Imputation Strategies for Missing Binary Outcomes in Cluster Randomized Controlled Trials

Jinhui Ma, Noori Akhtar-Danesh, Lisa Dolovich, Lehana Thabane, for the CHAT group

September 10th, 2010

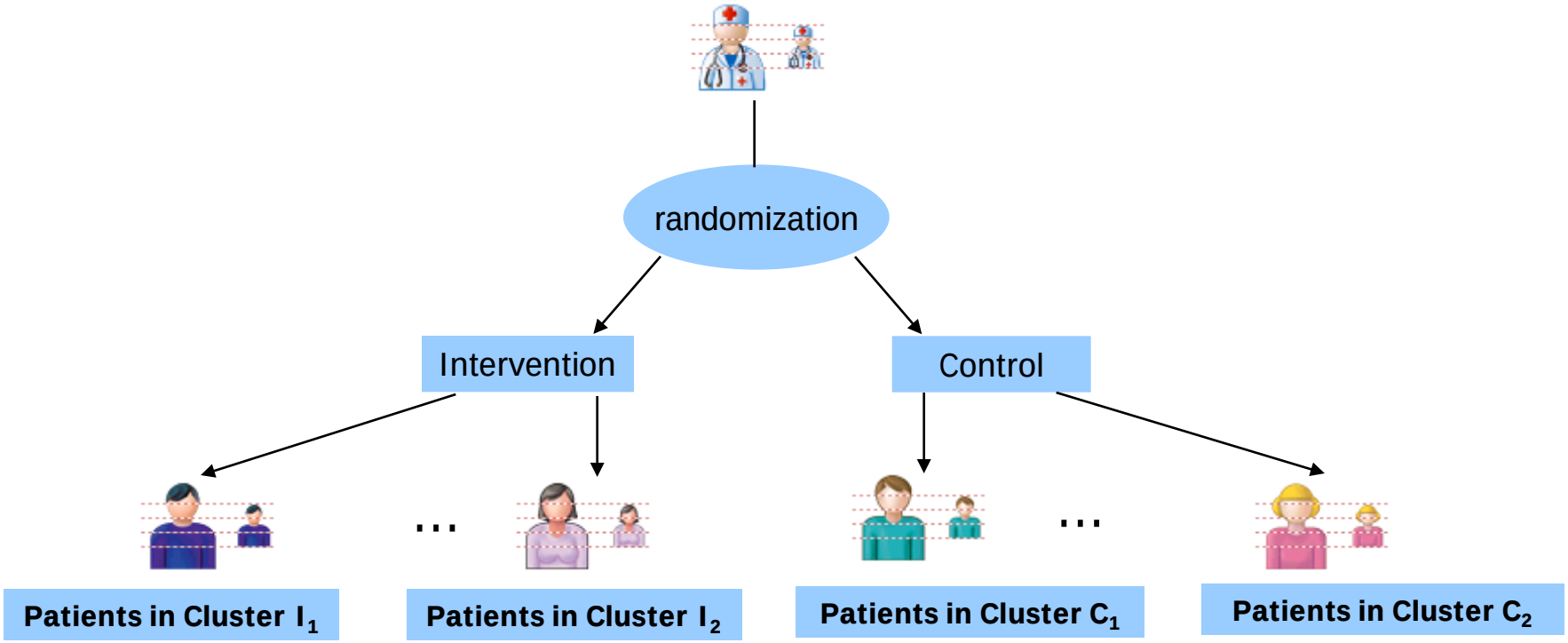
Department of Clinical Epidemiology & Biostatistics
McMaster University

Outline

- Introduction of missing data issues in cluster randomized controlled trials
- Multiple imputation methods
- Simulation study based on the Community Hypertension Assessment Trial
- Results
- Conclusions and limitations

Cluster Randomized Trial

- Cluster randomized trials (CRTs) are increasingly used in health research.



Missing Data in CRT

- Missing data is a problem in CRTs
  - Cause potential bias depending on why data are missing
  - weaken the power of the trial
- Additional concern with missing data in CRTs
  - Entire clusters may be missing

Missing Data in CRT

- Limited attention has been paid to issues on missing binary or categorical data in CRTs.
- Some investigations are done for missing continuous data
  - Taljaard *et al* compared several different imputation strategies for missing continuous outcome in CRTs under the assumption of MCAR.
  - Green *et al* stratified participants into groups that were more homogeneous with respect to the predicted outcome.
  - Etc.

Green SB, Corle DK, Gail MH, Mark SD, Pee D, Freedman LS, Graubard BI, Lynn WR: **Interplay between design and analysis for behavioral intervention trials with community as the unit of randomization.** *Am J Epidemiol* 1995, **142**(6):587-593.

Taljaard M, Donner A, Klar N: **Imputation strategies for missing continuous outcomes in cluster randomized trials.** *Biom J* 2008, **50**(3):329-345.

How to Deal with Missing Data

- Listwise deletion
  - ➢ complete-case analysis

Patient ID	Outcome	Potential Predictors	
	Blood Pressure Controlled	Age	Sex
101	0	65	F
102	0	68	F
103	x	78	M
104	1	70	M
105	0	69	M
106	x	82	F
107	1	67	M
108	1	71	M
.....			
281	0	80	M
282	x	77	F
283	x	73	F
284	1	79	M
285	1	70	F
286	x	81	F

How to Deal with Missing Data

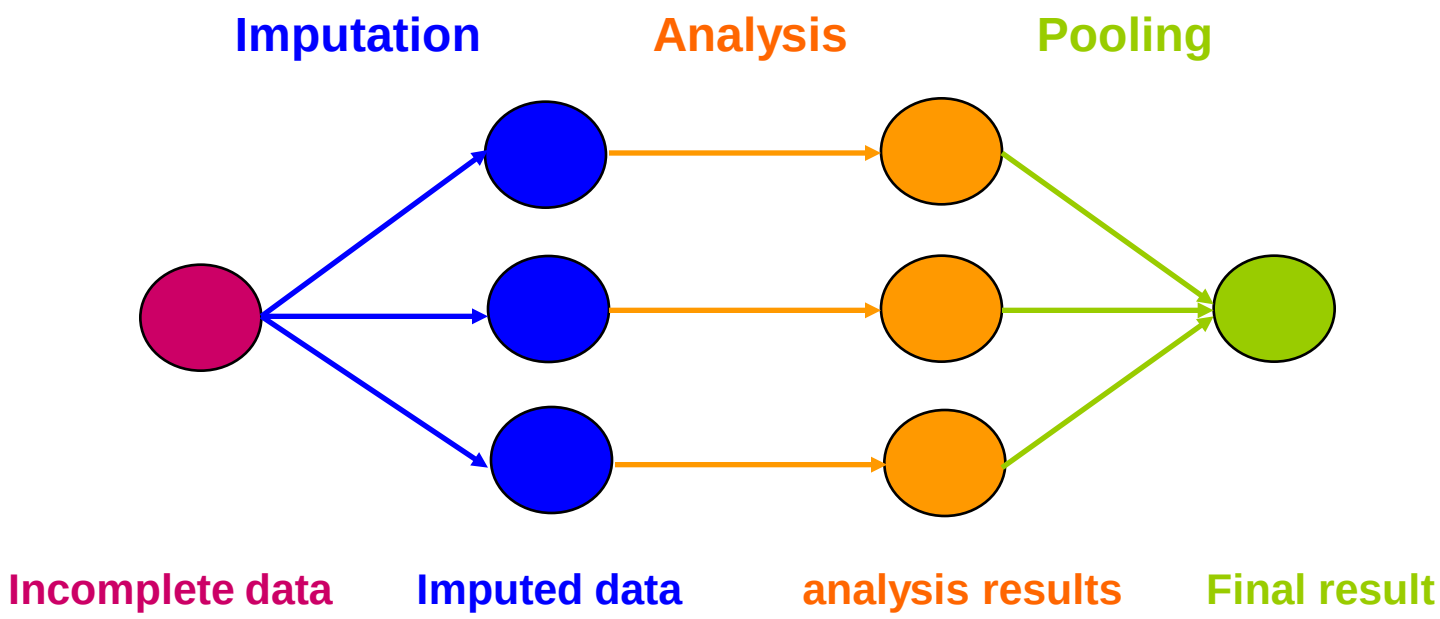
- Single imputation methods
  - Hot deck
  - Cold deck
  - Mean imputation
  - Regression technique
  - Last observation carried forward (LOCF)
  - Composite method

Subject ID	Outcome	Potential Predictors	
	Income	Age	Sex
101	35000	45	F
102	38120	38	F
103	x	28	M
104	40268	50	M
105	68420	49	M
106	x	52	F
107	31050	37	M
108	75000	41	M
.....			
281	32800	30	M
282	x	47	F
283	x	63	F
285	43500	59	M
287	39500	55	F
288	x	51	F

Mean: 44850

Multiple Imputation

- What is Multiple imputation?



http://www.multiple-imputation.com

Multiple Imputation

- Why Multiple imputation (MI)?
  - It is good for both ignorable and non-ignorable missing data
  - Considered as the best technique for imputation
- Relative efficiency of MI?

$$\text{Relative efficiency} = \left(1 + \frac{\lambda}{m}\right)^{-1}$$

m (number of datasets)	λ (percentage of missing) 10	20	30	50	70
3	0.968	0.938	0.909	0.857	0.811
5	**0.980**	**0.962**	**0.943**	**0.909**	**0.877**
10	0.990	0.980	0.971	0.952	0.935
20	0.995	0.990	0.985	0.976	0.966

Rubin, D. B. 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc., New York, NY. P.114

