Australian National University

# Developing a Quantitative Data Analysis Plan for Observational Studies

Developed by Emily Banks, Ellie Paige and Tanya Mather for the Chronic Disease Epidemiology Group

Last updated 25 November 2013

Research School of Population Health

ANU College of Medicine, Biology & Environment

# Table of Contents

## Overview

A Data Analysis Plan (DAP) is about putting thoughts into a plan of action. Research questions are often framed broadly and need to be clarified and funnelled down into testable hypotheses and action steps. The DAP provides an opportunity for input from collaborators and provides a platform for training. Having a clear plan of action is also important for research integrity and quality; it guards against data-driven results and allows analyses to be reproduced. The DAP should build on the research protocol and is typically written after the protocol has been developed and before commencing any statistical analysis.

## Principles

A DAP provides a map of your planned analysis and developing this map can assist you to work through, step-by-step, important pieces of information without getting lost. It can help if you can visualise the outcomes of your study: what is the main picture you are trying to convey? What are the main figures/tables that illustrate your outcome? These questions should be clearly addressed in your plan.

The DAP end product is a document which is similar to a recipe; it outlines which variables you will be including in the analysis, and a step-by-step methodology for how you will approach the research questions and hypotheses. As such, it is considered best practice to have a solid DAP.

The DAP is a tool that can be used within teams; the level of detail required may differ depending on the team's needs. The exact contents of methods section may depend on your team. The DAP, and in particular the methods section, is iterative and is a living document which should be updated over time. Start with your 'best idea' of what the analysis will be, though this may change. To assist in developing an analysis plan you may need to have a brief look at the data set. This could include doing some basic frequency tables and graphs. Data cleaning also needs to be undertaken. Data cleaning procedures are not covered in this guide but are a very important part of preparing to start data analysis.

## Components

There are four main components of a DAP: background; aims; methods; and planned (dummy) tables and figures. Each research group may have different expectations of what to include or the level of detail required, but these basic components form a solid base for a DAP.

### Background

The background should present an overview of the relevant literature and a clear rationale for study. The rationale should justify your research questions and your choice of analysis. When deciding on your approach to the analysis it is important to look at what has been done in other studies.

### Aims

The study aims and research questions need to be clearly defined and translated into testable hypotheses. The hypotheses are the bridge between the ideas and the data; data will be able to confirm or refute a hypothesis.

## Methods

The methods section is the main component of the data analysis plan. The methods should include details on:

- Data sources
- Study population: include a definition and outline the inclusion/exclusion criteria
- Study measures: detail definitions and derivations (including categorisation used, if any) of study measures including:
    - Main exposure variables
    - Outcome variables
    - Other covariates, including potential confounders and mediators
- Sub-groups: you may wish to examine if the main effect varies by sub-groups of participants.
- Missing data: Include details about methods used for dealing with missing data (complete case analysis, coding missing values as separate categories, imputation methods and/or sensitivity analyses)
- Sensitivity analyses: detail any sensitivity analyses to be undertaken.
- Sequence of planned analyses including: statistical methods; how hypotheses will be tested; and how potential confounding and bias will be assessed and addressed. The sequence often includes:
    - Outline of main comparison groups
    - Frequency and cross-tabulations of main variables
    - Basic analysis model (usually age- and sex-adjusted)
    - Final analysis model (including adjustment for other confounders)

  Statistical methods ideally include planned model building approach, methods used to verify statistical assumptions, alternate methods to be adopted in case of violations of assumption and choice significance level. Clinical significance levels would need to be pre-specified, if relevant.

- Analysis software: outline the software and version number you will be using for the analysis.

## Planned tables and figures

Planned tables and figures (also called dummy tables) are basically an outline of a table or figure which will used to present the result. The dummy table has empty cells which are to be populated after the data analysis. The planned tables and figures bring into focus of what you are doing and how you will display your results. Planned tables and figures can also be a useful talking point for discussing the analyses with collaborators and allow for refinement of your research intentions. Further, the planned tables and figures can be copied directly into the results section of a paper or chapter and the cells populated after analysis.

# Example

## Screen time and obesity

1. **Background**
   - increasing sedentary time, especially screen-time
   - obesity a mounting public health problem
   - increasing obesity with increasing screen-time already observed
   - focus on episodic physical activity but suggestion that total activity more important
   - lack of evidence regarding screen-time/obesity relationship in different population subgroups

2. **Aims**
   - to quantify the relationship of screen-time and other sedentary behaviours to obesity in a large cohort study of older Australian adults overall and within a range of population subgroups. Particular attention will be paid to how observed relationships vary with age, disability, work status and lifestyle factors.

