# Novel Statistical Methods for Data Harmonization

**Edwin van den Heuvel**
**Professor in Statistics**

**March 2018**

Technische Universiteit
**Eindhoven**
University of Technology

**Where innovation starts**

# Contents

- IPD Meta-analysis
- Harmonization approaches
- Standardization methods
- Simulation study
  - Data generation
  - Analysis method
  - Results
  - Conclusions

# IPD Meta-Analysis

- The gold standard in meta-analysis is “individual participant data (IPD)” meta-analysis
- Advantages of IPD
  - Correcting for confounders at individual level
  - Testing interactions of covariates with exposure is more powerful at individual level
  - Subgroup analysis to investigate heterogeneity or consistency
- There are two analysis forms of IPD:
  - Two-stage IPD
  - One-stage IPD

# IPD Meta-Analysis

## Two-stage Analysis:

- Similar to an analysis of aggregate data (AD) meta-analysis
- Same statistical model is fitted to each study
- Each study fits the same covariates
  - Estimates can be corrected when not all covariates are available (Fibrinogen Study, 2009)
- Parameter estimates are combined via regular meta-analysis techniques
  - Weighted averages (e.g. DerSimonian-Laird)
  - Maximum likelihood estimates

# IPD Meta-Analysis

## One-stage Analysis:

- Is one statistical analysis of all data addressing random and fixed effects
  - Studies are considered random
- Requires more sophisticated statistical tools and may be numerically more challenging
  - Analysis may require several hierarchical structures of random effects (especially for longitudinal data)
  - In some cases the analysis must be executed in a ***federated*** or ***distributed*** way (data can not be pooled into one location)

# IPD Meta-Analysis

## Federated Meta-Analysis:

- Expectation-maximization algorithm
  - Step 0: Choose starting values of all parameters
  - Step 1: E-Step: use the estimates from the previous step to estimate random effects

    M-Step: Using the result from the E-step determine fixed effects parameters
  - Evaluate: how much the estimates has changed
    - If the changes are small enough → convergence
    - If the changes are still to large → conduct step 1 using the last available estimates
  - EM uses study summary statistics

# IPD Meta-Analysis

## Federated Meta-Analysis: Example

- Response: Systolic blood pressure
- Exposure: Noise
- Confounders: Age, Sex, PM10 (particulate matter)
- Two cohorts: HUNT and LifeLines

| | $\beta_0$ | $\beta_{AGE}$ | $\beta_{SEX}$ | $\beta_{PM10}$ | $\beta_{NOISE}$ | $\tau^2$ | $\sigma^2$ |
|---|---|---|---|---|---|---|---|
| EM-0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| EM-Final | 111.59 | 0.4141 | -7.255 | 0.04627 | -0.01351 | 1.8992 | 217.43 |
| EM-0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| EM-Final | 114.68 | 0.4143 | -7.2473 | -0.16617 | -0.00300 | 0 | 217.45 |

- EM is dependent on starting value and it makes a difference in estimation

# IPD Meta-Analysis

- One-stage IPD requires more effort since similar variables can not just simply be pooled
- Variables from different studies may not contain the exact same information
- Example on memory:

| Study | HUI | RAVLT | Free BRCP | Cued BRCP |
|---|---|---|---|---|
| CCHS | 5 | 15 | - | - |
| CSHA | - | 15 | 12 | 12 |
| NuAge | - | - | 16 | 16 |

  - RAVLT=Rey auditory verbal learning test
  - BRCP=Buschke cued recall procedure
  - HUI=Health utility score