Multiple Imputation

- Multiple imputation procedures are available in standard software such as SAS
- Assumes observations are independent, i.e. ignore the intra-cluster correlation
- Special strategies which account for the intra-cluster correlation have to be developed

Standard Multiple Imputation

- Predictive model (logistic regression)
  - Fitting logistic regression using observed data
  - Construct the posterior predictive distribution of the parameters
  - Fit new logistic regression using parameters simulated from the above posterior distribution to impute missing values

Standard Multiple Imputation

- Propensity score method
  - Calculate propensity score (the conditional probability of being missing) given the observed data

$$\text{logit}(p(y_{ijl} \text{ is missing})) = \beta_0 + \beta_1 sex_{ijl} + \cdots$$

  - Stratify observations into a number of strata based on propensity scores
  - Apply an approximate Bayesian bootstrap imputation to each stratum

    Step 1: First generating a pool of possible “donors” for the missing data by drawing with replacement from the observed data

    Step 2: Then drawing the imputed values with replacement from the donor pool

Standard Multiple Imputation

- Markov chain Monte Carlo method
  - Assuming a joint distribution $P(Y_{mis}, \theta \mid Y_{obs})$
  - Replace Y_{mis} by some assumed values, then simulate θ from the resulting complete data posterior distribution $P(\theta \mid Y_{obs}, Y_{mis})$
  - Let $\theta^{(t)}$ be current value of θ, then $Y_{mis}^{(t+1)} \sim P(Y_{mis} \mid Y_{obs}, \theta^{(t)})$
  - Conditioning on $Y_{mis}^{(t+1)}$, $\theta^{(t+1)} \sim P(\theta \mid Y_{obs}, Y_{mis}^{(t+1)})$
  - Repeat above procedure until $\{(\theta^{(t)}, Y_{mis}^{(t)}), t = 1, 2, ...\}$ converge in distribution to $P(Y_{mis}, \theta \mid Y_{obs})$

Proposed Methods

- Within-cluster MI
  - Predictive model (logistic regression)
  - Propensity score method
  - Markov chain Monte Carlo (MCMC) method

Cluster	Patient ID	Outcome	Potential Predictors	
		Blood Pressure	Age	Sex
1	101	0	65	F
1	102	0	68	F
1	103	x	78	M
1	104	1	70	M
1	105	0	69	M
1	106	x	82	F
1	107	1	67	M
1	108	1	71	M
	.....			
28	281	0	80	M
28	282	x	77	F
28	283	x	73	F
28	284	1	69	F
28	285	1	79	M
28	286	1	70	F
28	287	1	75	F
28	288	x	81	F

Proposed Methods

- Across-cluster MI
  - Propensity score method

$$\text{logit}(p(y_{ijl} \text{ is missing})) = \beta_0 + \beta_1 cluster_{ij} + \beta_1 sex_{ijl} + \cdots$$

y_{ijl} is the outcome for patient *l* within FP *j* in the treatment group *i.*

  - Random-effects logistic regression

$$\text{logit}(p(y_{ijl} = 1)) = \beta_0 + \beta_1 sex_{ijl} + \cdots + u_{ij}$$

$$u_{ij} \sim N(0, \tau^2)$$

CHAT Study

- The community hypertension assessment trial (CHAT)

Simulation Study

- Generate missing outcome completely at random
  - Let y_{ijl} be the outcome for patient l within FP j in the treatment group i.
  - Generate $m_{ijl} \sim \text{Bernoulli}(p_0)$, which indicates whether the value of newly generated outcome is missing or not.
  - Create new outcome z_{ijl}

$$z_{ijl} = \begin{cases} y_{ijl} & \text{if } m_{ijl} = 0 \\ . & \text{if } m_{ijl} = 1 \end{cases}$$

Simulation Study

- Generate missing outcome at random
  - Let p_{ijl} be the probability of missing outcome for patient l within FP j in the treatment group i.

$$p_{ijl} = \frac{\exp(\beta_0 + \log(1.2) * sex_{ijl} + \log(1.4) * bp\_base_{ijl})}{1 + \exp(\beta_0 + \log(1.2) * sex_{ijl} + \log(1.4) * bp\_base_{ijl})}$$

  - Set β_0 to make sure the overall missingness is p_0
  - Generate $m_{ijl} \sim Bernoulli(p_{ijl})$, which indicates whether the value of newly generated outcome is missing or not.
  - Create new outcome z_{ijl}

$$z_{ijl} = \begin{cases} y_{ijl} & \text{if } m_{ijl} = 0 \\ . & \text{if } m_{ijl} = 1 \end{cases}$$

Simulation Study

- Compare agreement between complete CHAT dataset and imputed datasets

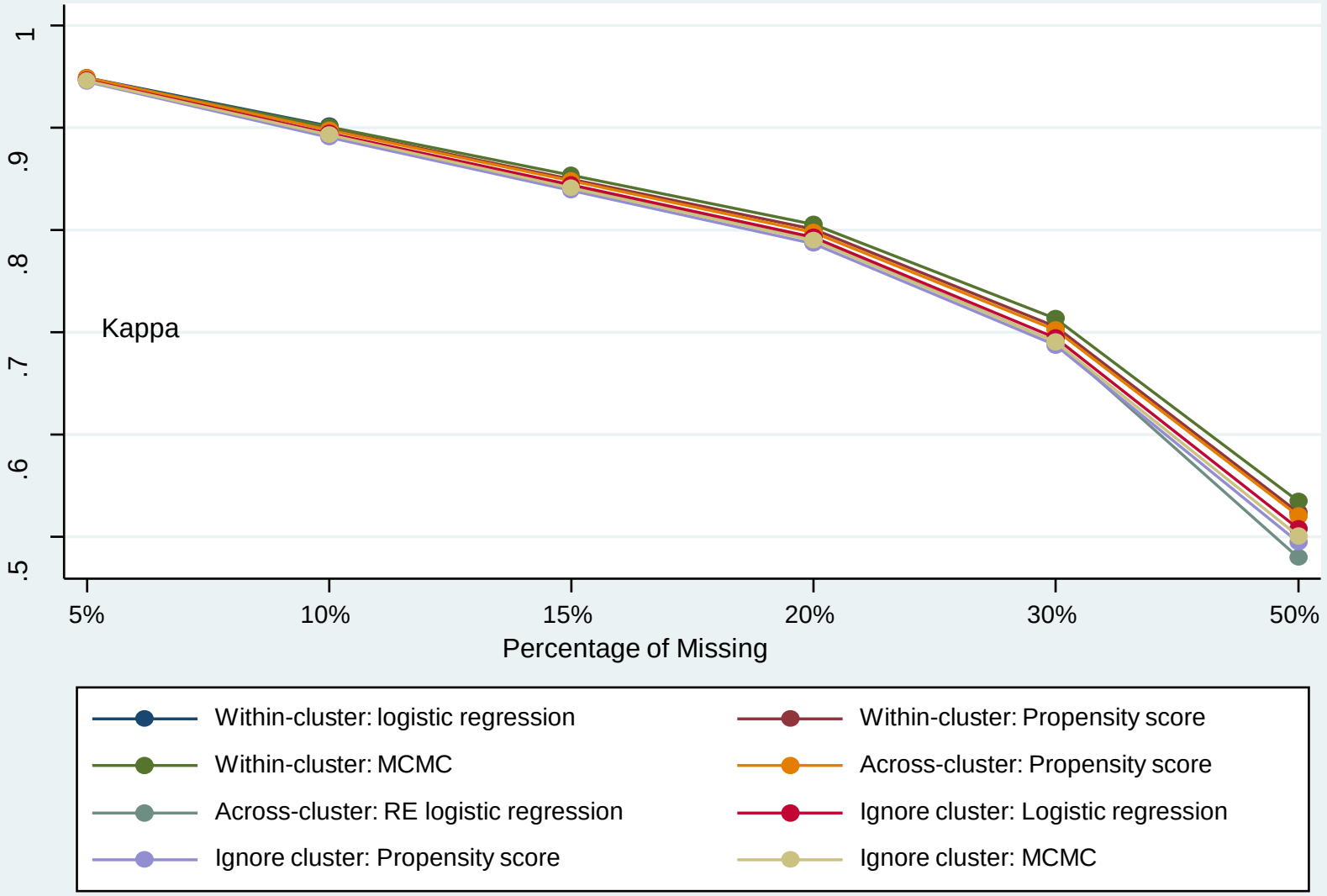
$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

