Conventional MAIHDA — ACIC 2022 Track 1a

Yiyang Gao

2025-10-07

Contents

1 1) Background: What is MAIHDA?

MAIHDA (Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy) studies how outcomes differ across **intersectional strata** (e.g., combinations of demographic or contextual attributes). Key ideas:

- **Random intercepts** for strata capture **baseline** differences between intersections with **partial pooling** (stabilises small groups).
- **Random slopes** (e.g., for treatment) let the **effect** vary across strata.
- We report **variance partition** (e.g., VPC/ICC) and, when modeling slopes, the **SD of the slope** and the **intercept–slope correlation**.

Here we use **ACIC 2022 Track-1a** (practice-randomised, patient outcomes by year).

---

2 2) Load & Merge ACIC Track-1a

```
# --- pick a replicate (1..1200 for Track 1a zip) ---
replicate_id_int <- 52
replicate_id <- sprintf("%04d", replicate_id_int)

# --- base path (edit to your local path) ---
base_dir <- "/Users/constanceko/Desktop/MAIHDA/ACIC_track1a_20220404"

# --- file paths ---
fp_patient_year  <- file.path(base_dir,"patient_year",  sprintf("acic_patient_year_%s.csv",  replicate_
fp_patient       <- file.path(base_dir,"patient",       sprintf("acic_patient_%s.csv",       replicate_
fp_practice      <- file.path(base_dir,"practice",      sprintf("acic_practice_%s.csv",      replicate_
fp_practice_year <- file.path(base_dir,"practice_year", sprintf("acic_practice_year_%s.csv", replicate_

# fast reader
rd <- function(p){
  stopifnot(file.exists(p))
  as.data.frame(fread(p, showProgress = TRUE))
}

# --- read ---
patient_year  <- rd(fp_patient_year)
patient       <- rd(fp_patient)
practice      <- rd(fp_practice)
practice_year <- rd(fp_practice_year)

# --- minimal typing / cleaning ---
patient_year  <- mutate(patient_year,  id.patient = as.integer(id.patient), year = as.integer(year))
patient       <- mutate(patient,       id.patient = as.integer(id.patient), id.practice = as.integer(id
practice      <- mutate(practice,      id.practice = as.integer(id.practice))
practice_year <- mutate(practice_year, id.practice = as.integer(id.practice), year = as.integer(year))

# drop practice-level Y if present (Track 1 outcome must be patient-level Y)
if ("Y" %in% names(practice_year))
  practice_year <- practice_year[ , setdiff(names(practice_year), "Y"), drop = FALSE]

# --- merge star schema (keep observed patient-years) ---
df <- patient_year %>%
  inner_join(patient,  by = "id.patient") %>%          # adds V1-V5, id.practice
  left_join(practice,  by = "id.practice")             # adds X1-X9

# append Z, post, n.patients, and any available practice-year aggregates (V*_C or V*_avg)
keep_cols <- intersect(
  c("id.practice","year","Z","post","n.patients",
    "V1_C","V2_C","V3_C","V4_C","V5_C","V1_avg","V2_avg","V3_avg","V4_avg","V5_avg"),
  names(practice_year)
)
df <- df %>%
  left_join(practice_year[, keep_cols, drop = FALSE], by = c("id.practice","year"))

# panel helpers and analysis variables
df <- df %>%
  arrange(id.patient, year) %>%
  group_by(id.patient) %>%
```

```r
  mutate(Y_lag = dplyr::lag(Y)) %>%
  ungroup() %>%
  mutate(
    # Exposure: treated practice in post period
    W            = as.integer(Z == 1 & post == 1),
    year         = factor(year),
    id.practice  = factor(id.practice),
    id.patient   = factor(id.patient)
  )

# quick peek
dplyr::glimpse(df)
```

```
## Rows: 1,262,729
## Columns: 27
## $ id.patient  <fct> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 5, 5, 6, 6, 7~
## $ year        <fct> 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 1, 2, 3, 4, 1, 2, 1~
## $ Y           <dbl> 378.8274, 634.0839, 654.8388, 664.2329, 302.2389, 296.0461~
## $ id.practice <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ V1          <dbl> 6.097, 6.097, 6.097, 6.097, 12.313, 12.313, 12.313, 12.313~
## $ V2          <int> 1, 1, 1, 1, 4, 4, 4, 4, 2, 2, 2, 2, 3, 2, 2, 2, 2, 3, 3, 4~
## $ V3          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1~
## $ V4          <dbl> 0.674, 0.674, 0.674, 0.674, 0.762, 0.762, 0.762, 0.762, -0~
## $ V5          <chr> "B", "B", "B", "B", "B", "B", "B", "B", "B", "B", "B", "B"~
## $ X1          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ X2          <chr> "B", "B", "B", "B", "B", "B", "B", "B", "B", "B", "B", "B"~
## $ X3          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ X4          <chr> "B", "B", "B", "B", "B", "B", "B", "B", "B", "B", "B", "B"~
## $ X5          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ X6          <dbl> 25.483, 25.483, 25.483, 25.483, 25.483, 25.483, 25.483, 25~
## $ X7          <dbl> 6.111, 6.111, 6.111, 6.111, 6.111, 6.111, 6.111, 6.111, 6.~
## $ X8          <dbl> 0.398, 0.398, 0.398, 0.398, 0.398, 0.398, 0.398, 0.398, 0.~
## $ X9          <dbl> 25.908, 25.908, 25.908, 25.908, 25.908, 25.908, 25.908, 25~
## $ Z           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ post        <int> 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0~
## $ n.patients  <int> 918, 1010, 1004, 1010, 918, 1010, 1004, 1010, 918, 1010, 1~
## $ V1_avg      <dbl> 11.005, 11.224, 11.304, 11.257, 11.005, 11.224, 11.304, 11~
## $ V2_avg      <dbl> 3.026, 3.045, 3.053, 3.023, 3.026, 3.045, 3.053, 3.023, 3.~
## $ V3_avg      <dbl> 0.619, 0.595, 0.584, 0.581, 0.619, 0.595, 0.584, 0.581, 0.~
## $ V4_avg      <dbl> 0.463, 0.337, 0.256, 0.201, 0.463, 0.337, 0.256, 0.201, 0.~
## $ Y_lag       <dbl> NA, 378.8274, 634.0839, 654.8388, NA, 302.2389, 296.0461, ~
## $ W           <int> 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0~
```

