

EACS HIV Summer School 2017

Introduction to collection of data

Valentina Cambiano,
Institute for Global Health, University College London

Conflict of Interests

I received personal fees from Merck Sharp & Dohme Limited in 2015.

Topics to be covered

- What data should you collect
 - Identifying information
 - Characteristics
 - Information related to your research question
 - Endpoints
 - Main exposures
 - Potential confounders or effect modifiers
- Type of data

Identifying Information

- Enable you to identify individuals within study
- Avoid people being included in study more than once
- May need to work back to correct errors in data
- Needed if you will be performing data linkage

Identifying Information

Hospital number

Co-morbidities

Age

Name

Ethnicity

Which are
identifying
information?

Co-medication

VL at study
entry

Date of birth

Sexual orientation

Study ID number

Gender

CD4 count at
study entry

Hospital name

Level of
education

Identifying Information

Hospital number

Age

Co-morbidities

Name

Ethnicity

Which are
identifying
information?

Co-medication

VL at study
entry

Date of birth

Sexual orientation

Study ID number

Gender

CD4 count at
study entry

Hospital name

Level of
education

Characteristics

- Describe the population studied
- In longitudinal studies, often collected at study entry and referred to as baseline characteristics
- Characteristics often collected:
 - **Socio-demographic:** Date of birth, Ethnicity, Gender, Level of Education
 - **HIV-related:** Viral load & CD4 at diagnosis (or study entry or ART initiation), AIDS diagnosis, cART regimen
 - **Others:** co-morbidities, co-medication

Study Endpoints

- They can be referred to as outcome, event of interest, disease, dependent variable
- A well defined study endpoint should:
 - Be defined in advance
 - Address the primary aim of the study
 - Have biological/clinical relevance
 - Be appropriate for the population included in the trial
- Well defined study endpoints (primary and secondary) are equally important for all study designs, whether RCTs or observational studies

Example - Primary Endpoint

- *"We wish to compare the efficacy of combination antiretroviral therapy (cART) in people who uses drugs (PVD) compared to non-drug users in previously ART-naive adults in an observational study"*

Example - Primary Endpoint

- *"We wish to compare the efficacy of first-line antiretroviral therapy (cART) drugs (PWD) compared to no treatment in previously ART-naïve adults in a randomised study"*
- **Clinical:** New AIDS-defining event, New non-AIDS defining event, Death
- **Virological:** Achieving VL<50 copies/ml at 1 year after starting cART, time to viral suppression, time to viral rebound
- **Immunological:** CD4>200 cells/mm³, time to CD4 increase >100 cells/mm³
- **Other:** on ART at 1 year, ART switches, adherence, quality of life, toxicity

Which primary endpoint would you choose?

Primary and Secondary Endpoints

- All clinical trial protocols should state one (sometimes two) primary endpoint
- Main conclusions should be based on the results from this endpoint
- Pre-defined secondary endpoints can also provide supportive data
- For event data (i.e. diagnosis of an illness or condition) it is important to record date of event as well as fact that event occurred

Main Exposures

- They can be referred to as predictors of interest, factors of interest, independent variables, ...
- They should ideally also be clearly defined in advance
- In an RCT the exposure is typically the interventions you are randomizing people to, so usually one or two
- In cohort studies, the exposures are the factors that you may want to evaluate whether they predict a certain endpoint. Therefore, there is more flexibility and you may have a number of exposures

Example - Main Exposures

- *"We wish to compare the efficacy of cART in PWD compared to non-drug users in previous cART naïve adults in an observational study"*

Which measure of exposure would you use?

