Defining variables and data management

Nicki Tiffin

RESPICE PROSPICE

SANBI
South African National
Bioinformatics Institute

A word on ethics

- Any project working with data from participants must have ethics approval from an Ethics Review Board

Recommended good practice:

- Before you receive your data, ask for a copy of the **ETHICS APPROVAL** and read it thoroughly
- Keep it on file

Nicki Tiffin

A word on ethics

- **INFORMED CONSENT** must have been obtained from every participant.

- Before you receive your data, ask for a copy of the informed consent form as well as the information provided to participants. Read it well and keep it on file.

Nicki Tiffin

A word on ethics

- Ethics approval and informed consent may be obtained just for a particular study

- Sometimes there is a clause that allows for **SECONDARY USE** of data or samples

Nicki Tiffin

A word on ethics

SECONDARY USE CONSENT

Secondary use consent allows the researcher to use the data/samples to answer research questions aside from the original research question, with the participants' agreement.

If you are not addressing the primary research question for which ethics was granted, ensure that there is ethics approval for secondary use.

Nicki Tiffin

Data storage

- Many datasets stored in EXCEL SPREADSHEETS
- This is not the best option, but it is manageable
- There are some potential pitfalls to be aware of:

Nicki Tiffin

Data storage

- Data can be altered, deleted, rearranged by mistake
- Files can be corrupted
- Excel sometimes autoformats numbers and dates and you lose information
- The fields you delete now might turn out to be important later

ALWAYS KEEP A MASTER COPY OF THE ORIGINAL DATA SOMEWHERE SECURE

Nicki Tiffin

Data storage

Anonymity

- There should be no information in the dataset that can be used to identify individuals UNLESS ethics approval specifically allows YOU to know who the participants are

- If necessary, request the PI or data generator to **DE-IDENTIFY**, or **ANONYMISE** individuals before you receive the data

Nicki Tiffin

Data storage

Security and backup

- Participant data needs to be well-secured
- Password-protect computer/files and secure hard copies (locked cupboard/filing cabinet)
- BACKUP all your work

Nicki Tiffin

Data storage

Data format

- In Excel, each data type, also called a **field**, is listed in a separate column.

- TAB-SEPARATED TEXT (.txt)

  The data are in a 'flat file' and fields are separated by a TAB ('\t')

  (You can open these in excel by right-click and 'open with' -> excel)

Nicki Tiffin

Data storage

Data format

- In Excel, fields are separated into columns
- COMMA-SEPARATED VALUES (.csv)

The data are in a ‘flat file’ and fields are separated by a COMMA (‘,’)

You can open these in excel by right-click and ‘open with’ -> excel

Notepad, and Notepad++ are good text editor programmes for working with these data files

Nicki Tiffin

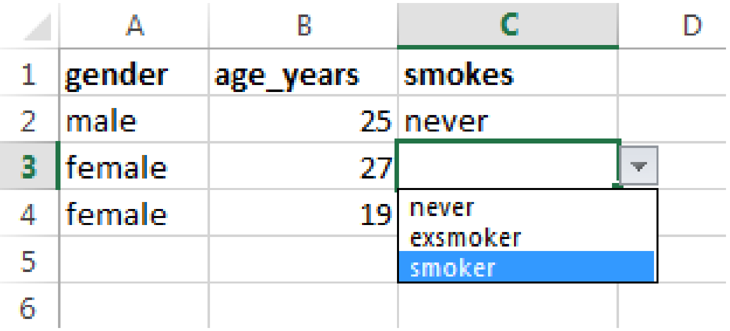
Data cleaning

A note on spaces

- In an excel file, some entries may contain extra spaces.
- Normally this is not a problem, but depending on what stats software you use it can get messy…
- Try and avoid spaces in your field names and data where you can. Rather use an underscore

  e.g. “never_smoked” instead of “never smoked”

Data cleaning



	A	B	C	D
1	gender	age_years	smokes	
2	male	25	never	
3	female	27		
4	female	19		
5				
6				

A note on capitalisation

- Some programs are case sensitive

e.g. , so ‘Male’ is not the same as ‘male’

- When you capture data, make sure that you capture each field in a standardized way to avoid spaces, capitalization issues, multiple ways of entering the same data

e.g. in Excel, you can make a drop-down box for the field you are capturing, that only allows certain options to be entered

‘YES’ and ‘NO’ options avoids entry of ‘Y’, ‘yes’, ‘ no’, e.t.c

Nicki Tiffin

Data cleaning

DATA CODING

- In what form is data actually stored?
- Make sure you are clear on the coding
- Be careful with greater than, less than and equal to
- IMPORTANT: Clarify missing data vs negative answer

e.g. yes=1, no=0, no_data=9
never=0, sometimes=1, often=2
50_or_less=1, greater_than_50=2

Nicki Tiffin

What is a VARIABLE?

