.5x and 3x. Note that this table is slightly different from the one presented in an earlier section.

## Results summary (1x level only):

**Table 5** Results summary for 1x level for Example 3.

| Target | Reference | Test | Difference | 95% CI for % Difference | Equivalence Assessment 95% CI within (-10%, +10%)? |
|---|---|---|---|---|---|
| CT | 57/60 (95.0%) | 58/60 (96.7%) | 1.70% | (−6.0, +9.4%) | Yes |
| NG | 56/60 (93.3%) | 56/60 (93.3%) | 0.00% | (−9.5, +9.5%) | Yes |
| TV | 58/60 (96.7%) | 59/60 (98.3%) | 1.70% | (−5.2, +8.5%) | Yes |
| MG | 55/60 (91.7%) | 57/60 (95.0%) | 3.30% | (−7.1, +13.8%) | No |
| HPV | 59/60 (98.3%) | 58/60 (96.7%) | −1.7% | (−8.5, +5.2%) | Yes |

**Overall assessment**: In this multiplex scenario, four of the five targets individually demonstrate equivalence at the ±10% margin. Detection patterns at 0.5x and 3x (not shown) were appropriate for all targets. The reformulated buffer preserves analytical sensitivity across four of the five targets tested.

**Statistical note**: No multiplicity adjustments were applied. Each target was evaluated independently against the pre-specified ±10% criterion. The consistent positive results across targets strengthen confidence that the formulation change has not completely adversely affected analytical sensitivity but must be checked for the MG target (in this example).

## Discussion

### The appropriate framework depends on assay type

This work demonstrates that the correct comparison framework depends fundamentally on whether the assay produces quantitative or qualitative data. This is not merely a technical distinction but reflects different measurement philosophies and analytical capabilities.

**Quantitative assays** yield rich information about detection probability as a continuous function of concentration. This allows estimation of a specific LoD value with confidence intervals, and comparisons focus on whether two numeric LoDs are equivalent. The advantage is precision; the disadvantage is that quantitative assays require sophisticated instrumentation and careful calibration.

**Qualitative assays** sacrifice granular information for simplicity and speed, producing only binary outcomes. This limits what can be inferred so that we cannot estimate a specific LoD, only test whether detection probabilities are equivalent at predefined concentrations. The advantage is practical simplicity; the disadvantage is reduced statistical information per test performed.

### Why overlapping confidence intervals are insufficient

Perhaps the most persistent misunderstanding in LoD comparison is the belief that overlapping confidence intervals for two LoDs constitute evidence of equivalence. This error appears repeatedly in regulatory submissions, journal manuscripts, and laboratory validation reports. It must be emphatically corrected.

The problem is that confidence intervals quantify uncertainty about individual parameters, not about their difference. Consider two LoDs:

I. Reference: 10 IU/mL (95% CI: 7-15 IU/mL)

II. Test: 14 IU/mL (95% CI: 9-21 IU/mL)

These intervals overlap substantially, yet the test LoD might be 40% higher than the reference, a clinically significant degradation. Conversely, confidence intervals for individual LoDs can fail to overlap even when the difference between them is not statistically significant or clinically meaningful, because the standard error of the difference is smaller than the sum of individual standard errors when measurements are positively correlated.

The correct approach is to construct a confidence interval for the *difference* (or ratio) and assess whether this interval lies within equivalence bounds. This directly quantifies uncertainty about the comparison of interest.

### Sample size and statistical power for qualitative assays

For qualitative assays at the 1x level, the required sample size is directly tied to the statistical objective and the pre-specified margin:

**JET (Just enough testing) / Non-inferiority designs:** These designs are statistically efficient and appropriate for internal R&D handoffs, where the goal is to demonstrate an assay is "good enough" for optimization. Power is typically set to 80% for a one-sided non-inferiority margin at a 5% significance level ($\alpha$=0.05):

I. ~12 to 25 replicates per arm for wider margins (e.g., ±20% to ±30%)

II. ~40 replicates per arm for a moderate margin (e.g., ±15%)

**Formal equivalence designs (validation/regulatory):** These designs are used for formal validation or regulatory submissions to demonstrate two-sided equivalence.

I. ~60 replicates per arm provides ~80% power for a standard ±10% equivalence margin.

**High-precision equivalence designs:** These are required for critical assays where sensitivity is paramount.

II. ~100 to 200 replicates per arm are needed for highly stringent ±5% to ±7.5% equivalence margins.

### Sample size and statistical power for quantitative assays

For quantitative assays, power depends on the precision of LoD estimates, which in turn depends on the number of concentration levels, replicates per level, and the steepness of the detection curve. Generally:

III. 6-8 concentration levels with 20-24 replicates per level provide good precision

IV. Steep curves (large $\beta_1$ parameter) yield more precise LoDs

V. Shallow curves require larger sample sizes to achieve equivalent precision

Pilot studies are valuable for estimating the detection curve slope, allowing sample size refinement for the pivotal comparison study.

