Experimental design dilemmas - discussion

This handout supports the NBI DTP Experimental Design workshop in Demeber 2022. Most of these issues are explored at greater length in the documents (books and reports) we discussed during the session, and in many other sources, and you should read around issues that might be relevant to your own work.

1. Early diagnosis (selection)

You want to study the effect of gut microbial composition on Parkinson's disease, using an observational design. Your plan is to compare the microbiomes of people with and without Parkinson's disease to look for associations.

How should you select your cases and controls?

Here the research question and associated theory are not specified well enough to design the study. We might be interested in whether microbiome composition causes PD, whether it contributes to Parkinsonian symptoms, or whether PD causes changes in the microbiome. Each of these causal hypotheses would require a different study design to test.

But in each case we should select our participants keeping in mind the goals of reducing bias (through selection bias or confounders) and reducing variance (by measuring or matching on factors that might affect the microbiome). We also need to keep in mind the sample size required, which will depend on the size of the differences we are expecting to detect, and the expected variability of the microbiome between individuals.

Suppose we wanted to test whether a particular feature of the microbiome cases PD. The ideal design would be a *cohort study*, whereby we recruit a cohort of participants without PD, measure their microbiomes, and then wait to see which participants develop PD in the future. However we would need to recruit (and retain) a lot of people in the study, probably thousands, to be reasonably sure of getting enough cases to analyse. It would also take a long time to conduct.

Instead we might decide to run a *case-control study*, whereby we select patients already with PD and compare them to people without PD. Clearly it is quicker and easier to recruit the people we want to compare in this way, but we can no longer measure the microbiome before disease incidence. So it will be more difficult to rule out confounding factors (such as treatments) or reverse causation (suppose early PD causes a dysbiosis).

We will also be faced with the question of what is a 'healthy control', and how they should be selected. Ideally for each case we should select one or more controls who are as similar as possible to the case, except that they do not have PD, and the cases should have an exposure pattern that mirrors that of the general population.

Typically studies will recruit household controls, although it will be difficult to match other characteristics such as age and sex exactly within a household, and it might be that the exposures that contribute to the PD microbiome also affect the household control, in which case we wouldn't observe the effect we are interested in.

2. Trust me, I'm a doctor (feasibility / human studies)

You worked with a statistician to calculate that you need 40 participants for your endoscopy study. Your second supervisor (a senior clinician) says this isn't a problem as their clinic is always busy with patients, so you shouldn't worry about it. However your inclusion criteria are quite strict and you aren't sure how easily you can recruit people for what is quite an invasive study.

What should you do?

Clinical collaborators can be optimistic when thinking about the numbers of patients they are likely to be able to recruit, and studies often recruit far fewer patients than they expected to. This can be a fatal problem for a study, and if it is to be the main source of data for your PhD then it's a problem for you as well. Scientists may have an optimistic view if they are used to working with healthy volunteer cohorts where a low response rate doesn't matter (because the pool of participants is very large), instead of smaller cohorts of potentially unwell patient groups where it does, and where there may be more resistance to study participation.

Ethics committees and grant funders *should* require evidence that the recruitment plans are feasible, yet ethics committees often do not do this, and PhD projects and ISP funded work are rarely scrutinised this closely at the individual project level.

To work out how many people you are likely to recruit you need to know (1) how many people are in your sampling frame (2) how many of these are likely to meet your inclusion criteria (3) how many will be approached and (4) how many are likely to participate.

Your collaborators should be able to give you the details of how many people go through their clinics, and what proportion of these will meet your criteria. The number that can be approached will depend on the staffing resource available (eg if you are doing the recruiting how much time can you realistically give it, subtracting holidays, any sick leave, other work etc) and how this is to be done.

It's not easy to work out how many people are likely to agree to participate, but if the clinic or CRF has run similar research before they might be able to estimate this, or you might find typical response rates from similar published studies. Finally, it can be helpful to pilot a study or to run a small focus group with patients to assess likely response, and to identify aspects of the study that might be off-putting and that you can correct in the main study to recruit more patients.

