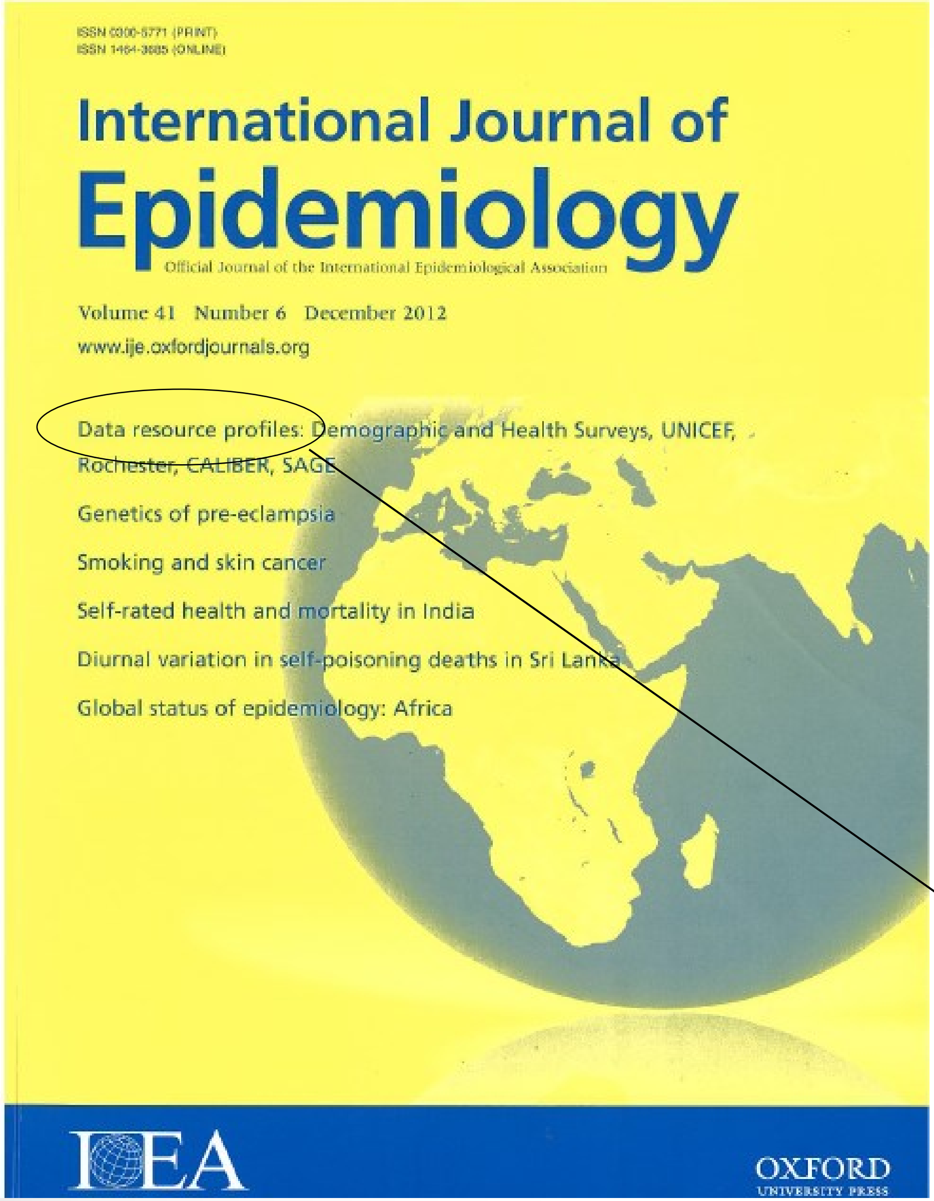
Using GUS data – looking at the scope of analysis possibilities

Dr Pamela Warner

Centre for Research on Families & Relationships,

University of Edinburgh



ISSN 0300-5771 (PRINT)
ISSN 1464-3685 (ONLINE)

International Journal of
Epidemiology

Official Journal of the International Epidemiological Association

Volume 41 Number 6 December 2012
www.ije.oxfordjournals.org

Data resource profiles: Demographic and Health Surveys, UNICEF, Rochester, CALIBER, SAGE

Genetics of pre-eclampsia

Smoking and skin cancer

Self-rated health and mortality in India

Diurnal variation in self-poisoning deaths in Sri Lanka

Global status of epidemiology: Africa

IEA

OXFORD
UNIVERSITY PRESS

"In God we trust, all others (must) bring data"

Lynch J & Stuckler D 2012; *IJE* **41**: 1503-1506

International Journal of Epidemiology *Editorial,* about their series

Data Resource Profiles

"We can't do epidemiology without data. Data are central to epidemiology's three main challenges:

to describe health states in populations;

to make inferences about their causes; and

to apply this knowledge to improve health.

The more high quality data we have to support these three tasks, the better."

"In God we trust, all others (must) bring data"

Lynch J & Stuckler D 2012;
IJE **41**: 1503-1506

International Journal of Epidemiology *Editorial about* their series

Data Resource Profiles

"We can't do epidemiology without data. Data are central to epidemiology's three main challenges:

to describe health states in populations;

to make inferences about their causes; and

to apply this knowledge to improve health.

The more high quality data we have to support these three tasks, the better."

'**Aye, aye!**'
we say

Lynch J & Stuckler D 2012;
IJE **41**: 1503-1506

They go on to say….

"Paradoxically, there is also a perception that we have too much data and not enough useful analysis…

Available data often go unused because they are not well enough documented, lack accessible how-to guides for their use, …

Some data may also require analytical skills that are in short supply; or people may simply be unaware of their existence or unable to access them."

Lynch J & Stuckler D 2012; *IJE* **41**: 1503-1506

They go on to say….

“Paradoxically, there is also a perception that we have too much data and not enough useful analysis…

Available data often go unused because they are not well enough documented, lack accessible how-to guides for their use, …

Some data may also require analytical skills that are in short supply; or people may simply be unaware of their existence or unable to access them.”

Lynch J & Stuckler D 2012; *IJE* **41**: 1503-1506

I say….

This perception is both paradoxical and disingenuous.

They go on to say….

"Paradoxically, there is also a perception that we have too much data and not enough useful analysis…

Available data often go unused because they are not well enough documented, lack accessible how-to guides for their use, …

Some data may also require analytical skills that are in short supply; or people may simply be unaware of their existence or unable to access them."

Lynch J & Stuckler D 2012; *IJE* **41**: 1503-1506

I say….

This perception is both paradoxical and disingenuous.

GUS data is well documented, including face-to-face workshops!

They go on to say….

"Paradoxically, there is also a perception that we have too much data and not enough useful analysis…

Available data often go unused because they are not well enough documented, lack accessible how-to guides for their use, …

Some data may also require analytical skills that are in short supply; or people may simply be unaware of their existence or unable to access them."

Lynch J & Stuckler D 2012; *IJE* **41**: 1503-1506

I say….

This perception is both paradoxical and disingenuous.

GUS data is well documented, including face-to-face workshops!

Re analytical skills, more anon..

They go on to say….

"Paradoxically, there is also a perception that we have too much data and not enough useful analysis…

Available data often go unused because they are not well enough documented, lack accessible how-to guides for their use, …

Some data may also require analytical skills that are in short supply; or people may simply be unaware of their existence or unable to access them."

Lynch J & Stuckler D 2012; *IJE* **41**: 1503-1506

I say….

This perception is both paradoxical and disingenuous.

GUS data is well documented, including face-to-face workshops!

Re analytical skills, more anon..
Awareness/ access are not an issue with GUS data.

Before proceeding we need to broaden out from the relatively narrow health focus of epidemiology….

We can't **achieve better understanding of outcomes for members of the population***, without data. Data are central to **population/social research's** three main challenges:

to describe **outcomes** in populations;

to make inferences about their causes; and

to apply this knowledge to improve **the ongoing situation, or at least outcomes for future cohorts**.

The more high quality data we have to support these three tasks, the better.

* *outcomes = health, well-being, education, work, income, etc…*

Adapted from - Lynch J & Stuckler D 2012; *IJE* **41**: 1503-1506

Outline of this presentation

- Is ‘lack of analysis skills’ the only problem?
- The secondary analysis research process
  - Differs from primary research
  - Need to iterate to a research aim that is both worthwhile & feasible
- Thinking about analysis possibilities (TAAP)
  - Some examples

Analysis skills

What about analysis skills then….?

- The analysis methods that can be applied *will* be limited by analysis skills

What about analysis skills then….?

- The analysis methods that can be applied *will* be limited by analysis skills
- But just as important, or ***more*** important, are:
  - a worthwhile & feasible research question (research aim)

worthwhile = the knowledge needed next in the topic

What about analysis skills then….?

- The analysis methods that can be applied *will* be limited by analysis skills
- But just as important, or ***more*** important, are:
  - a worthwhile & feasible research question (research aim)

worthwhile = the knowledge needed next in the topic

feasible = can be answered!

What about analysis skills then….?

- The analysis methods that can be applied *will* be limited by analysis skills
- But just as important, or ***more*** important, are:
  - a worthwhile & feasible research question (research aim)
  - the data needed to answer that question

worthwhile = the knowledge needed next in the topic

feasible = can be answered!

the ***specific*** data - not just some related variable(s)

What about analysis skills then….?

- The analysis methods that can be applied *will* be limited by analysis skills
- But just as important, or ***more*** important, are:
  - a worthwhile & feasible research question (research aim)
  - the data needed to answer that question
  - effective communication of the findings to the intended audience

worthwhile = the knowledge needed next in the topic

feasible = can be answered!

the ***specific*** data - not just some related variable(s)

effective communication – depends both on how 'said', and 'listener'

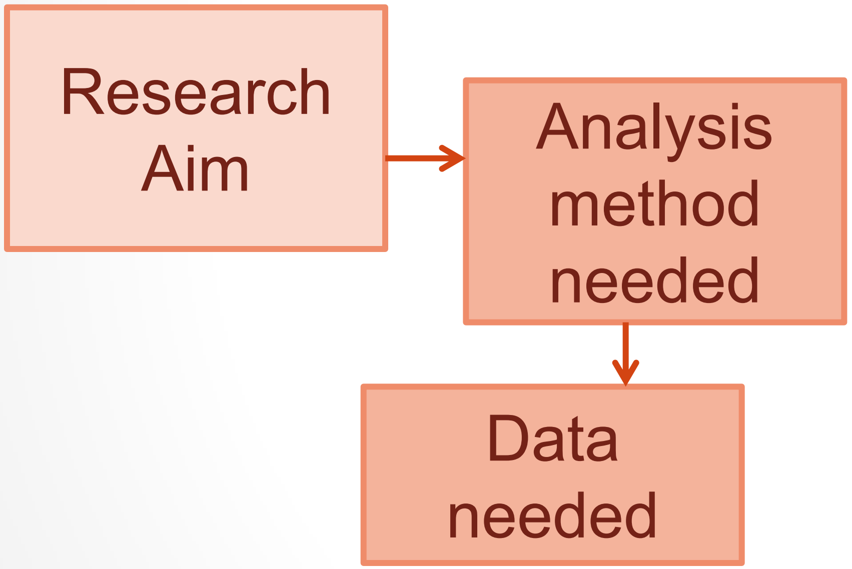
What is needed in addition to analysis skills ….?

