

Adjusting for selection bias in case control studie

S.Geneletti, S.Richardson, N.Best

Department of Epidemiology and Public Health, Imperial College

24/07/2008

OUTLINE

1. Examples
2. Hypospadias Study
3. What is a DAG?
4. Conditional Independence
5. SB in terms of DAGs
6. Odds ratios
7. Idea
8. Bias Breaking model
9. Hypospadias results
10. Simulations
11. Final Comments

SELECTION BIAS

Basic problem

- Selection bias comes about when there is differential selection of cases and controls
- and a variable that is associated to the exposure under investigation is implicated in the selection process
- Case control studies are particularly prone to this problem
- This is because in order to make valid comparisons the populations of cases and controls must come from the same target population
- It is a problem of internal validity
- We tackle the problem using **DAGs, Conditional independence and extra data**

SELECTION BIAS

Basic problem

- Selection bias comes about when there is differential selection of cases and controls
- and a variable that is associated to the exposure under investigation is implicated in the selection process
- Case control studies are particularly prone to this problem
- This is because in order to make valid comparisons the populations of cases and controls must come from the same target population
- It is a problem of internal validity
- We tackle the problem using **DAGs, Conditional independence and extra data**

SELECTION BIAS

Basic problem

- Selection bias comes about when there is differential selection of cases and controls
- and a variable that is associated to the exposure under investigation is implicated in the selection process
- Case control studies are particularly prone to this problem
- This is because in order to make valid comparisons the populations of cases and controls must come from the same target population
- It is a problem of internal validity
- We tackle the problem using **DAGs, Conditional independence and extra data**

SELECTION BIAS

Basic problem

- Selection bias comes about when there is differential selection of cases and controls
- and a variable that is associated to the exposure under investigation is implicated in the selection process
- Case control studies are particularly prone to this problem
- This is because in order to make valid comparisons the populations of cases and controls must come from the same target population
- It is a problem of internal validity
- We tackle the problem using **DAGs, Conditional independence and extra data**

SELECTION BIAS

Basic problem

- Selection bias comes about when there is differential selection of cases and controls
- and a variable that is associated to the exposure under investigation is implicated in the selection process
- Case control studies are particularly prone to this problem
- This is because in order to make valid comparisons the populations of cases and controls must come from the same target population
- It is a problem of internal validity
- We tackle the problem using **DAGs, Conditional independence and extra data**

SELECTION BIAS

Basic problem

- Selection bias comes about when there is differential selection of cases and controls
- and a variable that is associated to the exposure under investigation is implicated in the selection process
- Case control studies are particularly prone to this problem
- This is because in order to make valid comparisons the populations of cases and controls must come from the same target population
- It is a problem of internal validity
- We tackle the problem using **DAGs, Conditional independence and extra data**

HYPOSPADIAS CASE CONTROL STUDY

Story

- Hypospadias is a congenital malformation of newborn boys
- Is it associated to gestational age or smoking? [4, 5]
- Concern that **controls have a higher SES than cases-selection bias?**
- SES measured using the Carstairs score (C-score) - an area (ward) level index of deprivation ([6])

HYPOSPADIAS CASE CONTROL STUDY

Story

- Hypospadias is a congenital malformation of newborn boys
- Is it associated to gestational age or smoking? [4, 5]
- Concern that **controls have a higher SES than cases-selection bias?**
- SES measured using the Carstairs score (C-score) - an area (ward) level index of deprivation ([6])

HYPOSPADIAS CASE CONTROL STUDY

Story

- Hypospadias is a congenital malformation of newborn boys
- Is it associated to gestational age or smoking? [4, 5]
- Concern that **controls have a higher SES than cases-selection bias?**
- SES measured using the Carstairs score (C-score) - an area (ward) level index of deprivation ([6])

HYPOSPADIAS CASE CONTROL STUDY

Story

- Hypospadias is a congenital malformation of newborn boys
- Is it associated to gestational age or smoking? [4, 5]
- Concern that **controls have a higher SES than cases-selection bias?**
- SES measured using the Carstairs score (C-score) - an area (ward) level index of deprivation ([6])

