Contributions to Equivalence Testing

Younes Boulaguiem, PhD
Postdoctoral Researcher in Statistics

University of Geneva – Faculty of Medicine
Clinical Research Center, HUG

An overview of the key contributions from my PhD research.
Delivered August 2025

View Publications

Statistical Equivalence

What does "equivalence" mean?

- Not *"exactly the same"*, but *"so close that any difference is practically irrelevant"*.

Why is it important?

- We often need to **swap** a treatment, process, device, model, without losing performance. **Declaring equivalence** lets us do that **safely** and **efficiently**.

Why not rely on theory alone?

- Theory is only as good as its assumptions, but **assumptions break** because of invisible factors, noise, drifts...

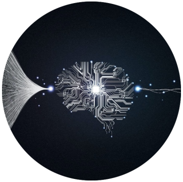
  → **Variability/uncertainty** is the rule.

Why statistics?

- It quantifies uncertainty, controls the risk of being wrong, and turns observations into decisions.

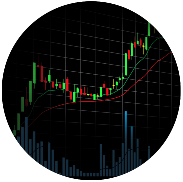
Domains of Application

Increasingly prevalent across fast-paced, high stakes industries:



AI/Software

Interchangeability of deployment pipelines



Finance

Equivalence of investment strategies.

Wearables

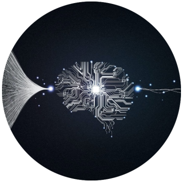
Essential to market as medical devices.

Generics

Crucial for market access.

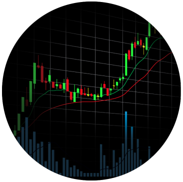
Domains of Application

Increasingly prevalent across fast-paced, high stakes industries:



AI/Software

Interchangeability of deployment pipelines



Finance

Equivalence of investment strategies.

Wearables

Essential to market as medical devices.

Our Focus!

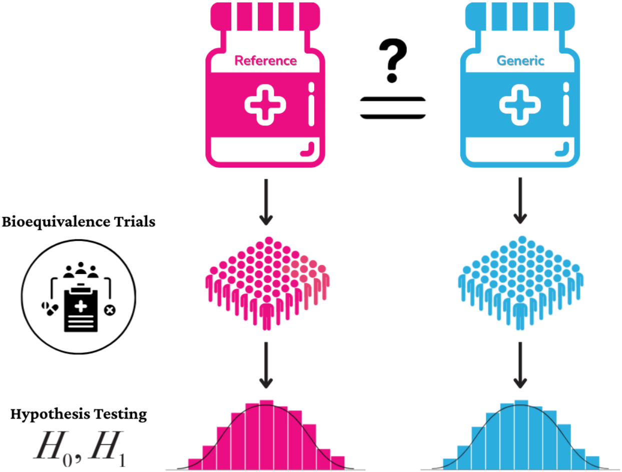
Generics

Also known as *bioequivalence*

Why?

- **Public health, highly regulated:** substituting with a non-equivalent product compromises consumer safety; regulators require statistical proof.
- **Massive economic stakes:** Over a 500-billion CHF market ⇒ missing true equivalence delays affordable, effective generics and wastes resources.
- **Topicals are pressing:** hard to measure and multidimensional.

Bioequivalence



Statistical Inference: In Everyday Life

Statistical hypothesis testing is a formalization of how we naturally make decisions in everyday life:

1. Begin with an assumption (denoted H_0 or the *null* hypothesis),
2. Look for evidence that could overturn it,
3. Change our mind (into H_1 or the *alternative* hypothesis) if evidence is strong enough to outweigh the risk of being wrong.

Example: Decide whether to take an umbrella outside in summer.

1. Start with the assumption it won't rain,
2. Glance outside at the sky,
3. Take the umbrella if it looks too cloudy.

Statistical Inference: Basic Principles

This process follows a few basic principles:

1. The initial assumption usually reflects the status quo and is the one we are most cautious about rejecting.

   → *it is super annoying to carry an umbrella on a dry summer day.*

2. The burden of proof is on the alternative hypothesis.

   → *I won't carry an umbrella unless I have strong evidence I should.*

3. We act under uncertainty.

   → We can't know for sure whether it will rain, but evidence helps us judge the likelihood

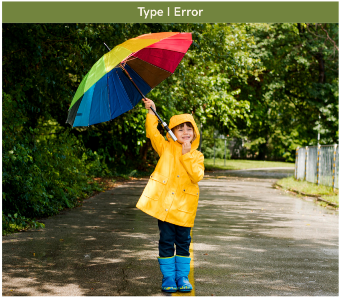
4. Two types of errors can be made:
   - **Type I error :** Occurs when we reject H_0 while it's true,
   - **Type II error :** Occurs when we fail to reject H_0 while it's wrong.

Type I and Type II Errors

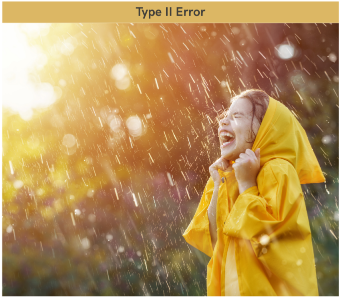
Type I error : Occurs when we $\underbrace{\text{reject } H_0}_{\text{take umbrella}}$ while $\underbrace{\text{it's true}}_{\text{it's a dry day}}$,

Type II error : Occurs when we $\underbrace{\text{fail ro reject } H_0}_{\text{don't take umbrella}}$ while $\underbrace{\text{it's wrong}}_{\text{it's a rainy day}}$.

Type I Error



Type II Error



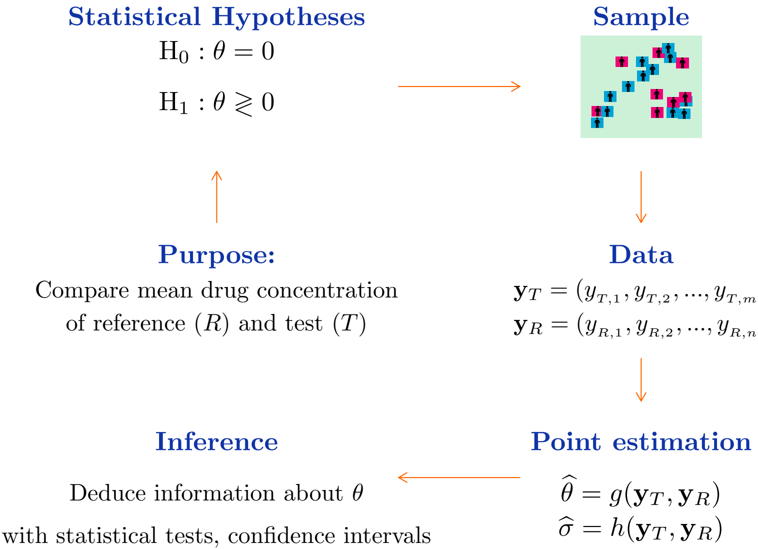
Recap and Terminology

- H_0 is the hypothesis we are cautious about rejecting; the burden of proof lies with H_1.
- We build a test and choose a rejection region, bounded by *critical values*, so that the Type I error rate equals α.
- Power can be increased without compromising Type I error rate by reducing uncertainty, which incurs additional costs and efforts.
- α is the *nominal significance level*; the actual Type I error rate is called the size of the test.

size $= \alpha$ $\Rightarrow$ **size**-α, i.e., exact test (ideal),

size $\leq \alpha$ $\Rightarrow$ **level**-α, i.e., conservative test (second best),

size $> \alpha$ $\Rightarrow$ liberal (we don't like...)



Purpose:

Compare mean drug concentration of reference (R) and test (T)

Statistical Hypotheses

$$H_0 : \theta = 0$$

$$H_1 : \theta \gtrless 0$$

Sample

Data

$$\mathbf{y}_T = (y_{T,1}, y_{T,2}, ..., y_{T,m})$$
$$\mathbf{y}_R = (y_{R,1}, y_{R,2}, ..., y_{R,n})$$

Point estimation

$$\widehat{\theta} = g(\mathbf{y}_T, \mathbf{y}_R)$$
$$\widehat{\sigma} = h(\mathbf{y}_T, \mathbf{y}_R)$$

Inference

Deduce information about θ

with statistical tests, confidence intervals

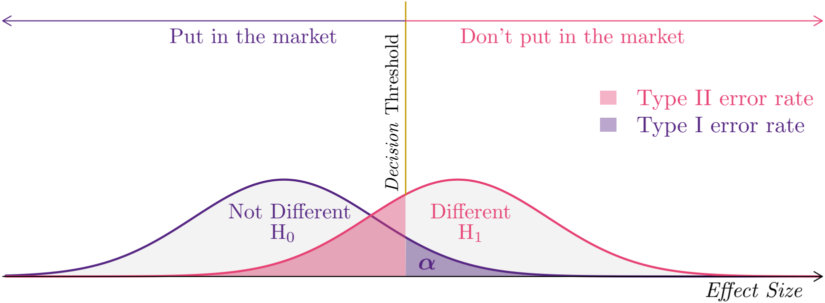
Limitation of the Classical Approach I

Drawbacks of the Classical Approach:

1. **Not rejecting the null does not mean no difference** $(\theta = 0)$.

