Big data in health research

LONDON
SCHOOL of
HYGIENE
&TROPICAL
MEDICINE

Opportunities and challenges

Dr Elizabeth Williamson

Medical Statistics Department
Faculty of Epidemiology & Population Health

CSM Big Data Symposium, 2017

Big data in healthcare

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

- There is huge excitement about the potential of big data in biomedical research

- Benchmark examples of big data are
  - online tracking of flu epidemics
  - genetic and genomic analyses of many human diseases
  - the analysis of millions of health and hospital records

- However, the explosion of big data applications has posed problems about how to best store, manage, analyze and integrate such ever-increasing data.

Big data in healthcare

- Healthcare is often said to be lagging behind other disciplines in its use of big data

- Issues complicating big data use in healthcare:
  - Privacy
  - Governance
  - Ethics
  - Ownership

- Computational
- Statistical

The 4 V’s of big data

- Commonly listed attributes of big data

- Volume
  - Scale of data
  - Big n (electronic health records), big p (e.g. bioinformatics)
- Variety
  - Different forms of data
  - Wearables, social media, twitter feeds, blogs, routinely collected data, images
  - Made (e.g. ‘omics data) vs found (e.g. administrative data) data
- Velocity
  - Speed of incoming data
  - Never have the “final analysis dataset”
  - (Semi-) automation of methods required
- Veracity
  - Quality of data

All models are wrong

“All models are wrong but some are useful”

George Box

**“All models are wrong…
… and increasingly you can succeed without them”**

Peter Norvig, Google's research director

Wired magazine: The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

With enough data, the numbers speak for themselves.

Models and algorithms

Key distinction between:

- Algorithms
  - E.g. Recommending the best book, given your previous purchases
- Modelling
  - E.g. Predicting the risk of myocardial infarction within 10 years
- Frequent tasks in biomedical research
  - Prediction
  - Estimation
  - Causal inference

Letting the data speak for themselves

"All models are wrong, and increasingly you can succeed without them"

- [Data veracity!] Norvig's website (http://norvig.com/fact-check.html) says "That's a silly statement, I didn't say it, and I disagree with it."
- "[Let's] instead embrace complexity, and use as much data as well as we can to help define (or estimate) the complex models we need for these complex domains"

"If data can speak for themselves, they can lie for themselves"

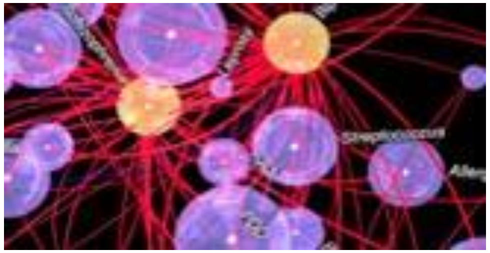
Prof David Hand
UCL Theory of Big Data conference, 2017

Centre for Statistical Methodology (CSM)

Aim

- to gather statistical and methodological expertise across the School in order to strengthen our research capacity in epidemiology and public health.
- and exploit the richness and interdisciplinarity of the work carried out within the School.

- 11 themes
- …one of which is Big Data



CSM Big Data theme

Aim

- to provide a collaborative environment for methodological development in big data problems…
- …and its dissemination across the LSHTM research community, and beyond.

Methods for big data

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

- Next few slides…
  - some examples of work that theme members are undertaking
- Not comprehensive
- Mainly focused on work I am involved with
  - Methods relevant to use of large-scale routinely collected data
- No ‘omics!

Assessing and improving data quality

- Methodological work to assess and improve the quality of data tends to be overlooked
- Better detection of errors, leading to enhanced chances of correcting erroneous data, is essential

- Longitudinal data within routinely collected primary care data
- Promising method: iterative approach
  - fitting mixed models,
  - identifying likely outliers, and
  - re-fitting the model after removal of outliers

Welch C, et al. Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. *Pharmacoepidemiology and Drug Safety*, 2012; 21:725-732.

Linkage

- "A merging that brings together information from two or more sources of data with the object of consolidating facts concerning an individual or an event that are not available in any separate record"

  (Organisation for Economic Co-operation and Development (OECD) Glossary of Statistical Terms).

- Opportunities:
  - Improving data quality (key outcomes, inconsistencies)
  - Cost-effective means of assembling a dataset
  - Address wider range of questions

- Challenges:
  - the lack of unique identifiers for linkage
  - data security considerations

Linkage

- Small amounts of linkage error can result in substantial bias
  - False matches (add variability, weaken associations)
  - Missed matches (reduce the sample size, potential selection bias).

