Bias-correction in adaptive trials using Stata

Christopher James Rose

Norwegian Institute of Public Health

NIPH
CENTRE FOR EPIDEMIC
INTERVENTIONS RESEARCH

Conflicts of interest

- I have no conflicts of interest to declare

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Outline

- Adaptive trials of public health and social measures (PHSM)
- Bias in adaptive trials
- A general approach to bias correction: conditional maximum likelihood
- Introduce `condml` — a generic command, not limited to adaptive trials
  - Focus on canonical-link GLMs, but `condml` works with any ML estimator
  - `condml` is under active development — some details may change
- Validation: a simulated group sequential trial
- Summary and limitations

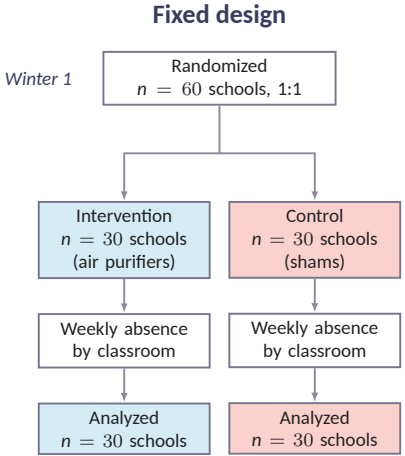
Introduction

Why adaptive trials?

- Public health and social measures (PHSM) widely used during COVID
  - School closures, masking, quarantine, physical distancing
  - Almost no reliable evidence on their benefits and harms
- Center for Epidemic Interventions Research established to run PHSM trials
  - These are often cluster trials (e.g., interventions often delivered to groups)
- Limited to winter respiratory disease season, burdensome interventions, cost
- Adaptive trials: design changes in response to accumulating data
- E.g., stop early, add/drop treatments, drop subgroups, change allocation
- Spread trials over winters, reuse clusters and expensive equipment, redeploy resources

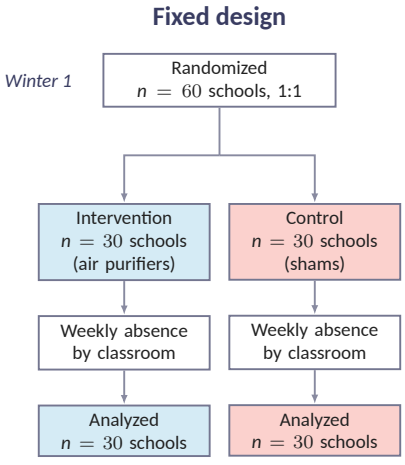
The group sequential design (GSD)

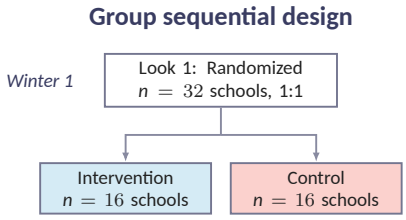
Based on a real cluster-randomized trial of air purifiers in schools (Solberg 2025)



The group sequential design (GSD)

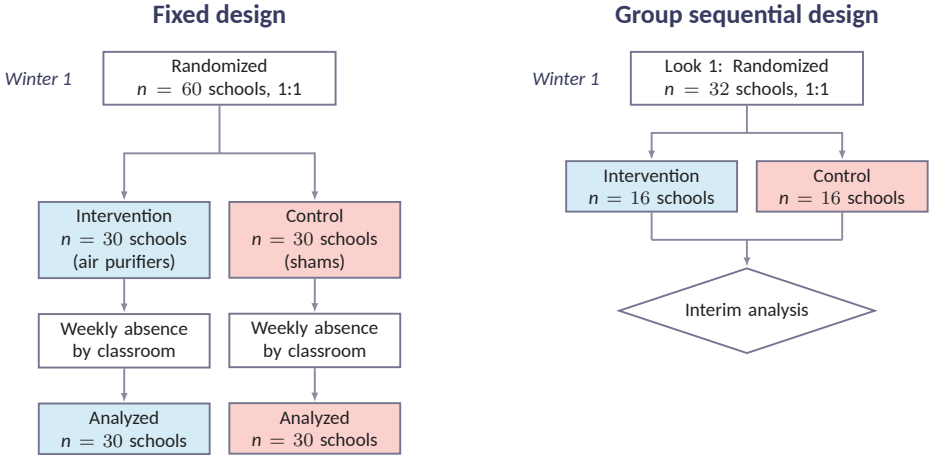
Based on a real cluster-randomized trial of air purifiers in schools (Solberg 2025)