→ Many times, just a lack of statistical power!

H$_0$: Not Different vs. H$_1$: Different



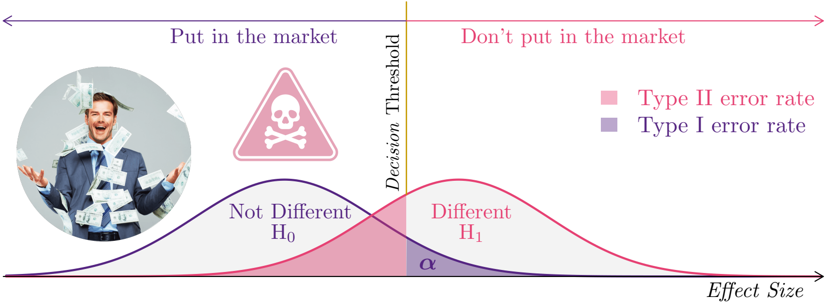
Limitation of the Classical Approach I

Drawbacks of the Classical Approach:

1. **Not rejecting the null does not mean no difference** $(\theta = 0)$.

→ Many times, just a lack of statistical power!

H_0 : Not Different vs. H_1 : Different



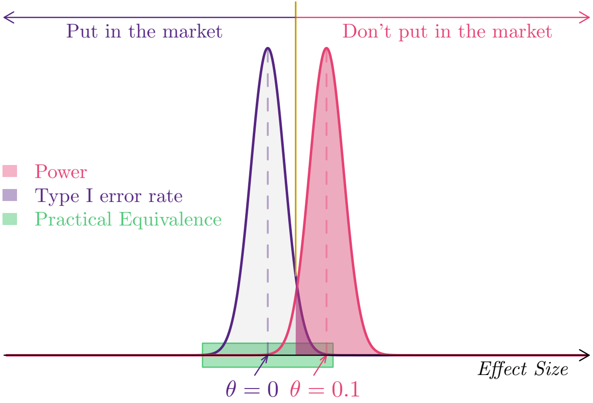
⚠ Drug manufacturers have no incentive to increase power (on the contrary!)
Potentially dangerous drugs make it to the market.

Limitation of the Classical Approach II

Drawbacks of the Classical Approach:

2. **Rejecting the null does not mean relevant effect size**.
→ In practice, small effects might not be considered "different" from 0.

H_0 : Not Different vs. H_1 : Different



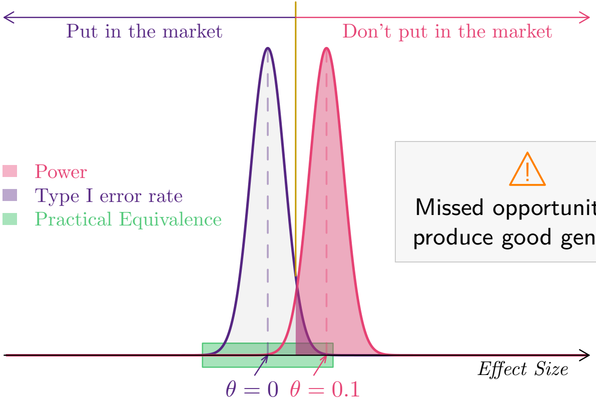
Limitation of the Classical Approach II

Drawbacks of the Classical Approach:

2. **Rejecting the null does not mean relevant effect size**.

→ In practise, small effects might not be considered “different” from 0.

H_0 : Not Different vs. H_1 : Different



⚠ Missed opportunity to produce good generics

Equivalence Testing

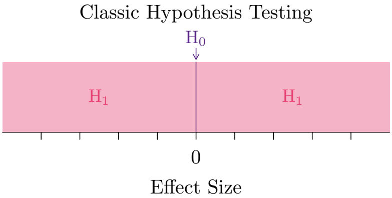
Solution:

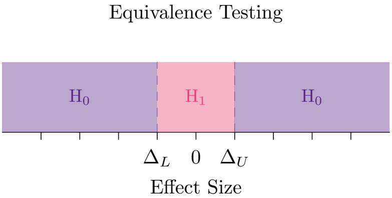
1. Put the burden of proof on the drug manufacturer to demonstrate equivalence:

   → Switch the null and the alternative.

2. Define a range for equivalence:

   → An open interval whose bounds correspond to what would be considered the “smallest” relevant effect sizes (in absolute value).



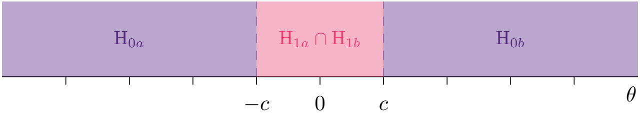


- In practice, equivalence bounds are often taken to be symmetrical around 0, so we define $c := \Delta_U = -\Delta_L$

Equivalence Testing Formally

Formally:

$$\begin{array}{ccc} \mathrm{H}_{0a}: \theta \leq -c & \textit{or} & \mathrm{H}_{0b}: \theta \geq c \\ \mathrm{H}_{1a}: \theta > -c & \textit{and} & \mathrm{H}_{1b}: \theta < c \end{array}$$



Decision rule by means of confidence intervals (interval inclusion principle):[1]

- Accept $\mathrm{H}_{1a}: \theta > -c$ **if** $\boldsymbol{I}_{\theta,U}^{1-\alpha} \not\ni -c$ (Upper one-sided CI),
- Accept $\mathrm{H}_{1b}: \theta < c$ **if** $\boldsymbol{I}_{\theta,L}^{1-\alpha} \not\ni c$ (Lower one-sided CI),

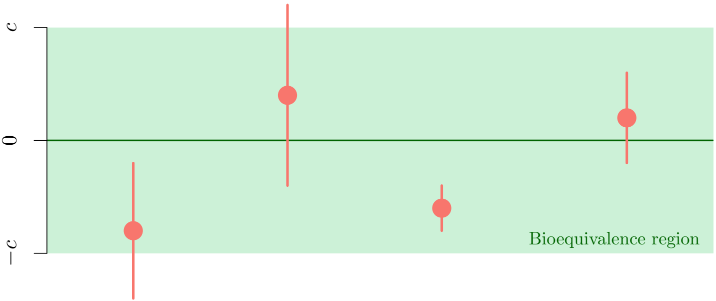
→ Accept $\mathrm{H}_1 := \mathrm{H}_{1a} \cap \mathrm{H}_{1b}$ **if** $\boldsymbol{I}_{\theta}^{1-2\alpha} := \boldsymbol{I}_{\theta,L}^{1-\alpha} \cap \boldsymbol{I}_{\theta,U}^{1-\alpha} \subset (-c, c)$.

Difference Testing vs. Equivalence Testing

Decision rule by means of confidence intervals (Interval Inclusion Principle):[1]

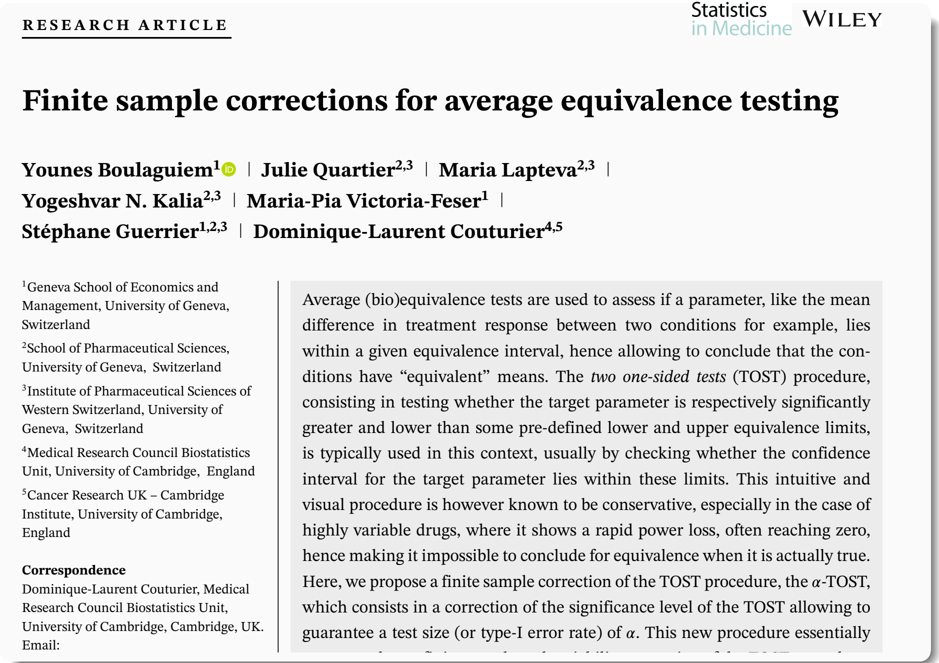
Case II: (a) Not different
(b) Nonequivalent

