Causality, confounders, and propensity scores

With examples from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort study at Bristol University, UK
http://www.bristol.ac.uk/alspac/

Roger B. Newson
r.newson@imperial.ac.uk
http://www.imperial.ac.uk/nhli/r.newson/

National Heart and Lung Institute
Imperial College London

Presented at Asthma Club, 11 December, 2008

What is causality?

- Suppose A and B are events. (A might be paracetamol exposure, B might be asthma diagnosis.)
- Then the statement

  "A causes B"

  is shorthand for

  "We can prevent B by **intervening** to prevent A".

- The event A is known as the **cause**, and the event B is known as the **effect**.
- Sometimes, A and B are defined as (mean or median) differences, or (mean, median or odds) ratios, between values of variables.
- For instance, A might be a unit increase in Vitamin D intake (the **exposure**), and B might be the corresponding odds ratio for asthma diagnosis (the **outcome**).
- *However*, causality means nothing unless we have a proposed (or fantasized) intervention in mind.

What is causality?

- Suppose A and B are events. (A might be paracetamol exposure, B might be asthma diagnosis.)
- Then the statement

  "A causes B"

  is shorthand for

  "We can prevent B by **intervening** to prevent A".

- The event A is known as the **cause**, and the event B is known as the **effect**.
- Sometimes, A and B are defined as (mean or median) differences, or (mean, median or odds) ratios, between values of variables.
- For instance, A might be a unit increase in Vitamin D intake (the **exposure**), and B might be the corresponding odds ratio for asthma diagnosis (the **outcome**).
- *However*, causality means nothing unless we have a proposed (or fantasized) intervention in mind.

What is causality?

- Suppose A and B are events. (A might be paracetamol exposure, B might be asthma diagnosis.)
- Then the statement

  "A causes B"

  is shorthand for

  "We can prevent B by **intervening** to prevent A".

- The event A is known as the **cause**, and the event B is known as the **effect**.
- Sometimes, A and B are defined as (mean or median) differences, or (mean, median or odds) ratios, between values of variables.
- For instance, A might be a unit increase in Vitamin D intake (the **exposure**), and B might be the corresponding odds ratio for asthma diagnosis (the **outcome**).
- *However*, causality means nothing unless we have a proposed (or fantasized) intervention in mind.

What is causality?

- Suppose A and B are events. (A might be paracetamol exposure, B might be asthma diagnosis.)
- Then the statement

  "A causes B"

  is shorthand for

  "We can prevent B by **intervening** to prevent A".

- The event A is known as the **cause**, and the event B is known as the **effect**.
- Sometimes, A and B are defined as (mean or median) differences, or (mean, median or odds) ratios, between values of variables.
- For instance, A might be a unit increase in Vitamin D intake (the **exposure**), and B might be the corresponding odds ratio for asthma diagnosis (the **outcome**).
- *However*, causality means nothing unless we have a proposed (or fantasized) intervention in mind.

What is causality?

- Suppose A and B are events. (A might be paracetamol exposure, B might be asthma diagnosis.)
- Then the statement

  "A causes B"

  is shorthand for

  "We can prevent B by **intervening** to prevent A".

- The event A is known as the **cause**, and the event B is known as the **effect**.
- Sometimes, A and B are defined as (mean or median) differences, or (mean, median or odds) ratios, between values of variables.
- For instance, A might be a unit increase in Vitamin D intake (the **exposure**), and B might be the corresponding odds ratio for asthma diagnosis (the **outcome**).
- *However*, causality means nothing unless we have a proposed (or fantasized) intervention in mind.

What is causality?

- Suppose A and B are events. (A might be paracetamol exposure, B might be asthma diagnosis.)
- Then the statement

  "A causes B"

  is shorthand for

  "We can prevent B by **intervening** to prevent A".

- The event A is known as the **cause**, and the event B is known as the **effect**.
- Sometimes, A and B are defined as (mean or median) differences, or (mean, median or odds) ratios, between values of variables.
- For instance, A might be a unit increase in Vitamin D intake (the **exposure**), and B might be the corresponding odds ratio for asthma diagnosis (the **outcome**).
- *However*, causality means nothing unless we have a proposed (or fantasized) intervention in mind.

What is causality?

- Suppose A and B are events. (A might be paracetamol exposure, B might be asthma diagnosis.)
- Then the statement

  "A causes B"

  is shorthand for

  "We can prevent B by **intervening** to prevent A".

- The event A is known as the **cause**, and the event B is known as the **effect**.
- Sometimes, A and B are defined as (mean or median) differences, or (mean, median or odds) ratios, between values of variables.
- For instance, A might be a unit increase in Vitamin D intake (the **exposure**), and B might be the corresponding odds ratio for asthma diagnosis (the **outcome**).
- *However*, causality means nothing unless we have a proposed (or fantasized) intervention in mind.

What are confounders?

Confounders are variables other than the outcome and exposure. There appears not to be a perfect consensus, even among statisticians, on anything else. *However*, I would propose the following informal criteria:

- A confounder is a predictor of "exposure–proneness".
- A confounder is a predictor of "outcome–proneness".
- A confounder is expected to be unchanged by the proposed (or fantasized) intervention.

All 3 of these criteria may be subject to the prior opinions of the investigators. *However*, the aim is that subjects with different exposure levels, and identical confounder levels, should be "exchangeable" (Greenland and Robins, 1986[2]).
Note that confounders defined in this way do not have to be "causally upstream" from the exposure and outcome, but they should *not* be "causally downstream" from either. (See Hernan *et al.*, 2002.[4])

What are confounders?

Confounders are variables other than the outcome and exposure. There appears not to be a perfect consensus, even among statisticians, on anything else. *However*, I would propose the following informal criteria:

- A confounder is a predictor of "exposure–proneness".
- A confounder is a predictor of "outcome–proneness".
- A confounder is expected to be unchanged by the proposed (or fantasized) intervention.

All 3 of these criteria may be subject to the prior opinions of the investigators. *However*, the aim is that subjects with different exposure levels, and identical confounder levels, should be "exchangeable" (Greenland and Robins, 1986[2]).
Note that confounders defined in this way do not have to be "causally upstream" from the exposure and outcome, but they should *not* be "causally downstream" from either. (See Hernan *et al.*, 2002.[4])

What are confounders?

Confounders are variables other than the outcome and exposure. There appears not to be a perfect consensus, even among statisticians, on anything else. *However*, I would propose the following informal criteria:

- A confounder is a predictor of "exposure–proneness".
- A confounder is a predictor of "outcome–proneness".
- A confounder is expected to be unchanged by the proposed (or fantasized) intervention.

All 3 of these criteria may be subject to the prior opinions of the investigators. *However*, the aim is that subjects with different exposure levels, and identical confounder levels, should be "exchangeable" (Greenland and Robins, 1986[2]).
Note that confounders defined in this way do not have to be "causally upstream" from the exposure and outcome, but they should *not* be "causally downstream" from either. (See Hernan *et al.*, 2002.[4])

What are confounders?

Confounders are variables other than the outcome and exposure. There appears not to be a perfect consensus, even among statisticians, on anything else. *However*, I would propose the following informal criteria:

- A confounder is a predictor of "exposure–proneness".
- A confounder is a predictor of "outcome–proneness".
- A confounder is expected to be unchanged by the proposed (or fantasized) intervention.

All 3 of these criteria may be subject to the prior opinions of the investigators. *However*, the aim is that subjects with different exposure levels, and identical confounder levels, should be "exchangeable" (Greenland and Robins, 1986[2]).
Note that confounders defined in this way do not have to be "causally upstream" from the exposure and outcome, but they should *not* be "causally downstream" from either. (See Hernan *et al.*, 2002.[4])

What are confounders?

Confounders are variables other than the outcome and exposure. There appears not to be a perfect consensus, even among statisticians, on anything else. *However*, I would propose the following informal criteria:

- A confounder is a predictor of "exposure–proneness".
- A confounder is a predictor of "outcome–proneness".
- A confounder is expected to be unchanged by the proposed (or fantasized) intervention.

All 3 of these criteria may be subject to the prior opinions of the investigators. *However*, the aim is that subjects with different exposure levels, and identical confounder levels, should be "exchangeable" (Greenland and Robins, 1986[2]).

Note that confounders defined in this way do not have to be "causally upstream" from the exposure and outcome, but they should *not* be "causally downstream" from either. (See Hernan *et al.*, 2002.[4])

What are confounders?

Confounders are variables other than the outcome and exposure.
There appears not to be a perfect consensus, even among statisticians, on anything else. *However*, I would propose the following informal criteria:

- A confounder is a predictor of "exposure–proneness".
- A confounder is a predictor of "outcome–proneness".
- A confounder is expected to be unchanged by the proposed (or fantasized) intervention.

All 3 of these criteria may be subject to the prior opinions of the investigators. *However*, the aim is that subjects with different exposure levels, and identical confounder levels, should be "exchangeable" (Greenland and Robins, 1986[2]).
Note that confounders defined in this way do not have to be "causally upstream" from the exposure and outcome, but they should *not* be "causally downstream" from either. (See Hernan *et al.*, 2002.[4])

Candidate confounders are *not* "innocent until proved guilty"

- In some circles, it is believed that confounders should be excluded from the model, unless they are *shown* to be "doing some confounding", and predicting the exposure *and* the outcome.
- This is typically the principle behind "stepwise" procedures, based *either* on significance levels *or* effect modification, which were critically reviewed by Greenland (1989)[3].
- In general, the theory of confidence intervals *does not* cover us for estimating the parameters in the same data as those in which we found the model.
- In particular, a large number of confounders may each have individually unconvincing effects, which may cumulatively add up to a lot of confounding.
- Davey Smith and Ebrahim (2002)[1] argued that inadequate confounder adjustment is probably more important than "data–dredging" in explaining the apparent high rate of "false positives" in nutritional epidemiology.

Candidate confounders are *not* "innocent until proved guilty"

- In some circles, it is believed that confounders should be excluded from the model, unless they are *shown* to be "doing some confounding", and predicting the exposure *and* the outcome.
- This is typically the principle behind "stepwise" procedures, based *either* on significance levels *or* effect modification, which were critically reviewed by Greenland (1989)[3].
- In general, the theory of confidence intervals *does not* cover us for estimating the parameters in the same data as those in which we found the model.
- In particular, a large number of confounders may each have individually unconvincing effects, which may cumulatively add up to a lot of confounding.
- Davey Smith and Ebrahim (2002)[1] argued that inadequate confounder adjustment is probably more important than "data–dredging" in explaining the apparent high rate of "false positives" in nutritional epidemiology.

Candidate confounders are *not* "innocent until proved guilty"

- In some circles, it is believed that confounders should be excluded from the model, unless they are *shown* to be "doing some confounding", and predicting the exposure *and* the outcome.
- This is typically the principle behind "stepwise" procedures, based *either* on significance levels *or* effect modification, which were critically reviewed by Greenland (1989)[3].
- In general, the theory of confidence intervals *does not* cover us for estimating the parameters in the same data as those in which we found the model.
- In particular, a large number of confounders may each have individually unconvincing effects, which may cumulatively add up to a lot of confounding.
- Davey Smith and Ebrahim (2002)[1] argued that inadequate confounder adjustment is probably more important than "data–dredging" in explaining the apparent high rate of "false positives" in nutritional epidemiology.