- Research question + data + analysis method are **interdependent**
- In primary research a typical temporal ordering is:

Research Aim

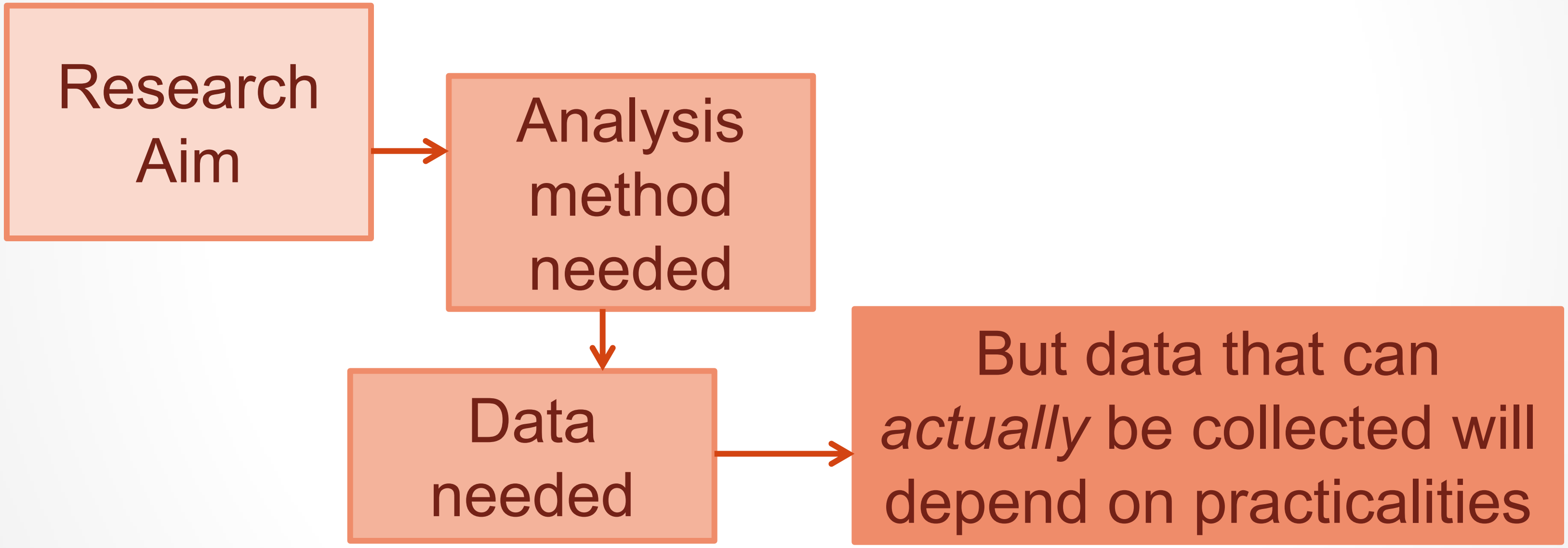
What is needed in addition to analysis skills ….?

- Research question + data + analysis method are **interdependent**
- In primary research a typical temporal ordering is:



What is needed in addition to analysis skills ….?

- Research question + data + analysis method are **interdependent**
- In primary research a typical temporal ordering is:



Research Aim

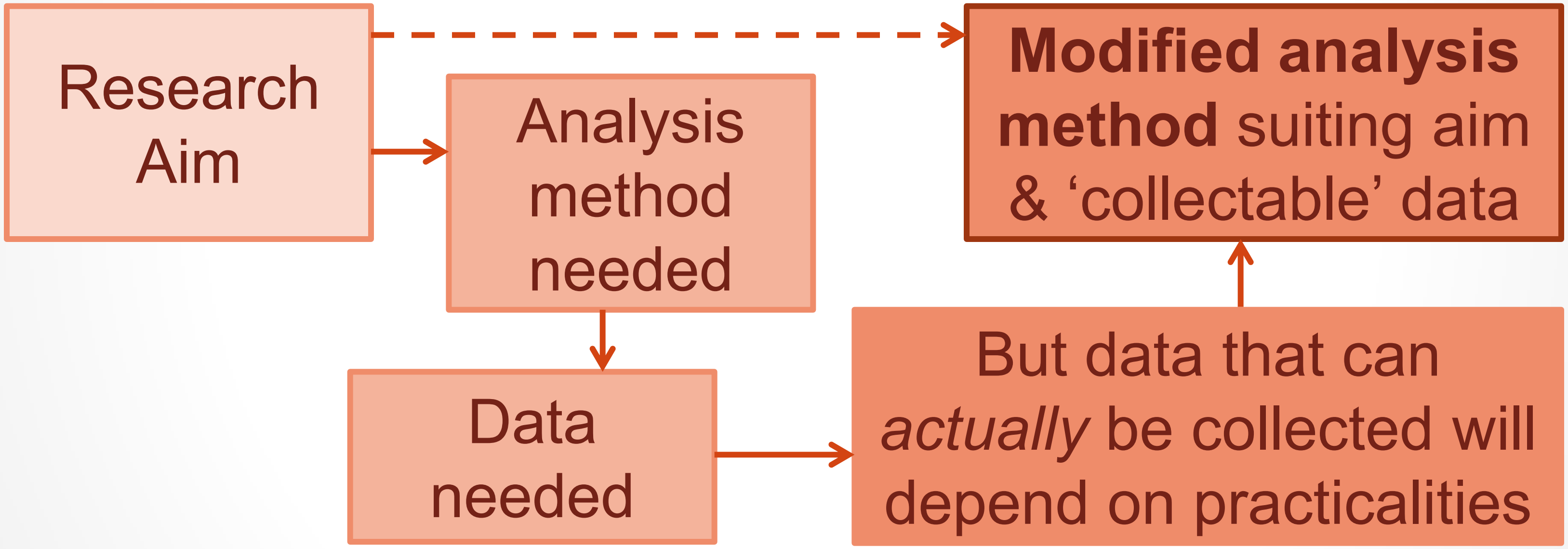
Analysis method needed

Data needed

But data that can *actually* be collected will depend on practicalities

What is needed in addition to analysis skills ….?

- Research question + data + analysis method are **interdependent**
- In primary research a typical temporal ordering is:



Research Aim

Analysis method needed

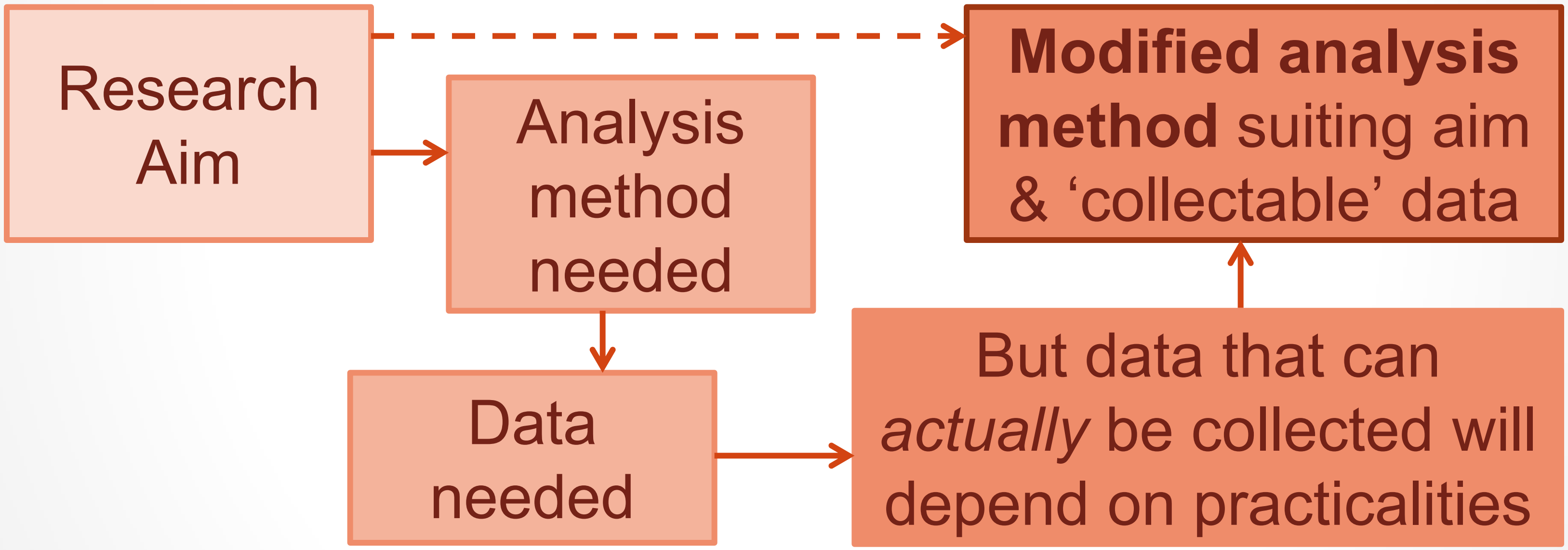
Data needed

But data that can *actually* be collected will depend on practicalities

Modified analysis method suiting aim & 'collectable' data

What is needed in addition to analysis skills ….?