Case IV: (a) Not different
(b) Equivalent



Case I: (a) Different
(b) Nonequivalent

Case III: (a) Different
(b) Equivalent



RESEARCH ARTICLE

Statistics
in Medicine
WILEY

Finite sample corrections for average equivalence testing

Younes Boulaguiem[1] | Julie Quartier[2,3] | Maria Lapteva[2,3] |
Yogeshvar N. Kalia[2,3] | Maria-Pia Victoria-Feser[1] |
Stéphane Guerrier[1,2,3] | Dominique-Laurent Couturier[4,5]

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

[2]School of Pharmaceutical Sciences, University of Geneva, Switzerland

[3]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Switzerland

[4]Medical Research Council Biostatistics Unit, University of Cambridge, England

[5]Cancer Research UK – Cambridge Institute, University of Cambridge, England

Correspondence
Dominique-Laurent Couturier, Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, UK.
Email:

Average (bio)equivalence tests are used to assess if a parameter, like the mean difference in treatment response between two conditions for example, lies within a given equivalence interval, hence allowing to conclude that the conditions have "equivalent" means. The *two one-sided tests* (TOST) procedure, consisting in testing whether the target parameter is respectively significantly greater and lower than some pre-defined lower and upper equivalence limits, is typically used in this context, usually by checking whether the confidence interval for the target parameter lies within these limits. This intuitive and visual procedure is however known to be conservative, especially in the case of highly variable drugs, where it shows a rapid power loss, often reaching zero, hence making it impossible to conclude for equivalence when it is actually true. Here, we propose a finite sample correction of the TOST procedure, the α-TOST, which consists in a correction of the significance level of the TOST allowing to guarantee a test size (or type-I error rate) of α. This new procedure essentially

Two One-Sided Tests (TOST)

Canonical form:

$$\widehat{\theta} \sim \mathcal{N}\left(\theta, \sigma_\nu^2\right) \quad \text{and} \quad \frac{\nu\widehat{\sigma}_\nu^2}{\sigma_\nu^2} \sim \chi_\nu^2,$$

where $\sigma_\nu^2 := \sigma^2/n$ and σ^2 denotes the asymptotic variance.

The TOST[2] is based on two test statistics:

$$T_L := \frac{\widehat{\theta}+c}{\widehat{\sigma}_\nu} \sim t_\nu\left(\frac{\theta+c}{\sigma_\nu}\right) \quad \text{and} \quad T_U := \frac{\widehat{\theta}-c}{\widehat{\sigma}_\nu} \sim t_\nu\left(\frac{\theta-c}{\sigma_\nu}\right).$$

Decision rule by means of the TOST:

- Accept $\mathrm{H}_{1a}: \theta > -c$ **if** $T_L \geq t_{\alpha,\nu}$
- Accept $\mathrm{H}_{1b}: \theta < c$ **if** $T_U \leq -t_{\alpha,\nu}$,

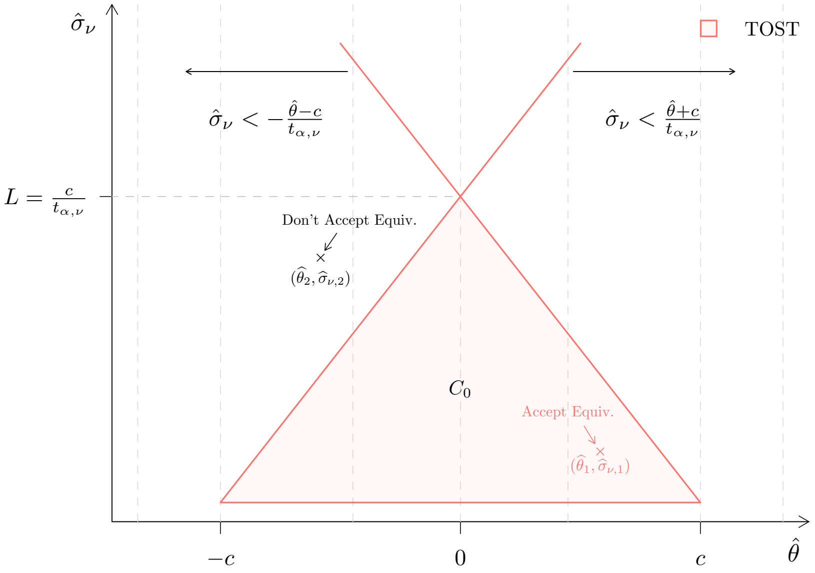
→ Accept $\mathrm{H}_1 := \mathrm{H}_{1a} \cap \mathrm{H}_{1b}$ **if** both tests reject their marginal nulls.

Rejection Region by rearranging the terms:

$$C_0 := \left\{\widehat{\theta} \in \mathbb{R}, \widehat{\sigma}_\nu > 0 \,\middle|\, |\widehat{\theta}| \leq c - t_{\alpha,\nu}\widehat{\sigma}_\nu\right\}.$$

Accept equivalence **if** $(\widehat{\theta}, \widehat{\sigma}_\nu) \in C_0$.

TOST: Rejection Region



$\hat{\sigma}_\nu$
□ TOST
$\hat{\sigma}_\nu < -\frac{\hat{\theta}-c}{t_{\alpha,\nu}}$
$\hat{\sigma}_\nu < \frac{\hat{\theta}+c}{t_{\alpha,\nu}}$
$L = \frac{c}{t_{\alpha,\nu}}$
Don't Accept Equiv.
$(\widehat{\theta}_2, \widehat{\sigma}_{\nu,2})$
C_0
Accept Equiv.
$(\widehat{\theta}_1, \widehat{\sigma}_{\nu,1})$
$-c$
0
c
$\hat{\theta}$

Size of the TOST

Power function:

$$p(\alpha, \theta, \sigma_\nu, \nu, c) := \Pr(\text{ Reject } \mathrm{H}_0 \text{ }) = \Pr\left(|\hat{\theta}| \leq c - t_{\alpha,\nu}\hat{\sigma}_1\right)$$

$$= \int_0^\infty I(\widehat{\sigma}_\nu t_{\alpha,\nu} < c) \left\{ \Phi\left(\frac{\theta}{\sigma_\nu} + \frac{c - t_{\alpha,\nu}\widehat{\sigma}_\nu}{\sigma_\nu}\right) - \Phi\left(\frac{\theta}{\sigma_\nu} - \frac{c - t_{\alpha,\nu}\widehat{\sigma}_\nu}{\sigma_\nu}\right) \right\} f_{\widehat{\sigma}_\nu}(\widehat{\sigma}_\nu | \sigma_\nu, \nu) d\widehat{\sigma}_\nu.$$

Size function:

$$\begin{aligned}\omega(\alpha, c, \sigma_\nu, \nu) :&= \Pr(\text{ Reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ }) \\ &= \sup_{\theta \notin (-c,c)} \Pr(\text{ Reject } \mathrm{H}_0 \text{ }) \\ &= p(\alpha, \text{ } c \text{ }, \sigma_\nu, \nu, c) \\ &< \alpha\end{aligned}$$

Size of the TOST

Power function:

$$p(\alpha, \theta, \sigma_\nu, \nu, c) := \Pr(\text{ Reject } H_0 \text{ }) = \Pr\left(|\hat{\theta}| \leq c - t_{\alpha,\nu}\hat{\sigma}_1\right)$$

$$= \int_0^\infty I(\widehat{\sigma}_\nu t_{\alpha,\nu} < c) \left\{ \Phi\left(\frac{\theta}{\sigma_\nu} + \frac{c - t_{\alpha,\nu}\widehat{\sigma}_\nu}{\sigma_\nu}\right) - \Phi\left(\frac{\theta}{\sigma_\nu} - \frac{c - t_{\alpha,\nu}\widehat{\sigma}_\nu}{\sigma_\nu}\right) \right\} f_{\widehat{\sigma}_\nu}(\widehat{\sigma}_\nu | \sigma_\nu, \nu) d\widehat{\sigma}_\nu .$$

Size function:

$$\omega(\alpha, c, \sigma_\nu, \nu) : = \Pr(\text{ Reject } H_0 \mid H_0 \text{ })$$

