



**QIMR Berghofer**  
Medical Research Institute

# **Collection and management of research data in Excel**

# Overview

- Where to store data
- Database considerations
- What is data cleaning and why do we do it?
- Practical Microsoft Excel tips for:
  - Data entry
    - Using data validation and conditional formatting
  - Data cleaning
    - Using formulae, figures and pivot tables

# Where to store data?

A spreadsheet of any kind, i.e. Excel, Access, what you feel comfortable managing.

- Each person is a row
- All measurements are contained in that row
- **Each person** has a **unique ID**
- **De-identified** i.e. no patient names

Consider a database like Access if your study:

- Is expected to expand in scope
- Has a multi-year time horizon
- Is not possible to validate line by line easily

# Database considerations

- **Who will be able to make changes?**
  - Create a master file that only one person can change
  - Keep a record of changes by creating a copy of the Excel sheet and highlighting any problems
- **Be careful manipulating the data**
  - Take care when sorting – all the data for one person should be in the same row
  - Take care when cutting and pasting
  - Take care when merging data sets

# What does a dataset look like in Excel?

Patient	Procedure	Survival_time	age	gender	no_of_prior_heart_conds	kidney_disease
1	1	4	999	f	0	no
2	2	1	86	m	3	yes
3	1	5	62	f	1	no
4	1	4	59	f	3	No
5	2	1	61	f	2	yes
6	2	5	86	m	1	yes
7	1	5	60	F	3	Yes
8	1	3	57	m	1	no
9	1	1	56	m	1	no
10	1	5	56	m	1	yes
11	2	20	90	m	1	no
12	2	1	89	f	3	no
13	1	4	62	f	1	no
14	1	3	75	m	0	yes
15	2	5	76	M	2	no
16	2	4	81	f	0	no
17	2	5	71	fm	3	no
18	1	4	66	m	0	no
19	2	3	57	m	3	no

# Data management: codebooks

- What is a Codebook?

It is a separate sheet in your Excel file that explains the data. It describes what each variable is, what units have been measured, valid range of values and how missing values are coded

- Ensures consistency during data entry and is required for data cleaning

# Codebook

Variable No.	Name	Label	Values
1	PatientID	Unique ID	30 patients in total
2	Procedure	Procedure type	1=angioplasty 2=bypass
3	Survival_time	length of survival since surgery	years, max no. of years is 5 years
4	age	Age at surgery	years, exclusion criteria: exclude those > 90
5	gender	Gender	m=male f=female
6	no_prior_heart_conds	number of prior heart conditions	numbers 0,1,2,... 999 means unknown
7	kidney_disease	Whether patient had kidney disease	yes no 999 means unknown

# Dataset: data entry

Patient	Procedure	Survival_time	age	gender	no_of_prior_heart_conds	kidney_disease
1	1	4	999	f	0	no
2	2	1	86	m	3	yes
3	1	5	62	f	1	no
4	1	4	59	f	3	No
5	2	1	61	f	2	yes
6	2	5	86	m	1	yes
7	1	5	60	F	3	Yes
8	1	3	57	m	1	no
9	1	1	56	m	1	no
10	1	5	56	m	1	yes
11	2	20	90	m	1	no
12	2	1	89	f	3	no
13	1	4	62	f	1	no
14	1	3	75	m	0	yes
15	2	5	76	M	2	no
16	2	4	81	f	0	no
17	2	5	71	fm	3	no
18	1	4	66	m	0	no
19	2	3	57	m	3	no



# Data entry: useful Excel tools

- Each person is a row and all measurements are contained in that row
- Ensure variable names do not have spaces
- Most statistical packages are case sensitive
- Missing data is important
- Tools:
  - Data validation (as you enter the data)
    - Highlight data cells to have rules applied to
    - Data tab -> Data tools -> Data validation
    - Enter validation criteria (rules for values that data can take)
  - Conditional formatting (for data which has already been entered)
    - Highlight data cells to have rules applied to
    - Home tab -> Styles -> Conditional formatting -> Highlight cell rules