Candidate confounders are *not* "innocent until proved guilty"

- In some circles, it is believed that confounders should be excluded from the model, unless they are *shown* to be "doing some confounding", and predicting the exposure *and* the outcome.
- This is typically the principle behind "stepwise" procedures, based *either* on significance levels *or* effect modification, which were critically reviewed by Greenland (1989)[3].
- In general, the theory of confidence intervals *does not* cover us for estimating the parameters in the same data as those in which we found the model.
- In particular, a large number of confounders may each have individually unconvincing effects, which may cumulatively add up to a lot of confounding.
- Davey Smith and Ebrahim (2002)[1] argued that inadequate confounder adjustment is probably more important than "data–dredging" in explaining the apparent high rate of "false positives" in nutritional epidemiology.

Candidate confounders are *not* "innocent until proved guilty"

- In some circles, it is believed that confounders should be excluded from the model, unless they are *shown* to be "doing some confounding", and predicting the exposure *and* the outcome.
- This is typically the principle behind "stepwise" procedures, based *either* on significance levels *or* effect modification, which were critically reviewed by Greenland (1989)[3].
- In general, the theory of confidence intervals *does not* cover us for estimating the parameters in the same data as those in which we found the model.
- In particular, a large number of confounders may each have individually unconvincing effects, which may cumulatively add up to a lot of confounding.
- Davey Smith and Ebrahim (2002)[1] argued that inadequate confounder adjustment is probably more important than "data–dredging" in explaining the apparent high rate of "false positives" in nutritional epidemiology.

Candidate confounders are *not* "innocent until proved guilty"

- In some circles, it is believed that confounders should be excluded from the model, unless they are *shown* to be "doing some confounding", and predicting the exposure *and* the outcome.
- This is typically the principle behind "stepwise" procedures, based *either* on significance levels *or* effect modification, which were critically reviewed by Greenland (1989)[3].
- In general, the theory of confidence intervals *does not* cover us for estimating the parameters in the same data as those in which we found the model.
- In particular, a large number of confounders may each have individually unconvincing effects, which may cumulatively add up to a lot of confounding.
- Davey Smith and Ebrahim (2002)[1] argued that inadequate confounder adjustment is probably more important than "data–dredging" in explaining the apparent high rate of "false positives" in nutritional epidemiology.

Example: Dietary scores and asthma in the ALSPAC cohort

- The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study, based at Bristol University, with 14060 subjects, born in the early 1990s.
- The **outcome** of primary interest is doctor–diagnosed asthma at 7 years of age.
- The **exposures** are 5 dietary patterns (or dietary scores), derived using principal components analysis by Northstone *et al.* (2008)[7] from a food frequency questionnaire (FFQ) completed by the mothers during pregnancy. These scores, expressed in standard deviation (SD) units, were named "Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian".
- The proposed **interventions** are presumably persuading pregnant mothers to eat a diet that generally includes more (or less) of health–conscious, traditional, processed, confectionery, or vegetarian items.
- The **candidate confounder list** was . . .

Example: Dietary scores and asthma in the ALSPAC cohort

- The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study, based at Bristol University, with 14060 subjects, born in the early 1990s.
- The **outcome** of primary interest is doctor–diagnosed asthma at 7 years of age.
- The **exposures** are 5 dietary patterns (or dietary scores), derived using principal components analysis by Northstone *et al.* (2008)[7] from a food frequency questionnaire (FFQ) completed by the mothers during pregnancy. These scores, expressed in standard deviation (SD) units, were named "Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian".
- The proposed **interventions** are presumably persuading pregnant mothers to eat a diet that generally includes more (or less) of health–conscious, traditional, processed, confectionery, or vegetarian items.
- The **candidate confounder list** was . . .

Example: Dietary scores and asthma in the ALSPAC cohort

- The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study, based at Bristol University, with 14060 subjects, born in the early 1990s.
- The **outcome** of primary interest is doctor–diagnosed asthma at 7 years of age.
- The **exposures** are 5 dietary patterns (or dietary scores), derived using principal components analysis by Northstone *et al.* (2008)[7] from a food frequency questionnaire (FFQ) completed by the mothers during pregnancy. These scores, expressed in standard deviation (SD) units, were named "Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian".
- The proposed **interventions** are presumably persuading pregnant mothers to eat a diet that generally includes more (or less) of health–conscious, traditional, processed, confectionery, or vegetarian items.
- The **candidate confounder list** was . . .

Example: Dietary scores and asthma in the ALSPAC cohort

- The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study, based at Bristol University, with 14060 subjects, born in the early 1990s.
- The **outcome** of primary interest is doctor–diagnosed asthma at 7 years of age.
- The **exposures** are 5 dietary patterns (or dietary scores), derived using principal components analysis by Northstone *et al.* (2008)[7] from a food frequency questionnaire (FFQ) completed by the mothers during pregnancy. These scores, expressed in standard deviation (SD) units, were named "Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian".
- The proposed **interventions** are presumably persuading pregnant mothers to eat a diet that generally includes more (or less) of health–conscious, traditional, processed, confectionery, or vegetarian items.
- The **candidate confounder list** was . . .

Example: Dietary scores and asthma in the ALSPAC cohort

- The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study, based at Bristol University, with 14060 subjects, born in the early 1990s.
- The **outcome** of primary interest is doctor–diagnosed asthma at 7 years of age.
- The **exposures** are 5 dietary patterns (or dietary scores), derived using principal components analysis by Northstone *et al.* (2008)[7] from a food frequency questionnaire (FFQ) completed by the mothers during pregnancy. These scores, expressed in standard deviation (SD) units, were named "Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian".
- The proposed **interventions** are presumably persuading pregnant mothers to eat a diet that generally includes more (or less) of health–conscious, traditional, processed, confectionery, or vegetarian items.
- The **candidate confounder list** was . . .

Example: Dietary scores and asthma in the ALSPAC cohort

- The Avon Longitudinal Study of Parents and Children (ALSPAC) is a birth cohort study, based at Bristol University, with 14060 subjects, born in the early 1990s.
- The **outcome** of primary interest is doctor–diagnosed asthma at 7 years of age.
- The **exposures** are 5 dietary patterns (or dietary scores), derived using principal components analysis by Northstone *et al.* (2008)[7] from a food frequency questionnaire (FFQ) completed by the mothers during pregnancy. These scores, expressed in standard deviation (SD) units, were named "Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian".
- The proposed **interventions** are presumably persuading pregnant mothers to eat a diet that generally includes more (or less) of health–conscious, traditional, processed, confectionery, or vegetarian items.
- The **candidate confounder list** was . . .

Candidate confounders

- . . . energy intake, gender, maternal age group, parity, gestational age at birth, prenatal tobacco exposure, maternal education, maternal housing tenure, birth weight, head circumference at birth, length at birth, maternal BMI, maternal ethnic origin, breast feeding in first 6 months, day care in first year, maternal pre–pregnancy atopic disease history (asthma, eczema, rhinoconjunctivitis), maternal infection history during pregnancy (colds/flu, urinary, other), maternal pre–pregnancy migraine history, multiple pregnancy, paracetamol use in late pregnancy, antibiotic use in late pregnancy, pets in first year, damp in home, weekend environmental tobacco exposure in first year, birth season, FFQ completion season, maternal financial difficulties, younger siblings at 7 years, child's BMI at 7 years.
- Most of these are not likely to be "causally upstream" from prenatal diet or asthma, but might indicate aspects of prenatal health or socio–economic state that might influence both. *However . . .*

Candidate confounders

- . . . energy intake, gender, maternal age group, parity, gestational age at birth, prenatal tobacco exposure, maternal education, maternal housing tenure, birth weight, head circumference at birth, length at birth, maternal BMI, maternal ethnic origin, breast feeding in first 6 months, day care in first year, maternal pre–pregnancy atopic disease history (asthma, eczema, rhinoconjunctivitis), maternal infection history during pregnancy (colds/flu, urinary, other), maternal pre–pregnancy migraine history, multiple pregnancy, paracetamol use in late pregnancy, antibiotic use in late pregnancy, pets in first year, damp in home, weekend environmental tobacco exposure in first year, birth season, FFQ completion season, maternal financial difficulties, younger siblings at 7 years, child's BMI at 7 years.
- Most of these are not likely to be "causally upstream" from prenatal diet or asthma, but might indicate aspects of prenatal health or socio–economic state that might influence both. *However . . .*

Candidate confounders

- . . . energy intake, gender, maternal age group, parity, gestational age at birth, prenatal tobacco exposure, maternal education, maternal housing tenure, birth weight, head circumference at birth, length at birth, maternal BMI, maternal ethnic origin, breast feeding in first 6 months, day care in first year, maternal pre–pregnancy atopic disease history (asthma, eczema, rhinoconjunctivitis), maternal infection history during pregnancy (colds/flu, urinary, other), maternal pre–pregnancy migraine history, multiple pregnancy, paracetamol use in late pregnancy, antibiotic use in late pregnancy, pets in first year, damp in home, weekend environmental tobacco exposure in first year, birth season, FFQ completion season, maternal financial difficulties, younger siblings at 7 years, child's BMI at 7 years.
- Most of these are not likely to be "causally upstream" from prenatal diet or asthma, but might indicate aspects of prenatal health or socio–economic state that might influence both. *However . . .*

Candidate confounders (continued)

- . . . it was thought that an intervention to change maternal diet in pregnancy might have incidental effects on some of these candidate confounders (birth weight, head circumference at birth, length at birth, child's BMI at 7 years).
- *Therefore*, we also defined a "non–causal" confounder set, excluding these particular confounders.
- So we had 3 nested alternative confounder sets: "None" (unadjusted), "Non–causal", and "All".

Candidate confounders (continued)

- . . . it was thought that an intervention to change maternal diet in pregnancy might have incidental effects on some of these candidate confounders (birth weight, head circumference at birth, length at birth, child's BMI at 7 years).
- *Therefore*, we also defined a "non–causal" confounder set, excluding these particular confounders.
- So we had 3 nested alternative confounder sets: "None" (unadjusted), "Non–causal", and "All".

Candidate confounders (continued)

- . . . it was thought that an intervention to change maternal diet in pregnancy might have incidental effects on some of these candidate confounders (birth weight, head circumference at birth, length at birth, child's BMI at 7 years).
- *Therefore*, we also defined a "non–causal" confounder set, excluding these particular confounders.
- So we had 3 nested alternative confounder sets: "None" (unadjusted), "Non–causal", and "All".

Candidate confounders (continued)

- . . . it was thought that an intervention to change maternal diet in pregnancy might have incidental effects on some of these candidate confounders (birth weight, head circumference at birth, length at birth, child's BMI at 7 years).
- *Therefore*, we also defined a "non–causal" confounder set, excluding these particular confounders.
- So we had 3 nested alternative confounder sets: "None" (unadjusted), "Non–causal", and "All".

Problem: The confounder space potentially has an infinite number of dimensions!

- The number of possible combinations of confounder values increases *exponentially* with the number of confounders, with each combination representing a small number of subjects.
- It is therefore not practicable to compare only subjects with exactly the same confounder values.
- Assuming a linear, additive model causes parameter numbers to increase *linearly* with the number of confounders.
- *However*, the number of parameters may still be uncomfortably large, possibly affecting the validity of the Central Limit Theorem.
- We would like to reduce the potentially infinite–dimensional confounder space to a manageable number of groups, within which subjects with different exposure levels are comparable.

Problem: The confounder space potentially has an infinite number of dimensions!

- The number of possible combinations of confounder values increases *exponentially* with the number of confounders, with each combination representing a small number of subjects.
- It is therefore not practicable to compare only subjects with exactly the same confounder values.
- Assuming a linear, additive model causes parameter numbers to increase *linearly* with the number of confounders.
- *However*, the number of parameters may still be uncomfortably large, possibly affecting the validity of the Central Limit Theorem.
- We would like to reduce the potentially infinite–dimensional confounder space to a manageable number of groups, within which subjects with different exposure levels are comparable.

