Imperial College London

Should we retire statistical significance?

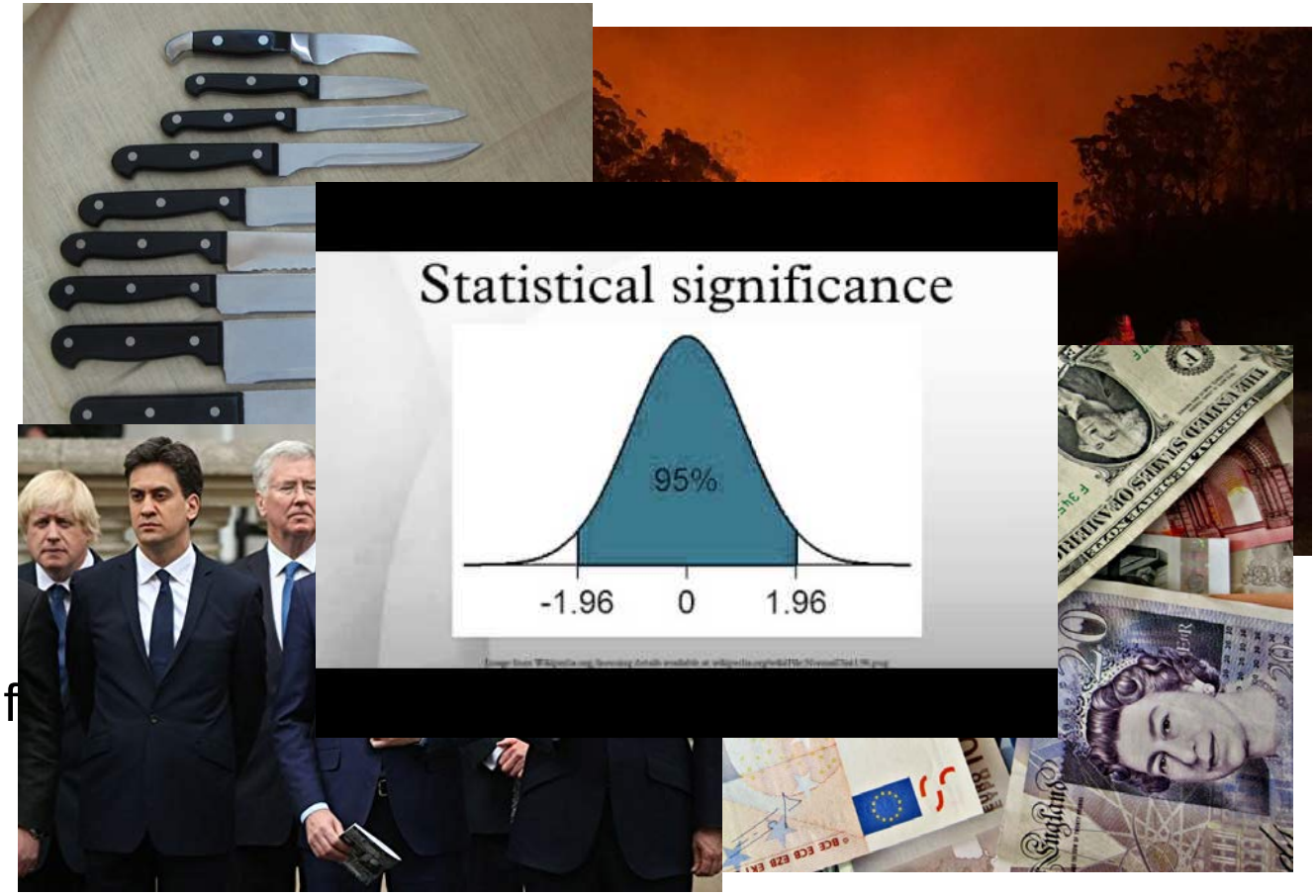
Professor Deborah Ashby
Director, School of Public Health,
Imperial College London
President, Royal Statistical Society

MiRoR Meeting, Paris
25 November, 2019

Imperial College London

What do the following have in common?