   **Hypotheses to be tested**
   - that individuals with reporting greater daily screen-time will be significantly more likely to be obese than those with less daily screen-time
   - that this relationship will not vary significantly according to age, sex, income, education, region, physical activity levels, work status, disability, smoking, alcohol, fruit/vegies
   - particularly interested in how much of screen-time/obesity relationship explained by physical activity and variation according to level of physical activity.

*While it is ok to include multiple sub-groups, you should have a rationale for doing so, and should report all results, not just significant ones*

3. **Methods**

   3.1 Data source: 45 and Up Study baseline questionnaire

   3.2 Study population
   - definition: Participants in the 45 and Up Study
   - inclusion/exclusion criteria: All participants in the 45 and Up Study, excluding those with missing data on height, weight and physical activity

   3.3 Study measures
   - exposure variables: screen-time, sitting time, standing time, physical activity
   - outcome variables: obesity (BMI ≥30kgm$^{-2}$)
   - covariates/potential confounding factors: many including age, sex, income, education, and physical activity
   - subgroups to be considered: many
   - definitions and derivations
   - how missing data will be dealt with: exclusion
   - comparison groups: level of physical activity. Work status etc.

   3.4 Data cleaning

   3.5 Sequence of planned analyses, including
   - cross-tabulation of relationship of screen-time to potential confounding factors
   - cross-tabulation of relationship of primary exposures to obesity
   - RR of obesity (generalised linear models) according to screen-time, sitting time, standing time, sleep time and physical activity
     - age and sex adjusted

- age, sex, income, education adjusted
- age, sex, income, education, physical activity adjusted

- RR of obesity according to
  - level of physical activity
  - work status
  - age, sex, income, education, region, physical activity levels, work status, disability, smoking, alcohol, fruit/vegies
  - test for statistical interaction (likelihood ratio test, weighted least-squares)

3.6 Analysis software: SAS v9.13

## 4. Dummy Tables and Figures

The below tables 1 and 2 were used to display analysis results. Note: dummy tables only show the row and column headings. The cells should be left blank. Figures 1 and 2 are examples of graphs used to present the results. For a DAP, these figures can be roughly
drawn without the inclusion of any specific axis values. In this example, dummy tables and figures were the main way of talking with the biostatistician.

**Table 1** Characteristics of the study population according to total daily screen-time

| | Hours of screen-time per day | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0–1·9 h | | 2·0–3·9 h | | 4·0–5·9 h | | 6·0–7·9 h | | >8 h | | |
| | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | *P* for trend |
| Male | | | | | | | | | | | |
| Urban resident | | | | | | | | | | | |
| Tertiary educated | | | | | | | | | | | |
| Annual income ≥$AU 70 000 | | | | | | | | | | | |
| In full-time paid work | | | | | | | | | | | |
| Current smoker | | | | | | | | | | | |
| Disabled | | | | | | | | | | | |
| Functional capacity (in lower third) | | | | | | | | | | | |
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | |
| Age (years) | | | | | | | | | | | |
| BMI (kg/m$^2$) | | | | | | | | | | | |
| Alcohol consumption (g/week) | | | | | | | | | | | |
| Vegetable intake (servings/d) | | | | | | | | | | | |
| Fruit intake (servings/d) | | | | | | | | | | | |
| Physical activity (sessions/week) | | | | | | | | | | | |
| Sitting (h/d) | | | | | | | | | | | |
| Standing (h/d) | | | | | | | | | | | |
| Sleeping (h/d) | | | | | | | | | | | |

**Table 2** Relative risk of obesity according to sedentary behaviours and physical activity

| | Total | Obese | Age- and sex-adjusted | | Adjusted | | Adjusted | |
|---|---|---|---|---|---|---|---|---|
| Total | *n* | % | RR | 95 % CI | RR | 95 % CI* | RR | 95 % CI† |
| Screen-time (h/d) | | | | | | | | |
| 0–1 | | | | | | | | |
| 2–3 | | | | | | | | |
| 4–5 | | | | | | | | |
| 6–7 | | | | | | | | |
| ≥ 8 | | | | | | | | |
| *P* for trend | | | | | | | | |
| Time spent sitting (h/d) | | | | | | | | |
| 0–1 | | | | | | | | |
| 2–3 | | | | | | | | |
| 4–5 | | | | | | | | |
| 6–7 | | | | | | | | |
| ≥ 8 | | | | | | | | |
| *P* for trend | | | | | | | | |
| Time spent standing (h/d) | | | | | | | | |
| 0–1 | | | | | | | | |
| 2–3 | | | | | | | | |
| 4–5 | | | | | | | | |
| 6–7 | | | | | | | | |
| ≥ 8 | | | | | | | | |
| *P* for trend | | | | | | | | |
| Time spent sleeping (h/d) | | | | | | | | |
| 0–5 h | | | | | | | | |
| 6–7 h | | | | | | | | |
| 8 h | | | | | | | | |
| 9–10 h | | | | | | | | |
| ≥ 11 h | | | | | | | | |
| *P* for trend | | | | | | | | |
| Sessions of physical activity per week | | | | | | | | |
| 0–3 sessions | | | | | | | | |
| 4–6 sessions | | | | | | | | |
| 7–11 sessions | | | | | | | | |
| 11–17 sessions | | | | | | | | |
| ≥ 18 sessions | | | | | | | | |
| *P* for trend | | | | | | | | |

*Adjusted for age, sex, income and education.
†Adjusted for age, sex, income, education and overall physical activity.