$\Pr(a)$ is the probability of observed agreement

$\Pr(e)$ is the probability of random agreement

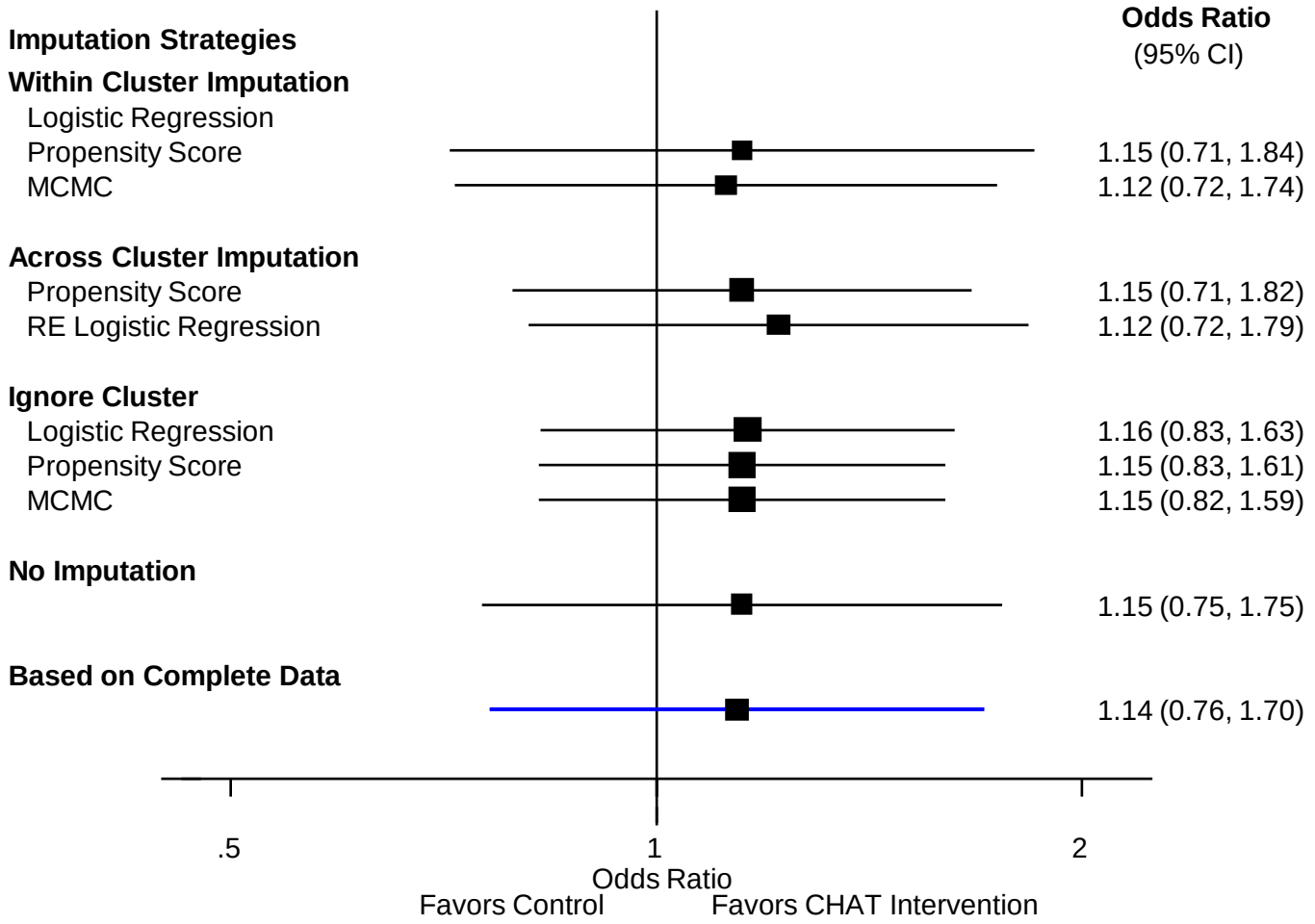
- Compare treatment effects estimated from complete CHAT dataset and multiple imputation
  - Generalized estimating equation (GEE)
  - Random-effects logistic regression (RE)

Results


1
.9
.8
.7
.6
.5
Kappa
5%
10%
15%
20%
30%
50%
Percentage of Missing
Within-cluster: logistic regression
Within-cluster: Propensity score
Within-cluster: MCMC
Across-cluster: Propensity score
Across-cluster: RE logistic regression
Ignore cluster: Logistic regression
Ignore cluster: Propensity score
Ignore cluster: MCMC

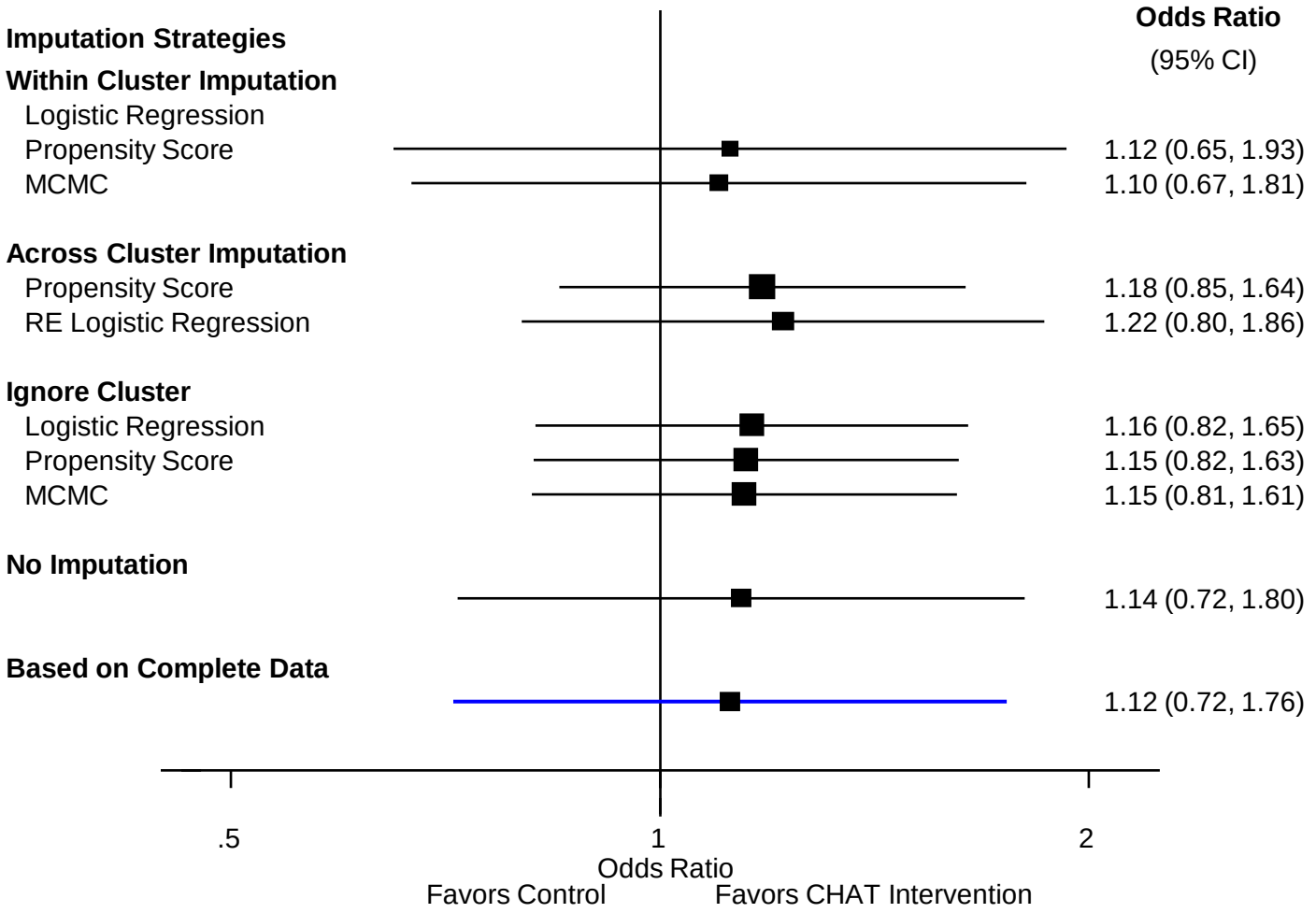
Results

Comparison of the estimate of treatment effect from GEE model for different imputation strategies when 30% data is missing at random



Results

Comparison of the estimate of treatment effect from RE model for different imputation strategies when 30% data is covariate dependent missing



Conclusions

- Investigate the missing data mechanism and pattern of missing before applying any imputation strategies
- Consider the intra-cluster correlation when the amount of missing or intra-cluster correlation are large
- Predictive model and MCMC methods are consistently better than propensity score method in imputing binary outcome from CRTs
- Results from GEE and RE model are unbiased if covariates associated with missing data mechanism are adjusted for

Limitations

- Imputation for a particular CRT design
  - Completely randomized design
  - Two level of clustering
  - ICC = 0.077; cluster size = 55; number of clusters = 28
- Did not consider imputation strategies for missing categorical outcome and missing covariates
- Did not consider the case of whole cluster missing

Thanks!

Prognostic Imbalance in Randomized Controlled Trials

Presenter: Rong (Rachel) Chu

Clinical Epi. & Biostats., McMaster University

CANNeCTIN Methodology Videoconference

September 10, 2010

Outline

- Prognostic imbalance (PI) in RCTs
- Measures of PI
- Impact of PI on effect estimates
- Simulation study
- Discussion

Introduction

- A prognostic factor (PF): a situation, condition, or a characteristic of a patient that affects his/her risk of the outcomes of interest or responsiveness to therapy
- In clinical trials, prognosis at the time of treatment assignment are of particular interest (“baseline covariates”)
- PFs measured during the course of disease may be affected by the choice of treatment and are of limited use

Introduction

- Randomization
  - On average, balances prognosis (known, unknown) between treatment groups
  - No guarantee of exact similarity of groups, depending on sample size and patient variation
  - Serious concern arises when large group differences occur in a baseline variable that has high prognostic value

Introduction

- Known PFs
  - Adjustment in design and/or analysis phase
    - Stratified randomization (w/ blocking), minimization
    - Stratified analysis, covariate adjustment via multiple regression
    - “Surprising” results on independent, balanced PFs…

