ACIC 2022 Track 1 (Patient–Year) Data Structure + Merge and Readable Strata Guide

Yiyang Gao

2025-10-07

Contents

1 The Four Core Files (Structure and Meaning)

Track 1 provides four CSVs per replicate (e.g., `xxxx = 0001`, public replicates often 0001–1200).

```
four_files <- tribble(
  ~File, ~`Level (Uniqueness)`, ~`Key(s)`, ~`Main contents`,
  "patient_xxxx.csv",        "Patient (time-invariant)",      "id.patient",                "One
  "patient_year_xxxx.csv",   "Patient-Year (time-varying)",   "id.patient, year",          "One
  "practice_xxxx.csv",       "Practice (time-invariant)",     "id.practice",               "One
  "practice_year_xxxx.csv",  "Practice-Year (time-varying)",  "id.practice, year",         "One
)

kable(
  four_files,
  format = ifelse(knitr::is_latex_output(), "latex", "simple"),
  booktabs = knitr::is_latex_output(),
  align = "l",
  caption = "Overview of ACIC Track 1 data files"
)
```

> **Track-1 outcome:** keep **patient-level** `Y` from `patient_year_xxxx.csv`; **drop** the practice-level `Y` from `practice_year_xxxx.csv`.

Table 1: Overview of ACIC Track 1 data files

File	Level (Uniqueness)	Key(s)	Main contents
patient_xxxx.csv	Patient (time-invariant)	id.patient	One row per patient. Patient covariates V1–V
patient_year_xxxx.csv	Patient–Year (time-varying)	id.patient, year	One row per patient per year. Outcome Y =
practice_xxxx.csv	Practice (time-invariant)	id.practice	One row per practice. Practice covariates X1-
practice_year_xxxx.csv	Practice–Year (time-varying)	id.practice, year	One row per practice per year: Z (treatment)

```
practice (X1-X9)
  ---> practice_year (Z, post, n.patients, aggregated V*, practice Y [drop in Track 1])
       ---> patient (V1-V5, id.practice)
            ---> patient_year (patient Y, by year)
> Practice-level Y exists but should be dropped for Track 1 outcome modelling.
```

This yields a **cluster-randomised** (practice-level assignment) longitudinal RCT.

2 Peek at the first rows (for each of the four files)

Below we show the **first 3 rows** of each file for `replicate = 0001`.

```r
# ---- File paths ----
base_dir <- params$data_dir
rid      <- sprintf("%04d", as.integer(params$replicate_id))

dir_patient       <- fs::path(base_dir, "patient")
dir_patient_year  <- fs::path(base_dir, "patient_year")
dir_practice      <- fs::path(base_dir, "practice")
dir_practice_year <- fs::path(base_dir, "practice_year")

fname <- function(prefix) glue("acic_{prefix}_{rid}.csv")
paths <- list(
  patient       = fs::path(dir_patient,       fname("patient")),
  patient_year  = fs::path(dir_patient_year,  fname("patient_year")),
  practice      = fs::path(dir_practice,      fname("practice")),
  practice_year = fs::path(dir_practice_year, fname("practice_year"))
)

# ---- A function to fall back to the provided sample ----
show_head_or_fallback <- function(path, fallback_block, n = 3) {
  if (fs::file_exists(path)) {
    cat(glue("**File:** `{fs::path_file(path)}`  \n"))
    x <- readr::read_csv(path, n_max = n, show_col_types = FALSE, progress = FALSE)
    print(knitr::kable(
      x,
      format = ifelse(knitr::is_latex_output(), "latex", "simple"),
      booktabs = knitr::is_latex_output(),
      align = "l"
    ))
  } else {
    cat(glue("**File not found locally, showing provided sample instead:** `{fs::path_file(path)}`  \n\
```

```
    cat("```\n")
    cat(fallback_block)
    cat("\n```\n\n")
  }
}

# ---- Provided samples (first lines) ----
sample_patient <- "id.patient\tid.practice\tV1\tV2\tV3\tV4\tV5
1\t1\t14.526\t3\t0\t1.85\tB
2\t1\t7.451\t4\t0\t0.07\tA
3\t1\t17.301\t1\t1\t1.583\tA"

sample_patient_year <- "id.patient\tyear\tY
1\t1\t3625.524311
2\t1\t395.6187556
3\t1\t611.5372043"

sample_practice <- "id.practice\tX1\tX2\tX3\tX4\tX5\tX6\tX7\tX8\tX9
1\t0\tA\t1\tA\t1\t20.774\t14.153\t0.161\t43.432
2\t0\tA\t0\tC\t0\t33.566\t3.285\t0.557\t12.722
3\t0\tC\t1\tA\t1\t57.283\t11.178\t0.257\t-7.353"

sample_practice_year <- "id.practice\tyear\tY\tZ\tpost\tn.patients\tV1_avg\tV2_avg\tV3_avg\tV4_avg\tV5_
1\t1\t765.9213094\t0\t0\t2716\t11.576\t2.977\t0.514\t-0.087\t0.762\t0.19\t0.047
1\t2\t976.9141846\t0\t0\t2782\t11.46\t2.964\t0.517\t-0.064\t0.767\t0.185\t0.048
1\t3\t1131.252587\t0\t1\t2595\t11.45\t2.981\t0.513\t-0.072\t0.765\t0.187\t0.048"

# ---- Print previews ----
cat("### `patient_XXXX.csv`\n\n")
```