HYPOSPADIAS CASE CONTROL STUDY

Data collection

- Ward (and hence C-score) and exposure measure of people who participated - **full participants** (indexed by f)
- Ward (and hence C-score) of people who were asked to participate but declined - **partial participants** (indexed by p)
- For partial participants we don't have exposure measure
- Finally, C-score of people who lived in the region the study was conducted **from census**

HYPOSPADIAS CASE CONTROL STUDY

Data collection

- Ward (and hence C-score) and exposure measure of people who participated - **full participants** (indexed by f)
- Ward (and hence C-score) of people who were asked to participate but declined - **partial participants** (indexed by p)
- For partial participants we don't have exposure measure
- Finally, C-score of people who lived in the region the study was conducted **from census**

HYPOSPADIAS CASE CONTROL STUDY

Data collection

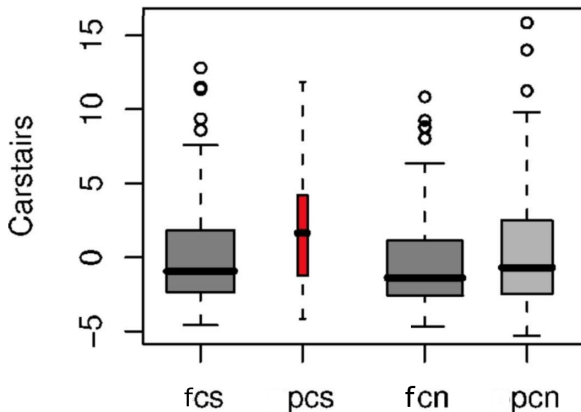
- Ward (and hence C-score) and exposure measure of people who participated - **full participants** (indexed by f)
- Ward (and hence C-score) of people who were asked to participate but declined - **partial participants** (indexed by p)
- For partial participants we don't have exposure measure
- Finally, C-score of people who lived in the region the study was conducted **from census**

HYPOSPADIAS CASE CONTROL STUDY

Data collection

- Ward (and hence C-score) and exposure measure of people who participated - **full participants** (indexed by f)
- Ward (and hence C-score) of people who were asked to participate but declined - **partial participants** (indexed by p)
- For partial participants we don't have exposure measure
- Finally, C-score of people who lived in the region the study was conducted **from census**

BOXPLOT

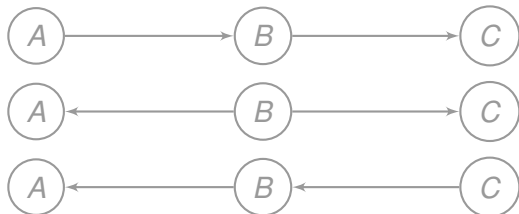


Is there also case selection bias? partial participant cases (pcs) have low SES (high Carstairs)

WHAT IS A DAG?

DAGs are *directed acyclic graphs*

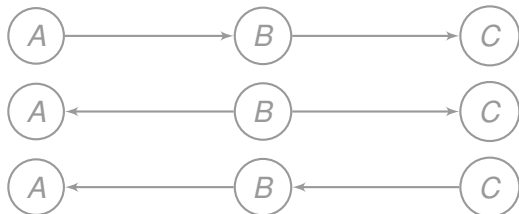
- All arrows have direction
- No cycles $A \rightarrow B \rightarrow A$
- DAGs are used to encode *conditional independence statements*
- $A \perp\!\!\!\perp C | B$ [1] means $p(A, C | B) = p(A | B)p(C | B)$
- Arrows are *not* causal unless extra assumptions made - time ordering, intervention



WHAT IS A DAG?

DAGs are *directed acyclic graphs*

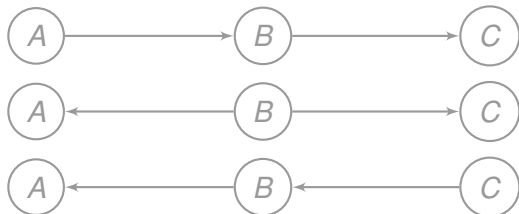
- All arrows have direction
- No cycles $A \rightarrow B \rightarrow A$
- DAGs are used to encode *conditional independence statements*
- $A \perp\!\!\!\perp C | B$ [1] means $p(A, C | B) = p(A | B)p(C | B)$
- Arrows are *not* causal unless extra assumptions made - time ordering, intervention



WHAT IS A DAG?

