Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

LINKO (Latent Information Normalization for Key Outcomes): A Framework for Evaluating the Validity of Meta-Analytic Pooling Across Heterogeneous RCT Data Structures

Onishi Tatsuki

bougtoir@gmail.com

Shiga University

Research Article

Keywords: meta-analysis, heterogeneity, information contribution ratio, evidence synthesis, principal component analysis, individual patient data, forest plot

Posted Date: June 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9338552/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

LINKO (Latent Information Normalization for Key Outcomes): A Framework for Evaluating the Validity of Meta-Analytic Pooling Across Heterogeneous RCT Data Structures

Onishi Tatsuki

Data Science AI Innovation Research Promotion Center, Shiga University

Corresponding author: Onishi Tatsuki, bougtoir@gmail.com

Abstract

Background: Meta-analysis pools effect sizes from multiple RCTs, but each RCT collects different numbers of variables, meaning the endpoint may represent vastly different proportions of the total information space. This structural heterogeneity is not captured by I-squared or tau-squared. We introduce the LINKO (Latent Information Normalization for Key Outcomes) framework, proposing the Information Contribution Ratio (ICR) as a diagnostic measure for assessing whether endpoint variables carry equivalent informational weight across studies in a meta-analysis.

Methods: We define ICR as the proportion of a study's total data information attributable to its endpoint. Two approaches are developed: (1) a variance-based approach (ICR_std = d/D) computable from published Table 1 statistics, and (2) a PCA-based approach (ICR_pca) using individual patient data. We introduce the Prism Forest Plot encoding ICR as color and size dimensions. Validation includes Monte Carlo simulation (ADEMP framework), two real-world domains, PCA validation using the IST dataset (19,435 patients, 8 country sub-studies), and leave-one-out sensitivity analysis.

Results: In simulation, heterogeneous ICR yielded higher I-squared (11.7% vs 11.0%). In real-world data, statin therapy (low ICRD = 0.009) showed I-squared = 0.0%, while glucose control (ICRD = 0.048) showed I-squared = 17.0%. PCA-based ICR varied 4-fold across IST sub-studies (CV = 0.36); regression-based ICR_pca correlated with 14-day mortality ($r = 0.90$, $p = 0.003$), robust in leave-one-out analysis ($r = 0.84$-0.95, all $p < 0.02$).

Conclusions: LINKO provides a computable diagnostic for assessing structural comparability in meta-analysis. We recommend reporting ICR discrepancy alongside I-squared and tau-squared.

Keywords: meta-analysis, heterogeneity, information contribution ratio, evidence synthesis, principal component analysis, individual patient data, forest plot

Background

Randomized controlled trials (RCTs) are the gold standard for evaluating treatment efficacy. Each RCT collects comprehensive patient-level data: demographic variables, baseline laboratory values, comorbidities, concomitant medications, and multiple outcome measures. A typical RCT may record 10-100+ variables per participant. Yet when these studies are synthesized through meta-analysis, the pooling is performed on a single endpoint (or a small number of endpoints), reducing each study's rich dataset to a single effect size estimate.

This raises a fundamental question: does the endpoint carry the same informational weight in every study? If Study A measures 10 variables and the endpoint is one of them (representing 10% of the data dimensionality), while Study B measures 50 variables with the same endpoint (representing 2% of the data dimensionality), are these effect sizes truly comparable for pooling?

Standard meta-analysis assesses between-study heterogeneity using Cochran's Q statistic and the I-squared index [1]. These measures quantify the proportion of variability in effect estimates that is due to true differences rather than sampling error. However, they do not account for differences in the informational context from which the endpoint was drawn. Sources of heterogeneity are typically attributed to clinical diversity (different populations, interventions, comparators), methodological diversity (study design, risk of bias), and statistical heterogeneity (unexplained variation). We propose that structural diversity -- differences in the data dimensionality and the endpoint's role within each study's data space -- represents an additional, previously unrecognized source of heterogeneity.

We introduce the LINKO (Latent Information Normalization for Key Outcomes) framework, centered on the Information Contribution Ratio (ICR) as a measure of how much of a study's total information is captured by its endpoint variables. We develop two complementary approaches: (1) a variance-based approach (ICR_v) computable from published Table 1 summary statistics, making retrospective analysis feasible for any published RCT; and (2) a PCA-based approach (ICR_pca) that captures the full covariance structure when individual patient data (IPD) are available. We also introduce the Prism Forest Plot, a novel visualization that extends the standard forest plot by encoding ICR information through color and size dimensions.

We hypothesize that when studies contributing to a meta-analysis have similar ICR values (low ICR Discrepancy, ICRD), the meta-analysis will exhibit lower heterogeneity. Conversely, when ICR varies substantially across studies (high ICRD), heterogeneity may increase because the endpoint captures different fractions of the total treatment effect in different studies.

Methods

Mathematical framework

Variance-based ICR (Approach 1: ICR_v)

For a study with D measured variables, the standardized ICR is defined as:

$$ICR_std = d / D$$

where d is the number of endpoint variables and D is the total number of variables. Under standardization (each variable scaled to unit variance), this represents the fraction of the data space occupied by the endpoint.