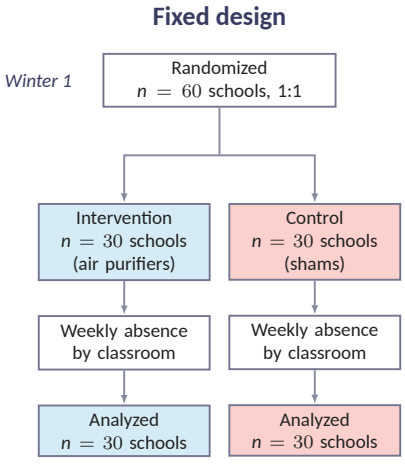
The group sequential design (GSD)

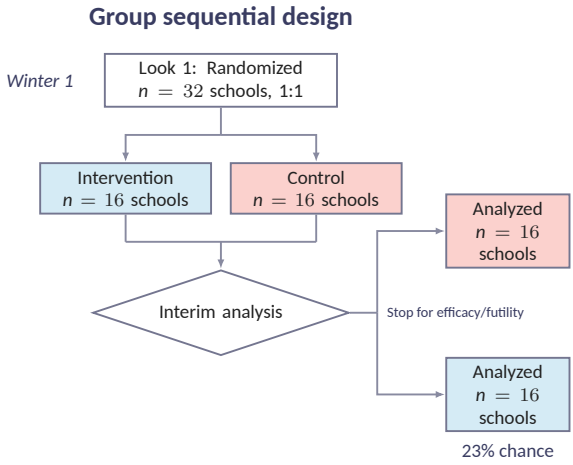
Based on a real cluster-randomized trial of air purifiers in schools (Solberg 2025)



The group sequential design (GSD)

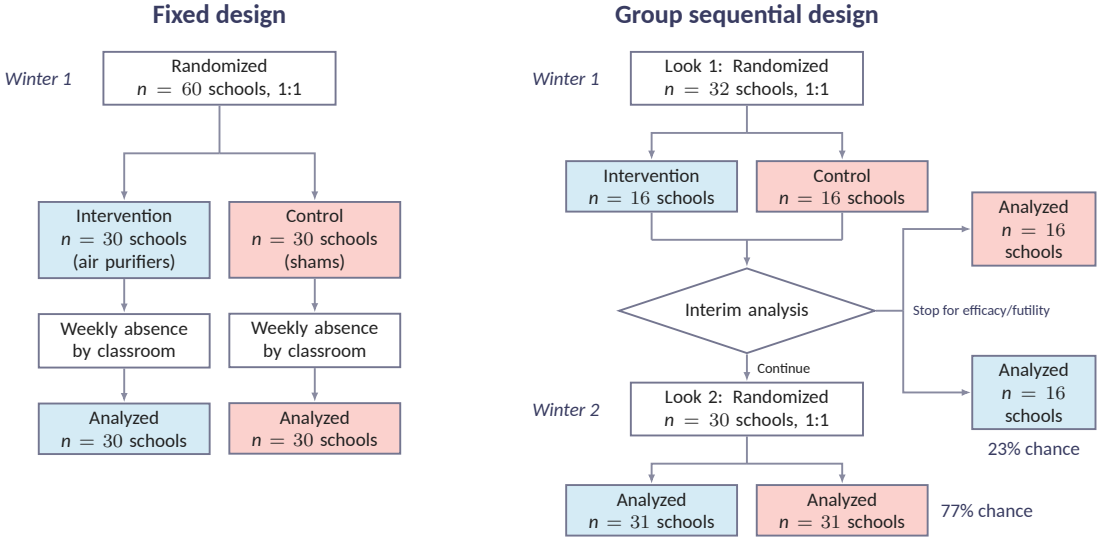
Based on a real cluster-randomized trial of air purifiers in schools (Solberg 2025)





