

Comparing Classic Crossover Trials and Aggregated N-of-1 Trials

A Methodological White Paper

2025-11-21

Contents

1	Abstract	2
2	Introduction	2
3	Background	2
3.1	Classic Crossover Trials	2
3.2	N-of-1 Trials	3
3.3	Aggregated N-of-1 Trials	3
3.4	Hybrid Designs	3
4	Core Differences Between Designs	3
4.1	Purpose and Estimands	3
4.2	Replication Structure	3
4.3	Homogeneity vs Heterogeneity	3
4.4	Carryover Effects	3
4.5	Biomarker-Treatment Interactions	4
5	Mixed-Model Specification	4
5.1	Classic Crossover Mixed Model	4
5.2	Aggregated N-of-1 Hierarchical Mixed Model	4
5.3	Hybrid Design Mixed Model	5
5.4	Why the Models Cannot Be the Same	5
6	Example: A PTSD (CAPS) Trial Designed as Either Crossover or Aggregated N-of-1	5
6.1	Clinical Assumptions	5
6.2	Option 1: Classic 2-Period Crossover Design	5
6.3	Option 2: Aggregated N-of-1 Trial	6
6.3.1	Why Fewer Participants Suffice	6
6.4	Interpretation Differences	6
6.5	Choosing Between the Designs for a PTSD Trial	7
7	Alternative Analyses for Testing Biomarker \times Treatment Interactions	7
7.1	The Core Insight: Interaction as Correlation	7
7.2	Signal-to-Noise Framework for Power	8
7.3	Implications for Study Design	8
7.4	Power Rules of Thumb	8
7.5	Pattern Recognition in Data	8
7.6	Comparison with Mixed-Effects Models	9
8	Conclusion	9
9	References	9

10 Appendix: R Code for Simulation and Analysis	11
10.1 Simulate a Classic 2-Period Crossover Dataset with Biomarker	11
10.2 Fit the Classic Crossover Mixed Model with Biomarker Interaction	12
10.3 Simulate an Aggregated N-of-1 Dataset with Biomarker and Carryover	13
10.4 Fit Aggregated N-of-1 Hierarchical Mixed Model with Biomarker Interaction	15
10.4.1 Advanced Model with Random Biomarker Interaction	16
10.5 Diagnostics and Visualization	17
10.6 Notes on Extensions and Real-Data Considerations	20
10.7 Alternative Analyses: Summary Statistic Approaches	21
10.7.1 Strategy 1: Summary Statistic Regression	21
10.7.2 Strategy 2: Correlation Test	22
10.7.3 Strategy 3: ANOVA with Biomarker Tertiles	22
10.7.4 Comparing Methods	23
10.7.5 Visualizing Interaction Patterns	23
10.8 Analytic Power Calculations	24
10.8.1 Power Function for Correlation	26
10.8.2 Computing Expected Effect Size from Parameters	26
10.8.3 Power Grid Across Parameters	28

1 Abstract

Classic crossover trials and aggregated N-of-1 trials share surface similarities, including repeated administration of treatments within individuals and frequent use of mixed-effects models. However, despite this overlap, the designs differ fundamentally in structure, inferential goals, estimands, replication, and the appropriate specification of linear mixed-effects models. This white paper expands on the conceptual and statistical distinctions between classic crossover designs and aggregated N-of-1 designs, includes narrative citations throughout, and provides a worked example involving a hypothetical PTSD trial using the Clinician-Administered PTSD Scale (CAPS) to illustrate how either design might be implemented.

2 Introduction

Repeated-treatment clinical trial designs reduce between-person variability by allowing each participant to serve as their own control. The best-known version of this approach is the classic crossover trial, widely used in chronic and stable conditions (Jones & Kenward, 2014). In contrast, N-of-1 trials, especially aggregated N-of-1 trials, are designed to evaluate individual-level responses and heterogeneity (Lillie et al., 2011; Guyatt et al., 2000). Despite frequent use of mixed-effects models in both designs, the correct statistical model depends on the inferential target and replication structure (Senn, 2002).

This white paper expands on how these two designs differ and why mixed-model analyses cannot be identical, even though they share an analytical framework. It concludes with an applied example involving a PTSD treatment trial using CAPS scores.

3 Background

3.1 Classic Crossover Trials

Classic crossover trials are structured in a small number of periods (typically 2–4), with treatment sequences such as AB/BA or ABBA (Araujo & Julious, 2014). Their primary purpose is to estimate a population average treatment effect while controlling for period and sequence effects (Mills et al., 2009). These designs assume relative stability over time, limited carryover, and generally homogeneous treatment effects across participants (Senn, 2002).

3.2 N-of-1 Trials

N-of-1 trials constitute repeated, randomized, within-person experiments using alternating treatment periods (Guyatt et al., 2000). They provide dense, within-person data suitable for estimating individual treatment effects, often with multiple AB pairs (e.g., ABABAB).

3.3 Aggregated N-of-1 Trials

Aggregated N-of-1 studies combine individual trials through hierarchical modeling to estimate both person-specific and population-level effects (Zucker et al., 2010; Punja et al., 2016). They explicitly target treatment-effect heterogeneity and leverage partial pooling (Deaton & Cartwright, 2018).

3.4 Hybrid Designs

Between classic crossover and full N-of-1 designs lies a spectrum of hybrid approaches. These designs extend crossover trials to 3–4 periods, allowing some within-person replication while maintaining feasibility. For example, a 4-path randomization (ABAB, ABBA, BAAB, BABA) provides two treatment contrasts per person—insufficient for robust individual effect estimation but enough to partially estimate heterogeneity (Senn, 2002). Hybrid designs represent a practical compromise when full N-of-1 density is infeasible but pure crossover replication is inadequate.

4 Core Differences Between Designs

4.1 Purpose and Estimands

The essential difference lies in the estimands, not the analytic framework.

- **Classic crossover:** population mean treatment effect (Mills et al., 2009).
- **Aggregated N-of-1:** individual treatment effects plus heterogeneity (Lillie et al., 2011).

4.2 Replication Structure

Replication determines what can be estimated.

- Crossover trials provide limited within-person replication—often only one treatment contrast—making random slopes for treatment unidentifiable (Brown, 1980).
- Aggregated N-of-1 trials provide substantial within-person replication, enabling estimation of random slopes and heterogeneity.

4.3 Homogeneity vs Heterogeneity

Crossover designs traditionally assume minimal heterogeneity (Senn, 2002). Aggregated N-of-1 designs assume and model heterogeneity directly (Punja et al., 2016).

4.4 Carryover Effects

Carryover—the persistence of treatment effects into subsequent periods—is handled fundamentally differently across designs:

- **Classic crossover:** Carryover is typically tested and assumed negligible. If present, it confounds period and treatment effects, potentially invalidating the design. Washout periods are employed to eliminate carryover, and statistical tests for carryover are often underpowered (Dwan et al., 2019; Senn, 2002).
- **Aggregated N-of-1:** Multiple treatment cycles enable explicit modeling of carryover as a parameter. Carryover can be incorporated as a fixed effect based on prior-period treatment, with magnitude

determined by pharmacokinetic half-life. This transforms carryover from a nuisance to an estimable quantity.

- **Hybrid designs:** With 3–4 periods, some carryover estimation becomes possible, though less robust than in full N-of-1 designs.

The key distinction: crossover designs require carryover to be absent, while aggregated N-of-1 designs can accommodate and estimate it.

4.5 Biomarker-Treatment Interactions

A critical extension beyond average treatment effects is the estimation of treatment effect modification by baseline or time-varying biomarkers. This addresses the personalized medicine question: “For whom does this treatment work best?”

- **Classic crossover:** Can include baseline biomarker \times treatment interactions as fixed effects. However, with limited within-person replication, only population-level effect modification is estimable. The model answers: “Do patients with high biomarker values respond differently on average?”
- **Aggregated N-of-1:** Dense within-person data enables estimation of both:
 1. Population-level biomarker \times treatment interactions
 2. Individual-specific biomarker \times treatment slopes (random slopes)

This allows the model to answer: “How does this specific patient’s biomarker value predict their treatment response?” Time-varying biomarkers can be incorporated, linking within-person biomarker fluctuations to within-person treatment response variation.

- **Hybrid designs:** Can estimate population-level biomarker \times treatment interactions with some precision, but individual-specific interaction terms remain difficult to identify.

The inclusion of biomarker interactions shifts the estimand from “Does treatment work?” to “For whom and under what conditions does treatment work?”—a fundamentally different scientific question that aggregated N-of-1 designs are uniquely positioned to answer.

