Advice for preparing a simple database for statistical analysis using Excel

1. Excel is probably the easiest and most user friendly package to use when preparing ***simple*** databases for analysis. Unfortunately, as data points are easily changed, Excel is also the most easily corruptible data entry package. To reduce mistakes due to improper sorting and accidental over-writing, consistently save backup copies of the data.

   The rule of thumb for choosing Excel rather than a database is, "Are you prepared to re-enter, from primary sources, all the accumulated data if you accidentally corrupt it?" If you're not, seek assistance with the design of a database.

2. Any dataset provided must have an accompanying document describing the columns in the worksheet and the coding of each field. This is often called a 'data dictionary'.

3. The most common way to structure data is to have one person (subject or patient) per row and one variable per column, with no blank rows or columns anywhere in the database.

4. Variable headings should be in a single row, the top row of the sheet, with each heading preferably limited to 8 characters or less. Avoid using blank spaces in the headings (an underscore can be used instead of a space), and avoid the following characters…
   i. ! @ # $ % ^ & * ( ) { } [ ] = - \ ; < > ?
   ii. … either in the headings or in other fields within the database.

5. In columns of numbers, all characters such as (>, \ / .?) must be removed. As a guide, Excel will right justify numbers and left justify characters. As > and < signs are not recognised by statistics packages, please think carefully about what single numerical value should be used to replace these values, and preferably supply two columns – one containing the original, mixed alpha-numeric data and one containing the numeric data.

6. Fields that contain missing data should be left completely blank (rather than NA, missing or ? or * or 999). Most statistical packages can cope with importing a blank field (and will replace it with the missing value code unique to the package).

7. Treatment groups and other groups (eg. Gender, Ethnicity, Marital Status) to which subjects belong should be indexed by categorical factors (each factor has its own column in the spreadsheet). Avoid putting more than one factor in a column (it complicates the importation of the data into a statistical package) and note that while colour coding can be helpful when visually checking the data, it cannot, unfortunately, be interpreted by most statistical packages – hence the need to include factors to index the groups of interest.

   Also avoid labelling the categories with textual codes that can be entered in multiple ways (e.g. Female, female, FEMALE) because Excel will read these different versions as distinct categories and they will be imported into statistical packages as such. More preferable are either numeric codes or single letter text codes, entered consistently

8. For ease of importing the data into a statistical analysis package, remove any superfluous data, summary tables and charts from the sheet that contains the "analysis data set".

9. Excel is quite finicky about date variables so it's important to use a consistent date format (and save the sheet consistently in a particular file-type). Cutting and pasting date fields from old to new versions of Excel (and vice versa) can create problems when importing the sheet into another package.

10. Repeated Measures: When only 2 time points exist, it is sometimes advantageous to structure repeats across the page (eg BPpre & BPpost). When there are more than 2 time points it is generally better to structure the data down the page, with an additional column indicating time (i.e. a "stacked" data sheet) and repeating ID codes for the subjects and their treatment groups. Statistical packages can easily switch the data structure from "wide" to "stacked".

11. For studies (e.g. clinical trials) with a complex visit structure we encourage you to seek the advice of professional data managers and database designers as well as a biostatistician. Generally these studies will have three styles of linked “sheets” or datasets. A demographic dataset with one row or record per subject, a visit dataset with multiple records per subject (one record for each visit, indexed by a visit or sample date), and running logs with multiple records per subject (each record corresponding to a new incident and each record with a start date and either a stop, or censor, date). The datasets are linked by subject ID and records are uniquely identified by dates within ID’s. Statistical packages can easily merge and link data from different sources but this facility relies on the sources being set up appropriately with variable names chosen carefully and linking variables defined clearly.

12. For ease of importation into statistical packages, where possible (e.g. with small studies) the final dataset for analysis should be contained in a single worksheet rather than spread over multiple short worksheets.

13. Although ideally researchers should provide a single, finalised dataset for analysis, in reality this may not occur because omissions, errors, corrections or new data may arise. In such circumstances it is imperative that updated versions of the dataset should be identified in the name of the file, the variables names should not change between versions, and the general layout of the dataset should not be altered.

Example of Excel data that is *unsuitable* for analysis

URN	Date Of Birth	Patient Age	Gender	start date	Current Smokers	NYHA	Systolic	Blood Pressure Pre	Blood Pressure Post		Marital status
9722-1	12/05/63	41YRS	Male	19/07/04	No	I	120	115	75		1
0651312	14/09/26	78	F	26/01/04	No	n/r	n/a	=90	50		2
0454545	7/12/33	70	M	n/a	N	II	.	140	70		3
0001111	21/05/35	69	m	n/a	N	III	?	130	80		2
0011111	5/02/44	60 +3months	F	29/07/04	N	II	<90	140	80		3
0106574	10/11/36	67	F	2/01/04	N	II	70 (under	120	70		1
1066329	19/09/46	58	f	n/a	N	III	>170	170	100		3
0537720	1/09/51	53	F	n/a	Y	II	115	120	80		2

Example of Excel data that is *correctly prepared* for analysis

Diabetic	UR	DOB	Age	Male	StartDat	CurSmoke	NYHA	Syst	BPpre	BPpost	Bpdiff	Marital	Married
1	0009722	12-May-63	41	1	19-Jul-04	0	1	120	115	75	40	married	1
0	0651312	14-Sep-26	78	0	26-Jan-04	0			90	50	40	single	0
1	0454545	07-Dec-33	70	1		0	2		140	70	70	divorced	0
1	0001111	21-May-35	69	1		0	3		130	80	50	married	1
1	0011111	05-Feb-44	60	0	29-Jul-04	0	2	89	140	80	60	single	0
0	0106574	10-Nov-36	67	0	2-Jan-04	0	2	70	120	70	50	divorced	0
1	1066329	19-Sep-46	58	0		0	3	171	170	100	70	married	1
1	0537720	01-Sep-51	53	0		1	2	115	120	80	40	single	0

The way to structure repeated data down the page

Diabetic	UR	DOB	Age	Male	Time	Bp	Sodium	Troponin
1	0009722	12-May-63	44	1	1	75	40	2.3
1	0009722	12-May-63	44	1	2	50	40	1.9
1	0009722	12-May-63	44	1	3	70	70	12.8
1	0009722	12-May-63	44	1	4	80	50	3.4
1	0009722	12-May-63	44	1	5	80	60	0
0	4444444	1-Jan-02	5	0	1	67	23	0.87
0	4444444	1-Jan-02	5	0	2	68	25	0.32