**Statistical analysis plan template for observational studies**

**Preamble**

This template has been designed both to strengthen the quality of observational 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.
* Advance planning 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 and agree upon the analysis plan before undertaking the analyses.

**Proposed paper**

**Key information**

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

Working title:

Working author list:

Target journal(s):

**Background/Rationale**

*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?).*

**Research question(s)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Specific research question** | **Type of research question**  *(Select one only for each question.*  *See examples below.)* | | |
| **Descriptive** | **Predictive** | **Causal** |
| **1.** | **🞏** | **🞏** | **🞏** |
| **2.** | **🞏** | **🞏** | **🞏** |
| **3.** | **🞏** | **🞏** | **🞏** |
| **4.** | **🞏** | **🞏** | **🞏** |

*Examples of different types of question:*

* *Descriptive (aims to characterise the distribution of a feature or outcome across a population): E.g., Describe trends over time in the prevalence of a disease. Describe disease risk across different subgroups or as a function of a given factor (e.g., age). Describe clusters of participants with higher disease risk. Describe individual trajectories of disease.*
* *Predictive (aims to produce predictions of an outcome based on measured factors/predictors):* *E.g., Build a prediction model for disease prognosis for patients based on their measured characteristics.*
* *Causal (“what” if questions, aiming to estimate the causal effect of an exposure, treatment or intervention on an outcome, 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 compared to a clearly defined control condition): E.g., Any question aiming to inform what to intervene on or change in order to change/improve outcomes, even if down the track and just hypothetical at this stage, falls under this category.*

**Data sources**

*List and briefly describe the data source(s) to be used in your study, including an approximate sample size if known. If you are considering combining data from multiple cohort studies, provide theoretical justification for doing this. For multi-cohort studies, see O’Connor et al.* [*here*](https://www.sciencedirect.com/science/article/pii/S1040260822000399)*, for a general overview and helpful guidance.*

**Statistical analysis plan**

*In the next pages are three separate sections for guiding the planning of analyses to address descriptive, predictive, and causal research questions. Complete the section(s) that is relevant to your research question(s).*

**Statistical analysis plan for descriptive questions**

The table below provides a framework for guiding the planning of descriptive analyses, with step-by-step instructions below the table.

|  |  |  |
| --- | --- | --- |
| **TARGET ESTIMAND** | **DATA DECISIONS** | **POTENTIAL REMAINING BIASES** |
| **Target population:** | **Analytic sample selection:**  **Approach to handling missing data and other potential sources of selection bias:** |  |
| **Outcome:** | **Outcome measure (including timeframe):** |  |
| **Subgroup analyses (if relevant):** | **Subgrouping variable(s) measure(s):** |  |
| **Summary measure:** |

1. Use the “Target estimand” column of the table above to define key aspects of your descriptive research question, including:
   * Target population (or eligibility criteria): Define the target population, e.g., who is in, who is excluded, where, when.
   * Outcome: Define the outcome of interest, including timeframe.
   * Subgroup analyses (if relevant): Specify any subgroup analyses (i.e., stratified analyses) that are of interest.
2. Define the summary measure of the outcome distribution that is of interest. For continuous outcomes, this is often a mean or, for skewed outcomes, a median. For binary/categorical outcomes, it can be a prevalence or cumulative incidence (i.e., proportion of cases occurring over a fixed time interval amongst an at-risk population).
3. Use the “Data decisions” column of the table above to specify data decisions to answer your research question, including:
   * Analytic sample selection and approach to handling missing data and other potential sources of selection bias: Define the analytic sample and provide an indication of the sample size available. Describe how missing data (due to e.g., outcome non-response and/or loss to follow-up) and other potential sources of selection bias (due to e.g., non-consent) will be handled/adjusted for. Examples of methods are multiple imputation, standardisation or inverse probability weighting (IPW), considering covariates that are predictors of selection into the sample, missingness or incomplete variables.
   * Outcome measure (including timeframe): Describe how the outcome is measured in the sample.
   * Subgrouping variable measure(s) (if relevant): Describe how the subgrouping variable(s) are measured in the sample.
4. Describe the estimation method you will use, in the context of the approaches described in step 3 to handle potential selection biases. In the absence of the latter, it may be possible to estimate the summary measure using the sample equivalent, e.g., for continuous outcomes, the sample mean/median to estimate the population mean/median, or, for binary outcomes, the sample proportion to estimate the population prevalence.
5. Complete the final “Potential remaining biases” column of the table above, identifying any potential remaining sources of bias that cannot be completely mitigated via the planned analysis approach, e.g., selection bias due to unmeasured causes of non-participation or missing data, or measurement bias arising as a result of measurement error, which reflects a discrepancy between the measured value of a quantity and its true value.
6. Indicate any sensitivity analyses to be undertaken, for example, common sensitivity analyses to consider include:
   * Alternative measures of the outcome (e.g., different cut-offs for categorisation or alternative measurement scales), in particular to examine the potential impact of measurement error.
   * Other strategies to handle missing data or other sources of selection bias, such as available/complete case analysis.
7. 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*](https://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503)*). We recommend that you avoid basing conclusions on dichotomous interpretation (significant/non-significant) of p-values.*

