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

DOI: xxx/xxxx

RESEARCH ARTICLE

nABCD: A Normalized Metric for Comparing Effect Modifier Distributions in Multi-Regional Clinical Trials

Author One[1] | **Author Two**[2] | **Author Three**[1]

[1] Department Name, Institution Name, State, Country

[2] Department Name, Institution Name, State, Country

Correspondence
Corresponding Author, Department Name, Institution Name, Address.
Email: corresponding@institution.edu

Abstract

The ICH E17 guideline recommends regional pooling in multi-regional clinical trials (MRCTs) based on similarity of effect modifier (EM) distributions, but provides no specific methodology for quantifying such similarity. Existing approaches focus on location differences (standardized mean difference) or lack interpretable scales (Kolmogorov–Smirnov statistic). We propose the normalized Area Between Cumulative Distributions (nABCD), defined as the Wasserstein-1 distance between two distributions divided by twice the pooled interquartile range. Through its connection to the treatment effect heterogeneity bound, nABCD enables clinical calibration: the maximum potential regional treatment effect difference (Δ_{max}) can be expressed as a function of nABCD, the pooled IQR, and the CATE sensitivity of the EM, with bootstrap confidence intervals quantifying estimation uncertainty. Simulation studies across location, scale, and shape scenarios at sample sizes typical in MRCT regional subgroups demonstrated satisfactory bias and coverage of the bootstrap-based estimator. Unlike the standardized mean difference, nABCD captured variance and shape differences that are invisible to location-only metrics. Application to a diabetes MRCT illustrated that clinical calibration provides context-dependent interpretation: the same distributional difference carries different clinical implications depending on the strength of effect modification. The nABCD framework fills a methodological gap in ICH E17 implementation by translating distributional differences into potential treatment effect heterogeneity on the clinical scale, supporting evidence-based pooling decisions.

KEYWORDS

multi-regional clinical trial; ICH E17; effect modifier; Wasserstein distance; regional pooling; distributional similarity

1 | INTRODUCTION

Multi-regional clinical trials (MRCTs), conducted across multiple countries or regulatory regions under a single protocol, have become the standard paradigm for global pharmaceutical development.[1,2] This approach offers substantial benefits: accelerated timelines, broader generalizability, and earlier access to new therapies for patients worldwide. The International Council for Harmonisation (ICH) E17 guideline, adopted in 2017, established principles for planning and designing MRCTs, with a central assumption that treatment effects are generalizable across the target population.[3]

A key strategy for enhancing the ability to asess regional consistency in treatment outcomes is the regional pooling approach, wherein regions with similar patient characteristics are grouped for analysis.[3] The ICH E17 guideline explicitly recommends that pooling decisions be based on the similarity of effect modifier (EM) distributions:

> "Regions may be pooled for randomisation and/or analysis if subjects are thought to be *similar enough* with respect to intrinsic and/or extrinsic factors relevant to the disease and/or drug under study." (ICH E17, Section 2.2.5)

Abbreviations: ABCD, area between cumulative distributions; CATE, conditional average treatment effect; CDF, cumulative distribution function; CI, confidence interval; CKD, chronic kidney disease; EM, effect modifier; EU, European Union; ICH, International Council for Harmonisation; IQR, interquartile range; KS, Kolmogorov–Smirnov; MRCT, multi-regional clinical trial; nABCD, normalized area between cumulative distributions; NMPA, National Medical Products Administration; SMD, standardized mean difference; US, United States.

An effect modifier is a baseline patient characteristic—such as age, disease severity, or genetic marker—for which the treatment benefit differs across subgroups. For example, if younger patients respond better to treatment than older patients, age is an effect modifier. When such heterogeneity exists, even if the drug works identically at the individual level, regions with different patient compositions may observe different average treatment effects. A region with predominantly younger patients would show larger benefits than a region with predominantly older patients, not because the drug works differently, but because the patient mix differs. This fundamental relationship underscores why EM distributional similarity is critical to the validity of regional pooling.

Despite the regulatory importance of EM distributional similarity, current practice lacks a standardized quantitative methodology. The ICH E17 guideline provides no specific metric, threshold, or statistical procedure for determining when distributions are "similar enough." Recent regulatory guidance has highlighted this gap. Song et al., writing from the China NMPA perspective on ICH E17 implementation, note the challenge of operationalizing pooling criteria without quantitative tools.[4] Long et al. further discuss basic considerations for consistency evaluation under ICH E17.[5]

Current approaches to assessing distributional similarity have significant limitations (Table 1). The standardized mean difference, while widely used for baseline covariate comparisons,[6] fundamentally cannot detect differences in variance or distributional shape—precisely the types of differences that may drive treatment effect heterogeneity through effect modification.

TABLE 1 Limitations of current approaches to distributional similarity assessment.

Method	Limitation
Visual inspection	Subjective, not reproducible
Standardized mean difference (SMD)	Captures only location, ignores scale and shape
Kolmogorov–Smirnov statistic	No interpretable scale for decision-making

This paper addresses the methodological gap by proposing the *normalized Area Between Cumulative Distributions* (nABCD), a metric for comparing EM distributions across regions, together with a clinical calibration framework that translates distributional differences into potential treatment effect heterogeneity on the outcome scale. Our specific research question is:

> How can we estimate distributional similarity between regions in a scale-free manner, and translate that estimate into clinically interpretable information about potential treatment effect heterogeneity?

The emphasis is on *estimation and clinical interpretation*, not hypothesis testing. We seek to provide regulatory scientists with quantitative tools that inform deliberation, not with binary accept/reject rules. This design philosophy reflects the ICH E17 principle that pooling decisions require contextual judgment rather than mechanical application of thresholds.

The nABCD metric measures the total area between two cumulative distribution functions, normalized by the pooled interquartile range to achieve scale-free interpretation. The framework offers four contributions:

1. *Full distributional comparison*: The Wasserstein-1 distance captures differences in location, scale, and shape simultaneously.[7]
2. *Scale-free estimation*: Normalization by IQR yields a unitless measure, enabling comparisons across EMs measured on different scales. Bootstrap confidence intervals quantify estimation uncertainty.
3. *Clinical calibration*: Through the heterogeneity bound (Proposition 2), nABCD estimates can be translated into potential treatment effect differences ($\Delta_{\max}$) specific to each EM's clinical context, expressed on the primary outcome scale.[8]
4. *Sensitivity analysis*: Because the calibration depends on the CATE sensitivity parameter L, which may be uncertain, the framework naturally accommodates sensitivity analysis over L, providing a richer evidence base for decision-making than a single binary test.

The remainder of this paper is organized as follows. Section 2 presents the methodological framework. Section 3 describes a comprehensive simulation study. Section 4 illustrates application to an MRCT dataset. Section 5 discusses implications, limitations, and future directions.

2 | METHODS

2.1 | The nABCD Metric

An effect modifier (EM) is a baseline patient characteristic for which the treatment effect varies across subgroups. Formally, let the conditional average treatment effect (CATE) be denoted $\tau(x) = E[Y(1) - Y(0) \mid X = x]$, where X is the effect modifier. When this function is non-constant, the average treatment effect observed in region r depends on the distribution F_r of the EM in that region:

$$\bar{\tau}_r = \int \tau(x)\, dF_r(x). \tag{1}$$

To formalize the relationship between distributional differences and treatment effect heterogeneity, let the CATE function be bounded with Lipschitz constant L. The difference in regional average treatment effects can be bounded by:

$$|\bar{\tau}_1 - \bar{\tau}_2| \leq L \cdot W_1(F_1, F_2), \tag{2}$$

where $W_1(F_1, F_2)$ denotes the Wasserstein-1 distance between EM distributions in regions 1 and 2.

