How to write a data analysis plan

Steve Simon

April 4, 2018

Who is this guy talking to you?

- Steve Simon
- Department of Biomedical and Health Informatics, UMKC
- Enterprise Analytics, KUMC
- P.Mean Consulting

I have helped hundreds of researchers write grants. Most of the assistance was on very small grants ($25,000 or less). Many of the interactions were brief, but some were quite intense.

I've been funded on a half dozen big grants, but never as a principal investigator.

Who is this guy talking to you, continued.

I have co-taught short courses at the 2012 and 2014 International Conference on Integrative Medicine and Health on how to write a CAM grant.

I have attended dozens of continuing education opportunities on grant writing.

3/40

Research story

I was working with a colleague in the federal government who wanted some advice about a statistical approach that his collaborators at a different federal agency had proposed. He warned me that his collaborators were insisting on their method because it was delivered by their "statistical gods."

At first, I was jealous. If people thought of me as a god, as someone with supernatural power, then I could probably double my consulting rate. I think that a statistical god could get at least $250 an hour, don't you?

But then it bothered me a bit.

Research story, continued

How did I learn all of the Statistics that I know? It wasn't delivered on stone tablets and it did not come in a vision. I learned statistics the old fashioned way, by reading books and attending classes and (more recently) through continuing education.

There's nothing supernatural about Statistics. It takes work, but no more work than anything else.

So don't think about statistics as if it were some supernatural power that is beyond the reach of ordinary mortals like yourself.

Overview of today's talk

In today's talk you will learn how to approach the data analysis plan that you have to write for your research grant. What I cover here also applies for the data analysis plan that you have to write for an IRB submission, assuming that you are working with human subjects, or IACUC if you are working with animals.

This talk will cover writing, so there are no formulas.

Overview of today's talk, continued

I interpret the data analysis plan broadly to include

1. sample size justification
2. survey instruments
3. data management
4. data sharing plan

in addition to the data analysis itself.

1. Sample size justification - two don'ts

Two things you should never use to justify your sample size.

1. It's all we can afford given your (pathetically low) budget limitations.
2. It's all we can get given your (pathetically short) time limitations.

Instead, you need to demonstrate that you can accomplish great things within the constraints of the granting agency.

8/40

1. Sample size justification - the three things you need

Typically, there are three things you need to justify your sample size.

- Research hypothesis
- Standard deviation of your outcome measure
- Minimum scientifically important difference

Some complex analyses (e.g., multi-level models) may need more information, but for most simple studies these three are sufficient.

Make sure that you specify both the standard deviation and the minimum scientifically important difference in your write-up.

1. Sample size justification - binary outcomes

For binary outcomes, the standard deviation is replaced by the proportion observed in the control group.

Rare events require much larger sample sizes, in general. You should strive, as a rule of thumb, for 25 to 50 events in your control group.

10/40

1. Sample size justification - examples

Sample size was estimated using a target power of 80%, at a type I error rate of 0.05 and was calculated relative to the primary outcome measure; Pain severity subscale of the BPI. The statistical test assumed was an independent samples t test for group differences in the change from baseline to subsequent assessment, assuming that the randomisation ensures no systematic baseline or other covariate group differences. The minimal clinically significant difference for the interference scale of the BPI is 1 unit (standard deviation of improvement of 2 units) [40]. Calculation produced a suggested sample size of 64 per group. Allowing for potential attrition rate of 20% our final sample size is 80 participants per group.

1. Sample size justification - examples

Sample size calculation was based on the primary outcome (number of volunteers reaching the clinical endpoint of â‰¤4 sites with PD â‰¥ 5 mm at 1 year post-therapy). Considering a difference of 31 percentage points between groups (31% vs 62%) as regards the primary outcome [14], a significance level of 5%, and 90% power, 50 subjects per group would be necessary. Considering a 20% rate of loss to follow up, it would be necessary to include 60 volunteers per group (total 180 subjects).

12/40

1. Sample size justification - no research hypothesis

Question: I don't have a research hypothesis. How do I justify my sample size?

In a quantitative study without a research hypothesis, your goal is often estimation. You want to select a sample size so that your estimate is reasonably precise.

In a qualitative study, the justification of the sample size is often itself qualitative.

1. Sample size justification - don't ignore time

If you are collecting data prospectively, you should always specify a time frame for completing your study and provide data to support that you can collect the required sample size in that time frame.

Using UK national audit data (2010â€“2011), 55,452 patients (for MI, PCI, CABG) engaged in cardiac rehabilitation with one of 280 teams (an average of 200 patients per team per year). Assuming that 17% had depressive symptoms [27], this equated to 35 eligible patients per team each year.