Notes.

- Treatment is assigned at **practice** $\times$ **year** (`Z`), while the individual exposure is `W = 1` in post-treated practice-years.
- Outcome `Y` is **patient-year** monthly average expenditure (simulated).
- We keep patient-level `Y` and drop practice-level `Y`.

---

3 3) Build Intersectional Strata

We create tertiles (low/medium/high) for two patient covariates (V1, V2) and, if it varies, for a practice covariate (X1). If X1 has only one level in this replicate, we fall back to V1×V2 only.

```r
# Helper: 3-quantile binning for a numeric variable (with tiny jitter for ties)
bin3 <- function(x, nm) {
  xj <- x + rnorm(length(x), sd = 1e-8)
  qs <- unique(quantile(xj, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE))
  if (length(qs) < 4) return(factor(paste0(nm, "_L")))
  cut(xj, qs, include.lowest = TRUE, labels = paste0(nm, c("_L","_M","_H")))
}

# Discretise
df <- df %>%
  mutate(
    V1_q = if ("V1" %in% names(.)) bin3(V1, "V1") else factor("V1_L"),
    V2_q = if ("V2" %in% names(.)) bin3(V2, "V2") else factor("V2_L"),
    X1_q = if ("X1" %in% names(.)) bin3(X1, "X1") else factor("X1_L")
  )

# If X1_q has <2 levels, drop it from strata
nlev <- sapply(df[,c("V1_q","V2_q","X1_q")], \(x) nlevels(droplevels(factor(x))))
use_X1 <- nlev["X1_q"] >= 2

df <- df %>%
  mutate(
    strata = if (use_X1)
               interaction(V1_q, V2_q, X1_q, drop = TRUE)
             else
               interaction(V1_q, V2_q, drop = TRUE)
  )

# Main effects used as fixed effects (drop any single-level)
main_terms <- c("V1_q","V2_q", if (use_X1) "X1_q" else NULL)
main_terms <- main_terms[sapply(df[main_terms], \(x) nlevels(droplevels(factor(x))) > 1)]

table(df$strata)[1:10]
```

```
##
## V1_L.V2_L.X1_L V1_M.V2_L.X1_L V1_H.V2_L.X1_L V1_L.V2_M.X1_L V1_M.V2_M.X1_L
##          46385          46847          46843          46920          46731
## V1_H.V2_M.X1_L V1_L.V2_H.X1_L V1_M.V2_H.X1_L V1_H.V2_H.X1_L V1_L.V2_L.X1_M
##          46440          46615          46966          47163          46964
```

