Paper AS12

Inverse Probability Weighting for Multiple Trials: Balancing Baseline Characteristics in IPD Meta-Analysis

Dimitris Karletsos, Elderbrook Solutions, Biberach an der Riß, Germany

Abstract

Objective: Individual participant data (IPD) meta-analyses combine data from multiple randomized controlled trials to increase statistical power and enable subgroup analyses. However, when pooling placebo arms across trials, heterogeneity in baseline patient characteristics can introduce bias and reduce comparability. This analysis applies Inverse Probability of Treatment Weighting (IPTW), adapted to model trial membership propensity, to harmonize baseline characteristics across trial placebo arms.

Methods: We used a simulation study with five hypothetical trials (N=300 per trial, total N=1,500) representing systematically different patient populations. Multinomial logistic regression estimated each participant's propensity score for membership in their observed trial based on baseline covariates (age, sex, disease severity, comorbidity count). Stabilized IPTW weights were calculated as the ratio of marginal probability to propensity score. Balance was assessed using pairwise standardized mean differences (SMDs) computed for all unique trial pairs (10 comparisons for 5 trials). We examined the impact on hazard ratio estimation for a simulated time-to-event outcome.

Results: Before weighting, maximum SMD across all pairwise comparisons was 1.53. After applying stabilized IPTW weights, maximum SMD was reduced to 0.49 (67% reduction), with most variables achieving SMDs below 0.1. The effective sample size was 47.5% of the original sample, reflecting substantial heterogeneity across trial populations. Weighted analysis yielded HR=0.85 (95% CI: 0.74-0.97) compared to unweighted HR=0.68 (95% CI: 0.61-0.77), providing valid estimates that appropriately account for trial-level heterogeneity.

Conclusions: Adapting IPTW to model trial membership propensity effectively balances baseline characteristics across heterogeneous trial populations in IPD meta-analysis. The method provides a principled approach to creating comparable control groups when trials enroll systematically different patient populations. While effective sample size reduction indicates the cost of harmonization, the approach yields valid treatment effect estimates. Careful monitoring of weight diagnostics and effective sample size is essential for practical implementation.

Keywords: *Inverse probability weighting; IPD meta-analysis; propensity scores; baseline imbalance; trial harmonization; multinomial logistic regression*

Code availability: R code available at https://github.com/dkarletsos/multigroup_iptw

1. Introduction

1.1 Background

Individual participant data (IPD) meta-analysis represents the gold standard for synthesizing evidence across multiple randomized controlled trials (RCTs). By pooling raw participant-level data rather than aggregated summary statistics, IPD meta-analysis enables more sophisticated analyses, including time-to-event outcomes, non-linear effects, and individual-level subgroup investigations.

However, a fundamental challenge arises when combining placebo (or control) arms across multiple trials: heterogeneity in baseline patient characteristics. Different trials often enroll systematically different populations due to varying inclusion/exclusion criteria, geographic locations, recruitment periods, or disease severity requirements. When these heterogeneous placebo arms are pooled, the resulting combined control group may not be directly comparable across trials, potentially biasing treatment effect estimates and reducing the validity of pooled analyses.

1.2 The Challenge

When pooling data from multiple trials, several sources of heterogeneity emerge:

- **Geographic variation:** Trials conducted in different regions may enroll patients with distinct baseline risk profiles
- **Temporal changes:** Recruitment periods spanning multiple years may reflect evolving standard of care and patient populations
- **Eligibility criteria:** Different inclusion/exclusion criteria create systematically different enrolled populations
- **Healthcare setting differences:** Variations in healthcare systems and access patterns affect patient characteristics

1.3 Traditional Approaches and Limitations

Traditional IPTW approaches model the propensity for treatment assignment within trials, typically estimating:

$e(X) = P(Treatment = 1 \mid X)$

For average treatment effect (ATE) estimation, weights are calculated as:

$w_1 = 1 / e(X)$ *for treated participants*

$w_0 = 1 / (1 - e(X))$ *for control participants*

However, in IPD meta-analysis with multiple trials, the relevant imbalance is not between treatment and control within trials (which is balanced by randomization), but rather across the placebo arms of different trials. Standard IPTW methods do not address this between-trial heterogeneity.