The Wasserstein-1 distance (also known as the Earth Mover's Distance) between two cumulative distribution functions F and G is defined as:[9]

$$W_1(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)|\, dx. \tag{3}$$

Geometrically, this equals the total area between the two CDFs (Figure 1). Unlike the standardized mean difference, the Wasserstein distance responds to changes in variance, skewness, and other distributional features.[7] We use W_1 rather than higher-order Wasserstein distances because the Lipschitz bound in equation (2) requires the W_1 distance specifically. By the Kantorovich–Rubinstein duality, $W_1(F_1, F_2) = \sup_{\|f\|_{\text{Lip}} \leq 1} |\int f\, dF_1 - \int f\, dF_2|$, which directly connects Lipschitz-continuous functions (including the CATE) to distributional distance.[9] The W_2 distance, while admitting closed-form expressions for Gaussian families, does not possess this dual characterization via Lipschitz functions and therefore cannot provide the heterogeneity bound that is central to our clinical calibration framework.

The *normalized Area Between Cumulative Distributions* (nABCD) is defined as the Wasserstein-1 distance normalized by twice the pooled interquartile range:

$$\text{nABCD}(F_1, F_2) = \frac{W_1(F_1, F_2)}{2 \cdot \text{IQR}_{\text{pooled}}}, \tag{4}$$

where the pooled IQR is computed from the combined sample. The IQR-based normalization enables scale-free interpretation, is resistant to outliers, and expresses distributional differences in units of spread.

Proposition 1. (Non-negativity). *For distributions with finite IQR, nABCD ≥ 0, with equality if and only if $F_1 = F_2$.*

Proposition 2. (Connection to heterogeneity). *If the CATE function has Lipschitz constant L, then*

$$|\bar{\tau}_1 - \bar{\tau}_2| \leq 2L \cdot IQR_{\text{pooled}} \cdot nABCD(F_1, F_2). \tag{5}$$

Proof. Substituting (4) into (2) yields $|\bar{\tau}_1 - \bar{\tau}_2| \leq L \cdot W_1(F_1, F_2) = L \cdot 2 \cdot \text{IQR}_{\text{pooled}} \cdot \text{nABCD}(F_1, F_2) = 2L \cdot \text{IQR}_{\text{pooled}} \cdot \text{nABCD}(F_1, F_2)$. □

2.2 | Estimation

Given samples $\{X_{1,i}\}_{i=1}^{n_1}$ from region 1 and $\{X_{2,j}\}_{j=1}^{n_2}$ from region 2, nABCD is estimated using empirical distribution functions:

$$\widehat{\text{nABCD}} = \frac{\sum_{k=1}^{n_1+n_2-1} |\hat{F}_1(x_{(k)}) - \hat{F}_2(x_{(k)})| \cdot (x_{(k+1)} - x_{(k)})}{2 \cdot \widehat{\text{IQR}}_{\text{pooled}}}, \tag{6}$$

where $x_{(1)} < \cdots < x_{(n_1+n_2)}$ are the combined order statistics.

Computational complexity: $O((n_1 + n_2)\log(n_1 + n_2))$, dominated by sorting.

We employ the nonparametric percentile bootstrap for inference[10,11] with $B = 2{,}000$ replicates.

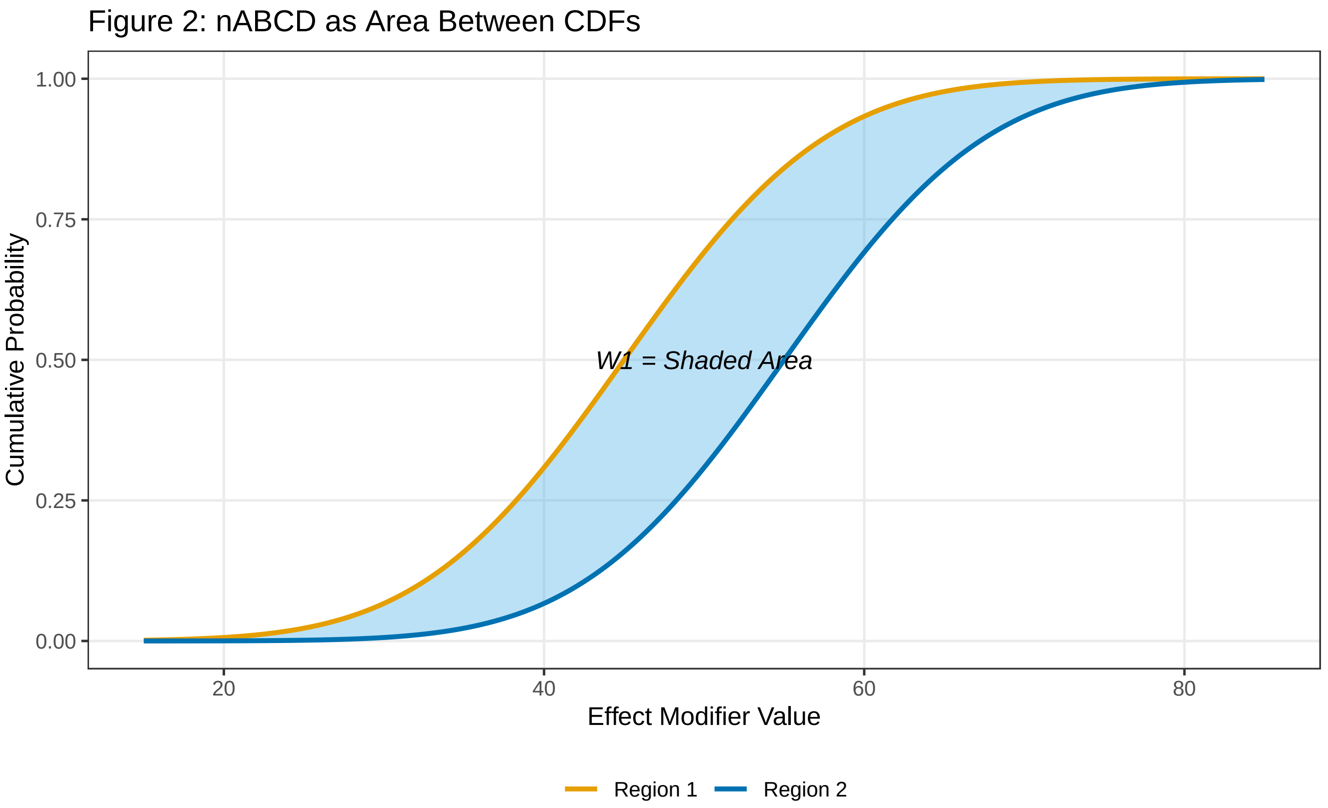

FIGURE 1 nABCD as the area between cumulative distribution functions. The shaded region represents the Wasserstein-1 distance $W_1(F_1, F_2)$, which equals the total area between the two CDFs. nABCD normalizes this area by twice the pooled IQR to achieve scale-free interpretation.

2.3 | Interpretation and Clinical Calibration

A central feature of nABCD is its connection to treatment effect heterogeneity through Proposition 2. This connection provides the basis for a clinically grounded interpretation framework rather than reliance on fixed thresholds alone.

2.3.1 | Clinical Calibration via the Heterogeneity Bound

From equation (5), the maximum potential difference in regional average treatment effects attributable to EM distributional differences is:

$$\Delta_{\max} = 2L \cdot \mathrm{IQR}_{\text{pooled}} \cdot \mathrm{nABCD}(F_1, F_2), \tag{7}$$

where L is the Lipschitz constant of the CATE function $\tau(x)$ with respect to the EM, reflecting the clinical sensitivity of treatment effect to that EM. This use of L as a sensitivity parameter follows the framework proposed by Armstrong and Kolesár,[12] who recommended reporting confidence intervals for a range of plausible Lipschitz constants. For a given EM, L can be estimated from prior knowledge, pilot data, or published subgroup analyses.