- A characteristic, number, or quantity that increases or decreases over time, or takes different values in different situations

i.e. the data types that have been collected from individuals in the study

Nicki Tiffin

What is a **VARIABLE**?

- The outcome you are investigating is called the **DEPENDENT VARIABLE**

  Often denoted as **Y**

- The exposures, or other variables that may or may not cause the outcome, are called the **INDEPENDENT VARIABLES**. There may be several

  Often denoted as $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$ …

Types of variables

Conceptual variable:

What is the ‘concept’ you are trying to measure?

Operational variable:

What do you actually measure?

Nicki Tiffin

Types of variables

Conceptual	Operational
Health status	
Obesity	

Nicki Tiffin

Types of variables

Conceptual	Operational
Health status	Blood pressure BMI No. of visits to doctor
Obesity	BMI Hip to waist ratio

Nicki Tiffin

Data types

NUMERICAL DATA = quantitative data

- Data that is a numerical measure, number.

CONTINUOUS NUMERICAL DATA

- Data that can fall anywhere along a range of numbers

  e.g. height, weight, age, blood pressure

Nicki Tiffin

Data types

NUMERICAL DATA = quantitative data

- Data that is a numerical measure, number.

DISCRETE NUMERICAL DATA

- Data that can only fall at specific numbers, and not at values in-between (often a count).

  e.g. number of children, number of cigarettes smoked, number of patients treated

Nicki Tiffin

Data types

CATEGORICAL DATA

- Data that does not have numerical value but can be sorted by category

  e.g. colour, gender, country of origin

- Can have implied order = **ORDINAL**

e.g. never/sometimes/always

below_20/20_to_50/above_50

Nicki Tiffin

Data types

CATEGORICAL DATA

- Many have no implied order = **NOMINAL**

  e.g. red/green/blue

  Africa/Asia/Europe

- Special case of only two categories

  Coded as 0/1 = **BINARY**

  e.g. Male=0, Female=1

  No=0, Yes=1

Nicki Tiffin

Data types

BINNING DATA

N.B. This does NOT mean 'throwing away'

- Sometimes with complex data it may make more sense to bin a continuous numerical value into categories

Nicki Tiffin

Data types

BINNING DATA

e.g. age_in_years (within range 0-100yrs) can be binned into three categories in age_category:

<20years, 20-50yrs, >50yrs

age_in_years is numerical continuous data

age_category is ordinal categorical data

Nicki Tiffin

Why do we care?

Different statistical tests are designed to handle different types of variables.

If you don’t know what type of variable you are dealing with, you may do the wrong type of analysis.

Nicki Tiffin

The variable table

- When designing a study, or at the latest before doing any analysis, summarise what type of data you have, in a table.
- Draw up a variable table before seeing a statistician, and they will be your friend.

Nicki Tiffin

The variable table

Variable_name	Description	Type_of_data	Coding
Age	Age in years at diagnosis	Numerical, continuous	Integer
Height	Height in metres at diagnosis	Numerical, continuous	Float (number with decimal point)
Gender	Male or female gender	Categorical, binary	Female=0, male=1
Children	Number of children at diagnosis	Numerical, discrete	Integer
Stage	Tumour severity graded by xyz criteria	Categorical, ordinal	0=stage 1 1=stage 2 2=stage 3 Empty field =no data

Nicki Tiffin

When can we do data sharing?

When the data can be compared/combined

- Meta analyses – joining datasets for statistical power
- Validation – comparing findings from one dataset in another

Data Standardisation:

- Collect your data in a standardised way which matches how other people collect these data

Data Harmonisation:

- Try and retrofit your datasets to report and combine the same data even though you collected the same metrics differently

Nicki Tiffin

When can we do data sharing?

