Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Equivalence Testing Under Privacy Constraints

Savita Pareek[1] | **Luca Insolia**[2] | **Roberto Molinari**[1] | **Stéphane Guerrier**[3]

[1]Department of Mathematics and Statistics, Auburn University, Alabama, United States

[2]Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland

[3]Faculty of Science, University of Geneva, Geneva, Switzerland

Correspondence
Email: robmolinari@auburn.edu

Abstract

Protecting individual privacy is essential across research domains, from socio-economic surveys to big-tech user data. This need is particularly acute in healthcare, where analyses often involve sensitive patient information. A typical example is comparing treatment efficacy across hospitals or ensuring consistency in diagnostic laboratory calibrations, both requiring privacy-preserving statistical procedures. However, standard equivalence testing procedures for differences in proportions or means, commonly used to assess average equivalence, can inadvertently disclose sensitive information. To address this problem, we develop differentially private equivalence testing procedures that rely on simulation-based calibration, as the finite-sample distribution is analytically intractable. Our approach introduces a unified framework, termed DP-TOST, for conducting differentially private equivalence testing of both means and proportions. Through simulation-based calibration and real-world applications, we demonstrate that the proposed method maintains type-I error control at the nominal level and achieves power comparable to its non-private counterpart as the privacy budget or sample size increases, while ensuring strong privacy guarantees. These findings establish a reliable and practical framework for privacy-preserving equivalence testing in high-stakes fields such as healthcare among others.

KEYWORDS

Difference in proportions, Bounded means, Confidence intervals, Equivalence test, Differential Privacy

1 | INTRODUCTION

Equivalence testing provides a statistical framework for demonstrating that two treatments, interventions, or systems yield outcomes sufficiently similar within a clinically meaningful margin [1]. Unlike conventional hypothesis testing, which treats the null hypothesis as no effect, equivalence testing reverses this logic by assuming non-equivalence as the null and concluding equivalence only when evidence indicates otherwise. The widely used two one-sided tests (TOST) procedure operationalizes this approach by constructing a $100(1-2\alpha)\%$ confidence interval for the parameter of interest and verifying that it lies entirely within a pre-specified equivalence margin [2]. Originating in the foundational work of [3] and [4], equivalence testing gained prominence through applications in pharmacokinetics, where it became central to assessing bioequivalence of generic drugs [5,6,7]. Its use now extends across diverse domains, including medicine [8,9,10], psychology [11], economics [12], political science [13], imaging [14], food science [15], and sports science [16]. Typical applications involve comparing treatment response rates, validating diagnostic devices, or analyzing bounded biological outcomes such as CD4 cell counts or symptom severity scores [17]. In pharmacokinetic studies, equivalence is established when the 90% confidence interval for the ratio of geometric means of test and reference formulations (e.g., AUC or C_{max}) lies within the 80–125% range, signifying no clinically meaningful difference [18,19,20,21,2].

Statistical procedures for equivalence testing have been well developed for both proportions and means. For binary outcomes, [22] introduced a score-based interval approach for paired-sample designs, while [23] developed asymptotic and exact procedures for assessing equivalence and noninferiority between two binomial proportions. For continuous outcomes, the TOST procedure remains a standard approach for testing mean equivalence. However, these classical approaches assume unrestricted access to summary statistics such as sample means or proportions, an assumption that raises critical privacy concerns in modern data environments. In fact, in today's data landscape, threats to sensitive information are intensifying due to unauthorized AI model

Abbreviations: TOST, Two One-Sided Tests; DP, Differential Privacy.

training on public datasets, large-scale cyberattacks on healthcare infrastructure, and evolving regulatory frameworks such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation. Recent reports of cyberattacks in healthcare systems[24] and concerns about unconsented AI model training on scraped web data[25] highlight the need for privacy-preserving statistical methods. Even when only aggregated or anonymized data are released, privacy risks persist. In the case of anonymized search logs, users were re-identified by linking query patterns with external sources[26]. The Netflix Prize dataset similarly enabled user de-anonymization through correlations with IMDb profiles[27]. In the biomedical domain,[28] demonstrated that publishing allele frequency summaries from genome-wide association studies could allow adversaries to detect individual participation, prompting the NIH to restrict access to dbGaP summary-level data. These cases highlight the limitations of releasing anonymized or aggregate data (such as means or proportions), often protected using techniques such as k-anonymity, l-diversity, or data masking, which lack formal privacy guarantees and remain vulnerable to re-identification via auxiliary information, thereby motivating the need for rigorous frameworks like differential privacy (DP).

To address the inherent vulnerabilities of conventional anonymization methods, DP has emerged as a leading framework for formal data protection, resolving the paradox of protecting individual data while still enabling useful insights about the population. Introduced by[29] and formally established by[30], DP offers quantifiable guarantees by ensuring that the inclusion or exclusion of any single individual's data has a provably limited effect on the outcome of an analysis, even when arbitrary external information is available. This is typically achieved by adding calibrated random noise to the output of statistical functions, such as sample means or proportions, where the noise magnitude is determined by the function's global sensitivity. Building on this foundation, several recent studies have extended classical inference procedures to the differentially private setting. Finite-sample confidence intervals for means have been developed within the DP framework[31], along with DP-compliant versions of the χ^2 test for goodness-of-fit and independence[32], DP confidence intervals for binomial data that rely on perturbing sufficient statistics followed by post-processing adjustments to ensure valid coverage[33] and for more general categorical data setting[34]. Although there has been advancement in delivering methods to perform uncertainty quantification and classical hypothesis testing in a DP framework, they have not directly addressed equivalence testing, which requires assessing whether a parameter lies within a predefined margin. To the best of the authors' knowledge, the only work close to this area are DP procedures (testers) to check identity and equivalence for discrete distributions, otherwise known as *goodness-of-fit* tests. Indeed, these are based on privatized χ^2 and collision-based statistics, establishing the first formal results for private equivalence testing in the distributional sense[35]. More recently, a DP equivalence tester for continuous distributions under the A_k-norm distance was proposed, employing randomized binning and L_2-based closeness testing[36].

While the few existing DP approaches focus on distributional equivalence, more specifically goodness-of-fit tests, these do not directly support testing parameter equivalence or the construction of DP confidence intervals in parametric settings, both of these being key components for conducting TOST-style inference. A first solution for the latter purpose would consist in adapting existing DP tests e.g.[31,33] but this remains challenging because the addition of privacy noise changes the sampling distribution of the test statistic, invalidating classical critical values. The DP noise interacts with data-dependent sensitivity and variance estimates, so type I error control cannot be guaranteed without recalibration. To retain valid statistical guarantees, one must explicitly account for both sampling and privacy randomness when defining thresholds or confidence intervals. To address this challenge with a flexible solution (i.e. adaptale to different test statistics), we develop a simulation-based framework for *differentially private equivalence testing* (DP-TOST), applicable to statistics with known (finite or asymptotic) distributions such as sample proportions and means. Our method builds upon the findings in[37] which delivers a general-purpose solution to achieve accurate inference under complex statistical settings using simulation-based methods such as indirect inference[38]. More specifically, we adopt this approach by simulating the statistics from their known (asymptotic) distribution and adding DP noise to them: we then employ a moment-matching procedure by comparing these simulated statistics to the original ones and finding the simulation parameters that allow to match the two quantities, thereby recovering solutions consistent with the observed privatized outputs. Using this Monte Carlo sampling approach, we construct valid confidence intervals that maintain DP guarantees while preserving statistical properties such as coverage and type I error control. Based on this new procedure, it is possible to perform equivalence testing across different sensitive domains where patient privacy is paramount. In particular, the utility of our approach is illustrated through extensive simulations and a real-world application to the ACTG175 clinical trial dataset[39], a landmark study in HIV research. Indeed, to test treatment equivalence in this study (as well as others), we should avoid disclosing sensitive information which could reveal whether a patient is infected or not, highlighting the need for the proposed testing procedure.

The remainder of this paper is therefore structured as follows. Section 2.1 and 2.2 introduces the statistical foundations of equivalence testing and DP. Section 3 presents the proposed DP-TOST methodology, including the moment-matching

and simulation framework. Section 4 provides simulation results comparing performance under varying privacy budgets. Section 5 demonstrates our method on ACTG175 data, and Section 6 concludes with future directions and implications for privacy-preserving medical research.

2 | PRELIMINARIES

2.1 | Equivalence Testing

Equivalence testing originated in the context of *average bioequivalence*, where the goal is to demonstrate that two formulations or treatments produce responses that are sufficiently similar within a pre-specified margin of clinical irrelevance[6,7,1]. This framework is primarily designed for assessing average effects, such as differences in population means, and naturally extends to proportions through large-sample approximations based on asymptotic normality[3,2]. The Two-One-Sided Tests (TOST) procedure[1] remains the standard framework for implementing such tests which consider a pre-specified equivalence margin $\Theta_1 := (-c_0, c_0)$, where $c_0 \in R^+$, within which a difference of interest (e.g., difference in proportions or means) is considered close enough to zero to declare two statistics (e.g. treatment outcomes) equivalent. Therefore, the hypotheses tested are the following:

$$H_0 : \theta \in \Theta \setminus \Theta_1, \quad H_1 : \theta \in \Theta_1 := (-c_0, c_0),$$

where Θ represents the entire parameter space for the difference of interest (e.g., $\Theta = [-1, 1]$ for a difference of proportions). The underlying idea of the TOST procedure (Figure 1) is to decompose the composite null hypothesis H_0 into two one-sided hypotheses: $H_{01} : \theta \leq -c_0$ and $H_{02} : \theta \geq c_0$[40]. Equivalence is established when both individual null hypotheses are simultaneously rejected at significance level α, ensuring that the parameter lies strictly within the equivalence bounds.

FIGURE 1 Illustration of hypothesis regions. (a) Classic two-sided null-hypothesis significance test, which seeks to detect any difference from zero. (b) Equivalence test using predefined bounds -c_0 and c_0 to demonstrate practical equivalence when the effect size lies entirely within this region.

The practical implementation of equivalence testing depends on the parameter of interest and the sampling distribution of its estimator. Under standard parametric assumptions, one assumes

$$\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2 \approx \mathcal{N}(\theta, \sigma^2),$$

with an independently estimated variance $\hat{\sigma}^2$ obtained from classical estimators. These assumptions underpin the TOST procedure for assessing mean or proportion equivalence[1,41,2].

Mean differences: For two independent samples of sizes n_1 and n_2, with sample means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations s_1 and s_2, the parameter of interest is $\theta = \mu_1 - \mu_2$. Welch's t-based TOST, which does not assume equal variances, is commonly used[42,43]. The test statistics are

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2 + c_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \qquad t_2 = \frac{\bar{x}_1 - \bar{x}_2 - c_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Equivalence is concluded if $t_1 > t_{\alpha,\mathrm{df}}$ and $t_2 < -t_{\alpha,\mathrm{df}}$, where $t_{\alpha,\mathrm{df}}$ is the upper α–quantile of a t-distribution with Welch–Satterthwaite degrees of freedom.

Differences in proportions: For two independent binomial samples with observed proportions $\hat{p}_1$ and $\hat{p}_2$, let $\theta = p_1 - p_2$. Under large-sample asymptotic normality,

$$\hat{p}_1 - \hat{p}_2 \xrightarrow{d} \mathcal{N}\left(p_1 - p_2,\ \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right),$$

where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$. The corresponding z-statistics are

$$z_1 = \frac{\hat{p}_1 - \hat{p}_2 + c_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \qquad z_2 = \frac{\hat{p}_1 - \hat{p}_2 - c_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Equivalence is established if $z_1 > z_\alpha$ and $z_2 < -z_\alpha$, where z_α denotes the upper α–quantile of the standard normal distribution. These asymptotic formulations perform well for large samples but can be conservative in small-sample regimes, where the normal approximation may overestimate tail probabilities[44,45].