Problem: The confounder space potentially has an infinite number of dimensions!

- The number of possible combinations of confounder values increases *exponentially* with the number of confounders, with each combination representing a small number of subjects.
- It is therefore not practicable to compare only subjects with exactly the same confounder values.
- Assuming a linear, additive model causes parameter numbers to increase *linearly* with the number of confounders.
- *However*, the number of parameters may still be uncomfortably large, possibly affecting the validity of the Central Limit Theorem.
- We would like to reduce the potentially infinite–dimensional confounder space to a manageable number of groups, within which subjects with different exposure levels are comparable.

Problem: The confounder space potentially has an infinite number of dimensions!

- The number of possible combinations of confounder values increases *exponentially* with the number of confounders, with each combination representing a small number of subjects.
- It is therefore not practicable to compare only subjects with exactly the same confounder values.
- Assuming a linear, additive model causes parameter numbers to increase *linearly* with the number of confounders.
- *However*, the number of parameters may still be uncomfortably large, possibly affecting the validity of the Central Limit Theorem.
- We would like to reduce the potentially infinite–dimensional confounder space to a manageable number of groups, within which subjects with different exposure levels are comparable.

Problem: The confounder space potentially has an infinite number of dimensions!

- The number of possible combinations of confounder values increases *exponentially* with the number of confounders, with each combination representing a small number of subjects.
- It is therefore not practicable to compare only subjects with exactly the same confounder values.
- Assuming a linear, additive model causes parameter numbers to increase *linearly* with the number of confounders.
- *However*, the number of parameters may still be uncomfortably large, possibly affecting the validity of the Central Limit Theorem.
- We would like to reduce the potentially infinite–dimensional confounder space to a manageable number of groups, within which subjects with different exposure levels are comparable.

Problem: The confounder space potentially has an infinite number of dimensions!

- The number of possible combinations of confounder values increases *exponentially* with the number of confounders, with each combination representing a small number of subjects.
- It is therefore not practicable to compare only subjects with exactly the same confounder values.
- Assuming a linear, additive model causes parameter numbers to increase *linearly* with the number of confounders.
- *However*, the number of parameters may still be uncomfortably large, possibly affecting the validity of the Central Limit Theorem.
- We would like to reduce the potentially infinite–dimensional confounder space to a manageable number of groups, within which subjects with different exposure levels are comparable.

Part of a solution: Propensity scores

- A **propensity score** (Imai and Van Dyk, 2004[5]) is a measure of the "exposure–proneness" of a subject, based on the confounder values.
- It is defined by fitting a statistical model, predicting the exposure from the confounders.
- For each subject, the propensity score is equal to the predicted exposure level for that subject, or sometimes to the corresponding linear predictor.
- Usually, we then group the subjects into a manageable number of similar–sized groups, based on their propensity scores.
- Having found our grouping in the exposure and confounder data, we can then estimate within–group effects of exposure on outcome, based on the within–group association of the outcome variable with the exposure variable.

Part of a solution: Propensity scores

- A **propensity score** (Imai and Van Dyk, 2004[5]) is a measure of the "exposure–proneness" of a subject, based on the confounder values.
- It is defined by fitting a statistical model, predicting the exposure from the confounders.
- For each subject, the propensity score is equal to the predicted exposure level for that subject, or sometimes to the corresponding linear predictor.
- Usually, we then group the subjects into a manageable number of similar–sized groups, based on their propensity scores.
- Having found our grouping in the exposure and confounder data, we can then estimate within–group effects of exposure on outcome, based on the within–group association of the outcome variable with the exposure variable.

Part of a solution: Propensity scores

- A **propensity score** (Imai and Van Dyk, 2004[5]) is a measure of the "exposure–proneness" of a subject, based on the confounder values.
- It is defined by fitting a statistical model, predicting the exposure from the confounders.
- For each subject, the propensity score is equal to the predicted exposure level for that subject, or sometimes to the corresponding linear predictor.
- Usually, we then group the subjects into a manageable number of similar–sized groups, based on their propensity scores.
- Having found our grouping in the exposure and confounder data, we can then estimate within–group effects of exposure on outcome, based on the within–group association of the outcome variable with the exposure variable.

Part of a solution: Propensity scores

- A **propensity score** (Imai and Van Dyk, 2004[5]) is a measure of the "exposure–proneness" of a subject, based on the confounder values.
- It is defined by fitting a statistical model, predicting the exposure from the confounders.
- For each subject, the propensity score is equal to the predicted exposure level for that subject, or sometimes to the corresponding linear predictor.
- Usually, we then group the subjects into a manageable number of similar–sized groups, based on their propensity scores.
- Having found our grouping in the exposure and confounder data, we can then estimate within–group effects of exposure on outcome, based on the within–group association of the outcome variable with the exposure variable.

Part of a solution: Propensity scores

- A **propensity score** (Imai and Van Dyk, 2004[5]) is a measure of the "exposure–proneness" of a subject, based on the confounder values.
- It is defined by fitting a statistical model, predicting the exposure from the confounders.
- For each subject, the propensity score is equal to the predicted exposure level for that subject, or sometimes to the corresponding linear predictor.
- Usually, we then group the subjects into a manageable number of similar–sized groups, based on their propensity scores.
- Having found our grouping in the exposure and confounder data, we can then estimate within–group effects of exposure on outcome, based on the within–group association of the outcome variable with the exposure variable.

Part of a solution: Propensity scores

- A **propensity score** (Imai and Van Dyk, 2004[5]) is a measure of the "exposure–proneness" of a subject, based on the confounder values.
- It is defined by fitting a statistical model, predicting the exposure from the confounders.
- For each subject, the propensity score is equal to the predicted exposure level for that subject, or sometimes to the corresponding linear predictor.
- Usually, we then group the subjects into a manageable number of similar–sized groups, based on their propensity scores.
- Having found our grouping in the exposure and confounder data, we can then estimate within–group effects of exposure on outcome, based on the within–group association of the outcome variable with the exposure variable.

Example: Dietary scores and asthma in the ALSPAC cohort

- In the ALSPAC cohort, the mothers of 12008 children completed the food frequency questionnaire (FFQ), and the 5 dietary pattern scores ("Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian") were computed.
- For each combination of the 5 diet scores and the 2 non–empty confounder sets ("All" and "Non–causal"), a linear regression model was fitted, predicting the diet score from the confounder set.
- The propensity score for each diet score, based on each confounder set, was the predicted value of that diet score, based on the regression model with those confounders.
- The 12008 children were grouped into 64 nearly–equal propensity groups, based on the appropriate propensity score.
- We then fitted a logistic regression model for the outcome "Doctor–diagnosed asthma" (recorded for 7625 subjects), with 64 baseline odds (1 per propensity group), and a common odds ratio for asthma per SD of the diet score.

Example: Dietary scores and asthma in the ALSPAC cohort

- In the ALSPAC cohort, the mothers of 12008 children completed the food frequency questionnaire (FFQ), and the 5 dietary pattern scores ("Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian") were computed.
- For each combination of the 5 diet scores and the 2 non–empty confounder sets ("All" and "Non–causal"), a linear regression model was fitted, predicting the diet score from the confounder set.
- The propensity score for each diet score, based on each confounder set, was the predicted value of that diet score, based on the regression model with those confounders.
- The 12008 children were grouped into 64 nearly–equal propensity groups, based on the appropriate propensity score.
- We then fitted a logistic regression model for the outcome "Doctor–diagnosed asthma" (recorded for 7625 subjects), with 64 baseline odds (1 per propensity group), and a common odds ratio for asthma per SD of the diet score.

Example: Dietary scores and asthma in the ALSPAC cohort

- In the ALSPAC cohort, the mothers of 12008 children completed the food frequency questionnaire (FFQ), and the 5 dietary pattern scores ("Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian") were computed.
- For each combination of the 5 diet scores and the 2 non–empty confounder sets ("All" and "Non–causal"), a linear regression model was fitted, predicting the diet score from the confounder set.
- The propensity score for each diet score, based on each confounder set, was the predicted value of that diet score, based on the regression model with those confounders.
- The 12008 children were grouped into 64 nearly–equal propensity groups, based on the appropriate propensity score.
- We then fitted a logistic regression model for the outcome "Doctor–diagnosed asthma" (recorded for 7625 subjects), with 64 baseline odds (1 per propensity group), and a common odds ratio for asthma per SD of the diet score.

Example: Dietary scores and asthma in the ALSPAC cohort

- In the ALSPAC cohort, the mothers of 12008 children completed the food frequency questionnaire (FFQ), and the 5 dietary pattern scores ("Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian") were computed.
- For each combination of the 5 diet scores and the 2 non–empty confounder sets ("All" and "Non–causal"), a linear regression model was fitted, predicting the diet score from the confounder set.
- The propensity score for each diet score, based on each confounder set, was the predicted value of that diet score, based on the regression model with those confounders.
- The 12008 children were grouped into 64 nearly–equal propensity groups, based on the appropriate propensity score.
- We then fitted a logistic regression model for the outcome "Doctor–diagnosed asthma" (recorded for 7625 subjects), with 64 baseline odds (1 per propensity group), and a common odds ratio for asthma per SD of the diet score.

Example: Dietary scores and asthma in the ALSPAC cohort

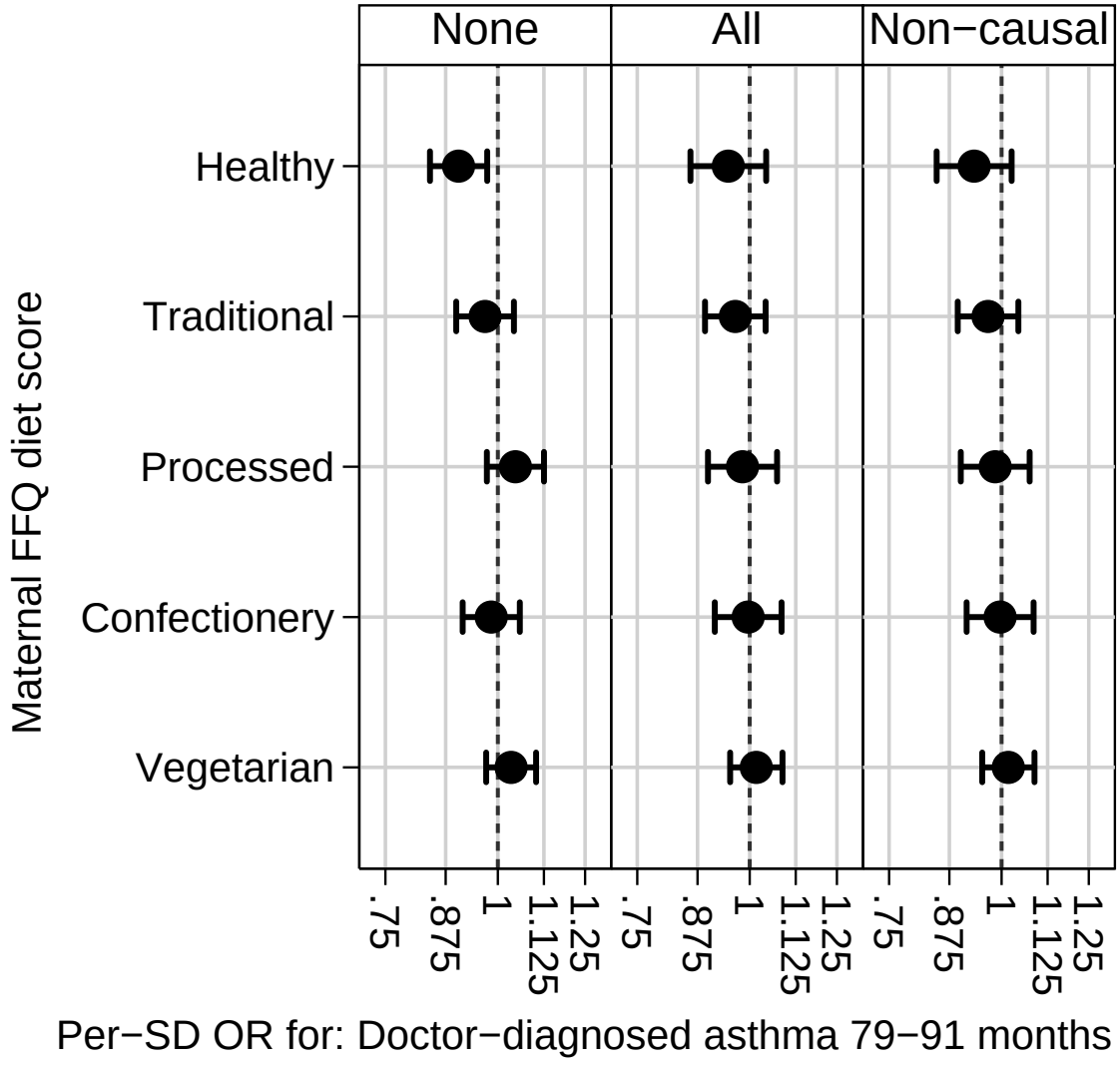
- In the ALSPAC cohort, the mothers of 12008 children completed the food frequency questionnaire (FFQ), and the 5 dietary pattern scores ("Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian") were computed.
- For each combination of the 5 diet scores and the 2 non–empty confounder sets ("All" and "Non–causal"), a linear regression model was fitted, predicting the diet score from the confounder set.
- The propensity score for each diet score, based on each confounder set, was the predicted value of that diet score, based on the regression model with those confounders.
- The 12008 children were grouped into 64 nearly–equal propensity groups, based on the appropriate propensity score.
- We then fitted a logistic regression model for the outcome "Doctor–diagnosed asthma" (recorded for 7625 subjects), with 64 baseline odds (1 per propensity group), and a common odds ratio for asthma per SD of the diet score.