‘patient_XXXX.csv‘

```
show_head_or_fallback(paths$patient, sample_patient)
```

File: ‘acic_patient_0001.csv‘
\begin{tabular}{lllllll}
\toprule
id.patient & id.practice & V1 & V2 & V3 & V4 & V5\\
\midrule
1 & 1 & 14.526 & 3 & 0 & 1.850 & B\\
2 & 1 & 7.451 & 4 & 0 & 0.070 & A\\
3 & 1 & 17.301 & 1 & 1 & 1.583 & A\\
\bottomrule
\end{tabular}

```
cat("### `patient_year_XXXX.csv`\n\n")
```

‘patient_year_XXXX.csv‘

```
show_head_or_fallback(paths$patient_year, sample_patient_year)
```

```
## **File:** ‘acic_patient_year_0001.csv‘
## \begin{tabular}{lll}
## \toprule
## id.patient & year & Y\\
## \midrule
## 1 & 1 & 3625.5243\\
## 2 & 1 & 395.6188\\
## 3 & 1 & 611.5372\\
## \bottomrule
## \end{tabular}
```

```
cat("### `practice_XXXX.csv`\n\n")
```

```
## ### ‘practice_XXXX.csv‘
```

```
show_head_or_fallback(paths$practice, sample_practice)
```

```
## **File:** ‘acic_practice_0001.csv‘
## \begin{tabular}{lllllllllll}
## \toprule
## id.practice & X1 & X2 & X3 & X4 & X5 & X6 & X7 & X8 & X9\\
## \midrule
## 1 & 0 & A & 1 & A & 1 & 20.774 & 14.153 & 0.161 & 43.432\\
## 2 & 0 & A & 0 & C & 0 & 33.566 & 3.285 & 0.557 & 12.722\\
## 3 & 0 & C & 1 & A & 1 & 57.283 & 11.178 & 0.257 & -7.353\\
## \bottomrule
## \end{tabular}
```

```
cat("### `practice_year_XXXX.csv`\n\n")
```

```
## ### ‘practice_year_XXXX.csv‘
```

```
show_head_or_fallback(paths$practice_year, sample_practice_year)
```

```
## **File:** ‘acic_practice_year_0001.csv‘
## \begin{tabular}{lllllllllllll}
## \toprule
## id.practice & year & Y & Z & post & n.patients & V1\_avg & V2\_avg & V3\_avg & V4\_avg & V5\_A\_avg &
## \midrule
## 1 & 1 & 1018.882 & 1 & 0 & 113 & 10.808 & 2.920 & 0.540 & 0.298 & 0.690 & 0.274 & 0.035\\
## 1 & 2 & 1607.136 & 1 & 0 & 109 & 10.768 & 2.872 & 0.532 & 0.274 & 0.706 & 0.266 & 0.028\\
## 1 & 3 & 1245.213 & 1 & 1 & 121 & 10.896 & 2.868 & 0.529 & 0.185 & 0.727 & 0.231 & 0.041\\
## \bottomrule
## \end{tabular}
```