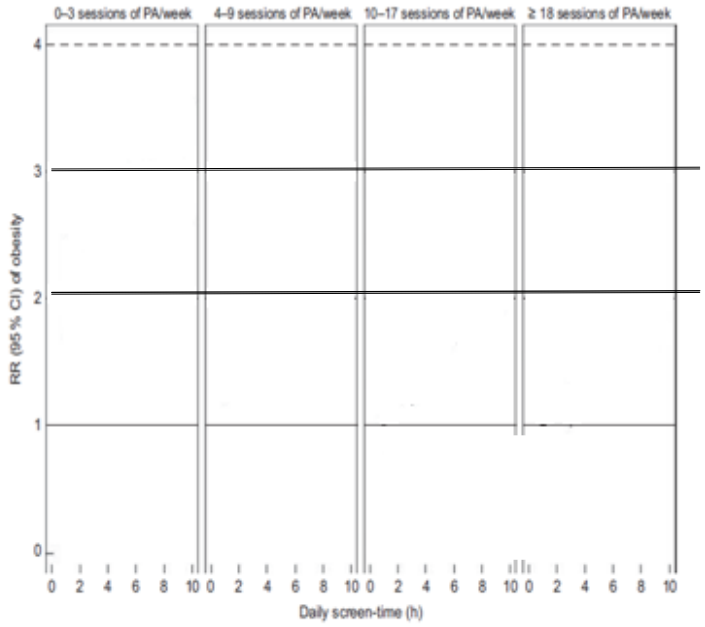
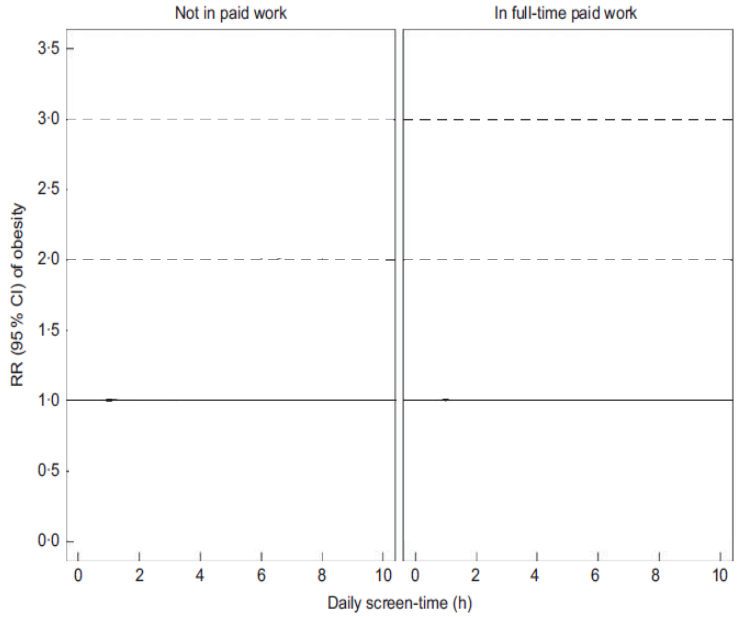



**Fig. 1** Relative risk (RR) of obesity according to hours of daily screen-time, in categories of overall physical activity (PA), ac
for age, sex, income and education (reference group: ≥18 sessions of PA per week, <2 h of daily screen-time); RR are
against the median daily screen-time for each category

*Put on a pair of 'statistical glasses' when looking at results. i.e. when looking at curves, take into account the confidence intervals rather than just concentrating on the point estimates.*




*Confidence intervals are very important and should be included in tables/graphs*

**Fig. 2** Relative risk (RR) of obesity according to hours of daily screen-time in study participants who are not in paid work and those in full-time paid work, adjusted for age, sex, income and education (reference group: not in paid work, <2 h of daily screen-time); RR are plotted against the median daily screen-time for each category

## Frequently Asked Questions

**Q:** A plan is boring. Why can't I just get on with my analysis?

**A:** Although it main seem laborious doing a DAP is actually time efficient. By planning out your analysis you can more quickly undertake the actual data cleaning and analysis and clearly answer your research questions. Different sections of text in the DAP can also be used to form the basis of papers or thesis chapters. A DAP also helps to sustain collaborator/supervisor relationships by avoiding mistakes and disagreements.