### Regulatory considerations

Regulatory agencies increasingly scrutinize LoD comparisons when evaluating assay modifications. Key expectations include:

I. **Pre-specification**: Equivalence margins, statistical methods, and decision criteria must be defined before data collection

II. **Justification**: Equivalence margins should be justified based on clinical risk, not chosen for statistical convenience

III. **Completeness**: Both point estimates and confidence intervals must be reported; pattern checks at multiple concentration levels strengthen conclusions

IV. **Transparency**: Negative results or concerning trends should be disclosed and addressed, not obscured by selective reporting

For multi-target assays, regulators typically expect target-by-target demonstration of equivalence, though exceptions may be granted when scientifically justified (e.g., when targets share common detection chemistry and formulation change is remote from target-specific components).

### When equivalence is not demonstrated

What should be done when a LoD comparison study fails to demonstrate equivalence? Several possibilities exist:

I. **Insufficient power**: The study may have been underpowered. If the point estimate suggests equivalence but confidence intervals are too wide, repeating with larger sample size is appropriate.

II. **True degradation**: The formulation change may genuinely affect sensitivity. Root cause investigation is warranted, potentially involving comparison of detection curves, cross-reactivity testing, or interferent studies.

III. **Inappropriate margin**: If the equivalence margin was overly stringent relative to clinical need, it may be appropriate (with proper justification and regulatory consultation) to redefine the margin. This should be exceptional, not routine.

IV. **Acceptance with conditions**: Minor failure to demonstrate equivalence might be acceptable if compensated by other factors (improved specificity, enhanced stability, and cost reduction). This requires thorough clinical risk-benefit analysis and regulatory dialogue.

### Practical recommendations for implementation

Based on extensive experience with LoD comparison studies, we offer the following practical guidance:

**Study planning**:

I. Conduct pilot studies to inform dilution level selection and sample size

II. Pre-specify all analysis methods and equivalence criteria in a detailed protocol

III. Ensure adequate sample volume to allow repeat testing if needed

IV. Plan for independent replication on separate days to assess reproducibility

**Execution**:

I. Randomize run order to eliminate temporal biases

II. Include frequent quality controls to verify assay performance stability

III. Document any deviations from protocol in real time

IV. Maintain strict blinding of operators to reference vs. test conditions

**Analysis**:

I. Report both point estimates and confidence intervals

II. Present data graphically (detection curves for quantitative assays, bar charts with error bars for qualitative assays)

III. Conduct sensitivity analyses (e.g., excluding outliers, alternative CI methods)

IV. Document all analysis code for reproducibility

**Interpretation**:

I. Consider biological plausibility alongside statistical results

II. Assess consistency across concentration levels and replicates

III. Evaluate practical significance, not just statistical significance

IV. Engage with clinical stakeholders to contextualize findings

### Future directions

The field of LoD comparison methodology continues to evolve. Areas of active development include:

I. **Bayesian approaches**: Bayesian inference provides a coherent framework for incorporating prior information (e.g., from earlier formulations) and directly estimating the probability that formulations are equivalent within specified bounds. This may be particularly valuable for multi-target assays where hierarchical models can borrow strength across targets.

II. **Adaptive designs**: Sequential and adaptive designs that allow early stopping (for equivalence or futility) could improve efficiency, particularly for expensive or sample-limited studies.

III. **Machine learning methods**: As assays become more complex (e.g., next-generation sequencing panels with hundreds of targets), machine learning approaches for pattern recognition and anomaly detection may complement traditional statistical methods.

IV. **Harmonization efforts**: International efforts to harmonize LoD comparison standards across regulatory jurisdictions would reduce redundancy and facilitate global market access for diagnostic innovations.[68,69]

## Conclusion

Comparing limits of detection between reference and test conditions requires distinct statistical frameworks depending on assay type and study design. For quantitative assays, direct comparison of numeric LoD estimates using confidence intervals for ratios or differences provides the appropriate evidence base. For qualitative assays, equivalence testing of detection proportions at pre-specified concentration levels, implemented through either the TOST procedure, offers a rigorous and interpretable approach or through using the JET procedure for R&D handoffs when adequate testing is warranted. Critical principles apply regardless of assay type: equivalence margins must be clinically justified and pre-specified; adequate sample sizes are non-negotiable for meaningful conclusions; overlapping confidence intervals are insufficient evidence of equivalence; and complete, transparent reporting builds regulatory confidence and scientific credibility. For multi-target assays, a target-by-target analysis

with independent equivalence assessments provides the most defensible approach, though pooled analyses may be appropriate when properly justified. The choice should be made prospectively based on the biological and analytical characteristics of the assay system and the nature of the formulation change. Correct application of these methods ensures regulatory compliance, scientific rigor, and ultimately, confidence that changes to diagnostic assays preserve the analytical sensitivity upon which patients and clinicians depend. As molecular diagnostics continue to advance in complexity and clinical importance, rigorous analytical validation, including properly executed LoD comparisons, remains the foundation of quality assurance.