- Fire
- Money
- Knives
- Power
- Statistical signif



Imperial College London

Inappropriate use of significance testing: bioequivalence

- Not significantly different does NOT imply equivalence
- Need to specify equivalence margins for parameter of interest
  eg 90% CI for ratio of AUCs within (80%, 125%)
  (Drug regulation mid-1990's)
- "To be inside the acceptance interval the lower bound should be ≥ 80.00% when rounded to two decimal places and the upper bound should be ≤ 125.00% when rounded to two decimal places."
  - (CHMP Guideline on the Investigation of Bioequivalence 2010)

Imperial College London

Inappropriate use of significance testing: baseline testing

From the ‘editors concerns’ from NEJM clinical trial submission 2016:

6) The editors generally request that tables of baseline characteristics include information about the baseline variables, since strictly speaking all baseline differences have presumably occurred at random. However, we consider it useful to calculate the (nominal) P values for these comparisons, and to indicate with an asterisk and footnote which (if any) of the baseline characteristics differ at P <0.05. Please provide this information for new Table 1 and 2.

So, on the basis of choosing our fights carefully, we ended up with a similar statement to the other statistician’s as footnotes to the first two tables. We declined to use the recommended asterisk, we did give the actual P-value. Table 1 had no sig diffs, Table 2 had one at P=0.02. Quelle surprise!

(Correspondence with Doug Altman, who was facing similar battles with NEJM)

Imperial College London

Appropriate use of significance testing: baseline testing

- Fraud detection in clinical trials

Imperial College London

Quantifying heterogeneity in a meta-analysis

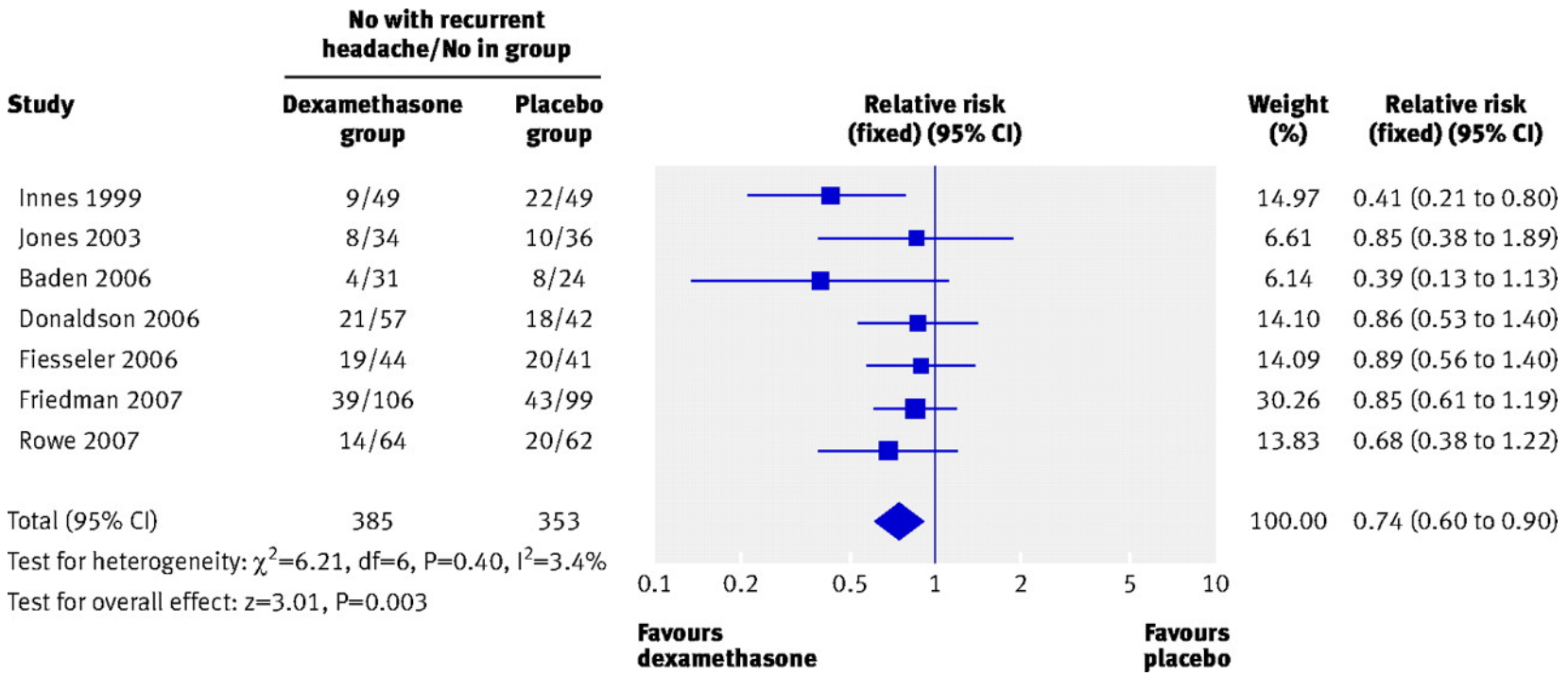
- I^2 much more informative than Q

(Julian Higgins and Simon Thompson SiM, 2002)

Imperial College London

Meta-Analysis: Forest plot

Forest plot of the effectiveness of dexamethasone in preventing the recurrence of acute severe migraine headache in adults compared with placebo.