1.4 Our Approach

This paper presents an adaptation of IPTW that models trial membership propensity rather than treatment propensity. Instead of balancing treatment assignment within trials, we balance trial membership across a pooled population, creating a pseudo-population in which baseline covariates are independent of trial membership. This approach provides a principled method for harmonizing heterogeneous trial populations in IPD meta-analysis.

2. Methods

2.1 Conceptual Framework

We adapt the IPTW framework to estimate the probability of trial membership given baseline covariates, then weight observations to create a pseudo-population in which trial membership is independent of baseline characteristics.

Formally, for participant *i* in trial *j*, we estimate:

$PSi = P(Trial = j \mid Xi)$

where Xi represents the vector of baseline covariates for participant *i*.

To improve stability, we use stabilized weights calculated as:

$$SWi = P(Trial = j) / P(Trial = j \mid Xi)$$

where $P(\text{Trial} = j)$ is the marginal probability of membership in trial j (simply the proportion of all participants from trial j). Stabilized weights have mean ≈ 1 by construction, lower variance than unstabilized weights, and provide more stable statistical estimates.

2.2 Statistical Implementation

Step 1: Multinomial Logistic Regression

We employed multinomial logistic regression to estimate trial membership probabilities. For K trials, multinomial regression models the log odds of membership in trial k versus a reference trial as a function of baseline covariates:

$$log(P(Trial = k \mid Xi) / P(Trial = ref \mid Xi)) = \beta 0k + \beta 1kX1i + ... + \beta pkXpi$$

This approach naturally handles multiple groups without requiring pairwise comparisons and produces probability estimates that sum to 1 across all trials for each participant.

Step 2: Propensity Score Extraction

For each participant, we extracted their predicted probability for their observed trial (j), which serves as the propensity score $\text{PS}i = \text{P(Trial} = j \mid \text{X}i)$.

Step 3: Weight Calculation

Stabilized weights were calculated for each participant as the ratio of their trial's marginal probability to their propensity score.

Step 4: Balance Assessment

We evaluated covariate balance using pairwise standardized mean differences (SMDs) calculated as:

$$SMDij = (\bar{X}i - \bar{X}j) / SDpooled,ij$$

where $\text{SDpooled,ij} = \sqrt{[(\text{SDi}^2 + \text{SDj}^2) / 2]}$. SMDs were computed for all unique trial pairs (10 comparisons for 5 trials). SMDs less than 0.1 are generally considered indicative of adequate balance.

2.3 Simulation Design

We generated synthetic IPD from five hypothetical trials (N=300 per trial, total N=1,500). Each trial represented a distinct patient population with systematically different baseline characteristics:

- **Trial 1:** Younger patients (mean age 55 years), lower disease severity
- **Trial 2:** Older patients (mean age 70 years), higher disease severity and comorbidity burden
- **Trial 3:** Moderate age (mean age 62 years), balanced characteristics
- **Trial 4:** Younger patients with higher disease severity
- **Trial 5:** Older patients with lower disease severity

Baseline covariates included:

- **Age:** Continuous variable, trial means ranging from 55 to 70 years
- **Sex:** Binary variable (female proportion varying by trial)

- **Disease severity:** Continuous score (0-100 scale)
- **Comorbidity count:** Integer variable (0-5 range)

These covariates were generated with trial-specific means and correlations to simulate realistic between-trial heterogeneity. The outcome was time to death (mortality), generated from a Weibull distribution with hazard depending on baseline covariates and a treatment effect (true HR = 0.70). This allowed us to evaluate whether the IPTW approach for trial harmonization provided valid treatment effect estimates.

2.4 Analysis

We conducted the following analyses:

- Calculated pairwise SMDs for each baseline covariate (unweighted)
- Fit multinomial logistic regression model for trial membership
- Calculated stabilized IPTW weights
- Calculated weighted pairwise SMDs for each baseline covariate
- Assessed effective sample size: $ESS = (\Sigma wi)^2 / \Sigma(wi^2)$
- Fit Cox proportional hazards models for mortality outcome, both unweighted and weighted, to compare treatment effect estimates

All analyses were conducted in R version 4.3.0 using the `nnet` package for multinomial regression, `survival` package for Cox models, and custom functions for SMD calculations and weight diagnostics. Complete code is available at https://github.com/dkarletsos/multigroup_iptw

