A multiple imputation approach to obtain causal effects under residual confounding, using coarse data models

Robert Grant

1st UK Causal Inference Meeting

Manchester, 14 May 2013

A real-life example

- Chen, Gilbert and Daling (1999) investigated risk factors for having a child with Down's syndrome.
- A simple analysis (crude effect) showed that smoking was protective!
- *Odds ratio=0.80, 95% CI 0.68-0.95*

A real-life example

- They were not so easily fooled. They knew that smoking would be *confounded* by maternal age:
  - Younger women were more likely to smoke
  - There is good evidence of a link between maternal age and Down's
  - So, if you ignore age, it will look like mothers who smoke are at lower risk.

A real-life example

- Adjusted age, number of previous children, and ethnicity.
- But age was categorised as <35 and 35+
- Smoking was still protective but not significant *Odds ratio=0.89, 95% CI 0.73-1.10*
- This sort of result usually leads to a call for further research.
- But it is still *residually confounded* because of large variation within each of the age groups

A real-life example

- When they included the precise age in years, the smoking effect completely disappeared: *Odds ratio=1.00, 95% CI 0.82-1.24*
- There was no causal smoking effect, it was the age effect in disguise all along.
- Similar problems lead to countless newspaper health scare stories.

Smith and Ebrahim. Int J Epidemiol (2001); 30: 1-11.

