Biostatistics Contact Form

Biostatistician: StellaMay Gwini
StellaMay.Gwini@barwonhealth.org.au
Tel: 03 4215 9624

Documents included:

1. Contact form
2. Guideline for preparing a data for statistical analysis
3. Guidelines for Determining Co-Authorship for Biostatisticians

Date of request:	
Name:	
Department:	
Position:	
Telephone:	
Email:	
Brief project description (e.g. background, objectives, research question)**:** *Also attach study protocol, research proposal, ethics application, questionnaires or any other documentation that may be useful in describing your project.*	
Description of support requested:	
Stage of your project (Check all that apply)	☐ Defining aims/planning/designing (data not yet collected) ☐ Preparing an HDR proposal ☐ Grant preparation ☐ Grant review ☐ Collecting data ☐ Analysing data/submitting (paper or dissertation) ☐ Other (please specify below) ……
Expected outcome of work (Check all that apply)**:**	☐ Conference presentation ☐ Peer-reviewed manuscript ☐ Grant application ☐ Exploratory analysis ☐ Other (specify) ……
Specific Statistical support requested (Check all that apply)**:**	☐ General Statistical advise *(will run own analysis)* ☐ Preparation of Statistical Analysis Plan ☐ Detailed statistical analysis *(Biostatistician to run analysis)* ☐ Sample size and power calculations

	☐ Statistician's check of statistical analysis in a manuscript ☐ Partnership/Collaboration in grant application ☐ Other (specify) ……
Comments	

Further information

In addition to providing the documents requested above, please review the two guidelines provided below. The first one provides guidelines to preparing data for statistical analyses while the second provides guidance for establishing level of co-authorship for the Biostatistician. For further information or clarification, you can contact the Biostatistician.

Guideline for preparing a data for statistical analysis

(Adapted and extended from an original document by A/Professor Michael Bailey, PhD, Epidemiology & Preventive Medicine, Monash University & A/Professor Dean Mckenzie, Epworth Healthcare)

Data available for statistical analyses is often collected using many different forms, including excel, paper questionnaires or software such as REDCap. However, regardless of its source, it is essential that data integrity is maintained and good data management strategies are in place before a study commences.

Here at Barwon Health we encourage all data to be collected or accumulated using the online tool REDCap. This software allows one to create questionnaires, administer them electronically and provide data format outputs that can easily be used with statistical software such as R, SAS, IBM SPSS and Stata. Importantly, data can be imported with all labels, therefore saving time for the data analyst. In addition, data can be transferred from paper questionnaires to REDCap allowing for better data management and storage. We strongly discourage the use of Excel spreadsheets for data collection as data can easily be entered incorrectly and/or cells can be accidently changed without a record of the changes.

Data collected using Microsoft Excel

However, if data are collected in Excel, the data needs to be formatted in such a way that it can be read by statistical packages such as IBM SPSS or Stata. Often when researchers use Excel spreadsheets, a lot of other data/information is included such as graphs, comments, different coloured text, formulae e.t.c. which cannot be read by the analytic software. Some of these can become problematic when the data needs to be analysed. Therefore, before sharing data spreadsheets with the Biostatistician, the spreadsheet should be

- Plain with column titles
- no graphs
- no summary statistics
- no notes

- no different colours to represent subgroups. If subgroups are required for the study, create a new column that indicates the subgroups.

Cleaning up spreadsheets and getting data into a usable shape can seriously delay the statistical analysis. If not sure on how to prepare your data collection tool, please contact the Biostatistician. Assuming that you would have used Excel to gather data, below is a summary on how to prepare your data for analyses.

1. The easiest way to set out data is to have one person per *row* of the data file or spreadsheet and one variable per *column*, with **no blank cells, blank rows or blank columns**. Otherwise, if there's a blank, it's not clear whether it is meant to represent 'missing' or whether something was meant to be entered but wasn't.

2. Any missing data should **be entered as a unique missing code** (e.g. -9 or -99) that cannot otherwise occur in the data or as a dot. If a difference between two scores is entered, for example, then negative scores may be possible and so -9 isn't a unique missing value code. Blanks should be used only if there's really no easier way.

