Challenges for the future: Big Data

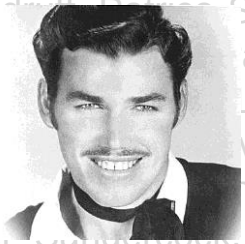
Malcolm Macleod

Professor of Neurology and Translational Neurosciences

University of Edinburgh

and

Honorary Consultant Neurologist, NHS Forth Valley



Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies

Starting points

- There are more papers published in a day than most people could read in a month
- Narrative expert review (and presumably expert opinion) is hopelessly confounded by bias
- Funding and translational decisions are taken on the basis of what's hot, and expert opinion
- Systematic reviews take so long to do that by the time they are published they are already out of date
- Even if research funders wanted to use systematic reviews to inform decision making, at present they could not

Rate of publication

- In 2013, 4700 new publications were added to PubMed every working day

Domain	Number
In vivo and in vitro	610
In vivo	350
Pharmacology	76
Neurosciences	52

- If a panelist for a neuroscience grant awarding body were to try to keep up to date, spending 30 minutes reading each paper, and did nothing else but read papers all year, they would get through 30% of the total

Systematic reviews v narrative reviews

- House dust mites and asthma
  - 63 of 70 review articles claimed efficacy for physical eradication measures (vacuum cleaning, bed covers, freezing …)
  - Most frequently cited study had 7 patients per group
  - Systematic review (Cochrane) identified 28 trials involving 939 patients
    - Found no effect of physical measures in improving outcome

Schmidt and Gotzsche, 2005 J Fam Practice

"Authors often use non randomised studies to create a false impression of consensus"

	Utility	Timeliness
Expert opinion	not known	immediate
Narrative review	low	intermediate
Systematic review	high	years

Stages of a systematic review

- Formulation of research question
- Secure funding
- Study protocol
- Literature search
- Screening against exclusion criteria
- Retrieve full text
- Screening against inclusion criteria
- Extraction of publication meta-data
- Extraction of outcome data
- Meta-analysis per study protocol
- Drafting of manuscript
- Submission and peer review
- Publication
- Dissemination

Stages of a systematic review

- Formulation of research question
- Secure funding
- Study protocol
- Literature search
- Screening against exclusion criteria
- Retrieve full text
- Screening against inclusion criteria
- Extraction of publication meta-data
- Extraction of outcome data
- Meta-analysis per study protocol
- Drafting of manuscript
- Submission and peer review
- Publication
- Dissemination

Stages of a systematic review

- Formulation of research question
- Secure funding
- Study protocol
- Literature search
- Screening against exclusion criteria
- Retrieve full text
- Screening against inclusion criteria
- Extraction of publication meta-data
- Extraction of outcome data
- Meta-analysis per study protocol
- Drafting of manuscript
- Submission and peer review
- Publication
- Dissemination

Stages of a systematic review

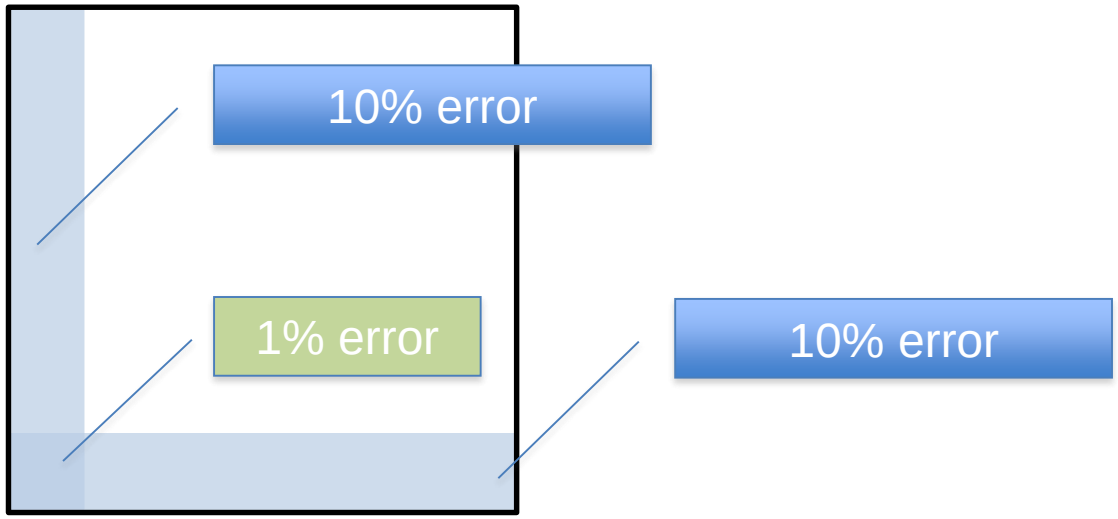
- Formulation of research question
- Secure funding
- Study protocol
- Literature search
- Screening against exclusion criteria
- Retrieve full text
- Screening against inclusion criteria
- Extraction of publication meta-data
- Extraction of outcome data
- Meta-analysis per study protocol
- Drafting of manuscript
- Submission and peer review
- Publication
- Dissemination