DAGs are *directed acyclic graphs*

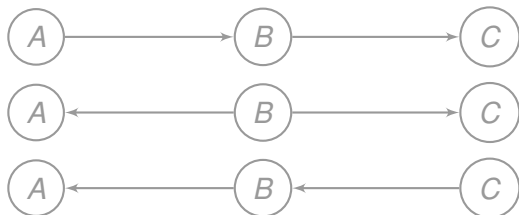
- All arrows have direction
- No cycles $A \rightarrow B \rightarrow A$
- DAGs are used to encode *conditional independence statements*
- $A \perp\!\!\!\perp C | B$ [1] means $p(A, C | B) = p(A | B)p(C | B)$
- Arrows are *not* causal unless extra assumptions made - time ordering, intervention



WHAT IS A DAG?

DAGs are *directed acyclic graphs*

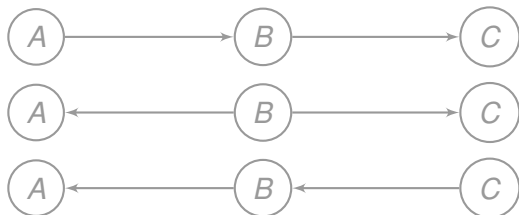
- All arrows have direction
- No cycles $A \rightarrow B \rightarrow A$
- DAGs are used to encode *conditional independence statements*
- $A \perp\!\!\!\perp C | B$ [1] means $p(A, C | B) = p(A | B)p(C | B)$
- Arrows are *not* causal unless extra assumptions made - time ordering, intervention



WHAT IS A DAG?

DAGs are *directed acyclic graphs*

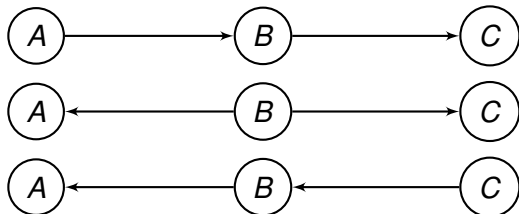
- All arrows have direction
- No cycles $A \rightarrow B \rightarrow A$
- DAGs are used to encode *conditional independence statements*
- $A \perp\!\!\!\perp C | B$ [1] means $p(A, C | B) = p(A | B)p(C | B)$
- Arrows are *not* causal unless extra assumptions made - time ordering, intervention



WHAT IS A DAG?

DAGs are *directed acyclic graphs*

- All arrows have direction
- No cycles $A \rightarrow B \rightarrow A$
- DAGs are used to encode *conditional independence statements*
- $A \perp\!\!\!\perp C | B$ [1] means $p(A, C | B) = p(A | B)p(C | B)$
- Arrows are *not* causal unless extra assumptions made - time ordering, intervention

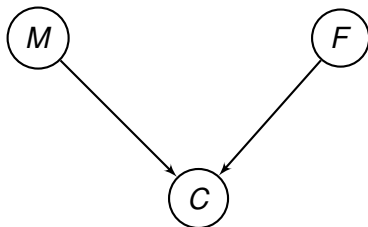


SIMPLE EXAMPLE - INHERITANCE



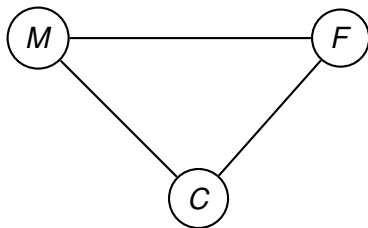
1. Male and female are independent $M \perp\!\!\!\perp F$

SIMPLE EXAMPLE - INHERITANCE



1. Male and female are independent $M \perp\!\!\!\perp F$
2. Then they meet and have a child

SIMPLE EXAMPLE - INHERITANCE

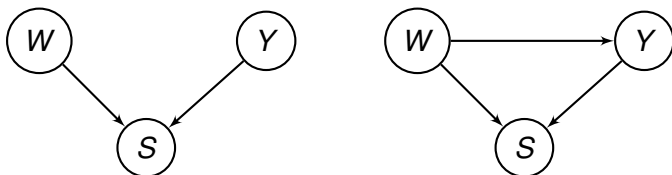


1. Male and female are independent $M \perp\!\!\!\perp F$
2. Then they meet and have a child
3. Now they are dependent through child $M \not\perp F | C$