The raw variance-based ICR captures scale-dependent information:

$$ICR_raw = SUM(Var(X_j) \text{ for } j \text{ in } E) / SUM(Var(X_j) \text{ for } j=1..D)$$

where E is the set of endpoint variables. For continuous variables, the pooled variance is reconstructed from group means, standard deviations, and sample sizes reported in Table 1. For binary variables, the variance is p(1-p). This allows retrospective ICR computation for any published RCT without requiring individual patient data.

PCA-based ICR (Approach 2: ICR_pca)

When individual patient data (IPD) are available, a more refined measure uses principal component analysis to capture the full covariance structure. We implement two PCA-based methods:

Loading-based method: Identifies principal components (PCs) where the endpoint variable has an absolute loading exceeding a threshold (|a_jk| >= 0.3), then sums their explained variance ratios:

ICR_pca_loading = SUM(lambda_k / SUM(lambda) for k in S_E)

where S_E is the set of endpoint-dominant principal components.

Regression-based method: Performs PCA on predictor variables only (excluding the endpoint), then regresses the endpoint on all PC scores. Since PCs are orthogonal, the regression coefficient for each PC is beta_k = Cov(Y, PC_k) / Var(PC_k). The ICR is computed as:

*ICR_pca_reg = SUM(beta_k^2 * lambda_k) / (SUM(lambda_k) + Var(Y))*

This measures the proportion of total data information (predictor eigenvalues plus endpoint variance) that "flows to" the endpoint through the principal component structure.

ICR Discrepancy (ICRD)

For a set of K studies in a meta-analysis, the ICR Discrepancy is:

ICRD = max(ICR_i) - min(ICR_i), i = 1, ..., K

We also compute the coefficient of variation of ICR (ICR_CV) as a complementary measure.

Simulation study design (ADEMP framework)

We report the simulation study following the ADEMP framework [3,4] for transparent reporting of simulation studies evaluating statistical methods.

Aims: To evaluate whether heterogeneity in ICR across studies affects the statistical heterogeneity (I-squared) of meta-analytic pooling, and whether adding studies with divergent ICR to an initially homogeneous set of studies increases I-squared.

Data-generating mechanisms: Each simulated RCT generates multivariate normal data with D dimensions, where the treatment has a direct effect (delta = 0.5, standardized) on the endpoint (dimension 0) and spillover effects on other dimensions proportional to their correlation with the endpoint. The spillover fraction was set to 0.3. The correlation matrix uses an exponential decay structure: rho_ij = 0.5^|i-j|. Each study has n = 200 subjects per arm.

Three scenarios were simulated:

Scenario A (Uniform ICR): 10 studies, each with D = 20. ICR_std = 1/20 = 0.05.

Scenario B (Heterogeneous ICR): 10 studies with D from {5, 10, 20, 40, 80}. ICR_std varies 0.0125-0.20.

Scenario C (Sequential with ICR Shift): 15 studies. First 5 with D = 20, then 10 with D from {5, 10, 40, 60, 80}.

Estimands: The primary estimand is the I-squared heterogeneity statistic from DerSimonian-Laird random-effects meta-analysis. Secondary: pooled effect, tau-squared, correlation between ICRD and I-squared.

Methods: DerSimonian-Laird random-effects meta-analysis [5]. For Scenario C, sequential (cumulative) meta-analysis.

Performance measures: Mean I-squared (SD) across 100 iterations, mean ICRD, Pearson correlation between ICRD and I-squared, mean pooled effect (bias from true delta = 0.5), proportion of simulations where adding heterogeneous-ICR studies increased I-squared.

Real-world data analysis

We selected two well-known meta-analysis domains:

Statin therapy for cardiovascular prevention: Five landmark RCTs (4S, WOSCOPS, CARE, LIPID, AFCAPS/TexCAPS) [6-10] with consistent mortality benefit. Similar variable sets (D = 10-11).

Intensive glucose control in type 2 diabetes: Four major RCTs (UKPDS 33, ACCORD, ADVANCE, VADT) [11-14] where heterogeneity emerged. Different variable counts (D = 8-13).

PCA-based ICR validation (individual patient data)

We used the International Stroke Trial (IST) dataset [15,16], comprising 19,435 patients across 36 countries. We encoded 25 analysis variables and treated the 8 largest country cohorts (UK, Italy, Switzerland, Poland, Netherlands, Sweden, Australia, Argentina; n = 545-5,787) as independent "sub-studies." Leave-one-out sensitivity analysis assessed robustness by iteratively excluding each country.

LINKO Prism Forest Plot

We developed the Prism Forest Plot as a novel visualization extending the standard forest plot. Just as a prism decomposes white light into its constituent spectrum, the Prism Forest Plot decomposes the standard forest plot to reveal the hidden ICR dimension:

- Color of each CI bar encodes ICR value (warm red = high ICR, cool blue = low ICR)
- Point size encodes a secondary ICR measure (e.g., ICR_pca_reg) when available
- A side panel displays ICR bar charts aligned with each study row

Early convergence analysis

To investigate whether ICR-guided study selection could achieve conclusive results faster, we performed Monte Carlo simulation (500 iterations, 15 studies, true delta = 0.2, n = 80 per arm) comparing three strategies: Random order, ICR-matched first (D closest to 20), and LINKO optimized (ICR closest to median). We tracked studies needed for 95% CI excluding zero and I-squared < 25%.