Data Standardisation:

Most common data standard used by clinicians?

Nicki Tiffin

When can we do data sharing?

Data Standardisation:

Most common data standard used by clinicians?

ICD-10

Nicki Tiffin

When can we do data sharing?

Data Standardisation:

Most common data standard used by clinicians?

ICD-10

Others you may be using frequently:

- ATC (pharmacy)
- Loinc – labs, universal
- Local labs data in South Africa – DISA and TRAK
- Moving health records around: FHIR spec, HL7

There are many more and you should check for common standards you can use when you plan your research

Nicki Tiffin

When can we do data sharing?

Ontologies include structure, hierarchies and relationships, with parent terms and child terms

- They can be shown as a diagram with nodes and edges
- A commonly used ontology in genetics research is the **gene ontology**

  www.geneontology.org

- Another example is the PHENEX ontology, for phenotype diversity

  Balhoff et al. (2010) PLOS ONE 5(5): e10500.

Nicki Tiffin

Structure of ontologies

Nodes:

Genes

Cellular structures

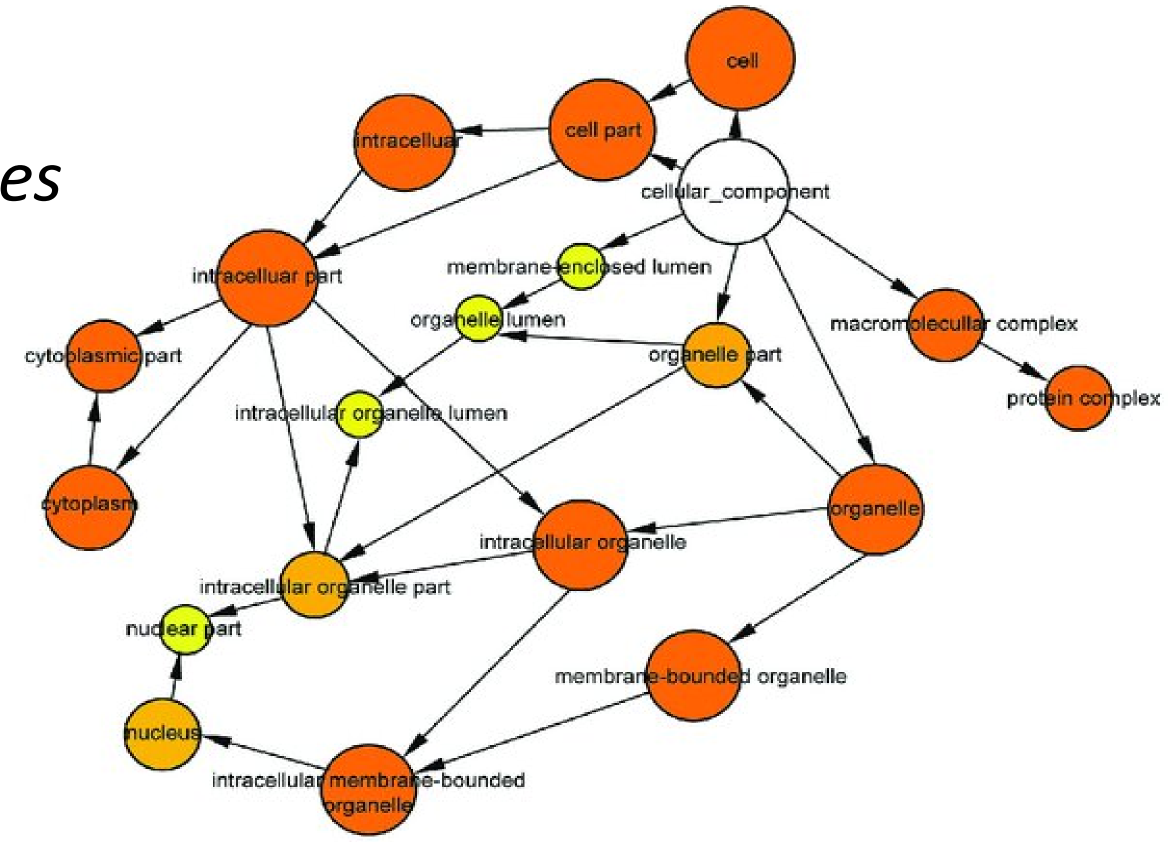
functions

Relations:

is a

is part of

regulates



Nicki Tiffin

Structure of ontologies

- Ontologies can also be used as **controlled vocabularies** – a common language for naming variables

E.g.