The clinical calibration procedure is as follows:

1. Identify candidate EMs and compute nABCD with bootstrap CIs for each EM across region pairs.
2. For each EM, estimate or bound L from prior knowledge (e.g., subgroup analyses showing treatment effect varies by $\Delta\tau$ per unit change in the EM, giving $L \approx \Delta\tau/\Delta x$).
3. Compute $\Delta_{\max}$ using equation (7). This represents the worst-case treatment effect difference attributable to the distributional difference.

4. Derive the bootstrap CI for $\Delta_{\max}$ from the nABCD confidence interval: $[\Delta_{\max,L}, \Delta_{\max,U}] = 2L \cdot \mathrm{IQR}_{\text{pooled}} \cdot [\mathrm{nABCD}_L, \mathrm{nABCD}_U]$. This expresses inferential uncertainty directly on the clinical scale.
5. Report $\Delta_{\max}$ and its CI alongside the clinical context—for example, the overall treatment effect size, the non-inferiority margin, or the minimal clinically important difference. This information enables trial designers and regulatory scientists to exercise clinical judgment about pooling, considering the totality of evidence rather than a binary accept/reject decision.

This estimation-centered approach deliberately avoids forcing pooling decisions into a binary hypothesis testing framework. The rationale is threefold. First, the ICH E17 guideline describes “similar enough” as inherently context-dependent, and reducing this judgment to a rejection threshold risks oversimplification. Second, the uncertainty in L means that $\Delta_{\max}$ itself is subject to sensitivity analysis (Section 4); a binary test cannot naturally accommodate this layered uncertainty. Third, consistent with recent calls to move beyond dichotomous statistical decisions,[13] presenting $\Delta_{\max}$ with its CI provides richer information for regulatory deliberation than a p-value or a reject/fail-to-reject outcome.

When regulatory agencies require a formal decision rule, the estimation framework can be adapted: declare pooling acceptable if the upper bound of the 95% CI for $\Delta_{\max}$ falls below a pre-specified clinical margin Δ_{clin}. However, we recommend this as a supplementary rather than primary use of the nABCD framework.

2.3.2 | Reference Benchmarks

When prior knowledge of L is unavailable, the benchmarks in Table 2 provide a convenience reference for initial assessment. These assume moderate CATE sensitivity and should be interpreted cautiously.

TABLE 2 Reference benchmarks for nABCD values when CATE sensitivity L is unknown.

nABCD Range	Interpretation	Suggested Action
< 0.05	Negligible difference	Pooling broadly supportable
0.05–0.15	Small difference	Pooling generally acceptable
0.15–0.30	Moderate difference	Proceed with clinical calibration
> 0.30	Large difference	Clinical calibration essential

These benchmarks are convenience references assuming moderate CATE sensitivity. Actual pooling decisions should use clinical calibration (equation 7) whenever possible.

3 | SIMULATION STUDY

We conducted simulation studies to evaluate the estimation properties of the nABCD estimator: bias, variability, and coverage probability of bootstrap confidence intervals. These properties are essential for the clinical calibration framework, as reliable estimation of nABCD directly determines the quality of $\Delta_{\max}$ estimates. We assessed performance across a range of scenarios relevant to MRCT applications.

3.1 | Simulation Design

3.1.1 | Scenarios

We designed two sets of scenarios. First, systematic scenarios for methodological validation examined controlled distributional differences: null (identical distributions), location shifts of 0.2, 0.5, and 1.0 standard deviations, scale difference (1.5-fold increase in standard deviation), and shape difference (Normal versus Gamma). Table 3 summarizes these scenarios with their true nABCD values.

Second, realistic clinical scenarios examined effect modifiers commonly encountered in MRCTs: BMI comparing Japan ($\mu = 23$, $\sigma = 3$) versus US ($\mu = 28$, $\sigma = 5$), age in elderly trials comparing Japan ($\mu = 72$, $\sigma = 8$) versus US ($\mu = 68$, $\sigma = 10$),

TABLE 3 Systematic simulation scenarios.

ID	Description	Distribution 1	Distribution 2	True nABCD
S01	Null	$N(50, 10^2)$	$N(50, 10^2)$	0.000
S03	Location 0.2σ	$N(50, 10^2)$	$N(52, 10^2)$	0.074
S04	Location 0.5σ	$N(50, 10^2)$	$N(55, 10^2)$	0.186
S05	Location 1.0σ	$N(50, 10^2)$	$N(60, 10^2)$	0.372
S06	Scale 1.5×	$N(50, 10^2)$	$N(50, 15^2)$	0.148
S08	Shape	$N(50, 10^2)$	Gamma(25, 0.5)	0.067

The Gamma distribution in S08 has shape parameter 25 and rate 0.5, yielding mean 50 and standard deviation 10.

eGFR in CKD populations, and HbA1c in diabetes trials. These parameters were informed by published literature on regional differences in patient characteristics.[14]

3.1.2 | Simulation Parameters

For each scenario, we generated samples of size n = 50, 100, and 200 per region, reflecting sample sizes typical in MRCT regional subgroups. We performed 10,000 replications per scenario–sample size combination to ensure stable estimates of operating characteristics, with Monte Carlo standard errors below 0.005 for all reported proportions. Bootstrap confidence intervals were computed using B = 2,000 resamples. All simulations were conducted in R version 4.3.3.

3.1.3 | Evaluation Metrics

Consistent with the estimation-centered framework, we evaluated:

1. *Bias*: $\text{Mean}(\widehat{\text{nABCD}}) - \text{true nABCD}$
2. *Root mean squared error (RMSE)*: $\sqrt{\text{Mean}[(\widehat{\text{nABCD}} - \text{true nABCD})^2]}$
3. *Coverage probability*: Proportion of 95% bootstrap CIs containing the true value
4. *CI width*: Mean width of bootstrap CIs, reflecting estimation precision

3.2 | Results

3.2.1 | Point Estimation and Coverage

Table 4 presents the bias of the nABCD estimator across scenarios and sample sizes. The estimator showed positive bias under the null hypothesis (S01), with bias decreasing from 0.093 at n = 50 to 0.047 at n = 200. This positive bias is attributable to the non-negative nature of the Wasserstein distance: even when true nABCD equals zero, sampling variability produces positive estimates.

TABLE 4 Bias of nABCD estimator by scenario and sample size.

Scenario	True nABCD	n = 50	n = 100	n = 200
S01 (Null)	0.000	0.093	0.066	0.047
S03 (0.2σ)	0.074	0.039	0.018	0.007
S04 (0.5σ)	0.186	0.004	−0.003	−0.004
S05 (1.0σ)	0.372	−0.038	−0.041	−0.043
S06 (Scale)	0.148	0.001	−0.012	−0.019
S08 (Shape)	0.067	0.029	0.003	−0.015

Bias is defined as $\text{Mean}(\widehat{\text{nABCD}}) - \text{true nABCD}$. Results based on 10,000 replications with B = 2,000 bootstrap resamples.