5 Mixed-Model Specification

5.1 Classic Crossover Mixed Model

A typical model (where i indexes participants, t indexes periods):

$$Y_{it} = \beta_0 + \beta_{treat} \times Treat_{it} + \beta_{period} \times Period_t + \beta_{bm} \times BM_i + \beta_{int} \times (Treat_{it} \times BM_i) + u_i + \varepsilon_{it}$$

- Random intercept only (u_i)
- Treatment effect treated as fixed (Araujo & Julious, 2014)
- Biomarker \times treatment interaction estimable at population level
- Random slopes for treatment generally not identifiable due to insufficient replication (Brown, 1980)

5.2 Aggregated N-of-1 Hierarchical Mixed Model

$$Y_{it} = \beta_0 + (\beta_{treat} + u_{i,treat}) \times Treat_{it} + \beta_{period} \times Period_t + \beta_{bm} \times BM_{it} + (\beta_{int} + u_{i,int}) \times (Treat_{it} \times BM_{it}) + u_i + \varepsilon_{it}$$

- Random intercept (u_i), random slope for treatment ($u_{i,treat}$), and optionally random slope for biomarker interaction ($u_{i,int}$)
- Individual treatment effect = $\beta_{treat} + u_{i,treat}$
- Individual biomarker interaction = $\beta_{int} + u_{i,int}$
- Heterogeneity quantified via $Var(u_{i,treat})$ and $Var(u_{i,int})$
- Time-varying biomarkers (BM_{it}) can link within-person biomarker changes to within-person treatment response

5.3 Hybrid Design Mixed Model

$$Y_{it} = \beta_0 + \beta_{treat} \times Treat_{it} + \beta_{period} \times Period_t + \beta_{bm} \times BM_i + \beta_{int} \times (Treat_{it} \times BM_i) + \beta_{carry} \times Carry_{it} + u_i + \varepsilon_{it}$$

- Random intercept with fixed effects for treatment, biomarker interaction, and carryover
- With 3–4 periods, limited random slope estimation may be attempted but often yields singular fits
- Carryover explicitly modeled as fixed effect based on prior-period treatment

5.4 Why the Models Cannot Be the Same

Even though both use LMMs, the models differ because of:

- Different estimands
- Different random-effect structures
- Different identifiability conditions
- Different inferential goals (Senn, 2002; Zucker et al., 2010)

6 Example: A PTSD (CAPS) Trial Designed as Either Crossover or Aggregated N-of-1

To illustrate the design differences, consider a hypothetical new fast-acting pharmacologic agent for reducing PTSD symptoms, measured using CAPS-5.

6.1 Clinical Assumptions

- Onset: 24–48 hours
- Washout: 48–72 hours
- Outcome measurement: daily CAPS or EMA-based PTSD symptom indices
- Condition: chronic, variable symptoms, suitable for repeated within-person comparisons

6.2 Option 1: Classic 2-Period Crossover Design

Design:

- Sequence AB or BA
- Treatment periods: 4 weeks each
- Washout: 1 week
- N = approximately 40 participants

Strengths:

- Well-aligned with regulatory expectations
- Simple interpretation of average treatment effect

Limitations:

- Only one treatment contrast per person
- Cannot estimate individual treatment effects
- Less suitable if heterogeneity is clinically important (Dwan et al., 2019)

Mixed Model:

- Random intercept only
- Fixed treatment effect
- Period and sequence as fixed effects

6.3 Option 2: Aggregated N-of-1 Trial

Design:

- Each participant undergoes 6 cycles of active vs placebo: ABABAB (or BAABAB, individualized randomization)
- Treatment periods: 1 week active, 1 week placebo
- Washout: same-week washout built in due to rapid clearance
- CAPS measured daily using EMA or twice weekly in clinic
- N = 20 participants (fewer needed due to dense data)

6.3.1 Why Fewer Participants Suffice

The N-of-1 design achieves adequate power with fewer participants because:

1. **Increased effective sample size:** Each participant contributes multiple treatment contrasts (6 cycles = 6 contrasts vs 1 contrast in crossover). With 20 participants \times 6 cycles, the effective number of treatment comparisons approaches 120.
2. **Reduced within-person variance:** Dense repeated measurements within periods average out measurement error and day-to-day fluctuations.
3. **Partial pooling:** Hierarchical models borrow strength across participants. Participants with fewer observations or more noise are shrunk toward the population mean, improving precision for both individual and population estimates (Zucker et al., 2010).
4. **Direct estimation target:** If the goal is individual treatment effects (not just population average), N-of-1 provides the data structure required—crossover designs cannot achieve this regardless of sample size.

Strengths:

- Estimates individual treatment effects
- Captures treatment-effect heterogeneity
- Estimates biomarker \times treatment interactions at individual level
- Can explicitly model carryover effects
- Partially pooled model improves precision (Lillie et al., 2011; Zucker et al., 2010)

Limitations:

- Higher participant burden
- Requires stable condition and rapid washout
- More complex data management

Mixed Model:

- Random intercept and random slope for treatment
- Biomarker \times treatment interaction (fixed or random)
- Carryover effect (fixed)
- Individual effects reported alongside pooled effect
- Heterogeneity explicitly estimated

6.4 Interpretation Differences

Question	Crossover	Hybrid	Aggregated N-of-1
Does the drug work on average?	Yes	Yes	Yes
Does the drug work for this individual?	No	Limited	Yes

Question	Crossover	Hybrid	Aggregated N-of-1
Is treatment effect heterogeneous?	Rarely estimable	Partially	Yes
Does biomarker predict response (population)?	Yes	Yes	Yes
Does biomarker predict response (individual)?	No	No	Yes
Can carryover be estimated?	No (assumed absent)	Partially	Yes
Does each person have replicated comparisons?	No	Limited (2-3)	Yes (6+)

6.5 Choosing Between the Designs for a PTSD Trial

- If the goal is regulatory approval or demonstration of a population-average effect → Crossover (Dwan et al., 2019).
- If the goal is to understand who responds, optimize personalized treatment, or examine heterogeneity → Aggregated N-of-1 (Lillie et al., 2011; Punja et al., 2016).
- If measurement is frequent and the drug acts quickly → N-of-1 is advantageous.
- If measurement is sporadic or drug has slow dynamics → Crossover preferred.

7 Alternative Analyses for Testing Biomarker \times Treatment Interactions

While mixed-effects models provide the most statistically efficient analysis of biomarker \times treatment interactions, they can obscure intuition about what drives statistical power. Alternative approaches based on summary statistics offer simpler, more interpretable heuristics for understanding when interactions are detectable and what data patterns indicate their presence.

7.1 The Core Insight: Interaction as Correlation

Testing a biomarker \times treatment interaction is mathematically equivalent to asking: **Does the individual treatment effect correlate with the biomarker?**

For each subject i , compute the treatment effect as the difference in mean outcomes:

$$\Delta_i = \bar{Y}_{i,active} - \bar{Y}_{i,placebo}$$

The interaction test then reduces to asking whether Δ_i varies systematically with BM_i . This can be assessed through:

1. **Pearson correlation** between Δ_i and BM_i
2. **Simple linear regression** of Δ_i on BM_i
3. **ANOVA** comparing mean Δ_i across biomarker groups

These approaches yield nearly equivalent p-values but provide different effect size metrics that aid interpretation.

7.2 Signal-to-Noise Framework for Power

Power to detect the interaction depends on a simple signal-to-noise ratio:

$$\text{SNR} = \frac{|\beta_{int}| \cdot \sigma_{BM}}{\sigma_{\Delta}}$$

This ratio equals the expected correlation r between biomarker and treatment response. The components are:

- **Signal** ($|\beta_{int}| \cdot \sigma_{BM}$): The interaction magnitude times biomarker spread. Larger interactions and greater biomarker variability increase the signal.
- **Noise** (σ_{Δ}): The standard deviation of individual treatment effects. This includes:
 - Measurement error: $\sqrt{2\sigma_{\varepsilon}^2/n_{obs}}$ where n_{obs} is observations per subject per treatment
 - Treatment effect heterogeneity: σ_{treat} (individual differences not explained by biomarker)

7.3 Implications for Study Design

This framework reveals why aggregated N-of-1 designs have superior power for detecting interactions:

1. **Dense measurements reduce σ_{Δ}** : With many observations per subject, the treatment effect estimate has lower variance. In the N-of-1 simulation (21 observations per treatment), measurement error contributes minimally to σ_{Δ} .
2. **Unexplained heterogeneity is the enemy**: Individual differences in treatment response that are *not* predicted by the biomarker (σ_{treat}) increase noise without adding signal. This heterogeneity cannot be reduced by more observations—it requires either larger samples or better predictors.
3. **Biomarker spread provides leverage**: Homogeneous biomarker values yield no ability to detect interactions regardless of sample size. Consider enrichment designs that ensure variability in the biomarker of interest.

7.4 Power Rules of Thumb

Using the correlation framework, standard power calculations apply:

Expected r	Required N (80% power)	Interpretation
0.10	>750	Rarely practical
0.20	~200	Large study needed
0.30	~85	Moderate study
0.40	~45	Feasible
0.50	~30	Small study sufficient

For a given design, compute the expected r from model parameters to determine required sample size.

7.5 Pattern Recognition in Data

Detectable interactions manifest as:

- **Non-parallel lines** in treatment \times biomarker interaction plots
- **Significant correlation** between biomarker and per-subject treatment effect
- **Monotonic trend** in mean treatment effect across biomarker tertiles
- **Positive R^2** when regressing treatment effect on biomarker

The summary statistic approach makes these patterns directly visible and interpretable, complementing the more complex mixed-model output.