# Data cleaning: what is it?

- Process of finding, correcting or removing data from a database that is incomplete, incorrect, formatted wrongly, duplicated and/or imprecise
- Data cleaning aims to have a database that is accurate, correct and complete

# Data cleaning: why do it?

- Our aim to conduct high-quality research
- Research integrity
  - Data needs to be valid and reliable
  - Research needs to be reproducible
- Poor quality data may lead to misleading results and incorrect conclusions
- Junk in = junk out

**Why spend all that time collecting your data to waste it on erroneous results?**

# Data cleaning: Excel tips

Knowing your data is important

- Each person should have a unique ID, eg URN
  - Check for duplicates
    - Highlight column -> Home tab -> Styles -> Conditional formatting -> Highlight cell rules -> Duplicate values -> OK
- Use the filter and refer back to the codebook to identify values that are not described
  - Turn on filter
    - Data tab -> Data tools -> Data validation -> Select Filter

# Data cleaning: Excel tips

- Most statistical packages are case sensitive
  - Use `=upper()` or `=lower()` to change case
- Check values against codebook. Are they compatible with clinical expectations?
  - Use formula
    - `=countif()` for categorical variables
    - `=min()`, `=max()`, `=average()` and `=stdev()` for quantitative variables
  - Create figures, eg scatter plots
    - Insert tab -> Choose chart type
  - Create pivot tables
    - Obtain frequencies, contingency tables, summary statistics
      - Highlight all of the data -> Insert tab -> PivotTable -> OK -> move variables into rows, columns and values and select value type

# Data cleaning: Excel tips

- Dates are finicky!
  - Date values should have a / not a . (e.g. dd/mm/yyyy)
  - Use find and replace to change all with . to /
  - All dates need to be in the same format
    - Examples of different formats:
      - 14/3/2001
      - 3/14/2001
      - 14/3
      - 14/3/01
      - 14/03/2001
      - 14-Mar-01
  - You can calculate the difference between dates and use conditional rules like `=if()`

# Dataset: Problems identified after data cleaning

Patient	Procedure	Survival_time	age	gender	no_of_prior_heart_conds	kidney_disease
1	1	4	999	f	0	no
2	2	1	86	m	3	yes
3	1	5	62	f	1	no
4	1	4	59	f	3	No
5	2	1	61	f	2	yes
6	2	5	86	m	1	yes
7	1	5	60	F	3	Yes
8	1	3	57	m	1	no
9	1	1	56	m	1	no
10	1	5	56	m	1	yes
11	2	20	91	m	1	no
12	2	1	89	f	3	no
13	1	4	62	f	1	no
14	1	3	75	m	0	yes
15	2	5	76	M	2	no
16	2	4	81	f	0	no
17	2	5	71	fm	3	no
18	1	4	66	m	0	no
19	2	3	57	m	3	no
20	2	3	58	f	10	yes
21	2	3	78	m	1	no
21	1	5	60	f	0	no
23	2	5	84	f	2	no
24	0	1	84	m	0	no
25	1	3	61	m	2	yes
26	2	2	77	m	2	no
27	1	5	84	f	0	No
28	2	5	62	m	2	yes
29	1	1	73	f	1	no
30	2	3	85	m	1	Yes

