Causal analyses of existing databases: The importance of understanding what can be achieved with your data before analysis (commentary on Hernán)

Tim P. Morris*
tim.morris@ucl.ac.uk
MRC CLinical Trials Unit at UCL
90 High Holborn, London
WC1V 6LJ, UK

Maarten van Smeden
Julius Center for Health Sciences and Primary Care
University Medical Center,
Utrecht University,
Utrecht, Netherlands

Published doi:10.1016/j.jclinepi.2021.09.026

What is new?

- Power calculations may not be necessary for all causal analyses of existing databases, but a similar level of planning is still necessary.
- Churning out analyses of existing data and hoping they will contribute meaningfully to meta-analyses rather understates the difficulty of meta-analysis.
- Skipping careful thought about sample size may lead to point-estimate-is-*the*-effect interpretations.

Hernán's short communication provocatively argues that power calculations should not be required for causal analyses of existing databases[1]. His view is thought-provoking and helpful: contrasting it with the opposite extreme helps us to consider where we might sit between the two. We agree with many of Hernán's points. However, the proposal that, regardless of sample size, available data should be analysed because results can be synthesised in later meta-analyses seems to depend on a scientific utopia that does not reflect reality. In this note we describe what is unsatisfactory about Hernán's proposal from a practical perspective. We emphasise that we are not wedded to conducting power analysis for every analysis but, practically, planning and understanding what can be achieved with currently-available data remains important.

Hernán's *Hypothetical example* of rare thrombotic events among vaccinated young people motivates observational causal analyses when pre-approval of a vaccine was based on randomised trials. We broadly agree on the point that 'The goal of this analysis is not to "detect" a causal effect, but to quantify it as unbiasedly and precisely as possible'. We note in passing that the hypothetical 'socially alarmed' groups in the *Hypothetical example* may simply be interested in the binary signal of whether or not the unusual thrombotic events in vaccinated young people were made less unusual by the vaccine. However, we agree with the notion that the plausible magnitude of such an effect is important. Hernán's proposed solution is for groups to conduct causal analyses of several existing available data sources, the results of which would be synthesised in a meta-analysis. This is a worthy goal. It also places a possibly-unbearable burden on systematic reviewers. To explain why, we continue to work through Hernán's hypothetical example.

One group (A) with a reasonably large dataset estimate the risk ratio as 5.0 with 95% compatibility interval [0.58, 43], as given by Hernán (we follow Hernán's use of *compatibility* rather than *confidence*[2]). Other groups note the uncertainty and are buoyed by Prof. Hernán's encouragement. Group B, with a smaller dataset, work through their identification logic and aim to estimate a relative risk. There are separation issues due to sparse data [3, 4] but these go undetected and so they report a risk ratio of 15 [0.1, 2250]. Another group (C) with a large dataset observe zero events. They estimate a risk difference of 0 and, despite a lot of information in their data, are unsure how to estimate a confidence interval and therefore do not attempt

to do so. Yet another group (D), whose data contain more events, use logistic regression to estimate a conditional odds ratio after adjustment for several continuous and categorical variables. They report it as 0.5 [0.28, 0.9].

To a systematic-reviewer, the results of groups A to D may feel like like turning up with a dustpan and brush after an earthquake because:

1. Inference for meta-analysis with rare events is notoriously difficult[5, 6, 7];

2. The analyst is expected to give each result the appropriate weight where group C's analysis, containing a lot of statistical information, gives no measure of uncertainty, and group B's result appears to be an artefact of data sparsity[3];

3. None of the four results targeted the same estimand and so attempting to meta-analyse the aggregate data would be combining apples with oranges[8].

The meta-analyst decides that this accumulated evidence is more like a pileup. They do nothing and the socially alarmed groups are left with four sets of equivocal and possibly contradictory results; arguably more alarming than no information at all.

Had all four groups read Hernán and Robins' book[9], some – but not all – of the issues might have been alleviated. A better situation might have been achieved in two further ways: first, systematic reviewers could be involved at the stage of the individual-study analyses, something that is also helpful in trials[10], at least as far as choosing a common estimand across studies; second, each study might have planned for the information they might reasonably expect.