3. **Do include a column for each possible response**, e.g. if it is possible to answer yes to two types of tablets taken, then include separate column for each and enter 1 if taken, 0 if not.

4. **Avoid long column names** (i.e. keep to around 12 characters or less), or names including embedded spaces (use underscores instead) or include characters other than letters or numbers. e.g. please use "age_at_onset" with underscores, rather than "age at onset" with blanks.

5. Try to **stick to numbers** to represent values if possible, e.g. enter 1 for Yes and 0 for No, enter 1 for Female and 0 for Male, rather than text.

6. Try to **avoid categorising data at entry**. Please enter actual values (e.g. actual age) if possible, as it can always be categorised or split into ranges) within the stats package, later on.

7. **Avoid colours, italics, symbols, graphs, formulae, notes, comments** etc in the Excel data spreadsheet *provided to the Statistician*. Please avoid use of spaces, brackets, slashes etc

(underscores and numbers ok, as long as starts with a letter) in column names, and column names should be on **one row** only.

8. **Avoid identifiable information** (names) in the file provided, but please keep and maintain a locked central list of names and other identifiable patient information. Try and avoid using URs but assign study numbers to each UR. If patients have multiple admissions, these can be distinguished by adding a column 'admission number'.

9. **Use standard date** e.g. DD-MM-YYY and time. Avoid mixing formats in one column e.g. using both the 12hr (11:30pm) & 24hr (23:30) notations.

10. **Include a Group column**. If there are separate groups, e.g. control and treatment, indicate group membership in a group column, e.g. 1 if Treatment, 0 if Control, **in the same spreadsheet** (i.e. separate tabs or files for each group aren't required).

11. **If it is possible to have two or more responses to a question (e.g. type of medications taken) then there should be separate columns for each response**, each of which can be coded absent (0) or present (1) as shown below. This actually makes it much easier to analyse than trying to come up with combinations such as AsproBex and AsproBexPanadol etc but instead present it as shown below.

aspro	*vincents*	*bex*	*panadol*	*other*
1	*0*	*0*	*0*	*1*

12. Variable headings or column names should be in the first row of the database/spreadsheet, begin with a letter a-z, with each name being around **10** to **12** characters or less in length **Please don't use spaces, brackets, slashes** in the column names/ headings or any of the following characters - !@#$%^&*(){}[]=-\;<>< >?. There should only be **one** row of column names, in row one, with data starting in the **second** row. Each column name must be unique, i.e. should only occur once in a database / spreadsheet. *(Most statistical packages will not allow names beginning with a number or containing non-standard characters).*

13. All columns in the spreadsheet or data file should be either **all numbers or all characters/strings**.. If <1, >1 and >5 are valid responses on your questionnaire you'll need to use specific codes, e.g. 1 for <1 and 2 for >1 and 3 for >5 and so on, keeping a key or data

dictionary in a separate spreadsheet or Word document to be supplied to the Biostatistician together with the data.

If data are already in text format, please make sure that **case and responses are consistent**, e.g. M or F, rather than M,m,male,Female,f.

14. **Repeated Measures / longitudinal / measurements across time**: In some types of designs, the same patients may be followed up over time (e.g. blood pressure measured across six weeks). When only 2 time points, it is easiest to structure repeats across the page or 'wide format' (e.g. BPpre & BPpost for blood pressure at pre-test and post-test).

    When more than 2 time points, or people can have a different number of entries (e.g. hospital admissions), it's easier to set out the data **down** the page ('long format') in **rows**. Include a column representing occasion/measure, e.g. 1,2,3,4 with each *row* representing a different occasion, and include date, and time of day if relevant. Patient 101 below has had her systolic blood pressure tested on three occasions, Patient 102 has had his blood pressure measured only twice.

Example of a spreadsheet with patients measured on multiple occasions ("long" format)

Id	Time point	Date	Systolic_Blood_Pressure
101	1	07-Jan-2016	122
101	2	09-Jan-2016	131
101	3	14-Jan-2016	126
102	1	08-Jan-2016	130
102	2	12-Jan-2016	128

15. Similarly, show when same patient has **multiple admissions / procedures / operations**. If a particular patient appears more than once in a data file (e.g. might have had several admissions for same or different illness, may have had both hips replaced, etc), such patients should have the same id (e.g. look up each UR number in the separate spreadsheet, **and assign the same study id to that patient each time he or she appears**. If this is not done, the variance estimates and hence p values and confidence intervals would be biased, as any possible dependence between measures needs to be taken into account.