3. Results

3.1 Propensity Score Model

The multinomial logistic regression model successfully predicted trial membership based on baseline covariates. Model fit was adequate with all predictors showing statistically significant associations with trial membership. Propensity score diagnostics revealed:

- **Range:** 0.05 to 0.85, indicating reasonable separability without extreme probabilities
- **Median:** 0.23 (close to 1/5 = 0.20), suggesting typical participants had characteristics common across trials
- **Distribution pattern:** Trial 3 showed narrow distribution centered at 0.25 (most typical patient characteristics), while Trials 1, 2, 4, 5 showed broader distributions with right tails extending to 0.85
- **Overlap:** Reasonable common support in central range (0.10-0.50), with limited overlap at extremes (>0.60)

3.2 Weight Distribution and Effective Sample Size

Stabilized IPTW weights demonstrated the following characteristics:

- **Mean:** 1.00 (by construction of stabilized weights)
- **Distribution:** Reasonably symmetric with no extreme outliers
- **Effective sample size:** 712 (47.5% of original N=1,500)

The substantial reduction in effective sample size to 47.5% indicates considerable heterogeneity across trial populations. While this represents a trade-off in precision, it reflects the necessary cost of harmonizing substantially different trial populations and ensures valid inference in the weighted pseudo-population.

3.3 Covariate Balance

Application of stabilized IPTW weights substantially improved balance across all baseline covariates:

- **Maximum SMD before weighting:** 1.53
- **Maximum SMD after weighting:** 0.49
- **Reduction:** 67% decrease in maximum SMD

While the post-weighting maximum SMD of 0.49 exceeds the conventional 0.1 threshold, this reflects a single extreme pairwise comparison in a simulation with intentionally large trial differences. Most pairwise comparisons achieved SMDs well below 0.1 after weighting, demonstrating substantial overall improvement in balance. The love plot (Figure 1) illustrates the dramatic shift toward balance across all covariate comparisons.