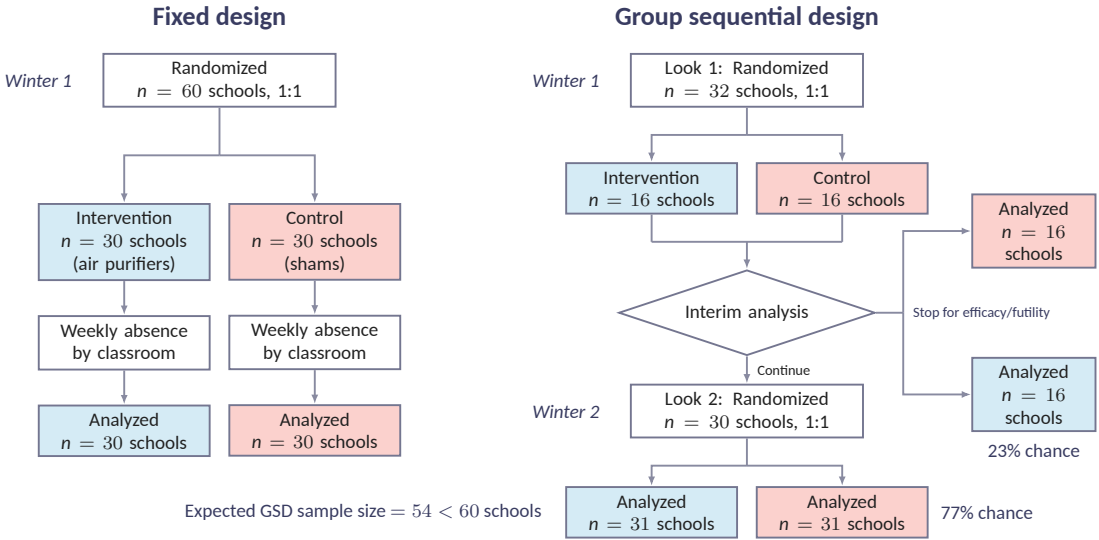
The group sequential design (GSD)

Based on a real cluster-randomized trial of air purifiers in schools (Solberg 2025)



The group sequential design (GSD)

Based on a real cluster-randomized trial of air purifiers in schools (Solberg 2025)



Estimating an odds ratio

Simple example — individually randomized (not a cluster trial)

- Group sequential trial, binary outcome, analyzed at the 402-participant interim
- Use Stata's `logit` to estimate an odds ratio and compute the *z* statistic

```
. logit outcome i.treatment , or nolog
Logistic regression                                     Number of obs =    402
                                                        LR chi2(1)    =  13.29
                                                        Prob > chi2   = 0.0003
Log likelihood = -267.19974                             Pseudo R2     = 0.0243
```

outcome	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
1.treatment	.4757282	.0978439	-3.61	0.000	.3179001	.7119131
_cons	1.05102	.1483124	0.35	0.724	.7970686	1.385883

Note: _cons estimates baseline odds.

- Trial must stop early: $z = -3.61$ is beyond the stopping boundary of -2.9626
- The point estimate for the treatment effect is an odds ratio of 0.48

Illustrative data.

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Bias in adaptive trials

Conventional analyses can give biased effect estimates

- Conventional analyses assume outcomes are statistically independent
- An adaptation is a form of **endogenous selection**
  - Some outcomes are measured/not measured based on outcomes of other units
- Adaptation breaks the independence assumption
- Conventional analysis methods can give biased estimates
- A substantial literature exists, reviewed by Robertson (2023a, 2023b)

Bias correction is known but almost never used

Systematic review	Included	Trials	Bias correction used/reported
Stevely (2015)	GSDs that stopped early	46	3 (7%)
Shimura (2020)	Sequential oncology trials	19	0 (0%)
Ben-Eltriki (2024)	Adaptive trials	317	0 (0%)

Why?

- Adaptive trials are usually designed to control type I and II error rates
- Pharma may use hypothesis testing for **go/no-go** decisions — unbiased estimation less important
- Limited software support for bias correction (SAS: SEQTEST Procedure; R: adestr)
- Closed-form corrections and software only cover simple designs and common outcome types

Design goals for a general solution

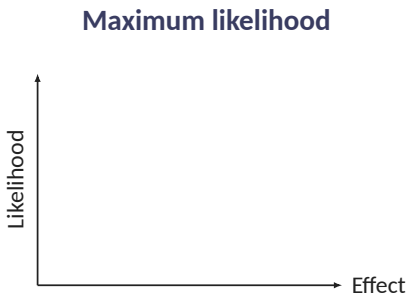
Make it easy to...