- Research question + data + analysis method are **interdependent**
- In primary research a typical temporal ordering is:



Research Aim

Analysis method needed

Data needed

But data that can *actually* be collected will depend on practicalities

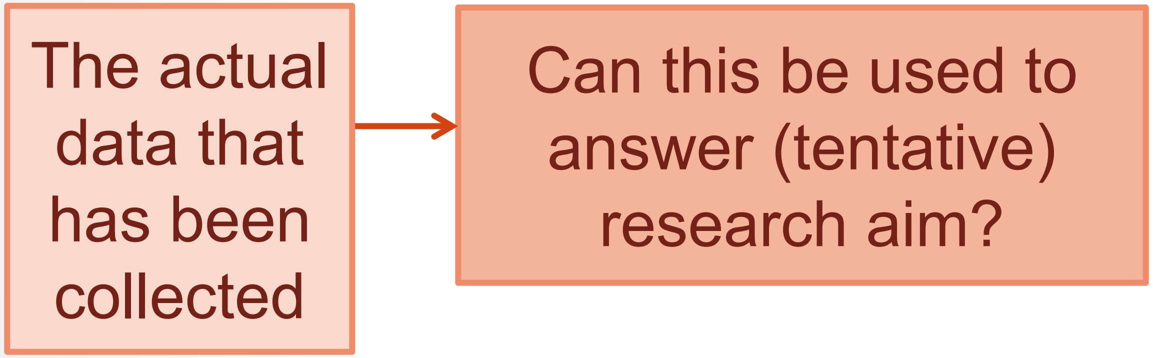
Modified analysis method suiting aim & 'collectable' data

- Lack of analysis skills might limit the complexity of analysis methods that you could apply, but that need not be a 'deal-breaker *(see later…)*

Secondary analysis process

Research process in secondary analysis

- In secondary analysis the data set has been collected long before your decision to undertake research (usually!)
- So a better way to view the 'secondary analysis' research process would be:

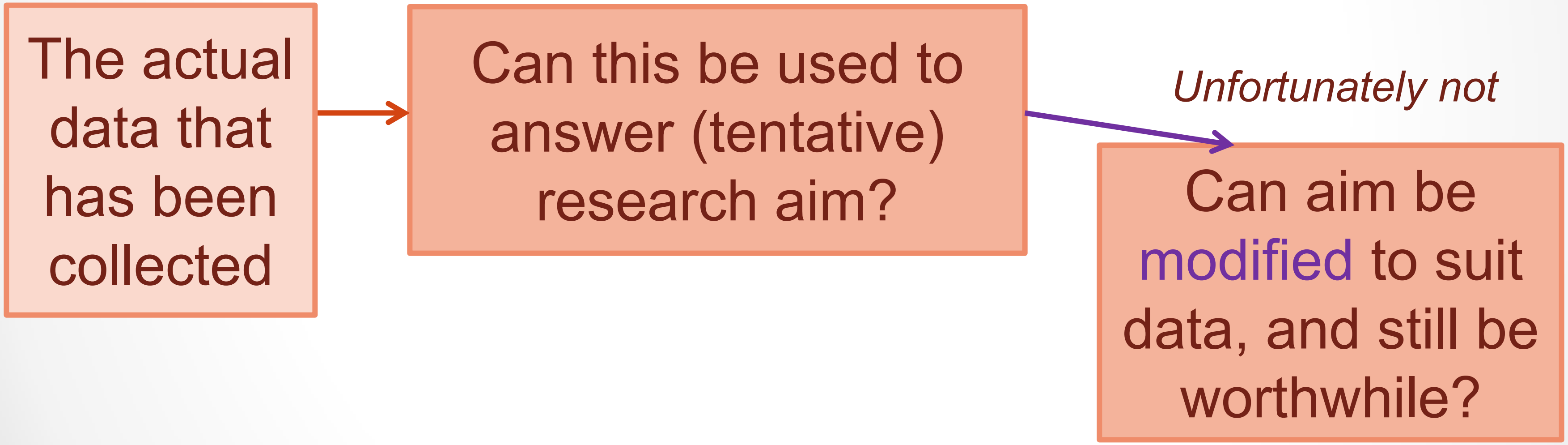


The actual data that has been collected

Can this be used to answer (tentative) research aim?

Research process in secondary analysis

- In secondary analysis the data set has been collected long before your decision to undertake research (usually!)
- So a better way to view the 'secondary analysis' research process would be:



The actual data that has been collected

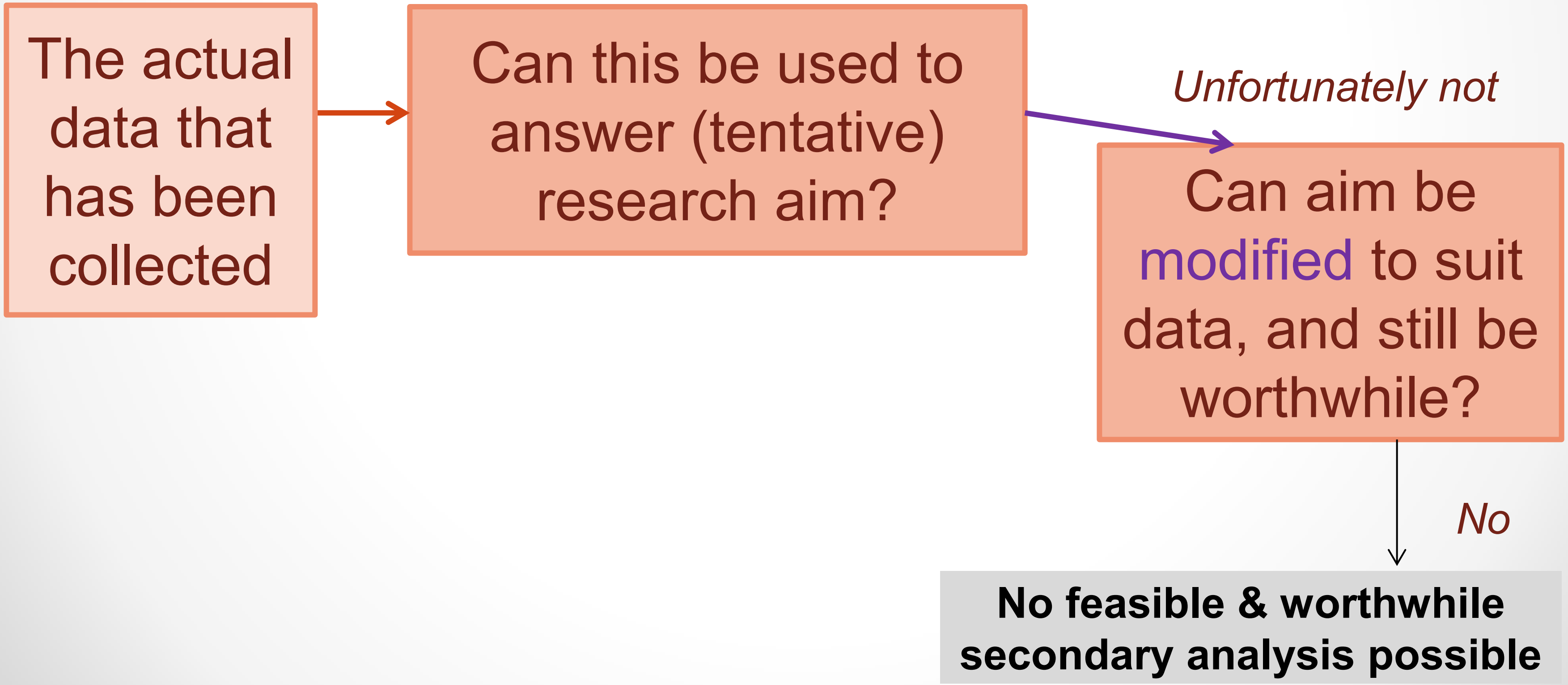
Can this be used to answer (tentative) research aim?

Unfortunately not

Can aim be modified to suit data, and still be worthwhile?

Research process in secondary analysis

- In secondary analysis the data set has been collected long before your decision to undertake research (usually!)
- So a better way to view the 'secondary analysis' research process would be:



The actual data that has been collected

Can this be used to answer (tentative) research aim?

Unfortunately not

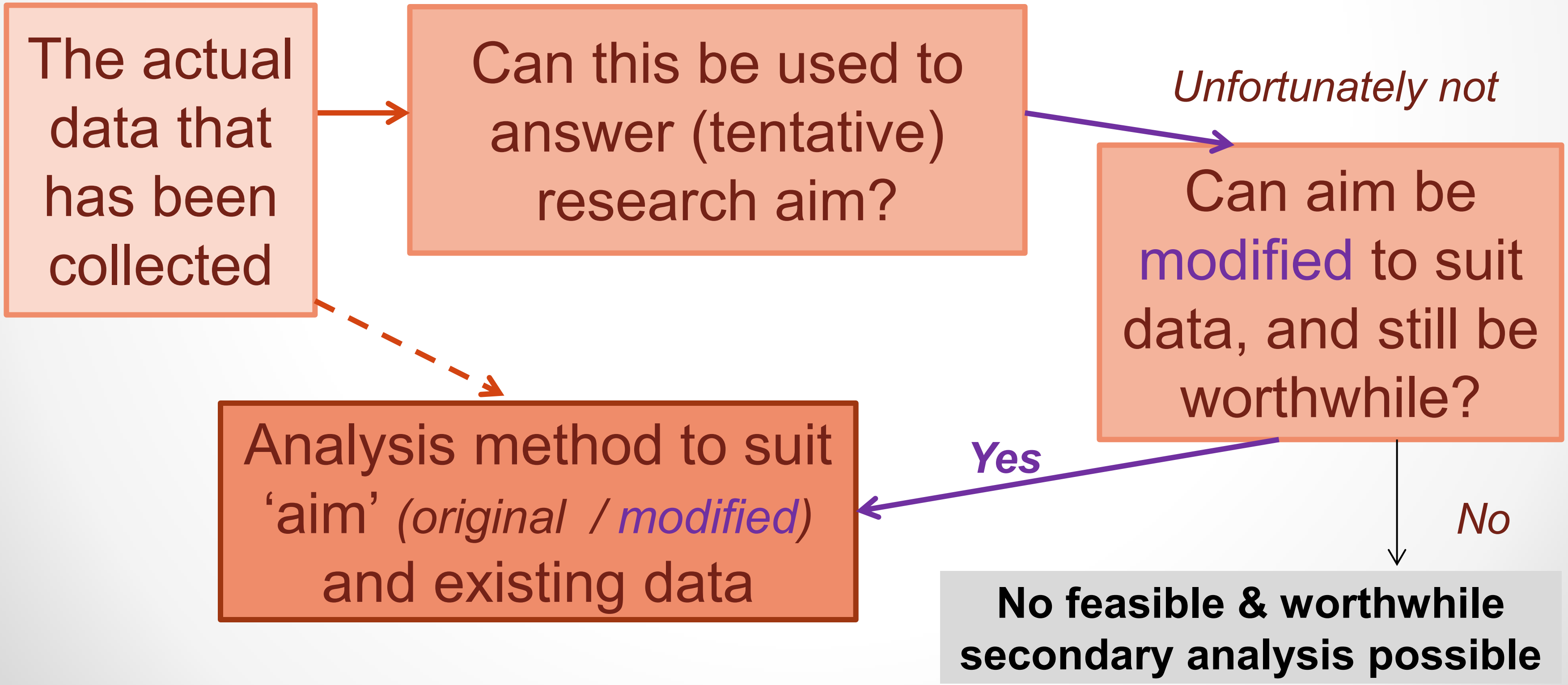
Can aim be modified to suit data, and still be worthwhile?

No

No feasible & worthwhile secondary analysis possible

Research process in secondary analysis

- In secondary analysis the data set has been collected long before your decision to undertake research (usually!)
- So a better way to view the 'secondary analysis' research process would be:



The actual data that has been collected

Can this be used to answer (tentative) research aim?