SELECTION BIAS DAG

Basic premise

Selection bias comes about by conditioning on a common child where we don't know distribution of child given parents



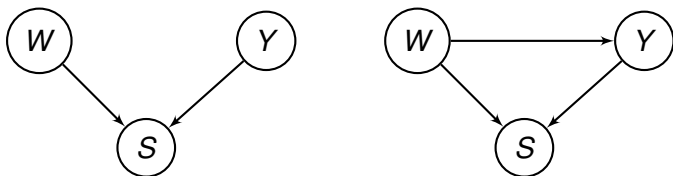
- Y is the outcome of interest, W the exposure, S the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

Problem - we don't know $p(S|Y)$

SELECTION BIAS DAG

Basic premise

Selection bias comes about by conditioning on a common child where we don't know distribution of child given parents



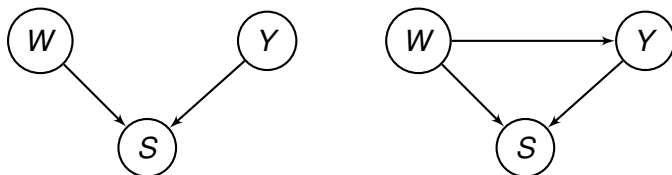
- Y is the outcome of interest, W the exposure, S the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

Problem - we don't know $p(S|Y)$

SELECTION BIAS DAG

Basic premise

Selection bias comes about by conditioning on a common child where we don't know distribution of child given parents



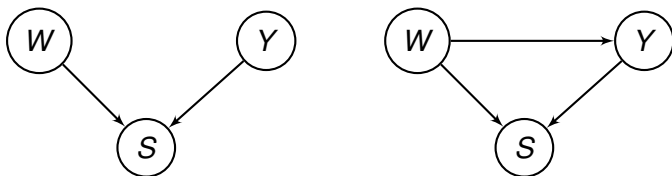
- Y is the outcome of interest, W the exposure, S the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

Problem - we don't know $p(S|Y)$

SELECTION BIAS DAG

Basic premise

Selection bias comes about by conditioning on a common child where we don't know distribution of child given parents



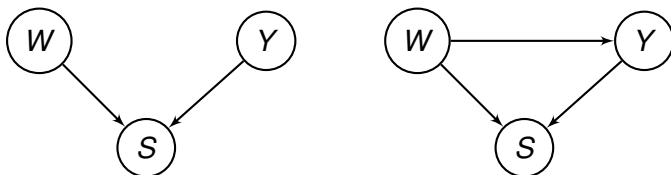
- Y is the outcome of interest, W the exposure, S the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

Problem - we don't know $p(S|Y)$

SELECTION BIAS DAG

Basic premise

Selection bias comes about by conditioning on a common child where we don't know distribution of child given parents



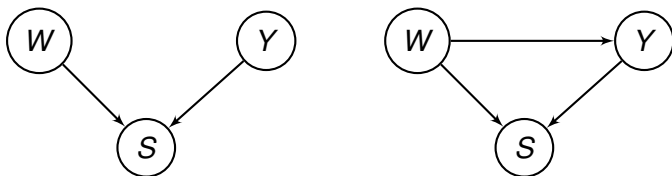
- Y is the outcome of interest, W the exposure, S the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

Problem - we don't know $p(S|Y)$

SELECTION BIAS DAG

Basic premise

Selection bias comes about by conditioning on a common child where we don't know distribution of child given parents



- Y is the outcome of interest, W the exposure, S the selection indicator.
- Left: conditioning induces relationship
- Right: conditioning distorts relationship
- Both share v-structure

Problem - we don't know $p(S|Y)$

ODDS RATIO

True Odds ratio

$$\begin{aligned}\psi &= \frac{p(Y = 1|W = 1)p(Y = 0|W = 0)}{p(Y = 0|W = 1)p(Y = 1|W = 0)} \\ &= \frac{p(W = 1|Y = 1)p(W = 0|Y = 0)}{p(W = 0|Y = 1)p(W = 1|Y = 0)}\end{aligned}\tag{1}$$

