Data-Driven Baseline Matching: Enhancing Indirect Comparisons With a Machine Learning-Informed Framework for Selecting Homogeneous Trial Sets

Saswata Paul Chaudhury[1], Sekhar Dutta[1], Subhajit Gupta[1]
[1]PharmaQuant Insights Pvt. Ltd., Kolkata, WB, India

MSR39

PharmaQuant
Focused Insights, Better Outcomes

BACKGROUND

- Indirect treatment comparisons (ITCs) and network meta-analyses (NMAs) are essential when head-to-head randomized controlled trials are unavailable, allowing cross-treatment inferences through shared comparators.[1]
- Violations of the transitivity assumption, such as imbalances in age, disease severity, biomarkers, or comorbidities across trials can lead to biased treatment-effect estimates and compromise the validity of indirect comparisons.[2]
- Population-adjustment methods like MAIC and STC attempt to address cross-trial imbalances through re-weighting or regression, but their performance degrades substantially when population overlap is limited.[2,3,4]
- The choice of which trials to compare is therefore a critical analytic decision that directly affects the credibility of results.
- Current trial selection relies on clinical judgment or pairwise standardized mean differences (SMDs) of individual covariates, lacking a systematic, multivariate framework for jointly evaluating all relevant characteristics.
- As candidate trial numbers grow, the combinatorial space of subsets expands exponentially ($2^n - 1$), making informal assessment impractical.
- Machine learning (ML) techniques such as principal component analysis (PCA) and unsupervised clustering can project high-dimensional covariate profiles into reduced spaces, enabling distance-based scoring and transparent ranking of all trial combinations by internal homogeneity.
- A unified, data-driven framework combining these techniques is currently lacking, motivating the present study.

OBJECTIVE

This study aimed to develop ML-informed methodology to quantify multivariate baseline similarity across trial combinations and provide a penalized, reproducible ranking of candidate trial pools for ITCs.

METHODS

This was a fully simulated methodological study in which synthetic patient-level data for 8 RCTs were generated and analyzed through a four-step workflow of trial-level summaries, PCA, clustering, and penalized subset similarity scoring.

Data simulation & trial-level summaries:

- Individual patient-level data was simulated for 8 RCTs (T1-T8; N = 150 to 400 per trial) with 1:1 treatment allocation.
- Ten baseline covariates were generated: age, BMI, baseline score, biomarker, sex, smoking status, stage III, lab value, tumor size, and prior treatment.
- Between-trial heterogeneity was introduced via trial-specific covariate distribution parameter.

Feature scaling & dimensionality reduction:[5]

- All features were standardized to zero mean and unit variance.
- PCA was applied to the standardized matrix; and components were retained to explain ≥ 80% cumulative variance (minimum 2 principal components [PCs]).

Clustering:[6,7]

- Hierarchical agglomerative clustering (Ward.D2, Euclidean distance) on the full feature space and a dendrogram were used to visualize trial grouping pathways.
- Gaussian mixture model-based clustering on the PC space was used as a complementary method.

Subset enumeration & similarity scoring (all 2^8 − 1 = 255 subsets):[8]

- The mean pairwise distance (MPD), defined as the average Euclidean distance among subset trials in the PC space, was calculated.
- The maximum pairwise distance (XPD), defined as the worst-case within-subset discordance, was calculated.
- The centroid Mahalanobis distance (CMD), defined as mean distance of trials to a sample-size-weighted centroid accounting for covariance structure, was calculated.
- These were then transformed to similarity scores on [0,1] via an exponential kernel: S = exp(−α·d), with α calibrated to the median pairwise distance.

RESULTS

Figure 1: Dendrogram of Hierarchical Agglomerative Clustering

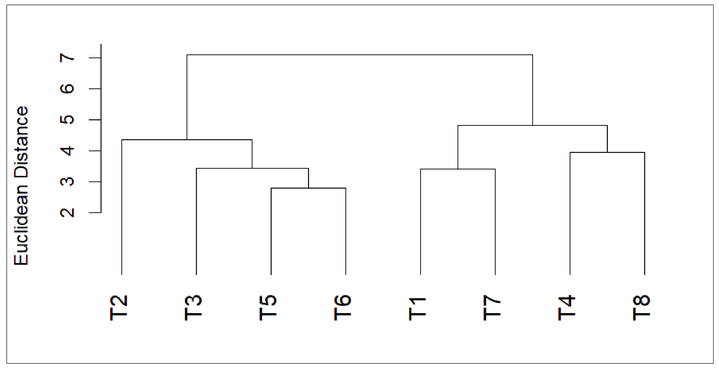
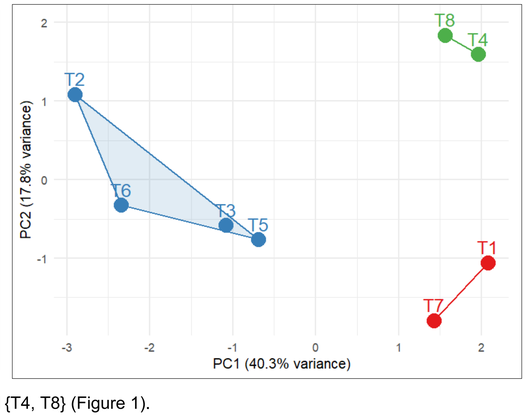


Figure 2: PCA Scatter Plot of Trial-Level Baseline Profiles



- Hierarchical clustering identified three distinct trial groups: {T2, T3, T5, T6}, {T1, T7}, and {T4, T8} (Figure 1).
- T5 and T6 merged at the lowest height (~2.8), indicating the most similar trial pair (Figure 1).
- Cluster 2 (blue: T2, T3, T5, T6) occupied the left region of PC space, confirming baseline similarity (Figure 2).
- Clusters 1 and 3 were spatially separated on the right, consistent with dendrogram findings (Figure 2).