Stages of a systematic review

- Formulation of research question
- Secure funding
- Study protocol
- Literature search
- Screening against exclusion criteria
- Retrieve full text
- Screening against inclusion criteria
- Extraction of publication meta-data
- Extraction of outcome data
- Meta-analysis per study protocol
- Drafting of manuscript
- Submission and peer review
- Publication
- Dissemination

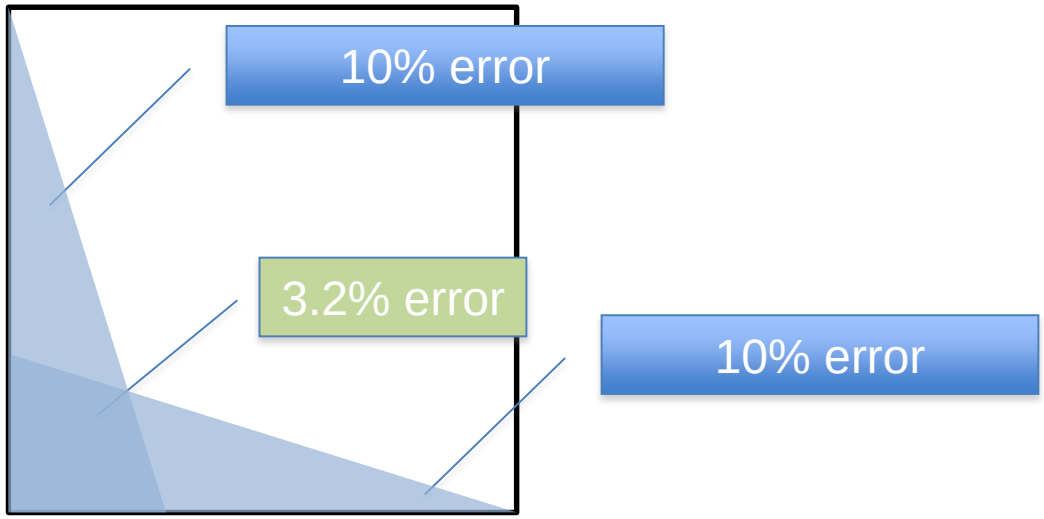
Screening against inclusion and exclusion criteria

- Neuropathic pain systematic review – 60,000 hits
  - Dual screening at 1 minute a paper takes 50 person weeks

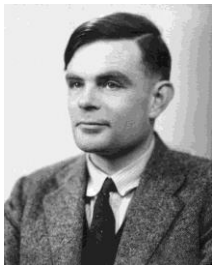


Screening against inclusion and exclusion criteria

- Neuropathic pain systematic review – 60,000 hits
  - Dual screening at 1 minute a paper takes 50 person weeks



Alan Turing



- The question is not whether computers can think, but can we make them do what we (as thinking entities) do?