An alternative yet equivalent decision rule is the *Interval Inclusion Principle* (IIP), which provides a conceptually transparent formulation of the TOST approach[46]. At significance level α, equivalence is concluded when the $100(1-2\alpha)\%$ confidence interval for θ lies entirely within the pre-specified equivalence bounds:

$$(-c_0, c_0) \supset CI_{1-2\alpha}(\theta).$$

In the DP setting, we adopt the same IIP but construct $CI_{1-2\alpha}(\theta)$ empirically from the quantiles of the sampling distribution of privatized statistics, as the finite-sample distribution of these test statistics is analytically intractable.

2.2 | Differential Privacy (DP)

DP provides a rigorous and quantifiable framework for protecting individual level information, such as medical diagnoses from health records, income data from census surveys, or user preferences from social media activity, during statistical analysis. As noted in Chapter 1 of[30],

> *"Differential privacy describes a promise, made by a data holder, or curator, to a data subject: 'You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources are available.'"*

This framing emphasizes DP not merely as a mathematical concept, but as a policy-level commitment to the data subject. At its best, DP enables accurate analysis of sensitive data without relying on restricted-access environments or complex data-use agreements. However, as[30] caution, the Fundamental Law of Information Recovery states that "overly accurate answers to too many questions will destroy privacy in a spectacular way." Thus, the aim of DP design is to balance utility and privacy, postponing this trade-off as effectively as possible

Formally, let $\mathcal{M} : \mathcal{X}^n \to \mathcal{R}$ be a randomized algorithm operating on a dataset $\mathbf{X} \in \mathcal{X}^n$, where $\mathcal{X}^n$ denotes the space of databases with n rows. We say that two databases $\mathbf{X}, \mathbf{X}' \in \mathcal{X}^n$ are *neighboring*, denoted $d(\mathbf{X}, \mathbf{X}') = 1$, if they differ in exactly one row. For parameters $\epsilon \geq 0$ and $\delta \in [0, 1]$, the mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if, for all neighboring databases x and x', and for all measurable subsets $S \subseteq \mathcal{R}$,

$$\Pr[\mathcal{M}(\mathbf{X}) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(\mathbf{X}') \in S] + \delta. \tag{1}$$

This definition in eq. (1), introduced by[47], guarantees that the inclusion or exclusion of a single individual has a bounded influence on the output, thereby preserving individual-level privacy. If $\delta = 0$, then $\mathcal{M}$ is said to satisfy ϵ-DP, also referred to as pure DP. Typical parameter choices include small constants $\epsilon \in [0.1, 1]$, and $\delta \leq 1/\text{poly}(n)$ [48].

The need for formal privacy guarantees is underscored by reconstruction attacks.[49] showed that an adversary issuing a sequence of aggregate queries can accurately reconstruct sensitive records when the added noise is insufficient or poorly calibrated. To mitigate such risks, DP mechanisms perturb query outputs using calibrated noise, with the scale determined by the *global sensitivity* of the function f (statistical query of interest, such as a mean or sum), defined as

$$\Delta f = \max_{\mathbf{X},\mathbf{X}'} \|f(\mathbf{X}) - f(\mathbf{X}')\|,$$

where $\|\cdot\|$ represents the a norm (usually the L_1-norm for the sensitivity), and $\mathbf{X}$ and $\mathbf{X}'$ are neighboring datasets differing in one individual's record. This quantity captures the maximum influence a single data point can have on the query outcome.

The most basic and widely used DP algorithm is the additive mechanism[47], which approximates a function $f : \mathcal{X}^n \to \mathbb{R}$ by adding random noise scaled to the function's global sensitivity. This mechanism is defined as

$$\mathcal{M}(\mathbf{X}) = f(\mathbf{X}) + U, \tag{2}$$

where U is a zero-mean random variable with variance proportional to $\Delta f/\epsilon$. When using a Laplace noise, this mechanism ensures $(\epsilon, 0)$-DP by design[30] Theorem 3.6. While the rest of our work uses the Laplace mechanism as the running example, the proposed methodology is compatible with a broader class of additive mechanisms that rely on symmetric, zero-mean noise distributions, including the Gaussian mechanism.

3 | DIFFERENTIALLY PRIVATE TOST

This section presents a unified framework for constructing DP confidence intervals and equivalence tests via a simulated moment-matching approach. This formulation supports valid inference for a wide range of outcomes, including rates, probabilities, and bounded clinical scores, commonly encountered in biomedical and public health applications. In particular, we employ the technique put forward and studied in[37] which takes from the indirect inference framework[38]. More in detail, supposing for simplicity that we are interested in a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ for which we have an estimator $\hat{\boldsymbol{\pi}}$ (we refer to this as an auxiliary estimator) which may be biased and/or even inconsistent with respect to the true parameter (which we denote as $\boldsymbol{\theta}_0$). In this case, under certain regularity conditions, a new moment-matching (minimum distance) estimator can be derived with good statistical properties as follows:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \|\hat{\boldsymbol{\pi}}(\boldsymbol{\theta}_0) - \hat{\boldsymbol{\pi}}^*(\boldsymbol{\theta})\|, \tag{3}$$

where $\hat{\boldsymbol{\pi}}(\boldsymbol{\theta}_0)$ represents the auxiliary estimator from the observed sample (generated from the unobserved $\boldsymbol{\theta}_0$) and $\hat{\boldsymbol{\pi}}^*(\boldsymbol{\theta})$ represents the auxiliary estimator computed on a simulated sample generated from a generic parameter value $\boldsymbol{\theta}$. The latter estimator will therefore depend on a simulated sample (say X^*) whose randomness is determined by a seed value (say ω^*) and for which one can then find a value of $\boldsymbol{\theta}$ that solves the problem in (3). If one generates multiple seeds (say $H \gg 0$) and consequently multiple samples,[37] show that the sequence of estimators $\{\hat{\boldsymbol{\theta}}^h\}_{h=1}^H$ can be used to perform inference using a percentile-based approach on this distribution. In a similar manner, if obtaining a statistic $\hat{\nu} = \varphi(\hat{\boldsymbol{\theta}})$, where $\varphi : \mathbb{R}^p \to \mathbb{R}$ is a continuous function, then the sequence $\{\hat{\nu}^h\}_{h=1}^H$ (where $\hat{\nu}^h = \varphi(\hat{\boldsymbol{\theta}}^h)$) can also be used to perform inference on the true function $\nu_0 = \varphi(\boldsymbol{\theta}_0)$.

In this work our auxiliary estimator $\hat{\boldsymbol{\pi}}$ is the privatized (DP) estimator defined in (2) where the function f is the estimator of the sample proportion or sample mean (with trimmed values) and the noise is calibrated to the sensitivity Δf of each of them respectively. As a result, since we are testing equivalence between two statistics, we will focus on the distribution of the parameter vector $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\theta}_2]$, where $\hat{\theta}_i \in \mathbb{R}$ is the estimator in (3) computed on sample i. More specifically, we will focus on the derived statistic $\hat{\nu} = \hat{\theta}_1 - \hat{\theta}_2$ which constitutes a continuous function of $\hat{\boldsymbol{\theta}}$. In this perspective, we firstly adapt this approach to assessing the equivalence between two population proportions and then extend it to general mean parameters with pre-defined bounds for the data over a fixed interval (l, u) (the latter being imposed to define global sensitivity and enable structured noise calibration under DP). In particular, among others, the use of the above framework allows to easily address more challenging settings including possible different sample sizes between groups. Hereinafter, we refer to this proposed framework (based on (3)) as the DP-TOST approach.

3.1 | DP-TOST for Proportions

To present the case for equivalence of proportions, let us start by focusing on one binary-valued sample that we denote as $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p_{10})$, after which we will then introduce the second binary-valued sample $Y_1, \ldots, Y_m \overset{iid}{\sim} \text{Bernoulli}(p_{20})$ to perform the two-sample test. With this in mind, the sample proportion is given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

To ensure DP we release adequately privatized versions of this statistic through the following additive mechanism:

$$\hat{p}_1 := \bar{X} + U_1,$$

where U_1 is an independent zero-mean random variable with variance proportional to the global sensitivity of the statistic. As mentioned earlier, in this work we will consider $U_1 \overset{iid}{\sim} \text{Laplace}(0, \lambda_1)$ which guarantees pure ϵ-DP and where the noise scale $\lambda_1 = 1/n\epsilon$ is based on the global sensitivity of sample proportion (i.e. $\Delta f = 1/n$). We now have that $\hat{p}_1$ plays the role of $\hat{\pi}(\theta_0)$ in (3) where $\hat{p}_1 = \bar{X}(p_{10}) + U_1$ is the DP proportion computed on the observed sample, with underlying non-private sample proportion $\bar{X}(p_{10})$ (and with p_{10} playing the role of θ_0). Following this, using $\|\cdot\|$ to denote the l_2-norm, we would therefore aim to solve the following matching problem:

$$\check{p}_1 = \underset{p_i \in [0,1]}{\text{argmin}} \, \|\hat{p}_1 - \hat{p}_1^*(p_1)\|,$$

where $\hat{p}_1^*(p_1) = \bar{X}(p_1) + U_1^*$. Given the relatively simple additive structure, we could decide to solve the matching explicitly by finding the value of p_1 such that $\hat{p}_1 = \hat{p}_1^*$. Indeed, expanding the term on the right of the latter equality and rearranging would lead to the solution:

$$\check{p}_1 = F^{-1}(\hat{p}_1 - U_1^*),$$

where F is the cumulative distribution function (CDF) of a Binomial random variable. The problem here is that inverting a discrete CDF gives interval solutions, therefore we use an approximation to find a point solution. More specifically, using the central limit theorem the privatized estimators approximately follow:

$$\hat{p}_1^* \overset{d}{\approx} p_1 + \sqrt{\frac{p_1(1-p_1)}{n}} Z_1 + U_1,$$

where $Z_1 \sim \mathcal{N}(0,1)$ is an independent standard normal variable. Hence we can now represent $\hat{p}_1^*$ through the above approximation and solve the matching explicitly. More in detail, by rewriting the matching problem we want to find the minimum in p_1 of the following function:

$$\check{p}_1 = \underset{p_1 \in [0,1]}{\text{argmin}} \left\| \hat{p}_1 - p_1 - \sqrt{\frac{p_1(1-p_1)}{n}} Z_1^* - U_1^* \right\|. \tag{4}$$

Since we now have an explicit (approximate) form for $\hat{p}_1(p_1)$, we can rewrite this as a quadratic equation and also find explicit solutions for p_1. To do so, we define the following quantities

$$\delta = \frac{Z_1^*}{\sqrt{n}}, \quad \gamma = \delta^2, \quad \Lambda = -4\hat{p}_1^2 + 4\hat{p}_1 + \gamma + 8\hat{p}_1 U_1^* - 4(U_1^*)^2 - 4U_1^*.$$