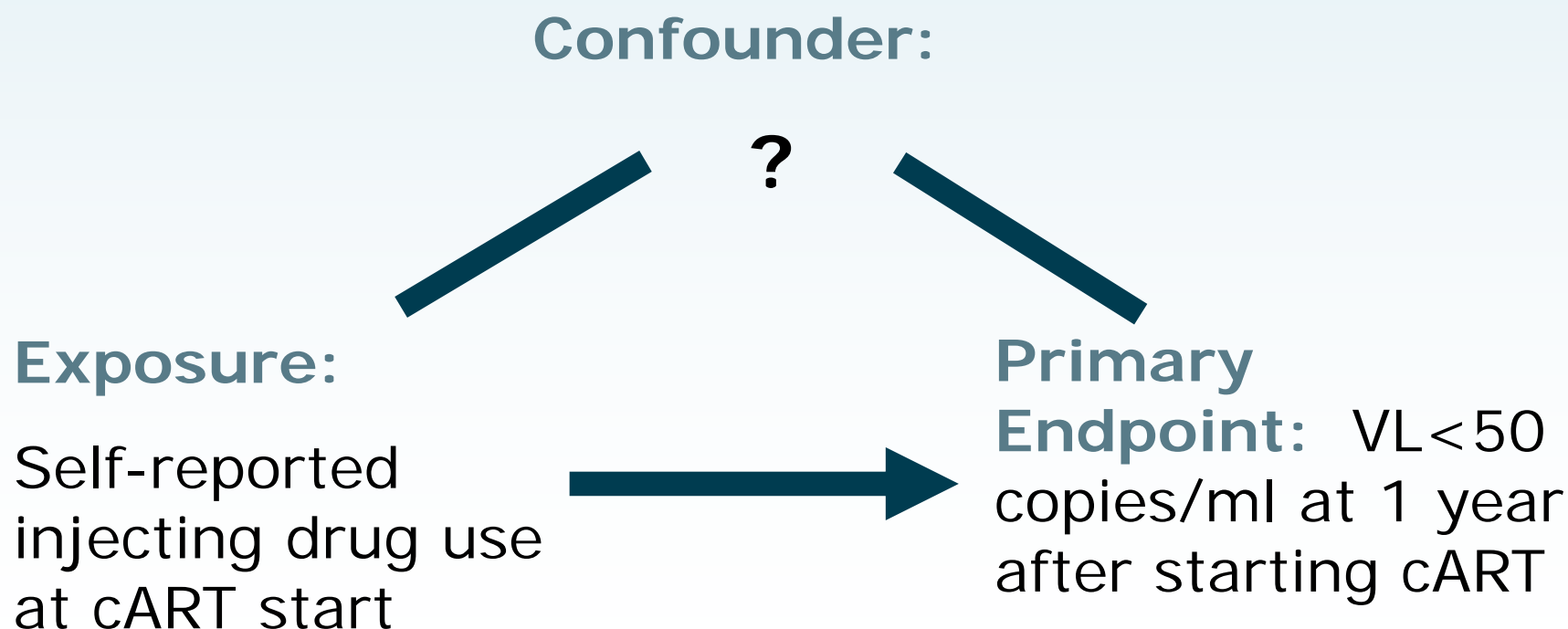
- Individual reports having ever taken drugs
- Individual reports current (at start of cART) injecting drug use
- Individual reports having ever injected illegal substances

Confounders (1)

- Confounding is particularly an issue in observational studies, as randomisation limits confounding in RCTs
- It occurs when a factor exists that is associated with **both** the **exposure** and **outcome** of interest

Example - Confounders (1)

"We wish to compare the efficacy of cART in PWD compared to non-drug users in previously ART-naïve adults in an observational study"



Confounders (2)



- Although one can never be certain that all have been accounted for, it is important to collect information on any known confounders
- It is possible to adjust for potential confounders using statistical (multivariable) models

Effect modifiers

- An **effect modifier** is a variable that differentially (positively and negatively) modifies the observed effect of an exposure on the endpoint
- An effect modifier is a type of interaction
- Effect modification is a phenomenon in which the exposure has a different impact in different circumstances

Example – effect modifiers

- Monoamine oxidase inhibitors (MAOI) are used to treat depression
- People who eat certain foods, such as cheese, are at higher risk of stroke if they take MAOI
- MAOI is an effect modifier
- MAOI is NOT associated with stroke, and so is NOT a confounder

Circumstance	Exposure		Endpoint
Taking MAOI	Cheese		Stroke
No MAOI	Cheese		Stroke

Exposures, confounders and effect modifiers

- As measurements may change over the study period, a patient's status should be re-assessed at regular times during the study
- The frequency at which each measurement is assessed will depend on the likelihood of it changing over time, as well as the reliability of the data sources
- Example: dietary factors, smoking status, alcohol consumption