Example: Dietary scores and asthma in the ALSPAC cohort

- In the ALSPAC cohort, the mothers of 12008 children completed the food frequency questionnaire (FFQ), and the 5 dietary pattern scores ("Health–conscious", "Traditional", "Processed", "Confectionery", and "Vegetarian") were computed.
- For each combination of the 5 diet scores and the 2 non–empty confounder sets ("All" and "Non–causal"), a linear regression model was fitted, predicting the diet score from the confounder set.
- The propensity score for each diet score, based on each confounder set, was the predicted value of that diet score, based on the regression model with those confounders.
- The 12008 children were grouped into 64 nearly–equal propensity groups, based on the appropriate propensity score.
- We then fitted a logistic regression model for the outcome "Doctor–diagnosed asthma" (recorded for 7625 subjects), with 64 baseline odds (1 per propensity group), and a common odds ratio for asthma per SD of the diet score.

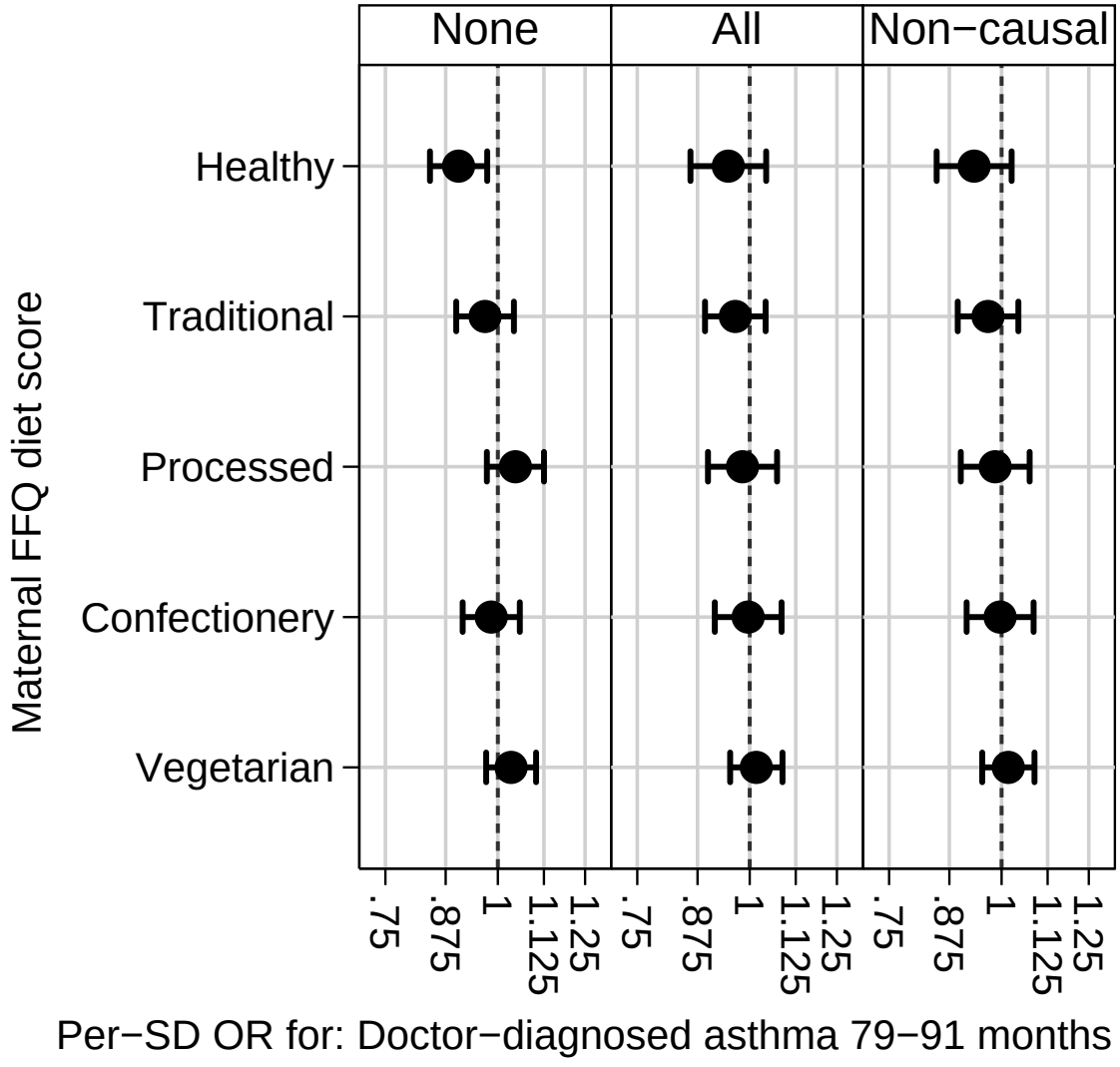
Odds ratios for asthma per standard–deviation change in diet scores

- Increasing the "Healthy" score by 1 SD seems to predict a modest fall in asthma odds (OR=0.90).
- But if we compare only "comparable" subjects, the OR is closer to 1, and the confidence interval is wider.
- *Therefore*, the association *may* arise because more "Healthy–diet–prone" subjects are also less "asthma–prone", whatever their mothers eat.



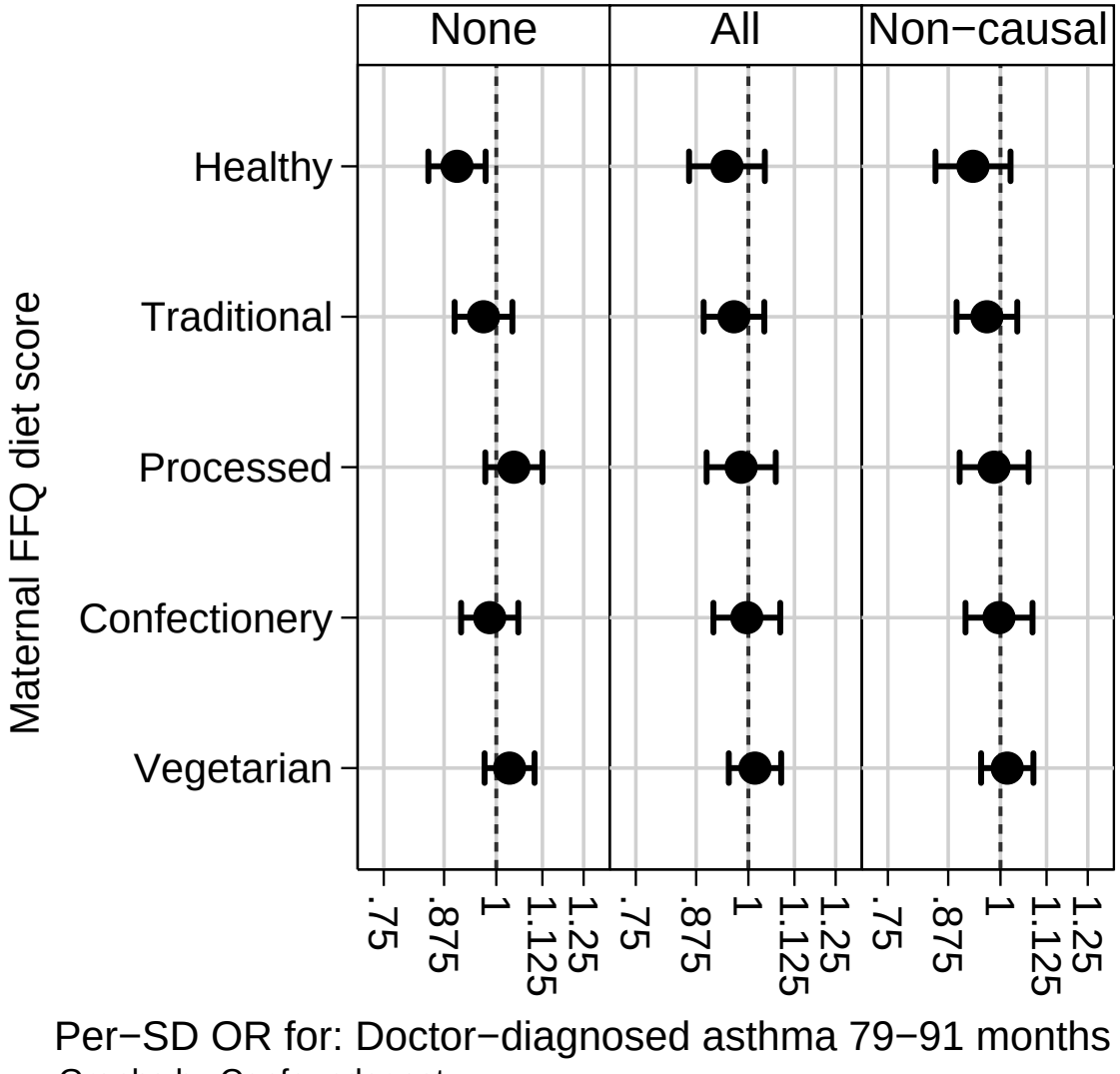
Odds ratios for asthma per standard–deviation change in diet scores

- Increasing the "Healthy" score by 1 SD seems to predict a modest fall in asthma odds (OR=0.90).
- But if we compare only "comparable" subjects, the OR is closer to 1, and the confidence interval is wider.
- *Therefore*, the association *may* arise because more "Healthy–diet–prone" subjects are also less "asthma–prone", whatever their mothers eat.