With these, if $\Lambda \geq 0$, the two candidate roots for the matching problem are given by:

$$\begin{aligned} \check{p}_1^{(1)} &= \frac{-\delta\sqrt{\Lambda} + (2\hat{p}_1 + \gamma - 2U_1^*)}{2(\gamma+1)}, \\ \check{p}_1^{(2)} &= \frac{\delta\sqrt{\Lambda} + (2\hat{p}_1 + \gamma - 2U_1^*)}{2(\gamma+1)}. \end{aligned} \tag{5}$$

Among the roots that lie in the required interval $[0,1]$, we evaluate the objective in (4) and select the solution with the smallest loss, ensuring that the corresponding $\check{p}_1$ allows to best match $\hat{p}_1^*(\check{p}_1)$ to the observed privatized estimator $\hat{p}_1$. By repeating this procedure B times, where each time we produce new draws of the variables Z_1^* and U_1^*, we obtain a collection of reconstructed values $\{\check{p}_1^{(b)}\}_{b=1}^B$ which can be used for inference on p_{10}. All of the above procedure can be applied in the exact same manner for the second sample $\{Y_1, \ldots, Y_m\}$ to obtain $\{\check{p}_2^{(b)}\}_{b=1}^B$. As a result, each pair $(\check{p}_1^{(b)}, \check{p}_2^{(b)})$ can be used to approximate the sampling distribution of $\hat{\nu}^* := \check{p}_1^* - \check{p}_2^*$ by defining $\hat{\nu}^{(b)} = \check{p}_1^{(b)} - \check{p}_2^{(b)}$ and using the sequence $\{\hat{\nu}^{(b)}\}_{b=1}^B$ as the approximate distribution of $\hat{\nu}^*$. The steps involved in generating this distribution are outlined in Algorithm 1 in Appendix A which, once run, can be used for the percentile-approach to deliver CIs as well as to determine p-values for one-sided hypothesis tests (moreover it provides an overview of the resampling strategy when solutions are not valid).

3.2 | DP-TOST for Means

We extend our framework to construct DP equivalence tests for two population means, where the outcome is assumed to be real-valued. In this setting, to guarantee privacy a common approach is to truncate (clamp) the data within pre-specified bounds in order to define global sensitivity and calibrate the magnitude of noise required for DP[30]. More specifically, let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ be independent variables: since the support of Gaussian variables is unbounded, the standard approach is to assume that these variables can be reasonably truncated within *a priori* known bounds, i.e., $X_i \in [a_1, b_1]$ and $Y_j \in [a_2, b_2]$ with constants $a_i, b_i \in \mathbb{R} \,\forall\, i$. Hence, let us denote the truncated samples as $\tilde{X}_i := \min(\max(X_i, a_1), b_1)$ and $\tilde{Y}_j := \min(\max(Y_j, a_2), b_2)$, and define:

$$\mathbf{T}_X = (\bar{X}, s_X), \quad \mathbf{T}_Y = (\bar{Y}, s_Y),$$

where $\bar{X}$ and $\bar{Y}$ are the sample means, and s_X, s_Y are the sample standard deviations of the bounded data. To achieve DP, calibrated noise is independently added to both the sample mean and standard deviation. The noise scale is determined by the global sensitivity of each statistic, which depends on the range $[a_i, b_i]$ and the sample size. The privatized summary statistics are:

$$\widehat{\mathbf{T}}_X = (\bar{X} + U_1,\ s_X + U_2), \quad \widehat{\mathbf{T}}_Y = (\bar{Y} + U_3,\ s_Y + U_4),$$

where $U_1, \ldots, U_4$ are independent zero-mean random variables representing the noise to guarantee DP, with each noise term calibrated using a scaled additive mechanism. Since for this work we consider pure ε-DP, to ensure that the joint release satisfies this requirement (and since the samples are independent) the privacy budget is divided equally between the mean and the standard deviation for each sample. The scale parameter is therefore given by

$$\tau_i = \frac{\Delta f_i}{\varepsilon/2},$$

where $\varepsilon > 0$ denotes the total privacy budget for the DP equivalence test and Δf_i is the global sensitivity of a statistic for sample i. Generally, for data bounded in $[a, b]$, the global sensitivities for sample mean and standard deviation are respectively

$$\Delta f_{\text{mean}} = \frac{b-a}{n} \qquad \text{and} \qquad \Delta f_{\text{sd}} = \frac{b-a}{\sqrt{n-1}}.$$

This calibration guarantees that the privatized statistics $\widehat{\mathbf{T}}_X$ and $\widehat{\mathbf{T}}_Y$ together satisfy ε-DP (other forms of DP can be guaranteed in a similar manner). Now, to recover estimates of the underlying population means from these privatized summaries, taking the random variable X_i as a reference, we solve the following matching problem:

$$\check{\mathbf{T}}_X^* = \underset{\mu_1, \sigma_1}{\operatorname{argmin}} \left\| \widehat{\mathbf{T}}_X - \widehat{\mathbf{T}}_X^*(\mu_1, \sigma_1) \right\|, \tag{6}$$

where $\widehat{\mathbf{T}}_X^*(\mu_1, \sigma_1)$ is the privatized simulated parameter vector defined as

$$\widehat{\mathbf{T}}_X^*(\mu_1, \sigma_1) = \left[\bar{X}^*(\mu_1, \sigma_1),\ \mathrm{s}_X^*(\mu_1, \sigma_1)\right] - [U_1^*, U_2^*],$$

where $\bar{X}^*(\mu_1, \sigma_1)$ and $\mathrm{s}_X^*(\mu_1, \sigma_1)$ are respectively the simulated sample mean and standard deviation implied by the parameters μ_1 and σ_1, while U_1^* and U_2^* are the simulated noise respectively added to the sample mean and standard deviation to guarantee DP. Similarly, for variable Y_i, we would have $\bar{Y}^*(\mu_2, \sigma_2)$ and $\mathrm{s}_Y^*(\mu_2, \sigma_2)$ representing the simulated mean and standard deviation under parameters μ_2 and σ_2, with U_3^* and U_4^* denoting the corresponding privatization noise.

Unlike the case for proportions, and also due to the truncation of the data, no closed-form solutions are available when both the mean and standard deviation are privatized (see discussion in Appendix B). For this problem therefore, for both samples X_i and Y_i, we employ numerical optimization to recover parameter estimates under privacy constraints. The complete procedure for constructing differentially private confidence intervals for bounded means is outlined in Algorithm 2 in Appendix A.

4 | SIMULATION STUDIES

To evaluate the performance of the proposed DP-TOST approaches we conducted simulation studies for both (i) the test for proportions and (ii) test for means. Each setting was examined separately for both the one- and two-sample cases, however in

this section we limit our analysis to the two-sample settings and leave the results for the one-sample settings in Appendix C. The computational setup and corresponding execution times are summarized in Appendix D. For all settings we evaluate the empirical behavior of the tests in terms of type-I error and power. The full simulation setup, including parameter values and design choices, is summarized in Table 1. More specifically, we consider different privacy budgets ($\epsilon = 0.5, 1, 10$) and sample sizes for type-I error ($n = 200, 300, \ldots, 1000$) and power ($n = 250, 500, 1000$) to determine how performance changes when the sample size increases. In particular, for proportions we consider a base proportion value that varies from 0.5 to 0.8 to check how the proposed approach is affected by values closer to the boundaries, whereas for means we consider different variances between samples as well as different forms of truncation of the data to study how robust our approach is with respect to these scenarios. For comparison, we benchmarked DP-TOST method against its non-private counterpart (TOST). More in detail, for proportions we used the pooled z-test, implemented making use of the $(1 - 2\alpha)$ confidence interval corresponding to two one-sided z-tests. For tests on means we applied the Welch two-sample t-test at level $(1 - 2\alpha)$, consistent with the standard TOST procedure. We report both empirical type I error and power across varying levels of the true difference and privacy budget. As highlighted in Table 1, each simulation is based on 4000 Monte Carlo replications, yielding a simulation standard error below 0.7 in all settings.

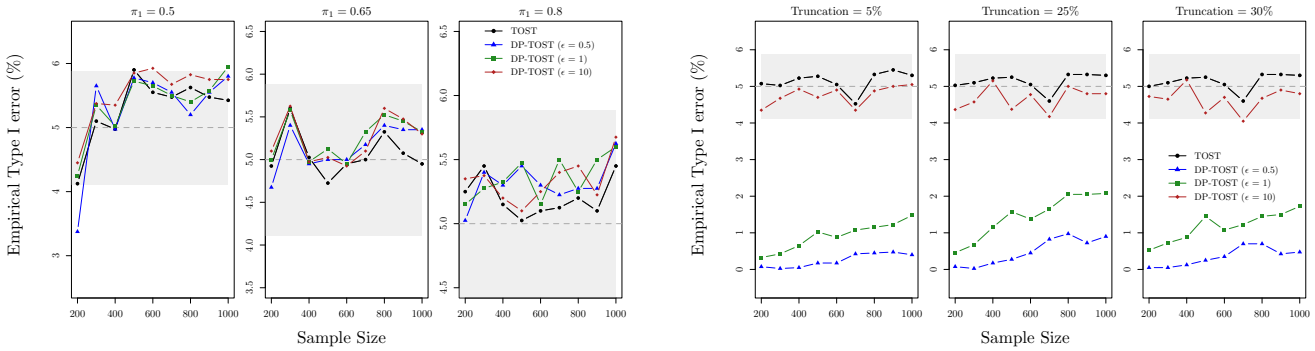
TABLE 1 Simulation settings for two-sample DP-TOST.

Parameter	Specification	Figure(s)
General Settings		
Significance level (α)	0.05	–
DP-TOST replications (B)	4×10^3	–
Monte Carlo simulations (H)	4×10^3	–
Privacy budget (ϵ)	0.5, 1, 10	–
Sample sizes ($n = m$)	200–1000, step 100 (for type I error)	Figure 2
	250, 500, 1000 (for power);	Figure 3
Proportion-based Simulations		
True proportion (π_1)	0.5, 0.65, 0.8	Figures 2a and 3a
True difference ($\pi_1 - \pi_2$)	13 values in [–0.12, 0.12]	–
Equivalence margin (c_0)	0.1	–
Bounded Mean-based Simulations		
True difference ($\mu_1 - \mu_2$)	13 values in [–0.6, 0.6] with $\mu_1 = 3.5$	–
Standard deviations (SD) (σ_1, σ_2)	(1, 1)	Figures 2b and 3b,
Unequal SD	(1, 2)	Figures A.2 and A.3,
Equivalence margin (c_0)	0.5	–
Shared truncation bounds	Identical bounds for both groups ($a_1 = a_2, b_1 = b_2$)	Figures 2b and 3b,
	(i) 5% symmetric: $a_1 = \mu_1 - 2\sigma_1, b_1 = \mu_1 + 2\sigma_1$	
	(ii) 25% symmetric: $a_1 = \mu_1 - 1.15\sigma_1, b_1 = \mu_1 + 1.15\sigma_1$	
	(iii) 30% asymmetric: $a_1 = \mu_1 - 0.84\sigma_1, b_1 = \mu_1 + 1.3\sigma_1$	
Separate truncation bounds	Bounds differ by group (a_i, b_i vary for $i = 1, 2$):	Figures A.2 and A.3
	$a_i = \mu_i - k_a\sigma_i, b_i = \mu_i + k_b\sigma_i$ for $i = 1, 2$,	
	where $(k_a, k_b) \in \{(2.0, 2.0), (1.15, 1.15), (0.84, 1.3)\}$.	