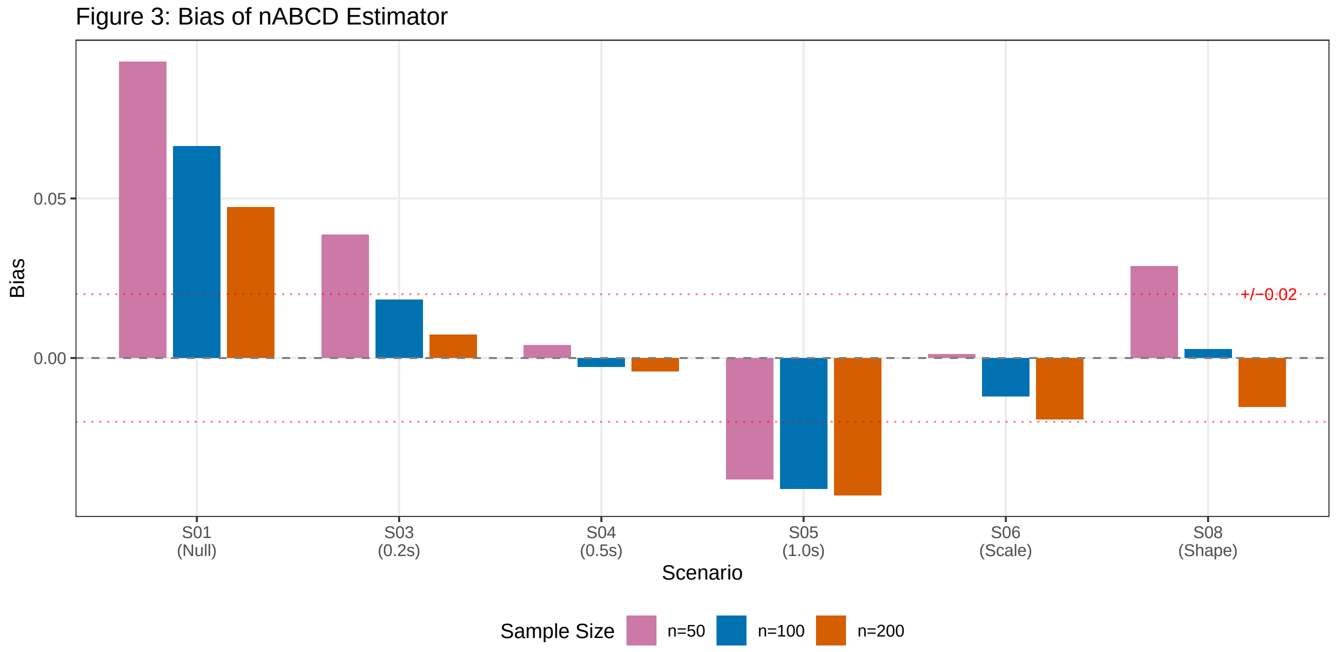

FIGURE 2 Bias of nABCD estimator by scenario and sample size. Horizontal dashed lines indicate ±0.02 bias threshold. For non-null scenarios excluding S05, bias is less than 0.02 at $n \geq 100$.

For non-null scenarios excluding S05, bias was less than 0.02 in absolute value at $n \geq 100$, indicating satisfactory point estimation performance at practical sample sizes. For S05 (1.0σ location shift), negative bias of approximately -0.04 persisted across all sample sizes, reflecting the bounded nature of nABCD near its theoretical upper range (Figure 2). These patterns were highly stable across the 10,000 replications.

Table 5 presents coverage probabilities of the 95% bootstrap confidence intervals. For $n \geq 100$, coverage was within 0.87–0.98 for most scenarios (S04, S06, S08), with moderate undercoverage for S03 at $n = 100$ (0.895). S05 showed progressive undercoverage with increasing n (0.867 at $n = 100$, 0.731 at $n = 200$), attributable to the persistent negative bias under large distributional differences. At $n = 50$, undercoverage was more pronounced, particularly for S08 (0.573) and S03 (0.672), indicating that this sample size is insufficient for reliable inference.

Coverage exhibited a non-monotonic pattern across sample sizes for the shape scenario (S08): 0.573 at $n = 50$, 0.945 at $n = 100$, and 0.996 at $n = 200$. The low coverage at $n = 50$ reflects large positive bias relative to CI width, while the overcoverage at $n = 200$ reflects the combination of shrinking bias (-0.015) with CIs that remain relatively wide for this small true value. This pattern suggests that coverage performance depends on the balance between bias magnitude and CI width, both of which change with n but at different rates.

We also evaluated BCa (bias-corrected and accelerated) confidence intervals. BCa intervals showed consistently lower coverage than percentile intervals across all scenarios, with the largest discrepancy in scale and shape scenarios (e.g., S06 at $n = 100$: percentile 0.976 vs. BCa 0.839). The BCa correction overcorrects for nABCD because the statistic is bounded below by zero, causing the acceleration parameter to distort the quantile adjustment. Based on these findings, we adopt percentile bootstrap as the primary inference method.

Monte Carlo standard errors for all reported coverage probabilities were below 0.005 at 10,000 replications, ensuring that observed differences across scenarios and sample sizes reflect genuine operating characteristic differences rather than simulation noise.

3.2.2 | Estimation Precision

Figure 3 summarizes the estimation quality across scenarios and sample sizes, presenting both coverage probabilities and mean CI widths.

TABLE 5 Coverage probability of 95% bootstrap confidence intervals.

Scenario	n = 50	n = 100	n = 200
S03 (0.2σ)	0.672	0.895	0.949
S04 (0.5σ)	0.956	0.950	0.949
S05 (1.0σ)	0.929	0.867	0.731
S06 (Scale)	0.963	0.976	0.939
S08 (Shape)	0.573	0.945	0.996

Coverage under S01 (Null) is not reported because the true value of 0 lies at the boundary of the parameter space.
For S05 (large effect), coverage decreased at larger sample sizes due to increased precision revealing the small negative bias.

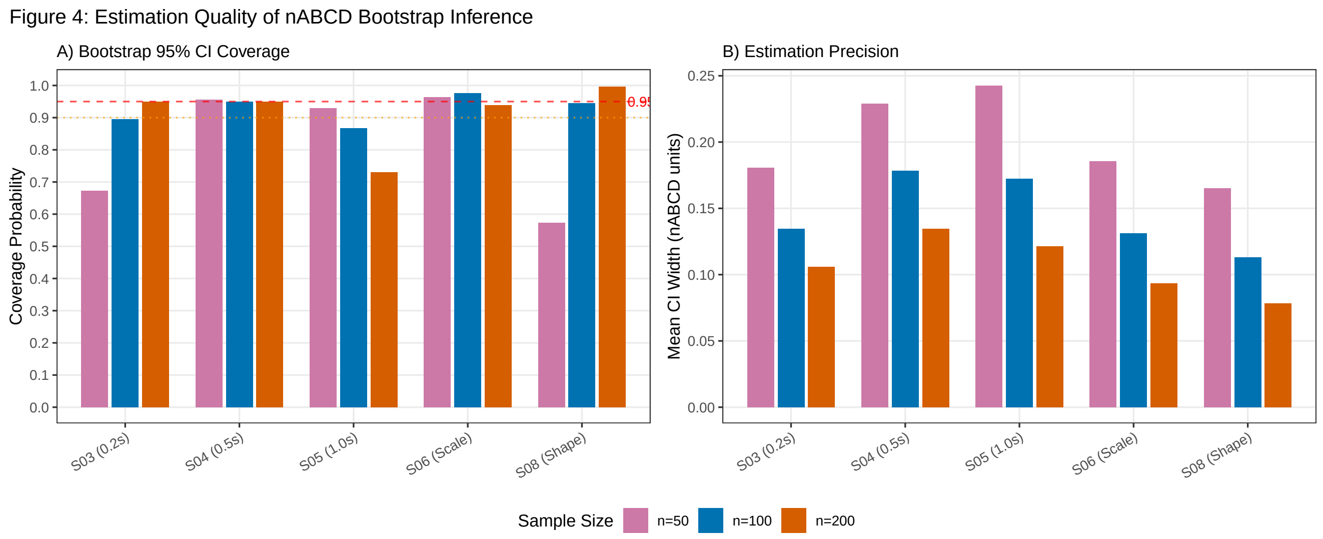

FIGURE 3 Estimation quality of nABCD bootstrap inference. Panel (A) shows coverage probability of 95% bootstrap confidence intervals; dashed and dotted lines indicate the nominal 0.95 and minimum acceptable 0.90 levels. Panel (B) shows mean CI width in nABCD units. Coverage exceeds 0.90 for most scenarios at $n \geq 100$; CI width narrows consistently with increasing n.