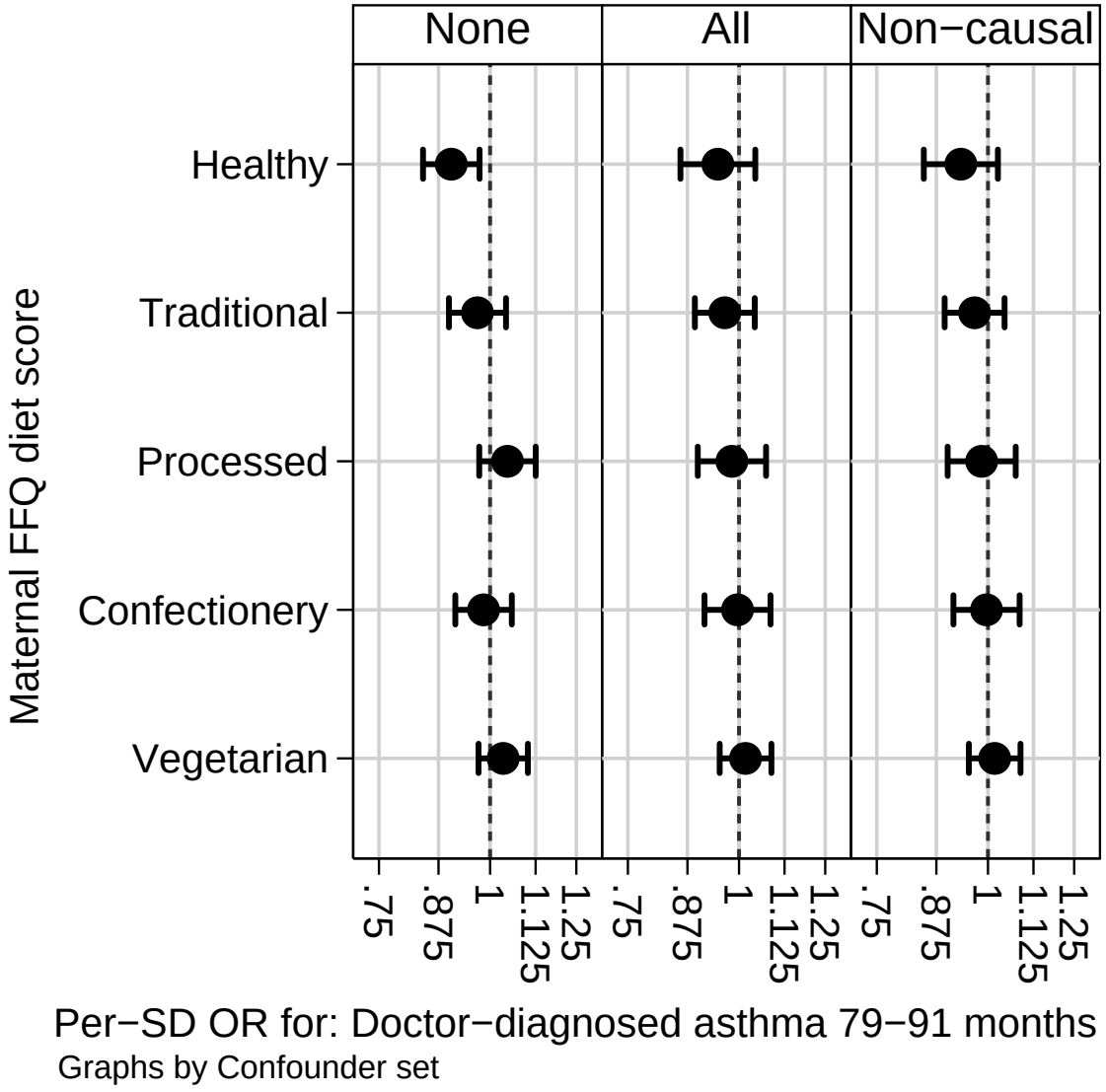
Odds ratios for asthma per standard–deviation change in diet scores

- Increasing the "Healthy" score by 1 SD seems to predict a modest fall in asthma odds (OR=0.90).
- But if we compare only "comparable" subjects, the OR is closer to 1, and the confidence interval is wider.
- *Therefore*, the association *may* arise because more "Healthy–diet–prone" subjects are also less "asthma–prone", whatever their mothers eat.



Odds ratios for asthma per standard–deviation change in diet scores

- Increasing the "Healthy" score by 1 SD seems to predict a modest fall in asthma odds (OR=0.90).
- But if we compare only "comparable" subjects, the OR is closer to 1, and the confidence interval is wider.
- *Therefore*, the association *may* arise because more "Healthy–diet–prone" subjects are also less "asthma–prone", whatever their mothers eat.



Advantages of propensity scores

- Statistical methods assume 2 “quasi–independent” stages in an analysis: “Choosing the parameters to estimate” and “Estimating the parameters”.
- It is *implicitly assumed* that we find the model in a different set of data from the dataset used to estimate its parameters.
- Using propensity groups, we find the model in the exposure and confounder data, and measure its parameters in the outcome data.
- The confidence interval formulas are based on the conditional distribution of the outcomes, given the exposures and confounders.
- This rigorous separation of the 2 stages implies that we can do various “interactive” checks on the joint distribution of the exposure, confounders and propensity groups, *before* looking at the outcomes. For example . . .

Advantages of propensity scores

- Statistical methods assume 2 “quasi–independent” stages in an analysis: “Choosing the parameters to estimate” and “Estimating the parameters”.
- It is *implicitly assumed* that we find the model in a different set of data from the dataset used to estimate its parameters.
- Using propensity groups, we find the model in the exposure and confounder data, and measure its parameters in the outcome data.
- The confidence interval formulas are based on the conditional distribution of the outcomes, given the exposures and confounders.
- This rigorous separation of the 2 stages implies that we can do various “interactive” checks on the joint distribution of the exposure, confounders and propensity groups, *before* looking at the outcomes. For example . . .

Advantages of propensity scores

- Statistical methods assume 2 “quasi–independent” stages in an analysis: “Choosing the parameters to estimate” and “Estimating the parameters”.
- It is *implicitly assumed* that we find the model in a different set of data from the dataset used to estimate its parameters.
- Using propensity groups, we find the model in the exposure and confounder data, and measure its parameters in the outcome data.
- The confidence interval formulas are based on the conditional distribution of the outcomes, given the exposures and confounders.
- This rigorous separation of the 2 stages implies that we can do various “interactive” checks on the joint distribution of the exposure, confounders and propensity groups, *before* looking at the outcomes. For example . . .

Advantages of propensity scores

- Statistical methods assume 2 “quasi–independent” stages in an analysis: “Choosing the parameters to estimate” and “Estimating the parameters”.
- It is *implicitly assumed* that we find the model in a different set of data from the dataset used to estimate its parameters.
- Using propensity groups, we find the model in the exposure and confounder data, and measure its parameters in the outcome data.
- The confidence interval formulas are based on the conditional distribution of the outcomes, given the exposures and confounders.
- This rigorous separation of the 2 stages implies that we can do various “interactive” checks on the joint distribution of the exposure, confounders and propensity groups, *before* looking at the outcomes. For example . . .

Advantages of propensity scores

- Statistical methods assume 2 “quasi–independent” stages in an analysis: “Choosing the parameters to estimate” and “Estimating the parameters”.
- It is *implicitly assumed* that we find the model in a different set of data from the dataset used to estimate its parameters.
- Using propensity groups, we find the model in the exposure and confounder data, and measure its parameters in the outcome data.
- The confidence interval formulas are based on the conditional distribution of the outcomes, given the exposures and confounders.
- This rigorous separation of the 2 stages implies that we can do various “interactive” checks on the joint distribution of the exposure, confounders and propensity groups, *before* looking at the outcomes. For example . . .

Advantages of propensity scores

- Statistical methods assume 2 "quasi–independent" stages in an analysis: "Choosing the parameters to estimate" and "Estimating the parameters".
- It is *implicitly assumed* that we find the model in a different set of data from the dataset used to estimate its parameters.
- Using propensity groups, we find the model in the exposure and confounder data, and measure its parameters in the outcome data.
- The confidence interval formulas are based on the conditional distribution of the outcomes, given the exposures and confounders.
- This rigorous separation of the 2 stages implies that we can do various "interactive" checks on the joint distribution of the exposure, confounders and propensity groups, *before* looking at the outcomes. For example . . .

Choosing the number of propensity groups

- Given an exposure, and a candidate confounder list, we would like to choose a sensible number of propensity groups.
- If there are too few, then there will be residual confounding.
- If there are too many, then there will be too few within–group comparisons between subjects with different exposure levels (losing power).
- The power of the propensity score to predict the exposure may be measured by calculating Somers' *D* of the propensity score with respect to the exposure (Newson, 2006[6]).
- Somers' *D* has a standard unstratified version, including all pairs of subjects with different exposure levels.
- It also has a stratified version, restricted to pairs of subjects with different exposure levels within the same propensity group.

Choosing the number of propensity groups

- Given an exposure, and a candidate confounder list, we would like to choose a sensible number of propensity groups.
- If there are too few, then there will be residual confounding.
- If there are too many, then there will be too few within–group comparisons between subjects with different exposure levels (losing power).
- The power of the propensity score to predict the exposure may be measured by calculating Somers' *D* of the propensity score with respect to the exposure (Newson, 2006[6]).
- Somers' *D* has a standard unstratified version, including all pairs of subjects with different exposure levels.
- It also has a stratified version, restricted to pairs of subjects with different exposure levels within the same propensity group.

Choosing the number of propensity groups

- Given an exposure, and a candidate confounder list, we would like to choose a sensible number of propensity groups.
- If there are too few, then there will be residual confounding.
- If there are too many, then there will be too few within–group comparisons between subjects with different exposure levels (losing power).
- The power of the propensity score to predict the exposure may be measured by calculating Somers' *D* of the propensity score with respect to the exposure (Newson, 2006[6]).
- Somers' *D* has a standard unstratified version, including all pairs of subjects with different exposure levels.
- It also has a stratified version, restricted to pairs of subjects with different exposure levels within the same propensity group.

Choosing the number of propensity groups

- Given an exposure, and a candidate confounder list, we would like to choose a sensible number of propensity groups.
- If there are too few, then there will be residual confounding.
- If there are too many, then there will be too few within–group comparisons between subjects with different exposure levels (losing power).
- The power of the propensity score to predict the exposure may be measured by calculating Somers' *D* of the propensity score with respect to the exposure (Newson, 2006[6]).
- Somers' *D* has a standard unstratified version, including all pairs of subjects with different exposure levels.
- It also has a stratified version, restricted to pairs of subjects with different exposure levels within the same propensity group.

Choosing the number of propensity groups

- Given an exposure, and a candidate confounder list, we would like to choose a sensible number of propensity groups.
- If there are too few, then there will be residual confounding.
- If there are too many, then there will be too few within–group comparisons between subjects with different exposure levels (losing power).
- The power of the propensity score to predict the exposure may be measured by calculating Somers' D of the propensity score with respect to the exposure (Newson, 2006[6]).
- Somers' D has a standard unstratified version, including all pairs of subjects with different exposure levels.
- It also has a stratified version, restricted to pairs of subjects with different exposure levels within the same propensity group.

Choosing the number of propensity groups

- Given an exposure, and a candidate confounder list, we would like to choose a sensible number of propensity groups.
- If there are too few, then there will be residual confounding.
- If there are too many, then there will be too few within–group comparisons between subjects with different exposure levels (losing power).
- The power of the propensity score to predict the exposure may be measured by calculating Somers' D of the propensity score with respect to the exposure (Newson, 2006[6]).
- Somers' D has a standard unstratified version, including all pairs of subjects with different exposure levels.
- It also has a stratified version, restricted to pairs of subjects with different exposure levels within the same propensity group.