The proposed DP-TOST procedures show a lower probability of rejecting H_0, reflecting their inherent conservativeness under the null and reduced power under the alternative. This behavior results from the additive zero-mean noise used to ensure ϵ-DP, which increases the variance of the test statistic without altering its expectation. Consequently, the DP-TOST is conservative but not liberal, with its performance converging to that of the standard non-private TOST as the privacy budget (ϵ) or sample size increases.

4.1 | Proportions

Figures 2a and 3a summarize the performance of the proposed DP-TOST for equivalence of proportions compared to the standard non-private TOST. Figure 2a shows how the non-private and DP-TOST perform in terms of type-I error across different privacy budgets and sample sizes for the various base proportion values for the first sample. We can observe that, for all privacy budgets, the DP-TOST controls the type-I error near the nominal 5% level across all settings with a performance comparable to the standard TOST (especially for $\epsilon = 1$ and $\epsilon = 10$). As expected, for the smaller privacy budget (i.e. $\epsilon = 0.5$) the DP-TOST shows minor deviations and is slightly conservative in small sample sizes. Indeed, for $\pi = 0.5$, this occurs despite the symmetry of the distribution, a known issue with the normal approximation in small samples[45]. As the sample size increases, the behavior becomes more stable and aligns with the nominal 5% level in all cases. For larger sample sizes the increased stability and alignment of DP-TOST with the standard TOST around the nominal level is confirmed (see Appendix Figures A.1a and A.1b).



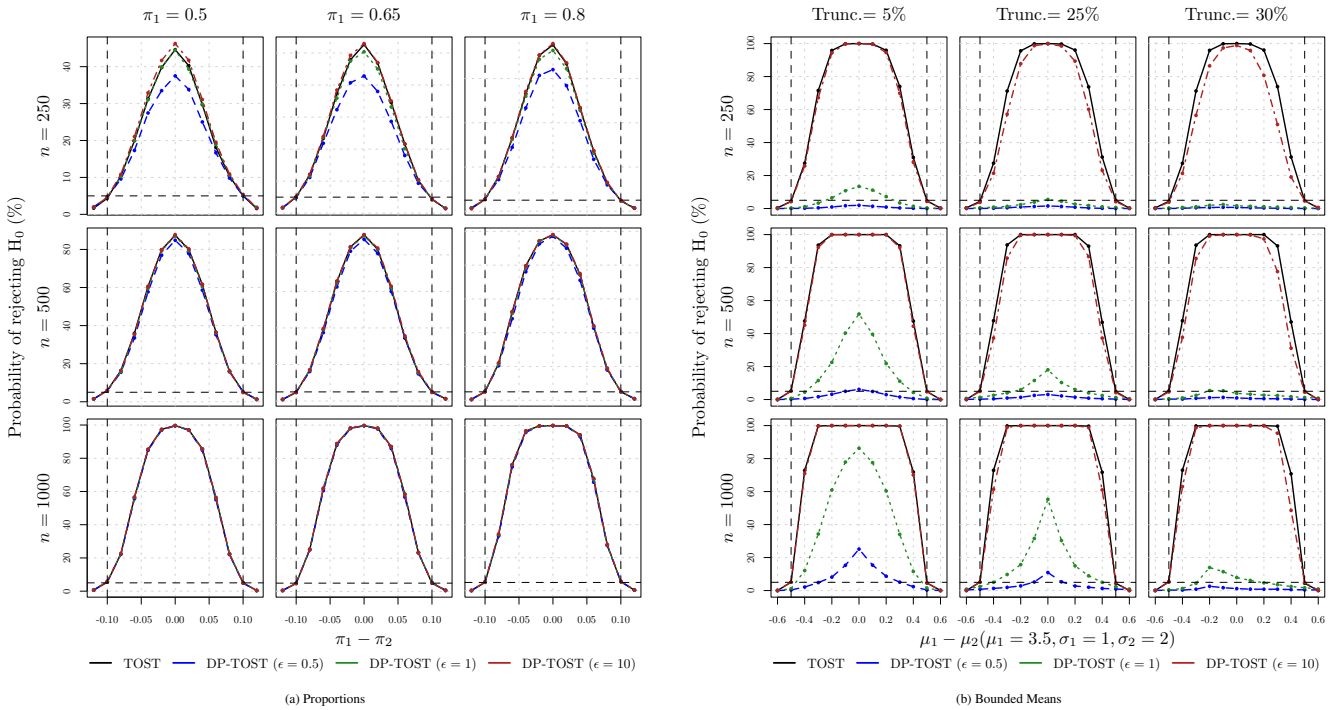
(a) Proportions (b) Bounded Means

FIGURE 2 Empirical Type I error rates for the standard TOST procedure and DP-TOST when assessing equivalence in the difference of (a) two proportions and (b) two means under shared bounds using $H = 4,000$ Monte Carlo replicates. Each column in left panel corresponds to a fixed value of $\pi_1 \in \{0.5, 0.65, 0.8\}$ and in the right panel corresponds to a specific truncation level. Empirical Type I error is calculated as the maximum rejection probability at the two boundary values $\pi_1 - \pi_2 = \pm 0.1$ (left panel) and $\mu_1 - \mu_2 = \pm 0.5$ (right panel). The shaded region denotes the nominal 5% level with simulation error. Curves are shown for the standard TOST and DP-TOST procedures with $\epsilon \in \{0.5, 1, 10\}$.

We now consider the empirical power of these approaches in the same settings which are shown in Figure 3a. As expected, for all approaches the power increases with larger n and, for the DP versions, for higher values of ϵ (which corresponds to lower levels of privacy). Indeed, the DP-TOST closely matches the performance of TOST for $\epsilon = 10$ while still showing good levels of power even for $\epsilon = 1$. At $\epsilon = 0.5$ the power appears to decline more noticeably, particularly for smaller sample sizes and effect sizes near the equivalence margin, reflecting the expected trade-off under strong privacy constraints. The results for the one-proportion equivalence case, presented in Appendix Figures A.4 and A.5, show similar patterns underscoring the robustness of the DP-TOST approach.

4.2 | Bounded Means

Figures 2b and 3b summarize the empirical type-I error and power, respectively, for the equivalence of means under shared bounds across a range of true differences and privacy budgets. Figure 2b indicates that DP-TOST maintains type-I error close to the nominal 5% level for the larger privacy budget $\epsilon = 10$ (which is a reasonable budget for these kind of settings). Moreover, Figure 3b shows that the power of DP-TOST increases with sample size and with larger privacy budgets, approaching the performance of the non-private TOST at $\epsilon = 10$ even for moderate n. The reduced power and conservative (sub-nominal) type I error observed at smaller privacy budgets ($\epsilon \in \{0.5, 1\}$) arise because Laplace noise is added to both the sample mean and standard deviation; the corresponding noise scales grow as ϵ decreases, inflating the variance of the test statistic. Consequently, larger sample sizes are required to recover nominal type I error and achieve competitive power in these settings. Results for the



(a) Proportions (b) Bounded Means

FIGURE 3 Empirical power of the standard TOST and the proposed DP-TOST procedures for testing equivalence in the difference of (a) two proportions and (b) two means under shared bounds using 4,000 Monte Carlo replicates. Two proportions case is evaluated across a range of true differences $(\pi_1 - \pi_2) \in [-0.10, 0.10]$. Two means case is evaluated across a range of true differences $(\mu_1 - \mu_2) \in [-0.6, 0.6]$ with $\mu_1 = 3.5$ and $\sigma_1 = \sigma_2 = 1$. Each column in left panel plot corresponds to a fixed value of $\pi_1 \in \{0.5, 0.65, 0.8\}$ and in right panel plot to a specific truncation level. Each row corresponds to a specific sample size $n \in \{250, 500, 1000\}$. Power is reported under varying privacy levels $\epsilon \in \{0.5, 1, 10\}$.

separate bounds setting, which follow similar trends with slightly reduced power under stricter truncation and tighter privacy constraints, are presented in Appendix Figures A.2 and A.3. Additional results for the one-sample mean case are also included in Appendix C (see Figures A.6 and A.7). In particular, in Figure A.7, for smaller privacy budgets the type-I error is below the nominal level at low truncation but reaches nominal as truncation and sample sizes increase. Although possibly counterintuitive, the larger truncation reduces sensitivity and consequent DP noise, improving calibration despite bias. In the two-sample test, the test statistic is based on the difference of sample means, $\hat{\mu}_1 - \hat{\mu}_2$; under the null, common bias from truncation cancels out, keeping type-I error stable.This interplay between bias, noise magnitude, and variance explains the effect, consistent with DP theory[32].

5 | CASE STUDY

To showcase the DP-TOST on real-world applications we utilize the "ACTG175" dataset, a publicly available HIV-1 clinical trial dataset consisting in a double-blind study to evaluate different treatments in infected adults whose CD4 cell counts ranged from 200 to 500 per cubic millimeter[50]. More in detail, the dataset contains 2,139 observations across 27 variables, including demographic characteristics, baseline clinical markers (e.g., CD4 counts), and treatment outcomes collected from four randomized arms comparing these different antiretroviral therapies. Based on this overview, it is straightforward to understand that this data can be extremely sensitive in terms of patient confidentiality requirements. Indeed, aside from the sensitivity of variables such as ethnicity and race, it is clear that simply the possibility of being identified as an HIV-1 infected adult can be an extremely problematic privacy breach for patient confidentiality. Prior work has in fact shown that even de-identified or aggregated health data can be vulnerable to re-identification through auxiliary information[51,52,53]. Also, one can imagine the possibility of different clinics applying different HIV-1 treatments which then need to be compared among each other for equivalence. In this perspective, the availability of a DP approach to assess equivalence of treatments, for example in terms of

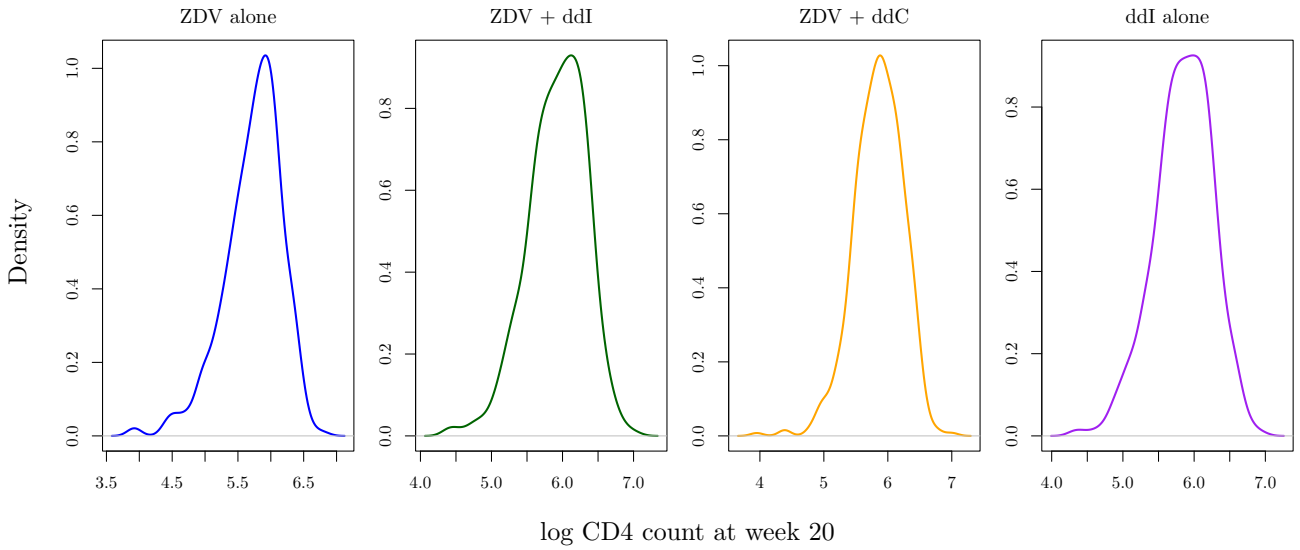


FIGURE 4 Density estimates of log-transformed CD4 T-cell counts at week 20 by treatment arm. Distributions are shown separately for ZDV alone, ZDV + ddI, ZDV + ddC, and ddI alone.

drop-out rates or in terms of CD4 cell counts, can be of considerable importance to perform these tests without compromising patient privacy.