**Q:** If it is so important, why don't all researchers do plans like this?

**A:** Groups do vary in expectations and more experienced researchers may cut corners but many usually do an outline of a DAP. Even if DAPs are not the norm in your research group, you will never be criticised for being too professional and thorough; there is little to be lost by writing a DAP. DAPs are particularly helpful for inexperienced researchers.

The degree of detail in the DAP may also depend on extent of background literature. If not much is known about your topic then a broader analysis looking at many variables may be more useful and may not include specific hypotheses to be tested.

**Q:** Why can't I just 'play' with the data?

**A:** Our brains see patterns easily, so a DAP helps guide interpretation and avoids a data-driven approach. If necessary, incorporate "exploratory" analyses into your plan, and be explicit about what is speculative.

**Q:** My analysis is standard. It's obvious what should be done so why should I do a DAP?

**A:** Some analyses are more straightforward than others and the requirements of different groups will vary. However, be aware of assumptions about shared assumptions – collaborators may differ on even simple ideas. If your analysis does turn out to be very straightforward, the DAP will be very quick to develop. There is little to be lost and plenty to be gained by doing a DAP.

*Horses for courses: data are fantastic and will answer a lot of questions but you need to ensure that the questions being asked are suitable for the data*

## General Advice

- Orient your work in terms of what you ultimately want to produce
- Figures and tables are a great way to think and communicate
- Good working relationships are critical to collaboration and communicating well and professionally is at the heart of these
- You can never be too clear or explicit about research
- Ask lots of questions
- Need to make sure the DAP is going where you want to as well as where your collaborators want it to
- The STROBE statement provides a checklist for the reporting of case-control, cohort and cross-sectional studies. These guidelines are useful when publishing results and should be read at the planning stage to ensure all necessary steps are undertaken during data analysis. The statement and checklists can be found at: http://www.strobe-statement.org/
  A similar statement for the reporting of RCTs is available (http://www.consort-statement.org/)

## Acknowledgments

This guide is based on a presentation given by Professor Emily Banks for a short course in the Master of Philosophy (Applied Epidemiology) at the Australian National University.

*"Love all, trust few, always paddle your own canoe"*

## Example Data Analysis Template

**Note:** This data analysis template is a modified version of the template created by the **Master of Philosophy (Applied Epidemiology) teaching team** at the Australian National University.

| **DATA ANALYSIS PLAN TEMPLATE** | | | |
|---|---|---|---|
| Reference No. | | Study name | |
| Date of plan | | Chief investigator | |
| Person conducting analysis | | Telephone | |
| | | Mobile | |
| | | Email | |
| Analysis team members | | | |

| **Background to the study and analysis (Please use plain language)** | | | |
|---|---|---|---|
| Provide an overview of the necessary background for the study including evidence of what is already known in the area of study and what the gaps are in the literature. Finish with a clear stated aim of the project. | | | |
| Number study participants | | Duration of study | |
| Study research question | | | |
| Specific hypothesis under study | | | |
| Endpoints or outcomes of interest | | | |

| **Data details (Please complete all that apply)** | |
|---|---|
| Study type | |
| Data sets used | |
| Analysis package | |
| Study population | |
| Inclusion/exclusion criteria for participants | |
| Exposure variables | |
| Outcome measures | |
| Covariates | |
| Sub-groups | |

| | |
|---|---|
| Approach to dealing with missing data | |

| **Please outline proposed analytical strategy** |
|---|
| Include:<br>– Outline of main comparison groups<br>– Frequency and cross-tabulations of main variables<br>– Basic analysis model (usually age- and sex-adjusted)<br>– Final analysis model (including adjustment for other confounders) |

| **Analysis dissemination strategy** |
|---|
| Outline the intended steps to be taken to disseminate the results of the study (i.e. will the results be published, presented at a conference etc) |

| **Interpretation** |
|---|
| Detail how you will interpret the results in the context of your stated hypothesis. I.e. if the results do/do not meet your hypothesis, what will you conclude? A concept map (see below) may assist with this. |

## Concept map or directed acyclic graph

Drawing a diagram of the ways in which the exposure might be related to the outcome will help to visualise your hypotheses as well as serving as a basis for clearly communicating this to your collaborators. The diagram should include the possible confounders or mediators of the relationship. This will require good knowledge of the background to the study.

Directed acyclic graphs are a type of causal graph. Further information about these graphs can be found through a Google search and in the paper “Causal Diagrams for Epidemiologic Research” by Sander Greenland, Judea Pearl, and James M. Robins (Epidemiology 1999;10:37-4)

## Dummy tables & Charts

Dummy tables and charts are empty skeleton tables and charts which show how the results will be presented but which do not contain any data/results.