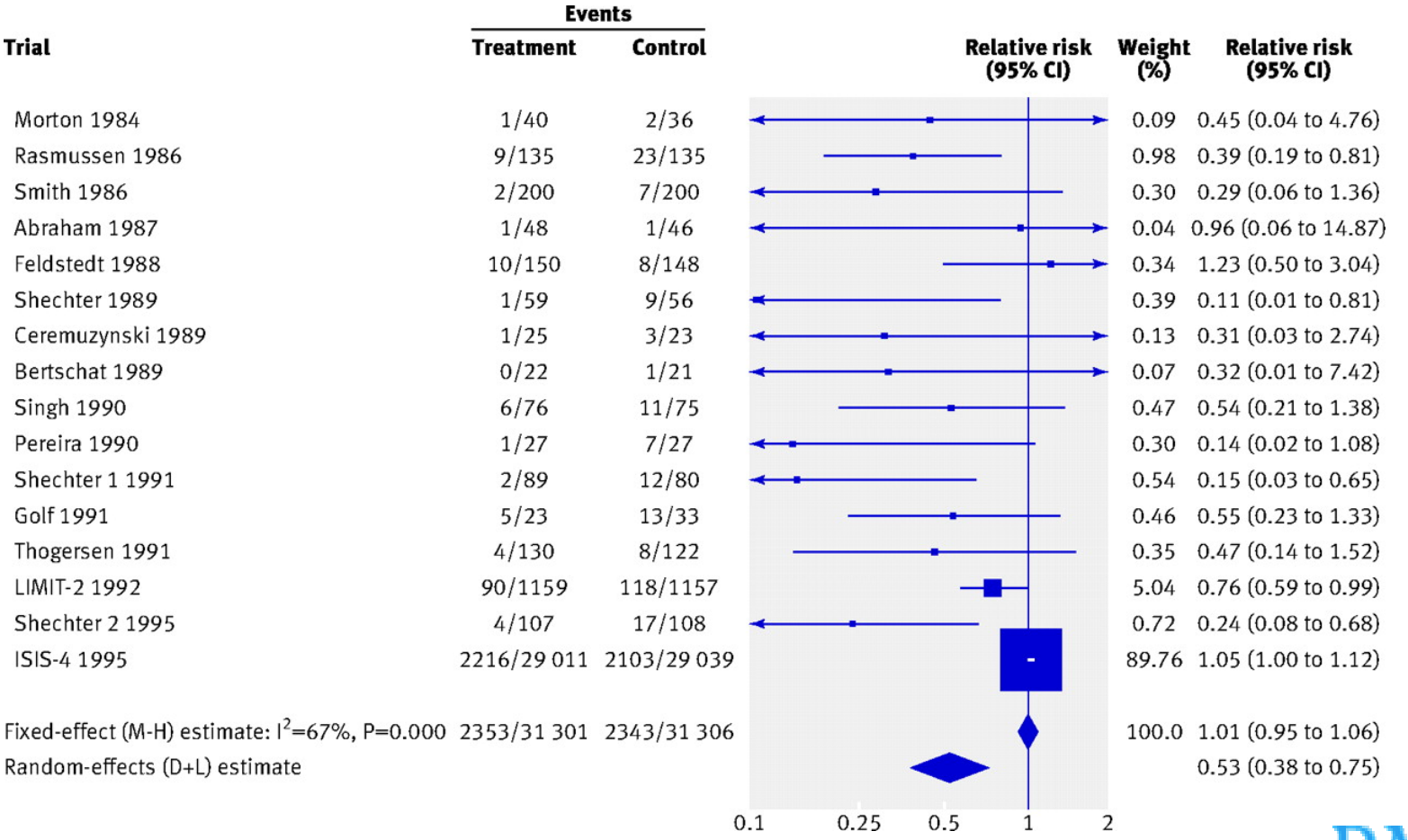


Sedgwick P BMJ 2011;341:bmj.d229

Imperial College London

Meta-Analysis: Forest plot

Comparison of fixed and random effects meta-analytical estimates of the effect of intravenous magnesium on mortality after myocardial infarction.