Table 6 presents the RMSE and mean CI width across scenarios. RMSE decreased consistently with increasing n, reaching values below 0.05 for all scenarios at $n = 200$. Mean CI width narrowed proportionally, with CIs at $n = 100$ typically spanning 0.11–0.18 in nABCD units. In the clinical calibration framework, this CI width propagates to Δ_{max}: for example, with $L = 0.3$ and IQR = 1.5, a CI width of 0.13 in nABCD corresponds to a CI width of $2 \times 0.3 \times 1.5 \times 0.13 = 0.12\%$ HbA1c—a level of precision sufficient for meaningful clinical calibration.

TABLE 6 RMSE and mean CI width by scenario and sample size.

	RMSE			Mean CI Width		
Scenario	n = 50	n = 100	n = 200	n = 50	n = 100	n = 200
S01 (Null)	0.099	0.071	0.050	0.16	0.11	0.08
S03 (0.2σ)	0.062	0.042	0.032	0.18	0.13	0.11
S04 (0.5σ)	0.066	0.049	0.036	0.23	0.18	0.13
S05 (1.0σ)	0.073	0.060	0.053	0.24	0.17	0.12
S06 (Scale)	0.045	0.035	0.030	0.19	0.13	0.09
S08 (Shape)	0.046	0.025	0.023	0.17	0.11	0.08

CI width is defined as mean of (upper − lower) across 10,000 replications.

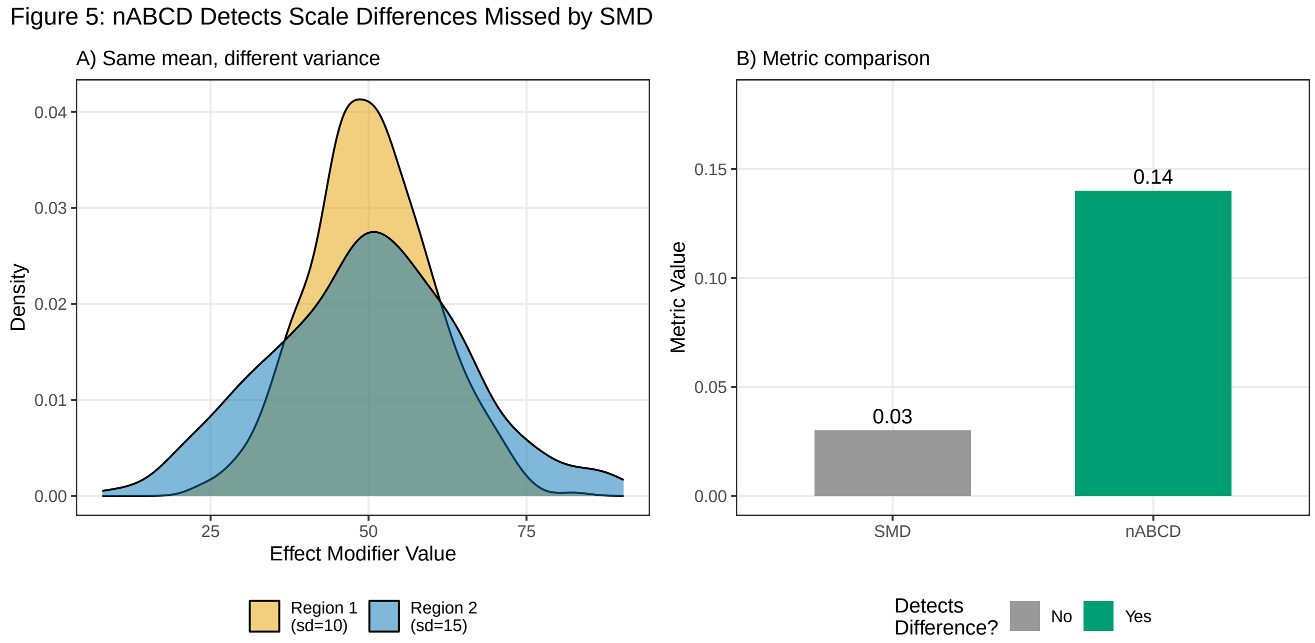

FIGURE 4 Comparison of nABCD and SMD across distributional difference types. nABCD responds to scale (S06) and shape (S08) differences where SMD remains near zero, demonstrating its advantage for full distributional comparison.

Under the null hypothesis (S01), the positive bias produces CIs that are shifted upward from zero. At n = 50, the mean estimate was 0.093 with mean CI width 0.16; at n = 200, the mean estimate decreased to 0.047 with mean CI width 0.08. This bias is relevant for clinical calibration: when the true distributional difference is zero, the estimator will suggest a small positive $\Delta_{\max}$. Practitioners should be aware of this conservative tendency, particularly at smaller sample sizes.

3.2.3 | Comparison with Standardized Mean Difference

Table 7 compares the sensitivity of nABCD and SMD to different types of distributional differences. While both metrics respond to location shifts, SMD is insensitive to differences in variance and shape—the very types of distributional differences that may drive treatment effect heterogeneity through non-linear CATE functions.

TABLE 7 Sensitivity comparison: nABCD versus SMD at n = 100.

Scenario	nABCD (mean ± SD)	SMD (mean ± SD)	Implication
S04 (Location)	0.183 ± 0.049	0.50 ± 0.14	Both detect
S06 (Scale only)	0.136 ± 0.033	0.00 ± 0.14	Only nABCD detects
S08 (Shape only)	0.070 ± 0.024	0.00 ± 0.14	Only nABCD detects

SMD depends only on the difference in means, making it insensitive to differences in variance and distributional shape. nABCD captures the full distributional difference.

In summary, the simulation study demonstrates that the nABCD estimator provides reliable estimation with the following properties at sample sizes typical in MRCT regional subgroups:

1. *Bias*: Less than 0.02 for non-null scenarios at $n \geq 100$, excluding S05 where persistent bias of -0.04 reflects boundary effects at large true values
2. *Coverage*: 0.87–0.98 at $n \geq 100$ for most scenarios
3. *Precision*: RMSE below 0.06 and CI width below 0.18 at $n \geq 100$
4. *Sensitivity*: Captures location, scale, and shape differences where SMD is limited to location only

Scenario S04 (0.5σ location shift, true nABCD = 0.186) exemplifies the operating characteristics at clinically relevant effect sizes: bias was negligible (-0.003 at n = 100), coverage was nominal (0.950), and CI width (0.18) translated to a $\Delta_{\max}$ CI width of $2 \times 0.3 \times 1.5 \times 0.18 = 0.16\%$ HbA1c when calibrated with L = 0.3 and IQR = 1.5—a level of precision sufficient for informed regulatory deliberation.

These estimation properties ensure that nABCD, when combined with the clinical calibration framework, provides $\Delta_{\max}$ estimates of sufficient quality for regulatory deliberation. We recommend $n \geq 100$ per region for reliable estimation and inference.

4 | APPLICATION

4.1 | Example: Type 2 Diabetes MRCT

We illustrate the nABCD clinical calibration framework using a hypothetical MRCT in type 2 diabetes with three regions: Japan (n = 150), United States (n = 200), and European Union (n = 180). The primary endpoint was change in HbA1c (%) at 24 weeks, with an overall treatment effect of -0.8% and a pre-specified non-inferiority margin of $\Delta_{\text{clin}} = 0.4\%$ HbA1c. Table 8 presents baseline characteristics by region.

TABLE 8 Baseline characteristics by region in the hypothetical diabetes MRCT.

Characteristic	**Japan** (n = 150)	**US** (n = 200)	**EU** (n = 180)
Age, mean (SD)	62.3 (10.2)	58.7 (11.5)	60.1 (10.8)
BMI, mean (SD)	24.8 (3.2)	32.1 (5.8)	29.4 (4.9)
HbA1c, mean (SD)	7.6 (0.9)	8.4 (1.3)	8.1 (1.1)

Table 9 presents pairwise nABCD values with 95% bootstrap CIs for the three candidate effect modifiers.

TABLE 9 Pairwise nABCD values with 95% bootstrap confidence intervals.