# IPD Meta-Analysis

| Correct # words | CCHS (n=7107) | | CSHA (n=1730) | | | NuAge (n=432) | |
|---|---|---|---|---|---|---|---|
| | HUI | Rey | Rey | Free B | Cued B | Free B | Cued B |
| 0 | 71.1 | 1.17 | 4.57 | 4.28 | 0.23 | 0.23 | 0 |
| 1 | 2.04 | 2.81 | 7.28 | 2.31 | 0.29 | 0 | 0 |
| 2 | 20.9 | 8.68 | 13.1 | 3.70 | 0.58 | 0.69 | 0 |
| 3 | 4.94 | 15.6 | 21.6 | 5.09 | 0.40 | 3.24 | 0 |
| 4 | 0.99 | 20.3 | 17.7 | 8.61 | 0.58 | 6.48 | 0.23 |
| 5 | 0.01 | 10.2 | 13.1 | 10.9 | 0.81 | 13.4 | 0 |
| 6 | Na | 15.3 | 5.32 | 14.0 | 1.27 | 14.1 | 0.69 |
| 7 | Na | 8.33 | 2.60 | 17.4 | 2.08 | 16.0 | 0.93 |
| 8 | Na | 4.54 | 1.05 | 14.3 | 2.66 | 14.8 | 0.93 |
| 9 | Na | 1.89 | 0.06 | 10.7 | 4.22 | 12.3 | 2.31 |
| 10 | Na | 0.79 | 0.17 | 6.42 | 7.11 | 8.80 | 1.16 |
| 11 | Na | 0.27 | 0 | 2.02 | 16.5 | 4.63 | 5.56 |
| 12 | Na | 0.11 | 0 | 0.29 | 63.2 | 3.24 | 7.64 |
| 13 | Na | 0.03 | 0 | Na | Na | 1.85 | 10.6 |
| 14 | Na | 0 | 0 | Na | Na | 0.23 | 18.8 |
| 15 | Na | 0 | 0 | Na | Na | 0 | 19.4 |
| 16 | Na | Na | Na | Na | Na | 0 | 31.7 |

# Harmonization Approaches

- Bridge variables are variables that make it possible to connect studies
  - RAVLT connects CCHS with CSHA
  - BCRP connects CSHA with NuAge
  - Thus all studies are connected
- Different Harmonization methods

**With bridge variables**

- Sequential calibration
- Latent variable models
- Imputation methods

**Without bridge variables**

- Algorithmic methods
- Standardization or normalization
- Imputation methods

# Harmonization Approaches

- Algorithmic harmonization:
  - Transform $y$ into categories (low, medium, high)
  - Thresholds are test and demographic specific
- Standardization or normalization:
  - Change the scores to a common scale:
  - Min-max scaling: $(y - y_{\min})/(y_{\max} - y_{\min})$
  - Z-scores: $(y - \bar{y})/s$
  - T-scores: $(y - \bar{y})/s$ corrected for covariates
  - C-scores: $(y - \bar{y}_C)/s_C$ corrected for covariates, with $\bar{y}_C$ an average of a well-defined control group
  - Quantile normalization: $F^{-1}(y)$

# Harmonization Approaches

- Calibration on responses
  - Sequential relations are formed: $y_2 = \psi(y_1)$
  - Requires “bridge variables” to connect studies
- Latent variable models:
  - Factor analysis or item response theory models
  - The latent variable is the harmonized variable
  - Requires “bridge variables” to connect studies
- Multiple imputation methods:
  - Must deal with sporadic and systematic missingness
  - Makes all variables available in all studies
  - Requires other bridge variables

# Harmonization Approaches

## Latent variable models

- Let $Z_i$ be the ability or latent variable on subject $i$
- Let $Y_{hi}$ be the score for test $h$ on subject $i$
- We assume that $Y_{hi}$ given $Z_i$ is binomial

$$Y_{hi}|Z_i \sim Bin\big(N_h, p_h(Z_i)\big)$$

- The different tests are harmonizable when there exist a function $\psi$ such that

$$p_1(Z_i) = p_2\big(\psi(Z_i)\big) \text{ or } p_2(Z_i) = p_1\big(\psi(Z_i)\big)$$

  - Calibration model on latent variable similar to inches and centimeters: 1 inch = 2.54 cm
  - We choose $\psi(z) = a + bz$