Sterne J A C et al. BMJ 2011;343:bmj.d4002

Imperial College London

Why is statistical significance so abused?

- Stems from a desire to make decisions
- Seems nicely binary with language of “reject” and “accept”
- cf dichotomisation
- We can always do better by focussing on estimation with its associated uncertainty

Imperial College London

Should we retire statistical significance?

Nature

Comment

20 March 2019

Scientists rise up against statistical significance

Valentic Armhein, Sander Greenland, Blakce McShance and more than 800 signatories call of an end to hyped claims and the dismissal of possibly crucial effects.

Imperial College London

RSS comments

- Deborah Ashby, RSS President

  “I understand the desire for a ‘simple’ rule of thumb – but a naïve interpretation of p-values can lead to seriously wrong conclusions such as whether medicines are effective of not. This is well understood by many but misused by many more. I’ve signed to help draw attention to the dangers of this outdated practice and promote the wider use of better alternatives.”

- David Spiegelhalter, RSS Past-President

  “I like p-values, but feel they are delicate things and should not be crudely split into “significant” and “not-significant”. I signed this article because I am fed up with researchers claiming a discovery when $p<0.05$, and claiming there is no effect when $p>0.05$.”

Imperial College London

RSS comments continued

- Guy Nason, RSS Vice-President for academic affairs

  “I signed the article because I agree with it! I am often surprised by how the outcomes of statistical methods are used to communicate results with unwarranted definiteness, based on assumptions, sometimes hidden, for which there is also usually considerable uncertainty. I particularly liked the article’s idea to talk about results’ compatibility with the data.”

- Stephen Senn, former RSS Council Member

  “Information is rarely dichotomous but decisions often are. Significance versus non-significance as a qualitative absolute distinction is ridiculous. As a threshold for action it may sometimes be justified but the appropriate standard will differ according to context.”

