Weighting Strategies in Stratified Binary Designs

When should sample size calculations account for stratification? If stratified, which weight yields the smaller sample size, and when?

Yujie Zhao, Keaven Anderson, and Devan Mehrotra
Merck & Co., Inc., Rahway, NJ, USA

Motivation

- **Prognostic & Predictive:**

  In stratified trials, strata can differ in baseline risk (prognostic) and/or treatment effect (predictive).

- **A Classic Example:**

  A small high-risk subgroup produces many events per patient, but the treatment effect may differ from the larger, lower-risk complement.

- **Question:**

  - When should sample size calculations account for stratification?
  - If stratified, which weight yields the smaller sample size, and when?

Two Sample Size Weighting Options in gsDesign2

We provide the following two weighting options for sample size calculations in stratified designs with binary outcomes where treatment effect is measured by risk difference (i.e., `gs_design_rd` and `gs_power_rd`):

Weights	Reference	Formula	Property
Inverse Variance (INVAR)	Mantel & Haenszel (1959)	$w_s = \frac{1/\sigma_s^2}{\sum_i 1/\sigma_i^2}$	Upweights strata with lower variance
Sample Size (SS)	Mehrotra & Railkar (2000)	$w_s = \frac{N_{C,s}N_{E,s}/(N_{C,s}+N_{E,s})}{\sum_i N_{C,i}N_{E,i}/(N_{C,i}+N_{E,i})}$	Targets a population-weighted average risk difference — interpretable regardless of heterogeneity

Scenario Framework

Today's examples are two-arm clinical trials (control vs. experimental) with a binary endpoint, stratifying by two factors (small stratum:large stratum prevalence = 20%:80%) with equal randomization (1:1) across strata. The treatment effect is measured by the risk difference. The design is powered at 90%.

We evaluated 3 scenarios to assess the impact of strata heterogeneity on the design:

- **Prognostic Only:** different control rate, same risk difference
- **Predictive Only:** same control rate, different risk difference
- **Prognostic and Predictive:** different control rate, different risk difference

Prognostic Only (Different Control Rate, Same Risk Difference)

This example assumes prognostic-only effect:

- The risk difference is constant across both strata at 5%.
- The event rates in the control arm differ by stratum, i.e., large stratum:small stratum = 10%:20%.

Sample Size for Targeted 90% Power						
			Total Sample Sizes			
Two strata	Event Rate (Ctrl:Exp)	Risk Diff	INVAR	SS	Unstratified	Best Weight
Large stratum (80%)	**10.0%** vs 5.0%	**5.0%**				
Small stratum (20%)	**20.0%** vs 15.0%	**5.0%**	**1,297**	1,414	1,441	INVAR

Conclusion:

- Stratification reduces the sample size.
- INVAR is optimal.

Predictive Only (Same Ctrl Rate, Large Stratum has Larger Trt Eff)

This example shows the predictive effect of the stratification factor:

- The event rate in the control arm is the same across both strata at 30%.
- The risk difference is heterogeneous: large stratum : small stratum = 10% : 2%.
- The large stratum has bigger treatment effect.

Sample Size for Targeted 90% Power						
			Total Sample Sizes			
Two strata	Event Rate (Ctrl:Exp)	Risk Diff	INVAR	SS	Unstratified	Best Weight
Large stratum (80%)	**30.0%** vs 20.0%	**10.0%**				
Small stratum (20%)	**30.0%** vs 28.0%	**2.0%**	**1,097**	1,134	1,136	INVAR

Conclusion:

- Low advantage for stratified designs.

Predictive Only (Same Ctrl Rate, Small Stratum has Larger Trt Eff)

This example shows the predictive effect of the stratification factor:

- The event rate in the control arm is the same across both strata at 30%.
- The risk difference is heterogeneous: large stratum : small stratum = 2% : 10%.
- The large stratum has weak treatment effect.

Sample Size for Targeted 90% Power						
			Total Sample Sizes			
Two strata	Event Rate (Ctrl:Exp)	Risk Diff	INVAR	SS	Unstratified	Best Weight
Large stratum (80%)	**30.0%** vs 28.0%	**2.0%**				
Small stratum (20%)	**30.0%** vs 20.0%	**10.0%**	**6,058**	6,551	6,562	INVAR

Conclusion:

- Some sample size saving with INVAR.

Prognostic and Predictive - The SS Sweet Spot

- The large stratum has a smaller treatment effect but lower variance.
- The small stratum has a bigger treatment effect but higher variance.