ODDS RATIO

True Odds ratio

$$\begin{aligned}\psi &= \frac{p(Y = 1|W = 1)p(Y = 0|W = 0)}{p(Y = 0|W = 1)p(Y = 1|W = 0)} \\ &= \frac{p(W = 1|Y = 1)p(W = 0|Y = 0)}{p(W = 0|Y = 1)p(W = 1|Y = 0)}\end{aligned}\quad (1)$$

Observed Odds ratio

$$\psi^o = \frac{p(Y = 1, W = 1|S = 1)p(Y = 0, W = 0|S = 1)}{p(Y = 0, W = 1|S = 1)p(Y = 1, W = 0|S = 1)}\quad (2)$$

BIAS BREAKING MODEL

- The problem can be addressed if we can find a **bias breaking** variable B
- s.t. we can **separate** exposure W from selection S

A1

$$W \perp\!\!\!\perp S | (Y, B) \quad (3)$$

- This means we can separate the **exposure-disease process** of interest from the **nuisance of the selection process**

A2 Case and control selection are independent

This is usually plausible as case and control recruitment processes are essentially different

Some assumptions for simplicity:

- S1* There is no selection bias in the cases i.e.
 $p(W = 1 | Y = 1, S = 1) = p(W = 1 | Y = 1)$.
- S2* Stratify B if it is not discrete

BIAS BREAKING MODEL

- The problem can be addressed if we can find a **bias breaking** variable B
- s.t. we can **separate** exposure W from selection S

A1

$$W \perp\!\!\!\perp S | (Y, B) \quad (3)$$

- This means we can separate the **exposure-disease process** of interest from the **nuisance of the selection process**

A2 Case and control selection are independent

This is usually plausible as case and control recruitment processes are essentially different

Some assumptions for simplicity:

S1 There is no selection bias in the cases i.e.

$$p(W = 1 | Y = 1, S = 1) = p(W = 1 | Y = 1).$$

S2 Stratify B if it is not discrete

BIAS BREAKING MODEL

- The problem can be addressed if we can find a **bias breaking** variable B
- s.t. we can **separate** exposure W from selection S

A1

$$W \perp\!\!\!\perp S | (Y, B) \quad (3)$$

- This means we can separate the **exposure-disease process** of interest from the **nuisance of the selection process**

A2 Case and control selection are independent

This is usually plausible as case and control recruitment processes are essentially different

Some assumptions for simplicity:

S1 There is no selection bias in the cases i.e.

$$p(W = 1 | Y = 1, S = 1) = p(W = 1 | Y = 1).$$

S2 Stratify B if it is not discrete

BIAS BREAKING MODEL

- The problem can be addressed if we can find a **bias breaking** variable B
- s.t. we can **separate** exposure W from selection S

A1

$$W \perp\!\!\!\perp S | (Y, B) \quad (3)$$

- This means we can separate the **exposure-disease process** of interest from the **nuisance of the selection process**

A2 Case and control selection are independent

This is usually plausible as case and control recruitment processes are essentially different

Some assumptions for simplicity:

S1 There is no selection bias in the cases i.e.

$$p(W = 1 | Y = 1, S = 1) = p(W = 1 | Y = 1).$$

S2 Stratify B if it is not discrete

BIAS BREAKING MODEL

- The problem can be addressed if we can find a **bias breaking** variable B
- s.t. we can **separate** exposure W from selection S

A1

$$W \perp\!\!\!\perp S | (Y, B) \quad (3)$$

- This means we can separate the **exposure-disease process** of interest from the **nuisance of the selection process**

A2 Case and control selection are independent

This is usually plausible as case and control recruitment processes are essentially different

Some assumptions for simplicity:

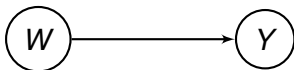
S1 There is no selection bias in the cases i.e.

$$p(W = 1 | Y = 1, S = 1) = p(W = 1 | Y = 1).$$

S2 Stratify B if it is not discrete

IDEA OF “SEPARATION”

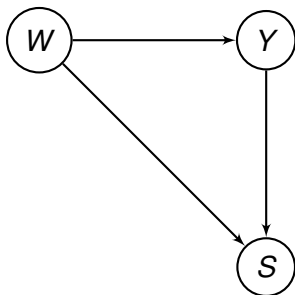
The conditional independence **A1** $W \perp\!\!\!\perp S | (Y, B)$ allows us to