Imperial College London

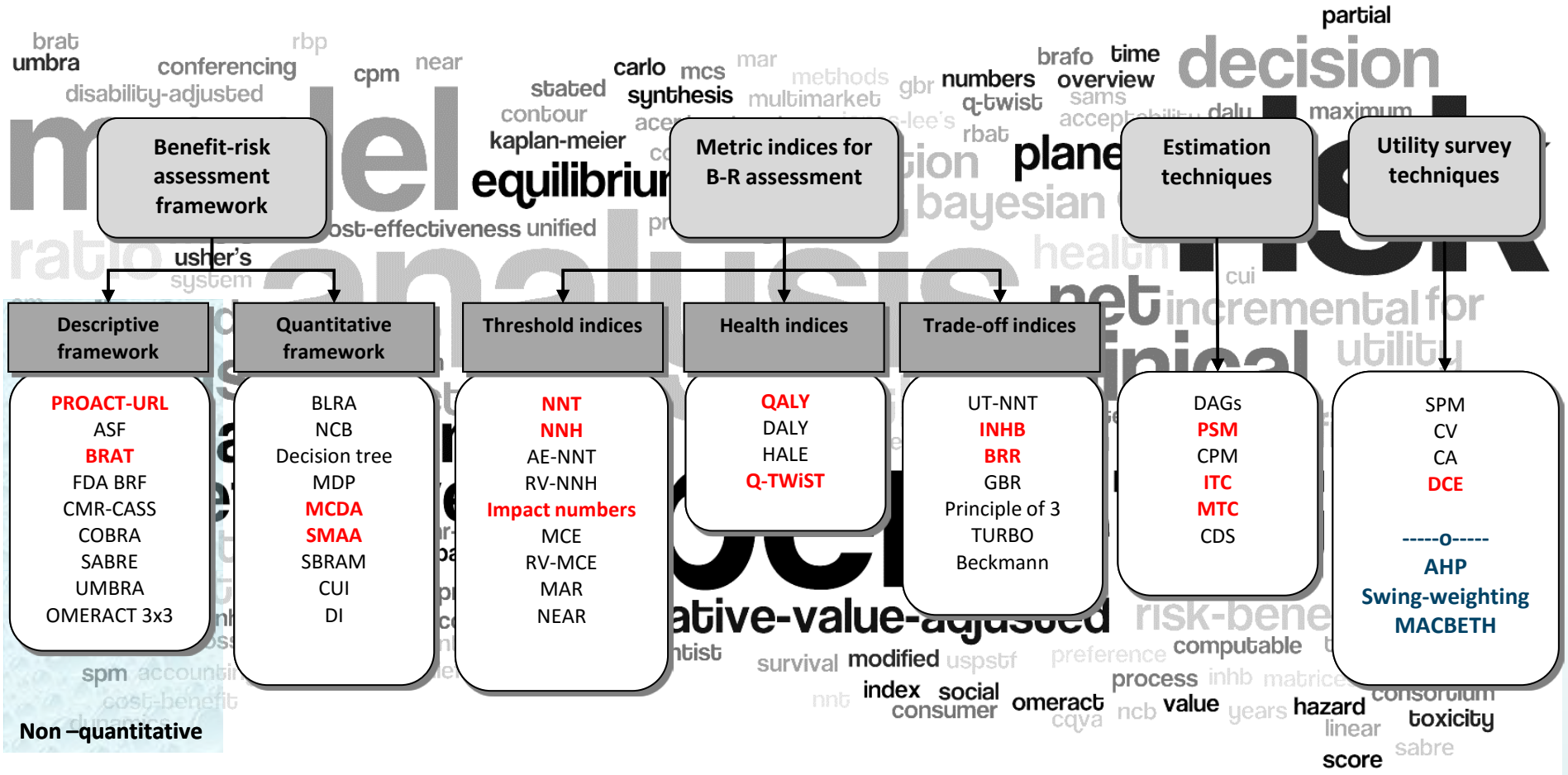
Decision making

- Proper approaches to decision making exist if we need to formalise it
- Decision-making under uncertainty closely allied with Bayesian statistics for decades, especially in health applications e.g. Raiffa, Schlaiffer, Cornfield, Lindley, Smith AFM, Smith J, Spiegelhalter, Berry, Parmigiani – see Ashby, SiM, 2006 for key references
- Extend uncertainty analysis in a probabilistic model
- Landscape for decisions through entire distributions
- Growing applications but there is still resistance

The licensing challenge

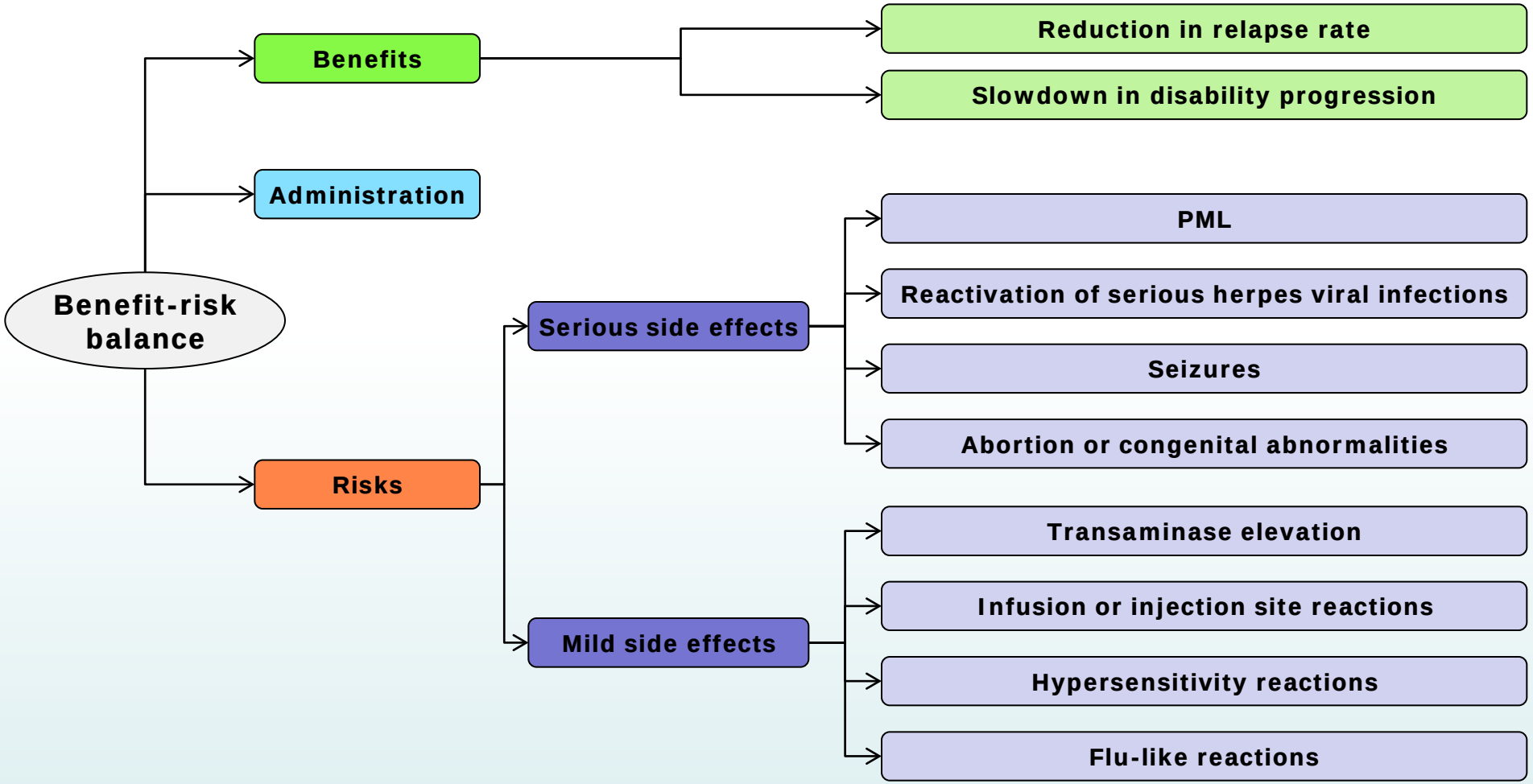
- The task of regulators (e.g. EMA, FDA) is to make a good and defensible decisions on which medicines should receive a license for which indications, based on the available evidence of risks and benefits
- It is increasingly important to be able to justify and explain these decisions to patients and other stakeholders.
- Can more formal approaches of decision-making, and especially more modern methods of graphical display help regulators do these better?