# Bad data practices

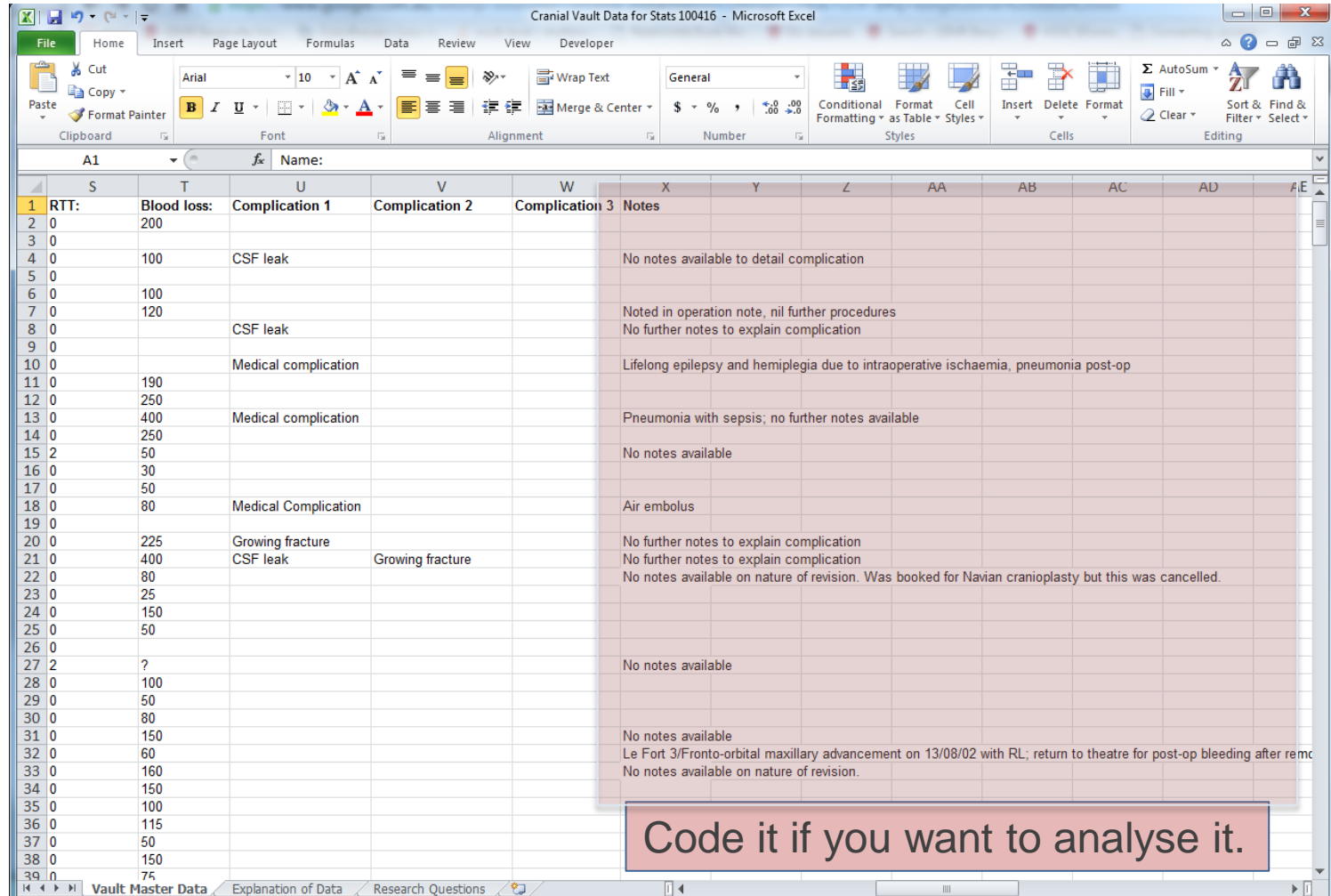
- More than one line for each subject

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I
1	<b>Patient #1</b>								
2									
3	<b>Hand Score:</b>								
4	<b>Left</b>				<b>Right</b>				
5									
6	Thumb	5			Thumb	4			
7	Fingers	5			Fingers	5			
8									
9	<b>Pinch Gauge:</b>						<b>units = kg</b>		
10	<b>Left</b>				<b>Right</b>				
11									
12		<b>1</b>	<b>2</b>	<b>3</b>		<b>1</b>	<b>2</b>	<b>3</b>	
13	Thumb	6	6	6	Thumb	6	5.5	5	
14	Index	4.5	4.5	4.5	Index	6	5.5	6.5	
15	Little	2.5	2.5	3	Little	2	3	3	
16									
17	<b>Split Hand Index:</b>								
18	<b>Left</b>				<b>Right</b>				
19									
20	APB	5.6			APB	3.7			
21	FDI	9.8			FDI	12			
22	ADM	9.9			ADM	9.1			
23									
24	<b>SI =</b>	$\frac{APB * FDI}{ADM}$			<b>SI =</b>	$\frac{APB * FDI}{ADM}$			
25									
26									
27	<b>SI =</b>	<b>5.54</b>			<b>SI =</b>	<b>4.88</b>			
28									