---

4 4) MAIHDA Models (A → B → C)

We follow a standard teaching flow (as in Leckie's tutorials):

- **Model A (fixed effects + practice RE):** baseline reference

- **Model B (add strata random intercept):** baseline intersectional heterogeneity
- **Model C (add treatment `W` + strata random slope):** treatment-effect heterogeneity across intersections

```r
# Build formulas robustly
fA <- as.formula(paste("Y ~", paste(main_terms, collapse = " + "),
                       "+ (1|id.practice)"))

fB <- as.formula(paste("Y ~", paste(main_terms, collapse = " + "),
                       "+ (1|id.practice) + (1|strata)"))

fC <- as.formula(paste("Y ~ W +", paste(main_terms, collapse = " + "),
                       "+ (1|id.practice) + (1 + W|strata)"))

mA <- lmer(fA, data = df, REML = TRUE)
mB <- lmer(fB, data = df, REML = TRUE)
mC <- lmer(fC, data = df, REML = TRUE)

summary(mB)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Y ~ V1_q + V2_q + X1_q + (1 | id.practice) + (1 | strata)
##    Data: df
##
## REML criterion at convergence: 23370241
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.7198 -0.3335 -0.2298 -0.0152 30.7037
##
## Random effects:
##  Groups      Name        Variance  Std.Dev.
##  id.practice (Intercept) 2.315e+04  152.137
##  strata      (Intercept) 4.735e+01    6.881
##  Residual                6.382e+06 2526.337
## Number of obs: 1262729, groups:  id.practice, 500; strata, 27
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 1294.276     10.444 123.928
## V1_qV1_M    -240.937      6.403 -37.630
## V1_qV1_H    -348.580      6.412 -54.364
## V2_qV2_M      -8.445      6.394  -1.321
## V2_qV2_H      -1.281      6.394  -0.200
## X1_qX1_M      12.360      7.469   1.655
## X1_qX1_H      10.814      9.376   1.153
##
## Correlation of Fixed Effects:
##          (Intr) V1_V1_M V1_V1_H V2_V2_M V2_V2_H X1_X1_M
## V1_qV1_M -0.307
## V1_qV1_H -0.307  0.502
## V2_qV2_M -0.306  0.001   0.001
## V2_qV2_H -0.306  0.000  -0.001   0.500
## X1_qX1_M -0.432  0.000   0.000   0.000   0.001
```

```
## X1_qX1_H -0.449  0.000   0.001   0.000   0.000   0.670
```

```r
summary(mC)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Y ~ W + V1_q + V2_q + X1_q + (1 | id.practice) + (1 + W | strata)
##    Data: df
##
## REML criterion at convergence: 23367932
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.7483 -0.3310 -0.2271 -0.0163 30.7093
##
## Random effects:
##  Groups      Name        Variance  Std.Dev. Corr
##  id.practice (Intercept)   22597.3  150.32
##  strata      (Intercept)     386.1   19.65
##              W              3940.9   62.78  -0.96
##  Residual                6370584.3 2524.00
## Number of obs: 1262729, groups:  id.practice, 500; strata, 27
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 1236.071     10.954 112.838
## W            334.500     14.016  23.865
## V1_qV1_M    -249.317      6.209 -40.151
## V1_qV1_H    -362.412      6.224 -58.231
## V2_qV2_M      -7.674      6.207  -1.236
## V2_qV2_H       1.781      6.207   0.287
## X1_qX1_M      12.145      7.272   1.670
## X1_qX1_H       8.918      9.258   0.963
##
## Correlation of Fixed Effects:
##          (Intr) W      V1_V1_M V1_V1_H V2_V2_M V2_V2_H X1_X1_M
## W        -0.352
## V1_qV1_M -0.283  0.000
## V1_qV1_H -0.284  0.001  0.500
## V2_qV2_M -0.284  0.000  0.000   0.000
## V2_qV2_H -0.283  0.000 -0.001  -0.001   0.500
## X1_qX1_M -0.405  0.001  0.000   0.001   0.001   0.001
## X1_qX1_H -0.423  0.002  0.000   0.001   0.000   0.000   0.672
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00528163 (tol = 0.002, component 1)
```

Why a practice random intercept? Track-1a is **cluster-randomised at practice** (over time). Including `(1|id.practice)` respects the assignment level and improves inference.

