From Design to Estimand: Stratification, Precision, and Randomization-Based Inference

PSI Belfast 2026

Diane Uschner

F. Hoffmann-La Roche, Basel

June 16th, 2026

Disclaimer

The views expressed in this presentation are my own and do not necessarily represent the views of my employer.

Overview

Stratification matters

- Randomization is the design feature that ensures the validity of the inference and makes causal claims credible in clinical trials.
- In practice, stratification is often treated as an operational detail rather than part of the inferential target.
- That separation is too simplistic: stratification affects precision, the implied estimand, and the valid reference set for inference.
- Key message: design choices and analysis choices should be aligned from the start.

Stratification

Why stratify at all?

- Important baseline covariates are often strongly prognostic for the endpoint.
- In a single realized trial, complete randomization can leave substantial chance imbalances.
- Stratification aims to control imbalance prospectively within clinically relevant subgroups.
- This is especially attractive for small and medium-sized trials, multicenter trials, and settings with planned subgroup or interim analyses.

Where does the precision gain come from?

- A useful working model is

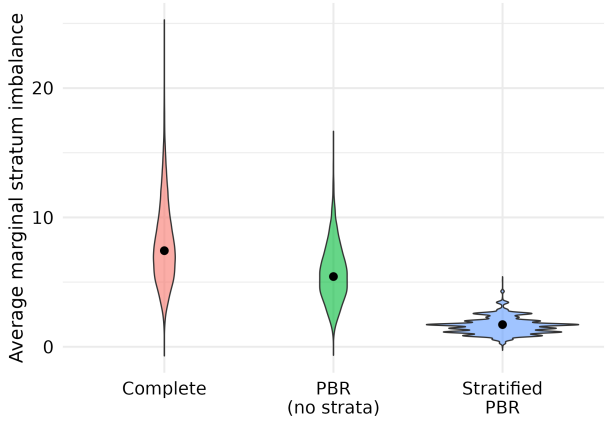
$$Y_i = \mu + \tau T_i + \beta^\top Z_i + \varepsilon_i,$$

where Z_i contains important prognostic covariates.

- Realized imbalance on important design covariates is one mechanism through which precision can be lost in a trial; see Senn, König, and Posch (2024).
- Conditional precision is best when the covariate is orthogonal to treatment assignment, in the sense emphasized by Senn (1994).
- Stratification reduces that realized association for selected covariates, and coherent covariate adjustment helps convert that balance into precision gain.
- The gain is largest when those same covariates are carried forward into the analysis.

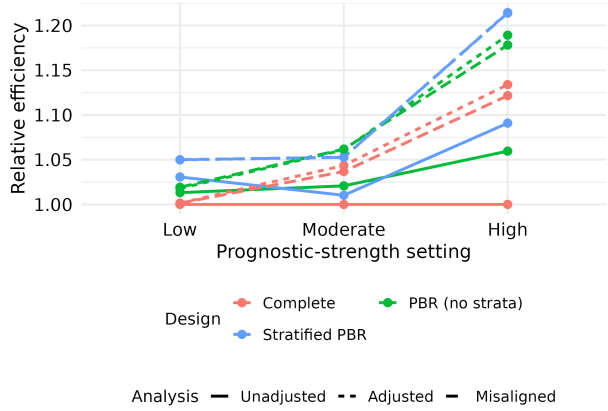
Realized marginal stratification balance

- For each level of sex, prior GLP-1 use, and region, compute the absolute treatment-count difference $|n_A - n_B|$.
- Average those absolute differences across all marginal levels to get one direct count-based balance measure per trial.
- Smaller values mean tighter realized balance on the variables that actually define the stratified randomization.



Precision depends on design-analysis alignment

- Relative efficiency compares the variance of a strategy with a baseline analysis under complete randomization.
- Values above 1 indicate greater precision than complete randomization with an unadjusted analysis.
- Low, Moderate, and High refer to the strength of the prognostic covariates in the outcome model.
- The "misaligned" analysis adjusts for sex, region, and BMI but omits prior GLP-1, so it does not fully match the stratified design.

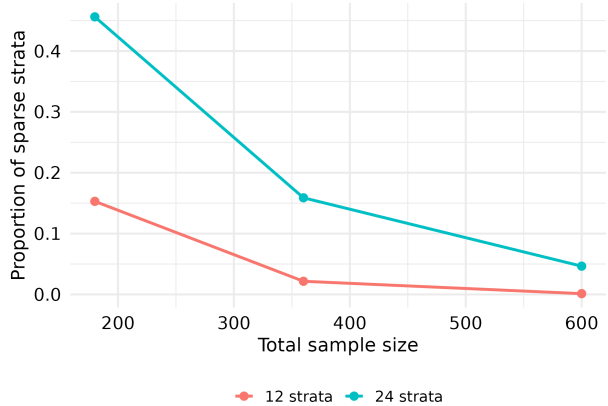


Over-stratification is a real finite-sample problem

- The number of strata grows multiplicatively with the number of factors and factor levels.
- As strata become sparse, many will end in incomplete blocks and within-stratum control weakens.
- Operational burden also rises: IRT implementation, data entry errors, monitoring, and amendment management all get harder.
- This is why guidance has long cautioned that more than a few stratification factors is rarely helpful.