1. separate the exposure disease mechanism of inferential interest
2. from the nuisance selection bias mechanism
3. by using B to separate these mechanisms

IDEA OF “SEPARATION”

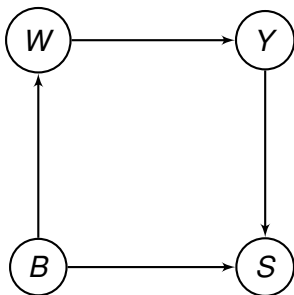
The conditional independence **A1** $W \perp\!\!\!\perp S \mid (Y, B)$ allows us to



1. separate the exposure disease mechanism of inferential interest
2. from the nuisance selection bias mechanism
3. by using B to separate these mechanisms

IDEA OF “SEPARATION”

The conditional independence **A1** $W \perp\!\!\!\perp S \mid (Y, B)$ allows us to



1. separate the exposure disease mechanism of inferential interest
2. from the nuisance selection bias mechanism
3. by using B to separate these mechanisms

BB MODEL

Now we can estimate $p(W = 1|Y = 0)$ as

$$\begin{aligned} p(W|Y = 0, S = 1, B) &= p(W|Y = 0, B) \\ \sum_B p(W|Y = 0, B)p(B|Y = 0) &= p(W|Y = 0) \end{aligned}$$

- **Focus is on finding estimates of $p(B|Y)$** as $p(W|Y, B)$ is estimated by stratum specific proportion of exposed cases/controls
- similar argument can be applied to case selection bias

BB MODEL

Now we can estimate $p(W = 1|Y = 0)$ as

$$p(W|Y = 0, S = 1, B) = p(W|Y = 0, B)$$
$$\sum_B p(W|Y = 0, B)p(B|Y = 0) = p(W|Y = 0)$$

- **Focus is on finding estimates of $p(B|Y)$** as $p(W|Y, B)$ is estimated by stratum specific proportion of exposed cases/controls
- similar argument can be applied to case selection bias

BB MODEL

Now we can estimate $p(W = 1|Y = 0)$ as

$$\begin{aligned} p(W|Y = 0, S = 1, B) &= p(W|Y = 0, B) \\ \sum_B p(W|Y = 0, B)p(B|Y = 0) &= p(W|Y = 0) \end{aligned}$$

- **Focus is on finding estimates of $p(B|Y)$** as $p(W|Y, B)$ is estimated by stratum specific proportion of exposed cases/controls
- similar argument can be applied to case selection bias

BB MODEL

Now we can estimate $p(W = 1|Y = 0)$ as

$$\begin{aligned} p(W|Y = 0, S = 1, B) &= p(W|Y = 0, B) \\ \sum_B p(W|Y = 0, B)p(B|Y = 0) &= p(W|Y = 0) \end{aligned}$$

- **Focus is on finding estimates of $p(B|Y)$** as $p(W|Y, B)$ is estimated by stratum specific proportion of exposed cases/controls
- similar argument can be applied to case selection bias

BB MODEL

Now we can estimate $p(W = 1|Y = 0)$ as

$$\begin{aligned} p(W|Y = 0, S = 1, B) &= p(W|Y = 0, B) \\ \sum_B p(W|Y = 0, B)p(B|Y = 0) &= p(W|Y = 0) \end{aligned}$$

- **Focus is on finding estimates of $p(B|Y)$** as $p(W|Y, B)$ is estimated by stratum specific proportion of exposed cases/controls
- similar argument can be applied to case selection bias

REMEMBER? HYPOSPADIAS CASE CONTROL STUDY

Data collection

- Ward (and hence C-score) and exposure measure of people who participated - **full participants**
- Ward (and hence C -core) of people who were asked to participate but declined - **partial participants**
- Finally, C-score of people who lived in the region the study was conducted **from census**

REMEMBER? HYPOSPADIAS CASE CONTROL STUDY

Data collection

- Ward (and hence C-score) and exposure measure of people who participated - **full participants**
- Ward (and hence C -core) of people who were asked to participate but declined - **partial participants**
- Finally, C-score of people who lived in the region the study was conducted **from census**

ESTIMATES OF $p(B|Y)$ FOR HYPOSP C-C STUDY

There are various options depending on the source of additional data to estimate $p(B|Y)$

Data sources

1. pooling Partial+Full study data on C-score (internal)
2. Census data to estimate regional distr of C-score (external).