## Acknowledgements

The author thanks Allen Hwang (Roche Diagnostics) for raising the important point about the expected detection probability at the 0.5x LoD level and for prompting a more precise treatment of the underlying Poisson assumptions. Any remaining errors or omissions are solely the author's responsibility.

## References

1. Clinical and Laboratory Standards Institute. EP17-A2: Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline – Second Edition. CLSI, Wayne, PA, 2012.
2. Clinical and Laboratory Standards Institute. EP17-Ed2: Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures. 2nd edn. Wayne, PA: Clinical and Laboratory Standards Institute; 2023.
3. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Stat Med*. 1998;17(8):873–890.
4. Forootan A, Sjöback R, Björkman J, et al. Methods to determine limit of detection and limit of quantification in quantitative real-time PCR (qPCR). *Biomol Detect Quantif*. 2017;12:1–6.
5. Oehlert GW. A note on the delta method. *Am Stat*. 1992;46(1):27–29.
6. Fieller EC. Some problems in interval estimation. *J R Stat Soc Series B Stat Methodol*. 1954;16(2):175–185.
7. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993. p. 1–11.
8. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm*. 1987;15(6):657–680.
9. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. *Am Stat.* 2001;55(3):182–186.
10. Cumming G, Finch S. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educ Psychol Meas*. 2001;61(4):532–574.
11. U.S. Food and Drug Administration. *Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests*. Silver Spring, MD: FDA; 2007.
12. European Medicines Agency. Guideline on similar biological medicinal products containing biotechnology-derived proteins as active substance: quality issues (revision 1). London: EMA; 2014. p. 1–7.
13. Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. *Clin Biochem Rev*. 2008;29 Suppl 1(Suppl 1):S49–S52.
14. Jenison R, Yang S, Haeberle E, et al. Interference-based detection of nucleic acid targets on optically coated silicon. *Nat Biotechnol*. 2001;19(1):62–65.
15. Bustin SA, Benes V, Garson JA, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem*. 2009;55(4):611–622.
16. International Conference on Harmonisation. ICH Q2(R2) Validation of Analytical Procedures. Geneva: ICH; 2023.
17. Petti CA, Polage CR, Schreckenberger P. The role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. *J Clin Microbiol*. 2005;43(12):6123–6125.
18. Valenstein PN, Meier FA. Outpatient order accuracy: a College of American Pathologists Q-Probes study of requisition order entry accuracy in 660 institutions. *Arch Pathol Lab Med*. 1999;123(12):1145–1150.
19. Wellek S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. 2nd edn. Boca Raton, FL: Chapman & Hall/CRC; 2010. p. 431.
20. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485.
21. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001;55(1):19–24.
22. International Council for Harmonisation. *ICH E9 Statistical Principles for Clinical Trials*. Geneva: ICH; 1998.
23. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350.
24. Payton ME, Greenstone MH, Schenker N. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *J Insect Sci*. 2003;3:34.
25. Gaydos CA, Van Der Pol B, Jett-Goheen M, et al. Performance of the Cepheid CT/NG Xpert Rapid PCR Test for detection of Chlamydia trachomatis and Neisseria gonorrhoeae. *J Clin Microbiol*. 2013;51(6):1666–1672.
26. Brendish NJ, Malachira AK, Armstrong L, et al. Routine molecular point-of-care testing for respiratory viruses in adults presenting to hospital with acute respiratory illness (ResPOC): a pragmatic, open-label, randomised controlled trial. *Lancet Respir Med*. 2017;5(5):401–411.
27. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther*. 2011;89(3):464–467.
28. Sehn LH, Berry B, Chhanabhai M, et al. The revised International Prognostic Index (R-IPI) is a better predictor of outcome than the standard IPI for patients with diffuse large B-cell lymphoma treated with R-CHOP. *Blood*. 2007;109(5):1857–1861.
29. Walker E, Nowacki AS. Understanding equivalence and noninferiority testing. *J Gen Intern Med*. 2011;26(2):192–196.
30. Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol.* 2000;25(2):169–193.
31. Kubista M, Andrade JM, Bengtsson M, et al. The real-time polymerase chain reaction. *Mol Aspects Med*. 2006;27(2-3):95–125.
32. Hosmer DW Jr, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd edn. Hoboken, NJ: Wiley; 2013. p. 1–518.
33. Agresti A. Categorical Data Analysis. 3rd edn. Hoboken, NJ: Wiley; 2013. p. 1–742.