Effect Modifier	**Japan vs. US**	**Japan vs. EU**	**US vs. EU**
Age	0.12 (0.07–0.18)	0.08 (0.04–0.13)	0.05 (0.02–0.09)
BMI	0.51 (0.44–0.58)	0.38 (0.31–0.45)	0.18 (0.12–0.24)
HbA1c	0.27 (0.20–0.34)	0.19 (0.13–0.26)	0.10 (0.05–0.16)

4.1.1 | Clinical Calibration

The nABCD values alone do not determine pooling decisions. We apply the clinical calibration procedure from Section 2.3 to translate distributional differences into potential treatment effect heterogeneity.

Step 1: Estimate CATE sensitivity L for each EM. From published subgroup analyses in diabetes trials, we adopt the following estimates for the Lipschitz constant of the CATE function:

- *Age*: $L_{\text{age}} \approx 0.01\%$ HbA1c per year. Subgroup analyses typically show modest age–treatment interactions, with treatment effect varying by approximately 0.1% HbA1c per decade of age.
- *BMI*: $L_{\text{BMI}} \approx 0.02\%$ HbA1c per kg/m^2. A regression analysis of DPP-4 inhibitor efficacy estimated a BMI coefficient of -0.02 per kg/m^2 (p = 0.024) for the relationship between BMI and HbA1c reduction.[15]
- *HbA1c*: $L_{\text{HbA1c}} \approx 0.3\%$ HbA1c per % HbA1c. Baseline HbA1c is the strongest effect modifier for most glucose-lowering therapies: meta-regression of 98 DPP-4 inhibitor trials estimated 0.4–0.5 percentage points greater HbA1c reduction per 1% higher baseline,[16] and a GLP-1 receptor agonist study estimated $\beta = -0.31$ per % HbA1c after correcting for regression to the mean.[17]

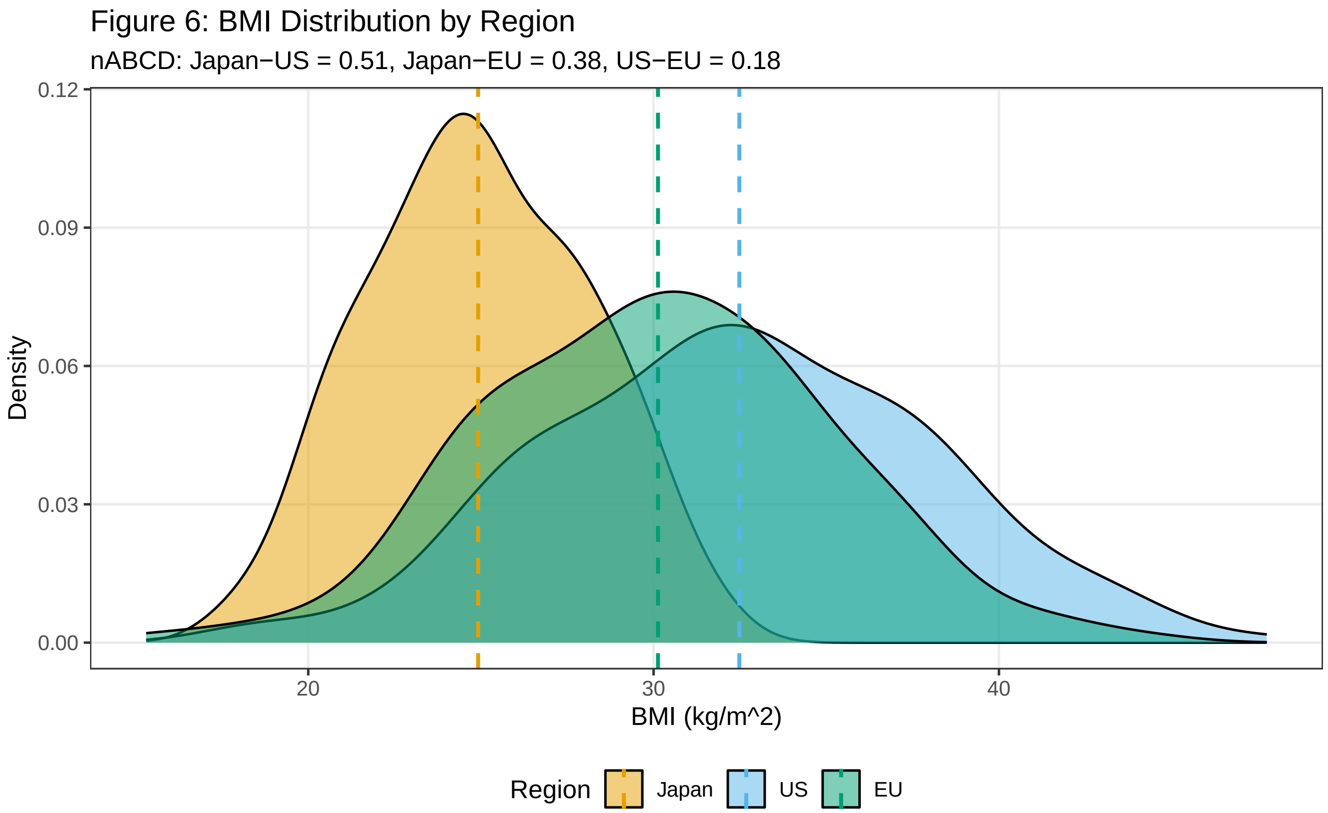

FIGURE 5 BMI distributions by region in the hypothetical diabetes MRCT. Japan shows substantially lower BMI compared to US and EU (nABCD = 0.51 for Japan–US). However, the clinical impact depends on BMI's role as an effect modifier, quantified through the heterogeneity bound $\Delta_{\max}$.

Step 2: Compute $\Delta_{\max}$ for the most discrepant region pair. Table 10 presents the clinical calibration for the Japan–US comparison, which showed the largest nABCD values across all EMs.

TABLE 10 Clinical calibration of nABCD for the Japan–US comparison.

EM	nABCD	L	IQR$_{\text{pooled}}$	$\Delta_{\max}$	vs. Δ_{clin}
Age	0.12	0.01	14.2 yr	0.03%	$\ll$ 0.4%
BMI	0.51	0.02	7.8 kg/m^2	0.16%	< 0.4%
HbA1c	0.27	0.30	1.5%	0.24%	< 0.4%

$\Delta_{\max} = 2 \times L \times \text{IQR}_{\text{pooled}} \times$ nABCD (equation 7). $\Delta_{\text{clin}} = 0.4\%$ HbA1c (non-inferiority margin).

Step 3: Interpret $\Delta_{\max}$ in clinical context.
The calibration reveals a nuanced picture that illustrates why clinical context, not nABCD alone, must drive interpretation:

- *Age* (nABCD = 0.12): $\Delta_{\max} = 0.03\%$, representing less than 4% of the overall treatment effect (−0.8% HbA1c) and less than 8% of the non-inferiority margin. Age is a weak effect modifier for this drug class, and the distributional difference between regions translates to negligible potential heterogeneity.
- *BMI* (nABCD = 0.51): Despite the large distributional difference, $\Delta_{\max} = 0.16\%$, or 20% of the overall treatment effect. BMI's low CATE sensitivity ($L = 0.02$) means that even substantial distributional differences have limited impact on treatment effect heterogeneity. Note that reference benchmarks alone would suggest "large difference"—an assessment that overlooks the weak relationship between BMI and treatment response for this drug class.

- *HbA1c* (nABCD = 0.27): $\Delta_{\max} = 0.24\%$, reaching 30% of the overall treatment effect and 60% of the non-inferiority margin. Although the nABCD is smaller than for BMI, baseline HbA1c's strong CATE sensitivity ($L = 0.30$) means the moderate distributional difference carries greater clinical weight. This comparison warrants careful deliberation.