Building the variable dictionary (‘code book’)

Deciding variable names and descriptions

Standardising to common terminology

Nicki Tiffin

When can we do data sharing?

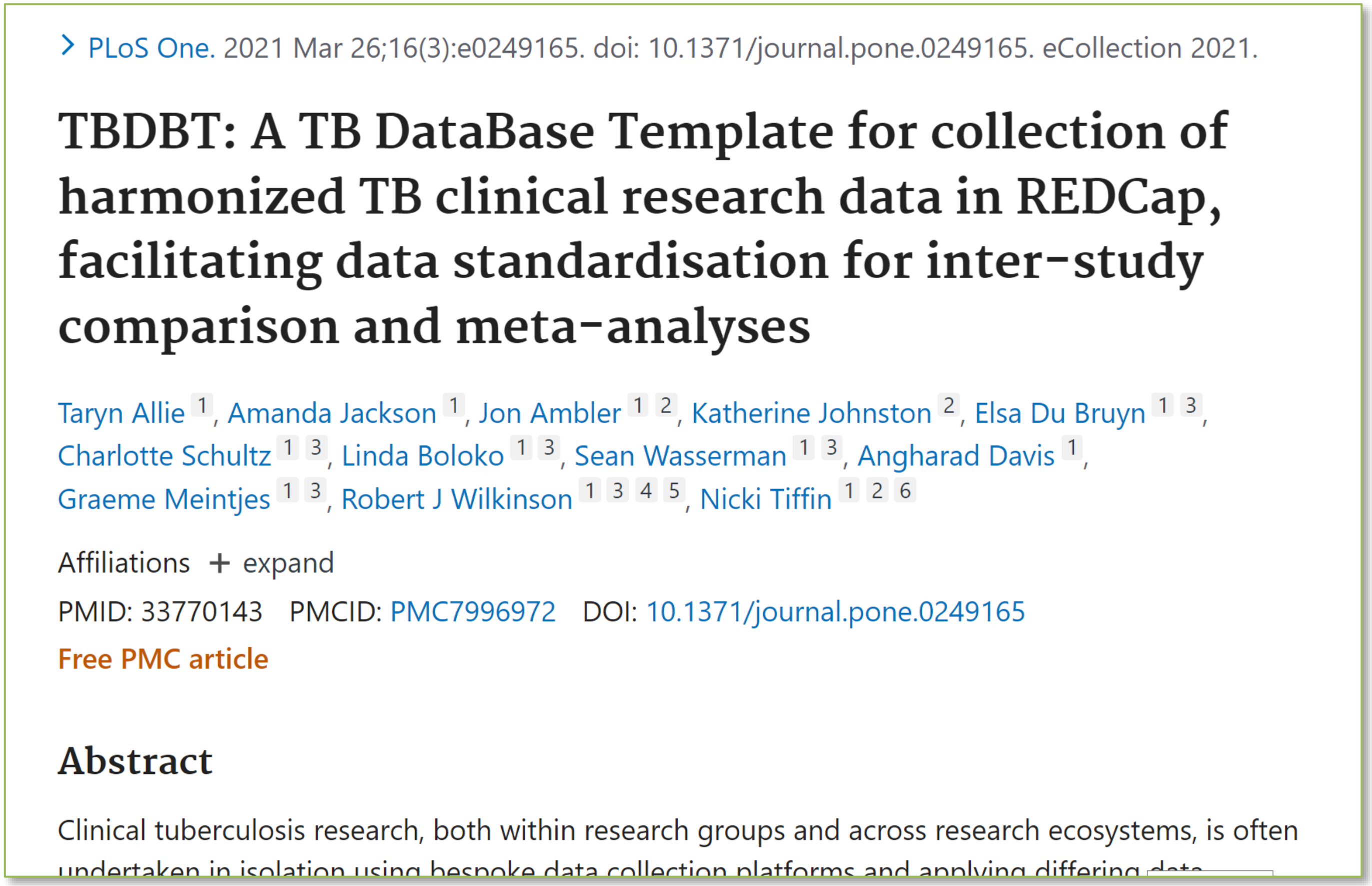
Data Harmonisation:

- Extremely time-consuming; often researchers don’t realise how much work this will require
- Sometimes the data just can’t be matched

Ever had TB
Has current TB
Presenting with TB
How many TB episodes?