Unfortunately not

Can aim be modified to suit data, and still be worthwhile?

Yes

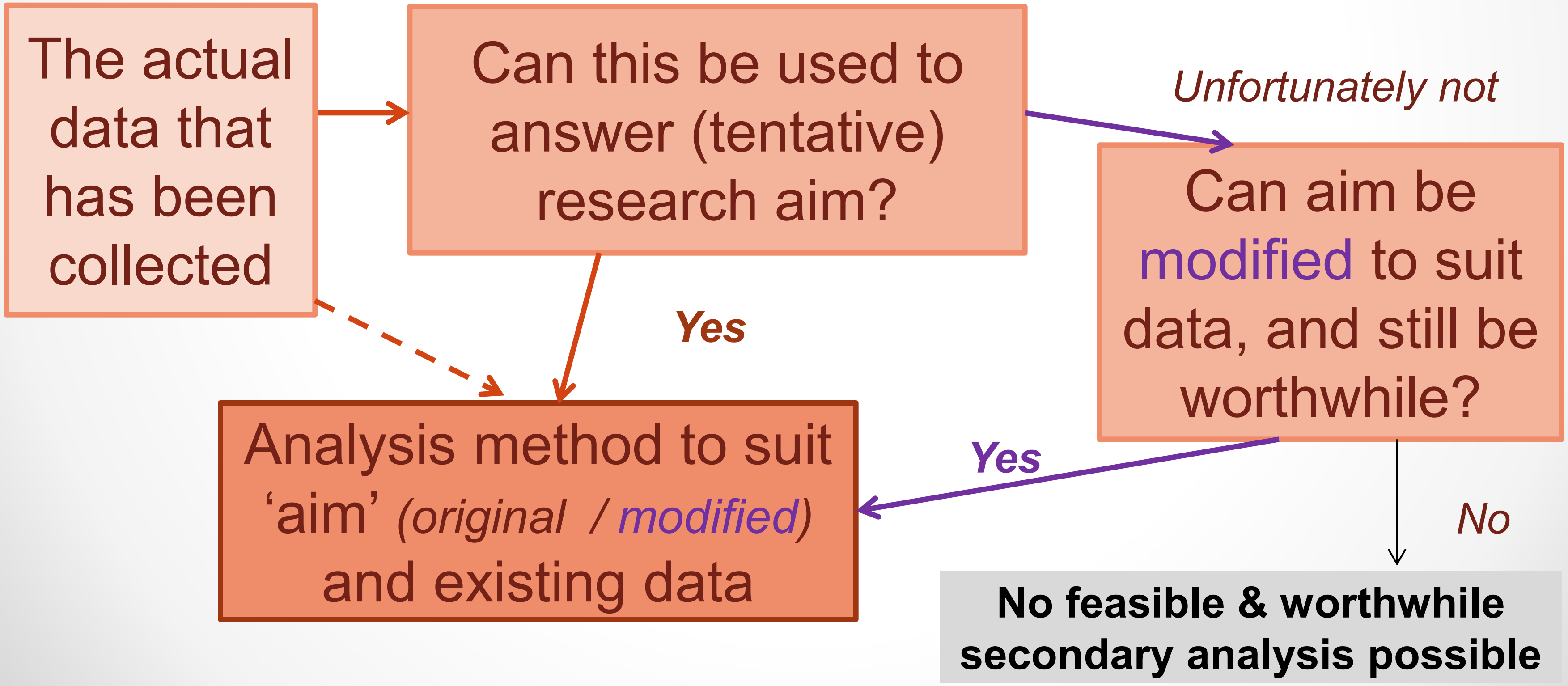
Analysis method to suit 'aim' *(original / modified)* and existing data

No

No feasible & worthwhile secondary analysis possible

Research process in secondary analysis

- In secondary analysis the data set has been collected long before your decision to undertake research (usually!)
- So a better way to view the 'secondary analysis' research process would be:



The actual data that has been collected

Can this be used to answer (tentative) research aim?

Unfortunately not

Can aim be modified to suit data, and still be worthwhile?

Yes

Yes

No

Analysis method to suit 'aim' *(original / modified)* and existing data

No feasible & worthwhile secondary analysis possible

Advantages & disadvantages of secondary analysis of GUS data

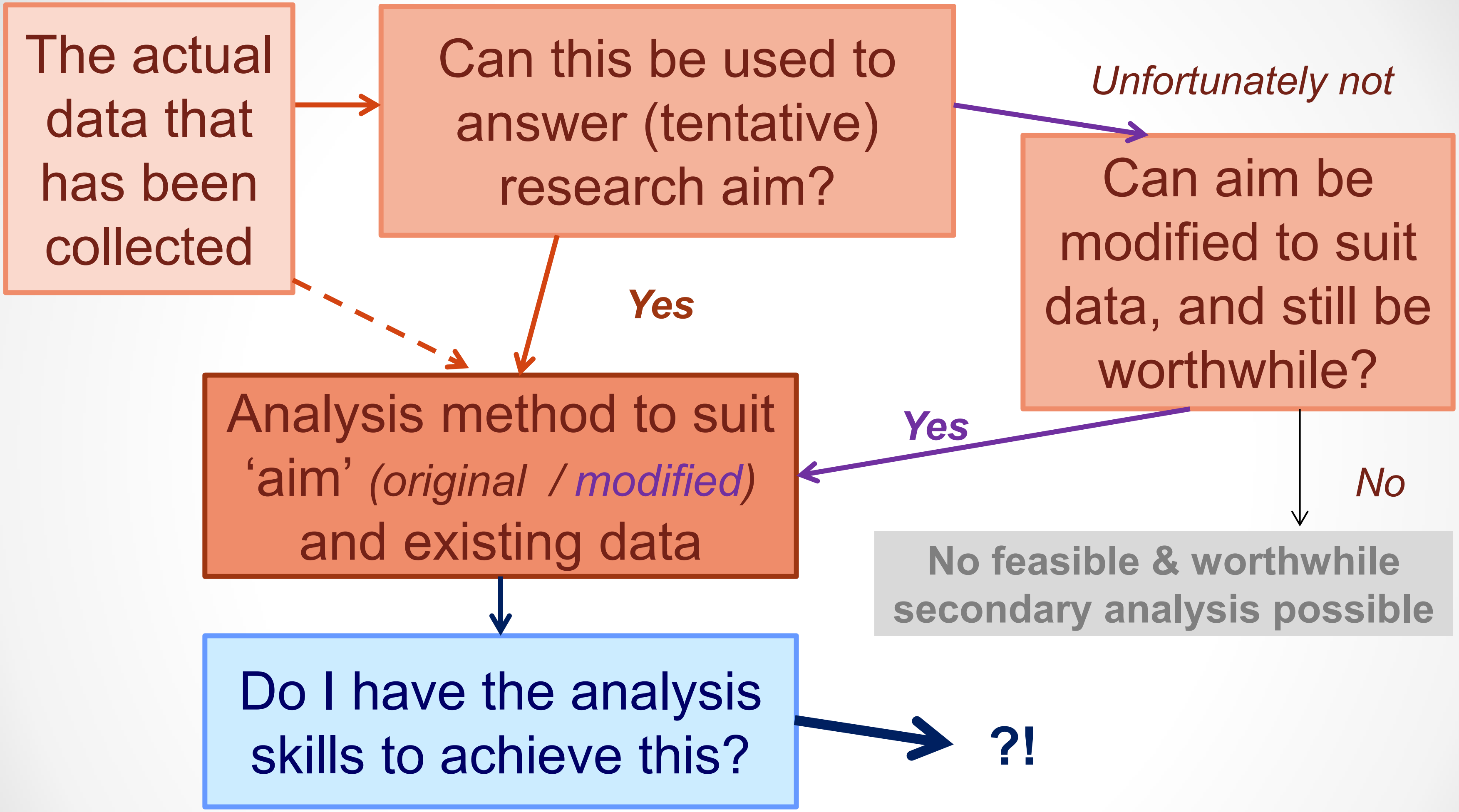
Disadvantages

- Too many unwanted variables
- Data not absolutely as needed for intended research
- Some data might be a bit out of date
- Maybe worthwhile analyses have all been done..?

Advantages

- **Well-thought-through variables** on **broad scope of topics**
- **High quality data** collection/entry all done for you
- **Representative** sample
- **Good response rate** avoids bias
- Large n provides **power even for quite complex analyses**
- **Longitudinal data, of key and enduring issues, available immediately and for free!**

*Where do **analysis skills** fit in to secondary analysis?*



The actual data that has been collected

Can this be used to answer (tentative) research aim?

Unfortunately not

Can aim be modified to suit data, and still be worthwhile?

Yes

Yes

No

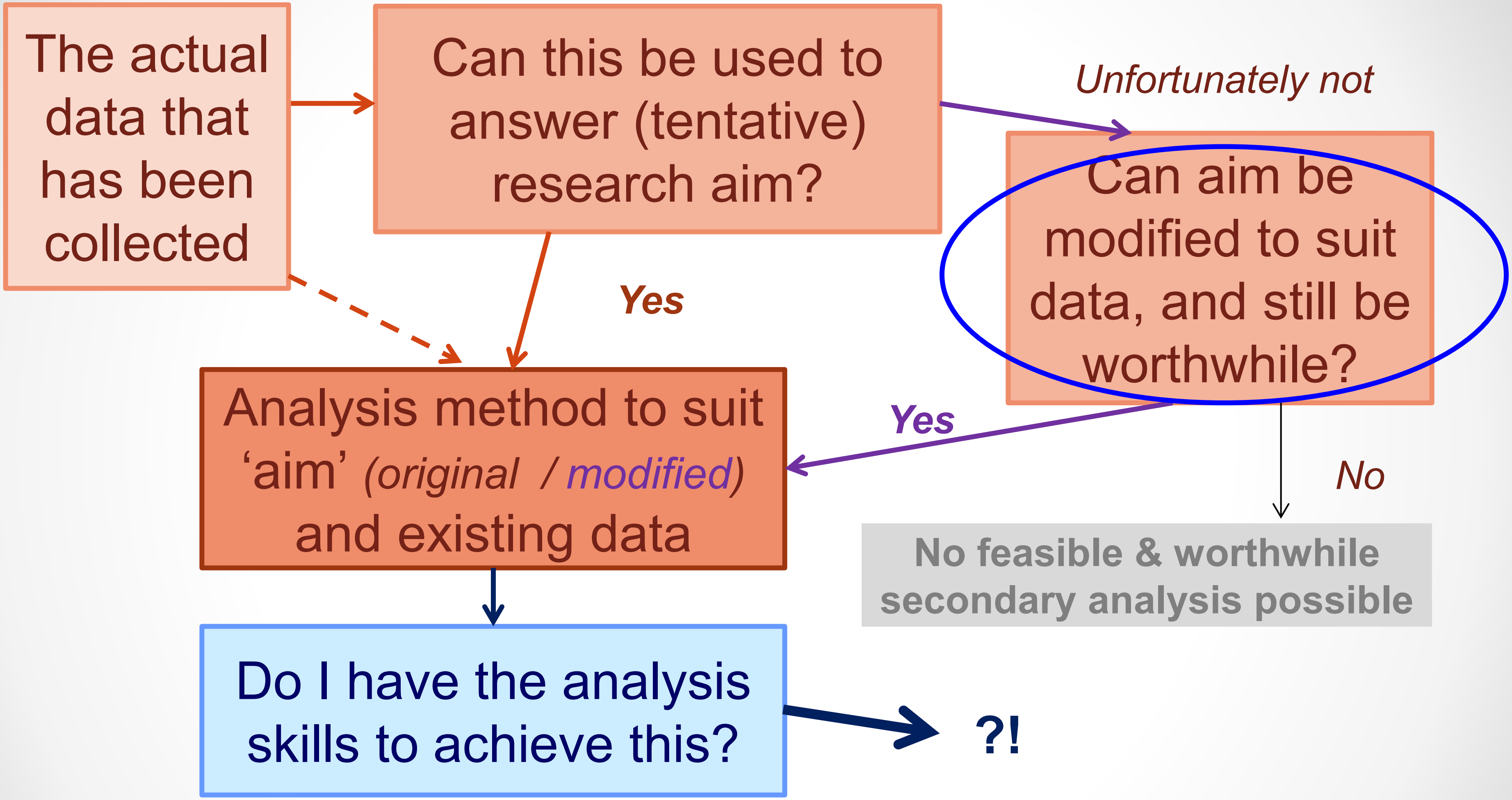
Analysis method to suit 'aim' *(original / modified)* and existing data

