Two-Phase Study Designs to Improve Efficiency of New Data Collection

GroupHealth

Sascha Dublin, MD, PhD
Group Health Research Institute

May 4, 2015

Outline

- Background and examples
- Introduction to methodology
- Analytic approaches
- Use of simulation studies to guide decisions

More information

- http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_Supplemental-Information_Two-Phase-Study-Designs.pdf

Background

- Bias can arise in studies using automated data when important measures are omitted or not accurate
- Sometimes there are opportunities to collect additional data on a subgroup
    - Medical record review
    - Surveys, interviews, biologic specimens, etc.
- How best to select that subgroup?

Example 1

- Healthy pregnant woman at 39 weeks asks: what are risks and benefits of inducing labor?
  - Inadequate data from RCTs
  - Observational studies suggest higher risk of cesarean delivery or newborn needing ICU care
- Many studies use automated data and/or birth records which contain inaccurate measures of induction and its indications
  - Algorithm for elective induction using automated data had PPV 36%
- Need better measures of exposure and key confounders (indications)

Example 2

GroupHealth

- Mini-Sentinel project: does saxagliptin (used for diabetes) increase risk of myocardial infarction compared to other therapies?
- Automated/claims data
  - Scant information about smoking, obesity, and other risk factors
- If a signal emerged, would likely want to review some medical records to validate outcomes and measure confounders

Introduction to Methodology

- Two-phase studies are used to estimate the association between an exposure and outcome when:
  - A large (phase 1) sample is available that contains outcome and exposure information; *and*
  - Additional information is needed and can be collected for a subsample (phase 2).
    - Can be about potential confounders, outcome or exposure.

A Simple Scenario

Data available at phase 1:

Exposure (X) and binary outcome (Y) are both observed without error

Data to collect at phase 2:

Confounder information (Z) that can only be obtained using more intensive data collection *(e.g., medical record review)*

Goal: Collect confounder information and estimate the exposure-outcome association using

$$\text{logit}(P(Y = 1|X, Z)) = \boxed{X\beta} + Z\beta_z$$

Phase 1

	Outcome (Y)	
Exposed (X)	Yes	No
Yes	N_1	N_2
No	N_3	N_4

- Phase 1 sample size is $N = N_1 + N_2 + N_3 + N_4$
- Phase 2 sample size is n drawn from N
  - Additional confounder data, Z, is collected for these n observations
- How should we select these n observations?

Study Design

- Simplest option: a random sample of n drawn from N

	Outcome (Y)	
Exposed (X)	Yes	No
Yes	N (N_1+ N_2 + N_3 + N_4)	
No		

- Other choices: stratified on outcome only (case-control) or exposure only

- 2-phase design needs to specify:
  - How will the phase 1 sample be stratified, and
  - How will the phase 2 sample be selected from these strata.

Usual Approach

- Sample based on both outcome and exposure:

	Phase 1			Phase 2	
	Outcome (Y)			Outcome (Y)	
Exposed (X)	Yes	No		Yes	No
Yes	N_1	N_2		n_1	n_2
No	N_3	N_4		n_3	n_4

- Stratify the phase 1 data on basis of both exposure and outcome, then take random sample from each of the four cells

Balanced Design

	Phase 1			**Phase 2**	
	Outcome (Y)			Outcome (Y)	
Exposed (X)	Yes	No		Yes	No
Yes	N_1	N_2		n	n
No	N_3	N_4		n	n

- Sample the same number from each stratum
- The probability of selection varies across strata. Patients in small phase 1 strata have a higher probability of selection.
- This oversampling of patients from small strata improves efficiency.

More on Simple Scenario

- Exposure (X) and binary outcome (Y) are both observed without error at phase 1
- The two-phase design, stratifying on both exposure and outcome, is at least as efficient* as other sampling designs.
- Efficiency gains are greatest when both the exposure and outcome are rare.

**Efficient* refers to the precision of an estimate. A more efficient design gives you greater precision for the same sample size than a less efficient design.

Other Scenarios

Data available at phase 1:

1. Error-prone exposure, outcome observed without error. Two-phase studies are an extension of case-control studies.
   - Most common in the statistical literature
   - Sampling on available exposure and outcome information is never a disadvantage in terms of efficiency
   - Larger efficiency gains when there is less error in exposure measure and the available exposure and outcome are more strongly associated.

Other Phase 2 Scenarios

Data available at phase 1:

2. Error-prone outcome, exposure observed without error.
   - Uncommon in the statistical literature
   - Analogous to Scenario 2 above
3. Both exposure and outcome are observed with error.
   - Very little statistical research in this area. New methodology development is needed.