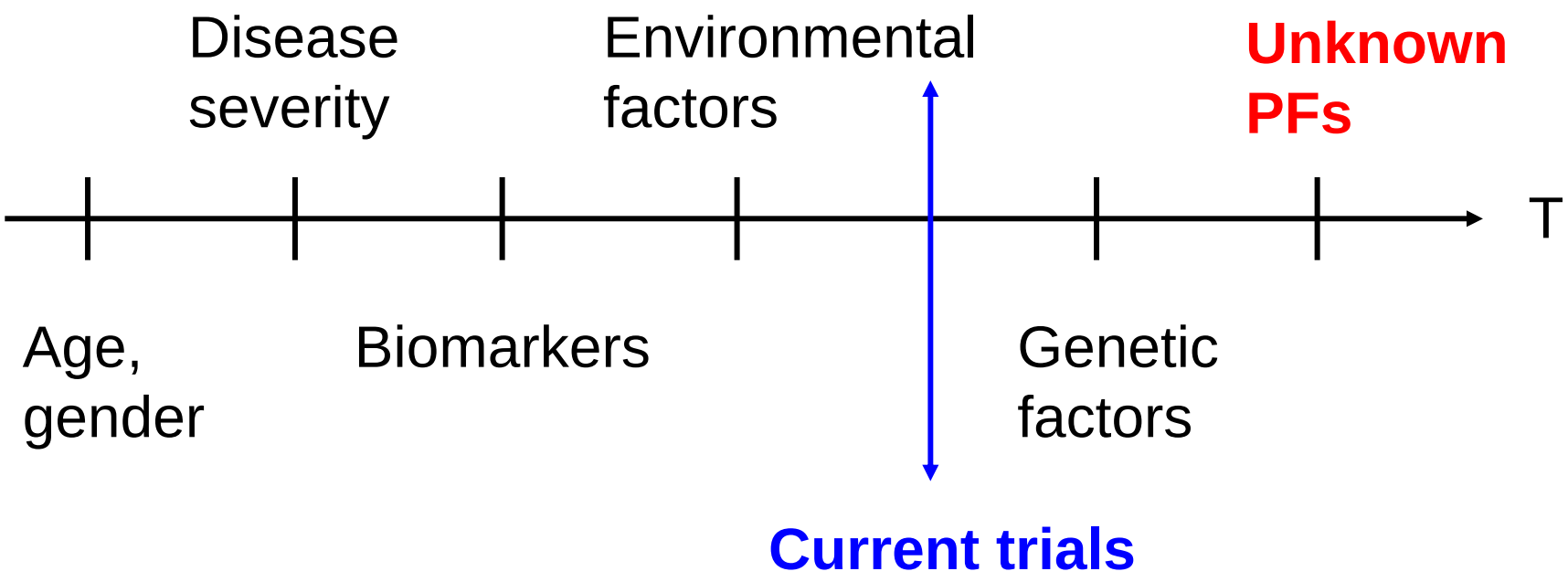
Yesterday Once More

Dr. Ramsay (2009). Bias in logistic regression due to omitted covariates

- Linear regression
  - omitting balanced, independent covariates doesn't bias effect estimates
  - including important covariates increases precision of effect estimate
- Logistic regression
  - omitting balanced, independent covariates DOES bias effect estimates (towards the null)
  - including these covariates DEREASES precision of effect estimate

Introduction

- Scientific evidence cumulates gradually



Introduction

- Unknown or unobservable PFs
  - Cannot be controlled at design or analysis phase
  - Rely on simple randomization
  - Creditability of well conducted, small – moderate size RCTs may suffer
  - Some advocate large, simple RCTs
  - Rare outcomes → huge sample size

Introduction

- E.g. Cardiovascular trials: moderate effect size (RRR=25%), 1% control event rate, 80% power, 5% type I error → about 22,000 patients per arm
- Many RCTs published in major clinical journals are underpowered
- Understand to how much unknown PFs can impact effect estimates from previous/current RCTs

Research Questions

- What is the probability that an imbalance of unknown PFs occur in simple randomized trials, when the PFs are truly present?
- What is the impact of ignoring prognostic imbalance on Rx effect estimation (due to unknown PFs), when the PFs are present?

Application

- Make statistical inference on individual trials
- Plan future studies
- Rate quality of evidence for systematic review
- Grade strength of recommendations in clinical guidelines

The Grading of Recommendations Assessment, Development and Evaluation GRADE

Factors in deciding on quality of evidence

Factors that might decrease quality of evidence

- Study limitations
- Inconsistency of results
- Indirectness of evidence
- Imprecision
- Publication bias

- Factors that might increase quality of evidence
- Large magnitude of effect
- Plausible confounding, which would reduce a demonstrated effect
- Dose-response gradient

http://www.gradeworkinggroup.org/index.htm

Likelihood of Prognostic Imbalance (PI)

- Simulations
  - Two treatment groups (X1): equal number of patients per arm
  - A single binary prognostic factor (X2) ~ Bernoulli (λ)
  - Prevalence of X2 (λ): 0.05 - 0.995
  - Simple randomization (X1 is uncorrelated to X1 in expectation)
  - Sample size: 125, 500, 2000 per arm
  - 10,000 replicates

Likelihood of Prognostic Imbalance (PI)

- Measures of PI
  - *Imb1* (absolute measure):

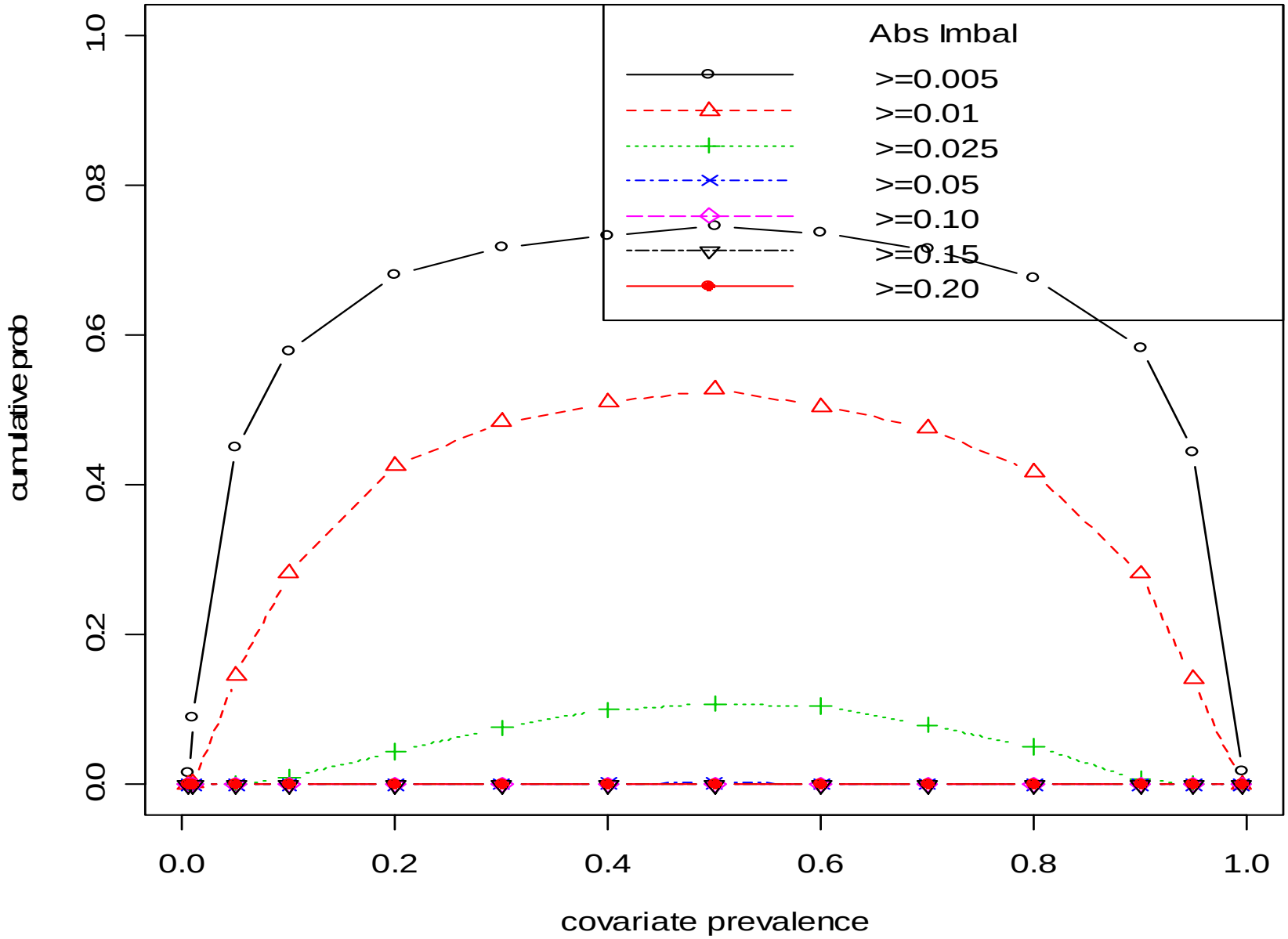
    |pt-pc|
  - *Imb2* (standardized measure):

    |pt-pc|/ sqrt(0.5*pt(1-pt)+0.5*pc(1-pc))
  - *Imb3* (relative measure):