TABLE 2 Descriptive statistics by treatment arm for log-transformed CD4 counts at week 20 ($CD4_{20}$) and proportion of treatment discontinuation before week 96 (Off-Treat). "SD" stands for standard deviation.

Treatment Arm	Sample Size	Mean log($CD4_{20}$)	SD log($CD4_{20}$)	Proportion Off-Treat
ZDV alone	532	5.78	0.34	0.41
ZDV + ddI	522	5.94	0.37	0.33
ZDV + ddC	524	5.87	0.34	0.39
ddI alone	561	5.87	0.36	0.33

Therefore, for our empirical evaluation we focus on two outcomes: (i) the proportion of participants who discontinued treatment before 96 weeks (Off-Treat); and (ii) the CD4 T-cell count at 20 ± 5 weeks ($CD4_{20}$). Each of these outcomes is measured in four distinct treatment arms: (a) 600 mg of zidovudine (ZDV); (b) 600 mg of ZDV plus 400 mg of didanosine (ddI); (c) 600 mg of ZDV plus 2.25 mg of zalcitabine (ddC); or (d) 400 mg of ddI alone see[50]. Summary statistics for the selected variables over these four different arms are provided in Table 2, where the logarithm of $CD4_{20}$ was taken to closer approximate the Normal distribution (see Figure 4 which suggests that this approximation can be considered reasonable).

For the analysis of mean CD4 T-cell counts at week 20, data were truncated using log-transformed values within clinically plausible bounds of log(100) to log(1500)[54], resulting in approximately 1–3% truncation. Consistent with simulation findings, smaller privacy budgets ($\epsilon < 1$) exhibited highly conservative behavior and limited power. Accordingly, we focus on ϵ values ranging from 1 to 3 for mean comparisons and from 0.1 to 2 for proportion comparisons, each in increments of 0.1. All tests are conducted at a significance level of $\alpha = 0.05$, with equivalence region widths of $c_0 = 0.1$ for Off-Treat proportions and $c_0 = \log(1.10)$ for $CD4_{20}$ means. To ensure ϵ-DP, Laplace noise is added according to the specified privacy budgets.

To evaluate robustness under ϵ-DP, we study the average behavior of DP-TOST by repeatedly applying the privacy mechanism to the original trial data. In a realistic application, only a single privatized release of the sufficient statistics would be available, yielding one noisy confidence interval and a single equivalence decision. In our study, however, we generated 500 perturbed realizations of the sufficient statistics by adding Laplace noise calibrated to each privacy budget. Each replicate represents a potential privatized release in practice, and averaging across replicates provided an estimate of the empirical probability of

rejecting the null hypothesis H_0 (in favor of equivalence). This design ensured that results reflected the expected operating characteristics of DP-TOST rather than artifacts of a single noisy draw.

Results for proportion comparisons (fig. 5) and mean comparisons (fig. 6) show consistent patterns. When the benchmark TOST declared equivalence (e.g., Arm 2 vs 3 in the means data and Arm 0 vs 2 or Arm 1 vs 3 in the proportions data), the DP-TOST rejection probability approached 1 as the privacy budget ϵ increased. Conversely, when the benchmark TOST indicated non-equivalence (e.g., Arm 0 vs 1, 0 vs 2, 0 vs 3 and 1 vs 3 in the means data and Arm 0 vs 1 or Arm 0 vs 3 and 2 vs 3 in the proportions data), the DP-TOST rejection probability converged to 0 with increasing ϵ, reflecting its asymptotic agreement with the non-private decision rule. Overall, ZDV monotherapy (Arm 0) was the least comparable, whereas ddI and ddC-based regimens, alone or combined with ZDV, showed similar $CD4_{20}$ means and Off-Treat proportions under both TOST and DP-TOST analyses.

TABLE 3 Benchmark non-private TOST results for two-sample proportion and mean comparisons. A **TRUE** decision indicates that the 90% confidence interval lies entirely within the specified equivalence margin $(-c_0, c_0)$.

	Proportions (c_0 = 0.10)		**Means** (c_0 = log(1.10))	
Comparison	90% CI	Decision	90% CI	Decision
Arm 0 vs 1	(0.024, 0.122)	FALSE	(-0.232, -0.145)	FALSE
Arm 0 vs 2	(-0.029, 0.070)	TRUE	(-0.161, -0.078)	FALSE
Arm 0 vs 3	(0.030, 0.126)	FALSE	(-0.156, -0.071)	FALSE
Arm 1 vs 2	(-0.101, -0.003)	FALSE	(0.028, 0.109)	FALSE
Arm 1 vs 3	(-0.042, 0.052)	TRUE	(0.033, 0.116)	FALSE
Arm 2 vs 3	(0.010, 0.105)	FALSE	(-0.034, 0.045)	TRUE

Note: These non-private TOST results serve as the benchmark reference for evaluating the DP-TOST procedure. A TRUE decision indicates that the 90% confidence interval lies entirely within the equivalence margin. The Arm 1 vs 2 comparison is excluded from DP-TOST analysis because its 90% confidence interval for proportions lies at the equivalence margin, causing unstable rejection probabilities.

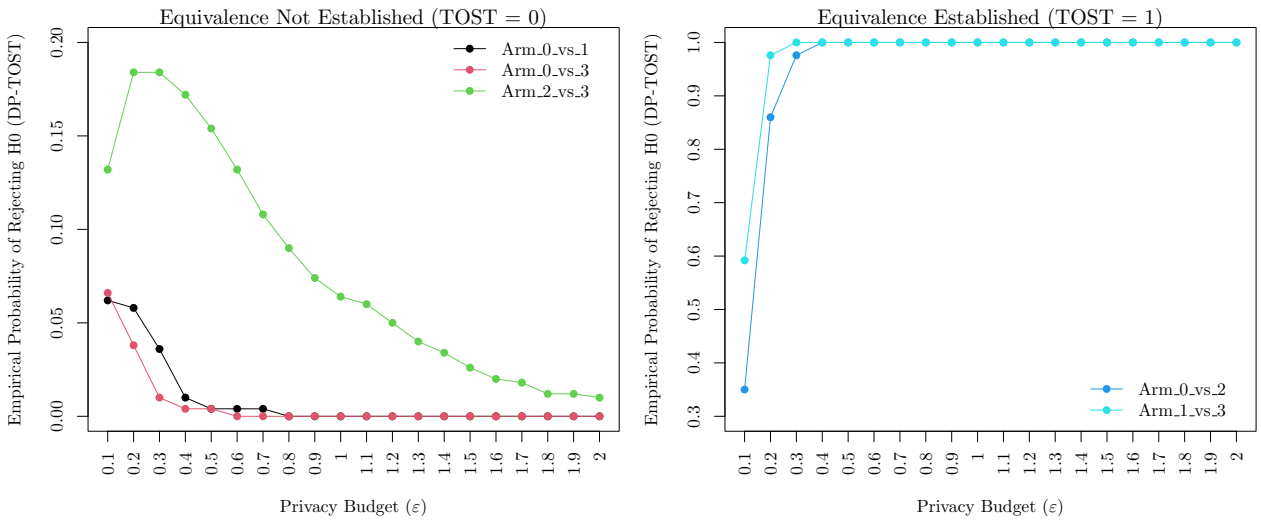


FIGURE 5 Empirical probability of rejecting H_0 under the DP-TOST procedure for two-sample proportion comparisons across varying privacy budgets ($\epsilon \in [0.1, 2]$) and $\alpha = 0.05$. Each curve corresponds to a specific arm comparison, with rejection probabilities averaged over 500 Monte Carlo replicates per ϵ. As ϵ increases, the DP-TOST results converge to their non-private TOST counterparts (see Table 3).

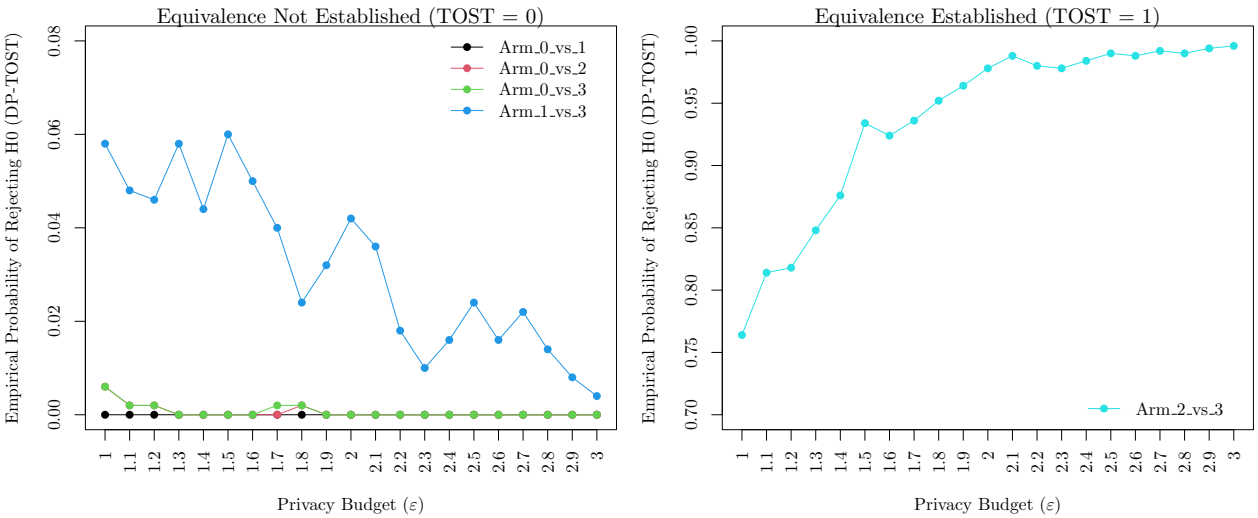


FIGURE 6 Empirical probability of rejecting H_0 under the DP-TOST procedure for two-sample mean comparisons across varying privacy budgets ($\epsilon \in [1, 3]$) and $\alpha = 0.05$. Each curve corresponds to a specific arm comparison, with rejection probabilities averaged over 500 Monte Carlo replicates per ϵ. As ϵ increases, the DP-TOST results converge to their non-private TOST counterparts (see Table 3).