Analysis of Two-Phase Data

Goal: estimate the association between exposure and outcome using logistic regression

$$\operatorname{logit}(P(Y=1|X,Z)) = X\beta + Z\beta_z$$

Three common estimation approaches are based on different formulations of the likelihood:

1. Weighted likelihood
2. Pseudo or profile likelihood
3. Maximum likelihood

Analysis of Two-Phase Data

1. Weighted likelihood
   - Simple but inefficient
   - Inversely weight observations based on selection probabilities
2. Pseudo or profile likelihood
   - Addresses selection probabilities by including offset terms (variables with coefficients set to 1)
   - Well developed in the statistical literature. Some work still needed for certain scenarios.
3. Maximum likelihood
   - Most efficient approach, but much more complicated to implement

Role of Simulations

- Repeated analysis of randomly generated datasets used to examine the operating characteristics of a statistical procedure in a hypothesized setting
- Useful for complex settings where established procedures may have uncertain behavior
- Can explore the potential benefits of a 2-phase study and also consequences of different design choices.
  - How to stratify phase 1 data
  - Sample size for phase 2
  - Different analytic approaches

Process

GroupHealth

- Generate hypothetical dataset
- Perform the analysis
- Repeat many times
- Analyze results:
  - Bias: does the process on average yield parameters equal to the true value?
  - Coverage probability: how often do 95% CIs from these analyses contain the true value?
  - Power

Example 2: Saxagliptin

- Suppose the Mini-Sentinel surveillance efforts detected a signal: higher risk of MI with use of saxagliptin
- Might want to review medical records, using 2-phase design
- Simulation can examine:
  - Who and how many to sample for Phase 2?
  - What is the estimated bias reduction?
  - How precise might estimates be?

Simulation Parameters

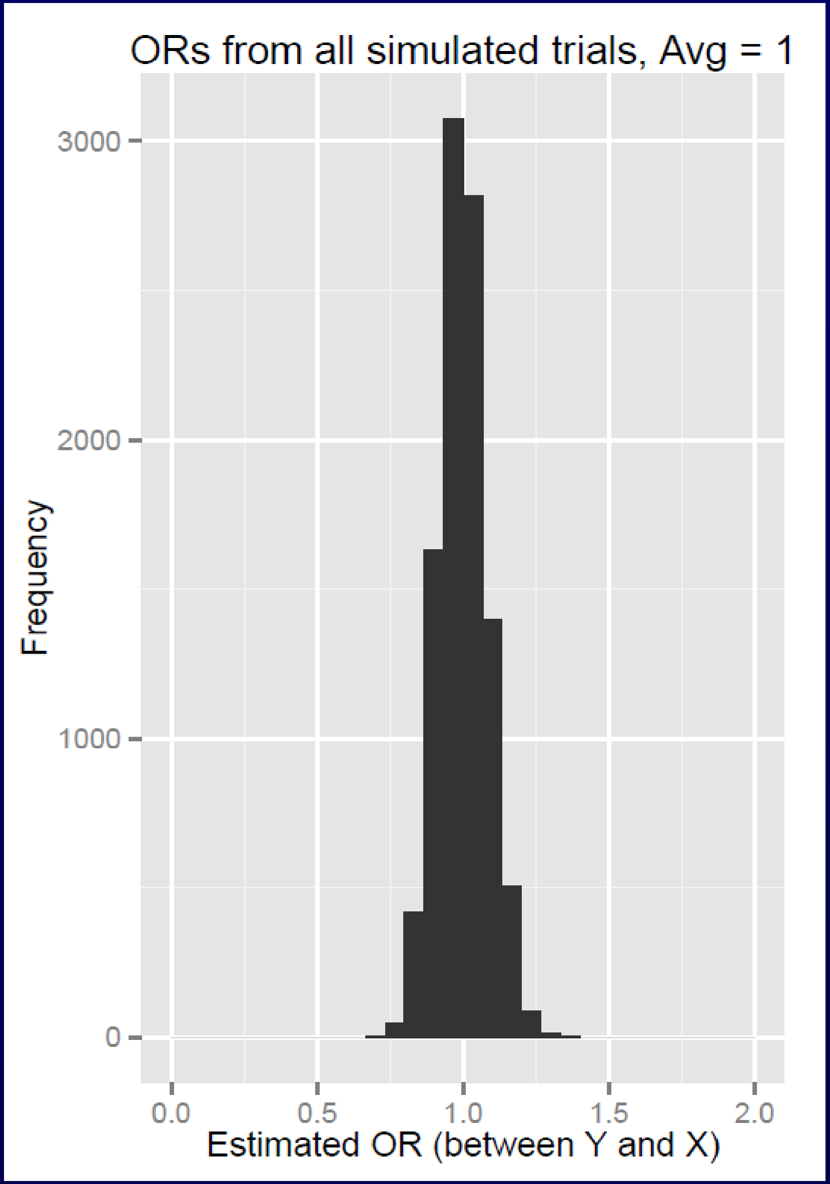
- Population of 150,000 including 20% using saxagliptin
- Outcome incidence: 1/100
- Assumed no true association between exposure and outcome (OR 1.0)
- Confounders: smoking and obesity
  - Assumed prevalence and association with MI based on the literature
  - Assumed no information from administrative data
- These would yield a OR of 1.44 (95% CI 1.28-1.61) – spuriously high due to confounding