- Analyze any adaptive trial design (reduce/eliminate bias)
- Correct estimators commonly used to analyze trials (`logit`, `poisson`, `glm`, ...)
- Compare estimators via Monte Carlo — not covered today

A general approach to bias correction

Conditional maximum likelihood

The idea



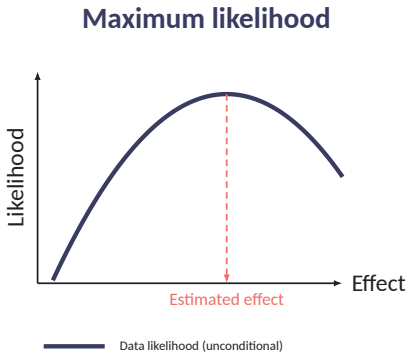
Conditional maximum likelihood

The idea



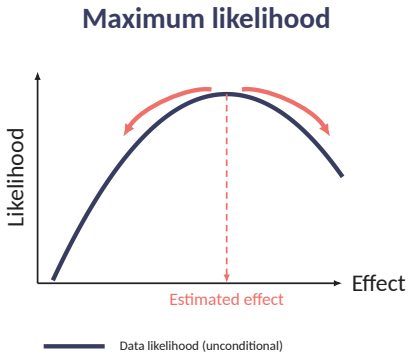
Conditional maximum likelihood

The idea



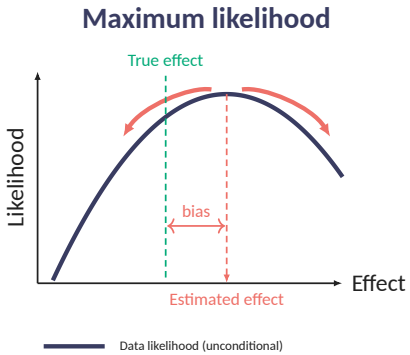
Conditional maximum likelihood

The idea



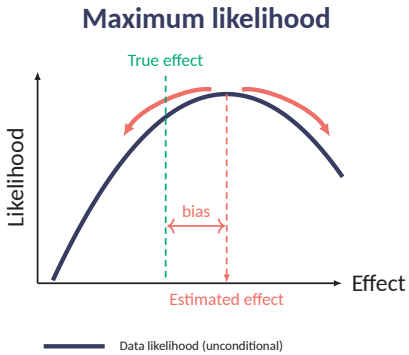
Conditional maximum likelihood

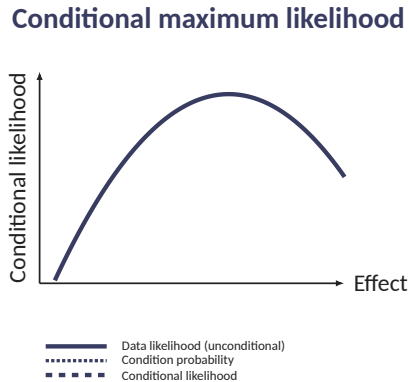
The idea



Conditional maximum likelihood

The idea

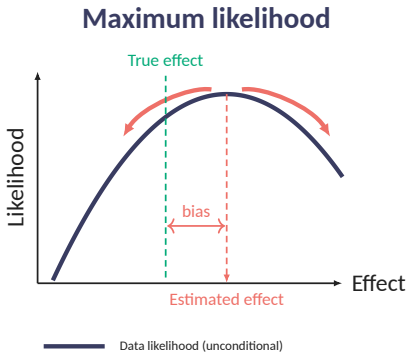


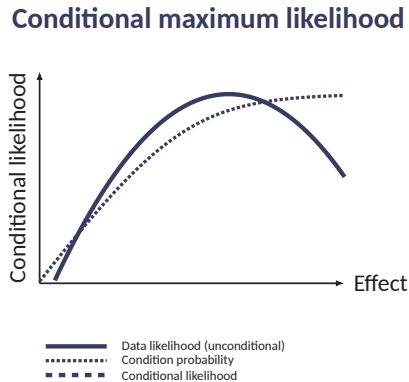


NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Conditional maximum likelihood

The idea

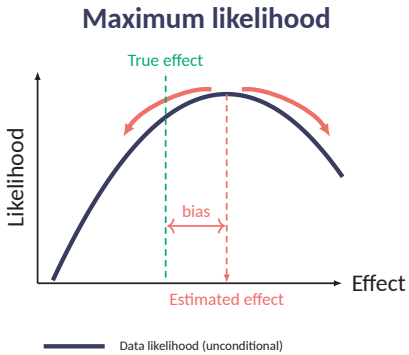


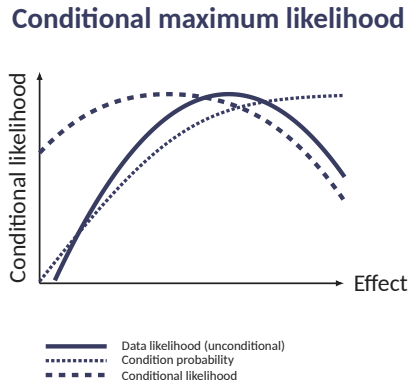


NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Conditional maximum likelihood