# Harmonization Approaches

## Latent variable models

- The distribution of $Z_i$ is considered normal
  - Mean: $\beta_0 + \sum_{m=1}^{p} \beta_m x_{m,i}$
  - Variance: $\gamma_0 + \sum_{m=1}^{p} \gamma_m x_{m,i}$
- The probability $p_h$ is $logit\big(p_h(z)\big) = \mu_h + \eta_h z$

| Variable | Level | CCHS (n=7107) | CSHA (n=1730) | NuAge (n=432) | P-value |
|---|---|---|---|---|---|
| Age [mean(sd)] | Numeric | 73.2 (5.85) | 79.7 (6.96) | 73.7 (3.95) | <0.001 |
| Sex [%] | Male | 42.3% | 37.3% | 46.3% | <0.001 |
| | Female | 57.7% | 62.7% | 53.7% | |
| Education [%]% | Low | 19.0% | 48.6% | 15.3% | <0.001 |
| | Medium | 37.7% | 35.8% | 39.6% | |
| | High | 43.3% | 15.6% | 45.1% | |

# Harmonization Approaches

## Latent variable models:

| Parameters | Mean | | | Variance | | |
|---|---|---|---|---|---|---|
| | CCHS | CSHA | NuAge | CCHS | CSHA | NuAge |
| Intercept | 0.955 | 2.653 | 1.009 | -3.124 | -2.161 | -2.263 |
| HUI | 2.813 | NA | NA | Zero (due to weak invariance) | | |
| RAVLT | NA | -1.444 | NA | | | |
| Cued Buschke | NA | 2.693 | 2.291 | | | |
| Med. Education | 0.201 | 0.049 | 0.160 | -0.063 | 0.106 | 0.207 |
| High education | 0.356 | 0.176 | 0.357 | -0.010 | 0.144 | 0.184 |
| Sex | 0.210 | -0004 | 0.449 | 0.148 | 0.286 | 0.018 |
| Age | -0.029 | -0033 | -0.022 | 0.023 | 0.021 | 0.020 |

- Model fit: $\chi^2/df = 1.114$

# Harmonization Approaches

- Without bridge variables there is no way to check “content equivalence”
  - BMI and IQ can be standardized, normalized or categorized, but they do not become exchangeable
- Bridge variables could check calibration or latent variable invariance principles
  - Consistency of the relation between free and cued BCRP can be verified for CSHA and NuAge
  - Sparsity restricts the verification of all invariances
  - Harmonization invariance is unequal to factorial measurement invariance but it is similar to differential item functioning

# Standardization Methods

- Compare standardization methods
  - T-score (corrected for confounders)
  - C-score (corrected for confounders)
- Harmonization setting
  - Response: Memory construct
  - Exposure: physical activity (low, medium, high)
  - Confounders: Age, gender, and education
- Two stage IPD meta-analysis
  - Pooled effect size: Hedges $g$ (adjusted and unadjusted for confounders)
  - Measure of heterogeneity: $I^2$

# Standardization Methods

## Calculations of T- and C-score

Let $Y_i$ be memory scale for participant $i$ in one study

- T-score

$$Z_i = 10 + 3(Y_i - \bar{Y})/S$$

$$Z_i = \alpha_0 + \sum_{k=1}^{p} \alpha_k x_{ik} + e_i, \quad e_i \sim N(0, \sigma_e^2)$$

$$T_i = 50 + 10\left(Z_i - \hat{Z}_i\right)/\hat{\sigma}_e$$

  - With $\bar{Y} = \sum_{i=1}^{n} Y_i$, $S = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$, $\hat{Z}_i$ the predicted $Z_i$, and $\hat{\sigma}_e$ an estimate of $\sigma_e$
- C-score: $C_i = (Y_i - \bar{Y}_C)/S_C$
  - With $\bar{Y}_C$ and $S_C$ the average and standard deviation of 70 to 74 years old and high educated females