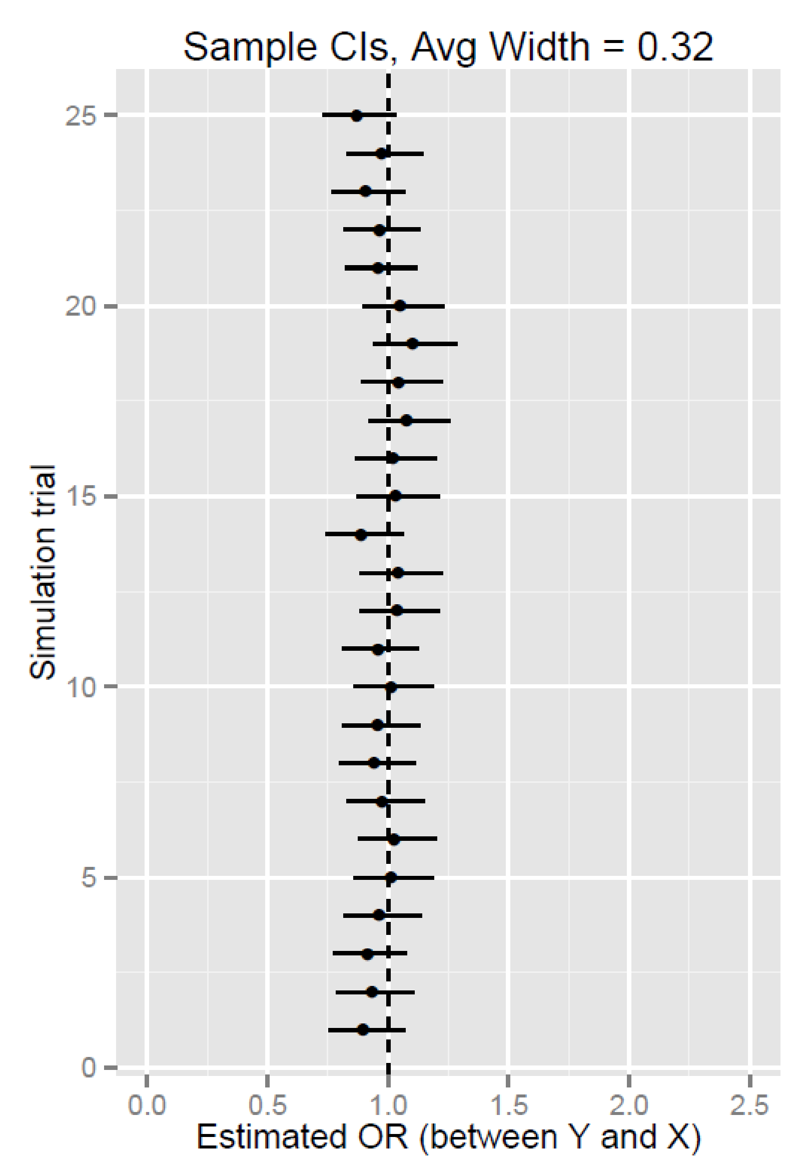
Simulation Question

- If we selected 1000 people for medical record review using a balanced 2-phase design, would this be helpful?

