STRATOS INITIATIVE

Update on p-values project

Michal Abrahamowicz, James Carpenter, Victor Kipnis

27th March 2023

James.Carpenter@lshtm.ac.uk

Aim:

- STRATOS paper/commentary on the p-value debate – ‘In defence of correct use of statistical significance’
  - Clarify issues (highlight inconsistencies in recent publications)
  - Present and interpret various ways p-values used
  - Underline the importance of reproducibility – highlighting that this does not mean reproducing the p-value
  - Discussion

Background/Rationale

- In March 2019 issue of ***Nature***:

  V. Amrhein, S. Greenland & B. McShane published the Comment "**Retire statistical significance**" [1], in which they recommended "*a stop to the use of P values in the conventional dichotomous way – to decide whether a result refutes or supports a scientific hypothesis*" and conclude: *"… **it's time for statistical significance to go**"*

- **The Comment was endorsed by >800 signatories**, mostly end-users of statistical methods, but also a few dozen well known statisticians, including a few STRATOS members **

  ** **Sampling properties of signatories selection are UNclear** ☺

- This <u>Comment has created a major confusion</u> among both:
  i. Non-statistical researchers, i.e. <u>End-users</u> (including Editors and Reviewers)
  ii. <u>Statisticians who Teach Applied Statistics</u>

[1] Amrhein *et al. Nature.* March 2019;567:305-307.

Selected Verbatim Citations from AGM's *Nature* Comment

- In the Opening 4 sentences Amrhein *et al* state:

  "When… you heard a… speaker claim there was 'no difference'… because the difference was 'statistically non-significant'? … We hope that… someone was perplexed if… ***a plot or table showed there actually was a difference*****. How do statistics so often lead scientists to deny* ***differences that those not educated in statistics can plainly see****?**"*

** AGM do NOT explain what is the Empirical Basis to establish that "*there* ***was a difference"*** or to ***"plainly see"*** such differences

Our and Other Statisticians' Concerns about AGM's "Black vs. White" recommendations

- **Removing the "gatekeeper" of statistical significance may open the floodgates toward an uncontrolled reporting of "associations" that may likely reflect just a combination of (i) sampling errors & (ii) Authors' wishful thinking**

- **Similar concerns** expressed (right after AGM Comment publication) **by other statisticians** [e.g. 2-5]:

  - E.g., **Julia Haaf *et al* state:** "… *when statistical testing is skipped, ... any differences between observations would be considered meaningful*" [3]

  - **John Ioannidis warns that removal of statistical significance**, a necessary "gatekeeper" to ensure **falsifiability** of the postulated scientific hypotheses [6], **may lead to** "*statistical anarchy*", and concludes "*Without clear rules for analyses, science and policy may rely less on data and evidence and more on subjective opinions and interpretations*" [5]

[2] Ioannidis, *Nature* 2019. [3] Johnson, *Nature* 2019. [4] Haaf et al, *Nature* 2019.
[5] Ioannidis, JAMA 2019. [6] Karl R. Popper, *The Logic of Scientific Discovery* 1959.

Examples of Impact in Empirical Studies that cite AGM's Comment

Panikkar *et al* [7], *Environmental Health* 2019 (IF = 4.7), state in Methods:

> "*To avoid placing too much emphasis on statistical significance, we emphasize the strength of associations in our results as well* [1]*."*
>
> (Similar statements in Methods of several other papers that cite AGM)

Then, in Results:

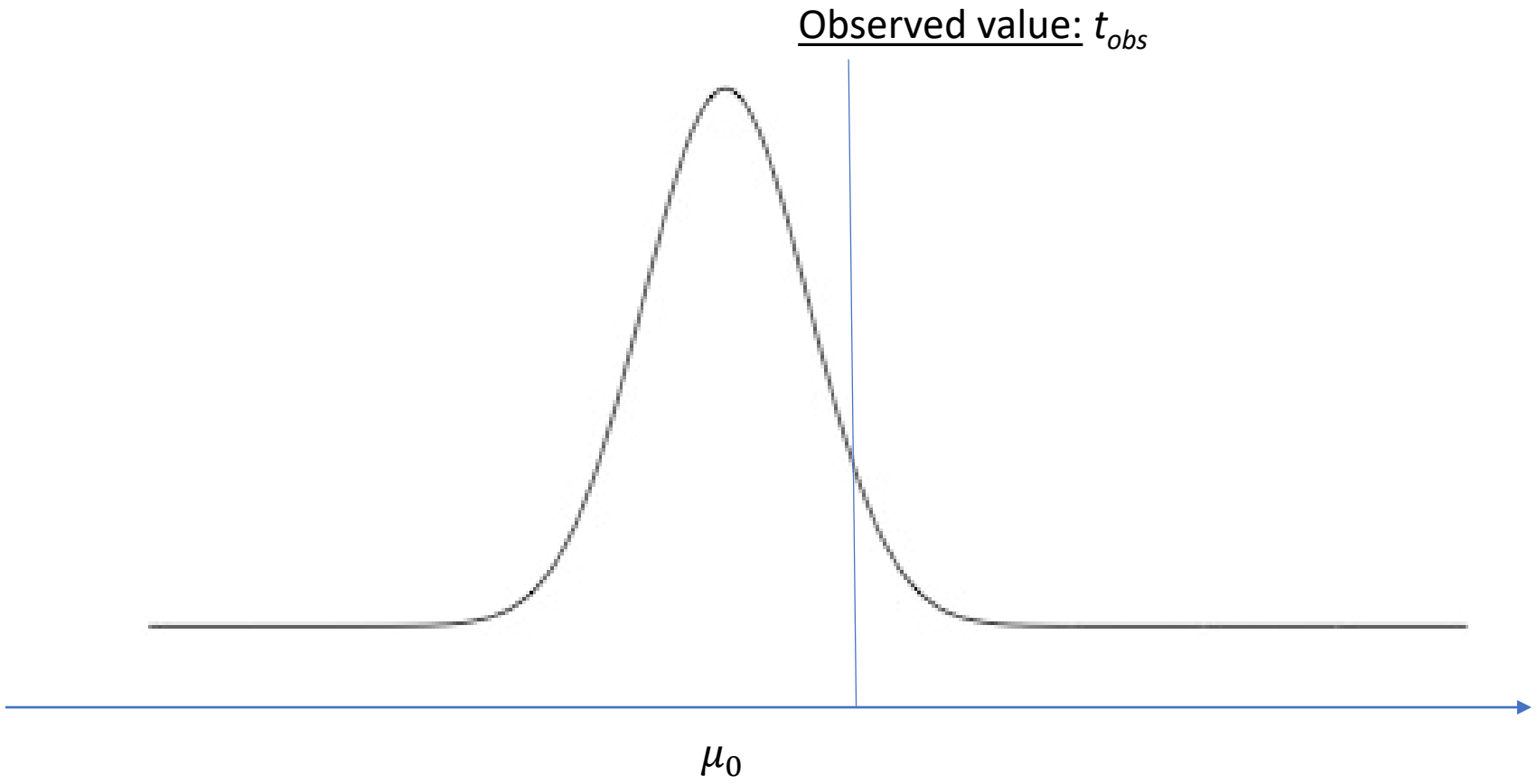
> "*Participants who had water filtration were also **close to 3 times more likely** to report developmental disorders (**OR = 2.960** (95%) CI: 0.7–12.8). … Residents who lived in Merrimack for 18–30 years (**OR = 4.966** 95% CI: 0.6–42.9) and over 30 years (**OR = 5.456** 95% CI: 0.3–90.6) were **5 times as likely** to report developmental problems.*" [7]

- Interpretating the point estimates as indicating "*close to 3 times*" or "*5 times*" risk increases illustrates the hazards of ignoring statistical (NON-)significance, and statistical inference in general
    i. All the three ORs would have a reasonable chance (>13% or >23%) of being observed even if there were no associations at all, with all **p-values >0.10 (0.14, 0.14 & 0.24)**
    ii. Furthermore, **the 95% CIs indicate that the point estimates are extremely imprecise**, and that the ranges of **ORs consistent with the observed results include even *important (up to 70%) risk reductions***!