    |pt-pc| / true prevalence of X2

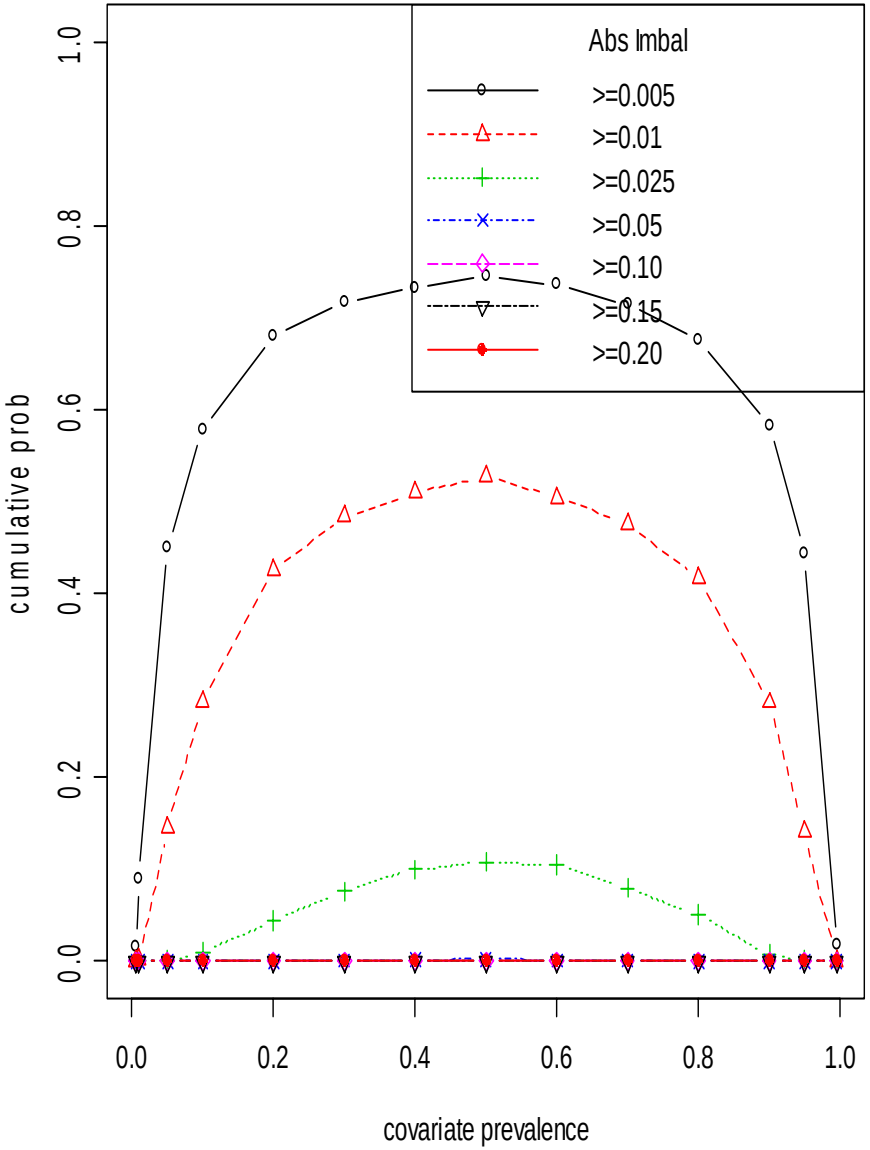
pt, pc: est. proportion of patients having the PF in the treatment or control group with continuity correction

Imb1 (absolute measure), 2000 per arm

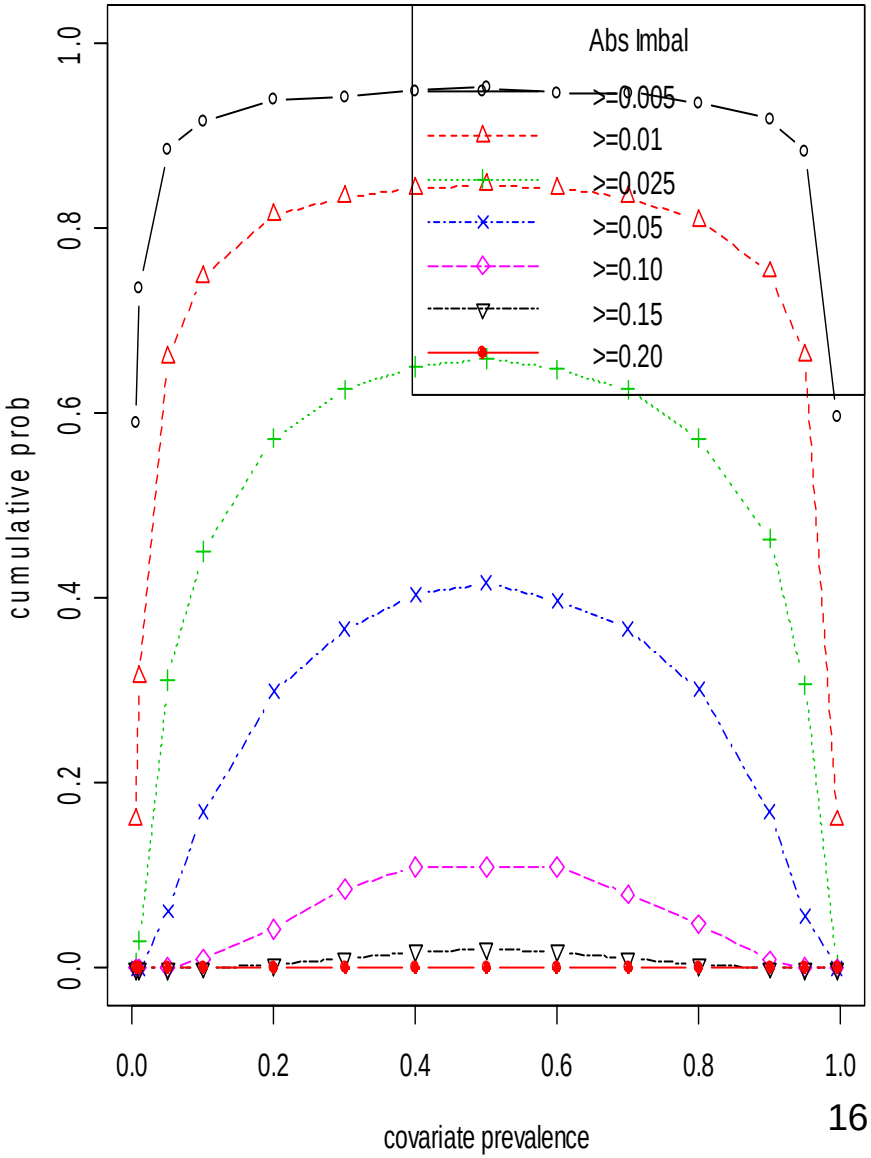


Imb1 (absolute measure)