- Evaluating the potential impact of linkage error on results is vital

Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Medical Research Methodology, 2014, 14:36.

Missing data

- The challenge of missing data is not restricted to the context of big data. However:
  - it is frequently an important issue in large routinely collected datasets
  - missing data in such settings raises complex and often novel challenges

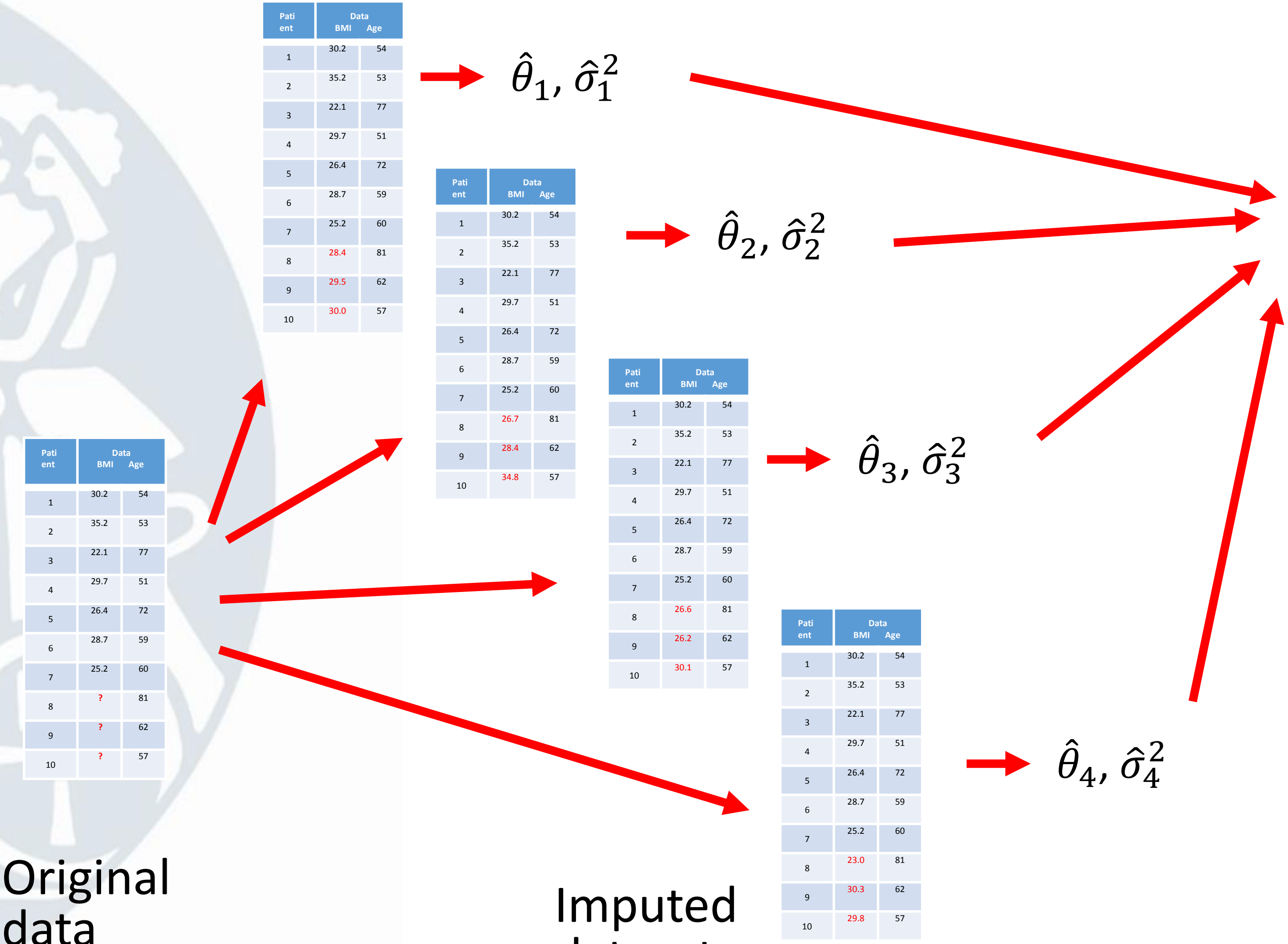
- Sheer volume of the data

- ‘Data dependent sampling’: the process you are trying to collect data (to some extent) controls the data you are able to collect.
  - E.g. participants leaving wearables at home on low activity days
  - E.g. in routinely collected primary care data, information is collected only when the patient chooses to visit their general practitioner