3. Optional stopping (p-hacking / Questionable Research Practices)

You have completed an experiment into the effect of a new growth medium on growth in wheat plants. You get a p-value for the effect of 0.07! Your supervisor suggests that if you grow a few more plants then analyse all the data together then you might get down to p<0.05 so that you can publish. Alternatively, you could switch the focus of the paper to a different outcome measure where you did see a significant difference.

Are these strategies acceptable? What should you do? What could you have done at the outset to avoid the issue?

There are a few issues to unpack here.

In short, both of these suggestions (adding more data following the result or changing outcome measure), while common, are examples of 'p-hacking', and should not be done. P-hacking occurs when our experiment did not work the way we hoped and so we change our experimental plan or analysis plan *post-hoc* until we find a way of looking at the data that does look like our expectation (or looks interesting enough to publish). We then imply (usually by omission) that this revised plan was our intention all along.

By giving ourselves more opportunities to conduct hypothesis tests (with different subsets of data or different outcome measures) we are creating more chances for false positive results. So, findings that are the result of this kind of p-hacking are much less likely to be real.

There are two potential remedies here. The first is to pre-register your experimental plan, and then stick to it. If the plan is written down and published it is much more difficult to deviate from without good reason.

Second, the idea that only 'significant' studies can be published should be resisted. A well-designed study with a negative result is still valuable, and we should disseminate its findings. Only publishing 'successful' studies badly distorts science and leads to publication bias, reinforcement of bad ideas and research waste, as well as the temptation to 'p-hack' the data to create a 'publishable' work.

4. Inadequate sample size (power)

You want to know whether a difference in bread starch content can reduce stomach bloating. The only feasible way to measure this is stomach volume measured by MRI scanner. You can afford to test 20 patients using this technology, but a sample size calculation suggests you need 80 to have a good chance of detecting the effect you are interested in.

What should you do?

An underpowered study has a low chance of detecting an effect, even when the effect exists and would be clinically important.

A corollary of this is that 'positive' findings from underpowered studies are very likely to be false positives since the chance of getting a 'true' positive is low. So the value of this study is questionable, and it might not be worth pursing since it will be difficult to interpret either a positive or a negative result. We know that too many research studies are conducted that have no real chance of finding anything useful, and it is probably better to choose another research question in this case.

Nevertheless, if it is impossible to get more resources or redesign the study, *and* the research is necessary, then it might be justifiable to conduct an underpowered study, even though any effect estimate coming from it will be very imprecise. In this case it is important to be honest with the study sponsors, ethics committees etc regarding the issue, and with readers in your write up, so they can judge whether the cost of the study and the risk to participants is worth it, and can interpret the results accordingly.

5. Replications give different results (replication / significance)

You run a colon model experiment into the effect of different substrates on short chain fatty acid production using stool samples from human donors. You see a significant improvement with your experimental substrate compared to a control.

However when the experiment is repeated (with exactly the same conditions), the difference was not statistically significant. Your supervisor suggests running a third replication to see which was correct, again this is not statistically significant.

What could have happened here? What would you conclude? What should you do? What could you have done differently at the outset?

This scenario is not unusual. P-values are not reliable, they will often come out differently even though study set-ups are identical. This does not reflect any failure of the experimental set up, just natural variation. So the idea that a 'significant' study if replicated should also lead to a 'significant' study, is simply wrong.

Contrary to popular belief, you cannot accumulate evidence for or against an effect by running lots of small studies and considering whether or not they individually have 'significant' p-values.

So what should you conclude? The best answer is probably to pool the data from your three studies and to analyse them together as a single experiment. Your analysis will need to account for the fact that your experiment was conducted in three 'blocks' but this isn't usually a huge problem for modern statistical software. In any case you should discuss this with a statistician, since the fact that the second and third studies were conducted conditional on the results of the first might also be an issue.

If you *must* analyse the studies separately, then a formal meta-analysis, whereby the individual results are pooled into a single result, is the appropriate way to combine the results.

This scenario would make me wonder if conducting two or three experiments like this was always the plan. If so, and if you have the resource to do it, then you should instead plan it as a single large experiment with blocks at the outset, rather than three small ones each to be analysed separately. This will be more powerful and estimates coming from it will be more precise.