Sample Size for Targeted 90% Power						
			Total Sample Sizes			
Two strata	Event Rate (Ctrl:Exp)	Risk Diff	INVAR	SS	Unstratified	Best Weight
Large stratum **(80%)**	15.0% vs 10.0%	**5.0%**				
Small stratum **(20%)**	60.0% vs 40.0%	**20.0%**	1,212	**896**	1,047	SS

Conclusion:

- INVAR is the worst; it overweights the stratum with small variance but weak treatment effect.
- SS upweights the big-treatment-effect stratum.
- Unstratified is in between of INVAR and SS.
- SS's sweet spot: there is a stratum with large treatment effect and large variance.

Prognostic and Predictive - The INVAR Sweet Spot

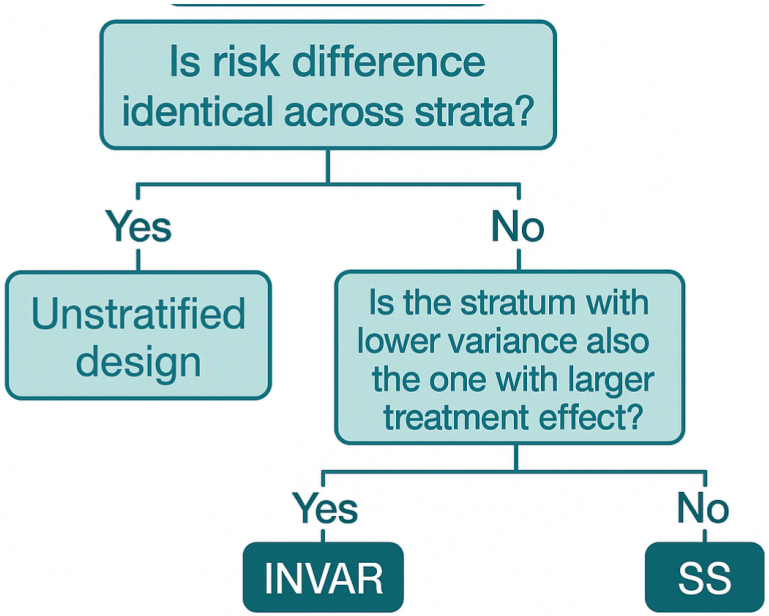
- The large stratum has a larger treatment effect and lower variance.
- The small stratum has a smaller treatment effect and higher variance.

Sample Size for Targeted 90% Power						
			Total Sample Sizes			
Two strata	Event Rate (Ctrl:Exp)	Risk Diff	INVAR	SS	Unstratified	Best Weight
Large stratum **(80%)**	15.0% vs 2.0%	**13.0%**				
Small stratum **(20%)**	60.0% vs 55.0%	**5.0%**	**240**	355	479	INVAR

Conclusion:

- INVAR upweights the stratum with large treatment effect and lower variance.
- Unstratified is worst.
- INVAR's sweet spot: there is a stratum with large treatment effect and lower variance.

When Does Each Weight Win?



Practical Recommendations

1. **Unstratified** design is adequate only when strata have same rates and same risk difference
2. **Big treatment effect in the low-variance stratum**: use INVAR weight
3. **Big treatment effect in the high-variance stratum**: use SS weight
4. In addition to INVAR and SS, adaptive weighting presents a flexible alternative.
5. **When uncertain**: Compute sample size under both INVAR and SS
6. **Beyond weighting**: When one stratum is expected to have a substantially stronger effect, consider alternative designs — e.g., testing only in the stronger subgroup, or a parametric graphical procedure that tests both the subgroup and overall population with multiplicity control

Software is Available!

```
gsDesign2::gs_design_rd(
  p_c = tibble(stratum = c("S1", "S2"), rate = c(.60, .15)),
  p_e = tibble(stratum = c("S1", "S2"), rate = c(.40, .10)),
  rd0 = 0,
  alpha = 0.025, beta = 0.2, ratio = 1,
  stratum_prev = tibble(
    stratum = c("S1", "S2"), prevalence = c(1, 4)
  ),
  weight = "ss",           # "ss", "invar", or "unstratified"
  info_scale = "h0_h1_info",
  upper = gs_b, lower = gs_b,
  upar = -qnorm(0.025), lpar = -Inf
)
```

Acknowledgement

- Keaven Anderson
- Devan Mehrotra
- Darcy Hille
- Claude Code
- Copilot

Thank you!