No feasible & worthwhile secondary analysis possible

Do I have the analysis skills to achieve this?

?!

Where do analysis skills fit in?



The actual data that has been collected

Can this be used to answer (tentative) research aim?

Unfortunately not

Can aim be modified to suit data, and still be worthwhile?

Yes

Yes

No

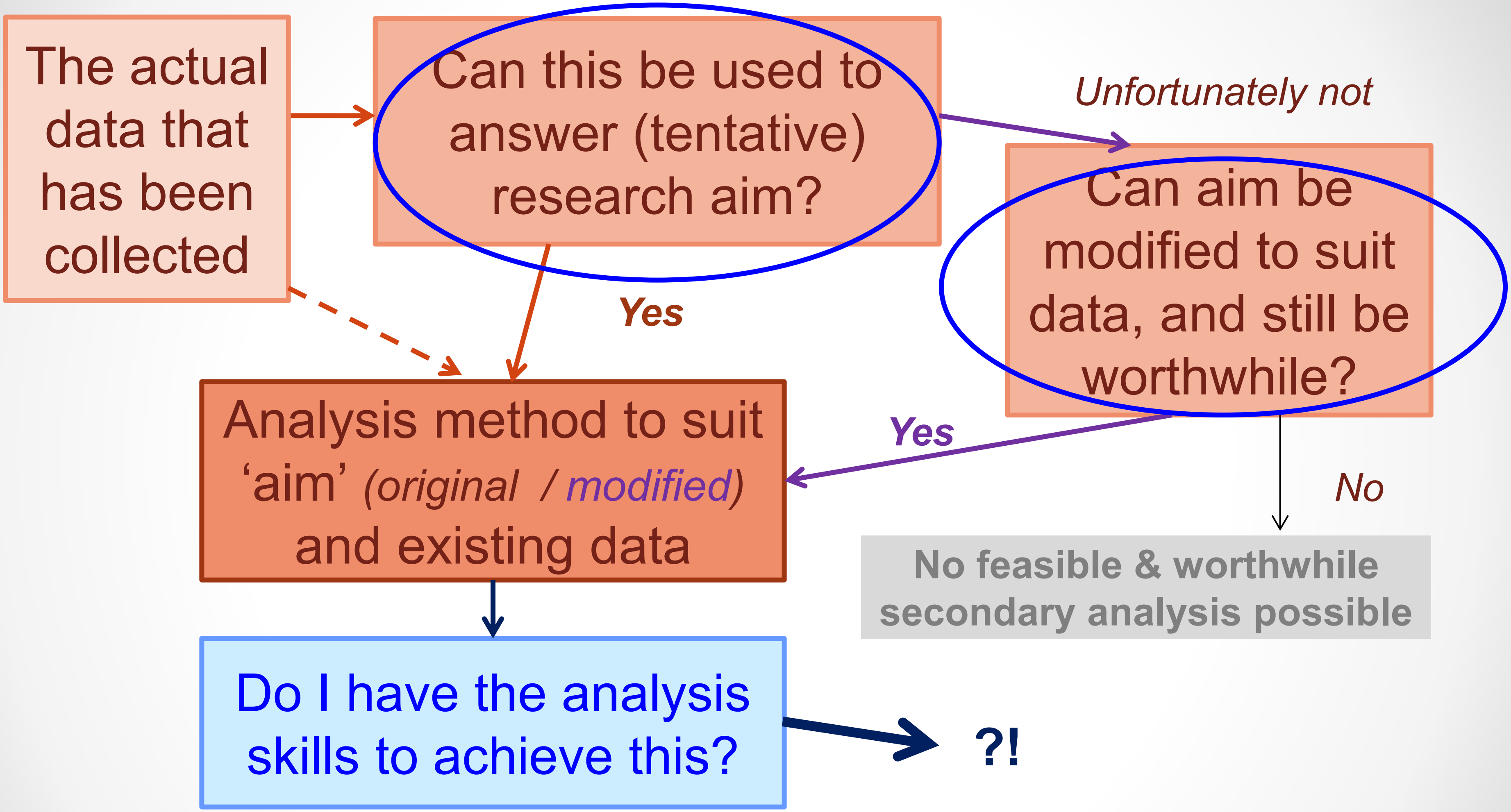
Analysis method to suit 'aim' *(original / modified)* and existing data

No feasible & worthwhile secondary analysis possible

Do I have the analysis skills to achieve this?

?!

Where do analysis skills fit in?



The actual data that has been collected

Can this be used to answer (tentative) research aim?

Unfortunately not

Can aim be modified to suit data, and still be worthwhile?

Yes

Analysis method to suit 'aim' *(original / modified)* and existing data

Yes

No

No feasible & worthwhile secondary analysis possible

Do I have the analysis skills to achieve this?

?!

- **It is essential that someone on the research team has adequate analysis skills, but expertise could be 'bought-in'**

Overview of secondary analysis

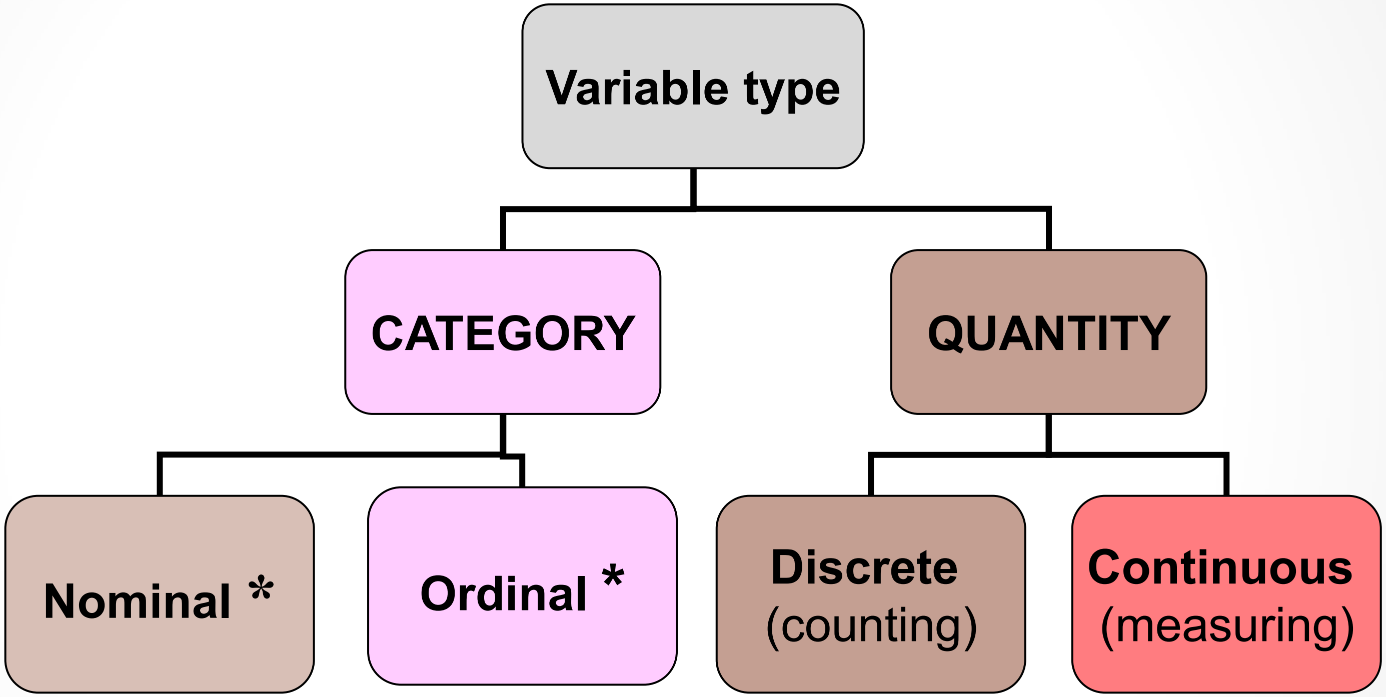
- There are huge advantages to using existing GUS data
- So it is good strategy to be a bit flexible, if possible, about your research aim, in order to ensure that some worthwhile research is possible
- Adequate analysis skills are important in 2 ways - lack of these might:
  - mean that the ‘ideal’ analysis for the GUS data and (tentative) aim is unrecognised, so that either suboptimal, or no analysis, is undertaken;
  - preclude successful execution & reporting of the analysis decided
- Therefor if expertise is to be ‘bought in’, it must be involved from design/planning through to reporting.

Thinking About Analysis Possibilities (TAAP)

Thinking about analysis possibilities

- What **type of variables** are involved?
- What is the **general thrust** of the research aim?
  - Describing the children, or testing hypotheses within the data
  - Groups - comparing or ‘discovering’
  - Associations between variables – pairs of variables, or multiple variables
- **How many variables** are involved?
- Is the intended research using data **at one point in time**, or measured at successive time points?