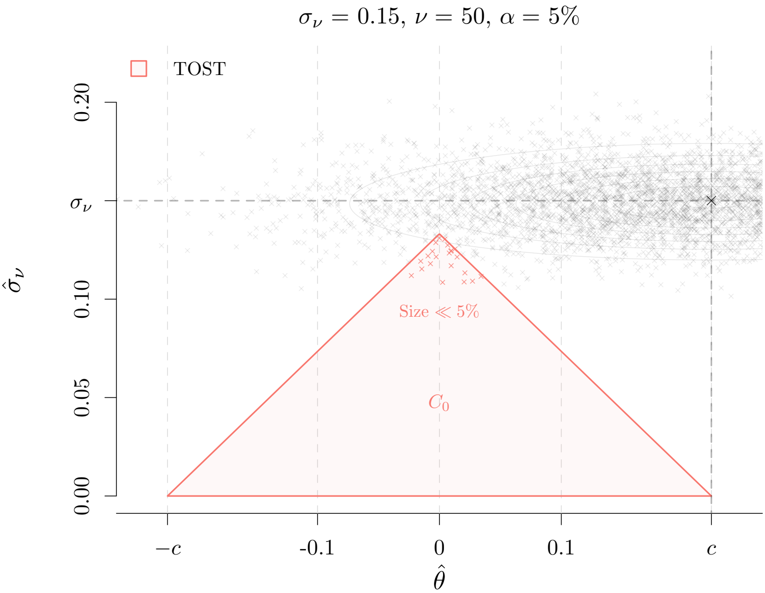
$$= \sup_{\theta \notin (-c,c)} \Pr(\text{ Reject } H_0 \text{ })$$

$$= p(\alpha, \; c \;, \sigma_\nu, \nu, c)$$

$$< \alpha$$

The **TOST** is conservative

Conservativeness of the TOST



Corrective Approaches

Recalling the size function, we can operate on two parameters:

$$\omega\left(\boxed{\alpha}, \boxed{c}, \sigma_\nu, \nu\right)$$

α-TOST:

$$\boxed{\alpha^*} := \underset{\gamma\in[\alpha,0.5]}{\text{argzero}} \left[\omega(\boxed{\gamma}, c, \sigma_\nu, \nu) - \alpha\right]$$

- A unique solution exits provided:

$$\sigma_\nu < \frac{2c}{\Phi^{-1}(\alpha + 0.5)},$$

- Converges exponentially fast:

$$\left|\alpha^{*(k+1)} - \alpha^*\right| < \frac{1}{2}\exp(-bk).$$

- Yields the rejection region:

$$C_1 := \left\{\hat{\theta} \in \mathbb{R}, \, \widehat{\sigma}_\nu > 0 \, \middle| \, |\hat{\theta}| \leq c - t_{\alpha^*,\nu}\widehat{\sigma}_\nu\right\}.$$

δ-TOST:

$$\boxed{\delta^*} := \underset{\delta\in[c,\infty)}{\text{argzero}} \left[\omega(\alpha, \boxed{\delta}, \sigma_\nu, \nu) - \alpha\right]$$

- A unique solution always exist:

$$\omega(c) < \alpha, \qquad \lim_{\delta\to\infty} \omega(\delta) = 1,$$

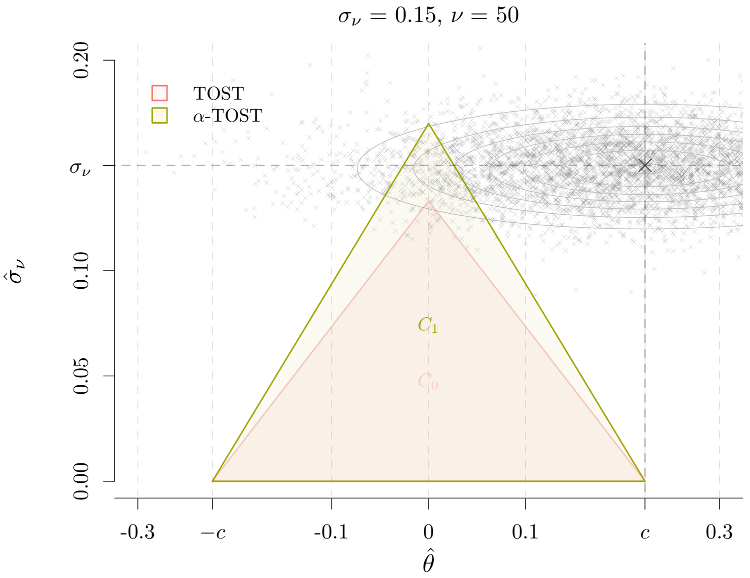
- No contraction.

→ Can be solved using general-purpose optimization methods.

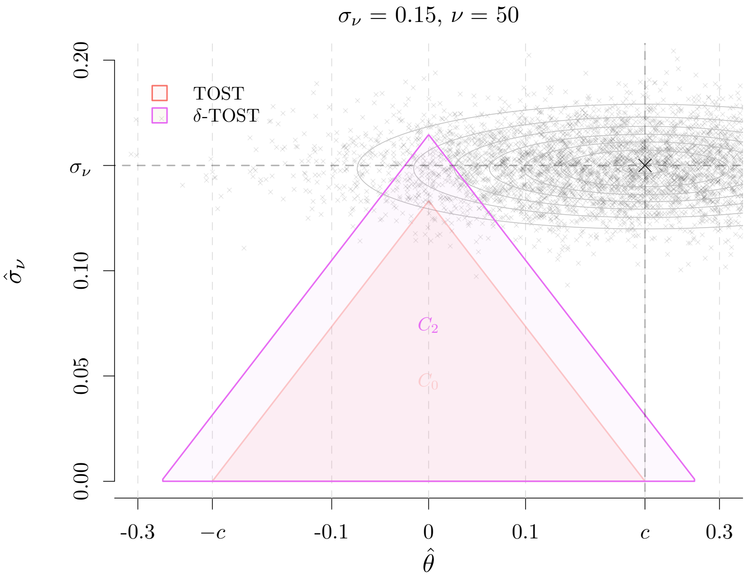
- Yields the rejection region:

$$C_2 := \left\{\hat{\theta} \in \mathbb{R}, \, \widehat{\sigma}_\nu > 0 \, \middle| \, |\hat{\theta}| \leq \delta^* - t_{\alpha,\nu}\widehat{\sigma}_\nu\right\}.$$

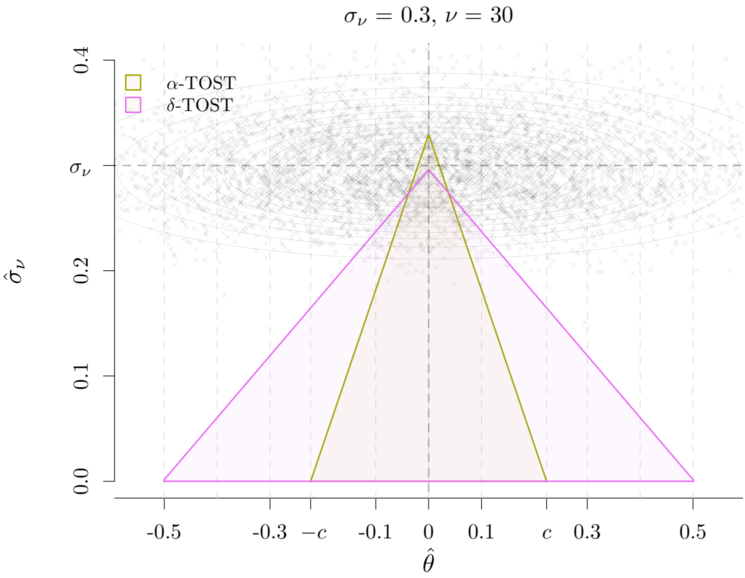
Rejection Region: α-TOST



Rejection Region: δ-TOST



Power at $\theta = 0$: α-TOST vs. δ-TOST



The α-TOST Empirically

When $\boldsymbol{\sigma_\nu}$ is unknown:

- Replace it by its empirical $\widehat{\sigma}_\nu$ and solve

$$\widehat{\alpha}^* := \alpha^*(\widehat{\sigma}_\nu) = \underset{\gamma\in[\alpha,0.5]}{\text{argzero}} \left[\omega(\gamma, c, \widehat{\sigma}_\nu, \nu) - \alpha\right].$$

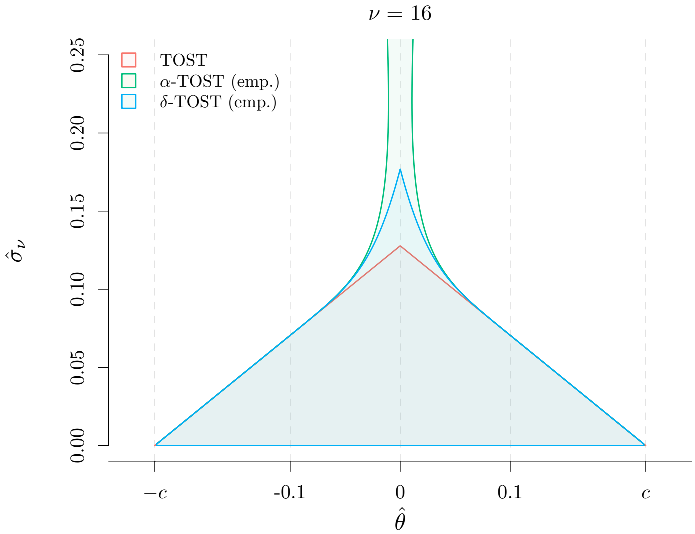
- Conditions for existence and computational efficiency remain unchanged.

Large sample behaviour:

$$\widehat{\alpha}^* = \alpha^* + o_p\left(\nu^{-1}\right), \;\; \widehat{\sigma}_\nu = \sigma_\nu + \mathcal{O}_p\left(\nu^{-1}\right), \;\; \widehat{\theta} = \theta + \mathcal{O}_p\left(\nu^{-1/2}\right).$$

→ Uncertainty related to $\widehat{\alpha}^*$ is asymptotically negligible compared to that of $\widehat{\theta}$ and $\widehat{\sigma}_\nu$.