---

5 5) Diagnostics & Visuals

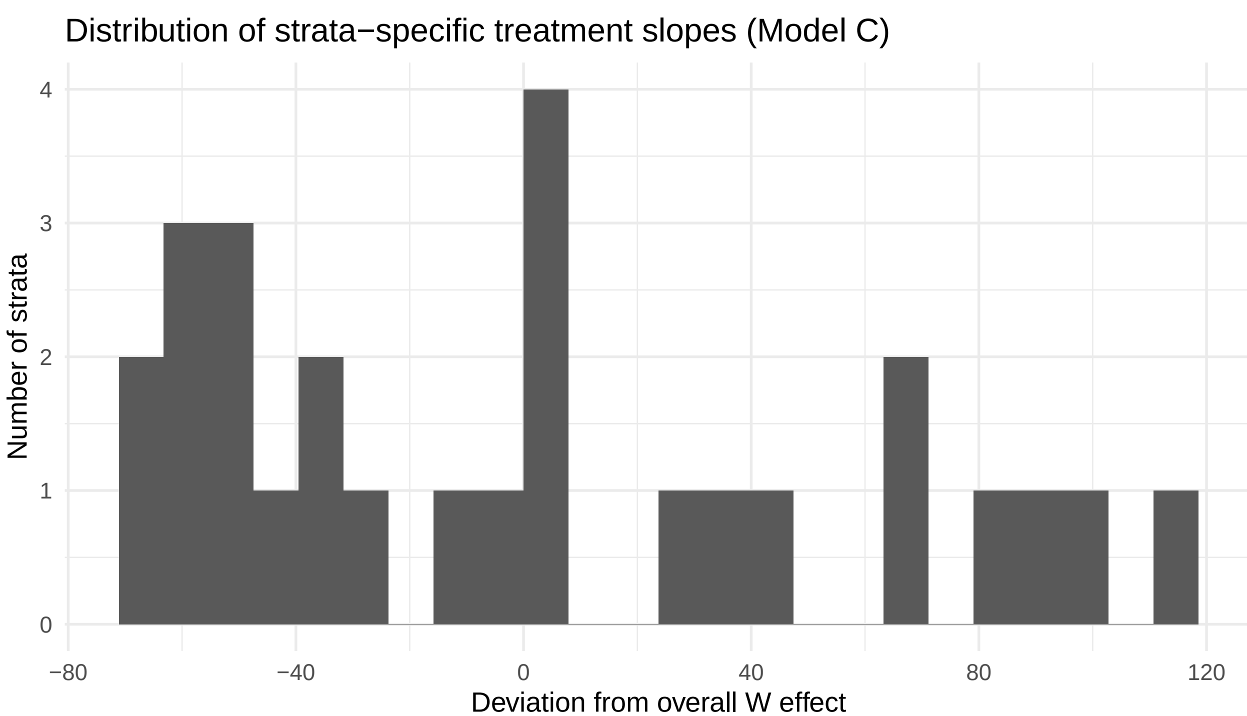
```
# VPC/ICC for strata in Model B
vcB <- as.data.frame(VarCorr(mB))
sigma_strata_B <- vcB$vcov[vcB$grp == "strata" & vcB$var1 == "(Intercept)"]
sigma_resid_B  <- sigma(mB)^2
VPC_B <- sigma_strata_B / (sigma_strata_B + sigma_resid_B)

cat(sprintf("VPC (strata; Model B) = %.4f\n", VPC_B))
```

```
## VPC (strata; Model B) = 0.0000
```

```
# Random slopes of W in Model C
reC <- ranef(mC)$strata
stopifnot(ncol(reC) >= 2) # columns: Intercept, W
colnames(reC)[1:2] <- c("Intercept_RE","W_slope_RE")

ggplot(as.data.frame(reC), aes(x = W_slope_RE)) +
  geom_histogram(bins = 24) +
  labs(
    title = "Distribution of strata-specific treatment slopes (Model C)",
    x = "Deviation from overall W effect", y = "Number of strata"
  ) +
  theme_minimal()
```