This example demonstrates that the clinical meaning of nABCD depends on the EM's role as an effect modifier. A large nABCD for a weak EM (BMI, $L = 0.02$) may be less consequential than a moderate nABCD for a strong EM (HbA1c, $L = 0.30$). When L is uncertain, sensitivity analysis over plausible values is essential. Table 11 illustrates this for baseline HbA1c.

TABLE 11 Sensitivity analysis: $\Delta_{\max}$ as a function of assumed CATE sensitivity L for baseline HbA1c (Japan vs. US, nABCD = 0.27, IQR = 1.5%).

L (assumed)	$\Delta_{\max}$	as % of treatment effect
0.10	0.08%	10%
0.20	0.16%	20%
0.30	0.24%	30%
0.40	0.32%	40%
0.50	0.41%	51%

Overall treatment effect: −0.8% HbA1c. The value $L^* = \Delta_{\text{clin}}/(2 \cdot \text{IQR} \cdot \text{nABCD}) = 0.49$ represents the CATE sensitivity at which $\Delta_{\max}$ equals the non-inferiority margin of 0.4%.

The sensitivity analysis provides trial designers and regulators with a transparent view: for this EM and region pair, at what level of CATE sensitivity does the distributional difference begin to matter clinically? This framing supports informed deliberation rather than a binary decision, consistent with the ICH E17 principle that "similar enough" requires contextual judgment.

The estimation precision from the simulation study (Section 3) propagates directly to the precision of $\Delta_{\max}$. For HbA1c (nABCD CI width ≈ 0.14 at $n = 150$–200), the CI width in $\Delta_{\max}$ units is $2 \times 0.30 \times 1.5 \times 0.14 = 0.13\%$ HbA1c—a level of uncertainty small enough to support meaningful clinical calibration. For BMI, where $L = 0.02$, the same nABCD CI width translates to only $2 \times 0.02 \times 7.8 \times 0.14 = 0.04\%$ HbA1c, demonstrating that estimation uncertainty in $\Delta_{\max}$ scales with CATE sensitivity.

The analysis illustrates that nABCD serves as a measuring instrument for distributional distance, while the clinical judgment of whether that distance matters requires integration with domain knowledge about the EM's CATE sensitivity. The role of the statistician is to provide $\Delta_{\max}$ and its uncertainty; the role of the clinical team is to evaluate that information against the therapeutic context.

5 | DISCUSSION

We developed and validated nABCD, a normalized metric for comparing effect modifier distributions in MRCTs. Our contributions include: (1) a principled metric combining Wasserstein-1 distance with IQR normalization; (2) a theoretical framework connecting nABCD to treatment effect heterogeneity through the heterogeneity bound; (3) a clinical calibration procedure that translates nABCD values into context-specific assessments using CATE sensitivity; and (4) rigorous validation through simulation.

The nABCD metric addresses limitations of current approaches. Compared to SMD, nABCD captures full distributional differences including variance and shape. Our simulation demonstrated that nABCD captured scale differences (S06) and shape differences (S08) where SMD estimates remained near zero. Compared to the KS statistic, nABCD provides a direct connection to treatment effect heterogeneity through the heterogeneity bound (equation 7), enabling clinical calibration rather than purely statistical assessment. Compared to visual inspection, nABCD is objective and reproducible.

The clinical calibration procedure introduced in Section 2.3 represents a departure from fixed-threshold approaches. Cohen's d benchmarks (small = 0.2, medium = 0.5, large = 0.8) have been widely criticized for context-free application, and we deliberately avoid this pattern. Instead, the heterogeneity bound provides a principled mechanism to translate nABCD into clinically meaningful units. This approach respects the ICH E17 principle that "similar enough" is context-dependent: the same nABCD value may support pooling for one EM (where CATE sensitivity is low) but warrant separate analysis for another (where CATE sensitivity is high), as demonstrated in Section 4. This connection draws on transportability theory[8,18] to relate distributional distance to potential treatment effect heterogeneity.

Based on our findings, we offer the following recommendations for practitioners:

1. Compute nABCD with bootstrap CIs ($n \geq 100$ per region) for each candidate EM across region pairs.
2. Translate nABCD into $\Delta_{\max}$ and its CI on the clinical outcome scale using equation (7), incorporating prior knowledge or estimates of CATE sensitivity L.
3. Report $\Delta_{\max}$ and its CI alongside the overall treatment effect, non-inferiority margin, or other clinically relevant benchmarks to support informed deliberation about pooling.
4. Conduct sensitivity analyses over plausible values of L when prior knowledge is uncertain (Table 11).
5. When L cannot be estimated, use the reference benchmarks in Table 2 as initial guidance, recognizing that clinical calibration provides a more rigorous basis for decision-making.

In practice, multiple candidate EMs may be evaluated simultaneously. The nABCD framework assesses each EM separately, and we recommend reporting $\Delta_{\max}$ and its CI for every candidate EM across all region pairs. To arrive at an overall pooling recommendation, trial designers may adopt a conservative approach based on the maximum $\Delta_{\max}$ across all EMs and region pairs, or they may take a totality-of-evidence approach in which the collection of $\Delta_{\max}$ values, together with their uncertainties and the reliability of each L estimate, informs a holistic judgment. The choice between these strategies depends on the regulatory context: for applications where a formal pooling criterion is required, the maximum $\Delta_{\max}$ provides a defensible worst-case assessment; for advisory settings where pooling decisions are informed by clinical deliberation, the full set of calibrated results offers a richer evidence base.

Several limitations should be acknowledged:

1. The current formulation applies to continuous EMs only; extensions to categorical or mixed-type EMs require further development.
2. nABCD evaluates each EM separately rather than jointly; a multivariate extension would address potential confounding among EMs.
3. Positive bias under the null at small sample sizes ($n < 100$) can inflate nABCD estimates, and persistent negative bias at large true values (S05, nABCD = 0.372) leads to progressive undercoverage (0.731 at $n = 200$). These boundary effects are inherent to the non-negative, bounded nature of nABCD as a function of the Wasserstein distance. When the true value equals zero, the parameter lies on the boundary of the parameter space, where standard bootstrap consistency results may not apply;[19] we therefore do not report null-scenario coverage and recommend interpreting nABCD estimates near zero with caution. For non-null scenarios, the true value lies in the interior of the parameter space where standard bootstrap consistency holds, and our simulation confirms this with coverage of 0.87–0.98 at $n \geq 100$ for most scenarios. Coverage can also exhibit non-monotonic behavior: for the shape scenario (S08), coverage was 0.573 at $n = 50$, 0.945 at $n = 100$, and 0.996 at $n = 200$, reflecting how the balance between bias magnitude and CI width shifts with sample size. We use the percentile bootstrap, which is first-order accurate; bias-corrected methods such as BCa showed inferior performance for this bounded statistic (Section 3.2), and the studentized bootstrap would require variance estimation for the Wasserstein distance ratio, which adds substantial complexity. We recommend $n \geq 100$ per region for reliable point estimation and confidence intervals, and caution that coverage may degrade when the true nABCD exceeds approximately 0.3.
4. The clinical calibration requires estimation of CATE sensitivity L, which may not always be available from prior data. Reference benchmarks are provided as a fallback but should not replace context-specific calibration.

Future work should address multivariate extensions, bias correction methods for small samples, and empirical calibration of CATE sensitivity parameters using real MRCT data. Methods for estimating L from historical trial databases would strengthen the clinical calibration framework. The nABCD framework could also be extended to longitudinal EM profiles, enabling dynamic assessment of distributional similarity over time.

nABCD fills a methodological gap in ICH E17 implementation by providing a principled framework for assessing distributional similarity. Rather than reducing “similar enough” to a fixed threshold, the clinical calibration procedure translates distributional differences into context-specific assessments of potential treatment effect heterogeneity, enabling evidence-based and clinically grounded pooling decisions. Open-source R code is available to facilitate adoption.