Rejection Areas: emprical α-TOST vs. empirical δ-TOST



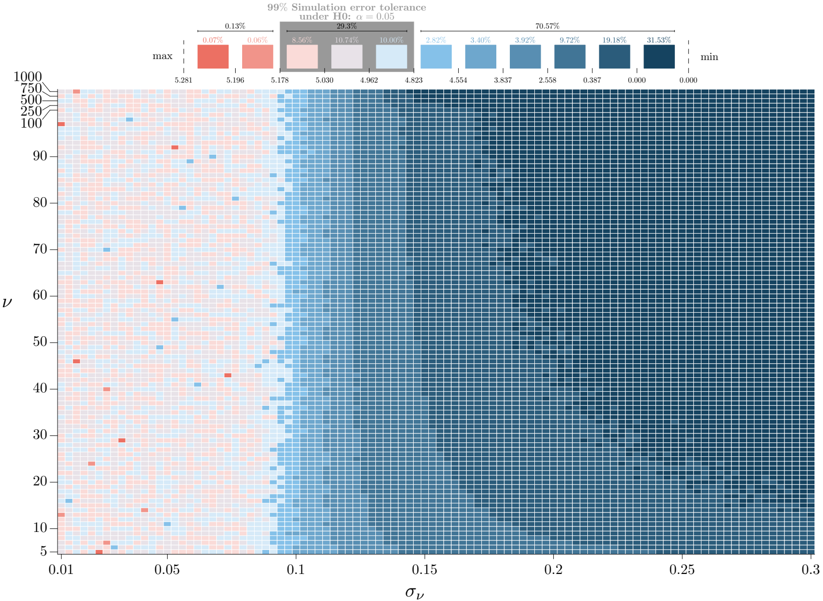
The α-TOST empirical rejection region **entirely includes** the other two.

→ The α-TOST is **uniformly more powerful**.

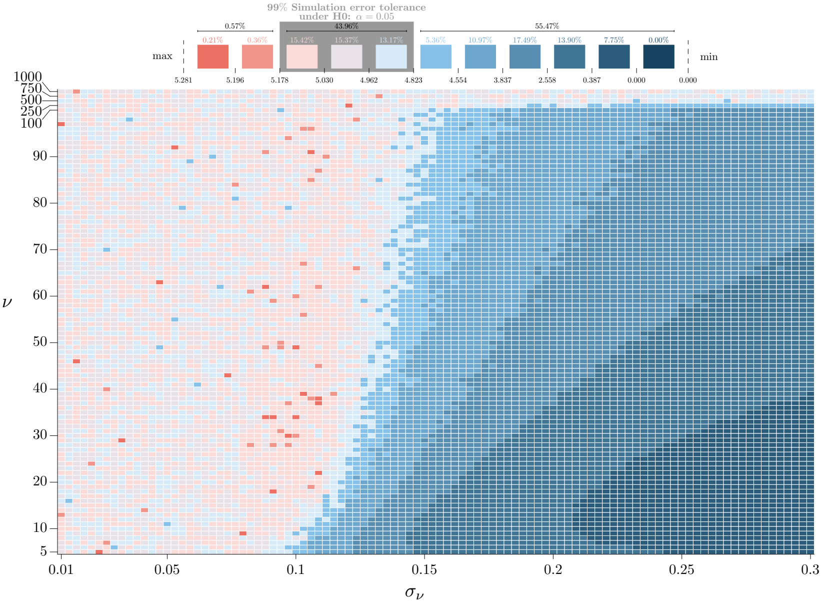
Simulation Study

	Simulation settings	
	Size	**Power**
c	$\log(1.25) \approx 0.2231$	
ν	5, 6, 7, … , 100, 250, 500, 750, 1000 (100 values)	
σ_ν	100 evenly spaced values between 0.01 and 0.3	
θ	c	0
α	0.05	
B	10^5	
Design	General (canonical form)	
Methods	TOST, α-TOST and δ-TOST	

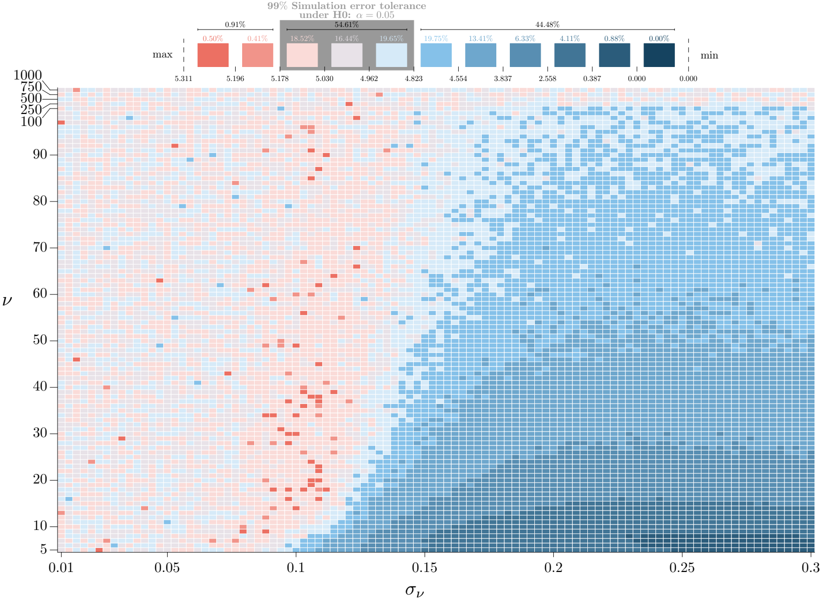
Simulation Study: TOST Empirical Size



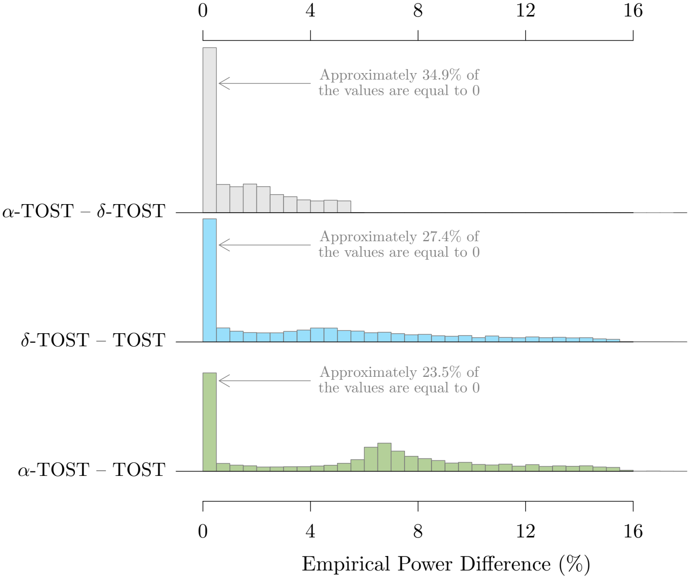
Simulation Study: δ-TOST Empirical Size



Simulation Study: α-TOST Empirical Size



Simulation Study: Empirical Power at $\theta = 0$

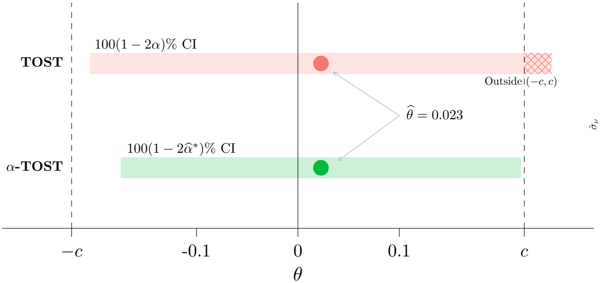


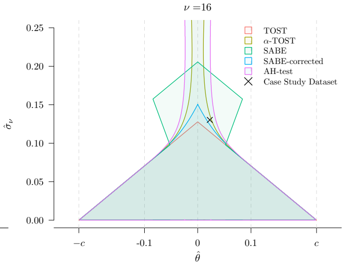
Case Study: Econazole Nitrate Deposition

- **Objective:** Assess bioequivalence of econazole nitrate skin deposition from a reference and an already approved generic cream.
- **Design:** 17 **paired** porcine skin samples.
- **Purpose:**
  1. Our method is **design-agnostic**, unlike other methods that require replicate designs.[3]
  2. Our method can produce **confidence intervals**, unlike other methods like the AH-test, Brown et al., and Berger & Hsu.[4]

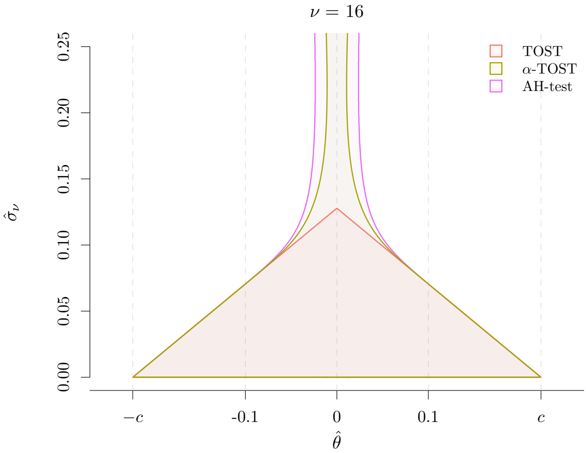
Case Study: Econazole Nitrate Deposition