Choosing the number of propensity groups

- Given an exposure, and a candidate confounder list, we would like to choose a sensible number of propensity groups.
- If there are too few, then there will be residual confounding.
- If there are too many, then there will be too few within–group comparisons between subjects with different exposure levels (losing power).
- The power of the propensity score to predict the exposure may be measured by calculating Somers' D of the propensity score with respect to the exposure (Newson, 2006[6]).
- Somers' D has a standard unstratified version, including all pairs of subjects with different exposure levels.
- It also has a stratified version, restricted to pairs of subjects with different exposure levels within the same propensity group.

Example: Diet scores in the ALSPAC study

- We wanted to choose a sensible number of propensity groups to stratify the propensity score for each diet score.
- For each of the 5 diet scores, we used the unstratified Somers' D of its propensity score with respect to the diet score to measure the power of the confounders (collectively) to predict the diet score.
- We used the group–stratified Somers' D to measure the same predictive power, *within* propensity groups.
- We tried various numbers of propensity groups, from 2 to 128.
- Increasing the number of propensity groups decreased the stratified Somers' D. With 64 propensity groups, little of the original association remained.

Example: Diet scores in the ALSPAC study

- We wanted to choose a sensible number of propensity groups to stratify the propensity score for each diet score.
- For each of the 5 diet scores, we used the unstratified Somers' D of its propensity score with respect to the diet score to measure the power of the confounders (collectively) to predict the diet score.
- We used the group–stratified Somers' D to measure the same predictive power, *within* propensity groups.
- We tried various numbers of propensity groups, from 2 to 128.
- Increasing the number of propensity groups decreased the stratified Somers' D. With 64 propensity groups, little of the original association remained.

Example: Diet scores in the ALSPAC study

- We wanted to choose a sensible number of propensity groups to stratify the propensity score for each diet score.
- For each of the 5 diet scores, we used the unstratified Somers' D of its propensity score with respect to the diet score to measure the power of the confounders (collectively) to predict the diet score.
- We used the group–stratified Somers' D to measure the same predictive power, *within* propensity groups.
- We tried various numbers of propensity groups, from 2 to 128.
- Increasing the number of propensity groups decreased the stratified Somers' D. With 64 propensity groups, little of the original association remained.

Example: Diet scores in the ALSPAC study

- We wanted to choose a sensible number of propensity groups to stratify the propensity score for each diet score.
- For each of the 5 diet scores, we used the unstratified Somers' D of its propensity score with respect to the diet score to measure the power of the confounders (collectively) to predict the diet score.
- We used the group–stratified Somers' D to measure the same predictive power, *within* propensity groups.
- We tried various numbers of propensity groups, from 2 to 128.
- Increasing the number of propensity groups decreased the stratified Somers' D. With 64 propensity groups, little of the original association remained.

Example: Diet scores in the ALSPAC study

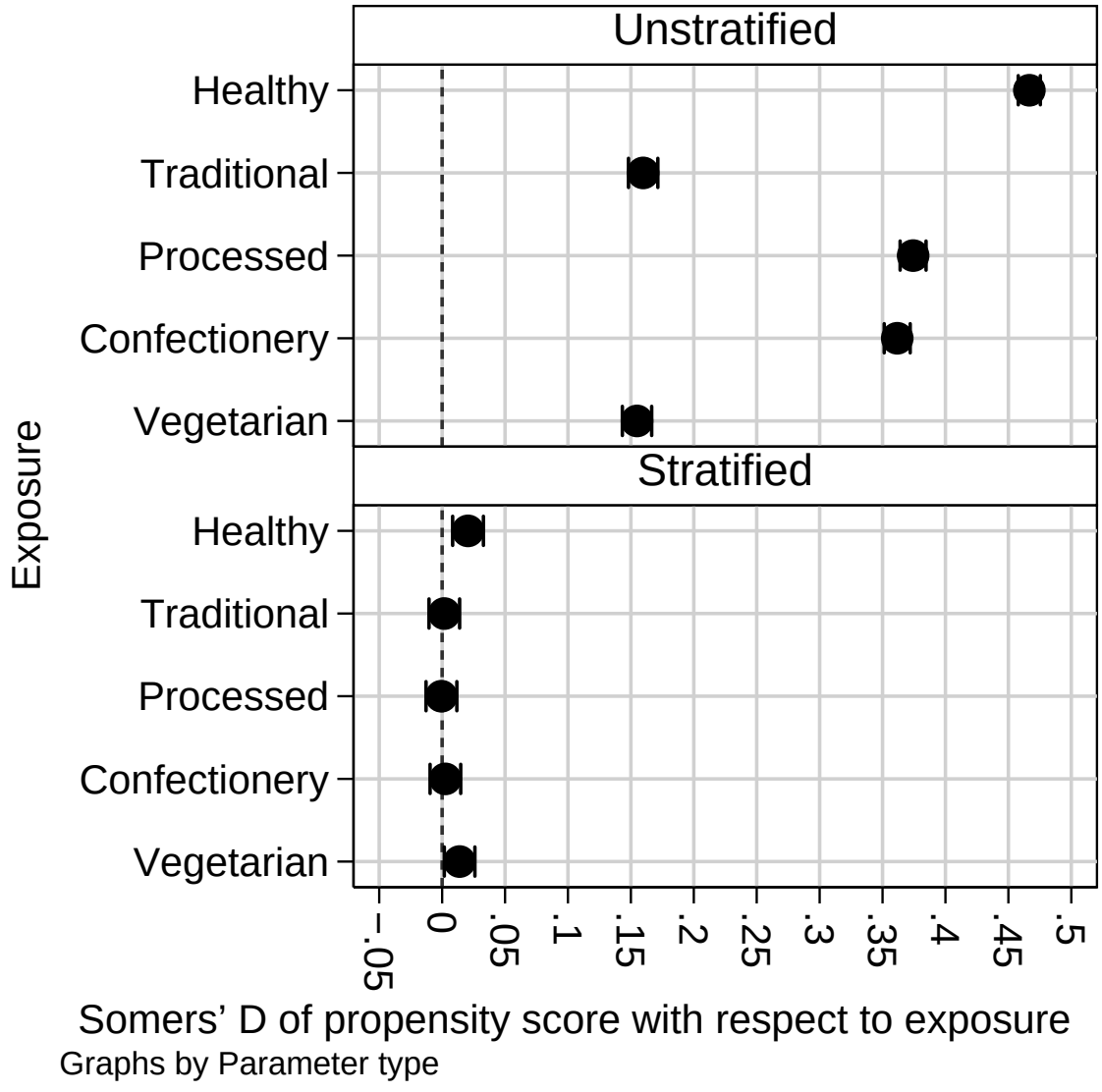
- We wanted to choose a sensible number of propensity groups to stratify the propensity score for each diet score.
- For each of the 5 diet scores, we used the unstratified Somers' D of its propensity score with respect to the diet score to measure the power of the confounders (collectively) to predict the diet score.
- We used the group–stratified Somers' D to measure the same predictive power, *within* propensity groups.
- We tried various numbers of propensity groups, from 2 to 128.
- Increasing the number of propensity groups decreased the stratified Somers' D. With 64 propensity groups, little of the original association remained.

Example: Diet scores in the ALSPAC study

- We wanted to choose a sensible number of propensity groups to stratify the propensity score for each diet score.
- For each of the 5 diet scores, we used the unstratified Somers' D of its propensity score with respect to the diet score to measure the power of the confounders (collectively) to predict the diet score.
- We used the group–stratified Somers' D to measure the same predictive power, *within* propensity groups.
- We tried various numbers of propensity groups, from 2 to 128.
- Increasing the number of propensity groups decreased the stratified Somers' D. With 64 propensity groups, little of the original association remained.

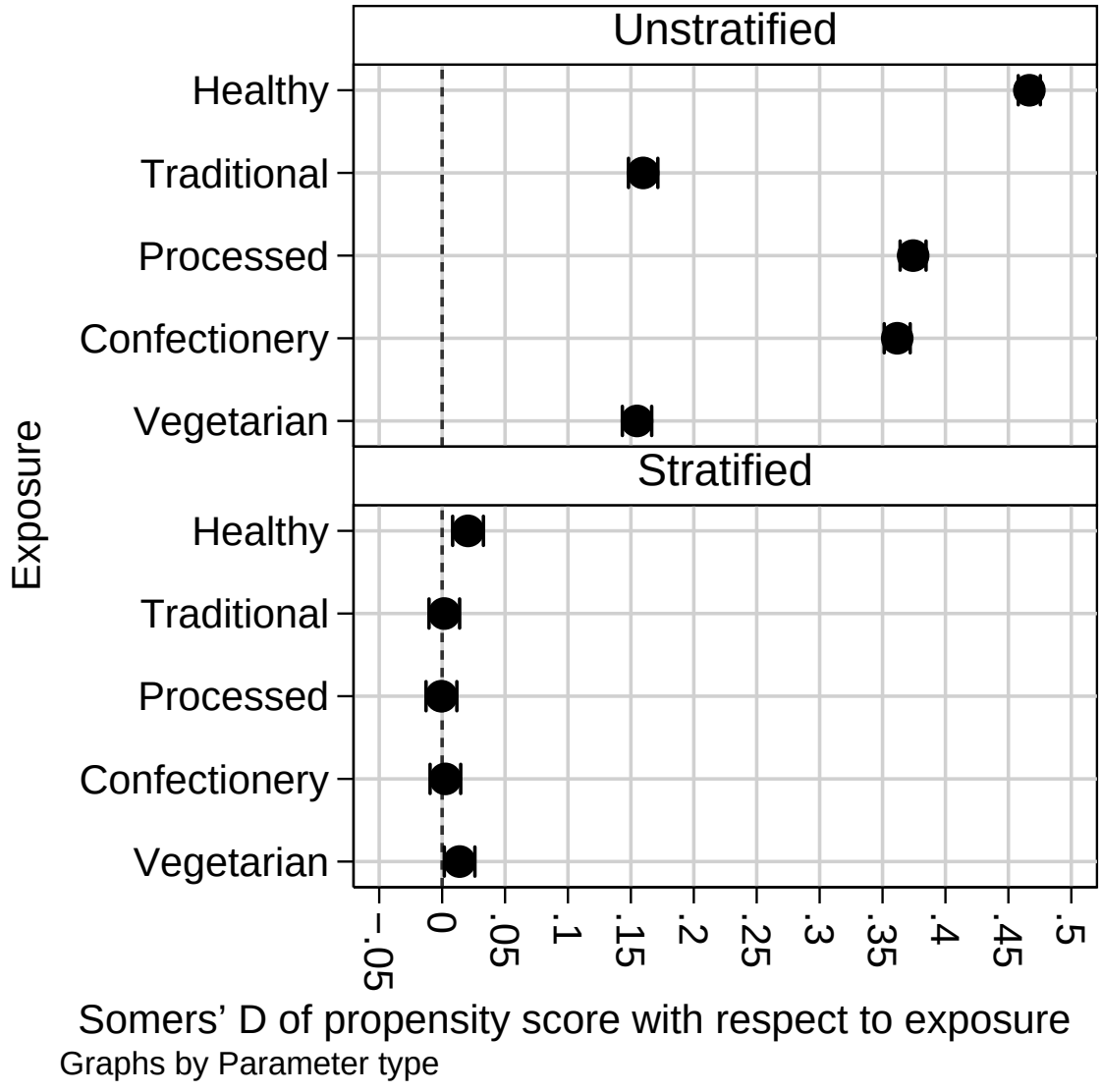
Predictive power of the propensity scores for the "Non–causal" confounder list

- The unstratified Somers' *D* values show that the counfounders predict the diet scores. (But not *too* well.)
- The stratified Somers' *D* values show that, within 64 propensity groups, this predictive power is not much of a problem.
- *Therefore*, 64 propensity groups should be sufficient for modelling out confounding.