Nicki Tiffin

Allie T et al. **TBDBT: A TB DataBase Template for collection of harmonized TB clinical research data in REDCap, facilitating data standardisation for inter-study comparison and meta-analyses.** PLoS One. 2021 Mar 26;16(3):e0249165



› PLoS One. 2021 Mar 26;16(3):e0249165. doi: 10.1371/journal.pone.0249165. eCollection 2021.

TBDBT: A TB DataBase Template for collection of harmonized TB clinical research data in REDCap, facilitating data standardisation for inter-study comparison and meta-analyses

Taryn Allie[1], Amanda Jackson[1], Jon Ambler[1,2], Katherine Johnston[2], Elsa Du Bruyn[1,3], Charlotte Schultz[1,3], Linda Boloko[1,3], Sean Wasserman[1,3], Angharad Davis[1], Graeme Meintjes[1,3], Robert J Wilkinson[1,3,4,5], Nicki Tiffin[1,2,6]

Affiliations + expand

PMID: 33770143 PMCID: PMC7996972 DOI: 10.1371/journal.pone.0249165

Free PMC article

Abstract

Clinical tuberculosis research, both within research groups and across research ecosystems, is often undertaken in isolation using bespoke data collection platforms and applying differing

Nicki Tiffin

As datasets get larger:

- Unlikely to look at all of raw data – often too large to open, or may not be people-readable

- How do you know your data are clean?

Nicki Tiffin

As datasets get larger:

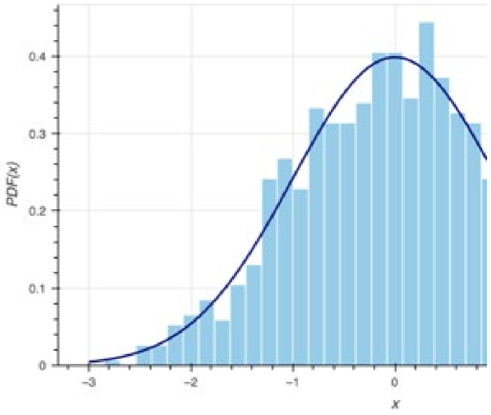
- Unlikely to look at all of raw data – often too large to open, or may not be people-readable
- How do you know your data are clean?

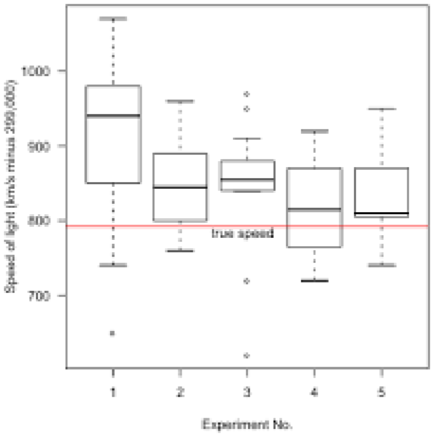
Nicki Tiffin

Data Exploration

Getting a sense of the data

- Plots and distribution figures





- Aggregate data

  - means, medians, counts

  - ranges, max and min values

Nicki Tiffin

Sanity checks:

Looking for nonsensical data

- Date of birth and date of death
- Current age
- Sex-based phenotypes
- Upper and lower limits of biological measures
- Identify outliers and inspect those data elements

Nicki Tiffin

Sanity checks:

Assess missing data:

- Missing completely at random

  Which records are missing is independent of observed and unobserved variables, there are no outside influencers of missingness

- Missing at random

  Probability of being missing is random within the observed data

- Missing not at random

  Systematically related to the unobserved data, that is, the missingness is related to events or factors which are not measured by the researcher – the missingness is directly related to the value of the missing variable.

Decide which fields and records to include/exclude

Nicki Tiffin

Thank you

ntiffin@uwc.ac.za

South African National Bioinformatics Institute
University of the Western Cape
South Africa

Acknowledging funding from:
Bill & Melinda Gates Foundation, Calestous Juma Fellowship
UKRI/MRC