- **Objective:** Assess bioequivalence of econazole nitrate deposition from a reference and an already approved generic cream.
- **Design:** 17 **paired** porcine skin samples.
- **Purpose:**
  1. Our method is **design-agnostic**, unlike SABE-like methods that require replicate designs.[3]
  2. Our method can produce **confidence intervals**, unlike other methods like the AH-test, Brown et al., and Berger & Hsu.[4]



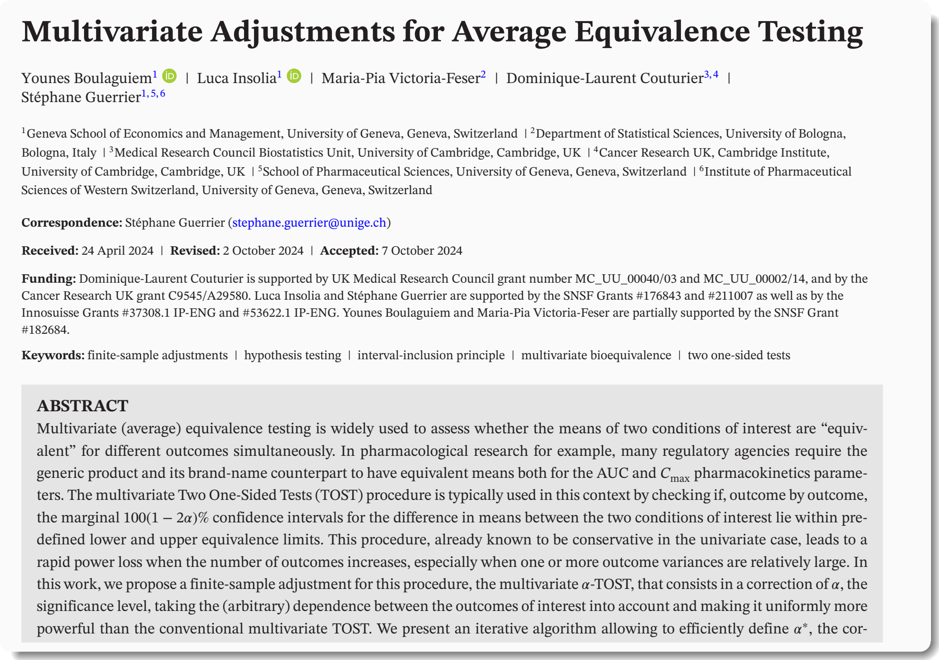


α-TOST: The Right Compromise



The AH-test is know to be quite liberal. In his TOST paper,[2] *Schuirmann* said the following after comparing the two methods:

> *"[...] The best procedure to use may therefore turn out to be a compromise between the two procedures."*



Multivariate Adjustments for Average Equivalence Testing

Younes Boulaguiem[1] | Luca Insolia[1] | Maria-Pia Victoria-Feser[2] | Dominique-Laurent Couturier[3,4] | Stéphane Guerrier[1,5,6]

[1]Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland | [2]Department of Statistical Sciences, University of Bologna, Bologna, Italy | [3]Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, UK | [4]Cancer Research UK, Cambridge Institute, University of Cambridge, Cambridge, UK | [5]School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland | [6]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland

Correspondence: Stéphane Guerrier (stephane.guerrier@unige.ch)

Received: 24 April 2024 | **Revised:** 2 October 2024 | **Accepted:** 7 October 2024

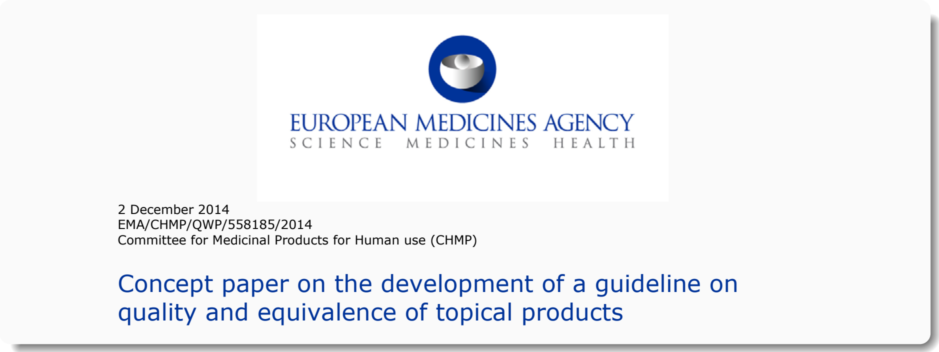
Funding: Dominique-Laurent Couturier is supported by UK Medical Research Council grant number MC_UU_00040/03 and MC_UU_00002/14, and by the Cancer Research UK grant C9545/A29580. Luca Insolia and Stéphane Guerrier are supported by the SNSF Grants #176843 and #211007 as well as by the Innosuisse Grants #37308.1 IP-ENG and #53622.1 IP-ENG. Younes Boulaguiem and Maria-Pia Victoria-Feser are partially supported by the SNSF Grant #182684.

Keywords: finite-sample adjustments | hypothesis testing | interval-inclusion principle | multivariate bioequivalence | two one-sided tests

ABSTRACT

Multivariate (average) equivalence testing is widely used to assess whether the means of two conditions of interest are "equivalent" for different outcomes simultaneously. In pharmacological research for example, many regulatory agencies require the generic product and its brand-name counterpart to have equivalent means both for the AUC and $C_{\max}$ pharmacokinetics parameters. The multivariate Two One-Sided Tests (TOST) procedure is typically used in this context by checking if, outcome by outcome, the marginal $100(1-2\alpha)\%$ confidence intervals for the difference in means between the two conditions of interest lie within predefined lower and upper equivalence limits. This procedure, already known to be conservative in the univariate case, leads to a rapid power loss when the number of outcomes increases, especially when one or more outcome variances are relatively large. In this work, we propose a finite-sample adjustment for this procedure, the multivariate α-TOST, that consists in a correction of α, the significance level, taking the (arbitrary) dependence between the outcomes of interest into account and making it uniformly more powerful than the conventional multivariate TOST. We present an iterative algorithm allowing to efficiently define α^*, the cor-

Equivalence Assessment of Topical Products



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

2 December 2014
EMA/CHMP/QWP/558185/2014
Committee for Medicinal Products for Human use (CHMP)

Concept paper on the development of a guideline on quality and equivalence of topical products

Challenges:

1. Blood-based PK methods fail: drug concentration often **untraceable in blood** (only side effects remain).

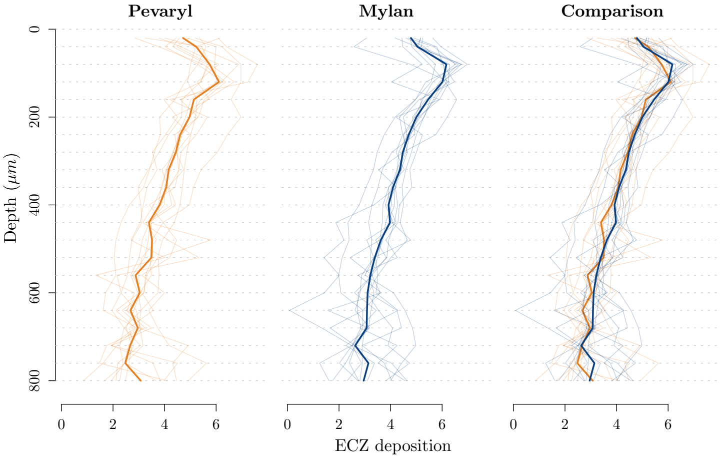
→ Need to assess bioavailability directly across **multiple skin targets**.

2. Need for appropriate statistical tools.

First Solution: Cutaneous Biodistribution

Cutaneous Biodistribution[5] addresses the first challenge:

→ Provides high resolution spatial distribution of the drug across the skin.



Multivariate Equivalence Testing

Let $\boldsymbol{\theta}$ be the vector of parameters:

$$\mathrm{H}_0 : \boldsymbol{\theta} \notin \boldsymbol{\Theta}_1 \quad \text{vs.} \quad \mathrm{H}_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1,$$

where $\boldsymbol{\Theta}_1 := \left\{x \in \mathbb{R}^m \,\middle|\, -c < x_j < c,\, j = 1, \ldots, m\right\}$

Decision rule by means of the Interval Inclusion Principle:

1. **Norm-based:** define hyperellipsoidal confidence regions (e.g., Mahalanobis, Tseng-Brown, Casella-Hwang, etc.).[4]
2. **Multivariate TOST:** defines hyper-rectangular confidence regions.