6 6) Results (auto summary + interpretation)

```r
# Fixed effect (ATE proxy) for W in Model C
coefs <- broom.mixed::tidy(mC, effects = "fixed")
ate_row <- coefs[coefs$term == "W", ]
ATE_W   <- ate_row$estimate
SE_W    <- ate_row$std.error
t_W     <- ate_row$statistic

# Variance components for strata in Model C
vcC <- as.data.frame(VarCorr(mC))
sd_W  <- sqrt(vcC$vcov[vcC$grp=="strata" & vcC$var1=="W" & vcC$var2=="W"])
sd_int<- sqrt(vcC$vcov[vcC$grp=="strata" & vcC$var1=="(Intercept)" & vcC$var2=="(Intercept)"])
cov_iw<- vcC$vcov[vcC$grp=="strata" & vcC$var1=="(Intercept)" & vcC$var2=="W"]
corr_iw <- cov_iw / (sd_W * sd_int)

# Re-compute VPC for Model B (printed above)
VPC_B <- as.numeric(VPC_B)

cat(glue("
**Auto-reported summary (Model C)**
- Fixed effect of W (ATE proxy): {round(ATE_W, 2)} (SE {round(SE_W, 2)}, t = {round(t_W, 2)})
- Strata VPC (Model B): {round(VPC_B, 4)}
- SD of W random slope (strata): {round(sd_W, 2)}
- Corr(Intercept, W) at strata: {round(corr_iw, 2)}
"))
```

```
## **Auto-reported summary (Model C)**
## - Fixed effect of W (ATE proxy): 334.5 (SE 14.02, t = 23.86)
## - Strata VPC (Model B): 0
## - SD of W random slope (strata): NA
## - Corr(Intercept, W) at strata: NA **Auto-reported summary (Model C)**
## - Fixed effect of W (ATE proxy): 334.5 (SE 14.02, t = 23.86)
## - Strata VPC (Model B): 0
## - SD of W random slope (strata): NA
## - Corr(Intercept, W) at strata: NA
```

```r
# One-paragraph interpretation using the numbers above
low  <- ATE_W - 2*sd_W
high <- ATE_W + 2*sd_W
cat(glue("
**Interpretation (reader-friendly):**
On average, exposure in treated post-periods (W = 1) is associated with an increase in the outcome of al
Baseline differences between intersectional strata are **{ifelse(VPC_B < 0.02, 'very small', ifelse(VPC
However, **treatment effects vary across strata**: the random-slope SD is about **{round(sd_W,1)}**, su
The **negative/positive** correlation between strata intercepts and W slopes (corr   {round(corr_iw,2)})
"))
```

Interpretation (reader-friendly):
On average, exposure in treated post-periods (W = 1) is associated with an increase in the outcome of about **335** units (SE 14).
Baseline differences between intersectional strata are **very small** (VPC   0%).

However, **treatment effects vary across strata**: the random-slope SD is about **NA**, suggesting many strata lie roughly between **NA** and **NA** around the average effect.
The **negative/positive** correlation between strata intercepts and W slopes (corr NA) indicates that strata with **NA**. **Interpretation (reader-friendly):**
On average, exposure in treated post-periods (W = 1) is associated with an increase in the outcome of about **335** units (SE 14).
Baseline differences between intersectional strata are **very small** (VPC 0%).
However, **treatment effects vary across strata**: the random-slope SD is about **NA**, suggesting many strata lie roughly between **NA** and **NA** around the average effect.
The **negative/positive** correlation between strata intercepts and W slopes (corr NA) indicates that strata with **NA**.

---

7 7) Notes & Tips

- If a binned covariate collapses to one level in a replicate (e.g., `X1_q`), we automatically drop it from fixed effects and from the strata definition.
- If a model becomes **singular** (a variance 0), that's informative: the data do not support that random component for this replicate.
- Given Track-1a's design, keeping a **practice random intercept** is good practice.