3 Merge Rules

1. **Add outcomes to patients:** join `patient` → `patient_year` by `id.patient`.
2. **Attach practice covariates:** join with `practice` by `id.practice`.

3. **Attach practice-year treatment and context:** join with `practice_year` by `c(id.practice, year)`.
4. **Drop practice-level `Y`** to avoid ambiguity.

```
# ---- Parameters ----
base_dir <- params$data_dir
rid      <- sprintf("%04d", as.integer(params$replicate_id))  # enforce "0001" format
read_engine <- getOption("acic.read_engine", "readr")         # "readr", "vroom", or "fread"

# ---- Paths ----
dir_patient       <- fs::path(base_dir, "patient")
dir_patient_year  <- fs::path(base_dir, "patient_year")
dir_practice      <- fs::path(base_dir, "practice")
dir_practice_year <- fs::path(base_dir, "practice_year")

fname <- function(prefix) glue("acic_{prefix}_{rid}.csv")
files <- list(
  patient       = fs::path(dir_patient,       fname("patient")),
  patient_year  = fs::path(dir_patient_year,  fname("patient_year")),
  practice      = fs::path(dir_practice,      fname("practice")),
  practice_year = fs::path(dir_practice_year, fname("practice_year"))
)

# ---- Column specs ----
colspec_patient_readr <- readr::cols(
  id.patient  = readr::col_integer(),
  id.practice = readr::col_integer(),
  V1 = readr::col_double(),
  V2 = readr::col_integer(),
  V3 = readr::col_integer(),
  V4 = readr::col_double(),
  V5 = readr::col_character()
)

colspec_patient_year_readr <- readr::cols(
  id.patient = readr::col_integer(),
  year       = readr::col_integer(),
  Y          = readr::col_double()
)

colspec_practice_readr <- readr::cols(
  id.practice = readr::col_integer(),
  X1 = readr::col_integer(),
  X2 = readr::col_character(),
  X3 = readr::col_integer(),
  X4 = readr::col_character(),
  X5 = readr::col_integer(),
  X6 = readr::col_double(),
  X7 = readr::col_double(),
  X8 = readr::col_double(),
  X9 = readr::col_double()
)

colspec_practice_year_readr <- readr::cols(
```

```r
  id.practice = readr::col_integer(),
  year        = readr::col_integer(),
  Y           = readr::col_double(),   # to be dropped later
  Z           = readr::col_integer(),
  post        = readr::col_integer(),
  n.patients  = readr::col_integer(),
  V1_avg      = readr::col_double(),
  V2_avg      = readr::col_double(),
  V3_avg      = readr::col_double(),
  V4_avg      = readr::col_double(),
  V5_A_avg    = readr::col_double(),
  V5_B_avg    = readr::col_double(),
  V5_C_avg    = readr::col_double()
)

# ---- Reader wrapper ----
`%||%` <- function(a, b) if (!is.null(a)) a else b
read_csv_smart <- function(path, spec_readr = NULL) {
  switch(read_engine,
    readr = readr::read_csv(path, col_types = spec_readr %||% readr::cols(), show_col_types = FALSE),
    vroom = vroom::vroom(path, altrep = TRUE),
    fread = data.table::fread(path, data.table = FALSE),
    stop("Unknown read_engine: ", read_engine)
  )
}

# ---- Read ----
patient       <- read_csv_smart(files$patient,       colspec_patient_readr)
patient_year  <- read_csv_smart(files$patient_year,  colspec_patient_year_readr)
practice      <- read_csv_smart(files$practice,      colspec_practice_readr)
practice_year <- read_csv_smart(files$practice_year, colspec_practice_year_readr)

# ---- Merge 1: keep observed patient-years only ----
df <- patient %>%
  inner_join(patient_year, by = "id.patient")

# ---- Merge 2: + practice (X1-X9) ----
df <- df %>%
  left_join(practice, by = "id.practice")

# ---- Merge 3: + practice_year (Z, post, aggregates) ----
df <- df %>%
  left_join(practice_year, by = c("id.practice", "year"), suffix = c("", ".practice"))

# ---- Keep only the patient-level Y ----
df <- df %>% select(-any_of("Y.practice"))

# ---- Basic checks ----
stopifnot(all(c("id.patient","id.practice","year","Y","Z","post") %in% names(df)))
stopifnot(!any(is.na(df$year)))
message(glue("Rows: {nrow(df)}, Patients: {dplyr::n_distinct(df$id.patient)}, Practices: {dplyr::n_dist

# Quick peek
```