7.6 Comparison with Mixed-Effects Models

The summary statistic approach:

- **Advantages:** Simpler computation, transparent power calculations, direct visualization of interaction patterns
- **Disadvantages:** Ignores within-subject correlation structure, potentially less efficient

In practice, the two approaches often yield similar power, particularly when within-subject correlations are modest. The summary statistic approach serves as a valuable complement for building intuition and planning studies, while mixed models remain preferred for final inference.

8 Conclusion

Although classic crossover and aggregated N-of-1 designs can both be analyzed using mixed-effects models, the correct models differ fundamentally. Crossover trials typically feature random-intercept models targeting a population-level treatment effect with minimal within-person replication. Aggregated N-of-1 designs feature hierarchical mixed models with random slopes for treatment, enabling estimation of individual effects and treatment heterogeneity. Hybrid designs occupy a middle ground, extending crossover to 3–4 periods for some heterogeneity estimation.

A critical extension is biomarker \times treatment interactions. All three designs can estimate population-level effect modification (does biomarker predict response on average?), but only aggregated N-of-1 designs provide sufficient within-person replication to estimate individual-specific biomarker interactions. This shifts the estimand from “Does treatment work?” to “For whom and under what conditions does treatment work?”—the fundamental question of personalized medicine.

Similarly, carryover effects—typically a nuisance to be eliminated in crossover designs—can be explicitly modeled and estimated in aggregated N-of-1 designs due to multiple treatment cycles.

The choice of design should be guided by:

- **Scientific aims:** Population average vs individual-level inference
- **Estimand:** Average treatment effect vs biomarker-stratified effects vs individual effects
- **Pharmacologic properties:** Onset, washout, and carryover characteristics
- **Feasibility:** Participant burden, measurement frequency, study duration

When the goal is regulatory approval of a population-average effect, crossover designs remain appropriate. When the goal is precision medicine—identifying who responds and why—aggregated N-of-1 designs are uniquely positioned to answer these questions.

9 References

- Araujo, A., & Julious, S. A. (2014). Understanding the assumptions of crossover trials. *Pharmaceutical Statistics*, 13(6), 341–350.
- Brown, H. (1980). The analysis of variance and covariance in crossover trials. *Biometrics*, 36(1), 69–79.
- Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21.
- Dwan, K., Li, T., Altman, D. G., & Elbourne, D. (2019). CONSORT 2010 statement: extension to randomised crossover trials. *BMJ*, 366, 14378.
- Guyatt, G. H., et al. (2000). The N-of-1 randomized controlled trial: Clinical usefulness. *Annals of Internal Medicine*, 112(4), 293–299.
- Mills, E. J., Chan, A. W., Wu, P., Vail, A., Guyatt, G. H., & Altman, D. G. (2009). Design, analysis, and presentation of crossover trials. *Trials*, 10, 27.
- Lillie, E. O., et al. (2011). The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine*, 8(2), 161–173.

- Punja, S., et al. (2016). N-of-1 trials for precision medicine: A systematic review. *Journal of Clinical Epidemiology*, 76, 1–13.
- Schmid, C. H., et al. (2013). Effect of statin therapy on muscle symptoms: An individual patient data meta-analysis. *JAMA Internal Medicine*, 173(16), 1–9.
- Senn, S. (2002). *Cross-over Trials in Clinical Research* (2nd ed.). Wiley.
- Zucker, D. R., Ruthazer, R., & Schmid, C. H. (2010). Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: Methodologic considerations. *Journal of Clinical Epidemiology*, 63(12), 1312–1323.

10 Appendix: R Code for Simulation and Analysis

Below are example R code chunks to simulate datasets for a classic crossover and an aggregated N-of-1 design, and to fit appropriate mixed-effects models. The code uses `lme4` for frequentist mixed models, `nlme` for correlation structures if desired, and `brms` for a Bayesian hierarchical alternative.

```
library(tidyverse)
library(lme4)
library(lmerTest) # Provides p-values for lmer models
library(nlme)
library(emmeans)
# library(brms) # Uncomment if using Bayesian models
```

10.1 Simulate a Classic 2-Period Crossover Dataset with Biomarker

```
set.seed(2025)

# Parameters
n_subj <- 80
periods <- 2

# True effects
mu <- 20 # baseline mean CAPS score
beta_treat <- -4 # average treatment effect (reduction)
beta_bm <- 2 # biomarker main effect
beta_int <- -1.5 # biomarker x treatment interaction
beta_period <- 0.5 # period effect
sigma_subj <- 6 # between-subject SD
sigma_resid <- 8 # residual SD

# Create subject-level intercepts and baseline biomarker
subj_df <- tibble(
  subject = 1:n_subj,
  u = rnorm(n_subj, 0, sigma_subj),
  bm = rnorm(n_subj, 0, 1)
)

# Assign half to sequence AB and half to BA
subj_df <- subj_df %>%
  mutate(sequence = rep(c("AB", "BA"), length.out = n_subj))

# Expand to periods
crossover <- subj_df %>%
  group_by(subject) %>%
  reframe(period = 1:periods, u = u, bm = bm, sequence = sequence) %>%
  mutate(
    treatment = case_when(
      sequence == "AB" & period == 1 ~ "A",
      sequence == "AB" & period == 2 ~ "B",
      sequence == "BA" & period == 1 ~ "B",
      sequence == "BA" & period == 2 ~ "A"
    ),
    trt_indicator = if_else(treatment == "B", 1, 0)
  )
```

```

# Simulate outcome with biomarker interaction
crossover <- crossover %>%
  mutate(
    Y = mu + u +
      beta_period * period +
      beta_treat * trt_indicator +
      beta_bm * bm +
      beta_int * trt_indicator * bm +
      rnorm(n(), 0, sigma_resid)
  )

# Quick glance
head(crossover)
#> # A tibble: 6 x 8
#>   subject period      u      bm sequence treatment trt_indicator      Y
#>   <int>   <int> <dbl> <dbl> <chr>      <chr>          <dbl> <dbl>
#> 1       1     1  3.72 -1.12 AB        A              0  24.2
#> 2       1     2  3.72 -1.12 AB        B              1  35.4
#> 3       2     1  0.214  1.47 BA        B              1  16.7
#> 4       2     2  0.214  1.47 BA        A              0  23.4
#> 5       3     1  4.64  0.205 AB        A              0  18.6
#> 6       3     2  4.64  0.205 AB        B              1   8.08

```

10.2 Fit the Classic Crossover Mixed Model with Biomarker Interaction

```

# lmer with random intercept and biomarker x treatment interaction
m_crossover <- lmer(
  Y ~ trt_indicator * bm + factor(period) + (1 | subject),
  data = crossover
)
summary(m_crossover)
#> Linear mixed model fit by REML. t-tests use Satterthwaite's method [
#> lmerModLmerTest]
#> Formula: Y ~ trt_indicator * bm + factor(period) + (1 | subject)
#> Data: crossover
#>
#> REML criterion at convergence: 1174.6
#>
#> Scaled residuals:
#>      Min       1Q   Median       3Q      Max
#> -2.87787 -0.58310  0.08426  0.63764  1.77324
#>
#> Random effects:
#> Groups Name          Variance Std.Dev.
#> subject (Intercept) 34.53     5.876
#> Residual              71.07     8.431
#> Number of obs: 160, groups: subject, 80
#>
#> Fixed effects:
#>              Estimate Std. Error      df t value Pr(>|t|)
#> (Intercept)    20.321     1.360 154.906  14.940  <2e-16 ***
#> trt_indicator   -3.328     1.358  77.000  -2.451   0.0165 *

```

```

#> bm                2.493        1.229 141.393    2.028    0.0444 *
#> factor(period)2    1.378        1.348  77.000    1.022    0.3101
#> trt_indicator:bm   -3.757        1.437  77.000   -2.615    0.0107 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>          (Intr) trt_nd bm      fct()2
#> trt_indictr -0.513
#> bm          0.207 -0.112
#> factr(prd)2 -0.510  0.029 -0.088
#> trt_ndctr:b -0.170  0.191 -0.585  0.151

# Estimated average treatment effect (at mean biomarker = 0)
fixef(m_crossover)["trt_indicator"]
#> trt_indicator
#>      -3.328289

# Estimated biomarker x treatment interaction (population-level)
fixef(m_crossover)["trt_indicator:bm"]
#> trt_indicator:bm
#>      -3.75727

# Treatment effect at different biomarker levels
emmeans(m_crossover, ~ trt_indicator | bm, at = list(bm = c(-1, 0, 1)))
#> bm = -1:
#>   trt_indicator emmean   SE  df lower.CL upper.CL
#>           0    18.5 1.53 141     15.5     21.5
#>           1    18.9 1.53 141     15.9     22.0
#>
#> bm = 0:
#>   trt_indicator emmean   SE  df lower.CL upper.CL
#>           0    21.0 1.17 141     18.7     23.3
#>           1    17.7 1.17 141     15.4     20.0
#>
#> bm = 1:
#>   trt_indicator emmean   SE  df lower.CL upper.CL
#>           0    23.5 1.85 141     19.8     27.2
#>           1    16.4 1.85 141     12.8     20.1
#>
#> Results are averaged over the levels of: period
#> Degrees-of-freedom method: kenward-roger
#> Confidence level used: 0.95

```

Interpretation: Can estimate population-level effect modification, but cannot estimate individual-specific interaction terms.