Tom Mitchell

“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”

An algorithm that learns from data, in contrast to an algorithm following programmed instructions

Machine learning

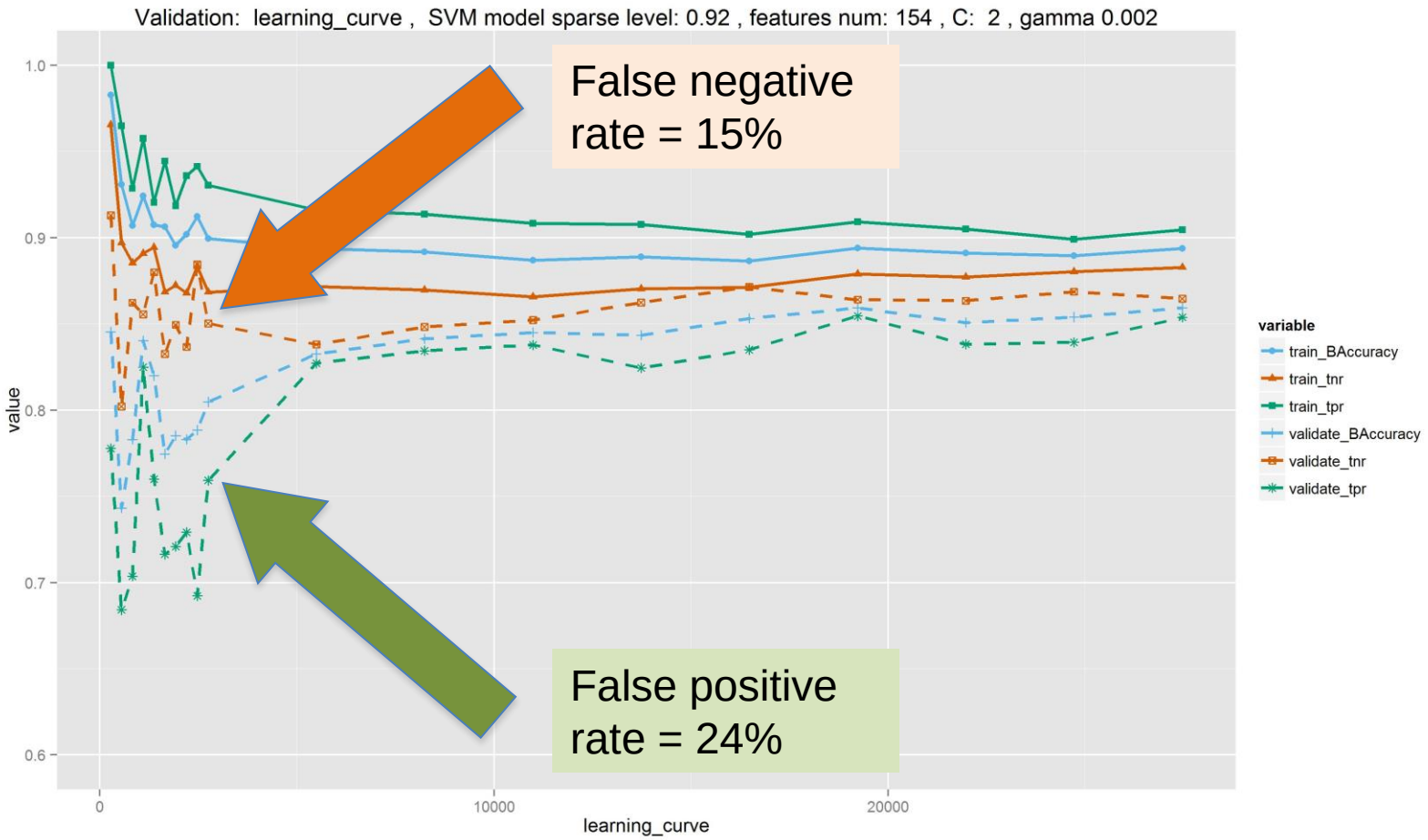
- Support vector machine
  - A form of machine learning where, when presented with a training set allocated to two or more categories, builds an algorithm to assign new items to be categorised

True positive	**False negative**
False positive	**True Negative**

Learning

- Neuropathic pain dataset
  - ~35,000 unique publications
  - Dual screened by GC, NS
  - ~6,500 included
- Randomly split
  - 60% as training subset
  - 20% as validation cohort
  - 20% reserve