6 | CONCLUSIONS

This paper introduces a unified framework for conducting equivalence testing under formal privacy guarantees. We develop DP analogues of the widely used TOST procedure for both differences in proportions and bounded means by constructing confidence intervals via simulation-based matching methods. Our approach offers statistical inference while adhering to ϵ-DP constraints, a critical requirement in modern biomedical and healthcare applications[30,55]. Through extensive simulation studies, while losing some performance for smaller samples and privacy budgets (as expected in DP scenarios), we demonstrate that the proposed DP-TOST procedures achieve acceptable empirical performance overall, maintaining type-I error control and competitive power relative to their non-private counterparts, particularly under common privacy budgets (e.g., $\epsilon \geq 1$). The methods remain reasonably robust across a variety of sample sizes and privacy levels, and perform comparably even in scenarios involving asymmetric truncation or separate bounds. Our real-data application to the ACTG175 HIV clinical trial illustrates the practical utility of the method. Equivalence conclusions obtained under DP-TOST generally mirror those from classical TOST, confirming the method's reliability in practice for the considered settings. The observed widening of confidence intervals under stronger privacy constraints (e.g., $\epsilon = 0.5, 1$) is expected but does not appear to considerably alter inference in most clinically relevant comparisons. Overall, these results indicate that DP-TOST provides the required balance between privacy and statistical validity, supporting its use in privacy-sensitive applications, such as comparing response or event rates in clinical trials.

This work lays the foundations for extending equivalence testing under differential privacy to more complex and clinically relevant settings. Indeed, the simulation-based nature of the proposed approach allows it to be adaptable to more general statistical procedures. These include procedures addressing outcomes such as ordinal or skewed clinical endpoints, longitudinal data structures with repeated measures over time, and multi-arm trial designs commonly encountered in comparative effectiveness research. Future work may also incorporate adaptive privacy mechanisms that balance utility and confidentiality based on data sensitivity, as well as alternative test statistics that go beyond mean-based summaries, such as rank-based or survival-based measures, to accommodate diverse outcome types. Multivariate formulations (e.g., $\hat{\boldsymbol{\theta}} \sim TN_K(\boldsymbol{\theta}, \sigma^2, a, b)$) may further enable applications in settings where equivalence testing is routinely applied, including bioequivalence, safety monitoring, and health outcomes research[41,56]. Overall, these developments highlight the feasibility of conducting statistically rigorous and privacy-preserving equivalence analyses in biomedical research, where protection of individual-level data remains both an ethical obligation and a regulatory requirement.

References

1. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics.* 1987;15(6):657–680.

2. Wellek S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Boca Raton, FL: CRC Press. 2nd ed., 2010.

3. Lehmann EL. *Testing Statistical Hypotheses*. New York: Wiley, 1959.

4. Bondy WH. A test of an experimental hypothesis of negligible difference between means. *The American Statistician.* 1969;23(5):28–30.

5. Metzler C. Bioavailability - a problem in equivalence. *Journal of Pharmaceutical Sciences.* 1974;30:309–317.

6. Westlake WJ. Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences.* 1972;61(8):1340–1341.

7. Westlake WJ. Symmetrical confidence intervals for bioequivalence trials. *Biometrics.* 1976;32:741–744.

8. Wehrle F, Bartal T, Adams M, et al. Similarities and differences in the neurodevelopmental outcome of children with congenital heart disease and children born very preterm at school entry. *The Journal of Pediatrics.* 2022;250:29–37.e1.

9. Sansone P, Giaccari L, Aurilio C, et al. Comparative efficacy of Tapentadol versus Tapentadol plus Duloxetine in patients with chemotherapy-induced peripheral neuropathy. *Cancers.* 2022;14:4002.

10. Branscheidt M, Ejaz N, Xu J, et al. No evidence for motor-recovery-related cortical connectivity changes after stroke using resting-state fMRI. *Journal of Neurophysiology.* 2022;127:637–650.

11. Lakens D, Scheel A, Isager P. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science.* 2018;1:259–269.

12. Feri F, Giannetti C, Guarnieri P. Risk-taking for others: An experiment on the role of moral discussion. *Journal of Behavioral and Experimental Finance.* 2023;37:100735.

13. Aggarwal M, Allen J, Coppock A, et al. A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout. *Nature Human Behaviour.* 2023:1–10.

14. Sureshkumar H, Xu R, Erukulla N, Wadhwa A, Zhao L. "Snap on" or not? A validation on the measurement tool in a virtual reality application. *Journal of Digital Imaging.* 2022;35:692–703.

15. Meyners M. Equivalence Tests - A Review. *Food Quality and Preference.* 2012;26:231–245.

16. Mazzolari R, Porcelli S, Bishop D, Lakens D. Myths and methodologies: The use of equivalence and non-inferiority tests for interventional studies in exercise physiology and sport science. *Experimental Physiology.* 2022;107:201–212.

17. Möllenhoff K, Dette H, Bretz F. Equivalence Tests for Binary Efficacy-Toxicity Responses. *arXiv preprint arXiv:1910.08769.* 2019.

18. U.S. Food and Drug Administration . Bioequivalence Studies with Pharmacokinetic Endpoints for Drugs Submitted Under an ANDA. tech. rep., Center for Drug Evaluation and Research (CDER); Silver Spring, MD, USA: 2021. Accessed: 2025-10-17.

19. European Medicines Agency . Guideline on the Investigation of Bioequivalence. tech. rep., Committee for Medicinal Products for Human Use (CHMP); London, UK: 2010. Accessed: 2025-10-17.

20. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) . M13A: Bioequivalence for Immediate-Release Solid Oral Dosage Forms. tech. rep., ICH; Geneva, Switzerland: 2024. Accessed: 2025-10-17.

21. World Health Organization . Annex 6: Multisource (Generic) Pharmaceutical Products: Guidelines on Registration Requirements to Establish Interchangeability. Tech. Rep. 992, World Health Organization; Geneva, Switzerland: 2017. Accessed: 2025-10-17.

22. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine.* 1998;17(8):891–908.

23. Chen JJ, Tsong Y, Kang S. Tests for Equivalence or Noninferiority Between Two Proportions. *Drug Information Journal.* 2000;34:569–578.

24. ElSayed Z, Abdelgawad A, Elsayed N. Cybersecurity and Frequent Cyber Attacks on IoT Devices in Healthcare: Issues and Solutions. *arXiv preprint arXiv:2501.11250.* 2025.

25. Solove D, Hartzog W. The Great Scrape: The Clash Between Scraping and Privacy. *California Law Review.* 2025;113:1521–1580.

26. Gotz M, Machanavajjhala A, Wang G, Xiao X, Gehrke J. Publishing search logs - a comparative study of privacy guarantees. *IEEE transactions on knowledge and data engineering.* 2011;24(3):520–532.

27. Narayanan A, Shmatikov V. Robust De-anonymization of Large Sparse Datasets. In: Proceedings of the IEEE Symposium on Security and Privacy. 2008; Oakland, CA, USA:111–125.

28. Homer N, Szelinger S, Redman M, et al. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genetics.* 2008;4(8).

29. Dwork C. Differential Privacy. In: Bugliesi M, Preneel B, Sassone V, Wegener I., eds. *Automata, Languages and Programming*. vol 4052 of *Lecture Notes in Computer Science*. Springer. 2006; Berlin, Heidelberg:1–12

30. Dwork C, Roth A. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science.* 2013;9(3–4):1-277.

31. Karwa V, Vadhan S. Finite Sample Differentially Private Confidence Intervals. In: Innovations in Theoretical Computer Science Conference (ITCS). Schloss Dagstuhl–Leibniz-Zentrum für Informatik 2018; Dagstuhl, Germany:44:1–44:9

32. Gaboardi M, Lim H, Rogers R, Vadhan S. Differentially Private Chi-Squared Hypothesis Testing: Goodness of Fit and Independence Testing. In: Proceedings of The 33rd International Conference on Machine Learning, PMLR. 2016:2111–2120.

33. Awan JA, Slavković A. Differentially Private Inference for Binomial Data. *Journal of Privacy and Confidentiality.* 2020;10(1).

34. Romanus OM, Boulaguiem Y, Molinari R. Fiducial Matching: Differentially Private Inference for Categorical Data. *arXiv preprint arXiv:2507.11762.* 2025.

35. Aliakbarpour M, Diakonikolas I, Rubinfeld R. Differentially Private Identity and Equivalence Testing of Discrete Distributions. In: International Conference on Machine Learning (ICML). PMLR 2018; Stockholm, Sweden:169–178.

36. Omer D, Sheffet O. Differentially Private Equivalence Testing for Continuous Distributions and Applications. In: Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS). 2024; Vancouver, Canada.

37. Orso S, Karemera M, Victoria-Feser MP, Guerrier S. An accurate percentile method for parametric inference based on asymptotically biased estimators. *arXiv preprint.* 2024.

38. Gourieroux C, Monfort A, Renault E. Indirect inference. *Journal of applied econometrics.* 1993;8(S1):S85–S118.

39. Hammer SM, Katzenstein DA, Hughes MD, et al. A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with CD4 Cell Counts from 200 to 500 per Cubic Millimeter. *New England Journal of Medicine.* 1996;335(15):1081–1090.

40. Lakens D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science.* 2017;8:355–362.

41. Hauschke D, Steinijans V, Pigeot I. *Bioequivalence Studies in Drug Development: Methods and Applications*. Chichester, UK: John Wiley & Sons, Ltd, 2007.

42. Welch BL. The generalization of Student's problem when several different population variances are involved. *Biometrika.* 1947;34(1-2):28–35.

43. NCSS, LLC . *Two-Sample T-Tests for Equivalence Allowing Unequal Variance*. PASS Sample Size Software; Kaysville, UT: 2025. Accessed: 2025-10-17.

44. Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. *Statistical Science.* 2001;16(2):101–133.

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

46. Berger R, Hsu J. Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets. *Statistical Science.* 1996;11(4):283–319.

47. Dwork C, McSherry F, Nissim K, Smith A. Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi S, Rabin T., eds. *Theory of Cryptography Conference*. vol 3876 of *Lecture Notes in Computer Science*. Springer. 2006; Berlin, Heidelberg:265–284

48. Alabi D, Vadhan S. Hypothesis Testing for Differentially Private Linear Regression. In: Advances in Neural Information Processing Systems 35 (NeurIPS). 2022:14196–14209.

49. Dinur I, Nissim K. Revealing information while preserving privacy. In: PODS '03: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2003:202–210.

50. Hammer SM, Katzenstein DA, Hughes MD, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine.* 1996;335(15):1081–1090.

51. Terry NP. Protecting patient privacy in the age of big data. *UMKC Law Review.* 2012;81:385–415.

52. Masood I, Wang Y, Daud A, Aljohani NR, Dawood H. Towards Smart Healthcare: Patient Data Privacy and Security in Sensor-Cloud Infrastructure. *Wireless Communications and Mobile Computing.* 2018;2018.

53. Xia W, Liu W, Wan Z, et al. Enabling realistic health data re-identification risk assessment through adversarial modeling. *Journal of the American Medical Informatics Association.* 2021;28(4):744–752.

54. Battistini Garcia SA, Zubair M, Guzman N. CD4 Cell Count and HIV. In: StatPearls [Internet]. StatPearls Publishing 2025; Treasure Island (FL).

55. Liu W, Zhang Y, Yang H, Meng Q. A Survey on Differential Privacy for Medical Data Analysis. *Annals of Data Science.* 2023:1–15.

56. Hoffelder T, Gössl R, Wellek S. Multivariate equivalence tests for use in pharmaceutical development. *Journal of Biopharmaceutical Statistics.* 2015;25(3):417–437.

57. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2024.

58. AUHPC . Auburn University High Performance Computing (HPC) Cluster. https://hpc.auburn.edu; 2024. Accessed: October 2025.

□

APPENDIX

A ALGORITHMS

In this appendix, we describe the steps to implement the proposed DP-TOST procedures for proportions and means. Algorithm 1 presents the Monte Carlo procedure for proportions under ϵ-DP, while Algorithm 2 describes the corresponding algorithm for bounded means using noise calibration and optimization-based reconstruction.