Figure 3: Heatmap of Pairwise Euclidean Distances

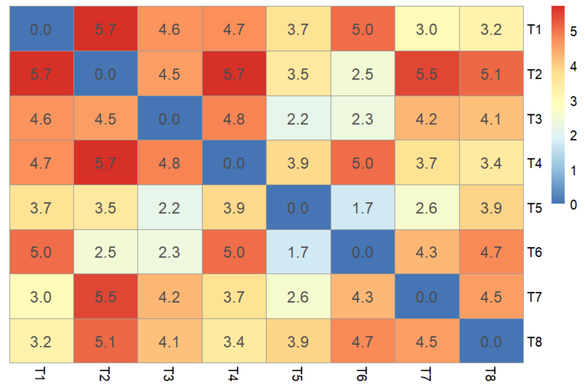
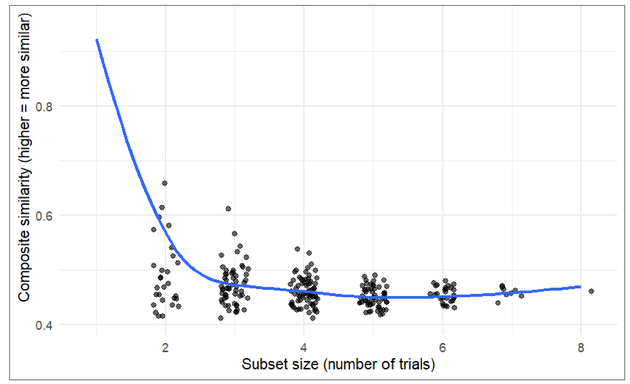


Figure 4: Penalized Composite Similarity Across Subset Sizes



- The T5-T6 pair showed the smallest pairwise distance (1.74), confirming the closest trial pair (Figure 3).
- Trials across clusters (e.g., T2-T4, T1-T6) showed the largest distances (>5.0) (Figure 3).
- Composite similarity decreased sharply from size 2 to 3, then plateaued at sizes 4-8 (Figure 4).
- Penalization successfully prevented singletons and small subsets from dominating rankings (Figure 4).
- Subsets of 2-3 trials achieved the highest composite similarity scores (Figure 4).

Table 1: Similarity Metrics for Top-Ranked Trial Subsets

Trials	Subset size	MPD	Penalized MPD	Composite similarity
T5,T6	2	1.7	2.6	0.7
T3,T5	2	2.2	3.2	0.6
T3,T6	2	2.3	3.5	0.6
T3,T5,T6	3	2.1	2.8	0.6
T2,T5,T6	3	2.6	3.4	0.6
T5,T6,T7	3	2.9	3.8	0.5
T2,T3,T5,T6	4	2.8	3.5	0.5
T3,T5,T6,T7	4	2.9	3.6	0.5
T3,T5,T6,T8	4	3.2	4.0	0.5
T2,T3,T5,T6,T7	5	3.3	4.0	0.5
T1,T3,T5,T6,T7	5	3.4	4.0	0.5
T3,T5,T6,T7,T8	5	3.5	4.2	0.5

MPD: Mean pairwise distance

- **Best 2-trial subset:** T5, T6 (MPD = 1.738; composite = 0.658)
- **Best 3-trial subset**: T3, T5, T6 (MPD = 2.076; composite = 0.611)
- **Best 4-trial subset:** T2, T3, T5, T6 (MPD = 2.794; composite = 0.537)
- All top-ranked subsets were drawn from Cluster 2, validating the clustering results.

CONCLUSION

- Current ITC workflows rely on clinical judgment for trial selection. This is subjective, hard to pre-specify, and difficult to scale as candidate trial numbers grow. This framework converts trial selection into a reproducible, auditable step. The penalized ranking can be pre-specified in HTA and regulatory submissions, replacing narrative justifications with a transparent quantitative process.
- The ranked subset list also creates possibilities for sensitivity analyses grounded in a quantitative, rather than narrative, rationale. In this analysis, clustering, PCA, and distance metrics converged on consistent groupings, reinforcing the robustness of the selection. The penalized composite index effectively balanced homogeneity against subset size, offering a transparent, reproducible alternative to subjective trial selection for ITCs.
- Future validation using real-world data is needed to assess any potential impacts on bias and precision in comparative effectiveness estimates.

REFERENCES

1. Phillippo DM, et al. *Methods for population-adjusted indirect comparisons in health technology appraisal.* Med Decis Making. 2018;38(2):200-211.
2. Signorovitch JE, et al. *Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research.* Value Health. 2012;15(6):940-947.
3. Caro JJ, Ishak KJ. *No head-to-head trial? Simulate the missing arms.* Pharmacoeconomics. 2010;28(10):957-967
4. Dias S, et al. *Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials.* Med Decis Making. 2013;33(5):641-656
5. Jolliffe IT, Cadima J. *Principal component analysis: a review and recent developments.* Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202.
6. Murtagh F, Legendre P. *Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?* J Classif. 2014;31(3):274-295.
7. De Maesschalck R, Jouan-Rimbaud D, Massart DL. *The Mahalanobis distance.* Chemom Intell Lab Syst. 2000;50(1):1-18.
8. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* 2nd ed. Springer; 2009

Presented at ISPOR US 2026; Philadelphia, Pennsylvania, United States; May 17-20, 2026