Analysis plan template for life-course cohort studies

Preamble

This template has been designed in conjunction with the MCRI's LifeCourse initiative both to strengthen the quality of life-course observational cohort studies and to aid in the efficient development of research manuscripts by addressing several key principles together, including:

- Explicit articulation of specific research questions, with recognition of their type from the beginning and of the implications of the type of question for analysis planning
- Planning in advance with the whole team, particularly to reduce the number of post hoc re-analyses
- Adopting best practice in the analysis, interpretation and reporting of observational studies

It is strongly advised that all co-authors review your analysis plan before you undertake analyses.

Proposed paper information

Provide as far as possible, no problem if not known or definitive.

Working title:

Working author list:

Target journal(s):

Rationale

Background

Very briefly, describe the problem you are studying, its significance and the state of the literature.

Broad aim(s) of the paper:

Very briefly, describe the gap in the literature that your research will address and its potential translational impact (what difference will it make?). Reflect on the expected strengths and limitations of your study.

Research question(s):

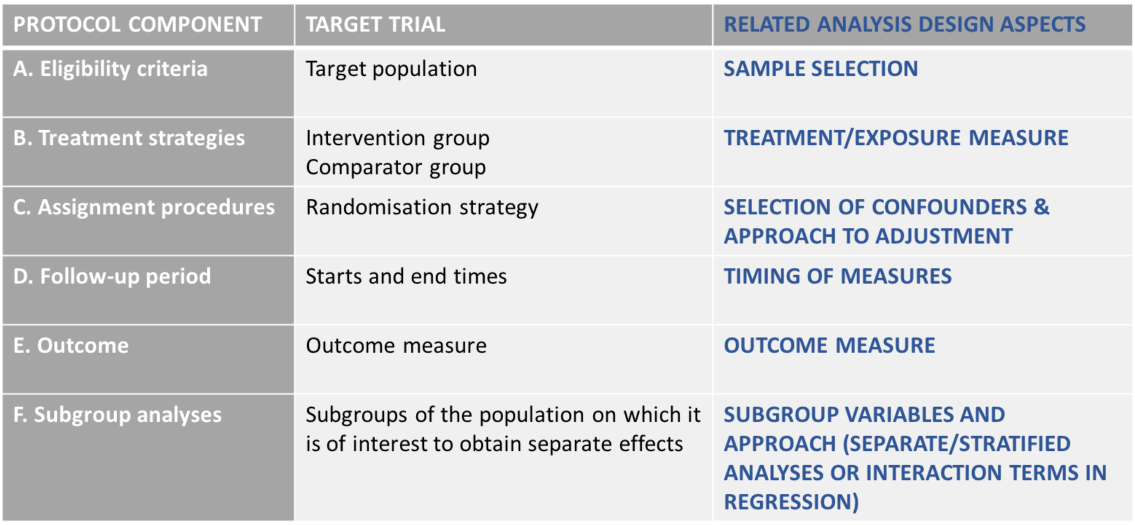
Specific research question	**Type of research question** *(Select one only for each question. See examples below.)*		
	Descriptive	**Causal**	**Prediction**
1.	□	□	□
2.	□	□	□
3.	□	□	□
4.	□	□	□

Examples of different types of question:

- *Descriptive: Describe trends over time in the prevalence of a disease in different subgroups. Examine the strength of correlation between a disease and each of multiple possible predictors (or "risk factors") separately. Cluster analyses. Describe class structures or trajectories. Data reduction.*
- *Causal: Estimate the causal effect of an exposure, treatment or intervention on a disease, where "effect" is understood to mean the effect that would be estimated in a hypothetical randomised controlled trial in which the exposure/treatment/intervention was allocated and compared to a clearly defined control condition. Any question aiming to inform future interventions, even if down the track and just hypothetical at this stage, falls under this category.*
- *Prediction: Build a prediction model for disease prognosis when a new patient arrives and has been examined and given a number of tests.*

Analysis plan

1. [For causal questions] Reflect on the target trial that you are aiming to emulate, in particular each of the protocol components as per below (see Hernan & Robins here). This will be key in guiding decisions on analysis design in the following points.