Algorithm 1 DP-TOST for Proportions

Input: Privatized estimates $\hat{p}_1, \hat{p}_2$; sample sizes n, m; privacy budget ϵ; number of Monte Carlo replicates B; significance level α; maximum resampling attempts K
Output: A sequence $\{\hat{\nu}^{(b)}\}_{b=1}^{B}$
for $i = 1, \ldots, B$ **do**
  Generate $(Z_1^{(b)}, Z_2^{(b)}) \overset{iid}{\sim} N(0,1)$
  Generate $U_1^{(b)}$ and $U_2^{(b)}$ independently drawn from Laplace$(0, 1/n\epsilon)$ and Laplace$(0, 1/m\epsilon)$ respectively
  Compute candidate solutions $(\check{p}_1^{(b)}, \check{p}_2^{(b)})$ using eq. (5)
  if at least one solution in $(\check{p}_1^{(b)}, \check{p}_2^{(b)})$ is invalid (e.g., NA or outside $[0,1]$) **then**
    $k = 0$
    while $k \leq K$ **do**
      Repeat Steps 4–6
      if Solutions $(\check{p}_1^{(b)}, \check{p}_2^{(b)})$ are both valid **then return** $(\check{p}_1^{(b)}, \check{p}_2^{(b)})$
      else
        $k = k + 1$
      end if
    end while
  end if
  Compute $\hat{\nu}^{(b)} = \check{p}_1^{(b)} - \check{p}_2^{(b)}$
end for
return $\{\hat{\nu}^{(b)}\}_{b=1}^{B}$

Algorithm 2 DP-TOST for Bounded Means

Input: Privatized estimates $(\hat{\mu}_1, \hat{\sigma}_1)$, $(\hat{\mu}_2, \hat{\sigma}_2)$; sample sizes n, m; data truncation bounds $[l_1, u_1]$, $[l_2, u_2]$; privacy budget ϵ; number of Monte Carlo replicates B; significance level α.
Output: A sequence $\{\hat{\nu}^{(b)}\}_{b=1}^{B}$
for $i = 1, \ldots, B$ **do**
  Generate $(Z_1^{(b)}, Z_2^{(b)})$ iid $\sim N(0,1)$
  Generate $(U_1^{(b)}, U_2^{(b)})$ iid $\sim$ Laplace$\left(0, \frac{u_1 - l_1}{n(\epsilon/2)}\right)$ and $(U_3^{(b)}, U_4^{(b)})$ iid $\sim$ Laplace$\left(0, \frac{u_2 - l_2}{m(\epsilon/2)}\right)$
  Compute candidate solutions $(\check{\mu}_1^{(b)}, \check{\mu}_2^{(b)})$ using eq. (6)
  Compute $\hat{\nu}^{(b)} := \check{\mu}_1^{(b)} - \check{\mu}_2^{(b)}$
end for
return $\{\hat{\nu}^{(b)}\}_{b=1}^{B}$

B OPTIMIZATION FOR MEANS

As mentioned in Section 3.2, the estimator in eq. (6) is obtained via numerical optimization since closed-form solutions are not available. In this section we provide a brief overview of why this is the case. Indeed, for bounded data the approximations similar to those for proportions may be derived by using the central limit theorem (CLT) for truncated normal distributions under both symmetric and asymmetric bounds. More specifically, we have that $X_i \sim TN(\mu, \sigma^2, a, b)$, where TN is a truncated normal distribution with bounds (a, b). Defining

$$\alpha = \frac{a-\mu}{\sigma}, \qquad \beta = \frac{b-\mu}{\sigma}, \qquad \text{and} \qquad Q = \Phi(\beta) - \Phi(\alpha),$$

the truncated mean and variance are given by

$$\mu_T = \mu + \sigma \frac{\phi(\alpha) - \phi(\beta)}{Q}, \qquad \sigma_T^2 = \sigma^2 \left[1 - \frac{\beta\phi(\beta) - \alpha\phi(\alpha)}{Q} - \left(\frac{\phi(\alpha) - \phi(\beta)}{Q} \right)^2 \right].$$

Let the estimators of these parameters be

$$\hat{\mu}_T = \tfrac{1}{n} \sum_{i=1}^{n} X_i, \qquad \hat{\sigma}_T^2 = \tfrac{1}{n} \sum_{i=1}^{n} X_i^2 - (\hat{\mu}_T)^2.$$

Then, as $n \to \infty$, we have that

$$\sqrt{n}\,(\hat{\mu}_T - \mu_T) \xrightarrow{d} N(0, \sigma_T^2), \qquad \sqrt{n}\,(\hat{\sigma}_T^2 - \sigma_T^2) \xrightarrow{d} N(0, \tau^2),$$

where $\tau^2 = \mu_4 - (\sigma_T^2)^2$, with $\mu_4 = \mathbb{E}[X_i^4]$. By the Delta method, we finally have

$$\sqrt{n}\,(\hat{\sigma}_T - \sigma_T) \xrightarrow{d} N(0, \xi^2), \qquad \xi^2 = \frac{\mu_4 - (\sigma_T^2)^2}{4\sigma_T^2}.$$

B.0.0.1 *Private moments.*
To ensure differential privacy under bounded data, we perturb the estimators with privacy noise:

$$\hat{\mu}_T^+ = \hat{\mu}_T + U_1, \qquad \hat{\sigma}_T^+ = \hat{\sigma}_T + U_2,$$

where, in the setting of this work,

$$U_1 \sim \text{Lap}\left(\tfrac{b-a}{n\varepsilon_1}\right), \qquad U_2 \sim \text{Lap}\left(\tfrac{b-a}{\sqrt{n-1}\varepsilon_2}\right), \qquad \varepsilon = \varepsilon_1 + \varepsilon_2.$$

B.0.0.2 *Approximate estimation.*
Based on these approximations, we jointly estimate (μ, σ) via

$$(\check{\mu}, \check{\sigma}) = \underset{\mu \in \mathbb{R},\, \sigma > 0}{\arg\min} \left\| \begin{pmatrix} \hat{\mu}_T^+ - \mu_T(\mu,\sigma) - \dfrac{\sigma_T(\mu,\sigma)}{\sqrt{n}} Z_1^* - U_1 \\ \hat{\sigma}_T^+ - \sigma_T(\mu,\sigma) - \dfrac{\xi(\mu,\sigma)}{\sqrt{n}} Z_2^* - U_2 \end{pmatrix} \right\| \tag{A.1}$$

where $Z_1^*, Z_2^* \sim N(0,1)$ are independent of the privacy noises (U_1, U_2), and

$$\xi(\mu,\sigma)^2 \;=\; \frac{\mu_4(\mu,\sigma) - \sigma_T(\mu,\sigma)^4}{4\,\sigma_T(\mu,\sigma)^2}.$$

B.0.0.3 *Symmetric bounds ($a = \mu - d\,\sigma,\ b = \mu + d\,\sigma$).*
For symmetric truncation, $\alpha = -d$, $\beta = d$, $Q = 2\Phi(d) - 1$, so $\mu_T(\mu,\sigma) = \mu$ and

$$\sigma_T(\mu,\sigma) = \sigma\, C(d), \qquad C(d) = \sqrt{1 - \frac{2d\,\phi(d)}{2\Phi(d) - 1}}\,.$$

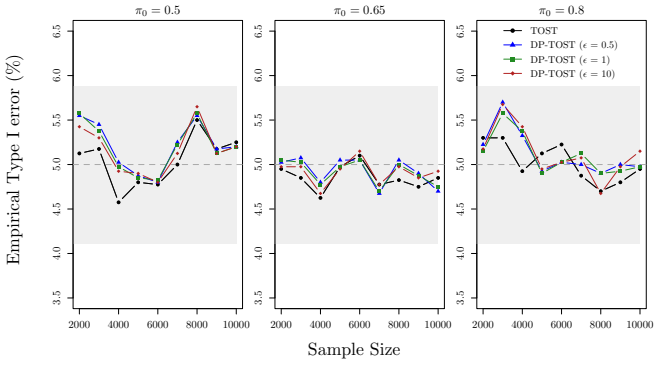
The joint estimator reduces to

$$(\check{\mu}, \check{\sigma}) = \underset{\mu\in\mathbb{R},\,\sigma>0}{\arg\min} \left\| \begin{pmatrix} \hat{\mu}_T^+ - \mu - \dfrac{\sigma\, C(d)}{\sqrt{n}} Z_1^* - U_1 \\ \hat{\sigma}_T^+ - \sigma\, C(d) - \dfrac{\xi(\mu,\sigma)}{\sqrt{n}} Z_2^* - U_2 \end{pmatrix} \right\| \tag{A.2}$$

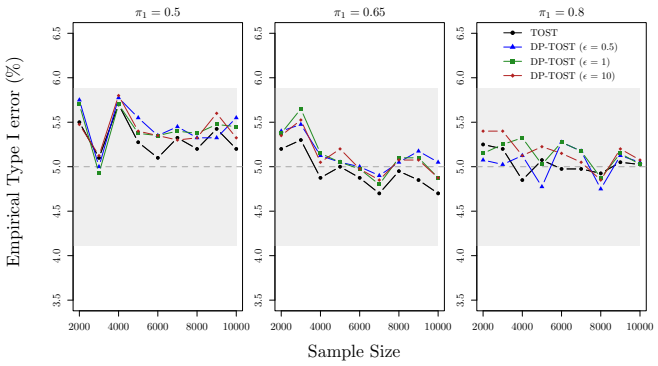
As can be observed in eqs. (A.1) and (A.2), it is not possible to obtain explicit solutions to the above minimization problems based on the approximations (knowledge of certain populations quantities would be required for this purpose and therefore would be unrealistic).

C ADDITIONAL RESULTS

This section presents additional simulation results, including: (i) type I error analysis for larger sample sizes (n = 200–1000) in the case of one-sample and two-sample proportion tests (Figures A.1a and A.1b); (ii) results for the two-sample bounded-mean setting with unequal variances ($\sigma_1^2 = 1$, $\sigma_2^2 = 2$) under separate truncation bounds (Figures A.2 and A.3); (iii) one-sample DP-TOST for proportions (Figures A.4 and A.5); and (iv) one-sample DP-TOST for bounded means (Figures A.6 and A.7).



(a) One Sample Proportions



(b) Two Sample Proportions

FIGURE A.1 Empirical Type I error rates for the standard TOST procedure and DP-TOST for larger sample sizes when assessing equivalence in (a) one proportion with respect to a fixed value and (b) two proportions using $H = 4,000$ Monte Carlo replicates. Each plot corresponds to a fixed value of the underlying proportion parameter (denoted as π_0 in one-sample settings and π_1 in two-sample settings), taking values in $0.5, 0.65, 0.8$. The shaded region denotes the nominal 5% level with simulation error. Curves are shown for the standard TOST and DP-TOST procedures with $\epsilon \in \{0.5, 1, 10\}$.