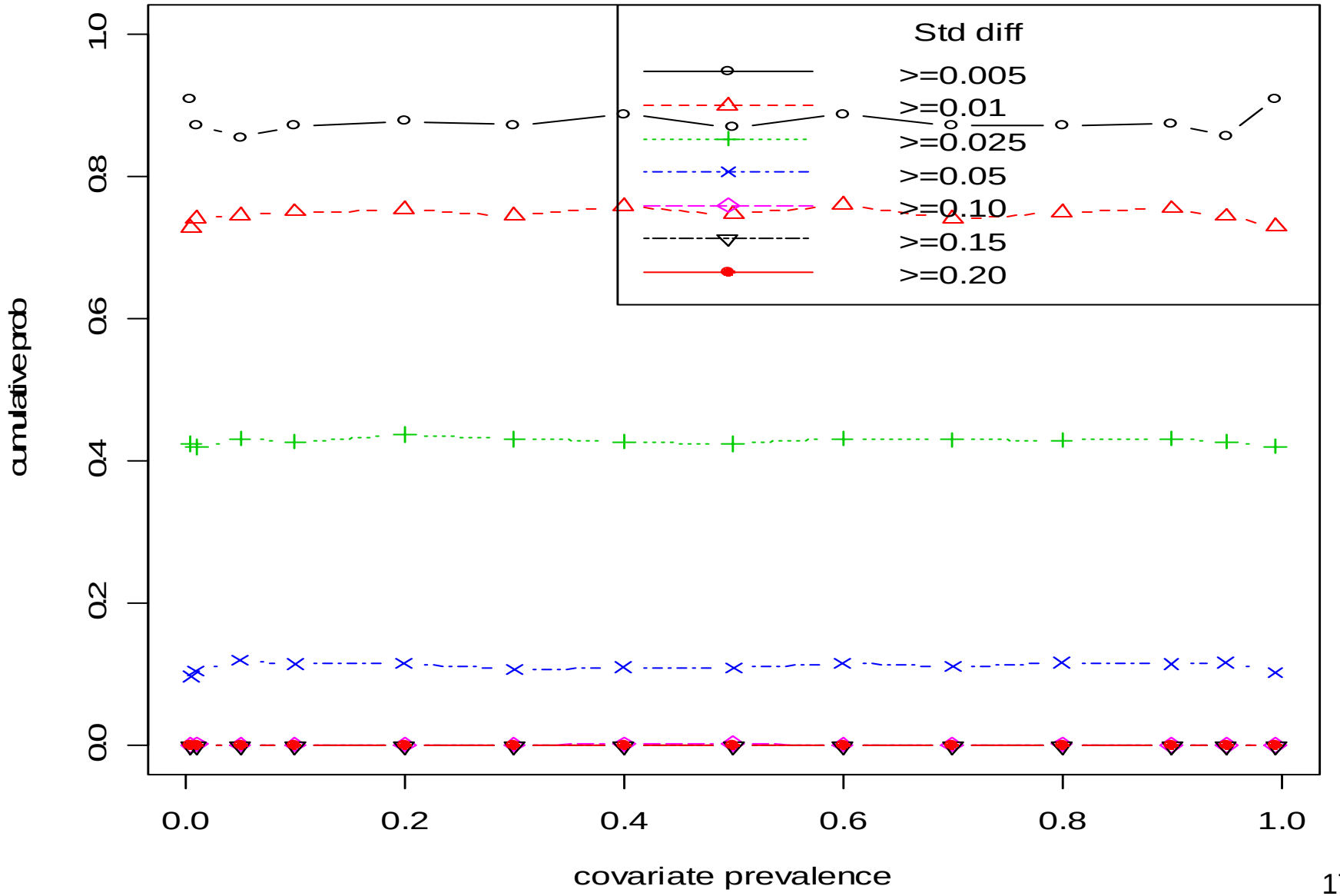
2000 per arm



125 per arm

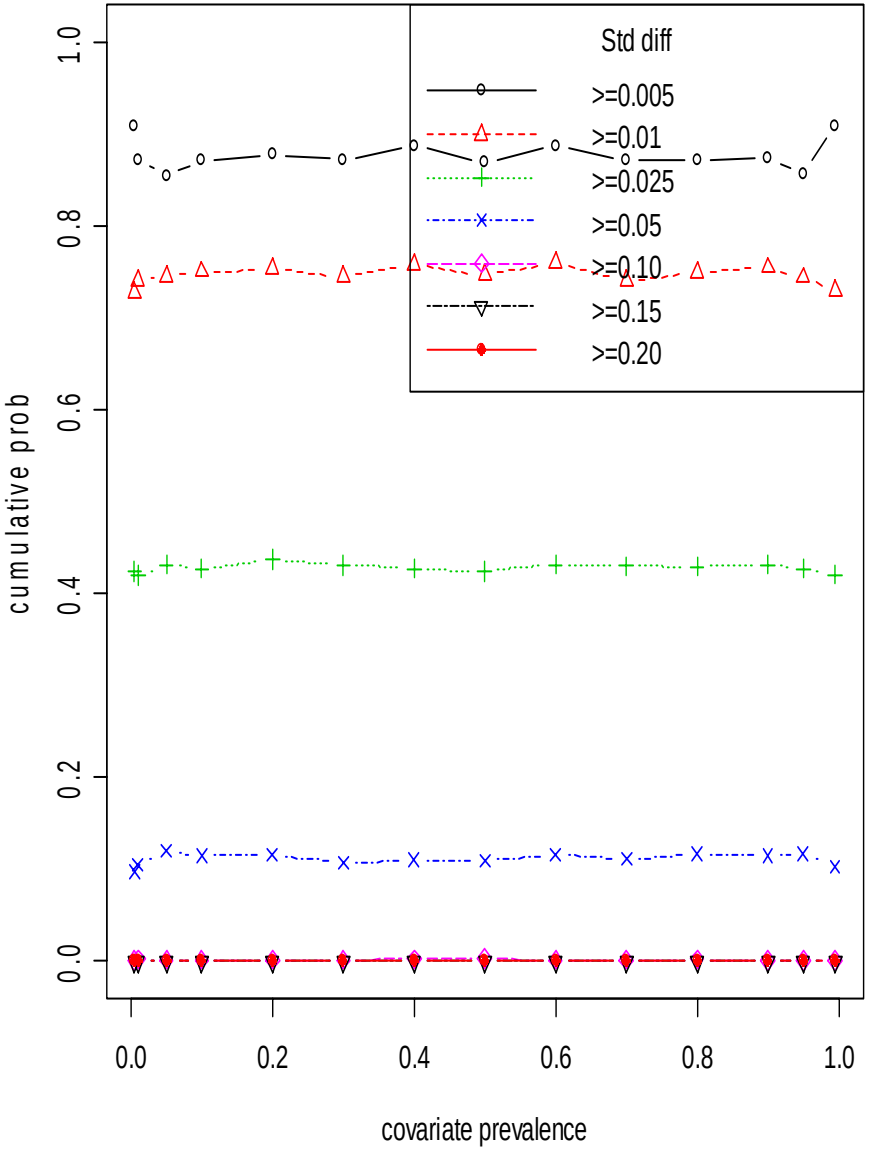


Imb2 (standardized measure), 2000 per arm

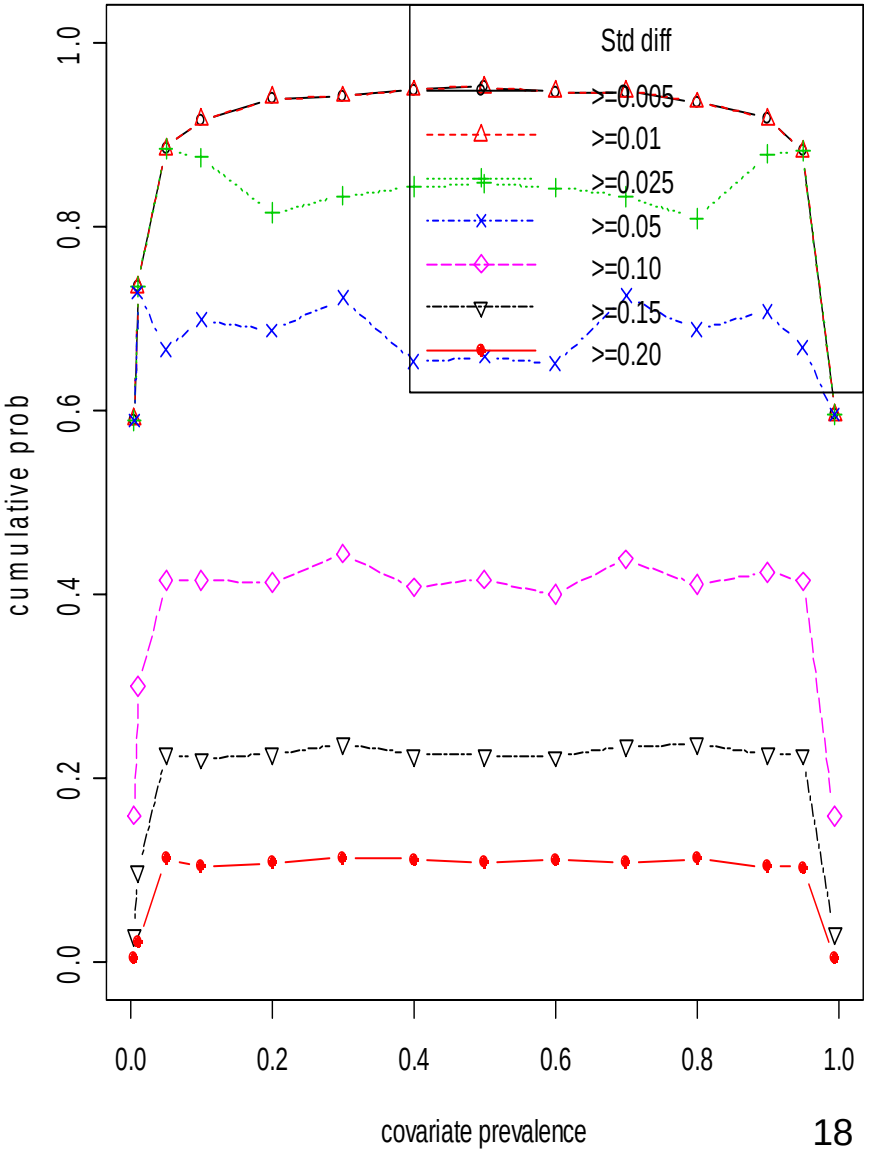


Imb2 (standardized measure)