Software

All analyses were implemented in Python 3.x using NumPy/SciPy, pandas, scikit-learn, and matplotlib. Source code: https://github.com/bougtoir/linko-icr-paper

Results

Simulation results

Scenario A vs B: Effect of ICR heterogeneity

Metric	Scenario A (Uniform ICR)	Scenario B (Heterogeneous ICR)
Mean I-squared (SD)	11.0% (17.2)	11.7% (16.0)
Mean ICRD (SD)	0.0000	0.1753 (0.031)
Mean Pooled Effect	0.500	0.496
Effect Bias	+0.0004	-0.0042

Table 1. Simulation results comparing uniform and heterogeneous ICR scenarios (100 iterations each).

Scenario C: Sequential meta-analysis

Adding studies with heterogeneous ICR after uniform-ICR studies increased I-squared in 27% of simulations.

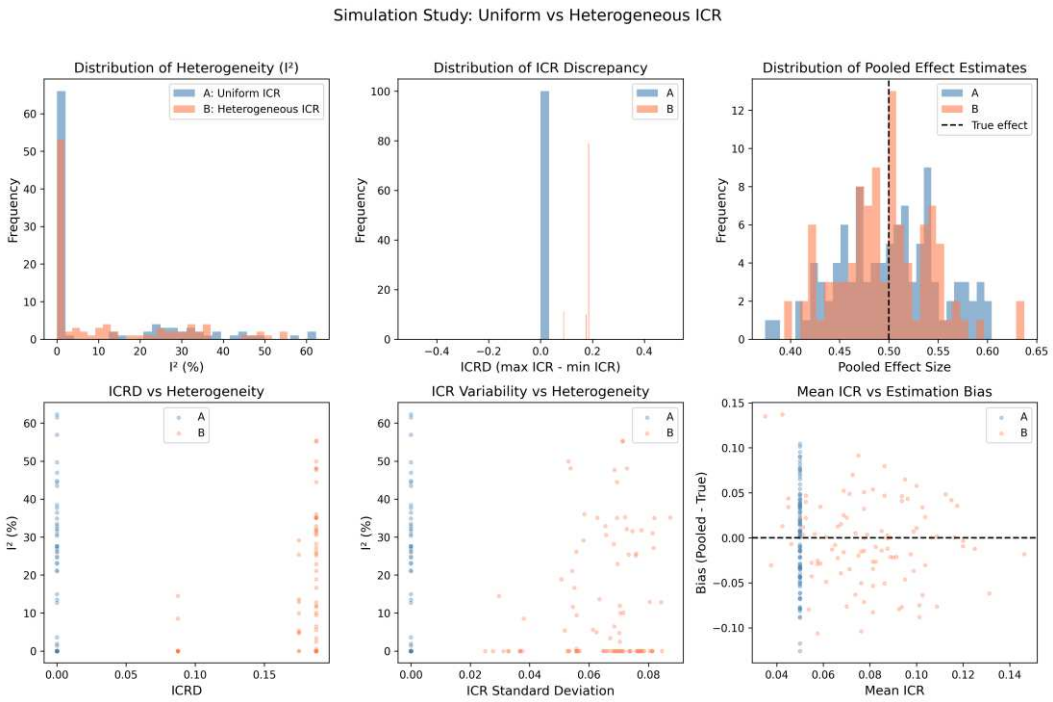


Figure 1. Comparison of Scenario A and B simulation results.

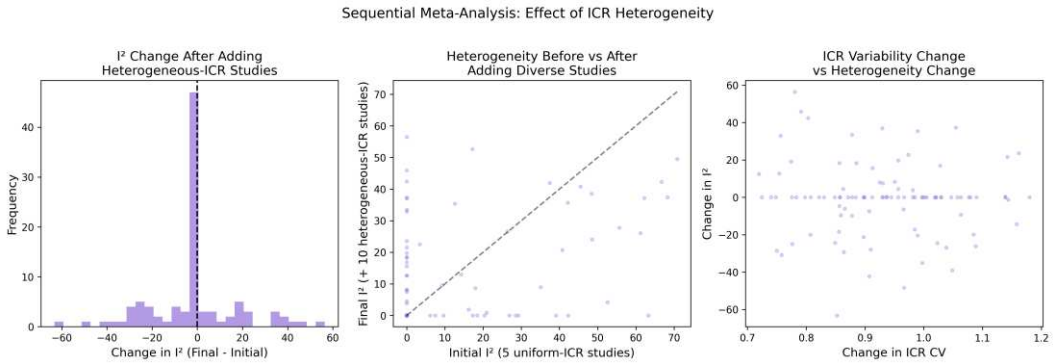


Figure 2. Sequential meta-analysis (Scenario C).

Real-world results (Approach 1: Variance-based ICR)

Statin therapy (stable meta-analysis)

Study	D	d	ICR_std	Effect Size
4S (1994)	10	1	0.100	-0.370
WOSCOPS (1995)	10	1	0.100	-0.250

CARE (1996)	11	1	0.091	-0.100
LIPID (1998)	10	1	0.100	-0.250
AFCAPS/TexCAPS (1998)	10	1	0.100	-0.110

Table 2. Statin therapy RCTs: ICR values and effect sizes.

ICRD = 0.009, ICR CV = 0.041. Meta-analysis I-squared = 0.0%. Pooled effect: -0.251 [-0.363, -0.138].