Table 2: Human-readable teaching semantics for ACIC variables

Variable	Meaning (teaching assumption)
V1	Baseline risk score (continuous) — recoded into quartiles: Low / Medium / High / Very High.
V2	Comorbidity grade (integer) — banded to: 0–1 / 2–3 / 4–5 / 6–7.
V3	Sex at registration (binary) — 0 = Male, 1 = Female.
V4	Deprivation index (continuous) — recoded into tertiles: Low / Mid / High.
V5	Insurance plan (categorical A/B/C) — labelled: Plan A / Plan B / Plan C.
X1	Rurality (binary) — 0 Urban, 1 Rural.
X2	Region class (A/B/C).
X3	Teaching practice (binary) — 0 No, 1 Yes.
X4	Ownership type (A/B/C).
X5	Urgent care facility (binary) — 0 No, 1 Yes.
X6–X9	Practice metrics (continuous) — used as controls (not in strata).
Z	Treatment assignment at practice–year.
post	Post-intervention period indicator.

```
dplyr::glimpse(df, width = 80)
```

4 Assign concrete meanings and recode (for readable strata)

We now **assign clear interpretations** to abstract variables and **recode** where appropriate. These are **teaching assumptions** for readability.

```
semantics <- tribble(
  ~Variable, ~`Meaning (teaching assumption)`,
  "V1", "Baseline risk score (continuous) - recoded into quartiles: Low / Medium / High / Very High.",
  "V2", "Comorbidity grade (integer) - banded to: 0-1 / 2-3 / 4-5 / 6-7.",
  "V3", "Sex at registration (binary) - 0 = Male, 1 = Female.",
  "V4", "Deprivation index (continuous) - recoded into tertiles: Low / Mid / High.",
  "V5", "Insurance plan (categorical A/B/C) - labelled: Plan A / Plan B / Plan C.",
  "X1", "Rurality (binary) - 0 Urban, 1 Rural.",
  "X2", "Region class (A/B/C).",
  "X3", "Teaching practice (binary) - 0 No, 1 Yes.",
  "X4", "Ownership type (A/B/C).",
  "X5", "Urgent care facility (binary) - 0 No, 1 Yes.",
  "X6-X9", "Practice metrics (continuous) - used as controls (not in strata).",
  "Z", "Treatment assignment at practice-year.",
  "post", "Post-intervention period indicator."
)

kable(
  semantics,
  format = ifelse(knitr::is_latex_output(), "latex", "simple"),
  booktabs = knitr::is_latex_output(),
  align = "l",
  caption = "Human-readable teaching semantics for ACIC variables"
)
```

```r
# This chunk assumes df exists from the merge step.

label_patient_vars <- function(dat){
  dat %>%
    mutate(
      # V1: baseline risk → quartiles (cut by sample quantiles)
      V1_band = cut(
        V1,
        breaks = quantile(V1, probs = seq(0, 1, 0.25), na.rm = TRUE),
        include.lowest = TRUE, dig.lab = 6
      ),
      V1_band = forcats::fct_recode(
        V1_band,
        "Risk: Low (Q1)"       = levels(V1_band)[1],
        "Risk: Medium (Q2)"    = levels(V1_band)[2],
        "Risk: High (Q3)"      = levels(V1_band)[3],
        "Risk: Very High (Q4)" = levels(V1_band)[4]
      ),

      # V2: comorbidity grade → bands
      V2_band = case_when(
        V2 <= 1          ~ "Comorbidity: 0-1",
        V2 %in% 2:3      ~ "Comorbidity: 2-3",