→ Multivariate TOST empirically outperforms alternatives.[6]

Canonical form:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}_m(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \quad \text{and} \quad \nu\widehat{\boldsymbol{\Sigma}} \sim \mathcal{W}_m(\nu, \boldsymbol{\Sigma}).$$

Rejection region:

$$C_0(\widehat{\boldsymbol{\Sigma}}) := \bigcap_{j=1}^{m} \left\{ |\hat{\theta}_j| \leq c - t_{\alpha,\nu}\hat{\sigma}_j \right\}.$$

Size and Power Functions

The power function:

$$p(\alpha, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \nu, \boldsymbol{c}) := \Pr\left\{ C_0(\widehat{\boldsymbol{\Sigma}}) | \alpha, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \nu, \boldsymbol{c} \right\}$$

$$= \int_0^{M_m^2} \cdots \int_0^{M_1^2} \int_{-M_{m-1}M_m}^{M_{m-1}M_m} \cdots \int_{-M_1 M_2}^{M_1 M_2}$$

$$\eta(\widehat{\boldsymbol{\Sigma}}) \times \Delta(\widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma} | \boldsymbol{\theta}) f_{\widehat{\Sigma}}(\widehat{\boldsymbol{\Sigma}} | \boldsymbol{\Sigma}) \; d\widehat{\Sigma}_{1,2} \ldots d\widehat{\Sigma}_{m-1,m} d\widehat{\sigma}_1^2 \ldots d\widehat{\sigma}_m^2,$$

for

$$\eta(\widehat{\boldsymbol{\Sigma}}) := I(|\widehat{\Sigma}_{1,2}| < \widehat{\sigma}_1 \widehat{\sigma}_2) \times \cdots \times I(|\widehat{\Sigma}_{m-1,m}| < \widehat{\sigma}_{m-1} \widehat{\sigma}_m),$$

The size of the multivariate TOST:

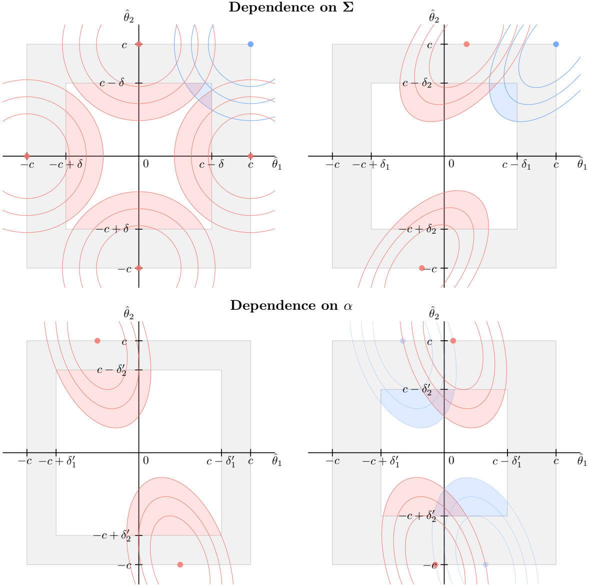
- Define:

$$\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^T \in \boldsymbol{\Lambda}(\alpha, \boldsymbol{\Sigma}) := \underset{\boldsymbol{\theta} \notin \Theta_1}{\operatorname{argsup}} \; p(\alpha, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \nu, \mathbf{c}).$$

- Therefore,

$$\text{Size} = \sup_{\boldsymbol{\theta} \notin \Theta_1} p(\alpha, \boldsymbol{\theta}, \boldsymbol{\Sigma}, \nu, \mathbf{c}) = p(\alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \nu, \mathbf{c}).$$

Dependence of $\boldsymbol{\lambda}$ on α and $\boldsymbol{\Sigma}$



Dependence on $\boldsymbol{\Sigma}$

Dependence on α

Size of the Multivariate TOST

Conservativeness:

$$\begin{aligned} p(\alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \nu, \boldsymbol{c}) &= \Pr\left(\bigcap_{j=1}^{m} \left\{|\widehat{\theta}_j| < c - t_{\alpha,\nu}\widehat{\sigma}_j\right\}\right) \\ &\leq \min_{j=1,\dots,m} \Pr\left(\left\{|\widehat{\theta}_j| < c - t_{\alpha,\nu}\widehat{\sigma}_j\right\}\right) \\ &\leq \Pr\left(\left\{|\widehat{\theta}_h| < c - t_{\alpha,\nu}\widehat{\sigma}_h\right\}\right) < \alpha, \end{aligned}$$

where h is such that λ_h is equal to c or $-c$.

- The conservativeness is further exacerbated as the number of dimensions increase.
- In extreme case with highly variable drugs for example, it is almost impossible to detect equivalence, the power curve is entirely flat.

Multivariate α-TOST

Definition:

$$\alpha^*(\mathbf{\Omega}) := \underset{\gamma\in[\alpha,0.5]}{\text{argzero}} \left[p(\gamma, \boldsymbol{\lambda}(\gamma, \mathbf{\Omega}), \mathbf{\Omega}, \nu, \mathbf{c}) - \alpha \right],$$

where $\mathbf{\Omega}$ represents the covariance matrix (true $\mathbf{\Sigma}$ or estimated $\widehat{\mathbf{\Sigma}}$).

Existence conditions:

- Depend on variability and dimension.
  - → e.g., under independence:

$$\sigma_{\text{max}} < \frac{2c}{\Phi^{-1}\left(\alpha^{1/m} + 1/2\right)}, \quad m < \log_2(\alpha^{-1}).$$

Asymptotic properties:

- Similar to the univariate case.

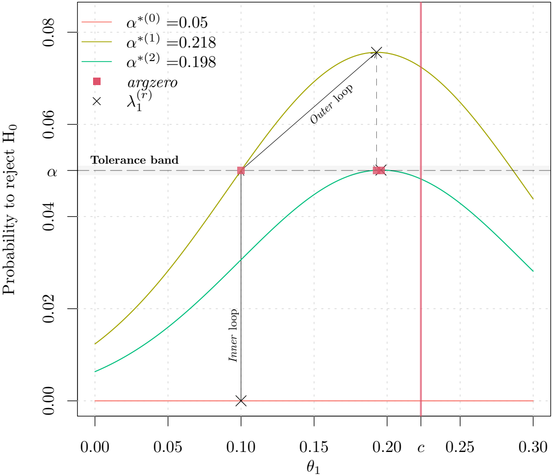
Algorithm:

- Trickier as $\boldsymbol{\lambda}$ depends on the significance level.

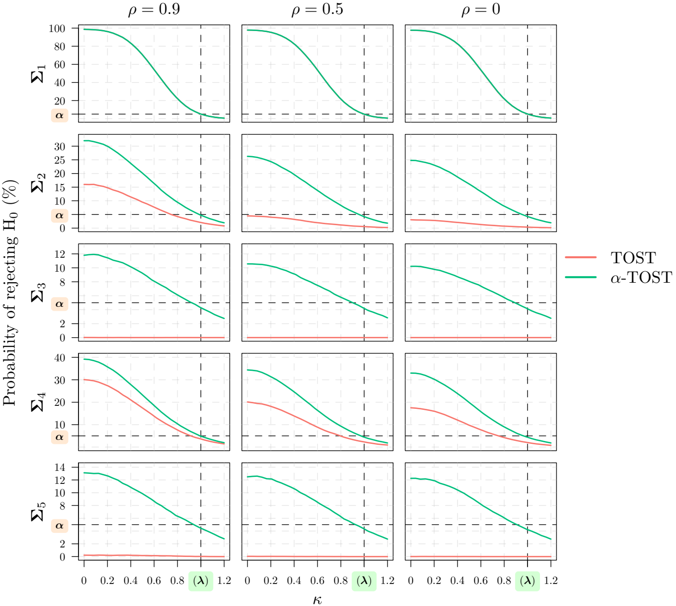
Multivariate α-TOST: Algorithm

Two-step update. For $r = 0, \ldots, r_{\max} - 1$:

1. **Outer loop:** given $\alpha^{*(r)}$, compute $\boldsymbol{\lambda}^{(r)}$.
2. **Inner loop:** given $\boldsymbol{\lambda}^{(r)}$, update $\alpha^{*(r+1)}$.