AUTHOR CONTRIBUTIONS

[Author 1] conceived the research question and led the project. [Author 2] developed the mathematical framework and conducted the simulation study. [Author 3] contributed to the application example and manuscript preparation. All authors reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

The authors thank [colleagues] for helpful discussions on ICH E17 implementation.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

R code for computing nABCD and reproducing the simulation study is available at [repository URL].

REFERENCES

1. Chen J, Quan H, Binkowitz B, et al. Assessing consistent treatment effect in a multi-regional clinical trial: a systematic review. *Pharm Stat.* 2010;9(3):242–253. doi: 10.1002/pst.438
2. Quan H, Li M, Chen J, et al. Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Inf J.* 2010;44(5):617–632. doi: 10.1177/009286151004400515
3. ICH E17 Expert Working Group . General Principles for Planning and Design of Multi-Regional Clinical Trials (E17). tech. rep., International Council for Harmonisation; Geneva, Switzerland: 2017.
4. Song J, Ji C, Chen M, Luo X, Quan H, Chen J. Basic Considerations for Data Pooling Strategy in Multi-Regional Clinical Trials (MRCTs). *Ther Innov Regul Sci.* 2025;59(2):359–364. doi: 10.1007/s43441-025-00744-8
5. Long M, Wu H, Liu X, Chen J. Basic Considerations for the Consistency Evaluation Based on ICH E17 Guideline. *Ther Innov Regul Sci.* 2025;59(2):328–336. doi: 10.1007/s43441-024-00737-z
6. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res.* 2011;46(3):399–424. doi: 10.1080/00273171.2011.568786
7. Panaretos VM, Zemel Y. Statistical aspects of Wasserstein distances. *Annu Rev Stat Appl.* 2019;6:405–431. doi: 10.1146/annurev-statistics-030718-104938
8. Pearl J, Bareinboim E. Transportability of causal and statistical relations: A formal approach. In: . 25. AAAI Press. 2011; Menlo Park, CA:247–254.
9. Villani C. *Optimal Transport: Old and New*. Springer, 2009
10. Barrio dE, Giné E, Matrán C. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann Probab.* 1999;27(2):1009–1071. doi: 10.1214/aop/1022677394
11. Sommerfeld M, Munk A. Inference for empirical Wasserstein distances on finite spaces. *J R Stat Soc Series B.* 2018;80(1):219–238. doi: 10.1111/rssb.12236
12. Armstrong TB, Kolesár M. Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness. *Econometrica.* 2021;89(3):1141–1177. doi: 10.3982/ECTA16907
13. Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. *Am Stat.* 2016;70(2):129–133. doi: 10.1080/00031305.2016.1154108
14. Sai K, Nakatani E, Iwama Y, Ono S. Efficacy Comparison for a Schizophrenia and a Dysuria Drug Among East Asian Populations: A Retrospective Analysis Using Multi-regional Clinical Trial Data. *Ther Innov Regul Sci.* 2021;55(3):523–538. doi: 10.1007/s43441-020-00246-9
15. Kim YG, Hahn S, Oh TJ, Kwak SH, Park KS, Cho YM. Predictive Factors for Efficacy of Dipeptidyl Peptidase-4 Inhibitors in Patients with Type 2 Diabetes Mellitus. *Diabetes Metab J.* 2015;39(4):342–348. doi: 10.4093/dmj.2015.39.4.342
16. Craddy P, Palin HJ, Johnson KI. Comparative effectiveness of dipeptidylpeptidase-4 inhibitors in type 2 diabetes: a systematic review and mixed treatment comparison. *Diabetes Ther.* 2014;5(1):1–41. doi: 10.1007/s13300-014-0061-3
17. Jones AG, McDonald TJ, Shields BM, et al. Should Studies of Diabetes Treatment Stratification Correct for Baseline HbA1c?. *PLoS One.* 2016;11(4):e0152428. doi: 10.1371/journal.pone.0152428
18. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci USA.* 2016;113(27):7345–7352. doi: 10.1073/pnas.1510507113
19. Andrews DWK. Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space. *Econometrica.* 2000;68(2):399–405. doi: 10.1111/1468-0262.00114

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website. Supporting materials include: complete R code for nABCD computation, simulation scripts, and additional figures for realistic clinical scenarios.

How to cite this article: Author One, Author Two, and Author Three. nABCD: A normalized metric for comparing effect modifier distributions in multi-regional clinical trials. Statist. Med. 2025;00:1–18.

APPENDIX

A PROOFS AND MATHEMATICAL DETAILS

A.1 Proof of Proposition 1

Non-negativity follows directly from the non-negativity of the Wasserstein-1 distance: $W_1(F_1, F_2) = \int |F_1(x) - F_2(x)|\,dx \geq 0$, with equality if and only if $F_1(x) = F_2(x)$ for all x. Since $\text{IQR}_{\text{pooled}} > 0$ for non-degenerate distributions, the ratio nABCD = $W_1/(2 \cdot \text{IQR}_{\text{pooled}}) \geq 0$.

A.2 Asymptotic Properties

Under standard regularity conditions, the empirical Wasserstein-1 distance $W_1(\hat{F}_1, \hat{F}_2)$ is a consistent estimator of $W_1(F_1, F_2)$.[10] Combined with the consistency of the empirical IQR, $\widehat{\text{nABCD}}$ is a consistent estimator of nABCD(F_1, F_2).

In one dimension, the empirical W_1 distance converges at rate $O(n^{-1/2})$ without logarithmic correction.[10] For two-sample estimation with $F_1 \neq F_2$ and finite second moments, $\sqrt{n}(W_1(\hat{F}_{1,n_1}, \hat{F}_{2,n_2}) - W_1(F_1, F_2))$ converges in distribution to a mean-zero Gaussian, where $n = n_1 + n_2$ with $n_1/n \to \lambda \in (0, 1)$. Since the empirical IQR also converges at rate $O(n^{-1/2})$ under standard density conditions, the delta method applied to the ratio $g(w, q) = w/(2q)$ yields asymptotic normality of $\widehat{\text{nABCD}}$:

$$\sqrt{n}\left(\widehat{\text{nABCD}} - \text{nABCD}(F_1, F_2)\right) \xrightarrow{d} N(0, \sigma^2_{\text{nABCD}}), \tag{A1}$$

where σ^2_{nABCD} depends on the asymptotic variances and covariance of the W_1 and IQR estimators. This result requires nABCD > 0 (i.e., $F_1 \neq F_2$); at the boundary $F_1 = F_2$, the parameter lies on the edge of the parameter space and standard asymptotics do not apply (see Section **??**).

The bootstrap provides valid inference for the Wasserstein distance under mild conditions.[11] The percentile bootstrap confidence interval achieves asymptotically correct coverage for non-degenerate distributions.

B R CODE FOR NABCD COMPUTATION

Listing 1 R function for computing nABCD with bootstrap confidence intervals.

```
compute_nABCD <- function(x1, x2) {
  pooled <- c(x1, x2)
  iqr_pooled <- IQR(pooled)
  if (iqr_pooled == 0) return(NA)
  all_vals <- sort(unique(pooled))
  F1 <- ecdf(x1); F2 <- ecdf(x2)
  w1 <- sum(diff(all_vals) *
    abs(F1(all_vals[-length(all_vals)]) -
        F2(all_vals[-length(all_vals)])))
  w1 / (2 * iqr_pooled)
}

nABCD_bootstrap <- function(x1, x2,
    B = 2000, conf = 0.95) {
  obs <- compute_nABCD(x1, x2)
  boot_vals <- replicate(B, {
    b1 <- sample(x1, replace = TRUE)
    b2 <- sample(x2, replace = TRUE)
    compute_nABCD(b1, b2)
  })
  alpha <- 1 - conf
  ci <- quantile(boot_vals,
    c(alpha/2, 1 - alpha/2), na.rm = TRUE)
  list(estimate = obs,
       ci_lower = ci[1], ci_upper = ci[2])
}
```