        V2 %in% 4:5      ~ "Comorbidity: 4-5",
        V2 >= 6          ~ "Comorbidity: 6-7",
        TRUE             ~ NA_character_
      ) |> factor(),

      # V3: sex (0=Male, 1=Female)
      sex = factor(if_else(V3 == 1, "Female", "Male"),
                   levels = c("Male","Female"),
                   labels = c("Sex: Male","Sex: Female")),

      # V4: deprivation index → tertiles
      V4_band = cut(
        V4,
        breaks = quantile(V4, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE),
        include.lowest = TRUE, dig.lab = 6
      ),
      V4_band = forcats::fct_recode(
        V4_band,
        "Deprivation: Low"  = levels(V4_band)[1],
        "Deprivation: Mid"  = levels(V4_band)[2],
        "Deprivation: High" = levels(V4_band)[3]
      ),

      # V5: plan labels
      plan = forcats::fct_recode(factor(V5),
                         "Plan: A" = "A",
                         "Plan: B" = "B",
                         "Plan: C" = "C")
    )
}
```

```r
# Apply patient-side labelling
df <- df %>% label_patient_vars()

# practice-side readable labels (useful for descriptives)
df <- df %>%
  mutate(
    rurality = factor(if_else(X1 == 1, "Rural", "Urban"),
                      levels = c("Urban","Rural"),
                      labels = c("Rurality: Urban","Rurality: Rural")),
    region   = forcats::fct_recode(factor(X2),
                           "Region: A" = "A",
                           "Region: B" = "B",
                           "Region: C" = "C"),
    teaching = factor(if_else(X3 == 1, "Yes", "No"),
                      levels = c("No","Yes"),
                      labels = c("Teaching practice: No","Teaching practice: Yes")),
    owner    = forcats::fct_recode(factor(X4),
                           "Ownership: A" = "A",
                           "Ownership: B" = "B",
                           "Ownership: C" = "C"),
    urgent   = factor(if_else(X5 == 1, "Yes", "No"),
                      levels = c("No","Yes"),
                      labels = c("Urgent care: No","Urgent care: Yes"))
  )

# Construct human-readable strata (patient-side only)
df <- df %>%
  mutate(
    strata_label = interaction(sex, V1_band, V2_band, V4_band, plan,
                               sep = " | ", drop = TRUE) |> forcats::fct_drop(),
    # A compact version for charts
    strata_short = glue::glue(
      "{ifelse(sex=='Sex: Female','F','M')} | ",
      "{forcats::fct_collapse(V1_band, Low='Risk: Low (Q1)', Med='Risk: Medium (Q2)', High=c('Risk: High
      "{stringr::str_replace(V2_band,'Comorbidity: ','C: ')} | ",
      "{stringr::str_replace(V4_band,'Deprivation: ','Dep: ')} | ",
      "{stringr::str_replace(plan,'Plan: ','P ')}"
    )
  )

# Quick checks
message("Preview of labelled variables and strata:")
df %>%
  dplyr::select(id.patient, year, Y, Z, post, sex, V1_band, V2_band, V4_band, plan, strata_label) %>%
  head(8) %>% print(n=8)
```