10.3 Simulate an Aggregated N-of-1 Dataset with Biomarker and Carryover

```

set.seed(2025)

n_subj <- 25
cycles <- 6

```

```

obs_per_period <- 7

# True effects
mu <- 20
beta_treat_pop <- -4
beta_bm <- 2
beta_int_pop <- -1.5
beta_carry <- 1.5

sigma_subj <- 6
sigma_treat_sd <- 3
sigma_int_sd <- 0.8
sigma_resid <- 5

# Create subject-level random effects
subj <- tibble(
  subject = 1:n_subj,
  u = rnorm(n_subj, 0, sigma_subj),
  u_trt = rnorm(n_subj, 0, sigma_treat_sd),
  u_int = rnorm(n_subj, 0, sigma_int_sd),
  bm_baseline = rnorm(n_subj, 0, 1)
)

# Build periods alternating A/B starting at random
nof1 <- subj %>%
  crossing(period = 1:cycles) %>%
  group_by(subject) %>%
  mutate(order_start = first(sample(c(0, 1), 1))) %>%
  ungroup() %>%
  mutate(trt_indicator = (order_start + period) %% 2)

# Expand to daily observations within each period
nof1 <- nof1 %>%
  crossing(day = 1:obs_per_period) %>%
  arrange(subject, period, day)

# Add time-varying biomarker
nof1 <- nof1 %>%
  mutate(bm = bm_baseline + rnorm(n(), 0, 0.3))

# Add carryover effect
nof1 <- nof1 %>%
  group_by(subject) %>%
  mutate(
    prior_trt = lag(trt_indicator, default = 0),
    carryover = if_else(period > 1 & day <= 2, prior_trt, 0)
  ) %>%
  ungroup()

# Simulate outcomes
nof1 <- nof1 %>%
  mutate(
    Y = mu + u +

```

```

    (beta_treat_pop + u_trt) * trt_indicator +
    beta_bm * bm +
    (beta_int_pop + u_int) * trt_indicator * bm +
    beta_carry * carryover +
    rnorm(n(), 0, sigma_resid)
  )

# Inspect
nof1 %>%
  group_by(subject) %>%
  summarise(n_obs = n(), n_periods = n_distinct(period)) %>%
  head()
#> # A tibble: 6 x 3
#>   subject n_obs n_periods
#>   <int> <int>   <int>
#> 1     1     42         6
#> 2     2     42         6
#> 3     3     42         6
#> 4     4     42         6
#> 5     5     42         6
#> 6     6     42         6

```

10.4 Fit Aggregated N-of-1 Hierarchical Mixed Model with Biomarker Interaction

```

# Random intercept + random slope for treatment + biomarker interaction + carryover
m_nof1 <- lmer(
  Y ~ trt_indicator * bm + carryover + (1 + trt_indicator | subject),
  data = nof1,
  REML = TRUE
)
summary(m_nof1)
#> Linear mixed model fit by REML. t-tests use Satterthwaite's method [
#> lmerModLmerTest]
#> Formula: Y ~ trt_indicator * bm + carryover + (1 + trt_indicator | subject)
#> Data: nof1
#>
#> REML criterion at convergence: 6448.2
#>
#> Scaled residuals:
#>    Min      1Q  Median      3Q     Max
#> -3.2778 -0.6249 -0.0146  0.6205  3.3358
#>
#> Random effects:
#> Groups Name Variance Std.Dev. Corr
#> subject (Intercept) 33.09 5.753
#> trt_indicator 10.86 3.295 0.02
#> Residual 23.89 4.887
#> Number of obs: 1050, groups: subject, 25
#>
#> Fixed effects:
#> Estimate Std. Error df t value Pr(>|t|)
#> (Intercept) 21.8463 1.1736 23.8224 18.614 1.07e-15 ***

```

```

#> trt_indicator      -3.9532      0.7293    23.6857   -5.420 1.50e-05 ***
#> bm                  1.5937      0.5273    182.1156    3.023 0.00287 **
#> carryover           0.7476      0.4672   1000.7994    1.600 0.10991
#> trt_indicator:bm    -1.2804      0.5776    50.7426   -2.217 0.03116 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Correlation of Fixed Effects:
#>          (Intr) trt_nd bm      crryvr
#> trt_indictr -0.039
#> bm           0.061 -0.042
#> carryover    -0.050  0.001 -0.046
#> trt_ndctr:b  -0.028  0.110 -0.463  0.020

# Extract fixed (population) effects
fixef(m_nof1)
#>      (Intercept)      trt_indicator          bm      carryover
#>      21.8463025      -3.9531711      1.5937459      0.7476118
#> trt_indicator:bm
#>      -1.2803543

# Population treatment effect (at mean biomarker = 0)
fixef(m_nof1)["trt_indicator"]
#> trt_indicator
#>      -3.953171

# Population biomarker x treatment interaction
fixef(m_nof1)["trt_indicator:bm"]
#> trt_indicator:bm
#>      -1.280354

# Carryover effect
fixef(m_nof1)["carryover"]
#> carryover
#> 0.7476118

# Extract subject-specific treatment effects (BLUPs)
subj_trt_effects <- ranef(m_nof1)$subject[["trt_indicator"]] +
  fixef(m_nof1)["trt_indicator"]
head(subj_trt_effects)
#> [1] -6.690817 -7.458577 -3.954958 -5.074274 -1.525257 -9.434112

```

10.4.1 Advanced Model with Random Biomarker Interaction

```

# This estimates individual-specific biomarker x treatment interactions
# May need more data or regularization for convergence
m_nof1_full <- lmer(
  Y ~ trt_indicator * bm + carryover +
    (1 + trt_indicator + trt_indicator:bm | subject),
  data = nof1,
  REML = TRUE,
  control = lmerControl(optimizer = "bobyqa")
)

```

```
summary(m_nof1_full)
```

10.5 Diagnostics and Visualization

```
# Plot individual subject treatment effects
subj_effects <- raneef(m_nof1)$subject %>%
  as.data.frame() %>%
  rownames_to_column(var = "subject") %>%
  mutate(
    subject = as.integer(subject),
    subj_trt = trt_indicator + fixef(m_nof1)["trt_indicator"]
  )

ggplot(subj_effects, aes(x = subj_trt)) +
  geom_histogram(binwidth = 0.5, fill = "steelblue", color = "white") +
  geom_vline(
    xintercept = fixef(m_nof1)["trt_indicator"],
    linetype = "dashed",
    color = "red"
  ) +
  labs(
    title = "Distribution of Individual Treatment Effects (BLUPs)",
    subtitle = "Red line = population average effect",
    x = "Individual treatment effect (CAPS change)",
    y = "Count"
  ) +
  theme_minimal()

# Plot biomarker x treatment interaction
bm_range <- seq(-2, 2, by = 0.1)
interaction_df <- tibble(
  bm = bm_range,
  trt_effect = fixef(m_nof1)["trt_indicator"] +
    fixef(m_nof1)["trt_indicator:bm"] * bm_range
)

ggplot(interaction_df, aes(x = bm, y = trt_effect)) +
  geom_line(linewidth = 1.2, color = "darkgreen") +
  geom_hline(yintercept = 0, linetype = "dotted") +
  labs(
    title = "Biomarker x Treatment Interaction",
    subtitle = "Treatment effect as function of biomarker level",
    x = "Biomarker (standardized)",
    y = "Treatment effect (CAPS change)"
  ) +
  theme_minimal()

# Observed trajectories for sample subjects
sample_subj <- sample(unique(nof1$subject), 6)

nof1 %>%
  filter(subject %in% sample_subj) %>%
  mutate(
```

