Supplementary Appendix

Technical Details and Implementation Guide

# Appendix A: Detailed Statistical Methodology

## A.1 Multinomial Logistic Regression for Propensity Score Estimation

For K trials, the multinomial logistic regression model estimates the probability that participant i belongs to trial j given their baseline covariates Xi:

The model uses a reference category (typically trial 1) and models the log odds of membership in each other trial relative to the reference:

log[P(Trial = k | Xi) / P(Trial = 1 | Xi)] = β₀k + β₁kX₁i + ... + βpkXpi

for k = 2, ..., K

The probabilities are recovered through the softmax transformation:

P(Trial = k | Xi) = exp(β₀k + Σβjk·Xji) / [1 + Σ(m=2 to K) exp(β₀m + Σβjm·Xji)]

with P(Trial = 1 | Xi) = 1 / [1 + Σ(m=2 to K) exp(β₀m + Σβjm·Xji)]

These probabilities sum to 1 across all trials for each participant. Maximum likelihood estimation is used to obtain parameter estimates.

## A.2 Weight Calculation and Stabilization

Unstabilized Weight: The basic IPTW weight for participant i in trial j is:

wi = 1 / P̂(Trial = j | Xi)

where P̂ denotes the estimated propensity score from the multinomial model. This weight creates a pseudo-population where trial membership is independent of measured covariates.

Stabilized Weight: To reduce variance, stabilized weights incorporate the marginal probability:

wi,stab = P(Trial = j) / P̂(Trial = j | Xi)

where P(Trial = j) = nj / N is simply the proportion of all participants from trial j. Stabilized weights have mean approximately equal to 1 and typically have lower variance than unstabilized weights, improving the stability of subsequent analyses.

## A.3 Properties of IPTW Weights

Theoretical Properties:

1. Under the assumption of positivity (0 < P(Trial = j | Xi) < 1 for all i, j), IPTW creates balance on measured covariates

2. In large samples, weighted estimators are consistent for causal effects if all confounders are measured

3. Stabilized weights preserve the total sample size: Σwi,stab = N

4. The effective sample size quantifies information loss due to weight variability

Practical Considerations:

• Extreme weights (very large or very small) indicate limited overlap in covariate distributions

• Weight truncation at specified percentiles can improve stability at the cost of some bias

• Effective sample size (ESS) < 50% of original sample may indicate problematic heterogeneity

# Appendix B: Diagnostic Measures

## B.1 Standardized Mean Difference (SMD)

For continuous variables, the SMD comparing trial j to the overall pooled sample is:

SMDj = (X̄j - X̄pooled) / SDpooled

For binary variables, the pooled standard deviation is:

SDpooled = √[p̄(1 - p̄)]

where p̄ is the overall proportion. After weighting, weighted means and standard deviations are used:

X̄weighted = Σ(wi · Xi) / Σwi

Interpretation: |SMD| < 0.1 is often used as a threshold for negligible imbalance, though this is a rule of thumb rather than a strict criterion. Cohen's d interpretations suggest |SMD| < 0.2 as small effect, 0.2-0.5 as medium, and > 0.5 as large.

## B.2 Effective Sample Size (ESS)

The effective sample size quantifies the information content of the weighted sample:

ESS = (Σwi)² / Σ(wi²)

Interpretation:

• ESS = N if all weights equal 1 (no weighting)

• ESS < N indicates some loss of precision due to weight variability

• ESS/N gives the proportion of original information retained

• ESS/N < 0.5 suggests substantial variability in weights and potential need for weight truncation or alternative methods

## B.3 Propensity Score Overlap

Visual inspection of propensity score distributions across trials helps assess covariate overlap:

• Strong overlap: Distributions substantially overlap with no extreme non-overlap regions

• Moderate overlap: Some regions of non-overlap but most participants have reasonable propensity scores for multiple trials

• Poor overlap: Clear separation between distributions, suggesting trials enrolled fundamentally different populations

Propensity scores very close to 0 or 1 indicate near-deterministic trial membership given covariates, which can lead to extreme weights and instability.

# Appendix C: Implementation in Statistical Software

## C.1 R Implementation

Key R code snippet for fitting the propensity score model:

library(nnet)

# Fit multinomial logistic regression

ps\_model <- multinom(trial\_id ~ age + female + severity + comorbidities,

data = ipd\_data, trace = FALSE)

# Predict propensity scores

ps\_matrix <- predict(ps\_model, type = 'probs')

# Extract PS for observed trial

ipd\_data$ps <- sapply(1:nrow(ipd\_data), function(i) {

trial <- as.numeric(ipd\_data$trial\_id[i])

ps\_matrix[i, trial]

})

# Calculate stabilized weights

marginal\_prob <- table(ipd\_data$trial\_id) / nrow(ipd\_data)

ipd\_data$sw <- marginal\_prob[ipd\_data$trial\_id] / ipd\_data$ps

## C.2 Python Implementation

Key Python code snippet using scikit-learn:

from sklearn.linear\_model import LogisticRegression

from sklearn.preprocessing import StandardScaler

# Prepare and standardize features

X = ipd\_data[['age', 'female', 'severity', 'comorbidities']].values

y = ipd\_data['trial\_id'].values

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# Fit multinomial logistic regression

ps\_model = LogisticRegression(multi\_class='multinomial', solver='lbfgs')

ps\_model.fit(X\_scaled, y)

# Predict propensity scores and extract for observed trial

ps\_matrix = ps\_model.predict\_proba(X\_scaled)

ipd\_data['ps'] = [ps\_matrix[i, y[i]-1] for i in range(len(y))]

# Calculate stabilized weights

marginal\_probs = ipd\_data['trial\_id'].value\_counts(normalize=True)

ipd\_data['marginal\_prob'] = ipd\_data['trial\_id'].map(marginal\_probs)

ipd\_data['sw'] = ipd\_data['marginal\_prob'] / ipd\_data['ps']

# Appendix D: Simulation Parameters

The simulation study used the following parameters to generate heterogeneous trial populations:

Sample Size:

• N = 1,500 total participants

• 300 participants per trial across 5 trials

Baseline Covariate Distributions by Trial:

Trial 1 (Young, low severity): Age ~ N(55, 10²), Severity ~ N(45, 15²), P(Female) = 0.45

Trial 2 (Old, high severity): Age ~ N(70, 8²), Severity ~ N(68, 12²), P(Female) = 0.55

Trial 3 (Moderate): Age ~ N(62, 12²), Severity ~ N(55, 18²), P(Female) = 0.50

Trial 4 (Young, high severity): Age ~ N(58, 11²), Severity ~ N(62, 14²), P(Female) = 0.48

Trial 5 (Old, low severity): Age ~ N(67, 9²), Severity ~ N(48, 16²), P(Female) = 0.52

Outcome Model:

• Survival times generated from Weibull distribution with shape = 1.2

• Hazard function depends on baseline covariates:

log(hazard) = -3.5 + 0.03×(age-60) + 0.02×(severity-50) + 0.15×comorbidities - 0.2×female

• Administrative censoring at 5 years

These parameters were chosen to create realistic heterogeneity while maintaining some overlap to allow propensity score estimation. The varying combinations of age and severity across trials simulate common patterns in real multi-center trials.