Topics to be covered

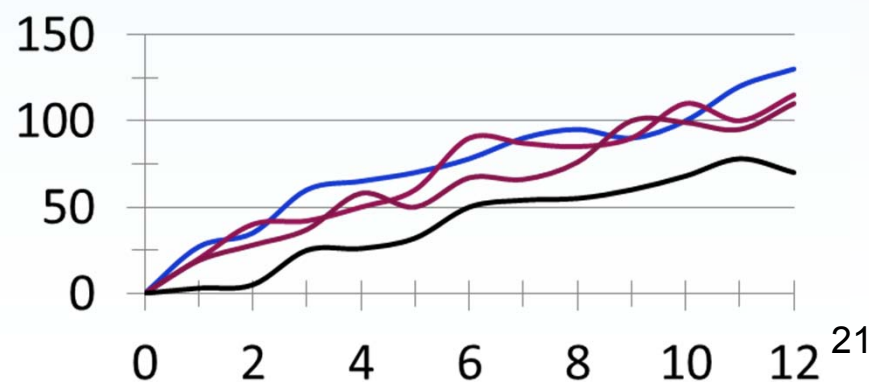
- What data should you collect
 - Identifying information
 - Baseline characteristics
 - Information related to your research question
 - Endpoints
 - Main exposures
 - Potential confounder or effect modifier
- Type of data

Types of data

- There are two main types of data
 - Categorical/qualitative



- Numerical/quantitative



Categorical data

- **Binary data**

- Two categories (yes/no, dead/alive, male/female)

- **Nominal data**

- More than two categories, no ordering to the groups (e.g. HIV exposure category, country of birth)

- **Ordinal data**

- More than two categories, some inherent ordering (e.g. CDC stage, education, some quality of life scores)

Numerical data

- **Discrete data**

- Can only take whole numbers within a given range (e.g. number of sexual partners)

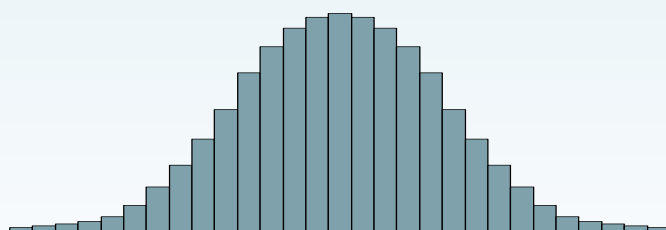
- **Continuous data**

- Can take any value in a range (e.g. height, CD4 count, total cholesterol). Continuous data can be censored (i.e. they can only be measured within a certain range), this includes:
 - Time to event data (can only assume positive values; e.g. survival from HIV diagnosis, time to rebound)
 - Lapse data (The lowest value is the limit of detection; e.g. HIV RNA data)
 - Proportions (can only assume values between 0 and 1)

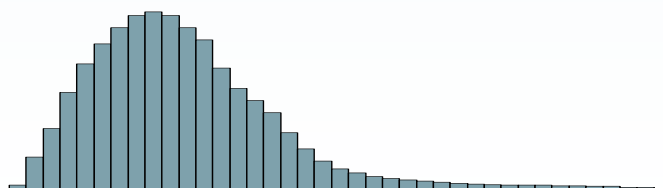
The distributions of quantitative data

The choice of summary statistics and the most appropriate analytical method will depend on the shape of the distribution

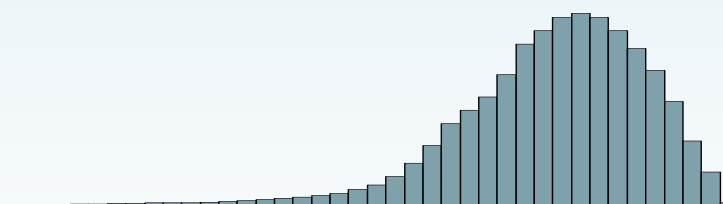
Symmetrical, bell-shaped
'Normal' distribution



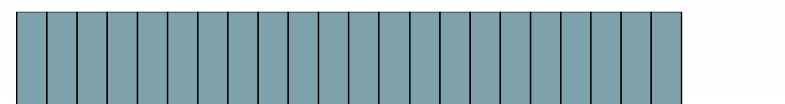
Positive skewness
common in laboratory data
eg. CD4 counts in HIV+ves



Negative skewness



Uniform distribution
Equal probability of taking
any value in the range



Summarising quantitative data

- We usually quote two measures:
 - A measure of the *average* value
 - A measure of how *variable* the data are

Type of data	Average	Variability
Numerical, normally distributed	Mean	SD/variance
Numerical, skewed	Median	Range/IQR
Categorical, nominal	Mode	No suitable measure – give % in each category
Categorical, ordinal, only a few categories	Mode	
Categorical, ordinal, reasonable number of categories	Median	

Summary

- It is important to consider study design and the research question to be addressed *before* beginning data collection
- A clear definition of exposure, endpoint and identification of potential confounders and effect modifiers prior to the start of the study means that information on these can be collected and adjusted for