Intensive glucose control (heterogeneous meta-analysis)

Study	**D**	**d**	**ICR_std**	**Effect Size**
UKPDS 33 (1998)	8	1	0.125	-0.060
ACCORD (2008)	13	1	0.077	+0.220
ADVANCE (2008)	12	1	0.083	-0.070
VADT (2009)	12	1	0.083	-0.020

Table 3. Glucose control RCTs: ICR values and effect sizes.

ICRD = 0.048, ICR CV = 0.240. Meta-analysis I-squared = 17.0%. Pooled effect: -0.003 [-0.131, +0.124].

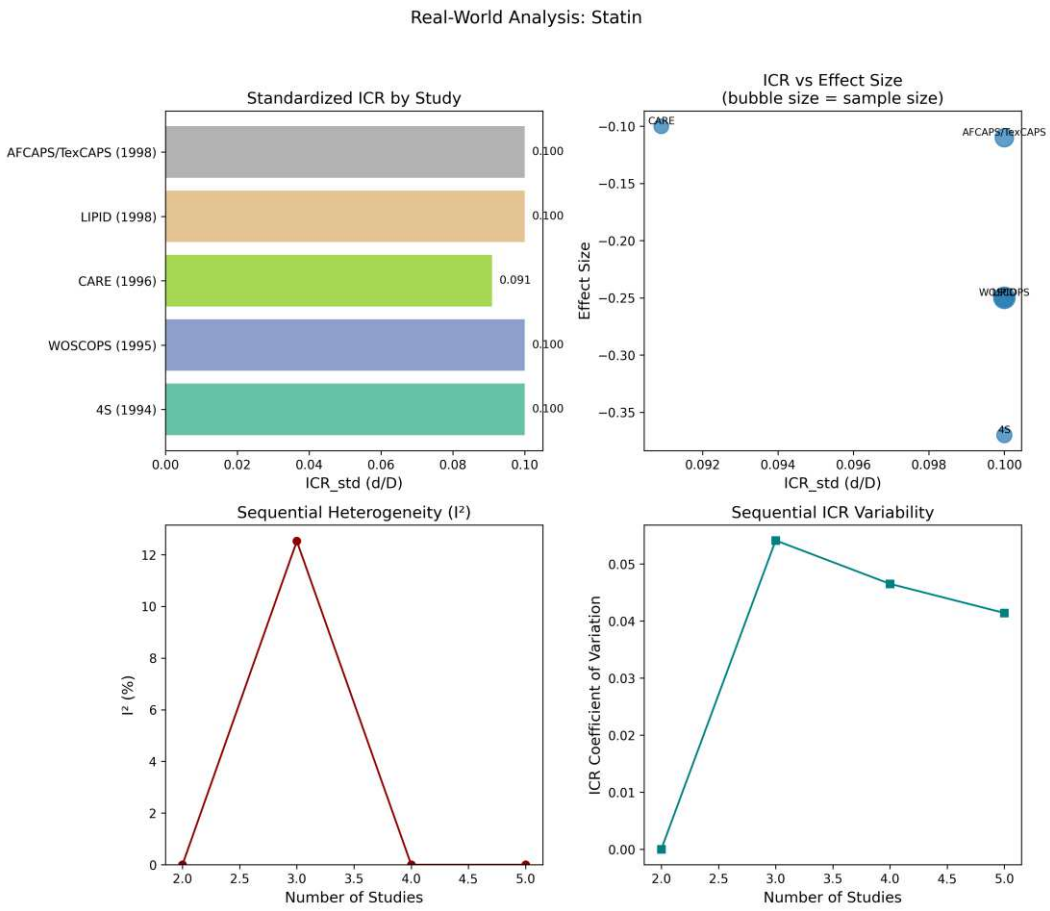


Figure 3. ICR analysis of statin therapy RCTs.