TAAP: Types of variables



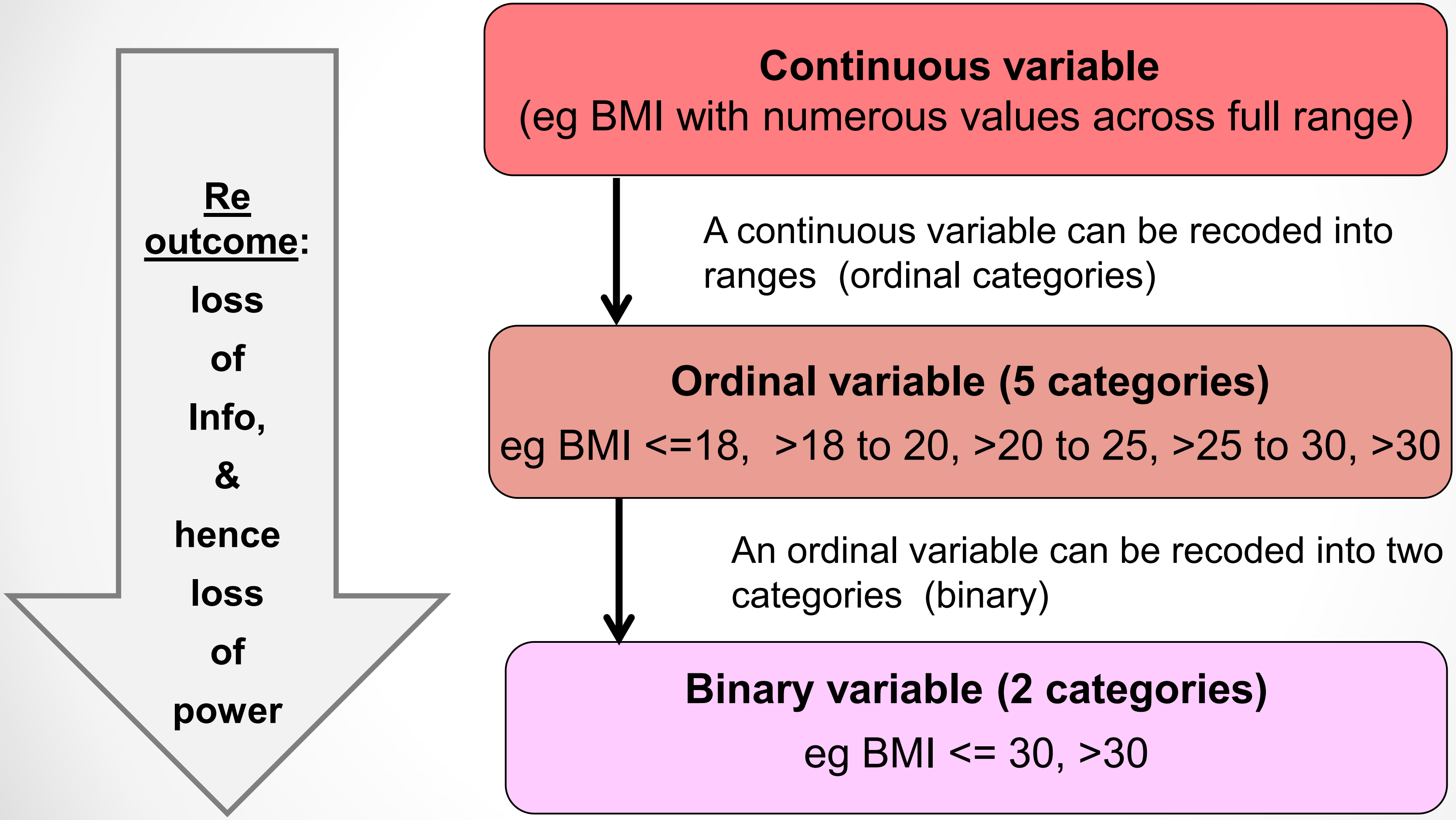
* ***Binary*** *variables are a special case with just 2 levels, so that there is no need for an ordinal/ nominal distinction*

Type of variable

This is important because it has impact on :

- **Choice of method for summary**
- **Choice of method for analysis**
- **Interpretation of results**
- **Power of analysis**

Variable ‘type’ is not necessarily fixed



Re outcome: **loss of Info, & hence loss of power**

Continuous variable
(eg BMI with numerous values across full range)

A continuous variable can be recoded into ranges (ordinal categories)

Ordinal variable (5 categories)
eg BMI <=18, >18 to 20, >20 to 25, >25 to 30, >30

An ordinal variable can be recoded into two categories (binary)

Binary variable (2 categories)
eg BMI <= 30, >30

Why recode if there is loss of info/ power?!

Because generally this simplifies analysis

- Less complex methods
- Fewer assumptions/ checks needed
- Interpretation of results is more straightforward, both for operator *and* audience

Simple as possible analysis can be a good thing. But only if no important subtlety/ power is lost because of the simplification. Inability to undertake the more appropriate (complex) analysis is not a justification!

TAAP: General thrust of research aim

- Describing the children, *or* testing hypotheses within the data

- Exploring (‘discovering’), *or* ‘confirming’ something envisaged in advance (from theory)

    - Groups:

        - Comparing (pre-defined) groups, in terms of some characteristic/ outcome

        - ‘Discovering’ groups of children similar in some respect

    - Associations between variables:

        - Estimating pre-specified associations from the data

        - Confirming whether associations, that had been theoretically-envisaged in advance, are present

General thrust - examples

- Describing, or testing
  - Describing: what are the childrens’ BMIs? – *average value & how widely spread*
  - Testing: *should we reject the Null Hypothesis that BMIs of boys are the same as girls?*
- Groups:
  - Comparing (pre-defined) groups – descriptively *(eg average BMIs for boys and girls)* or by testing *(see above)*
  - ‘Discovering’ groups of children similar in some respect – *what are 7 to 10 main types of family set up /circumstances for children at age 5*?

General thrust - examples

- Associations between variables:
  - Estimating pre-specified associations - *what is the association between exercise frequency/intensity and BMI?*
  - Confirming whether ‘theoretical’ associations are present – *is allergy by age 5 associated with formula feeding as an infant, as has been suggested in previous research?*
  - *Inter-relationships among many variables*

TAAP: How many variables?

It is almost unthinkable to undertake population/ social research with anything fewer than about 10 variables!

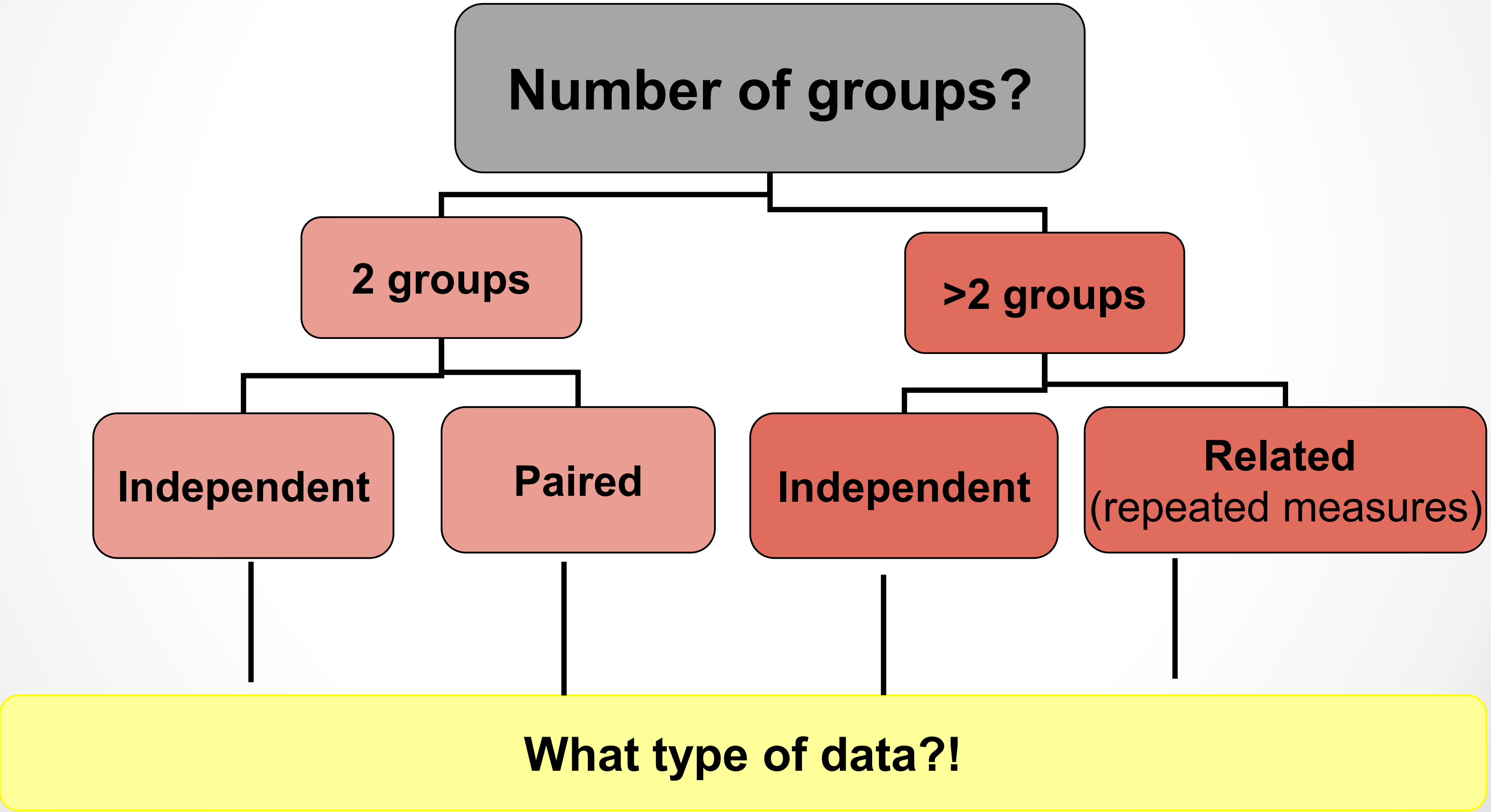
Therefore **multivariate analysis** will usually be needed.

TAAP: Single time-point or multiple

This issue relates to whether the analysis involves:

- Children in one cohort at one time-point *(cross-sectional or quasi-longitudinal)*
- Children in two different cohorts (born 6 years apart) measured at the same age – eg at age 5
- The same children re-measured across time (sweeps) – eg BMI at ages 3, 5 7 years
- More than one cohort measured at more than one point!

There is interplay between these – groups, repeats, data type: eg for Comparison



Number of groups?