# Standardization Methods

## Calculations of hedges g

- Unadjusted effect size of PA on memory per study
  - Three levels: Low-Medium, Low-High, Medium-High
- Let $(w_k, \tau_k, n_k)$ be the mean, standard deviation, and sample size for T- or C-score at PA level $k$
- Effect size for PA level $k$ versus $l$:
  - Mean difference: $d_{kl} = w_k - w_l$
  - Pooled standard deviation: $s_{kl}^2 = \frac{(n_k-1)\tau_k^2+(n_l-1)\tau_l^2}{n_k+n_l-2}$
  - Hedges g: $g_{kl} = \left(1 - \frac{3}{4(n_k+n_l)-9}\right)\frac{d_{kl}}{s_{kl}}$
  - Variance: $v_{kl}^2 = \frac{n_k+n_l}{n_k n_l} + \frac{g_{kl}^2}{2(n_k+n_l)}$

# Standardization Methods

## Calculations of hedges g

- Analysis of T- or C-score corrected for age, sex, education uses linear regression per study
  - Residual variance changes with PA levels and replaces $\tau_k^2$ and $\tau_k^2$
  - Least square means of PA in regression analysis are used as substitutes for difference $d_{kl}$
  - Samples sizes $n_k$ and $n_l$ remain the same
- DerSimonian & Laird to combine hedges g's
- Heterogeneity: $I_{kl}^2 = (Q_{kl} - m + 1)/Q_{kl}$

$$Q_{kl} = \sum_{r=1}^{m} v_{kl,r}^{-2}\left(d_{kl,r} - \bar{d}_{kl}\right)^2$$

# Simulation Study

## *Data generation*

- Age, gender, and education were independently drawn from normal and Bernoulli distribution
- Physical activity was simulated from logistic distribution with mean dependent on confounders
  - Then physical activity was set to three categories
- Latent memory followed a normal distribution
  - mean depend on physical activity and confounders
  - Variance depend on confounders
- Memory score given latent variable followed a binomial distribution with logistic link function
- Date generation was done per study

# Simulation Study

## *Data generation*

- Age $A \sim N(\mu, \sigma^2)$
- Gender $S \sim B(\pi_S)$
- Education $E \sim U(0,1)$
  - Low education: $E_L = 1$: $0 \leq U \leq \pi_L$
  - Medium education: $E_M = 1$: $\pi_L < U \leq \pi_M$
  - High education: $E_H = 1$: $\pi_M < U \leq 1$

| Parameters | Homogeneous populations | | | Heterogeneous populations | | |
|---|---|---|---|---|---|---|
| | Study 1 | Study 2 | Study 3 | Study 1 | Study 2 | Study 3 |
| $\mu_A$ | 75 | 75 | 75 | 70 | 80 | 75 |
| $\sigma_A$ | 6 | 6 | 6 | 6 | 7 | 4 |
| $\pi_S$ | 0.40 | 0.40 | 0.40 | 0.40 | 0.35 | 0.45 |
| $\pi_L$ | 0.30 | 0.30 | 0.30 | 0.20 | 0.50 | 0.15 |
| $\pi_M$ | 0.70 | 0.70 | 0.70 | 0.60 | 0.85 | 0.55 |

# Simulation Study
## *Data generation*

- Physical Activity $X \sim Logistic(\mu_{PA}, 1)$

$$\mu_{PA} = -0.03A - 0.5S - 0.6E_L - 0.3E_M$$

  - The influence of confounder on physical activity was the same in each study
- Physical activity was categorized into three categories:
  - Low PA: $P_L = 1$: $X \leq 1.0$
  - Medium PA: $P_M = 1$: $1.0 < X \leq 3.5$
  - High PA: $P_H = 1$: $3.5 < X$