Machine based learning



Other costs

- Dual extraction of meta-data
  - 6000 publications, 10 minutes each
    - 50 person weeks
- Dual extraction of outcome data
  - 6000 publications, 20 minutes each
    - 100 person weeks

- Total 200 person weeks

Systematic **L**iving **I**nformation **M**achine

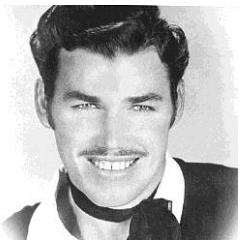
Formulate research question (protocol)

Conduct baseline systematic review and meta-analysis

Re-formulate research question, including stakeholder engagement, responses to baseline review (protocol)

Search	Apply Inclusion/Exclusion criteria	Risk of Bias assessment	Data extraction	Analysis
Develop and validate automated search strategy				Develop web based automated meta-analysis and reporting
	Develop and validate machine algorithm for Inclusion/ Exclusion	Develop and validate machine algorithm for Risk of Bias assessment	Develop and validate machine algorithm for Data Extraction	
	Develop and validate crowd sourcing for Inclusion/ Exclusion	Develop and validate crowd sourcing for Risk of Bias assessment	Develop and validate crowd sourcing for Data Extraction	

Enable continuously updated systematic reviews with meta-analysis in defined research areas



CAMARADES: Bringing evidence to translational medicine

Estimated savings

- Screening
  - Machine plus 1
    - Training set ~2000: < 2 weeks
    - Screening set (~58000): 24 person weeks
  - Machines 1 to n
    - Training set: <2 weeks

Estimated savings

- Meta-data extraction
  - Machine plus 1
    - Training set ~2000: 17 person weeks
    - Screening set (~4000): 17 person weeks
  - Machines 1 to n
    - Training set: 17 person weeks

Estimated savings

- Data extraction
  - Machine plus 1
    - Training set ~2000: ~ 34 weeks
    - Screening set (~4000): ~ 34 person weeks
  - Machines 1 to n
    - Training set: ~ 34 weeks
  - Crowd sourcing (assuming 2 per article)
    - 10 volunteers doing 3 a week: 400 weeks
    - 100 volunteers doing 3 a week: 40 weeks
    - 100 volunteers doing 6 a week: 20 weeks

For 60,000 hits
2000 required to train

	Current	machine +1	n machines	n machines plus crowd sourcing*	Once trained
Selection	50	26	2	2	0
Meta-data	50	34	17	17	0
Data	100	68	34	20	0
Total	200	128	53	49	0/49
saving		36%	75%	75%	100%

*assumes machines cannot do the data extraction

For 60,000 hits
500 required to train

	Current	machine +1	n machines	n machines plus crowd sourcing	Once trained
Selection	50	26	0.5	0.5	0
Meta-data	50	26	4	4	0
Data	100	54	8	20	0
Total	200	106	12.5	24.5	0/24.5
saving		42%	94%	88%	100%

*assumes machines cannot do the data extraction

Issues of concern

- Feasibility
  - Especially of data extraction
- Correction for sequential analyses
- Publication bias
- Dissemination and uptake
- Scale
  - Cost per paper included in a systematic review is currently 40 minutes; this might reduce it to 4 minutes
  - If only 10% of the *in vivo* literature is relevant to research questions of interest, reviewing these systematically would take, at present, 80 people working flat out
  - With the efficiencies predicted here, that might reduce to 8 people working flat out (if we can find a big enough crowd)

Additional benefits

- By knowing the precision with which we know things, across a range of dimensions of evidence, we will be more certain of what we don’t know
- This can inform “research economics” decisions about where research resources might most effectively be deployed
- It will also inform Reduction, avoiding unknowing replication

The SLIM consortium

- Andrew Rice — Imperial College London
- David Howells — Florey, Melbourne
- Emily Sena — University of Edinburgh
- Julian Elliot — Burnett Institute, Melbourne
- John Ioannidis — METRICS, Stanford
- Nicky Welton — Bristol University
- Doug Altman — University of Oxford
- Sophia Ananiadou — University of Manchester

Jing Liao, University of Edinburgh

If you are planning a systematic review or meta-analysis of animal data, CAMARADES are here to help: malcolm.macleod@ed.ac.uk