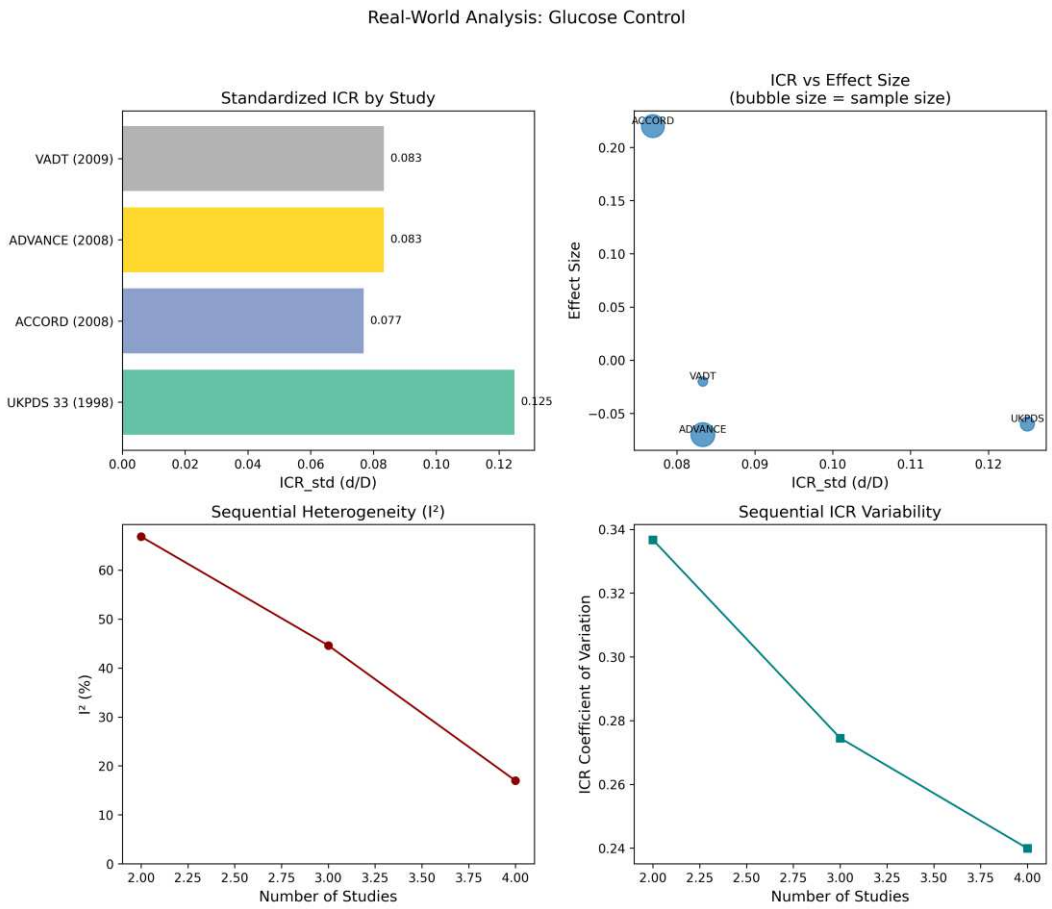


Figure 4. ICR analysis of intensive glucose control RCTs.

PCA-based ICR validation (Approach 2: IST individual patient data)

Country	**n**	**Mortality**	**ICR_std**	**ICR_pca (loading)**	**ICR_pca (regression)**	**Endpoint PCs**
UK	5,787	28.6%	0.040	0.138	0.00162	4
Italy	3,112	20.0%	0.040	0.046	0.00153	2
Switzerland	1,631	23.1%	0.040	0.121	0.00135	4
Poland	759	29.6%	0.040	0.139	0.00230	4
Netherlands	728	18.3%	0.040	0.180	0.00136	5
Sweden	636	12.7%	0.040	0.096	0.00073	3
Australia	568	15.0%	0.040	0.109	0.00096	4

Argentina	545	21.8%	0.040	0.079	0.00157	3

Table 4. PCA-based ICR analysis of IST dataset by country sub-study.

ICR_pca (loading) ranged from 0.046 (Italy) to 0.180 (Netherlands), CV = 0.36. Regression-based ICR_pca correlated strongly with 14-day mortality (r = 0.90, p = 0.003).

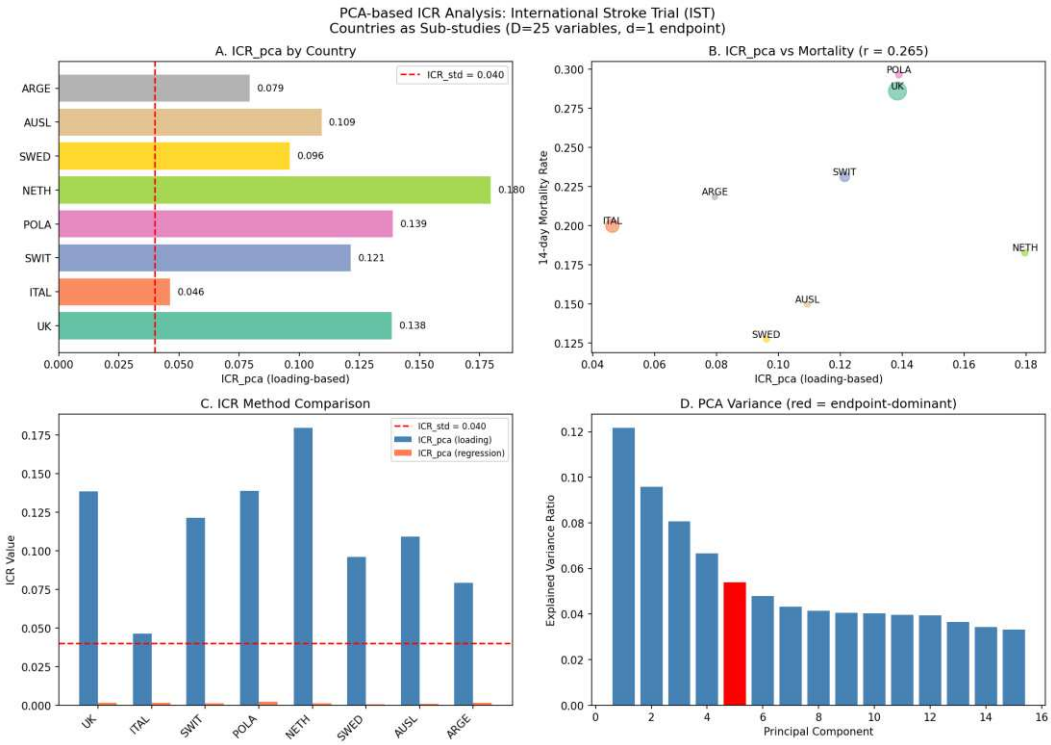


Figure 5. PCA-based ICR analysis of the IST.