2 groups

>2 groups

Independent

Paired

Independent

Related (repeated measures)

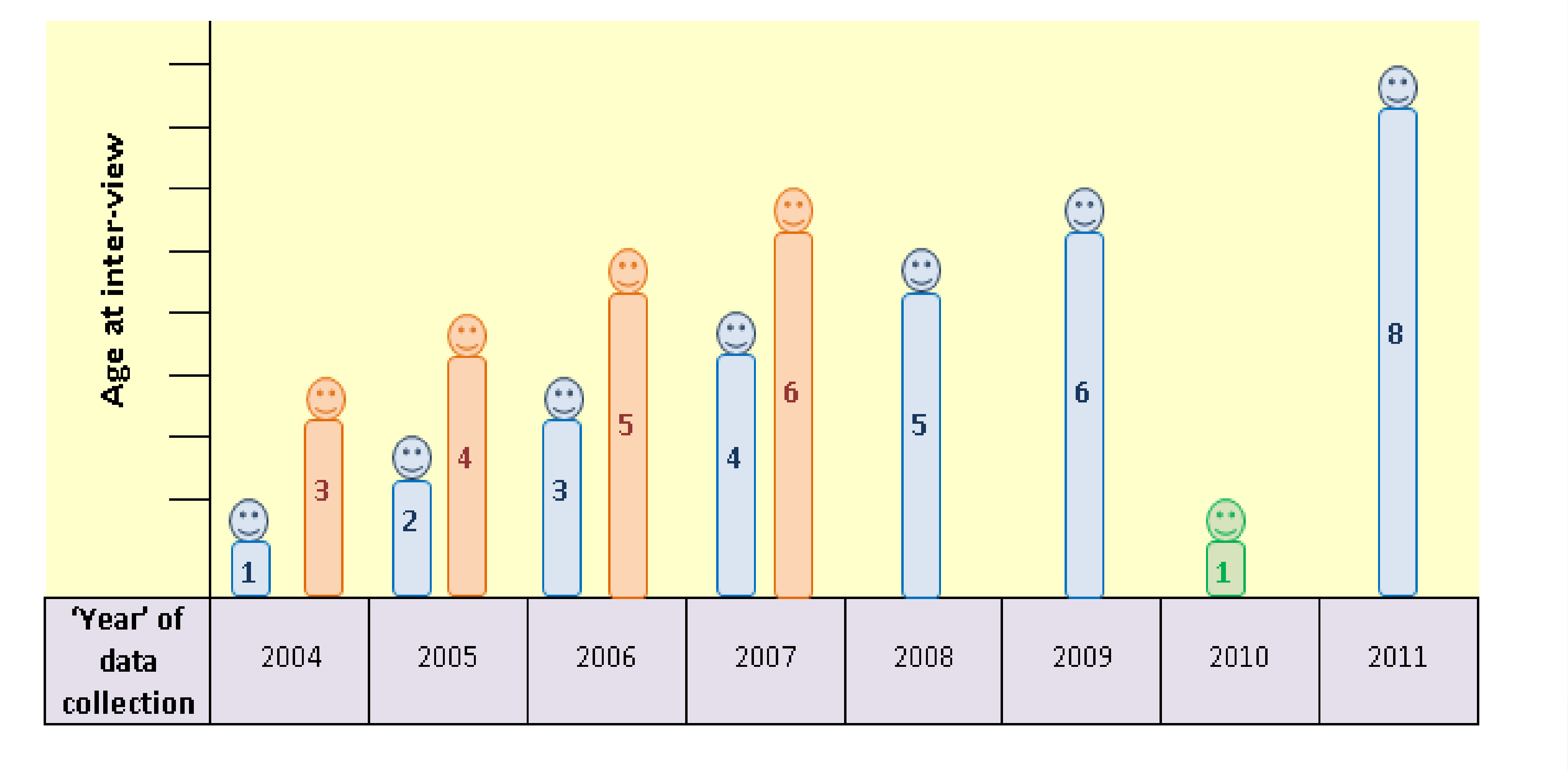
What type of data?!

What happens when we are interested in associations among multiple variables?

- We generally have to think in terms of MV modelling.
- There are many different methods, although they are often related mathematically one with another.
- Additional specific training will usually be needed, so as to:
  - Know the checks/ precautions that are needed
  - Execute, interpret and report properly
- The more variables involved, the greater the n needed, and the more caution needed

Thinking About Analysis Possibilities – some examples

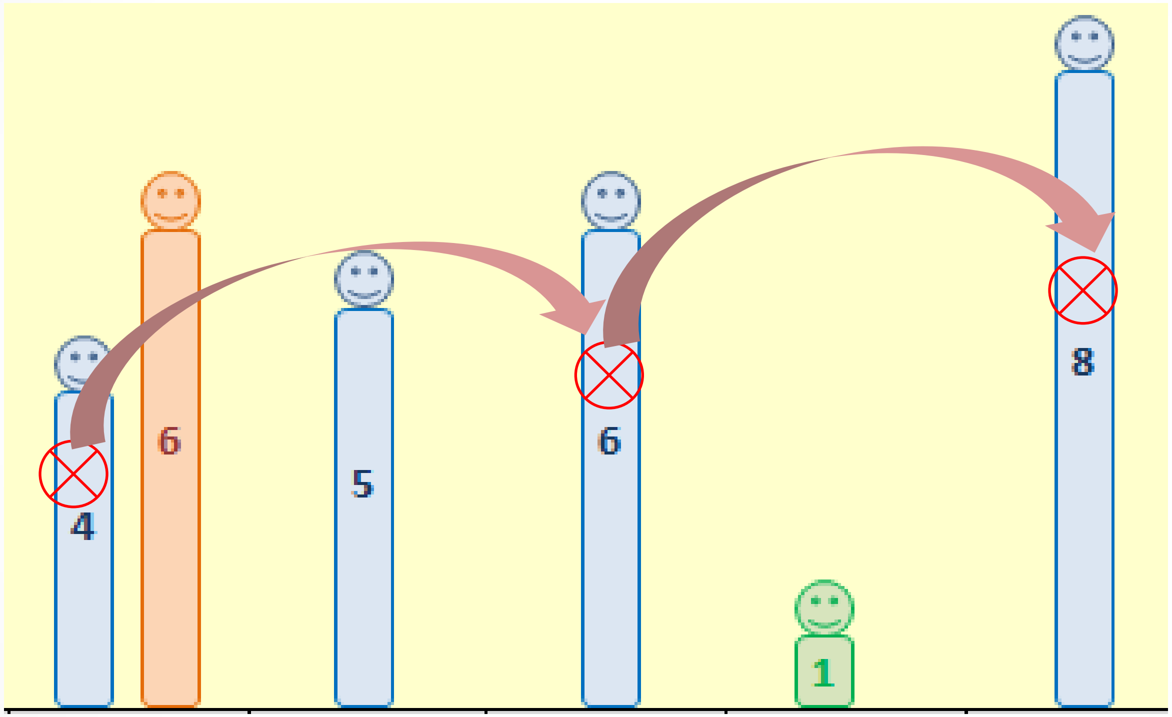
Diagrammatic representation of cohorts, calendar time and age



Note that there is a distinction between child 'time' (ie age) and calendar 'time' (ie year)

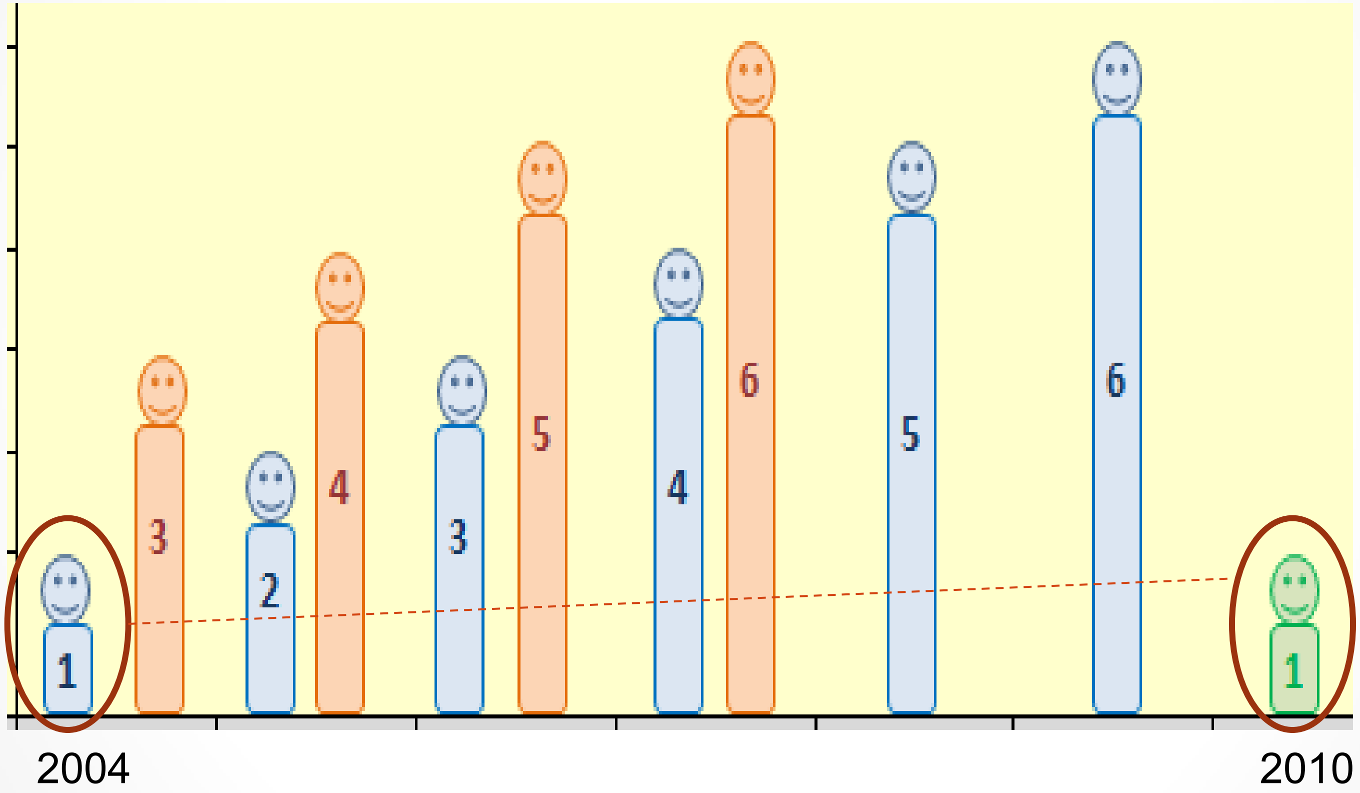
Within-cohort analysis of age effect

(longitudinal analysis)



e.g. within BC1, tracking some aspect such as height, at say ages 4, 6 and 8 years, and/ or examining how this is related to some other circumstance eg rural/urban living