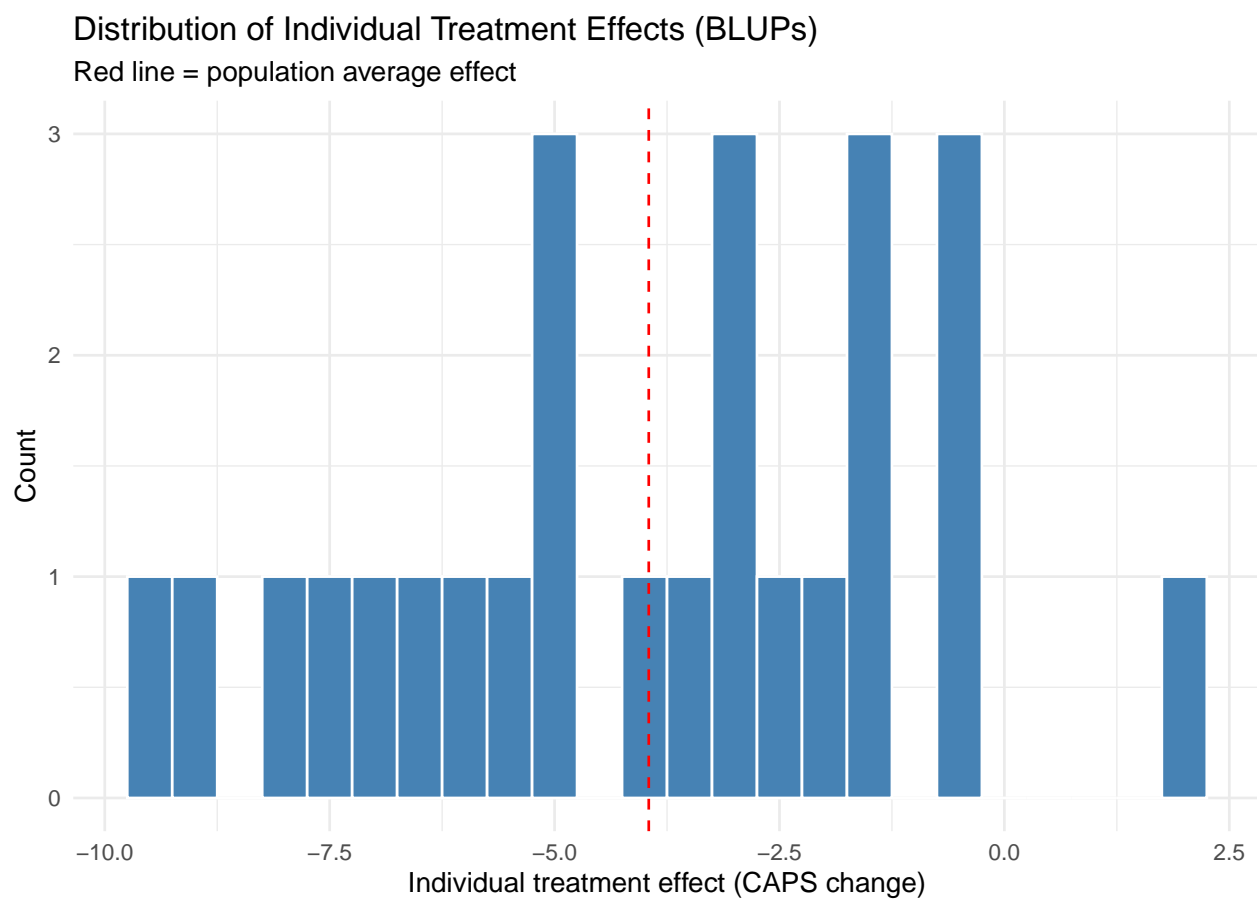


Figure 1: Distribution of individual treatment effects

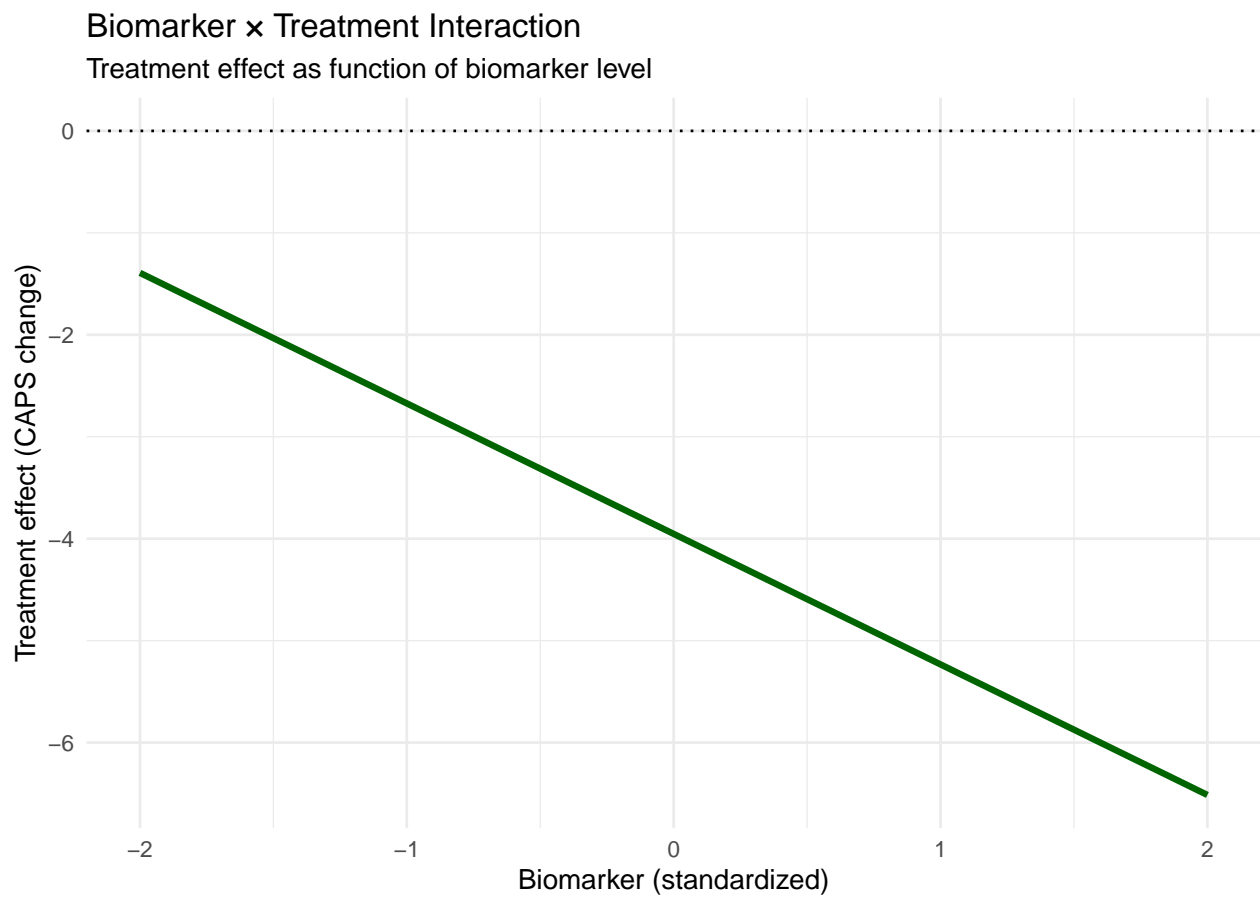


Figure 2: Biomarker by treatment interaction

```

time = (period - 1) * obs_per_period + day,
trt = if_else(trt_indicator == 1, "Active", "Placebo")
) %>%
ggplot(aes(x = time, y = Y, color = trt)) +
  geom_line(alpha = 0.7) +
  geom_point(size = 0.8) +
  facet_wrap(~subject, scales = "free_y") +
  labs(
    title = "Individual Patient Trajectories",
    x = "Day",
    y = "CAPS score",
    color = "Treatment"
  ) +
  theme_minimal()