The idea

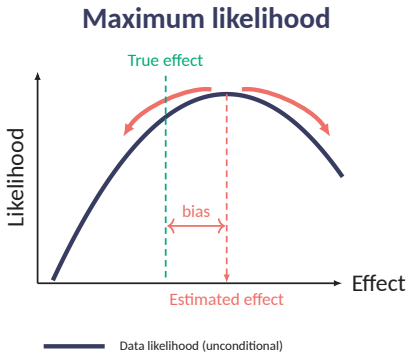


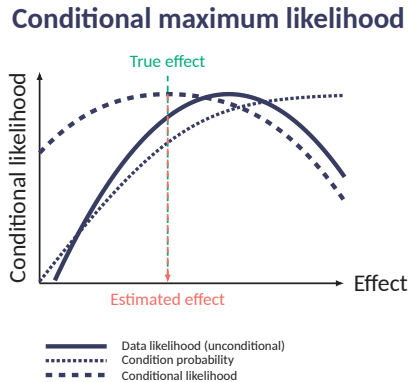


NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Conditional maximum likelihood

The idea

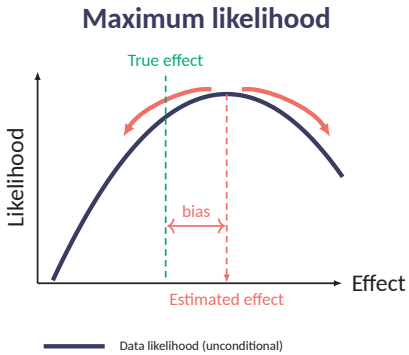


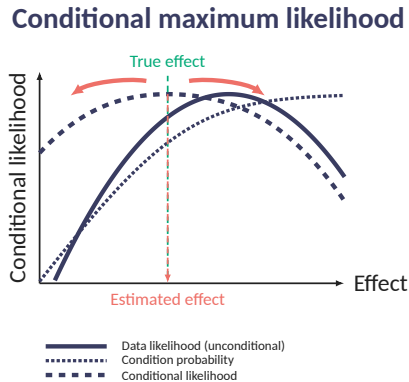


NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Conditional maximum likelihood

The idea





NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Estimating condition probability $P(C = c \mid \beta)$

During maximization

- Canonical-link GLMs are routinely used to analyze trial data, estimating mean differences, odds ratios, risk ratios, rate ratios
- $P(C = c \mid \beta)$ is an expectation over replicate trials, $E_\beta[\mathbf{1}\{C = c\}]$
- This expectation is generally not available in closed form
- We can choose a set of N values $\beta^{(j)}$, around the MLE in a region $\mathcal{R}$
- Simulate and refit one trial per $\beta^{(j)}$
- Each refit shows whether that trial's adaptation matched: $C^{(j)} = c$, or not
- GLMs facilitate simulation at any β (no user-written code needed)

The importance sampling estimator

- Treat the N values of $\beta^{(j)}$ as a sample from a mixture (compound) distribution
- Use importance sampling to estimate $E_\beta[\mathbf{1}\{C = c\}] = P(C = c \mid \beta)$
- Sampling from the mixture, refitting, and computing the mixture: one-off cost
- For canonical link GLMs, log condition probability $\log \hat{P}(C = c \mid \beta)$ is given by:

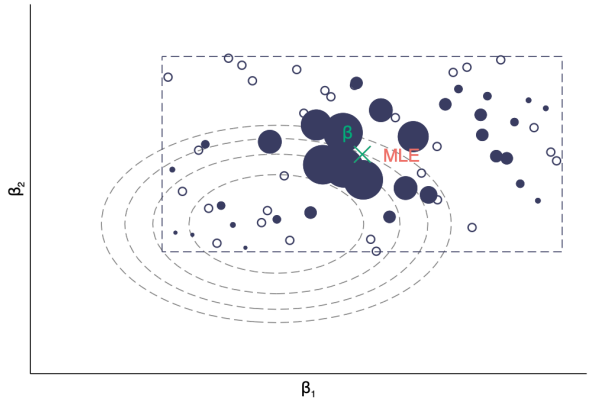
$$\log \sum_{j=1}^{N} \mathbf{1}\{C^{(j)} = c\} \exp\left(\beta^\top S_j - \log\left[\frac{1}{N}\sum_{k=1}^{N} \exp\left(\beta^{(k)\top} S_j - A(\beta^{(k)})\right)\right]\right) - \log N - A(\beta)$$