*Note 2:* *A variety of forms of reporting may be relevant, but 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.*

*Note 3:* *Table 1 would generally describe analytic sample characteristics to assess if reflective of target population.*

*Note 4:* *Table 2 would generally present descriptive estimates of interest, e.g., mean, median, prevalence, with 95% confidence intervals. It may be of interest to include p-values but consider their relevance, e.g., for a single measure, these are testing whether the measure equals 0, which is not usually relevant.*

*Note 5: Importantly, there is unlikely to be a sensible role for multivariable regression analysis in answering descriptive questions.*

**Statistical analysis plan for predictive questions**

The table below provides a helpful framework for guiding the planning of prediction analyses, with step-by-step instructions following the table.

|  |  |
| --- | --- |
| **KEY ASPECTS** | **DATA DECISIONS** |
| **Target population:** | **Analytic sample selection:** |
| **Outcome:** | **Outcome measure:** |
| **Predictors:** | **Predictor measures:** |

1. Use the “Key aspects” column of the table above to define key aspects of your predictive research question, including:
   * Target population (or eligibility criteria): Define the population in which the predictions are intended to be used, e.g., who is in, who is excluded, where, when.
   * Outcome: Define the outcome you are trying to predict.
   * Predictors: Define the predictors of interest.
2. Use the “Data decisions” column of the table above to specify data decisions to answer your research question, including:
   1. Analytic sample selection: Define the analytic sample that will be used, considering whether it is reflective of the target population in which the predictions are intended to be used. Provide an indication of the sample size available (see Riley *et al.* [here](https://www.bmj.com/content/368/bmj.m441) for guidance on the sample size required for developing a prediction model).
   2. Outcome measure: Describe how the outcome to be predicted is measured in the sample, considering whether it is measured using a valid, reliable method for all individuals.
   3. Predictor measures: Describe how each of the predictors are measured, considering whether each predictor is measured in an easily reproducible way.
3. Briefly appraise existing prediction models in your area of research and provide justification of the need for the development of a new prediction model as opposed to the validation of an existing model with your data.
4. Describe the methods to be used for model development including variable selection and functional form selection, and the approach to model fitting.