# We can not analyse “Notes”



The screenshot shows a Microsoft Excel spreadsheet titled "Cranial Vault Data for Stats 100416". The spreadsheet has columns labeled S, T, U, V, W, X, Y, Z, AA, AB, AC, AD, and AE. The data is organized into rows, with the first row (row 1) containing headers: "RTT:", "Blood loss:", "Complication 1", "Complication 2", "Complication 3", and "Notes". The "Notes" column contains various text entries, some of which are highlighted in a pink box. The text in the pink box is "Code it if you want to analyse it." The spreadsheet also shows a ribbon with various tabs like File, Home, Insert, Page Layout, Formulas, Data, Review, View, and Developer. The status bar at the bottom shows "Vault Master Data", "Explanation of Data", and "Research Questions".

	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	RTT:	Blood loss:	Complication 1	Complication 2	Complication 3	Notes							
2	0	200											
3	0												
4	0	100	CSF leak			No notes available to detail complication							
5	0												
6	0	100											
7	0	120				Noted in operation note, nil further procedures							
8	0		CSF leak			No further notes to explain complication							
9	0												
10	0		Medical complication			Lifelong epilepsy and hemiplegia due to intraoperative ischaemia, pneumonia post-op							
11	0	190											
12	0	250											
13	0	400	Medical complication			Pneumonia with sepsis; no further notes available							
14	0	250											
15	2	50				No notes available							
16	0	30											
17	0	50											
18	0	80	Medical Complication			Air embolus							
19	0												
20	0	225	Growing fracture			No further notes to explain complication							
21	0	400	CSF leak	Growing fracture		No further notes to explain complication							
22	0	80				No notes available on nature of revision. Was booked for Navian cranioplasty but this was cancelled.							
23	0	25											
24	0	150											
25	0	50											
26	0												
27	2	?				No notes available							
28	0	100											
29	0	50											
30	0	80											
31	0	150				No notes available							
32	0	60				Le Fort 3/Fronto-orbital maxillary advancement on 13/08/02 with RL; return to theatre for post-op bleeding after rem							
33	0	160				No notes available on nature of revision.							
34	0	150											
35	0	100											
36	0	115											
37	0	50											
38	0	150											
39	0	75											

# Can I believe these numbers?

- **Data checking**
  - High standards of measurement and data entry
  - Checks on range and logic
  - Compatibility with clinical expectations
- **Data summary**
  - Frequency and cross tabulation for categorical data
  - Mean, standard deviation and range for continuous data
  - Bivariate associations or univariate analysis
  - Simple tables and figures
- **Only the right individuals are included**
  - Satisfy inclusion and exclusion criteria
  - Are individuals representative of the reference population
- **Missing data**
  - Reason for missing related to the individual or independent
  - What proportion of data is missing
  - How will the computer package handle the missing data

# Summary

- Prepare your data collection sheet and codebook prior to commencing a study. Statisticians are happy to help
- Poor quality data may lead to misleading results and incorrect conclusions
- Collecting and managing data takes time but is worth the effort
- Please ask a statistician for guidance if you are unsure

# How to contact the Statistics Unit

- **Email:**

Statistical.Services@qimrberghofer.edu.au

- **Location:**

Level 12, Bancroft Building, QIMR Berghofer, 300 Herston Road, Herston

- **Website:**

<https://www.qimrberghofer.edu.au/our-research/scientific-services/qimr-berghofer-statistics-unit/>