- S_j and $A(\beta)$ have closed-form expressions for each supported distribution
- The estimate is cheap to compute for any candidate $\beta \in \mathcal{R}$
- Importance weighting is exact, accuracy improves as $N \to \infty$ (consistency)

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

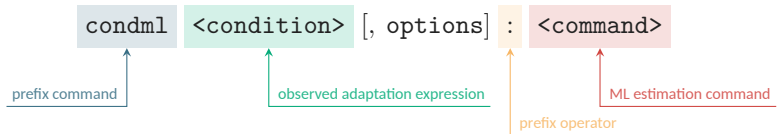
The importance sampling estimator

Visualized



The `condml` command

- `condml`: a prefix command (like `svy:`, `bootstrap:`) — corrects for the trial's adaptation
- Supports importance sampling for canonical-link GLMs and three flexible alternatives
- Simplified syntax:



- `<command>` used to determine outcome family and link, obtain data log-likelihood at any β

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Estimating a bias-corrected odds ratio

Applying condml to the example trial from earlier

- Condition: the trial stopped at the interim for efficacy, $z < -2.9626$

```
. condml _b[1.treatment] / _se[1.treatment] < -2.9626,                ///
>                       strata(treatment)                             ///
>                       seed(1234) nolog nodots eform(Odds ratio) :   ///
>                       logit outcome i.treatment , or
Conditional maximum likelihood (condml: logit)
Sampler: (automatic)                        Scale (SE)                      4.000
Samples: 1000                               Conditional log-likelihood   -266.754
Matches: 581/1000 (58.1%)                   Optimizer iterations                4
```

	Odds ratio	Std. err.	z	P>|z|	[95% conf.	interval]
outcome						
1.treatment	.5730117	.2011519	-1.59	0.113	.2879751	1.140176
_cons	.9552584	.1918961	-0.23	0.820	.6443593	1.416164

- `strata(treatment)` preserves the trial's 1:1 allocation
- `condml` corrected the naive OR from 0.48 to 0.57

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Monte Carlo validation: Group sequential trial

Method: validation experiment design

- Same GSD as before: individually randomized, binary outcome, one interim
- Four possible adaptations: stop for benefit, harm, futility, or continue
- Designed to detect an odds ratio of 0.75 at 80% power, two-sided α = 0.05
- Two scenarios — H_0: odds ratio = 1.00; Target: odds ratio = 0.75
- 1000 trial replications per scenario
- Compares naive and `condml`-corrected estimates

A larger validation experiment is planned.

Results: operating characteristics

	H_0	Target
Type I error / Power		
Naive	5.9% (4.5 to 7.5)	79.5% (76.9 to 82.0)
`condml`	5.3% (4.0 to 6.9)	60.8% (57.7 to 63.8)
Bias — stopped for benefit		
Naive	–	-0.114 (-0.126 to -0.102)
`condml`	–	-0.005 (-0.025 to 0.015)
Bias reduction	–	95.6%
Bias — stopped for futility		
Naive	0.003 (-0.004 to 0.010)	0.123 (0.107 to 0.140)
`condml`	0.005 (-0.008 to 0.017)	-0.014 (-0.044 to 0.017)
Bias reduction	NA	88.6%

Values in parentheses are 95% CIs, from 1000 simulated trials per scenario.

Bias is measured on the log odds ratio scale. Cells with <10 replications are suppressed.

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Results: key takeaways

- `condml`'s type I error is closer to nominal, but appears underpowered at 61%
- Incorrect interpretation: the trial is designed to protect type II error using the naive analysis hypothesis test — that estimator determines trial power
- `condml` cuts conditional bias by 89–96% across adaptations
- Bias causes naive estimator to have sub-nominal confidence interval coverage
- Bias correction restores coverage to 95%

Summary

Summary and limitations

- `condml` is a prefix command that works with any ML estimator
- Simple syntax for canonical-link GLMs (`logit`, `poisson`, `glm`, ...)
- Other estimators need a sampler program (not presented here)
- Validated by Monte Carlo across several adaptive designs
- ML estimators only — excludes `stcox`, common for time-to-event trials