```

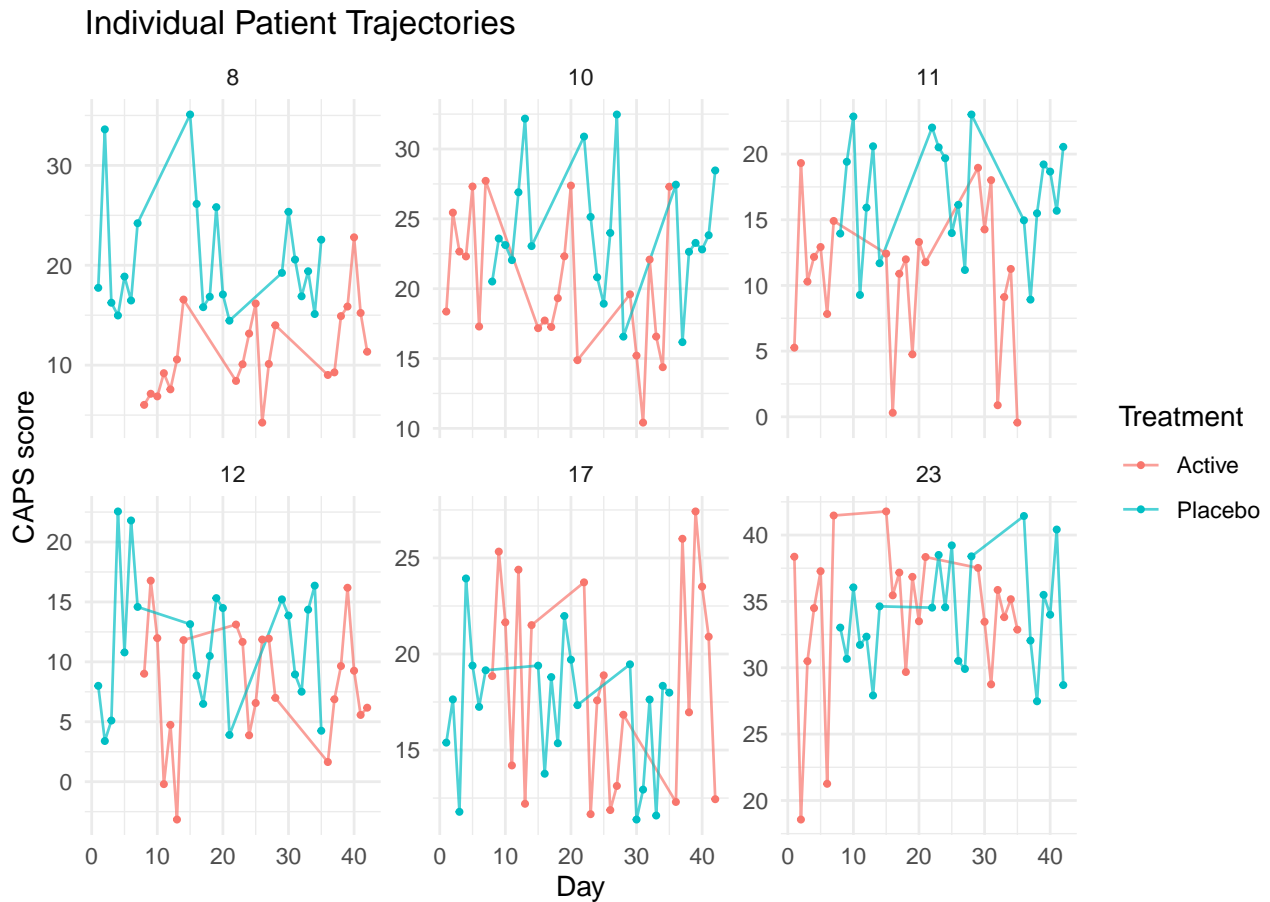


Figure 3: Individual patient trajectories

10.6 Notes on Extensions and Real-Data Considerations

- For real CAPS outcomes measured daily, expand the simulation to include within-period autocorrelation (e.g., AR(1)), measurement error, and missingness. Use `nlme::lme` or specify correlation structures for autocorrelation.
- If carryover is suspected, consider adding period-by-treatment interaction terms or explicitly modeling carryover terms; tests for carryover should be planned but interpreted cautiously (Dwan et al., 2019).

- In the aggregated N-of-1 context, Bayesian hierarchical models (e.g., `brms`) allow flexible priors and full posterior inference for individual effects (Zucker et al., 2010).
- For regulatory-facing analyses, prespecify the primary estimand (population mean vs individual responder analysis), multiplicity handling, and sensitivity analyses for missing data and carryover.

10.7 Alternative Analyses: Summary Statistic Approaches

These methods complement the mixed-model analysis by providing interpretable heuristics for understanding interaction patterns.

10.7.1 Strategy 1: Summary Statistic Regression

```
# Compute per-subject treatment effects from the N-of-1 data
subj_summary <- nof1 %>%
  group_by(subject, trt_indicator) %>%
  summarise(mean_Y = mean(Y), .groups = "drop") %>%
  pivot_wider(names_from = trt_indicator, values_from = mean_Y,
              names_prefix = "trt_") %>%
  mutate(delta = trt_1 - trt_0) # Active - Placebo

# Merge with baseline biomarker
subj_summary <- nof1 %>%
  distinct(subject, bm_baseline) %>%
  right_join(subj_summary, by = "subject")

# Simple linear regression of treatment effect on biomarker
reg_summary <- lm(delta ~ bm_baseline, data = subj_summary)
summary(reg_summary)

#>
#> Call:
#> lm(formula = delta ~ bm_baseline, data = subj_summary)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -6.513 -2.442  0.031  2.983  6.571
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  -3.9438      0.7379  -5.345 1.99e-05 ***
#> bm_baseline  -1.0667      0.7271  -1.467  0.156
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.652 on 23 degrees of freedom
#> Multiple R-squared:  0.08557,    Adjusted R-squared:  0.04581
#> F-statistic: 2.152 on 1 and 23 DF,  p-value: 0.1559

# The slope directly estimates the interaction
coef(reg_summary)["bm_baseline"]
#> bm_baseline
#> -1.066669
```

10.7.2 Strategy 2: Correlation Test

```
# Pearson correlation between biomarker and treatment effect
cor_result <- cor.test(subj_summary$bm_baseline, subj_summary$delta)
cor_result
#>
#> Pearson's product-moment correlation
#>
#> data: subj_summary$bm_baseline and subj_summary$delta
#> t = -1.4671, df = 23, p-value = 0.1559
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#> -0.6164098 0.1160127
#> sample estimates:
#> cor
#> -0.2925278

# Effect size: r 0.1 small, r 0.3 medium, r 0.5 large
```

10.7.3 Strategy 3: ANOVA with Biomarker Tertiles

```
# Discretize biomarker into tertiles
subj_summary <- subj_summary %>%
  mutate(
    bm_tertile = cut(
      bm_baseline,
      breaks = quantile(bm_baseline, c(0, 1/3, 2/3, 1)),
      labels = c("Low", "Medium", "High"),
      include.lowest = TRUE
    )
  )

# One-way ANOVA
anova_result <- anova(lm(delta ~ bm_tertile, data = subj_summary))
anova_result
#> Analysis of Variance Table
#>
#> Response: delta
#>          Df Sum Sq Mean Sq F value Pr(>F)
#> bm_tertile 2  48.53  24.265   1.8605 0.1793
#> Residuals 22 286.94  13.043

# Effect size: eta-squared
eta_sq <- anova_result["bm_tertile", "Sum Sq"] / sum(anova_result[, "Sum Sq"])
cat("Eta-squared:", round(eta_sq, 3), "\n")
#> Eta-squared: 0.145

# Cell means
subj_summary %>%
  group_by(bm_tertile) %>%
  summarise(
    n = n(),
    mean_delta = mean(delta),
```

```

sd_delta = sd(delta),
.groups = "drop"
)
#> # A tibble: 3 x 4
#>   bm_tertile      n mean_delta sd_delta
#>   <fct>      <int>      <dbl>   <dbl>
#> 1 Low           9      -1.97     4.27
#> 2 Medium        8      -5.17     2.69
#> 3 High          8      -4.46     3.59

```

10.7.4 Comparing Methods

```

# Extract p-value from lmer model (requires lmerTest)
lmer_coefs <- coef(summary(m_nof1))
lmer_pval <- lmer_coefs["trt_indicator:bm", "Pr(>|t|)"]

# Compile results
comparison <- tibble(
  Method = c("Mixed Model", "Summary Regression", "Correlation", "ANOVA"),
  Estimate = c(
    fixef(m_nof1)["trt_indicator:bm"],
    coef(reg_summary)["bm_baseline"],
    cor_result$estimate,
    eta_sq
  ),
  P_value = c(
    lmer_pval,
    summary(reg_summary)$coefficients["bm_baseline", "Pr(>|t|)"],
    cor_result$p.value,
    anova_result["bm_tertile", "Pr(>F)"]
  )
)

comparison %>%
  mutate(Estimate = round(Estimate, 3),
         P_value = format.pval(P_value, digits = 3))
#> # A tibble: 4 x 3
#>   Method      Estimate P_value
#>   <chr>      <dbl> <chr>
#> 1 Mixed Model    -1.28 0.0312
#> 2 Summary Regression -1.07 0.1559
#> 3 Correlation   -0.293 0.1559
#> 4 ANOVA         0.145 0.1793

```

10.7.5 Visualizing Interaction Patterns

```

ggplot(subj_summary, aes(x = bm_tertile, y = delta, fill = bm_tertile)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  scale_fill_brewer(palette = "Blues") +
  labs(

```

```

title = "Treatment Effect by Biomarker Group",
subtitle = "Trend indicates biomarker × treatment interaction",
x = "Biomarker Tertile",
y = "Individual Treatment Effect (Δ)"
) +
theme_minimal() +
theme(legend.position = "none")