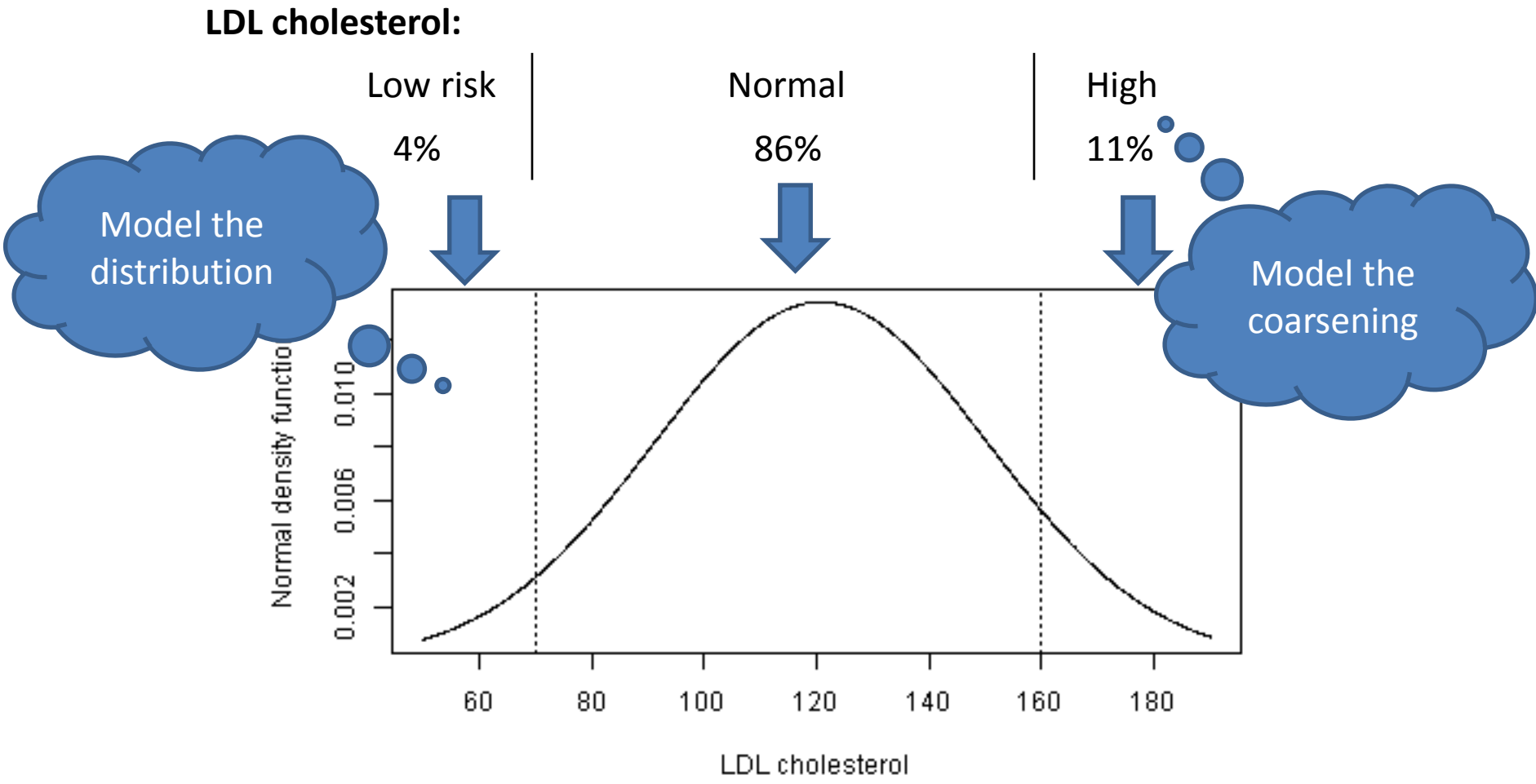
Residual confounding

- The problem occurs when the confounding variable is imperfectly measured, for example:
  - Age recorded only as <35 or 35+
  - Current smoker / ex / never
  - Toddlers' ages in years.
  - Using a Likert scale midpoint as "not applicable"
- Ethics, confidentiality, time constraints, secondary analysis can all lead to this for good reasons.
- The standard response is "better luck next time" – nothing can be done about it

Residual confounding

But not everyone recorded as “current smoker” has clocked up the same number of pack-years

Can we say anything about the true values?



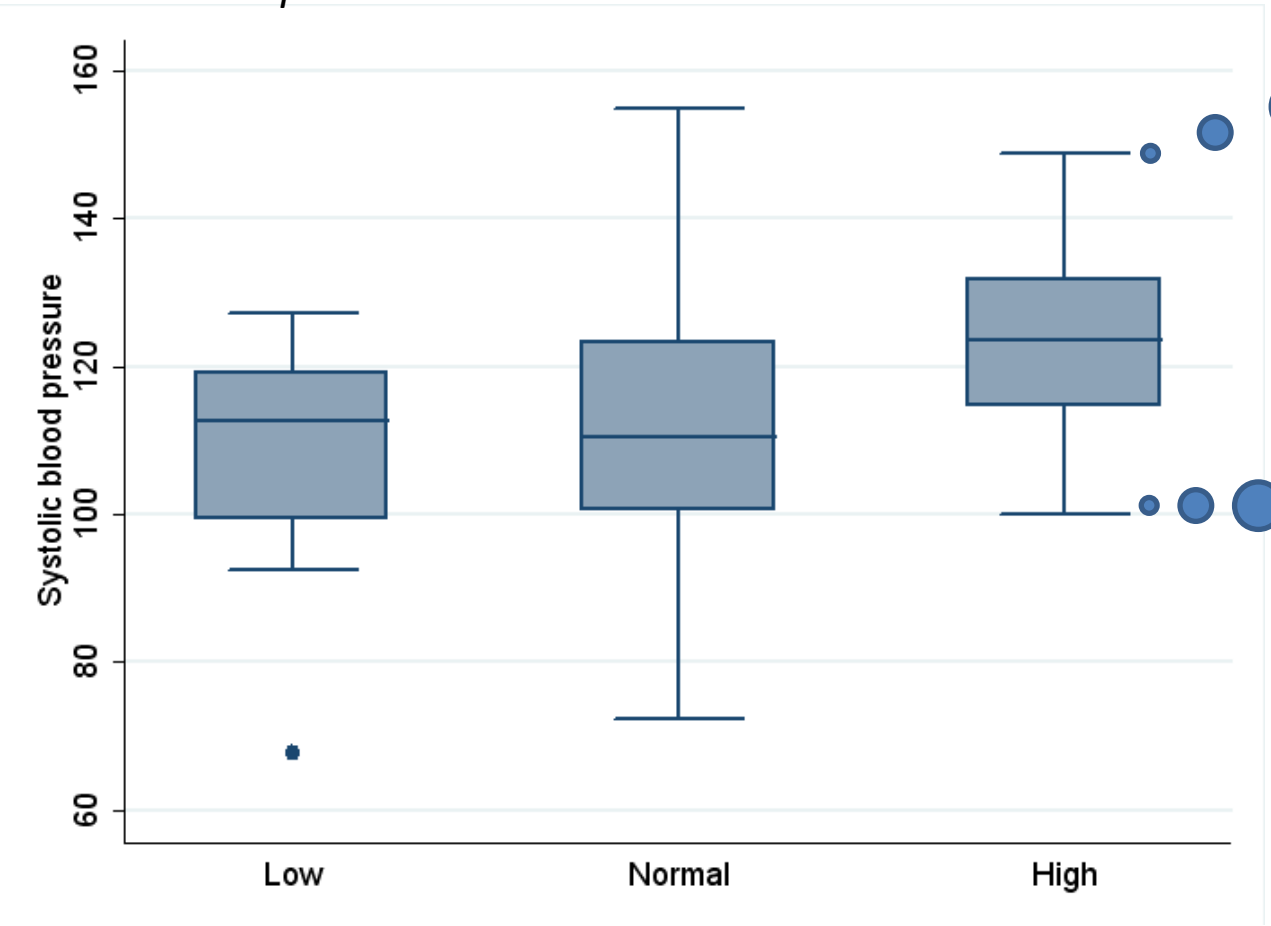
Artificial data based on statistics of Oda & Kawai. Diab Care (2009); 32(9): e113.