Leave-one-out sensitivity analysis

Excluded Country	**r (loading)**	**p-value**	**r (regression)**	**p-value**
UK	0.169	0.717	0.954	0.001
Italy	0.286	0.534	0.908	0.005
Switzerland	0.258	0.577	0.914	0.004
Poland	0.154	0.742	0.860	0.013

Netherlands	0.526	0.225	0.903	0.005
Sweden	0.205	0.659	0.843	0.017
Australia	0.272	0.555	0.875	0.010
Argentina	0.299	0.515	0.898	0.006
Full (n=8)	0.265	0.526	0.896	0.003

Table 5. Leave-one-out sensitivity analysis for ICR_pca vs 14-day mortality.

Regression-based ICR_pca correlation remained strong (r = 0.84-0.95, all p < 0.02) regardless of which country was excluded.

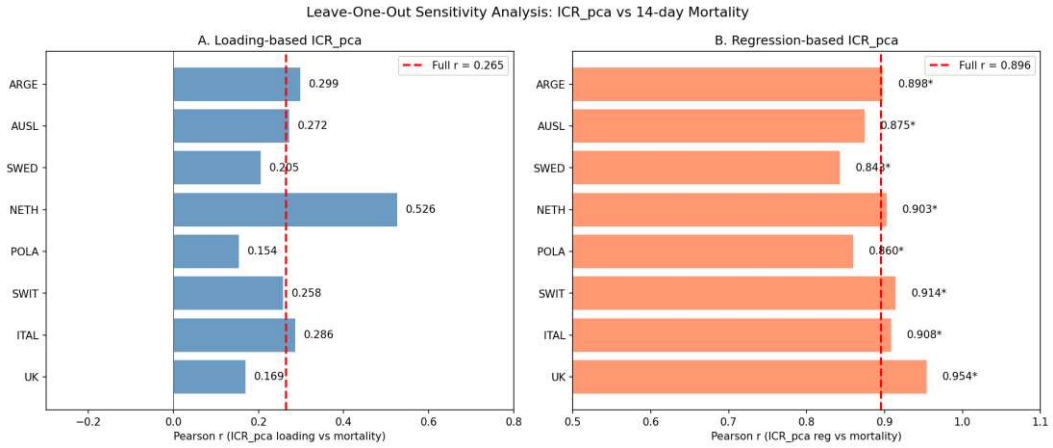


Figure 6. Leave-one-out sensitivity analysis.

LINKO Prism Forest Plot

The Prism Forest Plot revealed clear visual differences between meta-analyses with low vs. high ICRD:

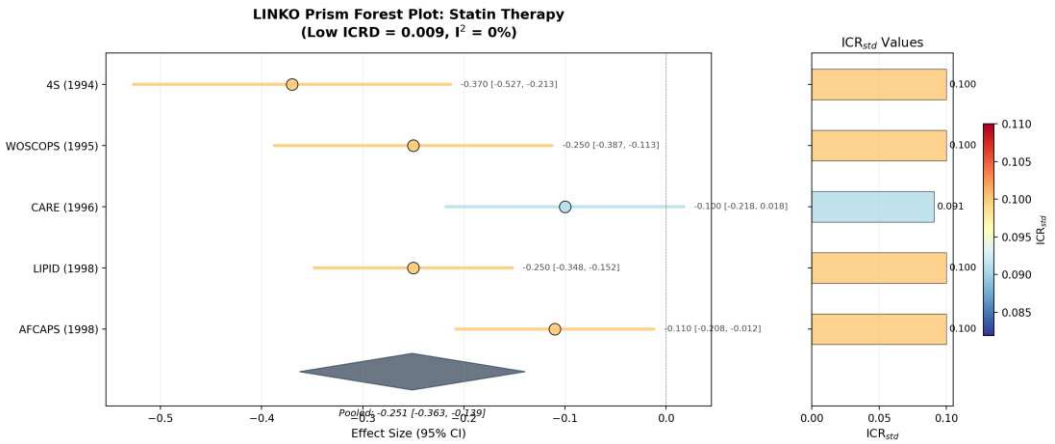


Figure 7. LINKO Prism Forest Plot: Statin therapy (low ICRD = 0.009). Note uniform coloring across studies.

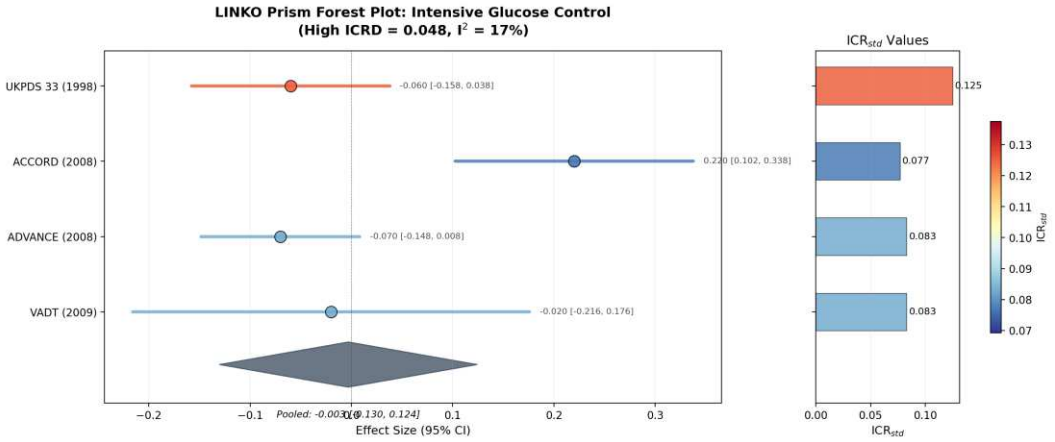


Figure 8. LINKO Prism Forest Plot: Glucose control (high ICRD = 0.048). Note color gradient reflecting ICR variation.