Multiple imputation

Patient	Data		Imputation 1		Imputation 2		Imputation 3		Imputation 4	
	BMI	Age	BMI	Age	BMI	Age	BMI	Age	BMI	Age
1	30.2	54	30.2	54	30.2	54	30.2	54	30.2	54
2	35.2	53	35.2	53	35.2	53	35.2	53	35.2	53
3	22.1	77	22.1	77	22.1	77	22.1	77	22.1	77
4	29.7	51	29.7	51	29.7	51	29.7	51	29.7	51
5	26.4	72	26.4	72	26.4	72	26.4	72	26.4	72
6	28.7	59	28.7	59	28.7	59	28.7	59	28.7	59
7	25.2	60	25.2	60	25.2	60	25.2	60	25.2	60
8	?	81	28.4	81	26.7	81	26.6	81	23.0	81
9	?	62	29.5	62	28.4	62	26.2	62	30.3	62
10	?	57	30.0	57	34.8	57	30.1	57	29.8	57

Multiple imputation



Patient	Data	
	BMI	Age
1	30.2	54
2	35.2	53
3	22.1	77
4	29.7	51
5	26.4	72
6	28.7	59
7	25.2	60
8	?	81
9	?	62
10	?	57

Original data

Patient	Data	
	BMI	Age
1	30.2	54
2	35.2	53
3	22.1	77
4	29.7	51
5	26.4	72
6	28.7	59
7	25.2	60
8	28.4	81
9	29.5	62
10	30.0	57

$\hat{\theta}_1, \hat{\sigma}_1^2$

Patient	Data	
	BMI	Age
1	30.2	54
2	35.2	53
3	22.1	77
4	29.7	51
5	26.4	72
6	28.7	59
7	25.2	60
8	26.7	81
9	28.4	62
10	34.8	57

$\hat{\theta}_2, \hat{\sigma}_2^2$

Patient	Data	
	BMI	Age
1	30.2	54
2	35.2	53
3	22.1	77
4	29.7	51
5	26.4	72
6	28.7	59
7	25.2	60
8	26.6	81
9	26.2	62
10	30.1	57

$\hat{\theta}_3, \hat{\sigma}_3^2$

Patient	Data	
	BMI	Age
1	30.2	54
2	35.2	53
3	22.1	77
4	29.7	51
5	26.4	72
6	28.7	59
7	25.2	60
8	23.0	81
9	30.3	62
10	29.8	57

$\hat{\theta}_4, \hat{\sigma}_4^2$

Imputed datasets

Combined (Rubin's rules) to give overall estimate

Two-fold imputation

- Large scale longitudinal data (e.g. primary care data)
- Intermittent patterns of missing data
- Question: How to deal with large numbers of variables repeatedly measured over time?

Some solutions:

- Divide time into blocks
  - Ignores temporal order – leads to bias
- Include all time blocks in the same MI model
  - Doesn't exploit temporal order – can break down in practice
- Full modelling of hierarchical and longitudinal structure
  - Computationally intractable
- Two-fold fully MI: conditions on measurements which are local in time

Welch C, Petersen I, Bartlett J, White IR, Marston L, Morris RW, Nazareth I, Walters K, Carpenter J. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statist. Med.* 2014, 33:3725–3737

Imputation in propensity score analyses

- With large routinely collected datasets, assessing harms and benefits of drugs is often undertaken using propensity score (PS) methods to handle large number of potential confounders
- In settings with many confounders, problems of missing data are exacerbated

- **Propensity score (PS) methods**
- The propensity score is an individual's probability of receiving the treatment conditional on measured confounders
- Basic idea: individuals with the same propensity score were effectively randomised to receive the treatment or not
- Therefore, comparing sub-groups of individuals with the same (or similar) propensity score provides an unbiased estimate of treatment effect
- Two-stage approach:
  - Estimate PS from data
  - Match, stratify on PS (& estimate treatment effect).