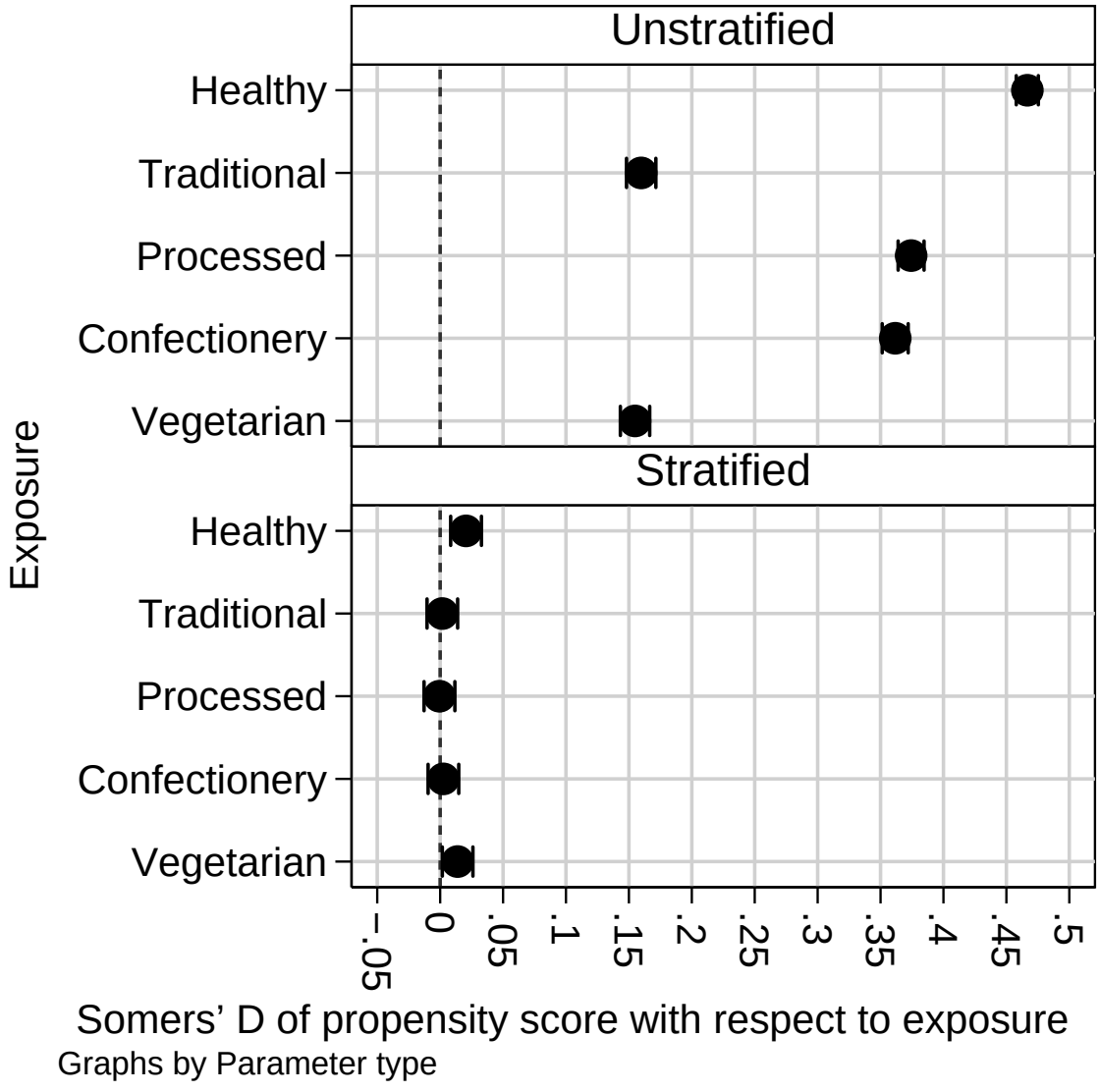
Predictive power of the propensity scores for the "Non–causal" confounder list

- The unstratified Somers' *D* values show that the counfounders predict the diet scores. (But not *too* well.)
- The stratified Somers' *D* values show that, within 64 propensity groups, this predictive power is not much of a problem.
- *Therefore*, 64 propensity groups should be sufficient for modelling out confounding.



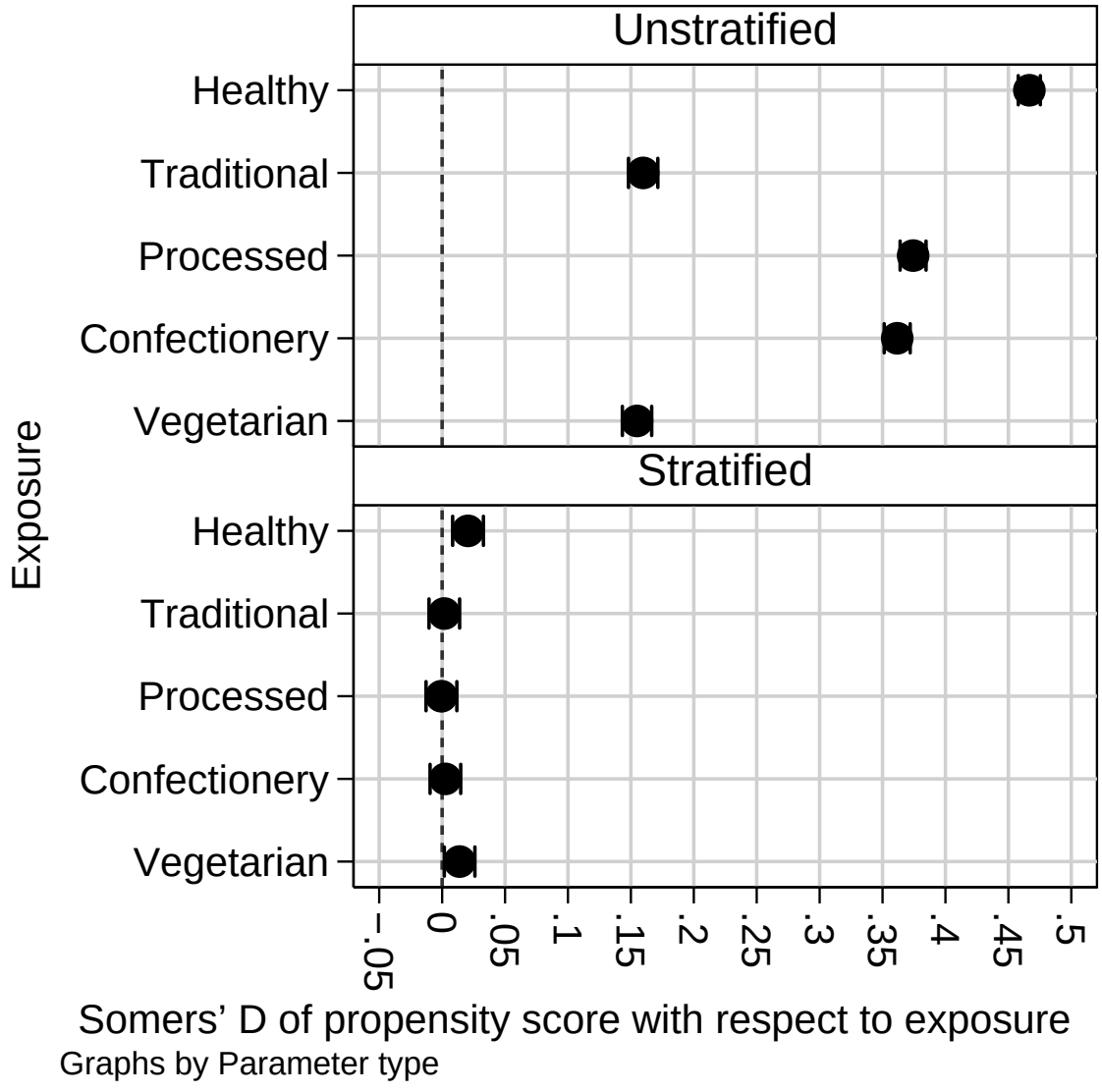
Predictive power of the propensity scores for the "Non–causal" confounder list

- The unstratified Somers' *D* values show that the counfounders predict the diet scores. (But not *too* well.)
- The stratified Somers' *D* values show that, within 64 propensity groups, this predictive power is not much of a problem.
- *Therefore*, 64 propensity groups should be sufficient for modelling out confounding.



Predictive power of the propensity scores for the "Non–causal" confounder list

- The unstratified Somers' *D* values show that the counfounders predict the diet scores. (But not *too* well.)
- The stratified Somers' *D* values show that, within 64 propensity groups, this predictive power is not much of a problem.
- *Therefore*, 64 propensity groups should be sufficient for modelling out confounding.



Limitations of propensity scores

- Propensity scores cannot control for *unobserved* confounders. (Nor can most alternatives.)
- Propensity scores do not tell us which confounders are "doing the confounding". (Nor can most alternatives, most of the time.)
- *However*, their *fundamental* limitation is that they only measure the "exposure–proneness" aspect of confounding.
- If we have prior reason to believe that a concomitant variable predicts "outcome–proneness" *within* each propensity group, then there is a case for including that concomitant variable, in addition to the propensity score.
- This can gain power by reducing residual variation. The concomitant variable is then acting as a blocking factor, rather than as a confounder properly so called.

Limitations of propensity scores

- Propensity scores cannot control for *unobserved* confounders. (Nor can most alternatives.)
- Propensity scores do not tell us which confounders are "doing the confounding". (Nor can most alternatives, most of the time.)
- *However*, their *fundamental* limitation is that they only measure the "exposure–proneness" aspect of confounding.
- If we have prior reason to believe that a concomitant variable predicts "outcome–proneness" *within* each propensity group, then there is a case for including that concomitant variable, in addition to the propensity score.
- This can gain power by reducing residual variation. The concomitant variable is then acting as a blocking factor, rather than as a confounder properly so called.

Limitations of propensity scores

- Propensity scores cannot control for *unobserved* confounders. (Nor can most alternatives.)
- Propensity scores do not tell us which confounders are "doing the confounding". (Nor can most alternatives, most of the time.)
- *However*, their *fundamental* limitation is that they only measure the "exposure–proneness" aspect of confounding.
- If we have prior reason to believe that a concomitant variable predicts "outcome–proneness" *within* each propensity group, then there is a case for including that concomitant variable, in addition to the propensity score.
- This can gain power by reducing residual variation. The concomitant variable is then acting as a blocking factor, rather than as a confounder properly so called.

Limitations of propensity scores

- Propensity scores cannot control for *unobserved* confounders. (Nor can most alternatives.)
- Propensity scores do not tell us which confounders are "doing the confounding". (Nor can most alternatives, most of the time.)
- *However*, their *fundamental* limitation is that they only measure the "exposure–proneness" aspect of confounding.
- If we have prior reason to believe that a concomitant variable predicts "outcome–proneness" *within* each propensity group, then there is a case for including that concomitant variable, in addition to the propensity score.
- This can gain power by reducing residual variation. The concomitant variable is then acting as a blocking factor, rather than as a confounder properly so called.