ESTIMATES OF $p(B|Y)$ FOR HYPOSP C-C STUDY

There are various options depending on the source of additional data to estimate $p(B|Y)$

Data sources

1. pooling Partial+Full study data on C-score (internal)
2. Census data to estimate regional distr of C-score (external).

... and also on the type of estimate:

Type of estimate

1. Conditional estimate - based on $p(B|Y)$ OR
2. Marginal estimate - based on $p(B)$ - when $p(B|Y = 0) \approx p(B)$.

ESTIMATES OF $p(B|Y)$ FOR HYPOSP C-C STUDY

There are various options depending on the source of additional data to estimate $p(B|Y)$

Data sources

1. pooling Partial+Full study data on C-score (internal)
2. Census data to estimate regional distr of C-score (external).

... and also on the type of estimate:

Type of estimate

1. Conditional estimate - based on $p(B|Y)$ OR
2. Marginal estimate - based on $p(B)$ - when $p(B|Y = 0) \approx p(B)$.

ESTIMATES OF $p(B|Y)$ FOR HYPOSP C-C STUDY

There are various options depending on the source of additional data to estimate $p(B|Y)$

Data sources

1. pooling Partial+Full study data on C-score (internal)
2. Census data to estimate regional distr of C-score (external).

... and also on the type of estimate:

Type of estimate

1. Conditional estimate - based on $p(B|Y)$ OR
2. Marginal estimate - based on $p(B)$ - when $p(B|Y = 0) \approx p(B)$.

ESTIMATES OF $p(B|Y)$ FOR HYPOSP C-C STUDY

There are various options depending on the source of additional data to estimate $p(B|Y)$

Data sources

1. pooling Partial+Full study data on C-score (internal)
2. Census data to estimate regional distr of C-score (external).

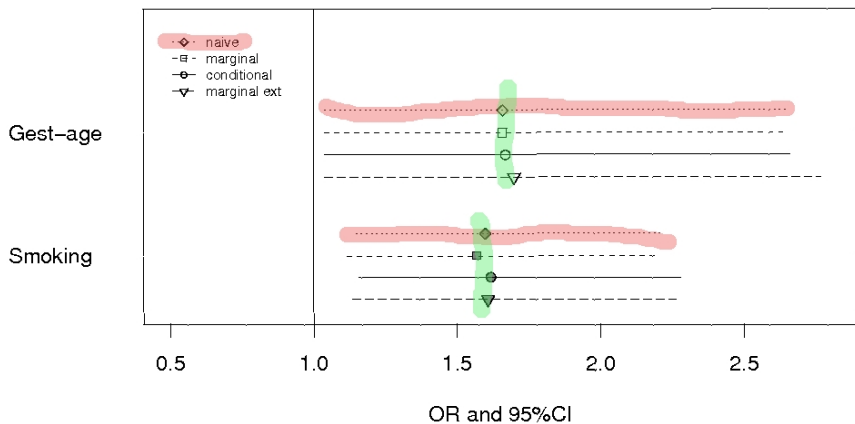
... and also on the type of estimate:

Type of estimate

1. Conditional estimate - based on $p(B|Y)$ OR
2. Marginal estimate - based on $p(B)$ - when $p(B|Y = 0) \approx p(B)$.

RESULTS

OR estimates: naive and adjusted



HYPOSPADIAS CASE CONTROL STUDY

Conclusions

- There appears to be no selection bias mediated by SES
- Naive and adjusted are all very similar
- Do not read too much into small differences
- Validates the study results

HYPOSPADIAS CASE CONTROL STUDY

Conclusions

- There appears to be no selection bias mediated by SES
- Naive and adjusted are all very similar
- Do not read too much into small differences
- Validates the study results

HYPOSPADIAS CASE CONTROL STUDY

Conclusions

- There appears to be no selection bias mediated by SES
- Naive and adjusted are all very similar
- Do not read too much into small differences
- Validates the study results

HYPOSPADIAS CASE CONTROL STUDY

Conclusions

- There appears to be no selection bias mediated by SES
- Naive and adjusted are all very similar
- Do not read too much into small differences
- Validates the study results