Simulation Results: Pseudo/Profile Likelihood

GroupHealth





Results: Coverage Probability

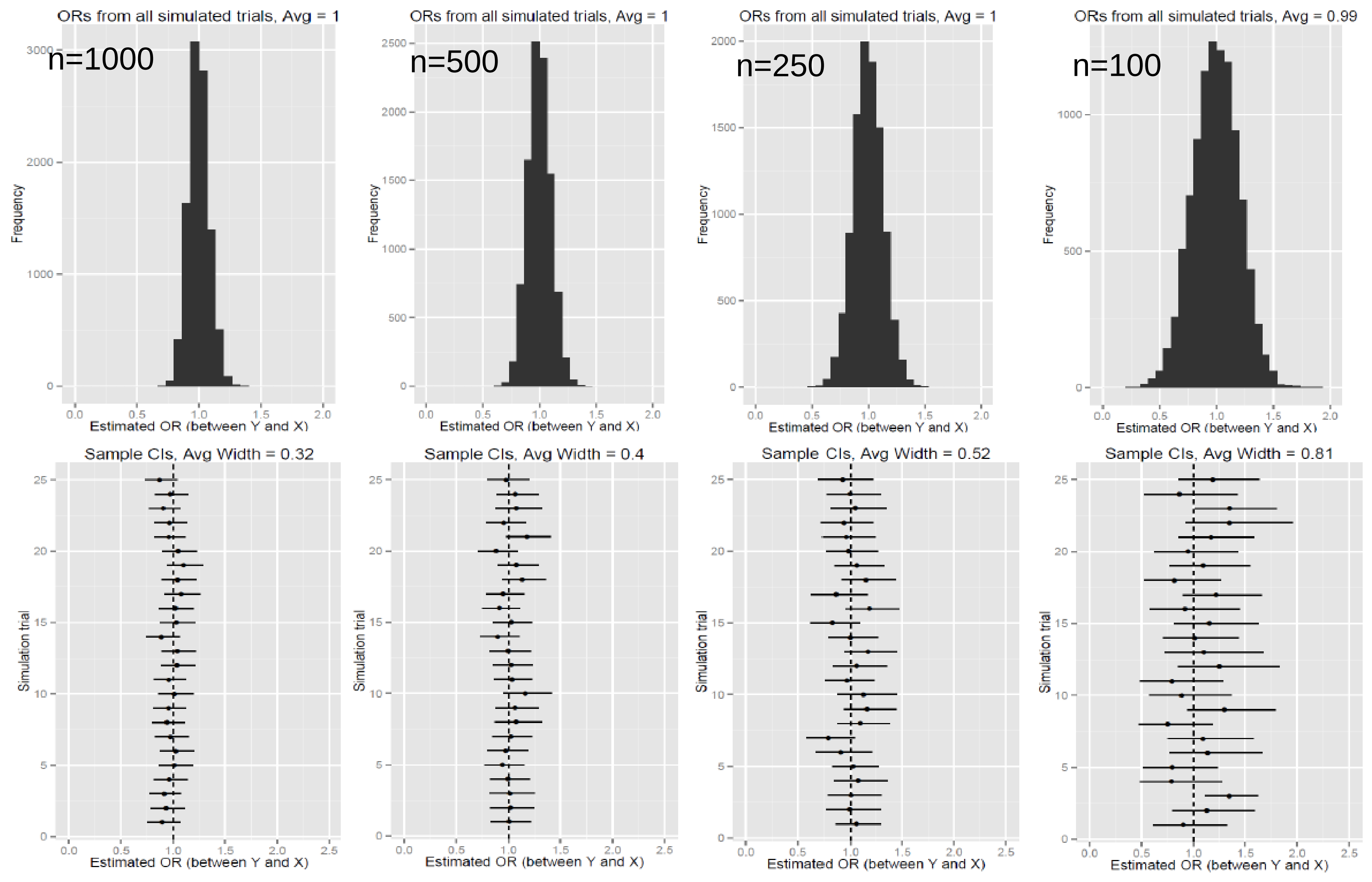
Odds Ratio	% of simulated CIs that excluded it
1.0	5
1.1	20
1.2	62
1.3	91
1.4	99
1.5	100

In this setting, a balanced two-phase design selecting 1,000 people for detailed review would probably be useful.

Exploring Alternatives

- Vary the size of the Phase 2 sample: 1000, 500, 250, or 100
- How might this affect bias and efficiency?

Results


ORs from all simulated trials, Avg = 1
n=1000
Frequency
0
1000
2000
3000
0.0
0.5
1.0
1.5
2.0
Estimated OR (between Y and X)
ORs from all simulated trials, Avg = 1
n=500
Frequency
0
500
1000
1500
2000
2500
Estimated OR (between Y and X)
ORs from all simulated trials, Avg = 1
n=250
Frequency
0
500
1000
1500
2000
Estimated OR (between Y and X)
ORs from all simulated trials, Avg = 0.99
n=100
Frequency
0
500
1000
Estimated OR (between Y and X)
Sample CIs, Avg Width = 0.32
Simulation trial
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
Estimated OR (between Y and X)
Sample CIs, Avg Width = 0.4
Simulation trial
Estimated OR (between Y and X)
Sample CIs, Avg Width = 0.52
Simulation trial
Estimated OR (between Y and X)
Sample CIs, Avg Width = 0.81
Simulation trial
Estimated OR (between Y and X)

Summary

GroupHealth

- Simulation can be used to examine the potential usefulness of conducting a 2-phase study in a particular setting
- Can explore the potential impact of design decisions
- Simulation code available in R – for more information, see Mini-Sentinel workgroup report

Conclusions

- 2-phase studies target the most informative people for review when supplemental data collection is needed
- Increases efficiency
- Key design elements include how to stratify Phase 1 sample and how to select the Phase 2 sample
- Simulations can provide some guidance

More information:

GroupHealth

- http://www.mini-sentinel.org/work_products/Statistical_Methods/Mini-Sentinel_Methods_Supplemental-Information_Two-Phase-Study-Designs.pdf