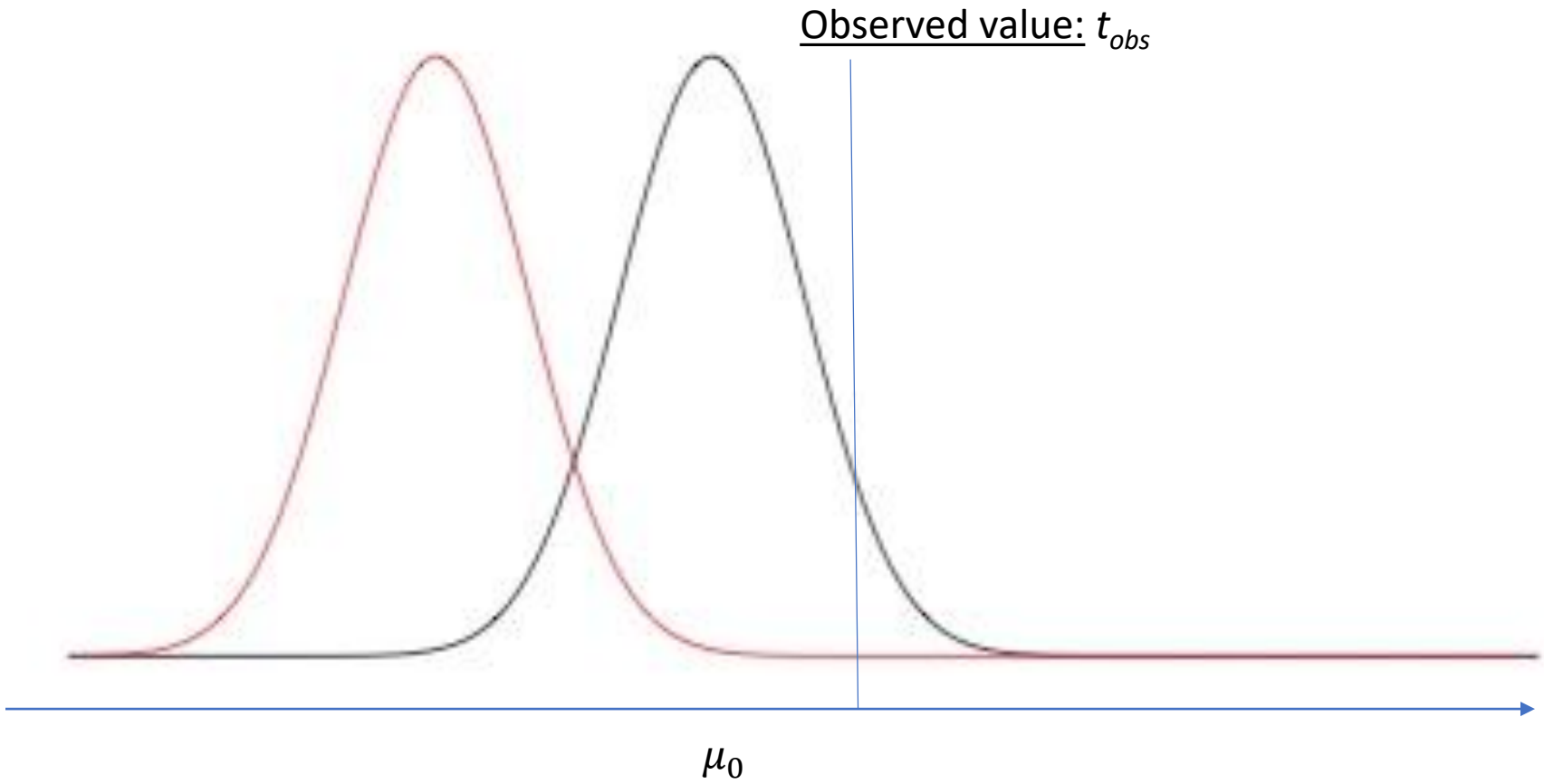
[1] Amrhein *et al. Nature* 2019. [7] Panikkar *et al,* Environ Health 2019.