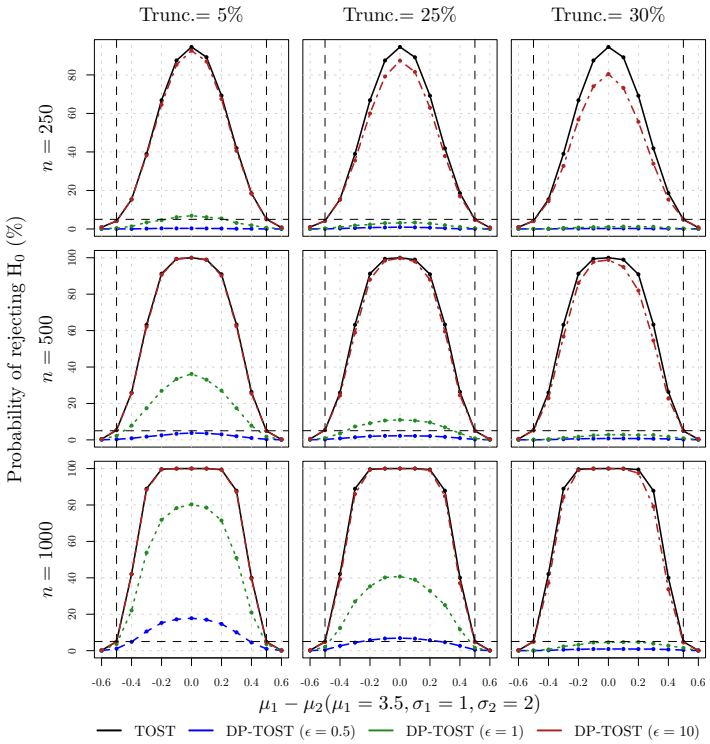


FIGURE A.2 Empirical power of standard TOST and proposed DP-TOST procedures for testing equivalence in difference of two means under separate bounds, evaluated across a range of true differences $(\mu_1 - \mu_2) \in [-0.6, 0.6]$ with $\mu_1 = 3.5$, $\sigma_1 = 1$, $\sigma_2 = 2$ and 4,000 Monte Carlo replicates. Each row corresponds to a specific sample size $n \in \{250, 500, 1000\}$ and each column to a truncation level. Power is reported under varying privacy budgets $\epsilon \in \{0.5, 1, 10\}$.

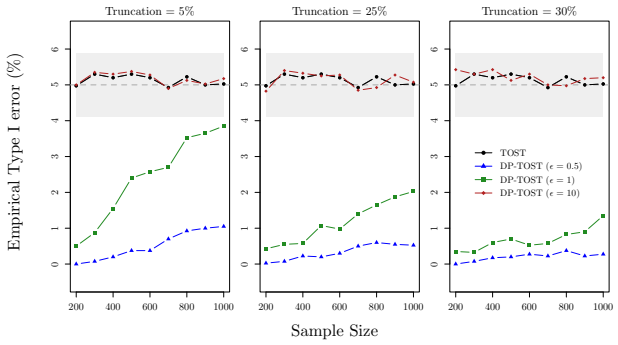


FIGURE A.3 Empirical Type I error rates for the standard TOST procedure and DP-TOST when testing equivalence in the difference of two means under separate bounds, under a nominal level of 5%. Results are shown for privacy budgets $\epsilon \in \{0.5, 1, 10\}$, $\sigma_1 = 1, \sigma_2 = 2$ and sample sizes $n \in (200, 1000)$. Empirical Type I error is calculated as the maximum rejection probability at the two boundary values $\mu_1 - \mu_2 = \pm 0.5$. Monte Carlo simulations were conducted with $H = 4,000$ replicates, and the shaded region denotes the nominal 5% level with simulation error.

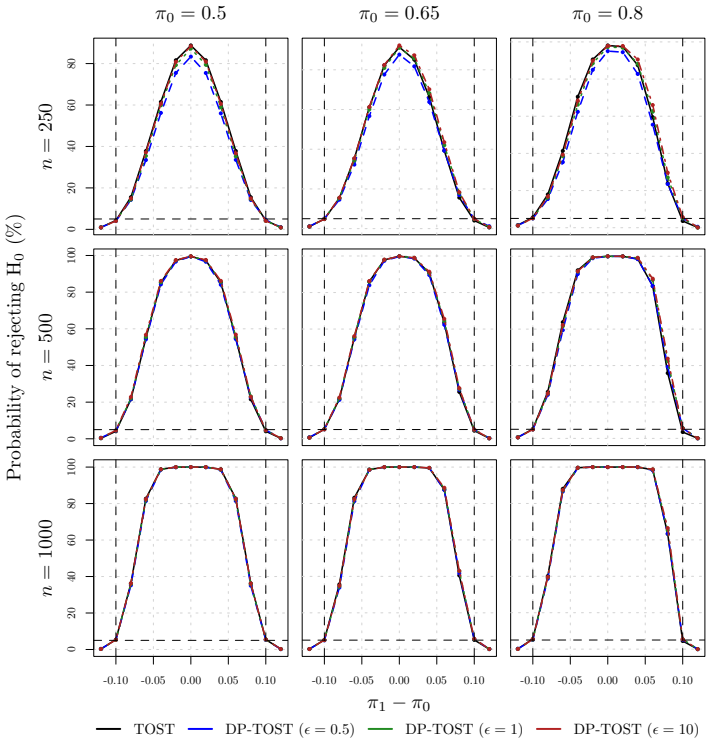


FIGURE A.4 Empirical power of the standard TOST and the proposed DP-TOST procedures for testing equivalence in a single proportion, evaluated across true differences $(\pi_1 - \pi_0) \in [-0.10, 0.10]$ using 4,000 Monte Carlo replicates. Each column corresponds to a fixed null proportion $\pi_0 \in \{0.5, 0.65, 0.8\}$, and each row to a sample size $n \in \{250, 500, 1000\}$. Power is reported under privacy levels $\epsilon \in \{0.5, 1, 10\}$.

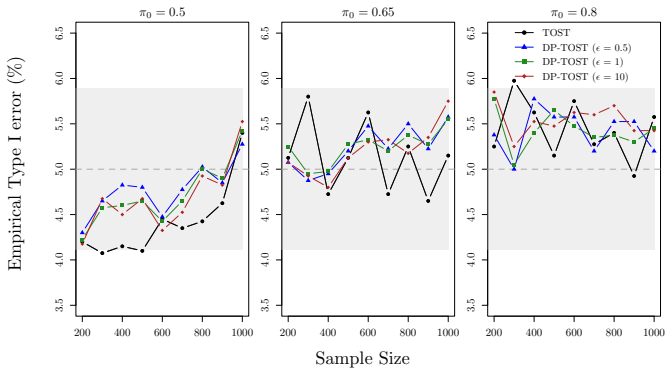


FIGURE A.5 Empirical Type I error rates for the standard TOST procedure and DP-TOST when assessing equivalence in a single proportion. Each panel corresponds to a different true null value $\pi_0 \in \{0.5, 0.65, 0.8\}$. Empirical Type I error is computed as the maximum rejection probability at the two equivalence boundaries $\pi_1 - \pi_0 = \pm 0.1$. Monte Carlo simulations were conducted with $H = 4,000$ replicates, and the shaded region denotes the nominal 5% level with simulation error. Results are shown for the standard TOST and DP-TOST procedures with privacy parameters $\epsilon \in \{0.5, 1, 10\}$.

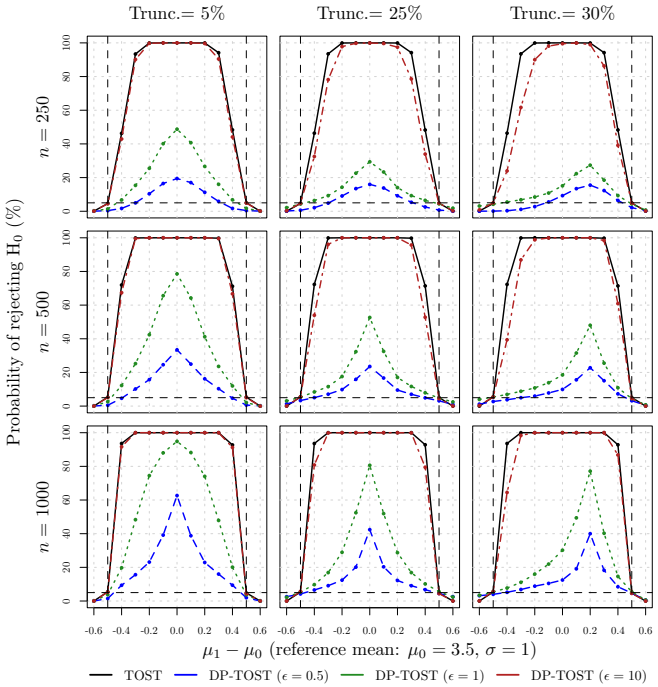


FIGURE A.6 Empirical power of the standard TOST and the proposed DP-TOST procedures for testing equivalence in a one-sample mean setting using 4,000 Monte Carlo replicates. Results are presented across a range of true differences ($\mu_1 - \mu_0$), where the reference mean is $\mu_0 = 3.5$ and standard deviation $\sigma = 1$. Each row corresponds to a sample size $n \in \{250, 500, 1000\}$, and each column represents a truncation level: 5%, 25%, and 30%. The 30% case includes both symmetric and asymmetric truncation bounds, with the latter extending the upper margin to improve sensitivity when $\mu_1 > \mu_0$.

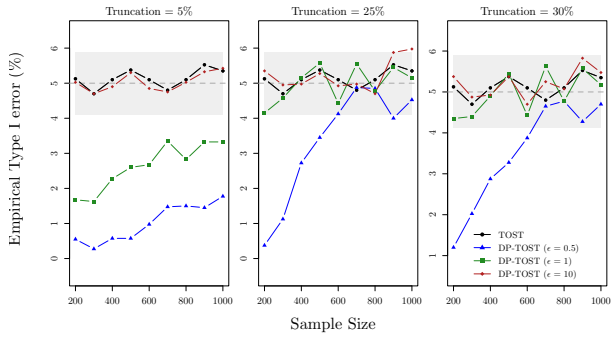


FIGURE A.7 Empirical Type I error rates for the standard TOST procedure and DP-TOST when testing equivalence in a one-sample mean setting, under a nominal level of 5%. Results are shown for privacy budgets $\epsilon \in \{0.5, 1, 10\}$, $\sigma = 1$, sample sizes $n \in (200, 1000)$ and true differences ($\mu_1 - \mu_0$), where the reference mean is $\mu_0 = 3.5$. Empirical Type I error is calculated as the maximum rejection probability at the two boundary values $\mu_1 - \mu_0 = \pm 0.5$. Monte Carlo simulations were conducted with $H = 4,000$ replicates, and the shaded region indicates the nominal 5% level with simulation error.

D COMPUTATIONAL DETAILS

All simulations and analyses were conducted in R[57], using both base and contributed packages to implement the proposed methods. Simulation specifics, including sample sizes, truncation levels, and privacy budgets, are detailed in Table 1. For the difference of proportions setting, a complete run generating all nine panels of Figure 3a required approximately 1.1 hours on a Dell desktop equipped with an Intel Core i7 processor using 7 parallel threads. In contrast, simulations for the difference of two bounded means were more computationally demanding due to the use of the `optim` function for parameter estimation. Generating a single panel of Figure 3b in a fixed configuration took approximately 50 minutes. To efficiently execute the 4000 Monte Carlo replications required for each scenario in multiple settings, we utilized the Auburn University High Performance Computing (AUHPC) cluster[58], which enabled parallel processing at scale. For the case study, the analysis for difference in proportions was completed in approximately 0.20 seconds, while Gaussian mean-based analysis required around 3.5 minutes, both executed on the same Dell system. All code and supplementary materials are made available at the GitHub repo in this link.