Challenges of handling big data in the context of CPRD

Adam Jacobs PhD MSc FICR CSci Dip IoD

Dianthus Medical Limited

> Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

Dan Ariely

Why am I talking about big data?

- Statistician with 20 years' clinical research experience
- I run a biometrics CRO
- Have worked on several GPRD projects
- For example:

Epidemiology, prescribing patterns and resource use associated with overactive bladder in UK primary care. Int J Clin Pract. 2006 Aug;60(8):949-58

Challenges of big data

- Privacy and confidentiality
- Statistical issues
- Practical issues

Privacy and confidentiality

- CPRD contains data on individuals
- Public awareness of privacy issues is increasing, because of care.data
- Anonymisation is a tricky concept for medical records

Advantages of big data

- Huge sample size
- Many variables
- Can support sophisticated analyses
- Real-life data

Disadvantages of big data

- No randomisation: can't draw causal inferences
- Routinely collected data is not of same quality as clinical trials
- May not include the variables you are interested in

What can you do with big data?

- Very good at answering “what happens?”
- Not so good at answering “why does it happen?”

Practical challenges

- Many separate datasets: need to be linked together carefully
- Enormous datasets! May be slow computationally
- How to handle missing data?
- Expert statistical advice is essential!

Questions

ajacobs@dianthus.co.uk

http://dianthus.co.uk/blog

@dianthusmed