34. Ver Hoef JM. Who invented the delta method? *Am Stat*. 2012;66(2):124–127.

35. Venzon DJ, Moolgavkar SH. A method for computing profile-likelihood-based confidence intervals. *J R Stat Soc Ser C Appl Stat*. 1988;37(1):87–94.

36. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press; 1997. p. 1–47.

37. Bland JM, Altman DG. The use of transformation when comparing two means. *BMJ*. 1996;312(7039):1153.

38. Franz VH. Ratios: A short guide to confidence limits and proper use. ArXiv. 2007. arXiv:0710.2024.

39. Montgomery DC. *Design and Analysis of Experiments*. 9th edn. Hoboken, NJ: Wiley; 2017. p. 1–196.

40. Burdick RK, LeBlond DJ, Pfahler LB, et al. Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry. New York: Springer; 2017.

41. Box GEP, Hunter WG, Hunter JS. Statistics for Experimenters: Design, Innovation, and Discovery. 2nd edn. Hoboken, NJ: Wiley-Interscience; 2005. p. 1–655.

42. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. 2nd edn. Hoboken, NJ: Wiley; 2011. p. 1–57.

43. Julious SA. Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Stat Med*. 2005;24(17):2745–2764.

44. Chow SC, Shao J, Wang H. *Sample Size Calculations in Clinical Research*. 2nd edn. Boca Raton, FL: Chapman & Hall/CRC; 2008. p. 480.

45. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials*. 1982;3(4):345–353.

46. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall; 1991. p. 624.

47. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci*. 2001;16(2):101–133.

48. Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. *Psychol Bull*. 1993;113(3):553–565.

49. Snapinn SM. Noninferiority trials. *Curr Control Trials Cardiovasc Med*. 2000;1(1):19–21.

50. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Stat Med*. 2003;22(2):169–186.

51. Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med.* 2012;31(4):328–340.

52. Pocock SJ, McMurray JJV, Collier TJ. Making sense of statistics in clinical trial reports: part 1 of a 4-part series on statistics for clinical trials. *J Am Coll Cardiol*. 2015;66(22):2536–2549.

53. Schachter J, Moncada J, Liska S, et al. Nucleic acid amplification tests in the diagnosis of chlamydial and gonococcal infections of the oropharynx and rectum in men who have sex with men. *Sex Transm Dis*. 2008;35(7):637–642.

54. Loeffelholz MJ, Tang YW. Laboratory diagnosis of emerging human coronavirus infections - the state of the art. *Emerg Microbes Infect.* 2020;9(1):747–756.

55. Caudle KE, Klein TE, Hoffman JM, et al. Incorporation of pharmacogenomics into routine clinical practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process. *Curr Drug Metab*. 2014;15(2):209–217.

56. Dmitrienko A, D'Agostino RB Sr. Tutorial in biostatistics: traditional multiplicity adjustment methods in clinical trials. *Stat Med*. 2013;32(29):5172–5218.

57. Hung HMJ, Wang SJ, O'Neill R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *J Biopharm Stat*. 2007;17(6):1201–1210.

58. Fitzmaurice GM, Laird NM, Ware JH. *Applied Longitudinal Analysis*. 2nd edn. Hoboken, NJ: Wiley; 2011.

59. Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. New York: Springer; 2000.

60. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press; 2007.

61. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Stat Med*. 2009;28(4):586–604.

62. Sozu T, Sugimoto T, Hamasaki T. *Sample Size Determination in Clinical Trials with Multiple Endpoints*. New York: Springer; 2015.

63. Westfall PH, Tobias RD, Wolfinger RD. *Multiple Comparisons and Multiple Tests Using SAS*. 2nd edn. Cary, NC: SAS Institute; 2011.

64. Proschan MA, Lan KKG, Wittes JT. Statistical Monitoring of Clinical Trials: A Unified Approach. New York: Springer; 2006.

65. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. *Control Clin Trials*. 2002;23(1):2–14.

66. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57(1):289–300.

67. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Hoboken, NJ: Wiley; 1987. p. 482.

68. Agresti A, Coull BA. Approximate Is Better Than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*. 1998;52(2):119–126.

69. Centers for Disease Control and Prevention (CDC), National Center for Health Statistics. Data Presentation Standards for Proportions. Series 2, Number 175. U.S. Department of Health and Human Services. 2017. p. 1–22.