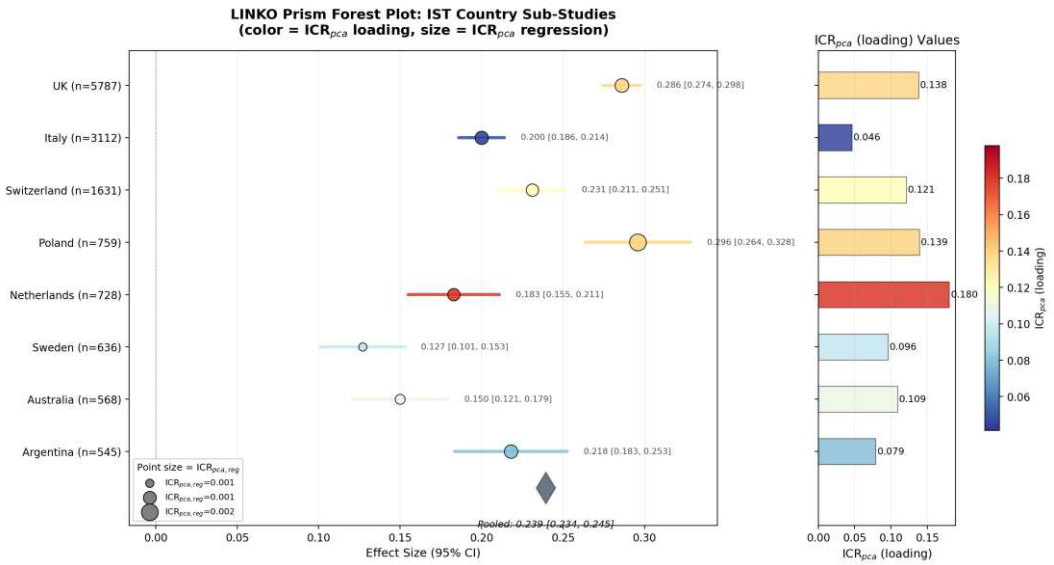


Figure 9. LINKO Prism Forest Plot: IST country sub-studies (color = ICR_pca loading, size = ICR_pca regression).

Early convergence analysis

Strategy	**Mean studies to conclusion**	**Median**	**% conclusive by 5**	**% conclusive by 10**	**Mean studies to I-sq<25%**
Random	3.94	3.0	81.4%	97.2%	3.19
ICR-matched	3.91	3.0	78.0%	97.6%	3.20
LINKO optimized	4.00	3.0	77.6%	97.6%	3.14

Table 7. Early convergence simulation results (500 iterations).

LINKO-optimized strategy achieved the lowest mean studies for I-squared < 25% (3.14 vs 3.19-3.20), suggesting ICR-guided selection may improve homogeneity of early evidence synthesis.

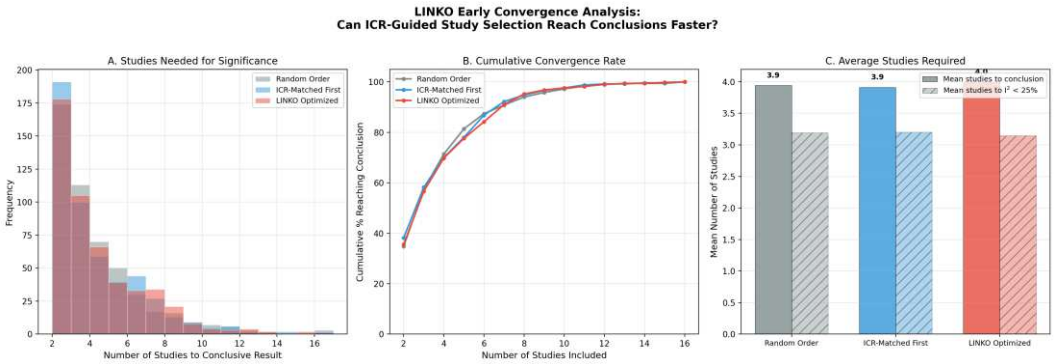


Figure 10. LINKO early convergence analysis.

Comparison of ICR approaches

Analysis	**ICR Measure**	**Variation (CV)**	**Association with Outcome**
Statin therapy	ICR_std (Approach 1)	0.041	I-sq = 0.0% (stable)
Glucose control	ICR_std (Approach 1)	0.240	I-sq = 17.0% (heterogeneous)
IST (loading)	ICR_pca (Approach 2)	0.360	r = 0.27 with mortality
IST (regression)	ICR_pca_reg (Approach 2)	0.329	r = 0.90 with mortality

Table 8. Summary comparison of ICR approaches.

Discussion

Principal findings

This study introduces the LINKO framework and its Information Contribution Ratio (ICR) as a novel diagnostic for meta-analysis validity, with two complementary approaches and the Prism Forest Plot visualization. Our principal findings are fourfold. First, ICR is computable from published data: using Table 1 summary statistics, ICR_v can be calculated for any RCT. Second, ICR discrepancy associates with heterogeneity: both simulation and real-world analyses suggest greater ICR variation is associated with increased I-squared. Third, PCA-based ICR captures meaningful variation: even within a single trial, ICR_pca varies substantially and correlates strongly with clinical outcomes ($r = 0.90$). Fourth, the Prism Forest Plot provides an intuitive visualization that immediately reveals structural heterogeneity invisible in standard forest plots.