References

- Ben-Eltriki, Mohamed *et al.* (2024). “Adaptive designs in clinical trials: a systematic review—part I”. In: *BMC Medical Research Methodology* 24.1, p. 229.
- Robertson, David S *et al.* (2023a). “Point estimation for adaptive trial designs I: a methodological review”. In: *Statistics in Medicine* 42.2, pp. 122–145.
- — (2023b). “Point estimation for adaptive trial designs II: practical considerations and guidance”. In: *Statistics in Medicine* 42.14, pp. 2496–2520.
- Shimura, Masashi *et al.* (2020). “Assessment of hazard ratios in oncology clinical trials terminated early for superiority: a systematic review”. In: *JAMA Network Open* 3.6, e208633.
- Solberg, Runar B *et al.* (2025). “Classroom air purifiers for reducing school absence: study protocol for a pragmatic, cluster-randomised, parallel, two-arm, group sequential superiority trial”. In: *BMJ Open* 15.12, e109659.
- Stevely, Abigail *et al.* (2015). “An investigation of the shortcomings of the CONSORT 2010 statement for the reporting of group sequential randomised controlled trials: a methodological systematic review”. In: *PLoS One* 10.11, e0141104.

Questions

NIPH
CENTRE FOR EPIDEMIC
INTERVENTIONS RESEARCH

Additional slides

Conditional maximum likelihood

We want the conditional log likelihood — $\log p(Y \mid C = c, \beta)$

the observed adaptation

the observed data

the parameter

1. $\log p(Y, C = c \mid \beta) = \log p(Y \mid C = c, \beta) + \log P(C = c \mid \beta)$	Definition of conditional probability
2. We know $\log p(Y, C = c \mid \beta) = \log p(Y \mid \beta)$	Outcome Y implies condition $C = c$
3. $\log p(Y \mid C = c, \beta) = \log p(Y \mid \beta) - \log P(C = c \mid \beta)$	Substitute and rearrange
4. $\hat{\beta} = \arg\max_\beta \log p(Y \mid \beta) - \log P(C = c \mid \beta)$	Bias-corrected estimate

- Data log-likelihood $\log p(Y \mid \beta)$: obtained by rerunning the estimation command at each candidate β
- Condition probability $P(C = c \mid \beta)$ also needs estimating while maximizing the conditional likelihood — that's next

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

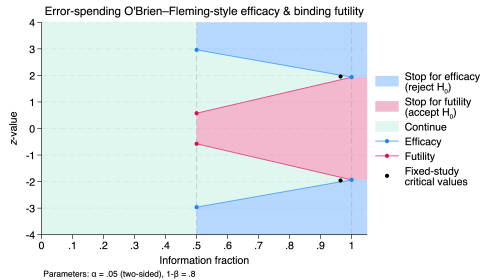
Designing a group sequential trial

Stopping boundaries for a GSD with a binary outcome with one interim look

```
 . gsbounds, efficacy(errobfleming) futility(errobfleming, binding) ///
>           nlooks(2) alpha(.05) power(.8) nopvalues graphbounds
Group sequential boundaries
Efficacy: Error-spending O´Brien-Fleming style
Futility: Error-spending O´Brien-Fleming style, binding
Study parameters:
      alpha = 0.0500   (two-sided)
      power = 0.8000
Info. ratio = 1.0371
Fixed-study crit. values = ±1.9600
Critical values for a group sequential design
```

Look	Info. frac.	Efficacy Lower	Efficacy Upper	Futility Lower	Futility Upper
1	0.50	-2.9626	2.9626	-0.5757	0.5757
2	1.00	-1.9347	1.9347	-1.9347	1.9347

```
Note: Critical values are for z statistics; otherwise,
      use p-value boundaries.
```



NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Estimating $P(C = c \mid \beta)$ at any β

Use importance sampling to estimate the condition probability

- We have trials at the $\beta^{(j)}$, but need $P(C = c \mid \beta)$ at any β for conditional ML
- Treat the N trials as a sample from the proposal mixture distribution $q(y) = \frac{1}{N} \sum_k p(y \mid \beta^{(k)})$
- Estimate $P(C = c \mid \beta) = E_\beta[\mathbf{1}\{C = c\}]$ using importance weights: $\frac{1}{N} \sum_j w_j(\beta) \cdot \mathbf{1}\{C^{(j)} = c\}$

  importance sampling weight
- Every canonical-link GLM's log-likelihood can be written:
  $\log p(y \mid X, \beta) = \beta^\top S(y) - A(\beta) + \log h(y)$
- So trial j's weight at any candidate β can be computed in closed form:

$$w_j(\beta) = \frac{p(y_j \mid \beta)}{q(y_j)} = \frac{\exp(\beta^\top S_j - A(\beta))\ h(y_j)}{\frac{1}{N} \sum_k \exp\left(\beta^{(k)\top} S_j - A(\beta^{(k)})\right)\ h(y_j)}$$

identical — cancels

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Families supported by importance sampling