6. **Mouse microbiome study with limited resources (confounding / pseudoreplication)**

You are conducting an experiment on germ-free mice, to test the effect of an experimental bacteria on stress hormones (compared to untreated control). You only have access to two isolators.

You know that you cannot house treated and untreated mice in the same isolator because there will be contamination between the groups. But if you house all of the treated mice together and all of the untreated mice together you will have completely confounded any isolator effect with the treatment effect, so the treatment effect cannot be identified.

What should you do?

In this scenario there is only effectively one independent experimental unit per treatment condition, and so there is no chance of estimating the treatment effect, unless you know there is no chance of an isolator effect or contamination between mice within the isolator, (this seems unlikely, we know that there are cage/isolator effects in many mouse studies), or that the effect you are expecting to see will be so great that it will overwhelm any possible cage effect.

If you can run your experiment over a longer time-period you could run several repeats of the experiment, randomly allocating the isolator to the treatment each time. This will be expensive and time consuming but is probably necessary in this case to make any meaningful inference.

Although there could be many mice in each isolator, they are not true 'biological' replicates because their outcomes will not be independent of one another. Mice within cages (or isolators) are examples of 'pseudo-replicates' which are commonly but wrongly analysed as if they are true replicates.

7. **Sample size calculation when no data is available (sample size)**

You are planning an experiment and need a sample size calculation. Your sample size calculator requires you to enter the variance of the outcome measure, and what is the smallest target effect size you want to be able to detect. Unfortunately, you have no prior information with which to calculate either of these things.

What should you do?

Without any information about how much your outcome will vary, or what kind of effect size you are looking for, it is difficult to design an experiment to test your effect. However it is rarely the case that there is absolutely no prior data on your outcome measure that you could use, either from your own lab or in the published literature.

For an analogy, think about trying to look for a physical object when you don't know how large it is, how precise your instrument is, or whether the place you are looking is lit well enough to see it anyway! You would never be able to conclude the object isn't there, even if you didn't see it. Some statisticians would argue that without this information you are in no position to even conduct an experiment, and you should run more pilot experiments to get the data on how variable your outcome is. However, to get a reliable estimate of the variance of an outcome might be more difficult and costly than running the main experiment itself, so this is unlikely to be practical, especially for small scale studies.

If you genuinely don't know anything about the outcome you are measuring or the likely effect size you are looking for, then you are conducting an <u>exploratory</u> experiment. You can use *Mead's resource equation* to calculate a sample size for a preliminary estimate of effect, but this is a very crude method. I also run a separate class on sample size calculation in March each year, but there is extensive literature on the subject for all kinds of study designs.

8. **Choosing between two analyses that give you different answers (p-hacking, pre-registration, test validity)**

You have conducted an experiment to check whether mice have better blood sugar on an experimental diet compared to their usual diet.

When using the conventional t-test that you had planned, there is no evidence of a difference. But a post-doc in your lab shows you a new statistical test (probably involving AI) that suggests the groups are significantly different.

What should you do? Which should you report?

Choosing different statistical tests until one gives you a significant answer is a good example of p-hacking, and should not be done. You should, in most cases, stick with the test you had planned to use before you saw any data, so long as its assumptions are met.

Still, all statistical tests have their strengths and limitations, and it might be worth exploring why one test gives you a different result to another, in case the newer test is more valid or efficient for your particular dataset or you think you might use it in future.

You should not report results from a statistical test without understanding whether it is valid for your particular dataset, so its important to understand how it works and what assumptions it makes about your data. There are usually many different approaches to solving any problem in statistics, and there are lots of new statistical methods under the label of ‘machine learning’ or ‘AI’, so do explore these but keep in mind that they are very new, may not be understood by your readers and might produce spurious results.

Also be wary of automatic statistical tests that are conducted as part of data analysis pipelines. While these are convenient and may work in some cases, it is important that the assumptions underlying a statistical test are checked against the specific data that are being used, and we do not just assume that since a test has worked for a prior dataset that it will continue to do so for your own data! In particular automated analysis often cannot incorporate experimental designs with covariates, blocks or pseudo-replication, so be aware of this.