Working with a Statistician

Topic: Key considerations when working with a statistician at various points of your study are discussed. Data formatting for analysis is also described.

Summary: A statistician can provide immense value to your research project. To maximize the value they bring, be sure to involve them in your project from the very beginning and to give them all the information they need.

At What Point Should You Contact a Statistician?

- **Ideally, a statistician should be involved from the study's inception.** Contact a statistician before you collect data, before you recruit participants, and before you submit a grant to fund the study. The moment you decide to start a study, you should think about who will provide the statistical expertise, should you need it.
- Benefits of including a statistician in the planning of a study:
  - Formulation of research questions in a clear and testable way
  - Verify that the planned procedures and sample sizes are adequate to address the research question
  - Guide selection and eventual implementation of methods for data analysis
  - Provide advice earned through experience about difficulties you may encounter

Meeting with a Statistician

- The statistician's expertise is not in medicine so you should expect they have little subject knowledge about your research. They may be familiar with certain variables, instruments, or acronyms, **but be sure to discuss what the statistician does and does not know about your research topic and clinical area of expertise.**
- You may need to explain clinical terms to the statistician in laymen's terms. Avoid jargon unless it is essential to the understanding of your project.
- Consider the following when meeting with a statistician:
  - What information about the phenomena being studied does the statistician need to know to help?
  - What are the assumptions I am making? Do they make scientific/statistical sense?
  - Is everyone who may need to work with the statistician at this meeting? If you have research associates or trainees, would it be beneficial for them to hear what the statistician has to say at this meeting?
- Statisticians may be working with several collaborators/clients at a given time. Remember to **provide the statistician with as much relevant information prior to meeting.** This will give the statistician time to think and research appropriate statistical issues. This is best done through providing a protocol for the statistician to review.

Providing a Protocol for Your Statistician

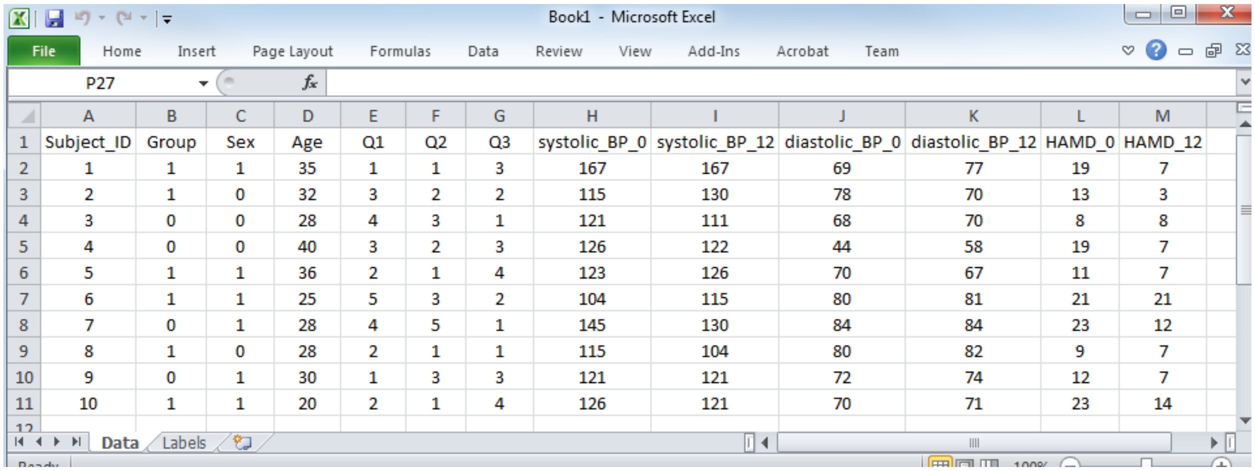
- **You must write a study protocol which the statistician can review.**
- Your protocol is a description of what you want to examine and what you hope to achieve.
- A study protocol should have sufficient information for the statistician to become familiar with your project. In particular, consider answering/including some of the following:
    - What is the motivation behind the study? Why are we researching this phenomenon? Why is it important?
    - What is the clinically important difference you hope to detect?
    - A clear identification of the exposure or intervention being studied. What is the intervention designed to do? Why might we give it to some patients and not others?
    - What is the question you are trying to answer? Is the intervention *better* than the standard of care? Is the intervention *no worse* than standard of care? Answers to these questions have statistical implications.
    - Who will be included in the study? Who will be excluded from the study? Why?
    - What are some relevant sources for the statistician to reference? Other papers on previous studies are an excellent resource for statisticians as they will allow the statistician to anticipate any troubles or biases that may burden the study.
- If you have an idea of what statistical procedure you would like to apply, consider answering the following questions in the protocol:
    - What covariates you are interested in recording and controlling for? Why do you think these covariates are important to control for? How will these be recorded? For example, if you are collecting age, will it be raw age or a bucketed age category?
    - Where will the data come from? Are you collecting it yourself, or will it come from an existing database?
    - Do you anticipate any of the data to be missing?
- Remember, **avoid excessive jargon in the protocol.**
- The protocol should include the following sections:
    - Background/Introduction/Previous Work Done
    - Proposed Research Question
    - Questions to Answer in the Study
    - Methods & Materials (if relevant at this point)
    - Resources and References
    -