Distribution	Canonical link	$S(y)$	$A(\beta)$	$h(y)$
Bernoulli/Binomial	Logit	$X^\top y$	$\mathbf{1}^\top \log(1 + e^{X\beta})$	$\binom{m}{y}$
Poisson	Log	$X^\top y$	$\mathbf{1}^\top e^{X\beta}$	$1/y!$
Gaussian	Identity	$X^\top y/\sigma^2$	$(X\beta)^\top (X\beta)/(2\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2/(2\sigma^2)}$
Gamma	Reciprocal	$-X^\top y/\phi$	$-\mathbf{1}^\top \log(X\beta)/\phi$	$\frac{y^{1/\phi - 1}}{\Gamma(1/\phi)\,\phi^{1/\phi}}$

$\log p(y \mid X, \beta) = \beta^\top S(y) - A(\beta) + \log h(y)$; *dispersion ($\sigma^2$, ϕ) held fixed at the original fit's value.*

NIPH
CENTRE FOR EPIDEMIC INTERVENTIONS RESEARCH

Beyond importance sampling

- Importance sampling covers canonical-link outcomes; others need an additional program
- Rao-Blackwellization: works from continuous statistics behind the adaptation
- E.g., an interim z-statistic — treated as approximately normal
- Computes $P(C = c \mid \beta)$ exactly from their fitted mean and covariance
- Two further condition-model estimators are available, for designs where that normal approximation doesn't hold

Bias-correction methods for adaptive trials

Method	Summary
Bias-adjusted MLE	Corrects the naive MLE using an estimate of its bias under the design's stopping rule
Median-unbiased estimator	Inverts a stagewise-ordered test so the estimate is unbiased in the median, not the mean
UMVUE (Rao–Blackwellized)	Uses the Rao–Blackwell theorem on a sufficient statistic; exact, but can be hard to compute
Shrinkage estimator	Shrinks the naive estimate toward a prior value, trading some bias for lower variance
CML estimation	Maximizes the likelihood given the observed adaptation; usually derived by hand for one design

UMVUE: uniformly minimum-variance unbiased estimator. CML: conditional maximum likelihood. Reviewed in Robertson (2023a).

Median unbiased estimator

- Ranks every possible trial outcome — the stage it stopped at, and its test statistic — from strongest to weakest evidence for the treatment
- The ranking (the *stagewise ordering*) is fixed by the design alone — first by stage (earlier stopping for efficacy ranks higher), then by the test statistic
- Defines $p(\beta) = P_\beta(Z \text{ at least as extreme as the observed } z)$, using that ranking
- The estimate $\hat{\beta}_{\text{MUE}}$ solves $p(\hat{\beta}_{\text{MUE}}) = \frac{1}{2}$
- Solving the same equation at other quantiles also gives exact confidence intervals
- Interpretation: $\hat{\beta}_{\text{MUE}}$ is equally likely to over- or under-estimate the true β
- Median-unbiased, not mean-unbiased — `condml`'s conditional MLE isn't exactly mean-unbiased either, only asymptotically so

UMVUE (Rao–Blackwellized)

- Average a naive estimator over every unobserved outcome consistent with what was actually observed
- Rao–Blackwellize a naive unbiased estimator $\tilde{\beta}$ (e.g. the treatment effect estimated using only units enrolled before the adaptation): replace it with $E[\tilde{\beta} \mid T]$, its expectation given a sufficient statistic T for β
- Sufficient statistic for β — e.g. an interim z-statistic — captures everything the data say about β: conditional on that statistic, nothing else adds more information
- This conditional expectation is still unbiased, with variance $\leq \tilde{\beta}$'s (Rao–Blackwell theorem)
- $\hat{\beta}_{\text{UMVUE}}$ is exactly unbiased, with the smallest variance among all unbiased estimators
- Disadvantage: computing $E[\tilde{\beta} \mid T]$ needs $\tilde{\beta}$'s conditional distribution given T, rarely available in closed form outside simple designs
- `condml`'s own Rao–Blackwellized condition model gets $P(C = c \mid \beta)$ exactly only for threshold rules on asymptotically normal statistics — it doesn't produce a UMVUE