Difficulties with interpreting p-values

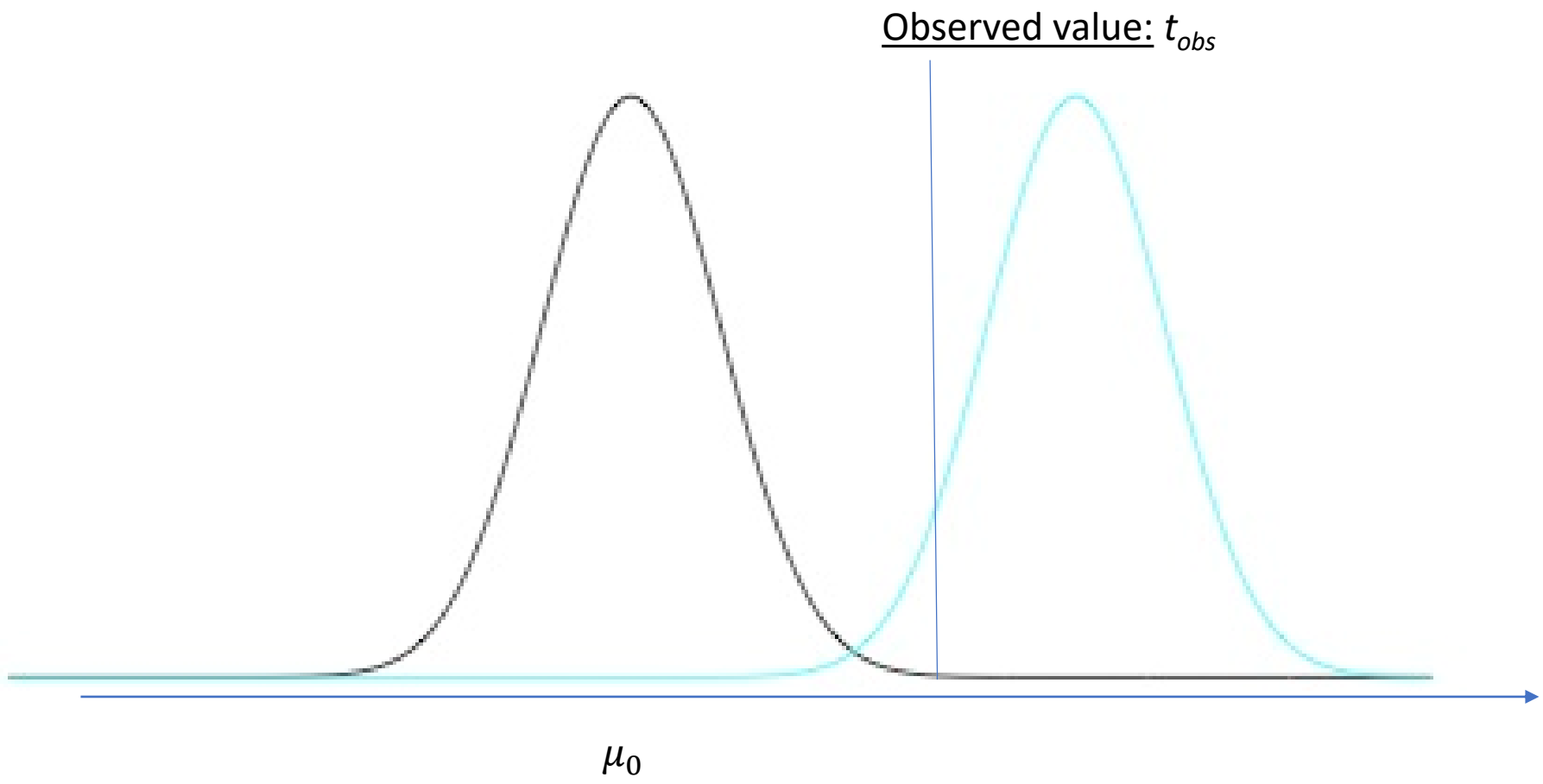
P-value (one sided) with no alternative



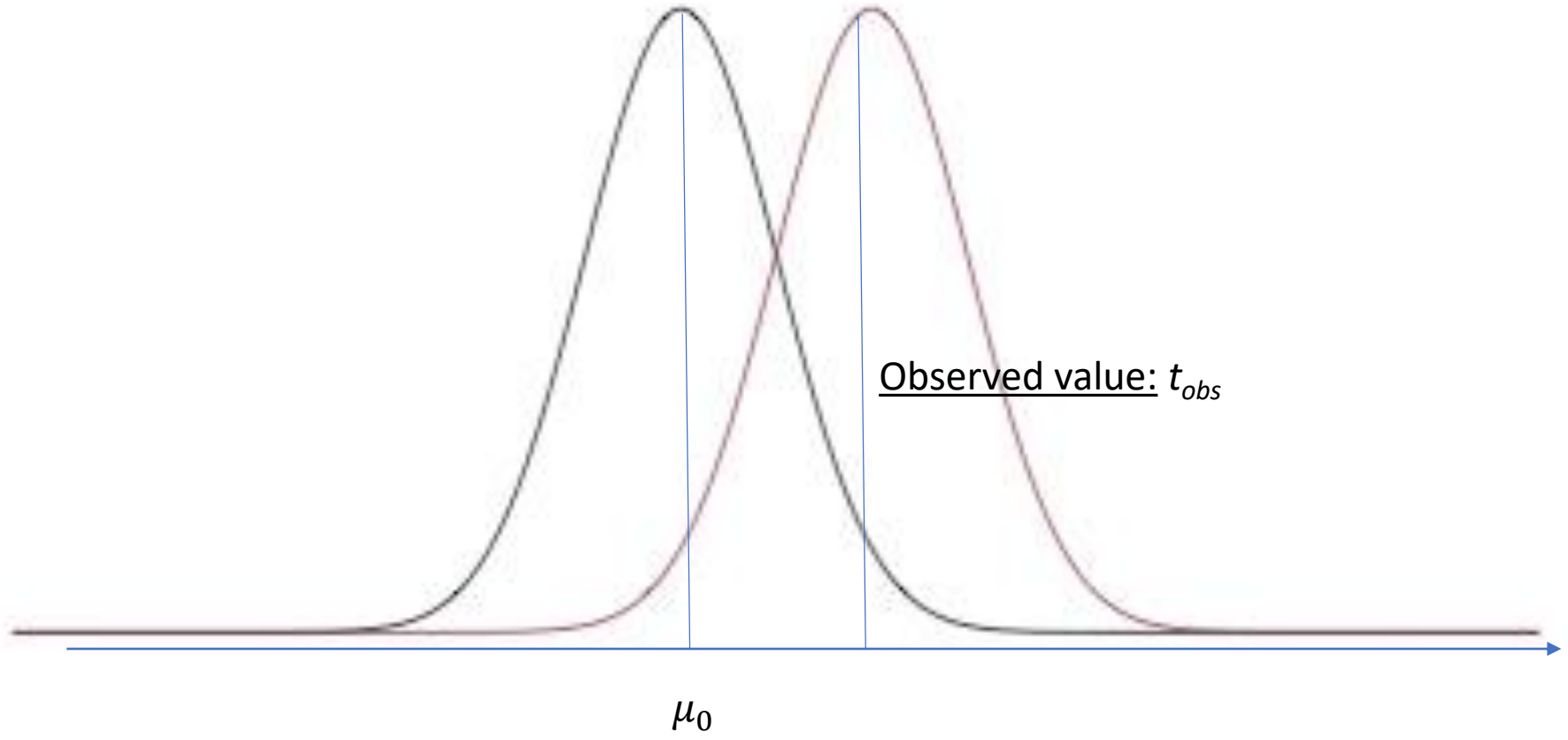
Alternative is less likely:



Alternative is more likely:



(informal) Bayesian interpretation:



Making decisions (NP vs Fisher)

- If we want to make decisions with defined operating characteristics, we need to specify the null and alternative distributions, and the decision boundary
  - We need to stick with this decision boundary
- Nevertheless, if we have a clear alternative in mind, we will feel less comfortable in rejecting H0 if we are just below the decision boundary
- In either case, reproducibility and meta-analysis are important

Reproducibility

Table: Probability of statistical significance ($p \leq 0.05$) upon repetition of an experiment as a function of the p-value of the initial experiment

Probability of $p \leq 0.05$ in a repeat experiment	
p-value of the initial experiment	$\mu = \bar{X}$ in the initial experiment
0.05	0.50
0.01	0.73
0.005	0.80

As the Table demonstrates, the replication probabilities are rather low,
have a weak relationship to the size of initial p-values, and are not in accord with
the informal credibility assigned to the null hypothesis when these p-values are observed.

Reproducibility – learning from trials

- Leading trials today are all registered on one of the established trials registers