Coarsening mechanisms

- Interval-censored or grouped
- Heaped
- Deterministic or stochastic
- CCAR, CAR, CNAR
- We might not know which data are coarsened
- May vary between individuals

Use other data to get a better guess at the true values

Spearman rho=0.18



This person's LDL is probably higher...

... than this one, even in the same "high" category

Building blocks

- *Multiple imputation* is being applied to problems beyond the original missing data setting, for example data linkage
- Several (10-100) different versions of the complete data are drawn from the conditional distribution and the results combined
- *Coarse data models* were set out and tested by Heitjan and Rubin. Despite lots of interest in censored outcomes, little has been done on coarsened covariates.

The new method

- Use Heitjan-Rubin to estimate the parameters for the distribution of the true values
- Use an extended version of their formula to impute multiple true values
- Analyse them as usual for the substantive model, and combine the results with “Rubin’s rules”

Formulas

Heitjan-Rubin:

*For a variable X and its coarsened values X**

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid X^*) = f(X^* \mid \boldsymbol{\theta}, \boldsymbol{\gamma}) = \int_X f(X, X^* \mid \boldsymbol{\theta}, \boldsymbol{\gamma})\, dX$$

$$= \int_X f(X^* \mid X, \boldsymbol{\gamma}) f(X \mid \boldsymbol{\theta})\, dX$$

New extended formula:

$$f(C \mid C^*, X, Y, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{f(C^* \mid C, X, Y, \boldsymbol{\gamma}) f(C \mid X, Y, \boldsymbol{\alpha})}{\int_C f(C^* \mid C, X, Y, \boldsymbol{\gamma}) f(C \mid X, Y, \boldsymbol{\alpha})\, dC}$$

Whole formula gives conditional distribution

Denominator gives parameters

Coarsening

True distribution

MLE for heaped Poisson

- Confounder is a Poisson-distributed count

$$f(C \mid X, Y, \boldsymbol{\alpha}) = \frac{(e^{-\lambda_i}\lambda_i^{c_i})}{c_i!} \quad \text{, where} \quad \lambda_i = e^{\alpha_0 + \alpha_1 x_i + \alpha_2 y_i}$$