Methodologies available



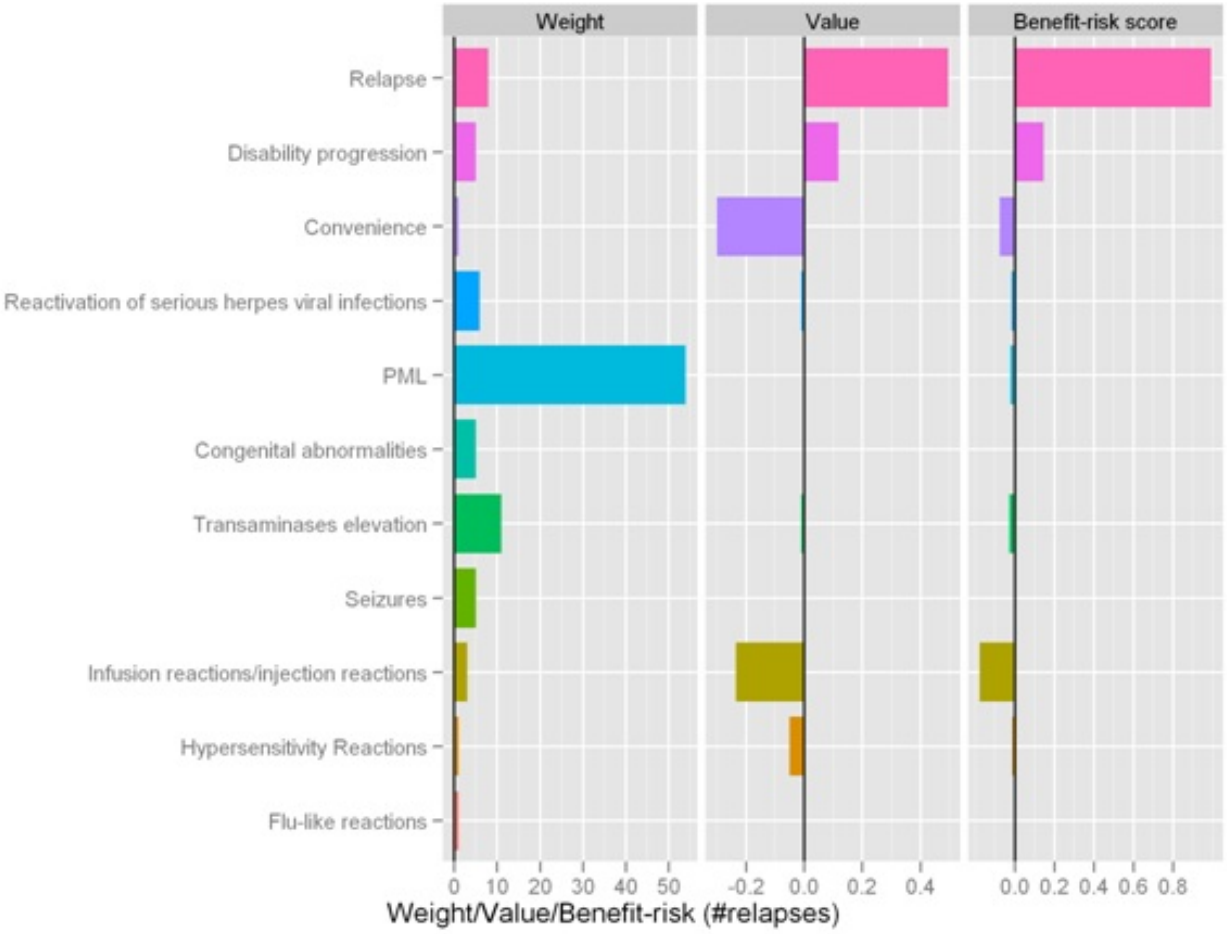
Mt-Isa *et al*. Balancing benefit and risk of medicines: a systematic review and classification of available methodologies. Pharmacoepidemiology and Drug Safety 2014. DOI: 10.1002/pds.3636.