Making the Most of Your Time with a Statistician

- Start your project with early involvement by a statistician.
- Be aware that statisticians have other collaborators/clients, and some have their own research, so reach out for statistical expertise time early on.
- **It is the role of the Research Assistant or Research Coordinator to clean the data.** Ensure your data is in a format that you and your statistician have agreed upon. If you have not met with a statistician prior to starting data collection, see the data format section below. If the data is in an unanalyzable format, the statistician may return the data asking it to be properly formatted before they conduct analysis.
- Do not assume the statistician will support your entire project. If your statistician is acting in a consulting capacity in the planning stage of a project, do not assume they will do the analysis. More services may require more funds as they will take up more of the statistician's time. Have an explicit conversation surrounding what you need from your statistician and the timeline.

Data Format (or, How to Keep Your Statistician Happy)

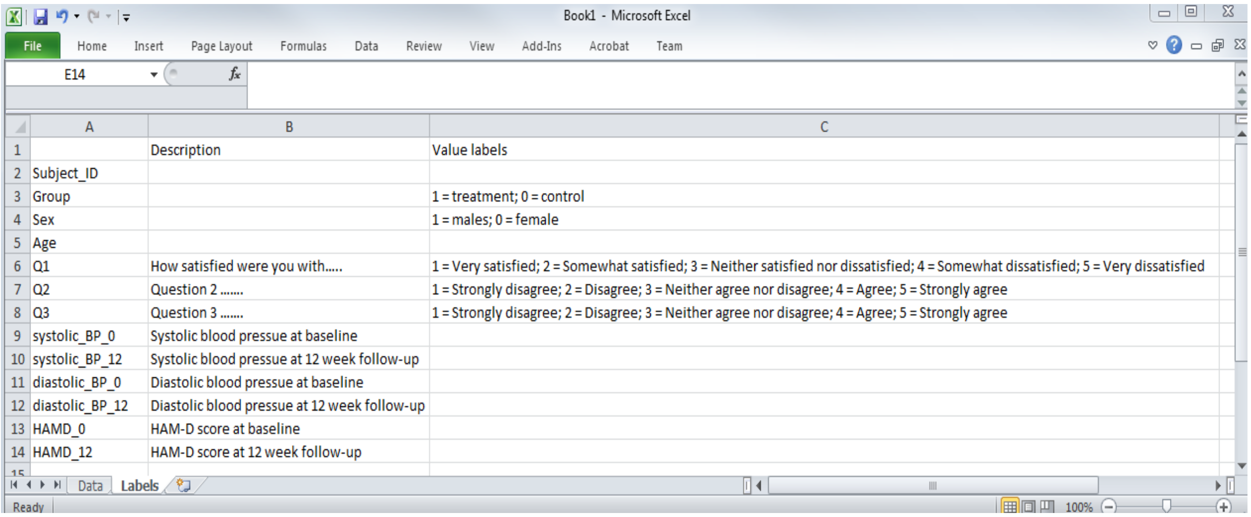
- Nothing makes a statistician more frustrated than receiving data in a format which is not conducive to expedient and reproducible analysis.
- The standard format for data is to be in *tidy format* (Hadley Wickham, 2014). Under this format:
  - Each variable forms a column
  - Each observation forms a row
  - Each type of observational unit forms a table.
- Shown below is an excellent example of "tidy data". Notice that each subject's data (or equivalently, each observation) is in a single row. Each column is a single variable.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Subject_ID	Group	Sex	Age	Q1	Q2	Q3	systolic_BP_0	systolic_BP_12	diastolic_BP_0	diastolic_BP_12	HAMD_0	HAMD_12
2	1	1	1	35	1	1	3	167	167	69	77	19	7
3	2	1	0	32	3	2	2	115	130	78	70	13	3
4	3	0	0	28	4	3	1	121	111	68	70	8	8
5	4	0	0	40	3	2	3	126	122	44	58	19	7
6	5	1	1	36	2	1	4	123	126	70	67	11	7
7	6	1	1	25	5	3	2	104	115	80	81	21	21
8	7	0	1	28	4	5	1	145	130	84	84	23	12
9	8	1	0	28	2	1	1	115	104	80	82	9	7
10	9	0	1	30	1	3	3	121	121	72	74	12	7
11	10	1	1	20	2	1	4	126	121	70	71	23	14
12													

- Note also that the data have no superfluous formatting. There are no bold titles, no borders, no colors, no shapes, and no plots. This is for ease of loading into statistical software. Formatting is for human readability, **when handing data to a statistician, remove all formatting. Emulate the above example as closely as possible.**