Sparse-strata burden in 12 vs 24 strata

- The 12-strata design is Sex $\times$ Prior GLP-1 Use $\times$ Region; adding BMI category doubles the design to 24 strata.
- A sparse stratum has fewer than one full block at the realized sample size, including empty strata.
- Higher values mean more strata where within-stratum control is weak and incomplete blocks become common.
- The practical question is whether the extra factor is prognostic enough to justify the added complexity.



From Design to Estimand

Stratification creates a family of effects

- Let S denote the baseline stratum.
- A natural stratum-specific causal effect is

$$\theta_s = E\{Y(1) - Y(0) \mid S = s\}.$$

- Any overall treatment effect is then a weighted average of these stratum-specific effects.

$$\theta(w) = \sum_s w_s \theta_s.$$

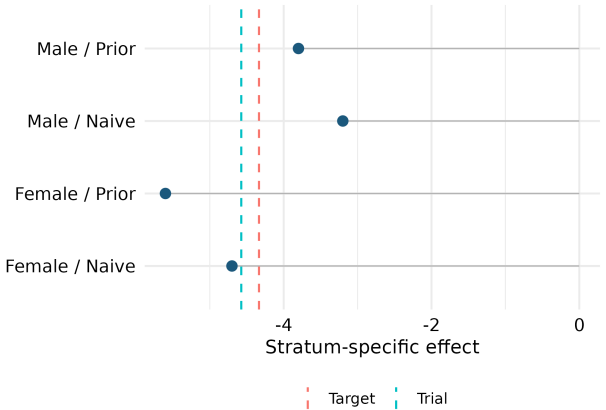
- The statistical question is therefore not only "what is the effect?" but also "for which population", or in other words, "for which weights?"

Which overall effect do we mean?

- If treatment effects are homogeneous across strata, many reasonable analyses target nearly the same quantity.
- If treatment effects differ by stratum, the choice of weights matters.
- Common options include:
  - the enrolled trial population,
  - a planned target population,
  - or a policy-relevant external population.
- A protocol that names stratification factors but does not clarify the weighting population leaves the estimand underspecified.

Weighting defines the overall effect

- The stratum-specific effects are held fixed in this placeholder calculation.
- Trial-weighted means weighting by the realized enrolled stratum mix: here 70% female and 40% prior GLP-1.
- Target-weighted means weighting by an external target-population mix: here 60% female and 30% prior GLP-1.
- For the four plotted strata, the weights shift from 18%, 42%, 12%, and 28% to 28%, 42%, 12%, and 18%.



Randomization-Based Inference

RBI starts from the actual design

- In a randomization test, the observed outcomes are held fixed.
- The null distribution comes from the randomization mechanism rather than from a "superpopulation" model.
- For a test statistic $T(R, Y)$, a design-aligned p-value is

$$p = P_{\Omega_{\text{design}}}\left(|T(R, Y)| \geq \left|T(R^{obs}, Y)\right|\right),$$

where Ω_{design} is the assignment space induced by the randomization procedure.

- The appeal is direct alignment between trial design and inferential uncertainty, as well as robustness to model misspecification.

What changes under stratified restricted randomization?

- The relevant reference set is not the set of all allocations with the same final totals.
- It must respect the actual restrictions used within strata:
  - realized stratum memberships,
  - block sizes or MTI constraints,
  - unequal allocation ratios,
  - and any conditioning implied by the design.
- Ignoring these restrictions means sampling from impossible treatment sequences.
- That can distort p-values and confidence intervals because the wrong sampling space is being used.

How the reference set is constructed

- Under separate randomization within strata, the design reference set factorizes across strata.
- Conceptually,

$$\Omega_{\text{strat}} = \prod_{s} \Omega_s,$$

with probability weights inherited from the stratum-specific randomization rules.

- Exact enumeration may be feasible for small problems.
- For realistic trials, Monte Carlo sampling from the valid reference set is often the pragmatic solution.

Takeaways

Key takeaways

- Stratification is not just a balancing device; it is part of how we define and estimate the treatment effect.
- Precision gains are most credible when important design covariates are also handled coherently in the analysis.
- Over-stratification trades theoretical balance for sparse strata, unfinished blocks, and operational cost.
- Inference after restricted randomization should use the reference set generated by the actual design, not a convenient approximation.
- The disciplined workflow is: choose prognostic covariates, define the estimand, then build inference that respects the design.