Imputation in propensity score analyses

- Surprisingly, there is little research about how best to implement imputation in the propensity score context

Questions:

- Should the PS or the treatment effect estimates be combined across imputed datasets?
- Should the outcome be included in the imputation model?
    - Missing data literature (**YES**) versus PS literature (**NO**)

- Widely read and cited paper suggested combining PS and omitting the outcome from the imputation model

Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res. 2016, 25(1):188-204.

Imputation in propensity score analyses

Our recent work showed that:

- omitting the outcome from the imputation model leads to substantial bias
- combining the PS across imputed datasets leads to bias, but in practice can sometimes be small
- combining the treatment effects across imputed datasets results in consistent estimators

Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J, Resche-Rigon M, Carpenter JR, Williamson EK. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Research, 2017, Epub ahead of print

Data dependent sampling

Example

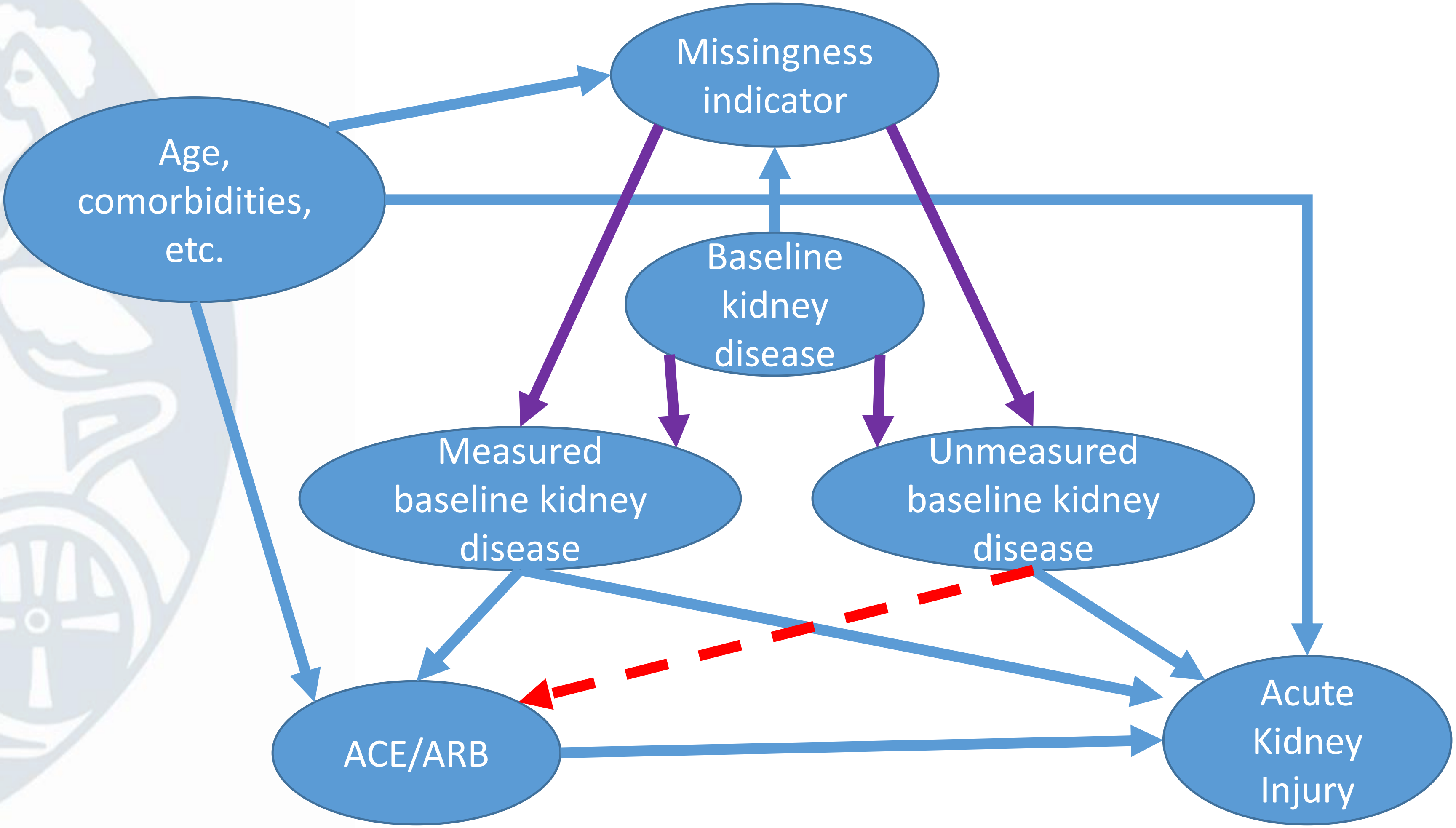
- Cohort study assessing the effect of angiotensin-converting enzyme inhibitors and angiotensin receptor blockers on acute kidney injury using routinely collected primary care data
  - Baseline severity of chronic kidney disease a key confounder
    - Large amounts of missing values
    - More likely to be measured if more severe