```

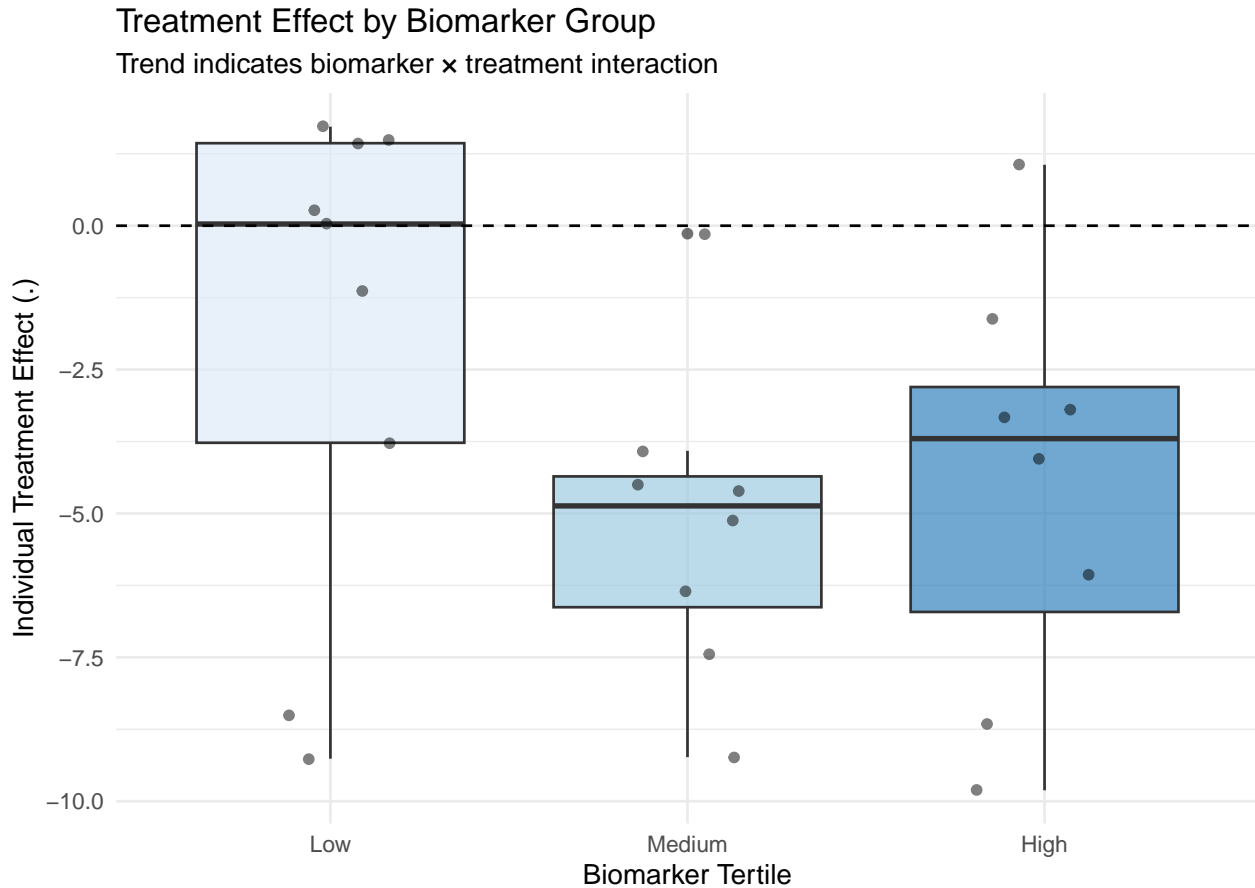


Figure 4: Treatment effect by biomarker group

```

ggplot(subj_summary, aes(x = bm_baseline, y = delta)) +
  geom_point(alpha = 0.7, size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "darkblue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(
    title = "Biomarker vs Individual Treatment Effect",
    subtitle = paste("r =", round(cor_result$estimate, 2)),
    x = "Baseline Biomarker",
    y = "Individual Treatment Effect (Δ)"
  ) +
  theme_minimal()

```

10.8 Analytic Power Calculations

Closed-form power formulas help plan studies and understand parameter impacts.

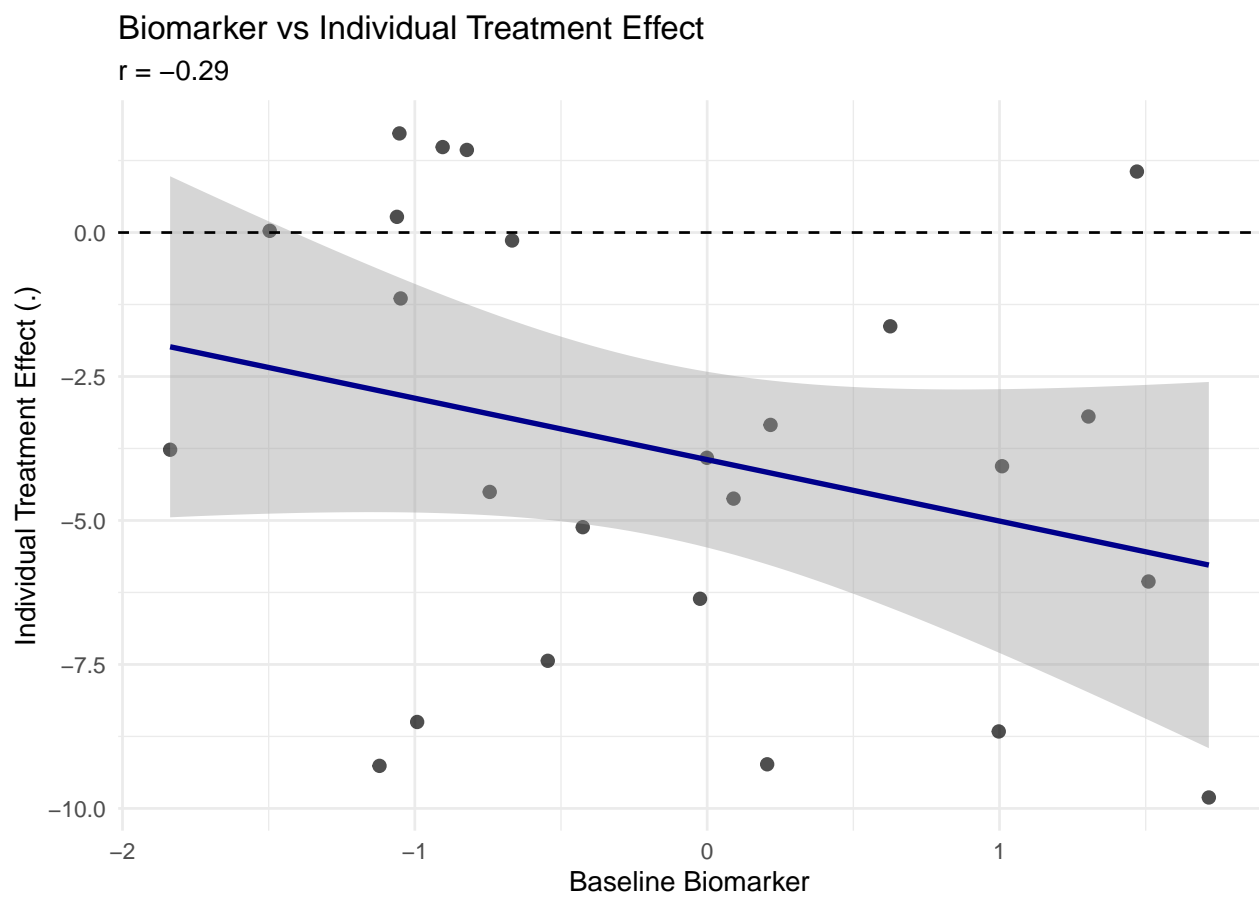


Figure 5: Biomarker vs treatment effect

10.8.1 Power Function for Correlation

```
power_correlation <- function(r, n, alpha = 0.05) {
  ncp <- abs(r) * sqrt(n - 2) / sqrt(1 - r^2)
  t_crit <- qt(1 - alpha/2, n - 2)
  power <- pt(t_crit, n - 2, ncp, lower.tail = FALSE) +
    pt(-t_crit, n - 2, ncp, lower.tail = TRUE)
  return(power)
}

# Power table
power_table <- expand_grid(
  r = c(0.1, 0.2, 0.3, 0.4, 0.5),
  n = c(20, 40, 70, 100, 150)
) %>%
  mutate(power = map2_dbl(r, n, power_correlation))

power_table %>%
  pivot_wider(names_from = n, values_from = power, names_prefix = "N=") %>%
  mutate(across(starts_with("N="), ~round(., 2)))

#> # A tibble: 5 x 6
#>       r `N=20` `N=40` `N=70` `N=100` `N=150`
#>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1  0.1  0.07  0.09  0.13  0.17  0.23
#> 2  0.2  0.13  0.23  0.38  0.52  0.69
#> 3  0.3  0.24  0.47  0.72  0.87  0.97
#> 4  0.4  0.42  0.75  0.94  0.99  1
#> 5  0.5  0.64  0.93  1      1      1

ggplot(power_table, aes(x = n, y = power, color = factor(r))) +
  geom_line(linewidth = 1) +
  geom_point() +
  geom_hline(yintercept = 0.8, linetype = "dashed", alpha = 0.5) +
  scale_color_viridis_d(option = "plasma", end = 0.8) +
  labs(
    title = "Power to Detect Biomarker × Treatment Interaction",
    x = "Number of Subjects",
    y = "Power",
    color = "Effect Size (r)"
  ) +
  theme_minimal()
```

10.8.2 Computing Expected Effect Size from Parameters

```
compute_expected_r <- function(beta_int, bm_sd, sigma_resid,
                               obs_per_trt, sigma_treat_sd = 0) {
  var_delta <- 2 * sigma_resid^2 / obs_per_trt + sigma_treat_sd^2
  r <- beta_int * bm_sd / sqrt(var_delta)
  return(r)
}

# For our simulation:
expected_r <- compute_expected_r(
  beta_int = beta_int_pop,
```