- Typically, the trial protocol is published (aims, details of population, design, primary and secondary outcomes, analysis) aim of the study, details of the population, the design, the primary and secondary outcomes and how they will be measured.
- Increasingly, protocols are making use of the structured *estimands* framework, in which the researchers explicitly specify:

(a) the population under study

(b) the exposure

(c) the outcome measure

(d) the summary statistic

(e) how to deal with post-randomisation events.

Further…

- An agreed statistical analysis plan is also very important.
- This should specify
  - (i) the design (including why it was chosen);
  - (ii) the causal diagram;
  - (iii) the strategy for handling confounding (when not derived from the design) – e.g. propensity score weighting/matching;
  - (iv) the strategy for model selection, and selecting which confounders to adjust for (favouring pre-specification where possible) and
  - (v) clearly distinguish between primary, secondary and hypothesis generating analyses.
- Alongside this, consideration needs to be given about whether to attempt to control the overall type -1 error across a range of related analyses, and if so how.

Discussion

- In the next three months, we aim to produce a manuscript for circulation to a wider group covering these broad areas.

- We don’t think we will be able to reach complete agreement, but can nevertheless
  - Clarify the issues, and areas of broad agreement
  - Highlight some fallacies
  - Emphasize the importance of reproducibility
  - Clarify areas of disagreement