- Complete records analysis: loses >50% sample – potential for bias
- Multiple imputation: Key assumption unlikely to be satisfied

Mansfield KE, Nitsch D, Smeeth L, Bhaskaran K, Tomlinson LA. Prescription of renin-angiotensin system blockers and risk of acute kidney injury: a population-based cohort study. *BMJ Open*. 2016.

Causal diagram

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE



Missingness pattern approach

- If we can assume unmeasured kidney function does not affect ACE/ARB prescription decision…
  - i.e. baseline chronic kidney disease is only a confounder when measured

- Our work has shown that: Missingness pattern approach likely to be valid

- In essence:
  - Estimate the propensity score (probability of ACE/ARB) in people with and without baseline severity of chronic kidney disease
  - (strong connections to missingness indicator method)
  - Combine into one PS
  - Proceed as in standard PS analysis

Work by Helen Blake (PhD candidate), Clemence Leyrat, James Carpenter, Elizabeth Williamson, Laurie Tomlinson, Kate Mansfield.

Confounding

Example

- Primary care data – cohort study and self controlled case series
- Patients receiving clopidogrel and aspirin.
- Exposure: prescription of proton pump inhibitors
- Outcome: incident myocardial infarction.
- 24,471 patients prescribed clopidogrel and aspirin, 12439 (50%) prescribed PPI at some point

- Cohort study Hazard ratio: 1.30 (95% confidence interval 1.12 to 1.50)
- SCCS Hazard ratio: 0.75 (95% confidence interval 0.55 to 1.01)

Douglas IJ, Evans SJW, et al. Clopidogrel and interaction with proton pump inhibitors: comparison between cohort and within person study designs. *BMJ* 2012;345:e4388

Confounding in electronic healthcare records

- A possible solution: High-dimensional PS
- Algorithmic approach to identifying confounders from the wealth of information available in electronic healthcare records
  - identifying data dimensions, e.g. diagnoses, procedures, and medications,
  - empirically identifying candidate covariates
  - prioritizing and selecting covariates
  - estimating the propensity score (probability of exposure)
  - estimating the treatment effect

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology, 2009.

- Applied on US claims data
- Few applications in UK setting
- Validation, e.g. by simulation, is very complex

Confounding in electronic healthcare records

- An alternative solution: Targeted Maximum Likelihood Estimation

- TMLE allows for data-adaptive estimation while obtaining valid statistical inference
- Doubly robust estimation

- Applied to UK primary care data to investigate the association between statins and all-cause mortality
- The authors concluded that more understanding of the comparative advantages and disadvantages of this approach was needed

Pang M, Schuster T, Filion KB, Eberg M, Platt RW. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*. 2016; 27(4): 570–577.

TMLE resources

- Members of our theme (Dr Miguel Angel Luque) have developed an online tutorial to introduce TMLE for Causal Inference