Value tree: Natalizumab case study



Natalizumab: MCDA weighted utilities analysis

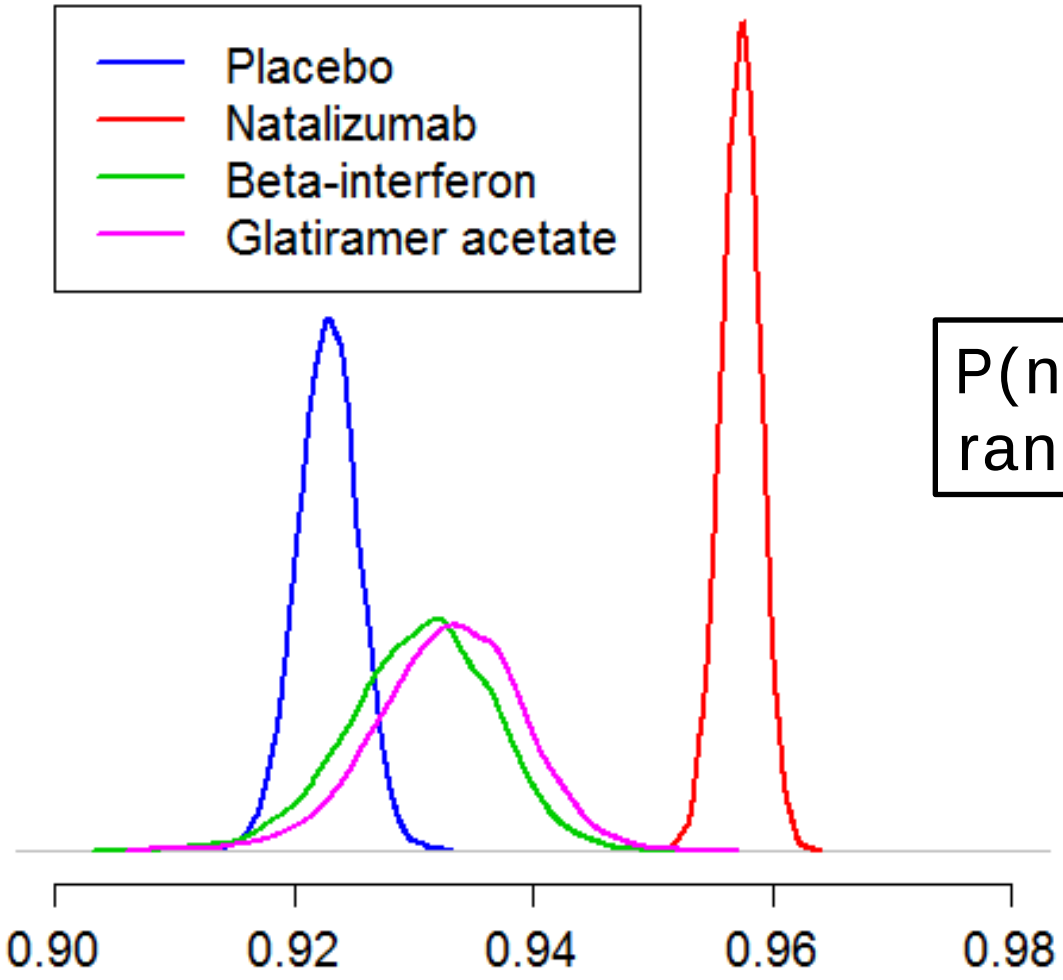
Contribution of each outcome for Natalizumab vs. placebo