- Note also that the column headers are short, informative, and no not contain any special characters (e.g. !&^%$,<>?;”).
    - Good column name examples: “subject_id”, “subject”, “sex”, “is_male”, “systolic_bp_mg”.
    - Bad column names include: “Subject Id #”, “Sex (M = 1, F = 0)”, “Blood Pressure (mmHg)”. Note that these all have spaces and/or special characters.
- Sometimes, column names cannot explain the variable completely. **You should always include a separate document/spreadsheet outlining the interpretation of the variables** in greater detail. This is called a **data dictionary**. Shown below is such a document:



	A	B	C
1		Description	Value labels
2	Subject_ID		
3	Group		1 = treatment; 0 = control
4	Sex		1 = males; 0 = female
5	Age		
6	Q1	How satisfied were you with.....	1 = Very satisfied; 2 = Somewhat satisfied; 3 = Neither satisfied nor dissatisfied; 4 = Somewhat dissatisfied; 5 = Very dissatisfied
7	Q2	Question 2	1 = Strongly disagree; 2 = Disagree; 3 = Neither agree nor disagree; 4 = Agree; 5 = Strongly agree
8	Q3	Question 3	1 = Strongly disagree; 2 = Disagree; 3 = Neither agree nor disagree; 4 = Agree; 5 = Strongly agree
9	systolic_BP_0	Systolic blood pressue at baseline	
10	systolic_BP_12	Systolic blood pressue at 12 week follow-up	
11	diastolic_BP_0	Diastolic blood pressue at baseline	
12	diastolic_BP_12	Diastolic blood pressue at 12 week follow-up	
13	HAMD_0	HAM-D score at baseline	
14	HAMD_12	HAM-D score at 12 week follow-up	

- Each column in the data set is listed and given a description. Value labels are included for interpretation purposes. If you code a category with a number (e.g. males = 1, females = 0), here is where you would elucidate your coding choices, as opposed to putting them in the columns). This is an ideal place to indicate units in which the measurements are made.
- If you transform the data in anyway (e.g. bucket a continuous measure, like age, into discrete categories) keep the original variable in an adjacent column for reference.
- If you have dates included in your data (e.g. date of enrollment, date of birth), please keep them in a consistent format. The preferred format is YYYY/MM/DD.

- Ensure that the outcome is easy for the statistician to find in the dataset. A good rule of thumb is to have the outcome of the study in the far most right column.

- Missing data is a fact of life. If data is missing, fill the cell with NA (for not available). Be sure to include somewhere in your data dictionary that you have used NA for missing data.
- There is a difference between data missing because it can't be collected for some reason (e.g. malfunction of a piece of equipment) and missing data because of non-response (e.g. patient refuses to answer a question). If a patient refuses to answer a question, code that explicitly as an option (e.g. Refused to answer).
- For more on the tidy data format, including examples of tidy and non-tidy data, see (Hadley Wickham, 2014).

Tidy Data Checklist

Use the checklist below to ensure that your data is in an appropriate format for your statistician:

- ✓ All formatting (e.g. bold, colors, plots) are removed
- ✓ Variables for columns
- ✓ Rows form observations
- ✓ Patient Identification numbers are encrypted
- ✓ Column names are short and informative
- ✓ There are no special characters in my column names
- ✓ Variables are described in an attached "Labels" document explaining each variable and the units it is measured in
- ✓ Categorical information (such as sex, ethnicity, etc) are numerically coded
- ✓ Any dates are in YYYY-MM-DD format, or at the very least in a consistent format
- ✓ Bucketed variables have their original variables in an adjacent column
- ✓ The outcome for the study is at the far most right column

Consultant Statistician vs. Team Statistician

- A **team statistician will be a part of your research team from the start of the project**. They can support study design, data collection, data analysis, and writing manuscripts for publication. A team statistician must be budgeted for in the initial funding application (e.g. stats trainee stipend on a CIHR grant).
- A **consulting statistician operates under a pay for service model**. They usually enter a project to advise on a particular aspect(s) of the project. Investigators should have explicit conversations about what is needed from the consulting statistician, and what the services will cost per hour. Consulting statisticians have no obligations to assist with writing manuscripts for publication unless the service is requested and paid for by the client.
- If you opt for a consulting statistician, ensure you have all the resources necessary to reproduce and recreate any analysis the consulting statistician may provide for you. This includes scripts, data, or any other resources the consulting statistician used.

- A team statistician comes at a higher financial cost, but will be able to support the project to completion. A consulting statistician may be a better option if investigators need help with a small portion of the project, but costs can add up quickly.

Summary

- It is always advisable to include a statistician in your projects. If you decide to include a statistician, the earlier you include them, the better.
- Remember that statisticians may need some background on the area of research. Their expertise is statistics, not your clinical area.
- Avoid pitfalls like asking for work to be done at the last minute. More time to think will lead to better and more robust results.
- Give your statistician as much information as you can. A protocol is the most appropriate way to do this.
- If your statistician will be doing your data analysis, be sure to have a conversation about the data and how it will be formatted. It is not the statistician's job to prepare your data.

References

Hadley Wickham. (2014). Tidy Data. *The Journal of Statistical Software*.