https://migariane.github.io/TMLE.nb.html#1_introduction

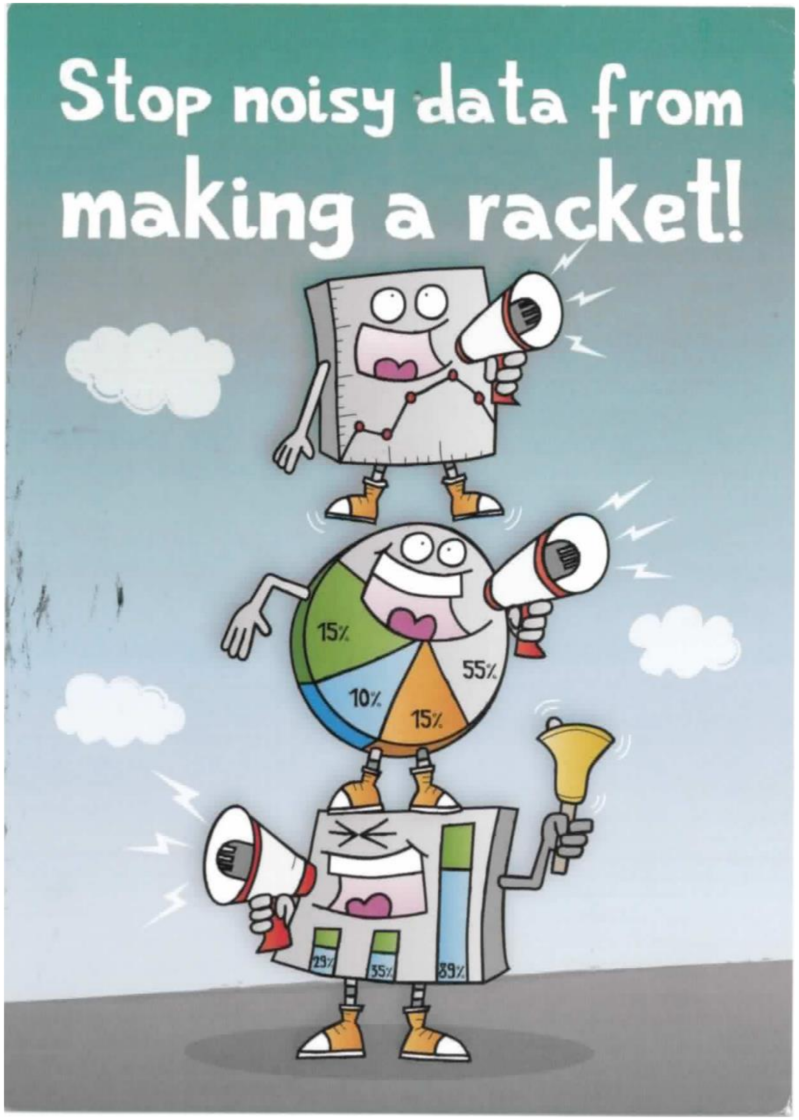
- They have also created and made available a free open source Stata program (ELTMLE) to implement double-robust methods for causal inference, including Machine Learning algorithms for prediction

https://github.com/migariane/meltmle

https://github.com/migariane/weltmle

Generalising RCTs

- Generalising from RCTs to wider populations
- Using RCTs to validate observational studies (e.g. TORCH)


Stop noisy data from making a racket!
15%
55%
10%
15%
29%
35%
89%

Some references

Welch C, Petersen I, Walters K, Morris RW, Nazareth I, Kalaitzaki E, White IR, Marston L, Carpenter J. Two-stage method to remove population- and individual-level outliers from longitudinal data in a primary care database. *Pharmacoepidemiology and Drug Safety*, 2012; 21:725-732.

Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC Medical Research Methodology, 2014, 14:36. **DOI:** 10.1186/1471-2288-14-36

Mansfield KE, Nitsch D, Smeeth L, Bhaskaran K, Tomlinson LA. Prescription of renin-angiotensin system blockers and risk of acute kidney injury: a population-based cohort study. *BMJ Open*. 2016. 21;6(12):e012690. doi: 10.1136/bmjopen-2016-012690

Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology, 2009. 20(4):512-522 doi: 10.1097/EDE.0b013e3181a663cc

Some references

Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley, 1987.

Welch C, Petersen I, Bartlett J, White IR, Marston L, Morris RW, Nazareth I, Walters K, Carpenter J. Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statist. Med.* 2014, 33:3725–3737

Mitra R, Reiter JP. A comparison of two methods of estimating propensity scores after multiple imputation. Stat Methods Med Res. 2016, 25(1):188-204.

Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J, Resche-Rigon M, Carpenter JR, Williamson EK. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Research, 2017, doi: 10.1177/0962280217713032. [Epub ahead of print]

Some references

Douglas IJ, Evans SJW, Hingorani AD, Grosso AM, Timmis Aa, Hemingway H, Smeeth L. Clopidogrel and interaction with proton pump inhibitors: comparison between cohort and within person study designs. *BMJ* 2012;345:e4388 doi: 10.1136/bmj.e4388

Pang M, Schuster T, Filion KB, Eberg M, Platt RW. Targeted maximum likelihood estimation for pharmacoepidemiologic research. *Epidemiology*. 2016; 27(4): 570–577.

Author: Dr Miguel Angel Luque:

https://migariane.github.io/TMLE.nb.html#1_introduction

https://github.com/migariane/meltmle

https://github.com/migariane/weltmle