14/40

1. Sample size justification - don't ignore time

Cancer.net. One in Five Clinical Trials for Adults with Cancer Never Finish â€“ New Study Examines the Reasons. Genitourinary Cancers Symposium. January 28, 2014. Available at http://www.cancer.net/one-five-clinical-trials-adults-cancer-never-finish-%E2%80%93-new-study-examines-reasons.

15/40

1. Sample size justification - pilot studies

Pilot studies are not a label you can slap on a when you can't properly justify your sample size. Here is a bad example.

Because of the lack of adequate preliminary studies, we adopted a pilot study design with 24 participants in each group, considering the limited research funds, study period, and recruitment opportunities.

16/40

1. Sample size justification - pilot studies

A pilot helps with the planning of a large scale study. It examines resource requirements, tests the ruggedness of novel aspects of the research, and identifies areas where Murphy's Law might strike.

Since the goals of a pilot study are often qualitative, your sample size justification for a pilot study is also often qualitative.

1. Sample size justification - pilot studies

Here's a good example of justifying a pilot study.

Secondary aims of our pilot trial are to inform our intervention content, delivery, technology as well as software and training requirements. We will also gather data to enhance our understanding of anticipated recruitment, intervention adherence and dropout and future sample size calculations and our choice of outcomes. We also aim to explore the links between our intervention and naming, functional communication, quality of life and other language impairments than naming at 4 months post randomization, and whether it is sustainable and feasible with regards to ethical, technical, logistic, patient and data safety aspects.

18/40

2. Questionnaires and surveys

Always try to use existing questionnaires without modification.

Modifications will hamper the properties of the questionnaire, including validity and reliability.

Types of modifications that raise concern

- translation to a different language,
- adding extra questions,
- dropping questions,
- changing the wording of a question
- changing the numeric scale (e.g., 5 point Likert scale)

2. Questionnaires and surveys

Modifications are sometimes necessary. Document the process of this modification.

The questionnaire was designed based on a literature review, prior case studies (Claro et al. 2003; Fischer et al. 2009), and also drawing from interviews with regional stakeholders (officers of the regional authority, representatives of PDO consortia, and members of trade associations).

20/40

2. Questionnaires and surveys

If you are foolish enough to develop a questionnaire from scratch, document this process.

Include a review by experts (other than yourself).

This initial list of items was subjected to expert judgment for redundancy, content validity, clarity, and readability (following Dornyei, 2003).

Conduct a pilot study of the questionnaire.

3. Data management - some basics

- If security is an issue, avoid USB sticks, local hard drives. Keep data on a network where access is controlled by passwords.
- Store sensitive data in a separate table and destroy this table at your first opportunity.
- Backup and backup your backup.
- Pay someone to do the data entry for you. Put this in your grant budget.
- Develop a data dictionary.

3. Data management - never in a spreadsheet

Spreadsheets, in general, and Microsoft Excel, in particular, are very bad choices for data management.

Clinical data are entered directly onto electronic spreadsheets (MicrosoftÂ® Excel 2011Â®, version 14.4.8, Copyright Â© 1990, Microsoft, Santa Rosa, California, USA), at the time of clinical examination, by a single investigator in each center. Data quality will be validated by checking missing data, out-of-range values and invalid responses.

23/40

3. Data management - never in a spreadsheet

For reasons, see

Jonathan Borwein and David H. Bailey. The Reinhart-Rogoff error - or how not to Excel at economics. The Conversation, April 22, 2013. Available at https://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646.

European Spreadsheet Risks Interest Group. EuSpRIG Horror Stories. Available at http://www.eusprig.org/horror-stories.htm.

Data management, a good example

REDCap (short for REsearch Data Capture) is a relational database available to all UMKC researchers at no charge.

REDCap is a secure, web-based application designed to support data capture for research studies, providing: 1) an intuitive interface for validated data entry; 2) functionality to track and send reminders to increase data completion; 3) audit trails for data manipulation and export procedures; 4) automated export procedures for seamless data downloads to common statistical packages; and 5) procedures for importing data from external sources.

3. Data management - modern guidance

There is a movement, reproducible research, that takes advantage of recently developed tools in software development and applies those tools to data management.

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. Available at https://doi.org/10.1371/journal.pcbi.1005510.

4. Data sharing plan

Funding agencies like to see a formal data sharing plan.

Barriers to a data sharing plan include

- ~~cost~~
- ~~time~~
- ~~lack of technical expertise~~
- confidentiality

4. Data sharing plan

Reasons why you should share your data

- ~~because you have to~~
- increases your citation count
- leads to additional co-authorships
- increases the number of researchers in your area

28/40

4. Data sharing plan

Although I strongly encourage data sharing, I have to admit that it is not without controversy.