The final dataset for analysis should ideally all be in th**e same, single Excel worksheet** rather than using different worksheets within files*. multiple SPSS files are ok, but should all have common ID.*

Example of a spreadsheet that is <u>not correctly prepared</u> for statistical analysis

URN	Date Of Birth	Patient Age	Gender	start date	Current Smoking /status	NYHA	blood Pressure Pre	blood Pressure Post		Marital status	ppainmed
9722-1	12/05/63	41YRS	Male	19/07/04	No☺	I	115	75		Married	Asp/panadol
0651312	14/09/26	78	F	26/01/04	No	n/r	=90	50		Singles	aspro
0011111	5/02/44	60 +3months	F	29/07/04	N	II	140	80		single	Aspirin/Panadol

Example of a spreadsheet that <u>is</u> correctly prepared for analysis

Ur	dob	age	female	startdate	cursmoke	NYHA	BPpre	BPpost	marital	aspirin	panadol
0651312	14-Sep-26	78	1	26-Jan-2014	0	-9	90	50	1	1	0
0454545	07-Dec-33	70	0	-9	0	2	140	70	3	-9	-9
0011111	05-Feb-44	60	1	29-Jul-2014	0	2	140	80	1	1	1
0106574	10-Nov-36	67	1	2-Jan-2014	0	2	120	70	3	-9	-9

Guidelines for Determining Co-Authorship for Biostatisticians

Biostatisticians are available to support staff undertaking research at Barwon Health. However, in many cases the contribution made by a biostatistician may warrant co-authorship. All researchers are required to consider established authorship guidelines in determining authorship, including the Barwon Health Guidelines for Collaborative Research and Authorship, and it is recommended that the role of the biostatistician is discussed and clarified at the beginning of a collaboration.

The role of the biostatistician and his/her contribution varies across different projects. Determining whether a statistical consultant should/could be a co-author should be negotiated on a case-by-case basis. Current internationally recognised guidelines advise that authorship should be based on the level of participation and should not be influenced by whether or not the consultant was paid. The basis for financial support and authorship are different as the former considers the time and effort put into the project whilst the later relies on intellectual input. The following internationally recognised guidelines should be seen as recommendations that may be helpful in determining co-authorship.

According to the International Committee of Medical Journal Editors (ICMJE Updated Recommendations, December 2018), all authors should satisfy each of the following four criterion:

1. a) substantial contributions to conception and design ***or***
   b) the acquisition, analysis or interpretation of data from the work
2. a) drafting the work ***or***
   b) revising it critically for important intellectual content
3. a) final approval of the version to be published
4. a) agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conditions 1, 2, 3 and 4 must all be met.

In addition to accountability for the parts of the work undertaken, an author should be able to identify which co-authors are responsible for specific other parts of the work. In addition, authors

should have confidence in the integrity of the contributions of their co-authors. All those designated as authors should meet all four criteria for authorship, and all who meet the four criteria should be identified as authors.

Considering a Biostatistician as a co-author: In keeping with the above guidelines, if the statistician participating in the study performs some activity in each of categories 1 and 2 above and also satisfies 3. and 4. then his or her contribution should be acknowledged as co-authorship. The most common contributions made by biostatisticians to projects, reports or manuscripts include:

- advising and collaborating on the design of the study (including determining the appropriate sample size),
- analysing the data collected, and interpreting the results for and with the principal investigator(s).
- writing the statistical methods section of the manuscript (or editing versions drafted by other investigators).
- writing and/or revising drafts of some or all of the results section.
- drafting or editing tables and figures in the manuscript.
- developing new statistical methods to meet the project's needs, and/or combining existing techniques in a novel manner.
- reviewing/providing critical analysis of drafts

Furthermore, some journals require that at least one author be responsible for each section of the manuscript that is essential to the conclusions made about the research. Therefore, if the statistician is not listed as a co-author, then one or more of the authors must take responsibility for the statistical analysis.