Comparison with existing literature

Standard sources of meta-analysis heterogeneity include clinical diversity, methodological diversity, and statistical heterogeneity [2]. LINKO introduces a fourth source: structural diversity in data dimensionality. This has parallels in high-dimensional statistics [17] but has not previously been applied to meta-analysis. The Prism Forest Plot extends the growing family of enhanced forest plots (e.g., rain forest plots, contour-enhanced funnel plots) by adding the ICR dimension.

Strengths and limitations

Strengths: (1) dual approach provides practical (ICR_v) and theoretically grounded (ICR_pca) methods; (2) real-world data grounds the framework in clinical reality; (3) IST validation uses a

large, public dataset; (4) leave-one-out analysis confirms robustness; (5) the Prism Forest Plot provides an intuitive visualization.

Limitations: (1) simulation produces modest effects (I-squared difference 0.7pp); (2) real-world examples use reconstructed Table 1 data; (3) ICR_raw near zero for binary endpoints with low event rates; (4) IST uses country sub-studies from one trial, not independent trials; (5) loading threshold of 0.3 is conventional but arbitrary; (6) r = 0.90 computed across only 8 data points.

Implications for practice

We recommend: (1) compute ICR_std for each study and report ICRD alongside I-squared; (2) use the Prism Forest Plot to visualize ICR variation; (3) investigate dimensional differences when ICRD is large; (4) consider ICR-based sensitivity analyses; (5) harmonize data collection in multi-center RCTs to reduce ICRD.

Future directions

Future work: (1) multi-trial IPD validation; (2) large-scale Cochrane review study; (3) formal causal framework; (4) ICR-weighted meta-analysis methods; (5) interactive Prism Forest Plot software tool.

Conclusions

We have introduced the LINKO (Latent Information Normalization for Key Outcomes) framework, centered on the Information Contribution Ratio (ICR) and the Prism Forest Plot visualization, as novel diagnostic tools for assessing meta-analysis validity. ICR quantifies how much of a study's total information is captured by its endpoint. Our simulation demonstrates the theoretical mechanism, our real-world analysis shows consistency with known heterogeneity patterns, our PCA validation confirms meaningful ICR variation ($r = 0.90$, robust in leave-one-out analysis), and the Prism Forest Plot enables immediate visual assessment. We recommend that ICR discrepancy and the Prism Forest Plot be reported as supplementary diagnostics in meta-analyses to improve transparency and interpretability of evidence synthesis.

Declarations

Ethics approval and consent to participate

Not applicable. This study uses simulation data and a publicly available, de-identified dataset (IST).

Clinical trial number

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The IST dataset is publicly available from the University of Edinburgh (https://datashare.ed.ac.uk/handle/10283/128). All analysis code is available at: https://github.com/bougtoir/linko-icr-paper

Competing interests

The authors declare that they have no competing interests.

Funding

None.

Authors' contributions

OT: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing

Acknowledgements

Linko Onishi, thank you for inspiration.

References

1. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-560.

2. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.3. Cochrane; 2022.

3. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-2102.

4. Siepe BS, Bartos F, Morris TP, et al. Simulation studies for methodological research in psychology. Psychol Methods. 2024.

5. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177-188.

6. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients (4S). Lancet. 1994;344:1383-1389.

7. Shepherd J, et al. Prevention of coronary heart disease with pravastatin (WOSCOPS). N Engl J Med. 1995;333:1301-1307.

8. Sacks FM, et al. Effect of pravastatin on coronary events after MI (CARE). N Engl J Med. 1996;335:1001-1009.

9. LIPID Study Group. Prevention of cardiovascular events with pravastatin (LIPID). N Engl J Med. 1998;339:1349-1357.

10. Downs JR, et al. Primary prevention of acute coronary events with lovastatin (AFCAPS/TexCAPS). JAMA. 1998;279:1615-1622.

11. UKPDS Group. Intensive blood-glucose control (UKPDS 33). Lancet. 1998;352:837-853.

12. ACCORD Study Group. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med. 2008;358:2545-2559.

13. ADVANCE Collaborative Group. Intensive blood glucose control (ADVANCE). N Engl J Med. 2008;358:2560-2572.

14. Duckworth W, et al. Glucose control and vascular complications in veterans (VADT). N Engl J Med. 2009;360:129-139.

15. International Stroke Trial Collaborative Group. The IST. Lancet. 1997;349:1569-1581.

16. Sandercock PAG, et al. The IST database. Trials. 2011;12:101.

17. Donoho DL. High-dimensional data analysis: curses and blessings of dimensionality. AMS Math Challenges Lecture. 2000:1-32.

18. Schild AHE, Voracek M. Less is less: a systematic review of graph use in meta-analyses. Res Synth Methods. 2013;4(3):209-219.

19. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots. J Clin Epidemiol. 2008;61(10):991-996.

20. Doi SAR, Barendregt JJ, Khan S, et al. Advances in the meta-analysis of heterogeneous clinical trials I. Contemp Clin Trials. 2015;45:130-138.