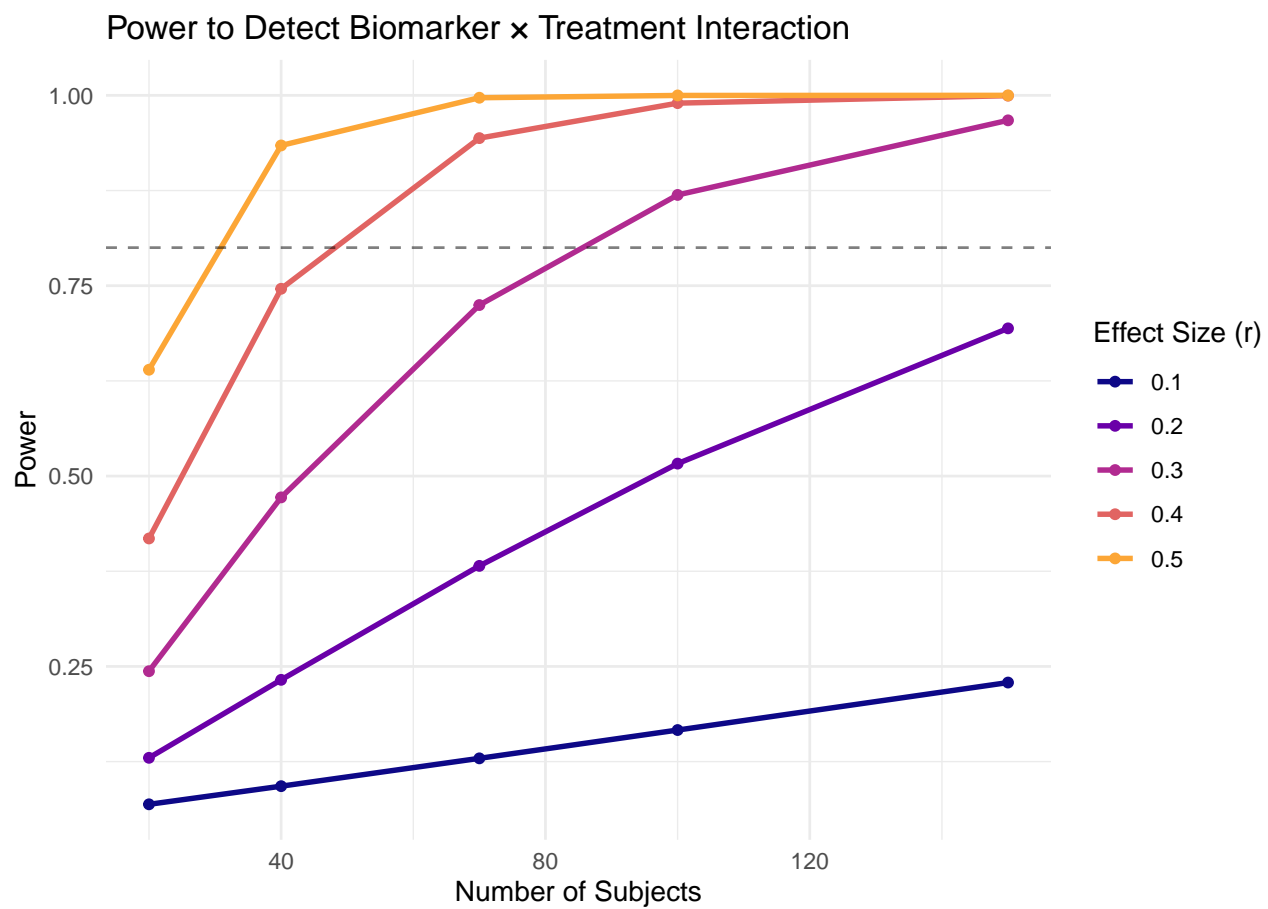


Figure 6: Power curves for interaction detection

```

bm_sd = 1,
sigma_resid = sigma_resid,
obs_per_trt = cycles * obs_per_period / 2,
sigma_treat_sd = sigma_treat_sd
)

cat("Expected r:", round(expected_r, 3), "\n")
#> Expected r: -0.445
cat("Power (N=25):", round(power_correlation(expected_r, n_subj), 3), "\n")
#> Power (N=25): 0.626

```

10.8.3 Power Grid Across Parameters

```

power_grid <- expand_grid(
  n_subj = c(20, 40, 70, 100),
  beta_int = c(-0.5, -1.0, -1.5, -2.0),
  obs_per_trt = c(3, 7, 21),
  sigma_treat_sd = c(0, 3)
) %>%
mutate(
  bm_sd = 1,
  sigma_resid = 5,
  expected_r = compute_expected_r(beta_int, bm_sd, sigma_resid,
                                  obs_per_trt, sigma_treat_sd),
  power = map2_dbl(abs(expected_r), n_subj, power_correlation)
)

ggplot(power_grid, aes(x = n_subj, y = power,
                      color = factor(beta_int),
                      linetype = factor(sigma_treat_sd))) +
  geom_line(linewidth = 0.8) +
  geom_point(size = 1.5) +
  geom_hline(yintercept = 0.8, linetype = "dashed", alpha = 0.3) +
  facet_wrap(~paste("Obs/trt =", obs_per_trt)) +
  scale_color_viridis_d(option = "plasma", end = 0.85) +
  labs(
    title = "Power Across Design Parameters",
    x = "Number of Subjects",
    y = "Power",
    color = "_int",
    linetype = "_treat"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")

```

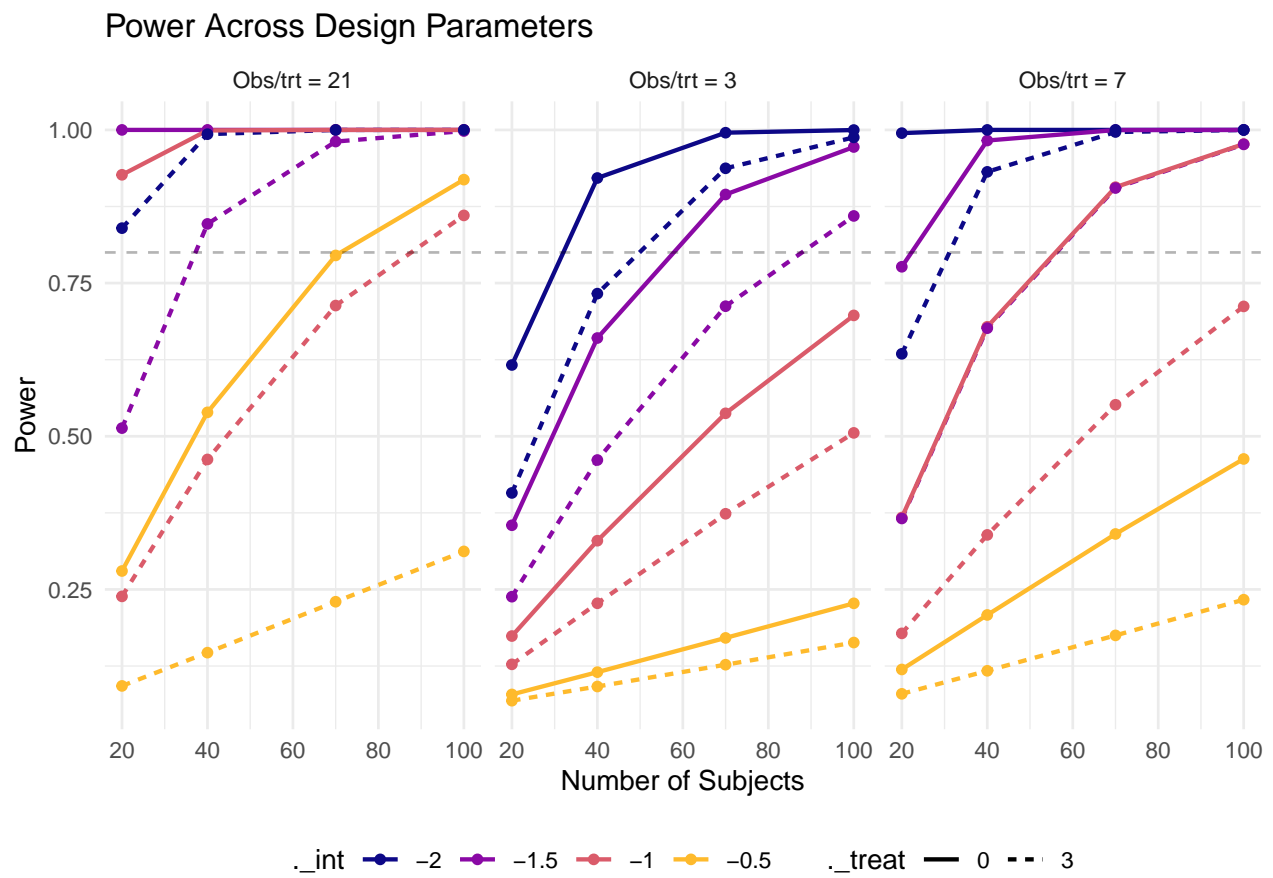


Figure 7: Power across design parameters