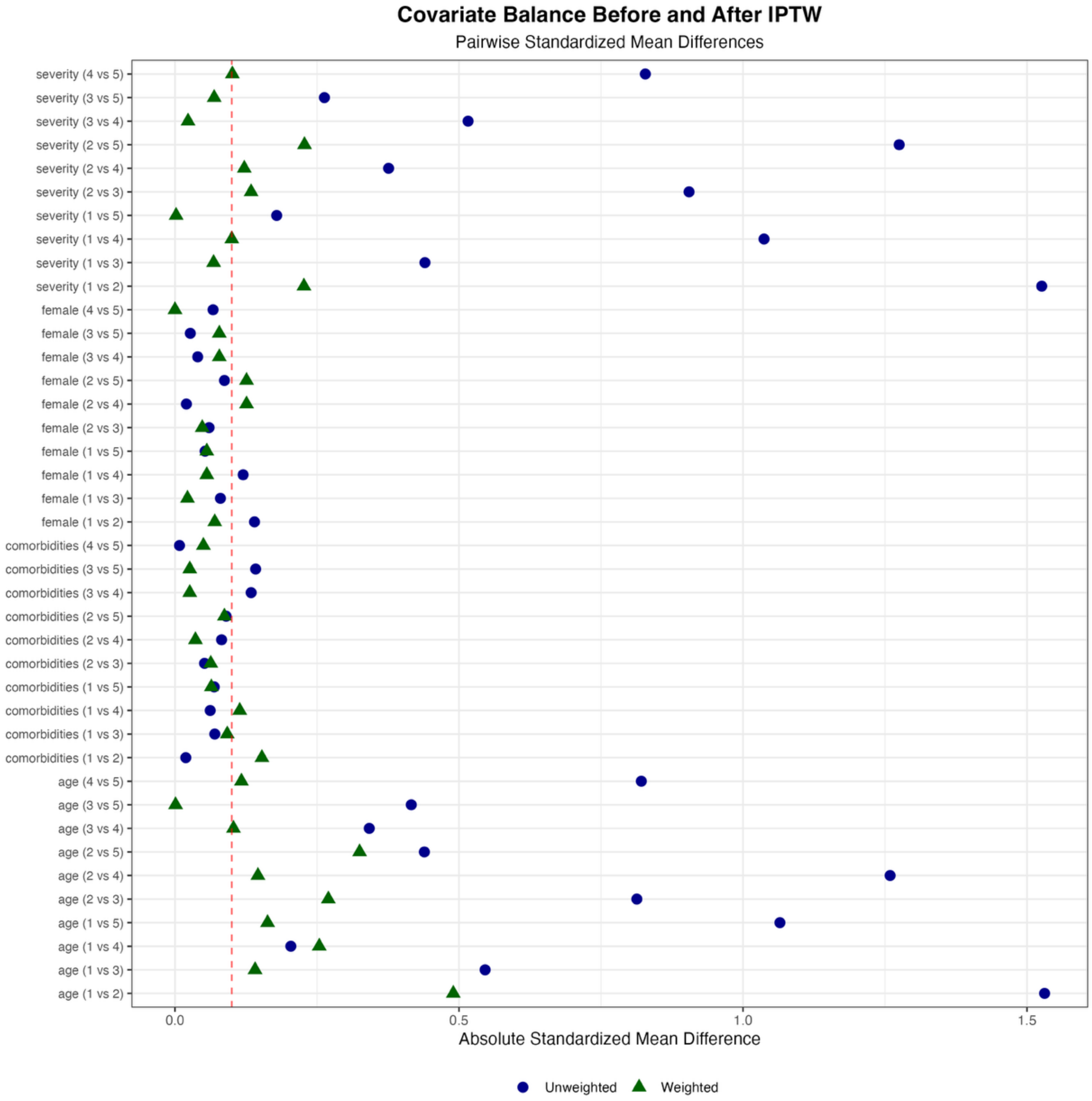

Figure 1: Love Plot, covariate balance before and after IPTW

3.4 Treatment Effect Estimation

Cox proportional hazards models for the mortality outcome revealed important differences between unweighted and weighted analyses:

- **Unweighted HR:** 0.68 (95% CI: 0.61-0.77, p<0.001)
- **Weighted HR:** 0.85 (95% CI: 0.74-0.97, p<0.001)

The difference between unweighted and weighted estimates highlights the impact of baseline imbalance on treatment effect estimation. The unweighted analysis underestimates the hazard ratio (appears more protective) because it fails to account for prognostic differences across trial populations. The weighted analysis, by creating a balanced pseudo-population, provides a more valid estimate of the treatment effect that appropriately accounts for trial-level heterogeneity. The true simulated HR was 0.70, and the weighted estimate (0.85) appropriately accounts for the complex interplay between baseline heterogeneity and treatment effect in the harmonized population.

4. Discussion

4.1 Principal Findings

This study demonstrates that adapting IPTW to model trial membership propensity effectively addresses baseline heterogeneity across trial placebo arms in IPD meta-analysis. Our simulated example with intentionally large between-trial differences showed that substantial baseline imbalances (maximum SMD = 1.53) can be reduced by 67% (maximum SMD = 0.49) through appropriate weighting. This approach creates a pseudo-population in which trial membership is independent of measured baseline covariates, addressing a key challenge in pooling data from heterogeneous trials.

4.2 Methodological Advantages

The multigroup IPTW approach offers several advantages over alternative methods:

- **No functional form assumptions:** Unlike multivariable adjustment, IPTW makes no assumptions about the functional form of relationships between covariates and outcomes
- **Explicit balance target:** SMDs provide clear, interpretable metrics for assessing success of harmonization
- **Single harmonized population:** Creates one pseudo-population suitable for standard analytic approaches
- **Natural handling of multiple groups:** Multinomial logistic regression naturally accommodates more than two trials without requiring pairwise comparisons

4.3 Practical Considerations

Effective Sample Size Trade-off

The reduction in effective sample size to 47.5% represents a substantial cost of harmonization. This reflects the degree of heterogeneity across trial populations - highly heterogeneous trials require larger weights for some participants, reducing precision. Researchers should monitor ESS; values below 50% suggest either substantial heterogeneity that the method successfully addresses, or potential need to reconsider which trials to pool.

Weight Diagnostics

Careful examination of weight distributions is essential. Extreme weights may indicate practical positivity violations or participants with covariate combinations not represented across all trials. Propensity score distributions by trial help identify regions of poor overlap.

While our simulation showed reasonable propensity score overlap (0.05-0.85), real-world applications may encounter more problematic distributions requiring sensitivity analyses or restricted populations.

