Missing Data in Clinical Research

Diagnose the pattern, preserve uncertainty and audit every imputation

Rverse Analytics

Missingness is part of the data-generating process. Do not choose complete-case analysis or multiple imputation from a percentage alone; consider why values are missing and which variables predict missingness.

Mechanism map

Mechanism	Working description	What observed data can establish
MCAR	Missingness unrelated to observed or unobserved values	can be challenged, rarely proven
MAR	Missingness explained by observed information	operational assumption for MI
MNAR	Missingness still depends on the unseen value	needs sensitivity analysis

Audit before modelling

```
library(dplyr)
library(mice)

miss <- tibble(
  variable = names(d),
  n_missing = colSums(is.na(d)),
  pct_missing = 100 * colMeans(is.na(d))
) |>
  arrange(desc(pct_missing))

md.pattern(d, rotate.names = TRUE)

# Is outcome missingness associated with observed variables?
d |>
  mutate(outcome_missing = is.na(outcome)) |>
  group_by(outcome_missing) |>
  summarise(across(c(age, baseline_score),
    list(mean = ~mean(.x, na.rm = TRUE), n = ~sum(!is.na(.x)))))
```

Check impossible values, structural missingness, skip patterns, duplicate records and whether missingness differs by treatment, centre, visit or outcome severity.

Choose an analysis route

Situation	Reasonable starting point	Main limitation
Outcome fully observed; covariate loss trivial	complete-case + sensitivity check	precision loss; selection bias possible
Repeated outcomes under likelihood model	mixed model using available outcomes	MAR still assumed
Several incomplete predictors/outcomes	multiple imputation	model must be compatible with analysis
Censored time-to-event data	survival model; impute incomplete covariates	do not impute censoring indicator casually
Plausible MNAR	delta/tipping-point sensitivity analysis	assumptions must be explicit

Multiple imputation with `mice`

```
vars <- d |>
  select(outcome, treatment, age, sex, baseline_score, center)

ini  <- mice(vars, maxit = 0, printFlag = FALSE)
meth <- ini$method
pred <- ini$predictorMatrix

# Never impute randomized treatment or immutable identifiers
meth[c("treatment", "center")] <- ""
pred[, "treatment"] <- 1
diag(pred) <- 0

# Examples: continuous / binary / unordered categorical
meth["baseline_score"] <- "pmm"
```

```r
meth["outcome"] <- "pmm"       # use "logreg" if binary
meth["sex"] <- "logreg"

imp <- mice(vars, m = 40, maxit = 20,
            method = meth, predictorMatrix = pred,
            seed = 20260711, printFlag = FALSE)

fit_mi <- with(imp, lm(outcome ~ treatment + baseline_score + age + sex))
pooled <- pool(fit_mi)
summary(pooled, conf.int = TRUE)
```

Use at least enough imputations to make Monte Carlo error negligible relative to the reported SE. Include the outcome, exposure, analysis covariates and good auxiliary predictors of missingness or values.

Diagnostics and completed data

```r
plot(imp)                        # convergence by iteration
densityplot(imp, ~ baseline_score)
stripplot(imp, baseline_score ~ .imp, pch = 20, cex = 0.5)

long <- complete(imp, action = "long", include = TRUE)
one  <- complete(imp, action = 1)  # inspect, not analyse alone
```

Red flags: drifting chains, implausible ranges, imputed categories that violate design, and distributions far from observed values without a scientific explanation.

Sensitivity analysis

```r
# Delta-adjust imputed outcomes to represent worse unobserved values
delta <- -3
long_delta <- complete(imp, "long", include = FALSE) |>
  mutate(outcome = if_else(.imp > 0 & is.na(d$outcome[.id]),
                           outcome + delta, outcome))

# Compare the estimand across a prespecified delta grid
delta_grid <- c(-5, -3, -1, 0, 1)
```

For formal MNAR work, define the departure from MAR, repeat the full pooled analysis, and report the value at which the conclusion changes.

Reporting checklist

- Give missing counts by variable, group and analysis time point.
- State the assumed mechanism, imputation model, methods, predictors, `m`, iterations and seed.
- Confirm that bounds, interactions, nonlinear terms, clustering and design variables were respected.
- Report pooled estimates and Rubin-rule uncertainty, not results from one completed dataset.
- Compare complete-case, MAR and MNAR-sensitive results on the same estimand.
