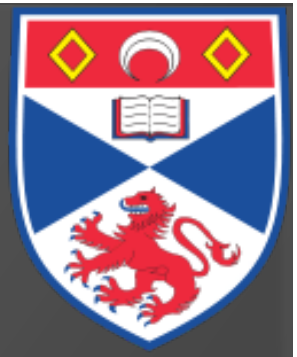

MACHINE SCIENCE IN BIOMEDICINE: PRACTICALITIES, PITFALLS AND POTENTIAL

Tom Kelsey – University of St Andrews
Hamish Wallace – University of Edinburgh

Machine Science

- A hot topic in the theory and philosophy of modern science
- Recent claims:
- “within a decade, even more powerful tools will enable automated, high-volume hypothesis generation to guide high-throughput experiments in biomedicine, chemistry, physics, and even the social sciences”
  - J Evans, A Rzhetsky, “**Machine Science**.” Science, 329 (5990); 399–400, Jul. 2010.

Examples

- The modern scientific literature is full of data-driven studies
- Providing hypotheses for further investigation
- Canonical “old-school” example is the double-helix structure of DNA
  - proposed by Watson & Crick after analysis of data from other researchers

Structure of DNA

- Watson & Crick, "**A Structure for Deoxyribose Nucleic Acid**", Nature, 1953
- Paper contains no experimental proofs
  - Contains only hypotheses
- They collected no data
  - Franklin, Wilkins, Gosling, Pauling, Chargaff, ...
  - They failed to fully acknowledge these sources
  - But won the race to publish

Methodology

- Find, retrieve & classify data
  - raw microarray freely available
  - other data from tables & plots
  - yet more from descriptive statistics
- Look for patterns, models & insights
- (Hopefully) validate
  - compare results against unseen data
  - perform specific laboratory experiments

Advantages

- Inexpensive
  - no wet lab resources required
- Provides more than just hypotheses
  - effect sizes
  - power, cohort sizes, ...
- Tool support
  - data mining
  - text mining
  - machine learning
  - non-linear regression, DE solvers, grid/cloud, ...

Practicality 1. Data availability

- The modern trend is for biomedical data to be freely available
  - data repositories
  - supplementary information in open access journals
- This is far from pervasive
  - and many important studies report
  - (or have reported)
  - summarised data

Practicality 1. Data availability

- Tabularised data is easily converted to CSV form using textual manipulation tools
- Chart data takes more effort
  - tools like Plot Digitizer allow calibration of axes, followed by retrieval of data points
  - repeated data points have to be derived from manual analysis of the descriptive statistics
  - or omitted

Practicality 1. Data availability

- What if we only have descriptive statistics?
- Equivalent datasets can be recreated
- Example: 58,673 measurements of ovarian volume (Pavlik et al.)
  - too many for a scatterplot
  - authors unwilling to grant access
  - log-adjusted means and SDs published

Practicality 1. Data availability

$$\mu = e^{x+\frac{1}{2}y^2}$$

$$\sigma = e^{2x+y^2}\left(e^{y^2}-1\right)$$

Solve for x and y to obtain log-unadjusted mean and SD

Repeat as many times as needed for k-fold cross-validation

Practicality 2. Data search

- All text-mining tools depend on search terms used
- Different terms can relate to the same things
  - MRI and NMR
  - AMH and MIS
- Natural language problems are routinely ignored
  - no obvious solution to this problem
  - unless the literature is translated to, say, English
  - which isn't going to happen

Practicality 3. Homgeneity

- Is is safe to combine or compare data from different studies?
- It should be, if enough information is given so that replication is possible
  - underlying assumption in science that results can be tested by others
- In practice, great care must be taken
  - assays used
  - inclusion/exclusion criteria
  - detection limits

Pitfalls

- Non-exhaustive search for data
- Inclusion of poor quality data
  - The first phase of search can easily be done by students
  - Several iterations will be needed to add missed data and exclude poor data
  - This has to be done by experts in the field
- Even an expert will miss information in a language he/she can’t understand

Pitfalls

- Over-reliance on full automation
  - ruling out important historic data & results
- Over-reliance on favoured analytical tools
  - Computer scientists and mathematicians make important contributions
  - But low-level human analysis is vital
  - Clear danger of computational tools looking for a biomedical application
    - with the main focus of research being the tool itself
- Where will the study be published?

Pitfalls

- Under analysis
  - Biomedical researchers seem all to have done the same course
  - p-values, ROC AUC, not much else
  - Confidence intervals & prediction limits are more useful
- Little or no independent validation
  - Ideally using wholly unseen data
  - Other validation techniques exist
  - How well would your results generalise to unseen data?

Pitfalls

- Assumption that full automation is possible
  - human experts have to read and understand the minutiae of methods sections
  - even for open access, shared data studies
  - so that the right conclusions are being based on the right tools applied to the right data
- Inertia in the scientific community
  - “What was your hypothesis?”
  - “Where is your own data?”
  - “Why have you done no experiments?”

Potential

- Time is on our side
  - more and better data & tools in the pipeline
  - fuller automation is a reality
- Small errors in data retrieval/re-creation can often be neglected
  - many biomedical measurement techniques have standard errors of 10% or more
  - so as long as analytic errors are not systematically skewed, the new data is as good as the original data

Potential

- Scope for interaction with wet labs
  - machine science based on hypothesis-led research
  - leading to hypotheses for testing
  - leading to data for machine science...
- Funding bodies broadly support the idea
  - NIH requires open access dissemination
  - data repositories are becoming mandatory
  - UK BBRC will not fund “lab only” grants

Summary

- Machine science is exciting & rewarding
  - important results are being published daily
- Multi-disciplinary teamwork is vital
  - I know little about angiogenesis regulation in rodent models
  - My colleagues know little about non-linear regression
- Inter-disciplinary co-operation is even more important
  - there are no “junior partners”

References

- *2010; W H B Wallace, T W Kelsey; "Human ovarian reserve from conception to the menopause"; PLoS ONE; 5(1): e8772. doi:10.1371/journal.pone.0008772*
- *2004; W H B Wallace, T W Kelsey; "Ovarian reserve and reproductive age may be determined from measurement of ovarian volume by transvaginal sonography"; Human Reproduction; 19(7):1612-1617;*
- *2011;* T W Kelsey, Wright, S M Nelson, R A Anderson, W H B Wallace*; "A validated model of serum anti-Mullerian hormone from conception to menopause"; PLoS Computational Biology; In press*
- *2011; S M Nelson, D A Lawlor; "Predicting live birth, preterm and low birth weight infant after in-vitro fertilisation: A prospective study of 144,018 treatment cycles"; PLoS Medicine; In press*