Limitations of propensity scores

- Propensity scores cannot control for *unobserved* confounders. (Nor can most alternatives.)
- Propensity scores do not tell us which confounders are "doing the confounding". (Nor can most alternatives, most of the time.)
- *However*, their *fundamental* limitation is that they only measure the "exposure–proneness" aspect of confounding.
- If we have prior reason to believe that a concomitant variable predicts "outcome–proneness" *within* each propensity group, then there is a case for including that concomitant variable, in addition to the propensity score.
- This can gain power by reducing residual variation. The concomitant variable is then acting as a blocking factor, rather than as a confounder properly so called.

Limitations of propensity scores

- Propensity scores cannot control for *unobserved* confounders. (Nor can most alternatives.)
- Propensity scores do not tell us which confounders are "doing the confounding". (Nor can most alternatives, most of the time.)
- *However*, their *fundamental* limitation is that they only measure the "exposure–proneness" aspect of confounding.
- If we have prior reason to believe that a concomitant variable predicts "outcome–proneness" *within* each propensity group, then there is a case for including that concomitant variable, in addition to the propensity score.
- This can gain power by reducing residual variation. The concomitant variable is then acting as a blocking factor, rather than as a confounder properly so called.

Example: Height, gender, and lung function outcomes

- In asthma research, outcomes are often lung function measures.
- Whatever exposure we are interested in, we are confident, *a priori*, that height and gender will predict lung function well, even within each exposure–propensity group.
- We therefore usually adjust for height and gender, whether or not these were included in calculating the propensity score.
- In fact, we are so confident that height and gender have an effect, that we usually use a *preliminary* regression model of lung function with respect to height and gender, and then enter the lung function outcome into the *main* analysis as a standardized residual.
- Note that we are not worried by the possibility that taller children have higher vital capacity only because their mothers were better fed.

Example: Height, gender, and lung function outcomes

- In asthma research, outcomes are often lung function measures.
- Whatever exposure we are interested in, we are confident, *a priori*, that height and gender will predict lung function well, even within each exposure–propensity group.
- We therefore usually adjust for height and gender, whether or not these were included in calculating the propensity score.
- In fact, we are so confident that height and gender have an effect, that we usually use a *preliminary* regression model of lung function with respect to height and gender, and then enter the lung function outcome into the *main* analysis as a standardized residual.
- Note that we are not worried by the possibility that taller children have higher vital capacity only because their mothers were better fed.

Example: Height, gender, and lung function outcomes

- In asthma research, outcomes are often lung function measures.
- Whatever exposure we are interested in, we are confident, *a priori*, that height and gender will predict lung function well, even within each exposure–propensity group.
- We therefore usually adjust for height and gender, whether or not these were included in calculating the propensity score.
- In fact, we are so confident that height and gender have an effect, that we usually use a *preliminary* regression model of lung function with respect to height and gender, and then enter the lung function outcome into the *main* analysis as a standardized residual.
- Note that we are not worried by the possibility that taller children have higher vital capacity only because their mothers were better fed.

Example: Height, gender, and lung function outcomes

- In asthma research, outcomes are often lung function measures.
- Whatever exposure we are interested in, we are confident, *a priori*, that height and gender will predict lung function well, even within each exposure–propensity group.
- We therefore usually adjust for height and gender, whether or not these were included in calculating the propensity score.
- In fact, we are so confident that height and gender have an effect, that we usually use a *preliminary* regression model of lung function with respect to height and gender, and then enter the lung function outcome into the *main* analysis as a standardized residual.
- Note that we are not worried by the possibility that taller children have higher vital capacity only because their mothers were better fed.

Example: Height, gender, and lung function outcomes

- In asthma research, outcomes are often lung function measures.
- Whatever exposure we are interested in, we are confident, *a priori*, that height and gender will predict lung function well, even within each exposure–propensity group.
- We therefore usually adjust for height and gender, whether or not these were included in calculating the propensity score.
- In fact, we are so confident that height and gender have an effect, that we usually use a *preliminary* regression model of lung function with respect to height and gender, and then enter the lung function outcome into the *main* analysis as a standardized residual.
- Note that we are not worried by the possibility that taller children have higher vital capacity only because their mothers were better fed.

Example: Height, gender, and lung function outcomes

- In asthma research, outcomes are often lung function measures.
- Whatever exposure we are interested in, we are confident, *a priori*, that height and gender will predict lung function well, even within each exposure–propensity group.
- We therefore usually adjust for height and gender, whether or not these were included in calculating the propensity score.
- In fact, we are so confident that height and gender have an effect, that we usually use a *preliminary* regression model of lung function with respect to height and gender, and then enter the lung function outcome into the *main* analysis as a standardized residual.
- Note that we are not worried by the possibility that taller children have higher vital capacity only because their mothers were better fed.

What do we mean by "over–adjusting"?

It is not always clear what is meant by this, except that we should have adjusted for fewer confounders. *However*, I would argue that it can mean one of two things:

- **Some of the confounders are causally downstream from the proposed intervention.** For instance, if we think the proposed dietary intervention during pregnancy will have the side effects of making children less small at birth, or less fat at 7 years, then we should not estimate its effect by comparing children with different maternal diets and the same birth weight and BMI.
- **The confounders collectively predict the exposure "too well".** This may happen if the number of confounders becomes too close to the number of subjects, and can cause loss of power. This should be checked by measuring the association between the exposure and the propensity score, using scatter plots and/or Somers' D and/or other methods being developed. If this is a problem, then we need to use a less ambitious confounder set.

What do we mean by "over–adjusting"?

It is not always clear what is meant by this, except that we should have adjusted for fewer confounders. *However*, I would argue that it can mean one of two things:

- **Some of the confounders are causally downstream from the proposed intervention.** For instance, if we think the proposed dietary intervention during pregnancy will have the side effects of making children less small at birth, or less fat at 7 years, then we should not estimate its effect by comparing children with different maternal diets and the same birth weight and BMI.
- **The confounders collectively predict the exposure "too well".** This may happen if the number of confounders becomes too close to the number of subjects, and can cause loss of power. This should be checked by measuring the association between the exposure and the propensity score, using scatter plots and/or Somers' D and/or other methods being developed. If this is a problem, then we need to use a less ambitious confounder set.

What do we mean by "over–adjusting"?

It is not always clear what is meant by this, except that we should have adjusted for fewer confounders. *However*, I would argue that it can mean one of two things:

- **Some of the confounders are causally downstream from the proposed intervention.** For instance, if we think the proposed dietary intervention during pregnancy will have the side effects of making children less small at birth, or less fat at 7 years, then we should not estimate its effect by comparing children with different maternal diets and the same birth weight and BMI.
- **The confounders collectively predict the exposure "too well".** This may happen if the number of confounders becomes too close to the number of subjects, and can cause loss of power. This should be checked by measuring the association between the exposure and the propensity score, using scatter plots and/or Somers' *D* and/or other methods being developed. If this is a problem, then we need to use a less ambitious confounder set.

What do we mean by "over–adjusting"?

It is not always clear what is meant by this, except that we should have adjusted for fewer confounders. *However*, I would argue that it can mean one of two things:

- **Some of the confounders are causally downstream from the proposed intervention.** For instance, if we think the proposed dietary intervention during pregnancy will have the side effects of making children less small at birth, or less fat at 7 years, then we should not estimate its effect by comparing children with different maternal diets and the same birth weight and BMI.
- **The confounders collectively predict the exposure "too well".** This may happen if the number of confounders becomes too close to the number of subjects, and can cause loss of power. This should be checked by measuring the association between the exposure and the propensity score, using scatter plots and/or Somers' D and/or other methods being developed. If this is a problem, then we need to use a less ambitious confounder set.

Summary

- Causal arguments are meaningless, unless we have an intervention in mind.
- Confounders must be variables that we do not expect to be changed by that intervention, and may (collectively) indicate "exposure–proneness" and "outcome–proneness".
- Usually, we cannot compare only subjects with *identical* confounder values, so we have to reduce the number of confounder groups somehow.
- Propensity scores are a good method for grouping the confounder value combinations by exposure–proneness.
- *However*, we can gain more power if we also know of ways to adjust for outcome–proneness, within each propensity group.

Summary

- Causal arguments are meaningless, unless we have an intervention in mind.
- Confounders must be variables that we do not expect to be changed by that intervention, and may (collectively) indicate "exposure–proneness" and "outcome–proneness".
- Usually, we cannot compare only subjects with *identical* confounder values, so we have to reduce the number of confounder groups somehow.
- Propensity scores are a good method for grouping the confounder value combinations by exposure–proneness.
- *However*, we can gain more power if we also know of ways to adjust for outcome–proneness, within each propensity group.

Summary

- Causal arguments are meaningless, unless we have an intervention in mind.
- Confounders must be variables that we do not expect to be changed by that intervention, and may (collectively) indicate "exposure–proneness" and "outcome–proneness".
- Usually, we cannot compare only subjects with *identical* confounder values, so we have to reduce the number of confounder groups somehow.
- Propensity scores are a good method for grouping the confounder value combinations by exposure–proneness.
- *However*, we can gain more power if we also know of ways to adjust for outcome–proneness, within each propensity group.

Summary

- Causal arguments are meaningless, unless we have an intervention in mind.
- Confounders must be variables that we do not expect to be changed by that intervention, and may (collectively) indicate "exposure–proneness" and "outcome–proneness".
- Usually, we cannot compare only subjects with *identical* confounder values, so we have to reduce the number of confounder groups somehow.
- Propensity scores are a good method for grouping the confounder value combinations by exposure–proneness.
- *However*, we can gain more power if we also know of ways to adjust for outcome–proneness, within each propensity group.

Summary

- Causal arguments are meaningless, unless we have an intervention in mind.
- Confounders must be variables that we do not expect to be changed by that intervention, and may (collectively) indicate "exposure–proneness" and "outcome–proneness".
- Usually, we cannot compare only subjects with *identical* confounder values, so we have to reduce the number of confounder groups somehow.
- Propensity scores are a good method for grouping the confounder value combinations by exposure–proneness.
- *However*, we can gain more power if we also know of ways to adjust for outcome–proneness, within each propensity group.

Summary

- Causal arguments are meaningless, unless we have an intervention in mind.
- Confounders must be variables that we do not expect to be changed by that intervention, and may (collectively) indicate "exposure–proneness" and "outcome–proneness".
- Usually, we cannot compare only subjects with *identical* confounder values, so we have to reduce the number of confounder groups somehow.
- Propensity scores are a good method for grouping the confounder value combinations by exposure–proneness.
- *However*, we can gain more power if we also know of ways to adjust for outcome–proneness, within each propensity group.

References

[1] Davey Smith G, Ebrahim S. Data dredging, bias or confounding. They can all get you into the BMJ and the Friday papers. *British Medical Journal* 2002; **325**: 1437–1438.

[2] Greenland S, Robins JM. Identifiability, exchangeability , and epidemiological confounding. *International Journal of Epidemiology* 1986; **15(3)**: 413–419.

[3] Greenland S. Modelling and variable selection in epidemiologic analysis. *American Journal of Public Health* 1989; **79(3)**: 340–349.

[4] Hernan MA, Hernandez–Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *American Journal of Epidemiology* 2002; **155(2)**: 176–184.

[5] Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association* 2004; **99(467)**: 854–866.

[6] Newson R. Confidence intervals for rank statistics: Somers' *D* and extensions. *The Stata Journal* 2006; **6(3)**: 309–334.

[7] Northstone K, Emmett P, Rogers I. Dietary patterns in pregnancy and associations with socio-demographic and lifestyle factors. *European Journal of Clinical Nutrition* 2008; **62**: 471–479.