- Heaped in some cases to multiples of 5 (*viz* cigarettes per day, child's age)

$$f(C^* \mid C, X, Y, \boldsymbol{\gamma}) = \pi_i^{g_i I(c_i^* = 5\lfloor c_i/5 \rfloor)} (1-\pi_i)^{(1-g_i) I(c_i^* = c_i)}$$

$$\pi_i = \text{expit}(\gamma_0 + \gamma_1 x_i + \gamma_2 y_i)$$

- Happily we know which have been coarsened

$$\frac{\frac{e^{-\lambda_i}\lambda_i^{c_i}}{c_i!}\pi_i^{g_i I(c_i^* = 5\lfloor c_i/5 \rfloor)}(1-\pi_i)^{(1-g_i) I(c_i^* = c_i)}}{\left(\pi_i \sum_{j=0}^{4} \frac{e^{-\lambda_i}\lambda_i^{(c_i^*+j)}}{(c_i^*+j)!}\right)^{g_i} \left(\frac{e^{-\lambda_i}\lambda_i^{c_i^*}}{c_i^*!}(1-\pi_i)\right)^{(1-g_i)}}$$

MLE for heaped Poisson

- Because of the discrete Poisson, the unpleasant integration is reduced to the sum of 5 possible values
- Knowing which values were coarsened is also extremely helpful.
- MLE (in Stata) on the denominator finds values for α and γ
- Plug these into the whole formula and draw imputed values from that

Other approaches

- An interval-censored normally distributed confounder is a nice special case because we can apply interval-censored (Tobit-esque) regression to get the parameters of the conditional distribution
- MCMC is very flexible; coarsened and uncoarsened values are combined in a mixture model, which can include multilevel structure (for human habits at data entry)

Results with artificial data (1)

- Poisson-distributed confounder, n=400
- Heaped down to multiples of 5
- Stochastic but we know which were coarsened
- Logistic regression is of substantive interest
- *True* *OR=1.01 (95% CI 0.98 to 1.03)*
- *Confounded* *OR=1.08 (95% CI 1.06 to 1.10)*
- *Res. conf.* *OR=1.05 (95% CI 1.03 to 1.07)*
- *After MI:* *OR=1.00 (95% CI 0.96 to 1.05)*

Results with artificial data (2)

- Normally distributed confounder, n=200
- Deterministically interval-censored
- Tobit-esque regression to get parameters of C
- Linear regression is of substantive interest
- *True* *Δy=0.36 (95% CI -0.68 to 1.40)*
- *Confounded* *Δy=1.95 (95% CI 1.33 to 2.57)*
- *Res. conf.* *Δy=1.27 (95% CI 0.35 to 2.19)*
- *After MI:* *Δy=0.69 (95% CI -0.71 to 2.09)*

Results with artificial data (3)

- Normally distributed confounder, n=1000
- Heaped at mean, stochastic and unknown, influenced by n_2=100 clusters
- MCMC estimates parameters and samples c_i
- Cox regression is of substantive interest
- *True* *HR=1.28 (95% CI 1.03 to 1.58)*
- *Confounded* *HR=1.03 (95% CI 0.89 to 1.19)*
- *Res. conf.* *HR=1.19 (95% CI 0.99 to 1.43)*
- *After MI:* *HR=1.31 (95% CI 1.06 to 1.61)*

Complications

- You might not know which observations are coarsened
- There might be more than one coarsening mechanism at work
- Coarsening might very between people, not individual observations
- Heitjan-Rubin has rarely been applied because the maths is often hard (i.e. impossible)
- All of these can be addressed but require assumptions

How do we get the models?

- Look at other research for the distributions
- Talk to study participants, or people who collected or cleaned the data; ask them what might have affected the recorded values.
- Study the raw data closely; look for strange patterns, maybe clustered by individual
- Unless most of the data are coarsened and the confounding is strong, approximate models will probably be adequate (I think...)

What if the models are wrong?

- Doubly robust estimation addresses the issue in missing data: the missingness mechanism and the conditional model can be wrong
- Coarsened data is even worse: you might not know which values are coarsened
- Is there such a thing as triply robust estimation?

Applications to human differences

- Survey on domestic violence – interviewers may get different levels of detail to very sensitive questions
- GP database study of incontinence and dementia – GPs may diagnose at different stages
- Attrition from longitudinal study – missing data, non-response and withdrawal are all interlinked, maybe coarsened data too