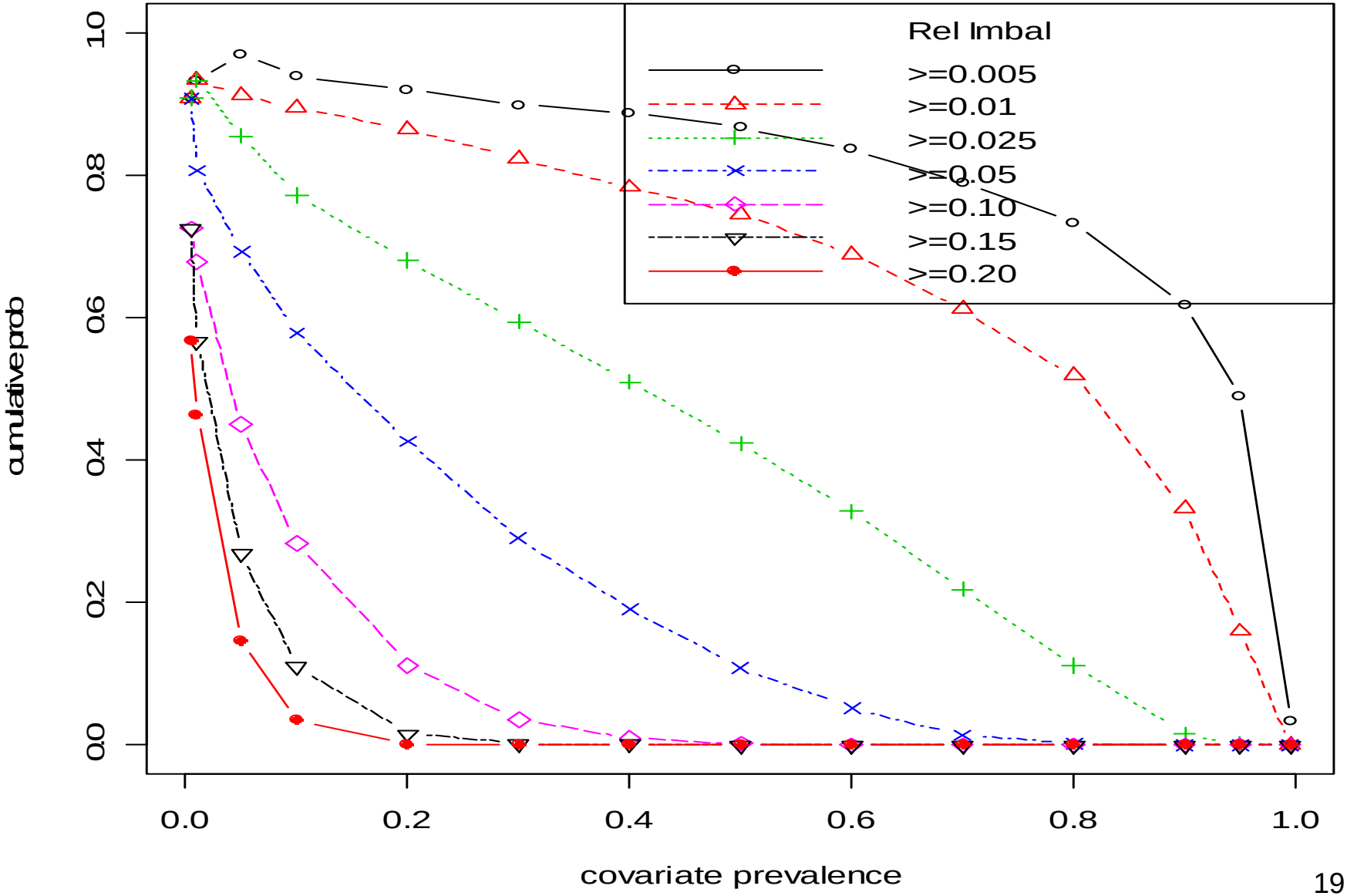
2000 per arm



125 per arm

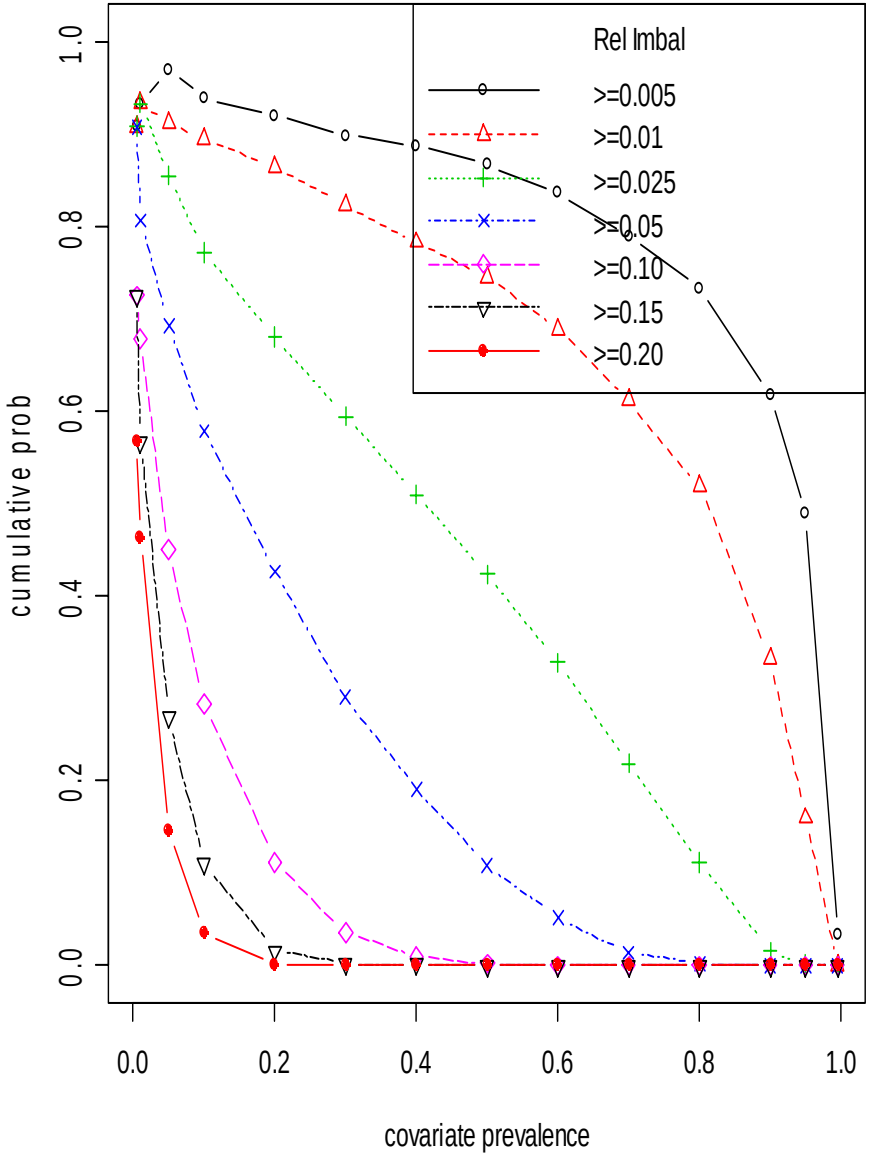


Imb3 (relative measure), 2000 par arm

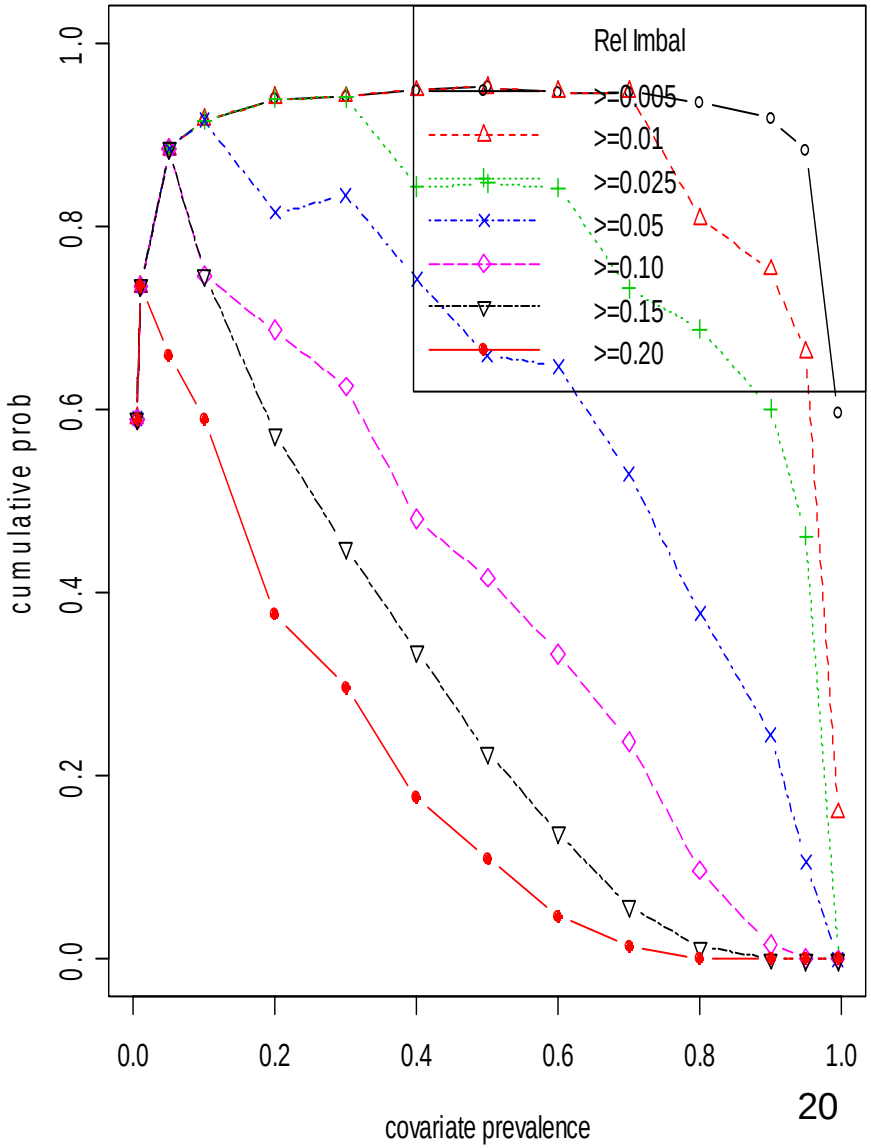


Imb3 (relative measure)

2000 par arm



125 per arm



Likelihood of Prognostic Imbalance (PI)

- Distribution of PI affected by choice of imbalance measure
- PI occurs more frequently as sample size reduces
- Absolute imbalance is more intuitive to understand/calculate → impact of PI on effect estimation
- Other alternatives?
- Which has better statistical property and helps to study the impact of PI (due to the unobserved PF) on Rx effect estimation?