# Simulation Study
## *Data generation*

- Latent Memory $Z \sim N(\mu_M, \sigma_M^2)$

$$\mu_M = \beta_0 + \beta_A A + \beta_S S + \beta_L E_L + \beta_M E_M + \beta_1 P_L + \beta_2 P_M$$

$$\log \sigma_M = \gamma_0 + \gamma_A A + \gamma_S S + \gamma_L E_L + \gamma_M E_M$$

- Memory scale $Y \sim Bin\big(N, p(Z)\big)$
  - Number of words tested $N$
    - Study 1: $N = 15$
    - Study 2: $N = 12$
    - Study 3: $N = 16$
  - Logistic function for probability

$$p(Z) = exp(Z)/\big(1 + exp(Z)\big)$$

# Simulation Study

## *Data generation*

- Parameters of confounders for memory

| Parameters | Homogeneous effects | | | Heterogeneous effects | | |
|---|---|---|---|---|---|---|
| | Study 1 | Study 2 | Study 3 | Study 1 | Study 2 | Study 3 |
| $\beta_0$ | 2.5 | 2.5 | 2.5 | 1.5 | 2.5 | 3.5 |
| $\beta_A$ | -0.025 | -0.025 | -0.025 | -0.025 | -0.035 | -0.020 |
| $\beta_S$ | 0.20 | 0.20 | 0.20 | 0.20 | 0 | 0.45 |
| $\beta_L$ | -0.25 | -0.25 | -0.25 | -0.35 | -0.15 | -0.35 |
| $\beta_M$ | -0.15 | -0.15 | -0.15 | -0.15 | -0.10 | -0.20 |
| $\gamma_0$ | -2.0 | -2.0 | -2.0 | -3.0 | -2.0 | -2.0 |
| $\gamma_A$ | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| $\gamma_S$ | 0.15 | 0.15 | 0.15 | 0.15 | 0.25 | 0 |
| $\gamma_L$ | -0.15 | -0.15 | -0.15 | 0 | -0.15 | -0.20 |
| $\gamma_M$ | -0.05 | -0.05 | -0.05 | 0 | -0.05 | 0 |

- Parameters of physical activity on memory

| Parameters | Homogeneous effects | | | Heterogeneous effects | | |
|---|---|---|---|---|---|---|
| | Study 1 | Study 2 | Study 3 | Study 1 | Study 2 | Study 3 |
| $\beta_1$ | -0.50 | -0.50 | -0.50 | -0.60 | -0.30 | -0.10 |
| $\beta_2$ | -0.25 | -0.25 | -0.25 | -0.30 | 0 | -0.10 |

# Simulation Study

## *Results*

### Unadjusted effect sizes

| Effect of PA | Population | Confounder Effect on Memory | Type of Outcome | Effect size | Power | $I^2$ |
|---|---|---|---|---|---|---|
| Homo | Homo | Homo | T-score | 0.57 | 100 | 16.6 |
| Homo | Homo | Homo | C-score | 0.53 | 100 | 17.1 |
| Homo | Homo | Hetero | T-score | 0.62 | 100 | 87.6 |
| Homo | Homo | Hetero | C-score | 0.57 | 100 | 82.7 |
| Homo | Hetero | Homo | T-score | 0.58 | 100 | 27.7 |
| Homo | Hetero | Homo | C-score | 0.54 | 100 | 30.0 |
| Homo | Hetero | Hetero | T-score | 0.62 | 100 | 91.9 |
| Homo | Hetero | Hetero | C-score | 0.57 | 100 | 89.8 |
| Hetero | Homo | Homo | T-score | 0.39 | 73.2 | 95.4 |
| Hetero | Homo | Homo | C-score | 0.36 | 56.2 | 95.6 |
| Hetero | Homo | Hetero | T-score | 0.46 | 19.9 | 98.2 |
| Hetero | Homo | Hetero | C-score | 0.41 | 13.5 | 98.0 |
| Hetero | Hetero | Homo | T-score | 0.40 | 62.8 | 96.0 |
| Hetero | Hetero | Homo | C-score | 0.37 | 46.9 | 96.0 |
| Hetero | Hetero | Hetero | T-score | 0.47 | 10.2 | 98.4 |
| Hetero | Hetero | Hetero | C-score | 0.42 | 7.0 | 98.3 |