- The Benefit-risk is the product of the weight and the value.
- Most of the Benefit-risk contribution is coming from prevention of relapses.
- Infusion site reactions are the worst risk

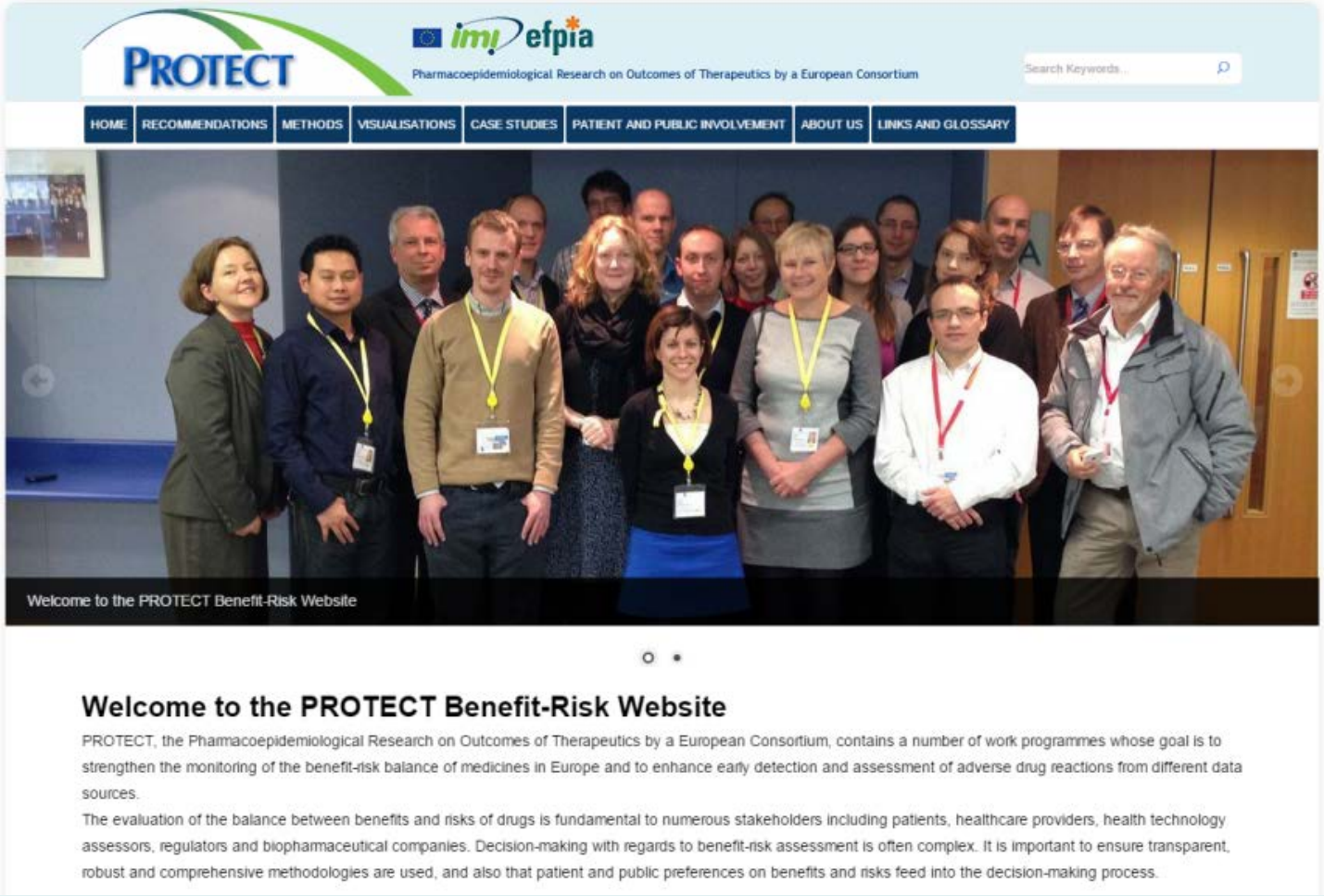
Natalizumab: Bayesian sensitivity analysis

Distribution of overall benefit-risk score



P(natalizumab ranked 1st) = 1

Dissemination and recommendations arising from PROTECT



http://PROTECTBenefitRisk.eu/

Imperial College London

Alternatives to statistical significance

- Estimate quantities of interest
- Show uncertainty from statistical modelling
- Acknowledge other sources of uncertainty e.g. uncertainty about model, biases
- Consider more formal approaches to decision making

Imperial College London

Significance from the RSS and ASA not to be retired!