Impact of PI on Effect Estimation

- Simulate binary response Y (in addition to X1, X2)
  - logit(q) = B0 + B1 * X1 + B2 *X2
  - Rx effect: RR|(X2=0)=0.75
  - Control group event rate:10%
  - Strength of prognosis: RR|(X1=0)=5 (X2=1: high risk)
  - Sample size: 1%, 5%, 10%, 25%, 50%,…,100% of adequate sample size (80% power, balanced PF)

Impact of PI on Effect Estimation

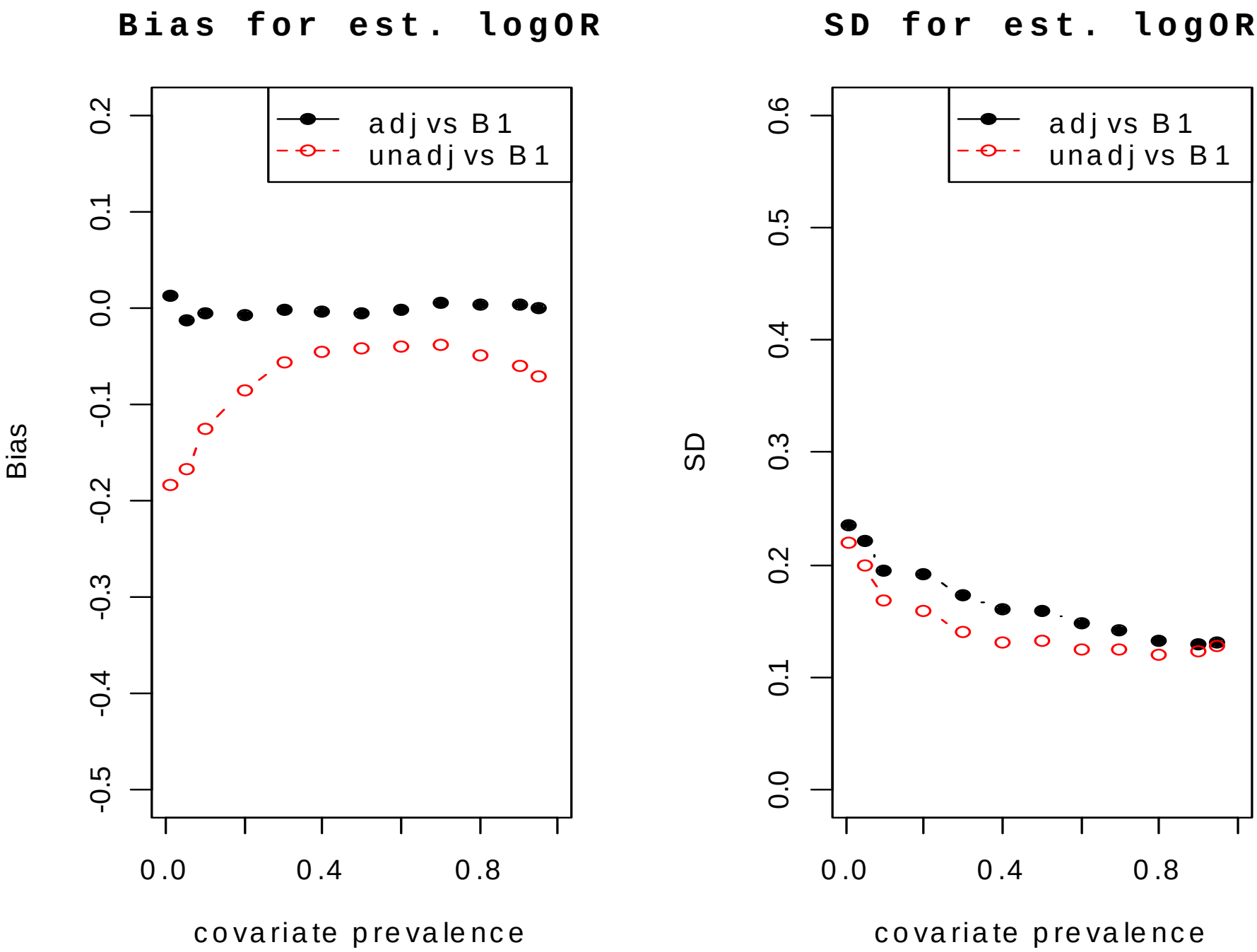
- Unadjusted model: logit(q) = A0 + A1 * X1
- Adjusted model: logit(q) = B0 + B1 * X1 + B2 *X2
- Assessment:
  - Bias, variance, MSE, coverage, empirical power

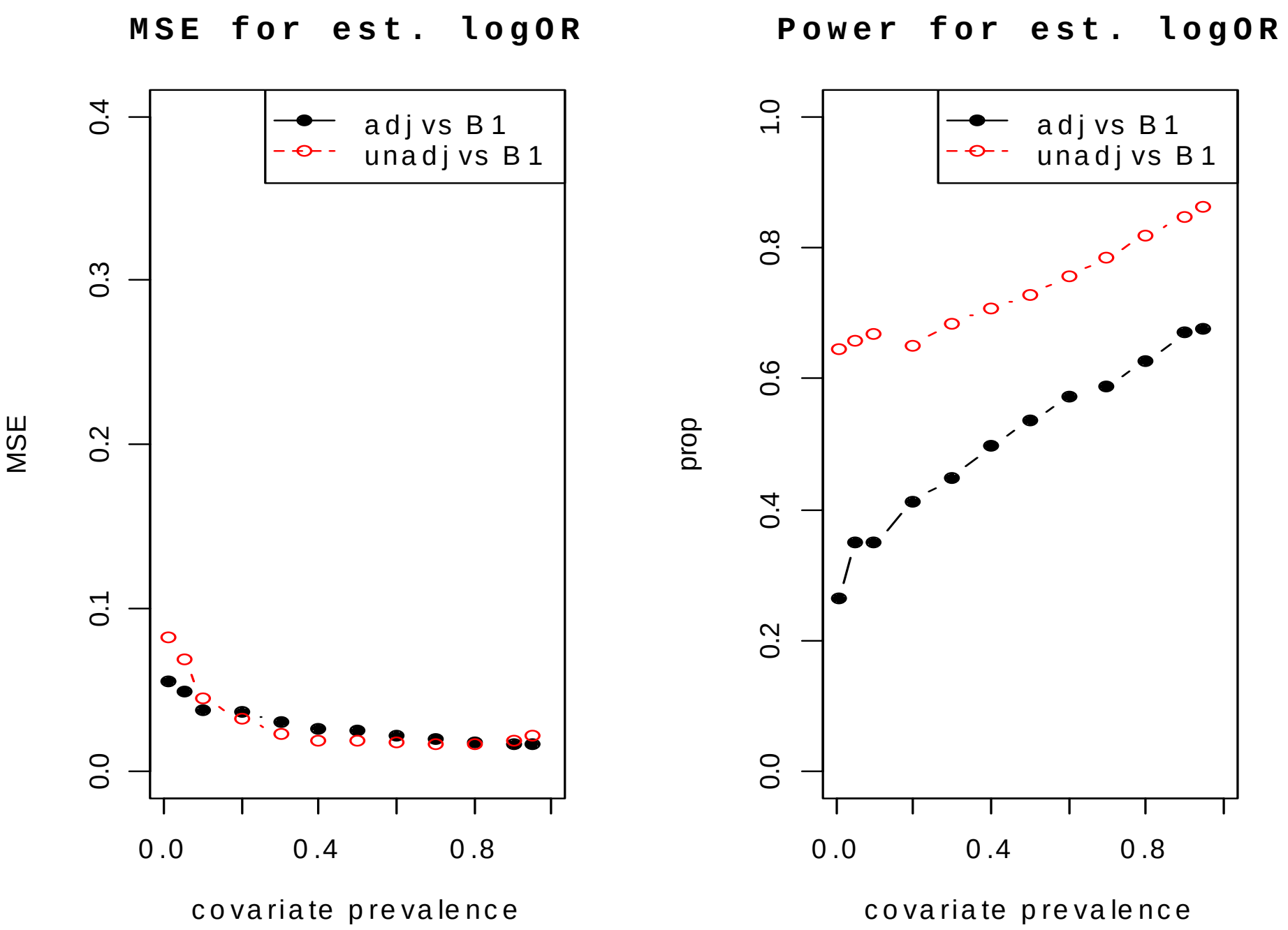
Impact of PI on Effect Estimation

- Unconditional setting
  - X1 fixed by design: two arms equal size
  - X2 ~ Bernoulli(λ) independent of X1
  - Results known (Gail 1984, Robinson & Jewell 1991)

- Conditional setting
  - Option 1: Generate data so that imbalance is fixed at specific levels (to obtain a reasonably large number of datasets to study model misspecification), e.g. 5%, 10%
  - Option 2: select replicates from unconditional setting where imbalance retains a minimum value, e.g. >=5%, >=10%

- 500 patients per arm (25% sufficient sample size)
- 25 more patients having X2=1 in the control group (abs. imbalance = 5%)


Bias for est. logOR
adj vs B1
unadj vs B1
Bias
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
0.0
0.4
0.8
covariate prevalence
SD for est. logOR
adj vs B1
unadj vs B1
SD
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.4
0.8
covariate prevalence


MSE for est. logOR
adj vs B1
unadj vs B1
MSE
0.0
0.1
0.2
0.3
0.4
0.0
0.4
0.8
covariate prevalence
Power for est. logOR
adj vs B1
unadj vs B1
prop
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.4
0.8
covariate prevalence

- Bias increases with effect of X2 on Y and level of imbalance
- Direction of bias depends on direction of the imbalance
- Difference in precision decreases with sample size
- Should measure and adjust for important prognostic factors given a binary outcome

Limitations

- Limited simulation scenarios
- Binary outcome
- One prognostic factor only

Discussion

- How to quantify prognostic imbalance (PI)?
- How to assess impact of imbalance due to unknown/unobserved PI?
- Can we control for impact of such PI in design and analysis phase?

Acknowledgement

- Dr. Stephen Walter
- Dr. Lehana Thabane
- Dr. Gordon Guyatt
- Dr. PJ Devereaux
- Mike Walsh
- Kristian Thorlund

References

- Robinson & Jewell(1991). Some surprising results about covariate adjustment in logistic regression models. Int Stat Rev; 58(2):227-240.
- Gail, Wieand & Piantadosi (1984).Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika;71(3):431-444.

Thank you for your attention!

chur@mcmaster.ca