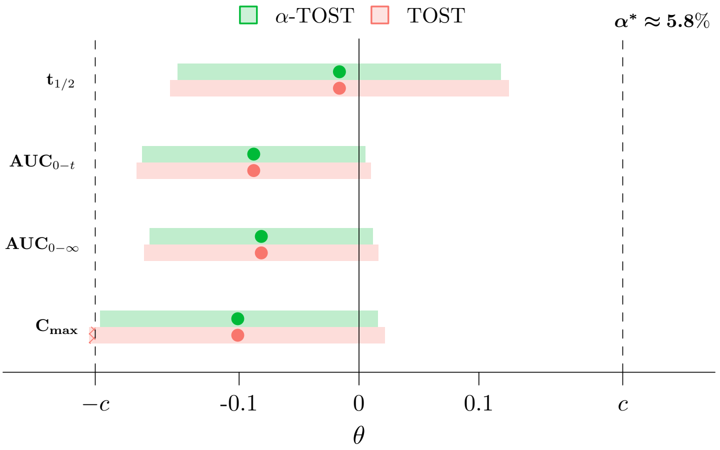


Empirical Power Curves for $m = 4$

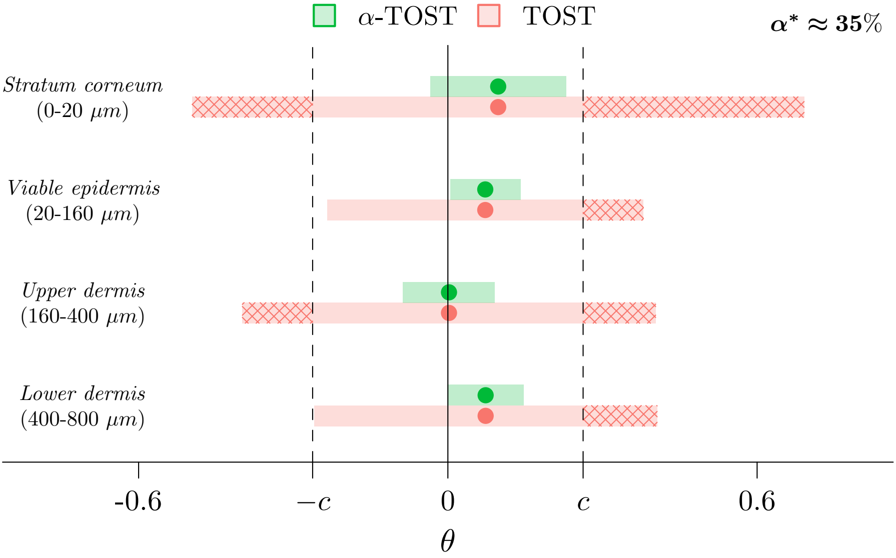


Case Study I: Ticlopidine Hydrochloride

Ticlopidine hydrochloride dataset:[7] 4 pharmacokinetic parameters measured in a 2×2×2 crossover design with n = 20 participants.



Case Study II: Econazole Nitrate



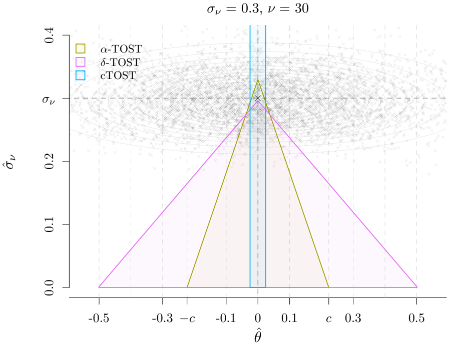
Towards Optimal Equivalence Testing

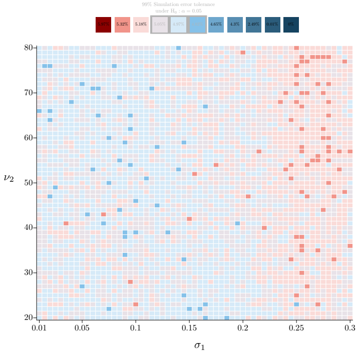
Can we do even better?

- Yes! By simultaneously optimizing for α and c:

$$[\alpha^*, c^*] \in \mathcal{A} = \underset{\substack{\gamma \in (0,1/2] \\ \delta \geq 0}}{\text{argzero}} \; \omega(\gamma, \delta, \sigma_\nu, \nu) - \alpha.$$

- We show that fixing $\alpha^* = 0.5$ and optimizing for c yields the uniformly most powerful test. We call it cTOST.





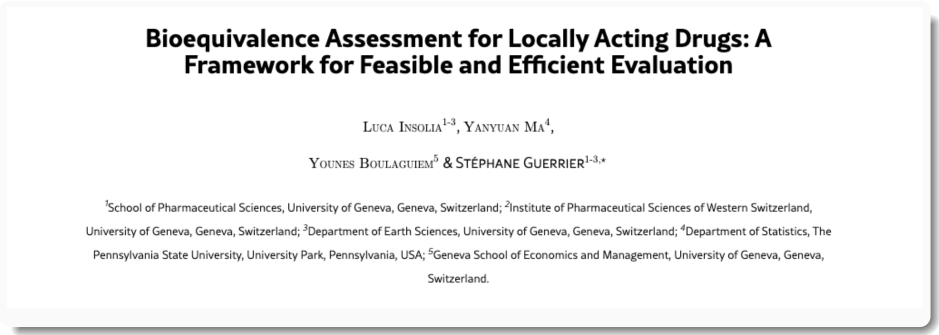
Towards Optimal Equivalence Testing

Can we do even better?

- Yes! By simultaneously optimizing for α and c:

$$[\alpha^*, c^*] \in \mathcal{A} = \underset{\substack{\gamma \in (0,1/2] \\ \delta \geq 0}}{\text{argzero}}\ \omega(\gamma, \delta, \sigma_\nu, \nu) - \alpha.$$

- We show that fixing $\alpha^* = 0.5$ and optimizing for c yields the uniformly most powerful test. We call it cTOST.



Bioequivalence Assessment for Locally Acting Drugs: A Framework for Feasible and Efficient Evaluation

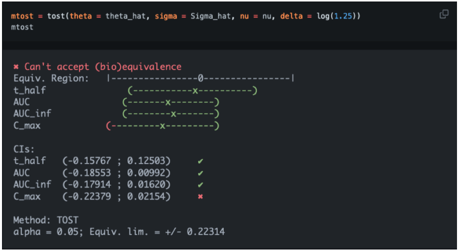
Luca Insolia[1-3], Yanyuan Ma[4],

Younes Boulaguiem[5] & Stéphane Guerrier[1-3,⋆]

[1]School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland; [2]Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland; [3]Department of Earth Sciences, University of Geneva, Geneva, Switzerland; [4]Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania, USA; [5]Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland.

cTOST R Package is in CRAN and GitHub!

→ Example of output with the *ticlopidine hydrochloride* data:

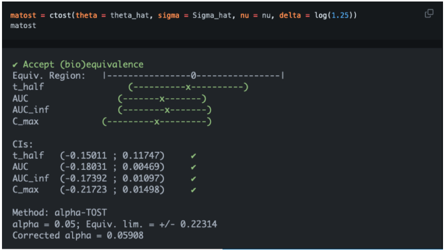


```
mtost = tost(theta = theta_hat, sigma = Sigma_hat, nu = nu, delta = log(1.25))
mtost
```

```
✖ Can't accept (bio)equivalence
Equiv. Region:    |----------------0----------------|
t_half                (-----------x----------)
AUC                  (-------x--------)
AUC_inf              (--------x-------)
C_max             (---------x---------)

CIs:
t_half   (-0.15767 ; 0.12503)     ✔
AUC      (-0.18553 ; 0.00992)     ✔
AUC_inf  (-0.17914 ; 0.01620)     ✔
C_max    (-0.22379 ; 0.02154)     ✖

Method: TOST
alpha = 0.05; Equiv. lim. = +/- 0.22314
```



```
matost = ctost(theta = theta_hat, sigma = Sigma_hat, nu = nu, delta = log(1.25))
matost
```

```
✔ Accept (bio)equivalence
Equiv. Region:    |----------------0----------------|
t_half                 (----------x----------)
AUC                  (-------x-------)
AUC_inf              (--------x-------)
C_max             (---------x---------)

CIs:
t_half   (-0.15011 ; 0.11747)     ✔
AUC      (-0.18031 ; 0.00469)     ✔
AUC_inf  (-0.17392 ; 0.01097)     ✔
C_max    (-0.21723 ; 0.01498)     ✔

Method: alpha-TOST
alpha = 0.05; Equiv. lim. = +/- 0.22314
Corrected alpha = 0.05908
```

In sum...

This work helps achieve:

- Safer swaps,
- Fewer arguments about “no significant difference”,
- Overall, **more correct decisions**.

In sum...

This works helps achieve:

- Safer swaps,
- Fewer arguments about “no significant difference”,
- Overall, **more correct decisions**.

References

[1] Stefan Wellek. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, 2010.

[2] Donald J Schuirmann. 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*, 15(6):657–680, 1987.

[3] D. Labes and H. Schütz. Inflation of type I error in the evaluation of scaled average bioequivalence, and a method for its control. *Pharmaceutical Research*, 33: 2805–2814, 2016.

[4] R. L. Berger and J. C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11:283–319, 1996.

[5] Julie Quartier, Ninon Capony, Maria Lapteva, and Yogeshvar N Kalia. Cutaneous biodistribution: a high-resolution methodology to assess bioequivalence in topical skin delivery. *Pharmaceutics*, 11(9):484, 2019.

[6] P. Pallmann and T. Jaki. Simultaneous confidence regions for multivariate bioequivalence. *Statistics in Medicine*, 36(29):4585–4603, 2017.

[7] Antonio Marzo, Lorenzo Dal Bo, Antonio Rusca, and Pierangelo Zini. Bioequivalence of ticlopidine hydrochloride administered in single dose to healthy volunteers. *Pharmacological Research*, 46(5):401–407, 2002.