# Simulation Study

## *Results*

## Adjusted effect sizes

| Effect of PA | Population | Confounder Effect on Memory | Type of Outcome | Effect size | Power | $I^2$ |
|---|---|---|---|---|---|---|
| Homo | Homo | Homo | T-score | 0.59 | 100 | 18.4 |
| Homo | Homo | Homo | C-score | 0.59 | 100 | 18.4 |
| Homo | Homo | Hetero | T-score | 0.65 | 100 | 88.5 |
| Homo | Homo | Hetero | C-score | 0.65 | 100 | 88.5 |
| Homo | Hetero | Homo | T-score | 0.60 | 100 | 29.9 |
| Homo | Hetero | Homo | C-score | 0.60 | 100 | 29.9 |
| Homo | Hetero | Hetero | T-score | 0.64 | 100 | 92.5 |
| Homo | Hetero | Hetero | C-score | 0.64 | 100 | 92.5 |
| Hetero | Homo | Homo | T-score | 0.41 | 72.7 | 95.8 |
| Hetero | Homo | Homo | C-score | 0.41 | 72.7 | 95.8 |
| Hetero | Homo | Hetero | T-score | 0.48 | 19.4 | 99.4 |
| Hetero | Homo | Hetero | C-score | 0.48 | 19.4 | 99.4 |
| Hetero | Hetero | Homo | T-score | 0.41 | 61.5 | 96.4 |
| Hetero | Hetero | Homo | C-score | 0.41 | 61.5 | 96.4 |
| Hetero | Hetero | Hetero | T-score | 0.48 | 9.2 | 98.6 |
| Hetero | Hetero | Hetero | C-score | 0.48 | 9.2 | 98.6 |

# Simulation Study
## *Conclusions*

- Pooled effects of PA on memory for T and C are
  - Different when scores are unadjusted
  - Almost identical for adjusted effect sizes
  - A bit affected by heterogeneity in confounders
- Unadjusted effect sizes show less study heterogeneity for C-scores than for T-scores
- Heterogeneity in homogeneous effect sizes of PA on memory with T- and C-scores
  - Is disturbed by heterogeneous effect sizes of the confounders on memory (adjusted and unadjusted)
  - Is only little affected by heterogeneity in confounders across studies

# Simulation Study
## *Conclusions*

MedCrave
Step into the World of Research
Biometrics & Biostatistics International Journal

# Statistical Harmonization Methods in Individual Participants Data Meta-Analysis are Highly Needed

## Editorial

Meta-analysis has a long history within the medical sciences and epidemiology [1-3]. The main goal of a meta-analysis is to improve the precision of a specific effect size of a treatment or an exposure on a clinical or disease outcome by pooling or combining multiple studies. It is frequently conducted within a systematic review of the scientific literature to guarantee that studies with appropriate information are not ignored or overlooked and to make sure that the pooled estimate represents an unbiased and precise estimate of the true effect size. Moreover, the pooled effect estimate is evaluated in the context of study heterogeneity. In case substantial heterogeneity in the effect sizes across studies is present, the pooled estimate is considered less reliable or even questionable. Thus not only the pooled estimate must be unbiased, also the estimate of the measure

Editorial

Volume 3 Issue 3 - 2016

E R van den Heuvel[1]* and Griffith LE[2]

[1]*Department of Stochastics, Eindhoven University of Technology, Netherlands*
[2]*Department of Clinical Epidemiology & Biostatistics, McMaster University, Canada*

***Corresponding author:** E R van den Heuvel, Department of Stochastics Biostatistics, Faculty of Mathematics and Computer Science, Eindhoven University of Technology, Netherlands,
Email: e.r.v.d.heuvel@tue.nl

**Received:** January 28, 2016 | **Published:** February 01, 2016