Point: Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016). Available at http://www.nature.com/articles/sdata201618.

Counterpoint: Longo DL, Drazen JM. Data Sharing. N Engl J Med 2016; 374:276-277. Available at http://www.nejm.org/doi/full/10.1056/NEJMe1516564.

4. Data sharing plan

For an example of how to write a data sharing plan, see

http://annals.org/aim/fullarticle/2630766

Data sharing systems inlcude

- BMC data notes, http://blogs.biomedcentral.com/bmcblog/2017/09/29/get-credit-for-your-data-bmc-research-notes-launches-data-notes/
- Dryad, http://datadryad.org/
- figshare, https://figshare.com/
- github, https://github.com/

5. Data analysis

Every data analysis is different, but there are certain themes that occur repeatedly. Here are some examples.

5. Data analysis - threshold for statistical significance

Some examples:

We considered a P <0.05 to be statistically significant for all of the analyses.

The significance level was set at .05.

Alpha was set at 0.05.

We calculated the mean difference and its 95% confidence between the periods and used independent t tests to compare both periods.

5. Data analysis - statistical significance

Always specify whether you are using one-sided or two-sided tests prior to data collection.

A two-tailed P value of less than 0.05 is considered to be statistically significant.

If you have multiple outcome measures, specify a hierarchy (e.g., primary outcomes, secondary outcomes) and/or talk about whether you use an adjustment (e.g., Bonferroni).

Correction for multiple comparisons was not applied to the analyses of secondary outcomes.

5. Data analysis - risk adjustment

If you are looking at a risk adjustment, control of confounding variables, etc., specify the variables used in that analysis.

Analyses were also performed with adjustment for prespecified covariates: site, age, sex, Glasgow Coma Scale score, CD4+ cell count, CSF fungal count at baseline, and ART status at baseline.

34/40

5. Data analysis - statistical software

Always cite the statistical software you will be using.

Some examples:

All analyses were carried out using R [30]. … (Bibliography) 30. R Core Team. R: A language and environment for statistical computing. 3.0.2 ed. Vienna, Austria: R Foundation for Statistical Computing; 2013.

We used the software Resampling Procedures 1.3 by D. C. Howell (freeware, www.uvm.edu/~dhowall/statPages/Resampling/ResamplingPackage.zip).

The R free source statistical package version 3.2.2 (The R Project for Statistical Computing, Vienna Austria), and the SPSS 21.0 (IBM SPSS, Chicago, IL, USA) were used in all of the analyses.

5. Data analysis - when to cite references

Commonly used statistical techniques require no citation.

These include:

- t-tests
- chi-square tests
- linear regression
- analysis of variance
- logistic regression

5. Data analysis - when to cite references

Common citations for more complex techniques:

Generalized linear models: McCullagh P, Nelder JA. Generalized Linear Models, Second Edition (1989) Chapman and Hall.

Kappa statistic: Landis JR, Koch GG. â€œThe measurement of observer agreement for categorical data.â€ Biometrics 33, 159â€“174 (1977)

Mediator/moderator variables: Baron RM, Kenny DA. â€œThe moderatorâ€“mediator variable distinction in social psychological-research â€” conceptual, strategic, and statistical considerationsâ€ J. Pers. Soc. Psychol. 51, 1173â€“1182 (1986)

37/40

5. Data analysis - when to cite references

Proportional hazards regression models: Cox, David R (1972). "Regression Models and Life-Tables". Journal of the Royal Statistical Society, Series B. 34 (2): 187â€“220.

Structural Equation Models: Bollen KA, Structural Equations with Latent Variables (1989) Wiley Interscience.

Time series (ARIMA) models: Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time Series Analysis: Forecasting and Control5th Edition (2015) Wiley and Sons.

5. Data analysis - how much detail?

A research grant (or an IRB submission) does not require you to document your data analysis in sufficient detail to allow reproducibility. Just show enough detail to establish that you are not a total noob.

Red flags for reviewers

- Failure to account for hierarchical design
- Choosing a continuous outcome model when your outcome is categorical
- Failure to recognize need for equivalence/non-inferiority testing
- Lack of control for multiple comparisons

In general, the research design is far more important.

Conclusion

A good data analysis plan should be more than "I plan to run a chi-squared test on the data." Other important factors are:

- sample size justification
- survey or questionnaire development
- data management
- data sharing

Copy of these slides at https://github.com/pmean/write-data-analysis-plan

40/40