Hernán notes that the complex nature of observational datasets means that there are never appropriate sample-size formulas[1]. Though there have been attempts[11], it may in some cases be impossible to come up with sensible and usable formulas, as in the context of clinical prediction models[12]. To take into account the complex nature of causal analysis, simulation studies may be required. This is also true for most randomised trials, though approximate formulas are arguably closer to adequate there. Simulation studies are a sensible step for several reasons beyond simply grappling with adequate sample size. Everyone would agree that such studies involve many assumptions with a lot of inherent uncertainty. However, conducting them to aid planning is still worthwhile. For example:

- The analyst is forced to think about the data structure they might expect, even if this involves simplifications of reality, and whether their estimators are in fact unbiased.
- They experience issues such as separation or sparse data bias, most likely to occur at small sample sizes, and are forced to consider how this might be handled in analysis of real data.
- It becomes clear that these issues can be exacerbated by the adjustments required[3], and they must therefore consider contingency plans.
- If the above are properly addressed, they are able to consider the range of effects with which their results may be compatible and whether it is worth the effort of obtaining this amount of information.

When there is a fixed amount of available data, there are only two possible decisions: to analyse the available data or not. However, when data that could be analysed continues to accumulate, a third option is to decide how much more data is needed before attempting any analysis. Although Hernán explicitly focuses on *existing* databases, the choices still arise with prospectively collected observational data. Once *any* data have been collected, can they be analysed? Should data collection then stop? If not, how long should it continue before an analysis?

Finally, we note a possible abuse of Hernán's thoughtful arguments. Careless readers may take it for a suggestion that *any* sample size is acceptable when making causal inferences about important questions. We believe this is a real risk. While relying on significance tests to judge a causal effect as 'zero' or 'non-zero' may be unpalatable, a uniformly worse interpretation is that the point estimate is the effect, since a point estimate corresponds to a 0% compatibility interval. Simply treating such a quantity as The Effect is nonsense, but the practice seems to be increasingly common, perhaps in reaction to the use of statistical significance. Extreme point estimates are simply more likely to occur in smaller datasets due to higher variance. Consequently, more small-dataset analyses may create more false-alarms about health risks and overoptimism about the effectiveness of new treatments, and dominate the health news in popular media, negatively affecting the reputation of our field.

In summary, we agree with several of Hernán's points but his proposed solution

would require a level of planning *beyond* simple power calculations rather than simply skipping that step. If the idea is for synthesis of several results to be the tool for summarising the overall results, involvement of systematic review professionals in the individual study analyses is advisable, as well as sharing of data to facilitate individual participant data meta-analysis[13].

Acknowledgements

Tim Morris was funded by MRC grant MC_UU_00004/07.

Declaration of interest

None

References

[1] M. A. Hernán. Causal analyses of existing databases: no power calculations required. *Journal of Clinical Epidemiology*, 2021.

[2] Z. Rafi and S. Greenland. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology*, 20(1), 2020.

[3] S. Greenland, M. Ali Mansournia, and D. G. Altman. Sparse data bias: a problem hiding in plain sight. *BMJ*, page i1981, 2016.

[4] M. van Smeden, J. A. H. de Groot, K. G. M. Moons, G. S. Collins, D. G. Altman, M. J. C. Eijkemans, and J. B. Reitsma. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1), 2016.

[5] M. J. Bradburn, J. J. Deeks, J. A. Berlin, and A. R. Localio. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine*, 26(1):53–77, 2006.

[6] F. C. Warren, K. R. Abrams, S. Golder, and A. J. Sutton. Systematic review of methods used in meta-analyses where a primary outcome is an adverse or unintended event. *BMC Medical Research Methodology*, 12(1), 2012.

[7] O. Efthimiou. Practical guide to the meta-analysis of rare events. *Evidence Based Mental Health*, 21(2):72–76, 2018.

[8] R. Daniel, J. Zhang, and D. Farewell. Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, 63(3):528–557, 2021.

[9] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[10] J. F. Tierney, D. J. Fisher, C. L. Vale, S. Burdett, L. H. Rydzewska, E. Rogozińska, P. J. Godolphin, I. R. White, and M. K. B. Parmar. A framework for prospective, adaptive meta-analysis (FAME) of aggregate data from randomised trials. *PLOS Medicine*, 18(5):e1003629, 2021.

[11] E. Demidenko. Sample size determination for logistic regression revisited. *Statistics in Medicine*, 26(18):3385–3397, 2006.

[12] R. D. Riley, J. Ensor, K. I. E. Snell, F. E. Harrell, G. P. Martin, J. B. Reitsma, K. G. M. Moons, G Collins, and M. van Smeden. Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368, 2020.

[13] R. D. Riley, J. F. Tierney, and L. A. Stewart. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Wiley, 2021.