5 Quick descriptive: strata size and mean outcome

```r
# Summarise strata sizes and mean outcome
strata_summary <- df %>%
```

```r
  group_by(strata_label) %>%
  summarise(
    n = n(),
    patients = n_distinct(id.patient),
    practices = n_distinct(id.practice),
    mean_Y = mean(Y, na.rm = TRUE),
    mean_Z = mean(Z, na.rm = TRUE)
  ) %>%
  arrange(desc(n))

kable(
  head(strata_summary, 15),
  format = ifelse(knitr::is_latex_output(), "latex", "simple"),
  booktabs = knitr::is_latex_output(),
  caption = "Top strata by size (n), with mean outcome and exposure rate"
)
```

6 Variable Roles in Modelling

```r
vrm <- tribble(
  ~Variable,       ~Level,                        ~Origin,                       ~Original,
  "Y",             "Patient-Year",                "patient_year_xxxx.csv",       "Patient annual outcor
  "Z",             "Practice-Year",               "practice_year_xxxx.csv",      "Cluster-level treatm
  "post",          "Practice-Year",               "practice_year_xxxx.csv",      "Post-intervention ti
  "V1",            "Patient",                     "patient_xxxx.csv",            "Continuous patient c
  "V2",            "Patient",                     "patient_xxxx.csv",            "Discrete numeric cov
  "V3",            "Patient",                     "patient_xxxx.csv",            "Binary indicator (0/
  "V4",            "Patient",                     "patient_xxxx.csv",            "Continuous numeric c
  "V5",            "Patient",                     "patient_xxxx.csv",            "Categorical variable
  "X1",            "Practice",                    "practice_xxxx.csv",           "Binary covariate (0/
  "X2",            "Practice",                    "practice_xxxx.csv",           "Categorical covariat
  "X3",            "Practice",                    "practice_xxxx.csv",           "Binary covariate (0/
  "X4",            "Practice",                    "practice_xxxx.csv",           "Categorical covariat
  "X5",            "Practice",                    "practice_xxxx.csv",           "Binary covariate (0/
  "X6-X9",         "Practice",                    "practice_xxxx.csv",           "Continuous practice-
  "V1_avg-V4_avg", "Practice-Year",               "practice_year_xxxx.csv",      "Practice-year means
  "V5_A_avg etc.", "Practice-Year",               "practice_year_xxxx.csv",      "Proportions of V5 ca
  "n.patients",    "Practice-Year",               "practice_year_xxxx.csv",      "Number of patients i
  "id.patient",    "Patient / Patient-Year",      "all joined tables",           "Unique patient identi
  "id.practice",   "Practice / Practice-Year",    "all joined tables",           "Unique practice (clus
  "year",          "Patient-Year / Practice-Year","patient_year / practice_year","Observation time ind
)

if (knitr::is_html_output()) {
  suppressPackageStartupMessages(library(DT))
  DT::datatable(
    vrm,
    rownames = FALSE,
    options = list(pageLength = 25, scrollX = TRUE),
    caption = htmltools::tags$caption(style = 'caption-side: top; text-align: left;',
                                      'Variable Roles in Modelling (Expanded Explanation)')
```

```
  )
} else {
  txt <- paste(capture.output(print(vrm, n = nrow(vrm), width = Inf)), collapse = "\n")
  cat("**Variable Roles in Modelling (Expanded Explanation)**\n\n")
  cat("```\n"); cat(txt); cat("\n```\n")
}
```

```