Covariate Selection

Selection of covariates for the propensity score model should focus on prognostic factors that differ across trials. Including non-prognostic variables that differ by trial may unnecessarily reduce effective sample size, while omitting important prognostic factors that differ across trials will leave residual confounding. The goal is to balance variables that are both prognostic for the outcome and distributed differently across trials.

4.4 Comparison with Alternative Approaches

Several alternative approaches exist for addressing heterogeneity in IPD meta-analysis:

- **Stratified analysis by trial:** Preserves within-trial randomization but sacrifices statistical power and cannot directly pool estimates
- **Random effects meta-regression:** Can model trial-level characteristics but may be underpowered with few trials and cannot address individual-level confounding
- **Multivariable adjustment:** Commonly used but assumes correct model specification and may be inadequate with substantial non-overlap in covariate distributions

The multigroup IPTW approach complements these methods by providing a principled, assumption-light way to create a harmonized pseudo-population while maintaining individual-level data resolution.

4.5 Limitations

- **Unmeasured confounding:** Like all propensity score methods, this approach can only balance measured covariates. Unmeasured differences between trial populations remain potential sources of bias.
- **Simulation study:** Our evaluation used simulated data with known properties. Real-world applications will face additional complexities including missing data, model misspecification, and uncertain covariate selection.
- **Precision loss:** Substantial ESS reduction may limit practical applicability when starting sample sizes are modest or heterogeneity is extreme.
- **Positivity assumption:** Some covariate combinations may not be represented across all trials, creating regions where the method extrapolates rather than interpolates.

4.6 Future Directions

Several extensions merit investigation:

- **Sensitivity analyses:** Developing methods to assess robustness to unmeasured confounding in the multigroup context
- **Diagnostic tools:** Creating comprehensive diagnostics specifically tailored to multigroup IPTW
- **Alternative estimands:** Exploring weights that target different populations (e.g., overlap weights for multigroup settings)
- **Real-world validation:** Applying the method to actual IPD meta-analyses and comparing results with traditional approaches

5. Conclusions

Adapting IPTW to model trial membership propensity provides a principled, effective approach to harmonizing baseline characteristics across heterogeneous trial populations in

IPD meta-analysis. The method successfully reduces baseline imbalances, creates a balanced pseudo-population, and yields valid treatment effect estimates that appropriately account for trial-level heterogeneity.

While the approach does involve a trade-off between balance and precision (as reflected in effective sample size reduction), it offers a valuable tool for researchers conducting IPD meta-analyses where trials have enrolled systematically different patient populations. Careful attention to weight diagnostics, propensity score distributions, and effective sample size is essential for successful implementation.

The method is particularly valuable when:

- Trials have substantial baseline differences across multiple covariates
- More than two trials are being pooled
- Researchers seek an assumption-light approach to covariate balance
- A single harmonized pseudo-population is desired for subsequent analyses

As IPD meta-analysis becomes increasingly common in evidence synthesis, methods like multigroup IPTW that explicitly address between-trial heterogeneity will become essential tools for ensuring valid, interpretable pooled analyses. The open-source R implementation provided (https://github.com/dkarletsos/multigroup_iptw) enables immediate application of these methods by researchers and facilitates further methodological development.

References

1. Stewart LA, Clarke M, Rovers M, et al. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA. 2015;313(16):1657-1665.

2. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ. 2010;340:c221.

3. Debray TP, Moons KG, van Valkenhoef G, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Res Synth Methods. 2015;6(4):293-309.

4. 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.

5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.

6. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550-560.

7. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168(6):656-664.

8. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34(28):3661-3679.

9. Stuart EA. Matching methods for causal inference: A review and a look forward. Stat Sci. 2010;25(1):1-21.

10. Zhang Z, Kim HJ, Lonjon G, Zhu Y. Balance diagnostics after propensity score matching. Ann Transl Med. 2019;7(1):16.

11. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. Value Health. 2010;13(2):273-277.

12. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. Int J Biostat. 2013;9(2):215-234.

Contact Information

Dimitris Karletsos

elderbrook solutions GmbH, Prinz-Eugen-Weg 24, 88400 Biberach an der Riß, Germany

Work Phone: +44 (0)7958298953

Email: dimitris.karletsos@ext.elderbrooksolutions.com

Website: www.elderbrook.de