Comparison between cohorts to examine calendar time (‘year’ effect)



e.g. focussing on 1-year-olds, and examining change in ‘1-yr-old’ weight between BC1 and BC2– ie between 2004 and 2010;
and/ or examining how this is relates to some other circumstance which might also have changed in the interim, such as maternal age

Logistic Regression

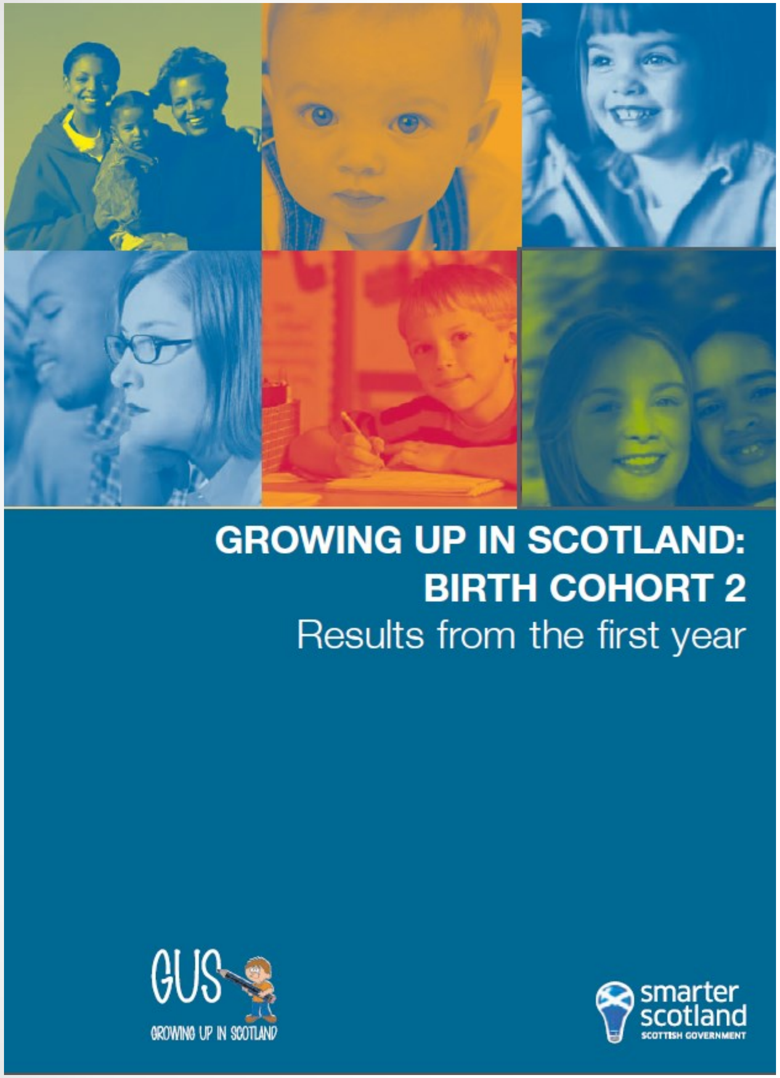
- There are relatively few continuous outcome variables in GUS data
- Perhaps, therefor, the most commonly used MV method used with GUS data is logistic regression – *which analyses counts in a corresponding huge (but unseen!) cross-tabulation*
- Logistic regression is good for understanding the association of a categorical outcome with a number of potential explanatory variables, jointly
- This enables:
  - assessment of association adjusted for all explanatory variables in the model
  - evaluation of confounding and effect modification

Binary Logistic Regression

- The most commonly used MV method with GUS data is **binary** logistic regression
- However, this often requires a multi-category outcome variable to be recoded as binary, to allow use of the method. And this which might involve loss of information…
- It would be good to see more use of ***ordinal*** logistic regression, instead of binary

Other MV methods of potential use

- Structural Equation modelling
- Latent class analysis
- Profile analysis
- Principal component analysis
- Multi-level models
- Time-to-event analysis
- Event history analysis



GROWING UP IN SCOTLAND:
BIRTH COHORT 2
Results from the first year

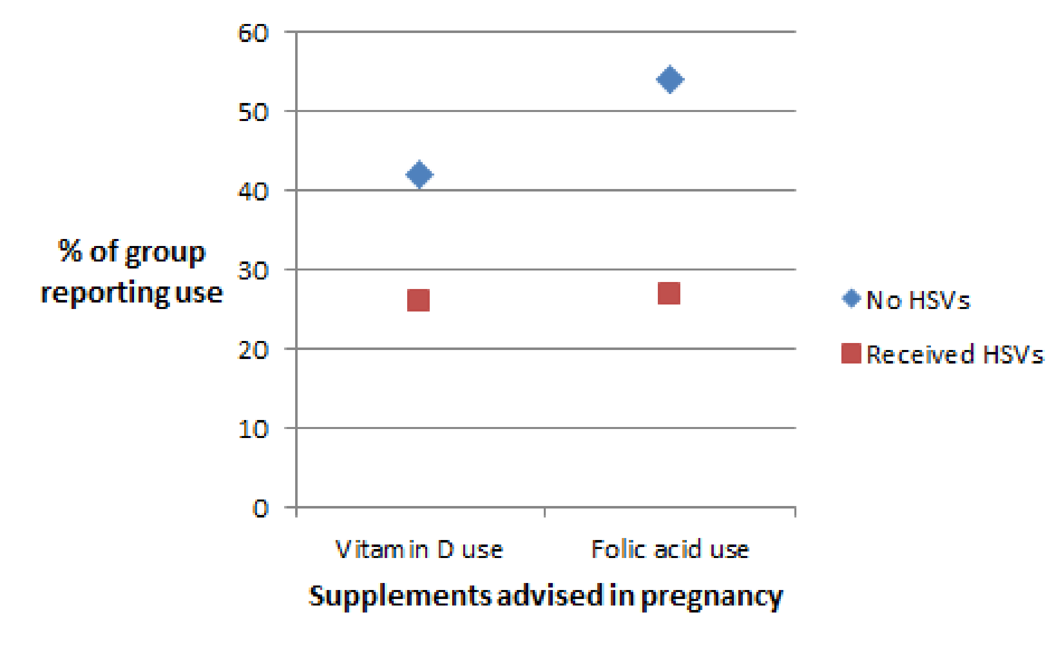
GUS
GROWING UP IN SCOTLAND

smarter
scotland
SCOTTISH GOVERNMENT

In the recent GUS report for BC2, (binary!) logistic regression was used in a number of analyses of reported behaviours during pregnancy and infant feeding.

Let’s look at recipients of **Healthy Start Vouchers**. These Vouchers were introduced to try to support very low income young women to have healthier diets and to be able to afford supplements that are considered beneficial in pregnancy – Vitamin D and folic acid.

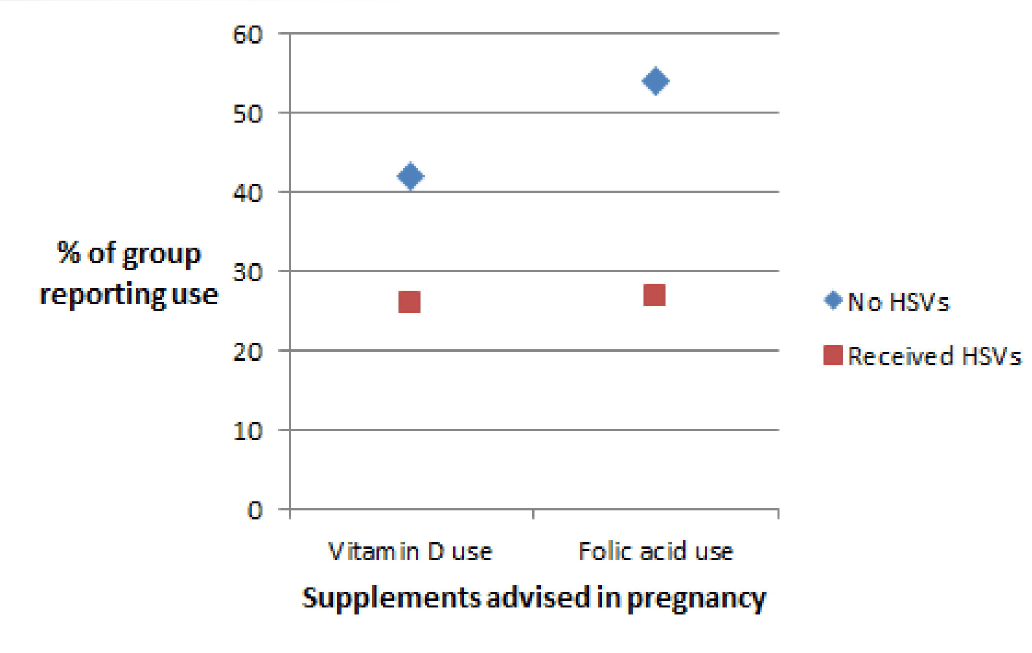
Findings for Folic Acid and Vitamin D



So simple analysis suggested that women receiving HSVs were very much ***less likely*** to use of Vitamin D or folic acid around their pregnancy.

Very disappointing findings for this policy initiative…..

Findings for Folic Acid and Vitamin D



However, women receiving HSVs are very disadvantaged, and disadvantage is generally predictive of lack of use of supplements….. So this might be an unfair comparison!

MV binary logistic regression, adjusting for socio-demographic variables, found that there was no difference between groups in use of Vitamin D *(OR close to 1)*

Concluding points

- GUS encompasses a glorious wealth of data
- The first key to success is to find a fit between a burning research question (you have) and data that has been collected
- Do not be put off because you feel you lack analytical skills
- There ***will*** be a suitable method of analysis, but you might need help:
  - to ‘discover’ it
  - in executing/ reporting it
- More complex analysis is not necessarily better - the simplest analysis ‘fit for purpose’ is best.
- The ‘quest’ to answer your research question will:
  - increase your analytical skills markedly
  - in addition, through your findings , extend knowledge about child outcomes, and factors influencing them
- *Onward!*

Bibliography

Singer JD & Willett JB (2003) Applied longitudinal data analysis – modelling change and event occurrence. Oxford University Press

Tabachnick BG, Fidell LS. (???) Using multivariate statistics. ?th edition Harper Collins Publishers

Long JD (2012) Longitudinal data analysis for the behavioural sciences, in R, Sage, London

Kleinbaum DG (1994) Logistic regression – a self-learning text. Sprimger

Kirkwood BR, Sterne JAC (2003) Essential Medical Statistics. Blackwell Science, Oxford

Moons KGM et al (2009) Prognosis and prognostic research: what, why and how? BMJ 338:1317-1320

Peng C-Y J et al (2002) An introduction to logistic regression analysis and reporting. Journal of Educational Research 6: 3 -14

Warner P (2008) Ordinal logistic regression. JFPRHC 34: 169-170