## **Variable Roles in Modelling (Expanded Explanation)**
##
## ```
## # A tibble: 20 x 5
##    Variable      Level                       Origin
##    <chr>         <chr>                       <chr>
##  1 Y             Patient-Year                patient_year_xxxx.csv
##  2 Z             Practice-Year               practice_year_xxxx.csv
##  3 post          Practice-Year               practice_year_xxxx.csv
##  4 V1            Patient                     patient_xxxx.csv
##  5 V2            Patient                     patient_xxxx.csv
##  6 V3            Patient                     patient_xxxx.csv
##  7 V4            Patient                     patient_xxxx.csv
##  8 V5            Patient                     patient_xxxx.csv
##  9 X1            Practice                    practice_xxxx.csv
## 10 X2            Practice                    practice_xxxx.csv
## 11 X3            Practice                    practice_xxxx.csv
## 12 X4            Practice                    practice_xxxx.csv
## 13 X5            Practice                    practice_xxxx.csv
## 14 X6-X9         Practice                    practice_xxxx.csv
## 15 V1_avg-V4_avg Practice-Year               practice_year_xxxx.csv
## 16 V5_A_avg etc. Practice-Year               practice_year_xxxx.csv
## 17 n.patients    Practice-Year               practice_year_xxxx.csv
## 18 id.patient    Patient / Patient-Year      all joined tables
## 19 id.practice   Practice / Practice-Year     all joined tables
## 20 year          Patient-Year / Practice-Year patient_year / practice_year
##    Original
##    <chr>
##  1 Patient annual outcome (continuous numeric).
##  2 Cluster-level treatment assignment (0/1).
##  3 Post-intervention time indicator.
##  4 Continuous patient covariate (abstract).
##  5 Discrete numeric covariate (small integer range).
##  6 Binary indicator (0/1).
##  7 Continuous numeric covariate.
##  8 Categorical variable (A/B/C).
##  9 Binary covariate (0/1).
## 10 Categorical covariate (A/B/C).
## 11 Binary covariate (0/1).
## 12 Categorical covariate (A/B/C).
## 13 Binary covariate (0/1).
## 14 Continuous practice-level covariates.
## 15 Practice-year means of V1-V4.
## 16 Proportions of V5 categories (A/B/C).
## 17 Number of patients in the practice that year.
## 18 Unique patient identifier.
```

```
## 19 Unique practice (cluster) identifier.
## 20 Observation time index (integer).
##    Teaching
##    <chr>
##  1 Outcome: monthly average medical expenditure for that patient and year; Trac~
##  2 Treatment indicator: whether the practice received the intervention (1 treat~
##  3 Post-period flag: 1 after intervention, 0 before; used for time interactions~
##  4 Baseline risk score: recoded into quartiles (Low / Medium / High / Very High~
##  5 Comorbidity grade: banded 0-1 / 2-3 / 4-5 / 6-7.
##  6 Sex at registration: 0 Male, 1 Female; labelled as factors.
##  7 Deprivation index: recoded into tertiles (Low / Mid / High).
##  8 Insurance plan: Plan A / Plan B / Plan C (used for strata).
##  9 Rurality: 0 Urban, 1 Rural.
## 10 Region class: Region A / Region B / Region C.
## 11 Teaching practice: 0 No, 1 Yes.
## 12 Ownership type: Ownership A / B / C.
## 13 Urgent-care facility: 0 No, 1 Yes.
## 14 Practice metrics: contextual measures (e.g., staff experience, mean income, ~
## 15 Practice-year averages of risk, comorbidity, deprivation, etc.; contextual c~
## 16 Practice-year category shares: proportion in Plan A / Plan B / Plan C.
## 17 Practice-year sample size: for weighting/scaling aggregated statistics.
## 18 Merge key; level-1 random effect in multilevel models.
## 19 Merge key; level-2 (cluster) random effect.
## 20 Year of observation: defines longitudinal sequences.
## ```
```

Common pitfalls

- Do **not** mix the two `Y`s — use **patient `Y`** for Track-1.
- Join keys must align: `id.patient`, `id.practice`, `year`.
- Patients do not appear in every year; this is **by design** (ageing in/out, death), not missingness.