*Note 1: For variable selection, there is no single optimal strategy. General guidance is to include variables with reasonable prior knowledge/evidence, omit highly correlated variables and avoid unnecessary dichotomisation of continuous variables.*

*Note 2:* *In an approach that is not data-adaptive (e.g., multivariable regression), the functional form of the model/algorithm is explicitly and manually pre-specified by the analyst. For continuous outcomes, it can be helpful to explore graphically using a “locally weighted scatterplot smoothing” (lowess) plot. When considering non-linear transformations and interactions, it is recommended to start with simple structures and extend to more complex functions. In a data-adaptive (a/k/a “machine learning”) approach (e.g., random forest) the full functional form of the model/algorithm is not fixed and adapts/learns from the data.*

1. Describe the internal validation strategy to evaluate model performance and iteratively update the prediction model as required. Model performance aspects include:
   * Discrimination: in the context of binary outcomes, it measures how well the predictions from the model distinguish between those who experience the outcome and those who do not. The concept extends to continuous outcomes.
   * Calibration: measures how well the predicted distribution matches the observed outcome distribution.
   * Overall performance: measures the distance between the predicted outcomes and actual outcomes, which is a combination of discrimination and calibration.

*Note: Using the same data for model development and validation can lead to over-optimistic performance measures so it is important to use cross-validation or bootstrap resampling methods for internal validation.*

1. Describe, if relevant, any plans for external validation to assess how well the prediction model can generalise to new data in new contexts (e.g., temporal, geographical, different patient groups/clinical conditions).

*Note: Model performance measures are the same as those used for internal validation.*

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

*Note 1: The transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement (see Collins et al.* [*here*](https://www.acpjournals.org/doi/10.7326/m14-0697#:~:text=The%20TRIPOD%20Statement%20aims%20to,TRIPOD%20explanation%20and%20elaboration%20document.)*) provides guidance on key information to report when developing or validating prediction models.*

*Note 2:* *Always present the full model equation (including the intercept) to enable independent external validation.*

*Note 3:* *Prediction models can be presented in different formats to aid in implementation in clinical settings (see Bonnet et al.* [*here*](https://pubmed.ncbi.nlm.nih.gov/30995987/)*).*

**Statistical analysis plan for causal questions**

The “target trial” approach is a powerful framework for guiding the planning and interpretation of causal analyses in observational data (see Hernán & Robins [here](https://academic.oup.com/aje/article/183/8/758/1739860)). The table below will help you to work through this process, with step-by-step instructions following the table.

|  |  |  |  |
| --- | --- | --- | --- |
| **PROTOCOL COMPONENT** | **TARGET TRIAL** | **EMULATION** | **POTENTIAL**  **REMAINING BIASES** |
| 1. **Eligibility**   **criteria** | **Target population:** | **Analytic sample selection:**  **Approach to handling missing data and other potential sources of selection bias:** |  |
| 1. **Treatment strategies** | **Treatment arms in the trial, i.e., intervention group, comparator group:** | **Treatment/Exposure measure:** |  |
| 1. **Assignment procedures** | **Randomisation strategy:** | **Selection of confounders:**  **Approach to confounding adjustment:** |  |
| 1. **Follow-up period** | **Start and end times:** | **Timing of measures:** |  |
| 1. **Outcome** | **Outcome:** | **Outcome measure:** |  |
| 1. **Subgroup analyses** | **Subgroups of the population of which it is of interest to obtain separate effects (if relevant):** | **Subgrouping variable(s) measure(s) and approach:** |  |
| 1. **Causal contrasts of interest and causal effect measure** |  |

1. Complete the “Target trial” column (protocol components A-F) of the table above. This defines the key protocol components of the ideal randomised experiment that would hypothetically be conducted to address the causal research question of interest. (see Hernán & Robins [here](https://academic.oup.com/aje/article/183/8/758/1739860) for further explanation and an example).

*Note: 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.*

1. Define the causal contrasts of interest (what treatment arms are being contrasted) and the causal effect measure (in what scale will the effect be defined) - target trial protocol component G in the table above. For continuous outcomes, the causal effect measure is generally a difference in mean potential outcomes between levels of the exposure/treatment/intervention variable or, for skewed variables, it could be a difference in medians. For binary outcomes, it can be a risk difference, risk ratio or odds ratio.
2. Outline the broad conceptual model for your observed data using a causal diagram (i.e., a directed acyclic graph, or DAG), indicating expected relationships amongst the variables, including exposure, outcome, confounders and indicators of selection into the study or analytic sample. This should be informed by prior evidence or knowledge and an understanding of the study’s sampling design and recruitment process. A DAG can help identify potential sources of bias in particular:
   * Selection bias: arises when the sample used for analysis is not representative of the target population due to, for example, individuals with certain characteristics being more likely to not participate or be lost from the study over time.
   * Confounding bias: arises from differences between exposure groups in terms of pre-exposure characteristics that are also related to the outcome.