PROTOCOL COMPONENT	TARGET TRIAL	RELATED ANALYSIS DESIGN ASPECTS
A. Eligibility criteria	Target population	SAMPLE SELECTION
B. Treatment strategies	Intervention group Comparator group	TREATMENT/EXPOSURE MEASURE
C. Assignment procedures	Randomisation strategy	SELECTION OF CONFOUNDERS & APPROACH TO ADJUSTMENT
D. Follow-up period	Starts and end times	TIMING OF MEASURES
E. Outcome	Outcome measure	OUTCOME MEASURE
F. Subgroup analyses	Subgroups of the population on which it is of interest to obtain separate effects	SUBGROUP VARIABLES AND APPROACH (SEPARATE/STRATIFIED ANALYSES OR INTERACTION TERMS IN REGRESSION)

2. [For causal questions] Outline the broad conceptual model for your observed data using a causal diagram (i.e. a directed acyclic graph, or DAG), indicating the exposure and outcome. Use this to inform the selection of the sample in point 3 below (to avoid collider bias) and the selection of confounding variables that need to be adjusted for based on prior evidence or knowledge, i.e. what variables can be used to block open backdoor paths between exposure and outcome.

   Note: Variables affected by the exposure must not be included in the adjustment set. If there is uncertainty as to whether a variable must be adjusted for, it is a good idea to plan on reporting the results of a set of models that are progressively adjusted by obvious confounders or a specific class of confounders (e.g. demographics) and then adding less obvious ones or other classes (e.g. environmental factors). If you are considering repeated measures of the exposure or outcome (i.e. time-varying exposure or outcome), or if you are considering a mediation analysis, you might need to discuss your DAG and plan further with a biostatistician with expertise in observational study methodology.

3. How will you select the analysis sample(s) and are there any pre-specified stratification variables (e.g. sex) and what is the justification? (for causal questions, refer back to points 1 and 2). Indicate the approximate sample size(s).

4. Are you pooling data from multiple cohorts and, if so, what is the theoretical justification for doing this (similar source populations etc.) and temporal (age/period) alignment?

   Note: For a multi-cohort pooled-data analysis, it will generally be a good idea to conduct cohort-specific analyses as a secondary analysis (see points 6 and 7 below).

5. List the specific (possibly derived) variables to be used in the analysis, how they were measured/derived, at what age/wave, their role (e.g. exposure/outcome/confounders for causal questions) and their type (continuous, categorical), along with descriptive statistics of these if possible including details about missing data.

   Note 1: For a multi-cohort analysis, indicate whether these variables are measured differently across datasets and how you will harmonise them.

   Note 2: For continuous variables, you should look at a histogram and check for symmetry. If it is not symmetrical you should describe the data using the median and interquartile range, and you might need to consider a transformation for the analyses, particularly if it is the outcome.

6. Describe the statistical analyses to be undertaken for each question and the rationale for the choice of methods, including planned strategy to deal with missing data (e.g. multiple imputation) if this is an issue.

   Note: For a multi-cohort analysis, "cohort" must be included as an adjustment variable in causal analyses, and it will also generally be a good idea to examine interactions by cohort and/or cohort-specific analyses as a secondary analyses (see points 3 and 7).

7. Indicate the secondary analyses to be undertaken, for example common secondary analyses include considering:

   (a) alternative measures of the exposure or outcome (e.g. different reporters or instruments, or binary/continuous versions, or measures at other waves/ages), in particular to examine potential impact of measurement error

   (b) other strategies to select the samples (e.g. analysis separately by cohort in multi-cohort analyses) or to handle missing data (available/complete case analysis, sensitivity analysis to missing not at random), in particular to examine potential impact of selection bias

   (c) analysis examining the potential impact of unmeasured confounding in causal questions

8. Describe the planned table and figure structure for the paper (main text and appendix).

   Note 1: In reporting your results, please avoid the misuse of p-values (see here). We recommend that you avoid basing conclusions on dichotomous interpretation (significant/non-significant) of p-values.

   Note 2: For causal questions,

   - *Table 1 would generally be as in a trial, describing the characteristics of the two exposure/treatment groups*
   - *Table 2 would generally present causal effect estimates. Beware of the "Table 2 Fallacy" (see here for an explanation and recommendations on how to avoid it).*

   Note 3: For descriptive questions,

   - *A variety of forms of reporting may be relevant, but in particular we encourage the use of graphics instead of large tables. This can help to reduce the focus on multiple p-values and encourage the description of overall patterns.*
   - *Importantly, there is unlikely to be a sensible role for multiple regression analysis in answering descriptive questions.*