SIMULATIONS

Set-up

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, B is also a confounder
- B has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1 | B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

SIMULATIONS

Set-up

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, B is also a confounder
- B has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1|B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

SIMULATIONS

Set-up

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, B is also a confounder
- B has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1|B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

SIMULATIONS

Set-up

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, B is also a confounder
- B has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1|B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

SIMULATIONS

Set-up

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, B is also a confounder
- B has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1|B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

SIMULATIONS

Set-up

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, B is also a confounder
- B has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1|B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

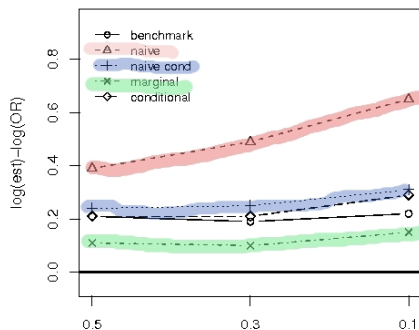
SIMULATIONS

Set-up

- True OR = 1, 2, 2.41 (only show 2 and 2.41)
- When OR=2.41, B is also a confounder
- B has 3 levels - imagine this is SES
- Introduce bias by changing the probability of being selected into study if in 3rd level ($p(S = 1|B = 3)$)
- for different probabilities of being in 3rd level. ($p(B = 3)$)
- Have two simulation studies, one emulates the Hypospadias case-control study with full and partial participants
- The second emulates the Hypospadias case-control study with full participants and census information

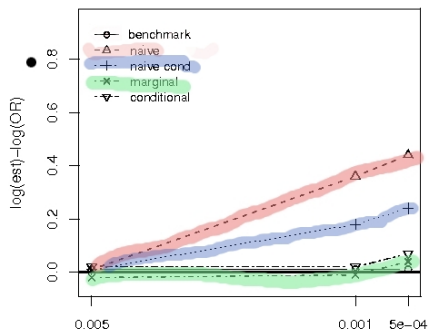
RESULTS

Sim Study 1, true OR=2.41



$p(S=1|B=3, Y=0)$, selection bias increases \rightarrow

Sim study 2, true OR=2



$p(S=1|B=3, Y=0)$, selection bias increases \rightarrow

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

FINAL COMMENTS

Conclusions

1. Our methods adjust well for selection bias
2. Marginal estimators in particular as they use more data than others
3. The estimators do not introduce bias when it is not present
4. Can be used for sensitivity analysis and validation
5. Note that we do not “tamper” with disease or exposure variables
6. Similar to post-stratification [7]
7. In current issue of Biostatistics
8. Have developed Bayesian version
9. Are applying it to EMF data from the US [8]

BIBLIOGRAPHY

- [1] A. P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 41(1):1–31, 1979.
- [2] R.I. Horwitz and A.R. Feinstein. Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine*, 299(20):1089–1094, 1978.
- [3] G. Mezei and L. Kheifets. Selection bias and its implications for case-control studies: a case study of magnetic field exposure and childhood leukaemia. *International Journal of Epidemiology*, 35:397–406, 2006.
- [4] G. Ormond, M.J. Nieuwenhuijsen, P. Nelson, N. Izatt, S. Geneletti, M. Toledano, and P. Elliott. Folate supplementation, endocrine disruptors and hypospadias: case-control study. under review in *BMJ*, 2008.
- [5] M. Nieuwenhuijsen, P. Nelson, and P. Elliott. Occupational exposure of pregnant women in the south east of England. *Epidemiology*, 15(4):S165, 2004.
- [6] V. Carstairs and R. Morris. *Deprivation and Health in Scotland*. Aberdeen University Press, Aberdeen, 1991.
- [7] A. Gelman. Struggles with survey weighting and regression modelling. *Statistical Science*, 22:153–164, 2007.
- [8] E.E. Hatch, R.A. Kleinerman, M.S. Linet, R.E. Tarone, W.T. Kaune, A. Anssi, B. Dasul, L.L. Robison, and S. Wacholder. Do confounding or selection factors of residential wire codings and magnetic fields distort findings of electromagnetic field studies? *Epidemiology*, (11):189–198, 2000.