1. Complete the “Emulation” column of the table above, detailing the strategy for emulating the target trial as closely as possible with the available data.

* Analytic sample selection and approach to handling missing data and other potential sources of selection bias: Define the analytic sample and provide an indication of the sample size available (see Hernán [here](https://pubmed.ncbi.nlm.nih.gov/34461211/) for a discussion on power calculations for observational studies). Describe how missing data (due to e.g., item non-response and/or loss to follow-up) and other potential sources of selection bias will be handled, e.g., via multiple imputation or inverse probability weighting (IPW), considering the predictors of selection into the sample, missingness and/or incomplete variables as identified in the DAG.
* Treatment/Exposure measure: Describe how the exposure/treatment/intervention is measured in the sample, including the measurement tool used, whether it is derived from multiple variables, at what age/wave, and cut-off scores for categorisation if applicable.
* Selection of confounders and approach to confounding adjustment: Include a sufficiently rich confounder set to minimise sources of confounding bias identified in the DAG. Describe how confounders are measured (as for treatment/exposure measure). Variables affected by the exposure (i.e., on the path from the exposure to outcome) must not be included in the adjustment set. Approaches to confounding adjustment include conditioning-based methods such as multivariable outcome regression, and standardisation-based methods such as g-computation and IPW using propensity scores.
* Timing of measures: Define at what time/age/wave follow-up time starts and stops for all participants, considering the timing of measurements and potential variability in this across participants.
* Outcome measure: Describe how the outcome is measured in the sample (as for treatment/exposure measure).
* Subgrouping variable(s) measure(s) and approach (if relevant): Describe how the subgroup variable(s) (if relevant) is measured (as for treatment/exposure measure). Subgroup analyses approaches include separate/stratified analyses or specification of interaction terms in regression.

*Note: For multi-cohort studies, complete a separate Emulation column for each contributing cohort and consider how to harmonise variables that are measured differently across contributing cohorts (see Downes et al,* [*here*](https://arxiv.org/abs/2206.11117) *for further guidance).*

1. Provide a brief summary of the planned statistical analysis approach outlined in the emulation strategy.
2. Complete the final “Potential remaining biases” column of the table above by considering the proposed emulation strategy in relation to the target trial, and identifying any potential remaining sources of bias that cannot be completely mitigated via the planned analysis approach, e.g., selection bias due to unmeasured causes of non-participation or loss to follow-up/missing data, confounding bias due to unmeasured confounding, or measurement bias arising as a result of measurement error, which reflects a discrepancy between the measured value of a quantity and its true value.
3. Indicate any sensitivity analyses to be undertaken, for example common sensitivity analyses to consider include:
   * + Alternative measures of the exposure or outcome (e.g., different reporters or instruments, or binary/continuous versions, or measures at other waves/ages), to examine the potential impact of measurement error.
     + 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), to examine the potential impact of selection bias.
     + A progressive adjustment approach to examine the potential impact of confounder selection, i.e., report 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); or analyses examining the potential impact of unmeasured confounding.
4. 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*](https://www.nature.com/news/statisticians-issue-warning-over-misuse-of-p-values-1.19503)*). We recommend that you avoid basing conclusions on dichotomous interpretation (significant/non-significant) of p-values.*

*Note 2:* *Table 1 would generally be as in a trial, describing the characteristics of the two exposure/treatment/intervention groups, with details on missing data.*

*Note 3:* *Table 2 would generally present causal effect estimates. Beware of the “Table 2 Fallacy” (see* [*here*](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3626058/) *for an explanation and recommendations on how to avoid it).*