**SAPI Expanded Checklist**

**Statistical Analysis Plan with Initial data analysis for observational studies**

|  |  |  |
| --- | --- | --- |
| **Section/Topic** | **Item** | **Checklist item** |
| **ADMINISTRATIVE INFORMATION** | | |
| Project title | 1.1 | **State the title of the analysis project** |
| Project documents | 1.2 | **Provide links to any documents (e.g. protocol) describing the analysis project for this SAPI, if available**   * For example, links to the research protocol or data management plan including version numbers |
| Ethics approval | 1.3 | **Provide details of any ethics approval**   * If ethics approval is required for this project, indicate the status (not yet submitted, submitted, approved, exempt) * Name the institutional research board(s) or ethics committee(s) relevant for conducting the analyses * Provide details of how this analysis project is covered by existing informed consent, if applicable * Alternatively, refer to the documents where this information is provided |
| Names and contact | 1.4 | **Provide names, affiliations, and contacts of key project team members**   * List all author(s) of this document * List all principal investigator(s) of this analysis project * List all data analyst(s) * List any oversight committee members, if applicable |
| **PROJECT BACKGROUND** | | |
| Research aims and objectives | 2.1 | **Describe the research aims and objectives of this analysis project**   * State whether the research objectives are descriptive, predictive, or causal * Refer to applicable project documents for background and rationale |
| Target population | 2.2 | **Describe the target population of interest for the research objectives**   * For example, explain health care context, geographic location, age, sex, gender, specific disease to which the results of the analysis project apply |
| **DESIGN AND DATA** | | |
| Data sources | 3.1 | **Describe the sources of data (e.g., routinely collected electronic health records, registries, surveys)**   * Describe the name and origin of each data source relevant for this analysis project * Describe the structure, design, or method of collection of each data source, e.g. survey, registry, routinely collected electronic health records, clinical measurements. * Describe information relevant for the analyses on methods for assessment and collection of data and corresponding instruments (e.g. questionnaires or laboratory tests) * Discuss potential for information or selection biases of the data source for the analysis project * Alternatively, refer to the documents where this information   is provided (study protocol, publication, DOI, website) |
| Design | 3.2 | **State the study design for this analysis project (e.g., cross-sectional, cohort, case-control, case-crossover, serial cross-sectional)**   * Define observation units, e.g. individuals, samples, or groups * Describe methods or variables for selecting observation units, e.g., specific subset or designs * Alternatively, refer to the documents where this information is provided |
| Eligibility criteria | 3.3 | **Specify eligibility criteria for the analysis project based on information available at baseline**   * Specify the observation windows for which data will be used with start and end dates. This may refer to periods for examinations, diagnoses, visits, events, or follow-up, for example. * Define the index condition for including observation units, for example, the first diagnosis of a disease, the beginning of a treatment, or a health examination. * Describe the index date which will be used to anchor observation windows between observation units in analyses, if applicable. For example, the index date could be the date of diagnosis, the date of treatment assignment, or the date at which predictors are measured. Define the index date for each research objective. Specify follow-up times, if applicable * Describe further eligibility criteria (e.g. variables used to select the analysis population), construct a flow diagram starting with the data source and successive steps for inclusion criteria and then exclusion criteria. * Describe the setting such as health care context or geographic location |
| Data sets | 3.5 | **Describe how the data is provided for this analysis project (e.g., format and content of data sets)**   * Describe how the data was processed from the data source for use in this analysis project, e.g., linking of datasets, deletion of cases and variables, data cleaning, assessing compliance with pre-specified structural and technical requirements is assessed. * Include a data dictionary or refer to documents where this information is available. This includes variable abbreviations, variable names, values or units of measurements, instruments (e.g. questionnaires or laboratory tests), data standards, expectations about the data, or link to data sources for specific variables, if applicable * Alternatively, refer to the documents and reproducible code where this information is provided, if available |
| **VARIABLES** | | |
| Variables used in the main data analysis | 4.1 | **Define all variable(s) and their roles in the analyses to answer research questions**   * Provide details relevant to the analyses on measurements or point(s) of measurements relative to the index date * For each objective, and according to the nature of the objective, define the roles of variables such as outcome, covariates, independent variables, predictors, exposure, mediators, or confounding variables * If appropriate, provide justification for the specification of confounding variables by including directed acyclic graphs (DAG) indicating the assumed causal relationships or selection variables * Include a table of variables aligned with each objective, if applicable |
| Other variables | 4.2 | **Define any variables that are not directly used to answer the research objectives, but which provide information about the observation units**   * These variables can be used for evaluating data quality, structuring IDA reports, or generating statistical weights. Examples are process variables such as centers, time stamps, or other design variables |
| **METHODS: MAIN DATA ANALYSIS (MDA)** | | |
| Description of observation units | 5.1 | **Describe the methods of analysis to summarize the characteristics of observation units.**   * For example, describe which variables will be included in such descriptions, type of summary statistics used, or graphical descriptions. A template for a table summarizing the characteristics can be included here. * Specify if these data descriptions will be stratified by any variable, and by which, if applicable * Specify any association analyses to be conducted, e.g. by means of correlation analysis, and the type of association measure to be computed or the graphical display of such associations |
| Main data analysis methods | 5.2 | **Describe the methods of analysis for each research objective, including the quantities to be estimated, the analysis models with variables and observation units, and methods to mitigate potential bias for non-random selection**   * Describe the quantities to be statistically estimated or predicted (estimands, predictimands) for each research objective, e.g., mean, cumulative incidence, probability of event, odds or hazard ratio, differences in means) * Specify the analysis model for each objective and/or the relevant estimator to estimate the quantities of interest, e.g., any focal regression coefficients or predictions or transforms thereof * Describe how variables are handled in each analysis, for example regarding transformations, functional forms, or interactions * Define observation units for the analyses * Discuss methods used to mitigate potential bias for non-random selection * Describe methods for model development, e.g. selection of variables and their functional form * Specify planned subgroup analyses, if applicable * Describe any further aspects of analysis that are relevant for full reproducibility |
| Assumptions and diagnostics | 5.3 | **State any statistical assumptions and diagnostics for each analysis.**   * Specify all measures and diagnostics used to evaluate statistical assumptions and appropriateness of analyses, including graphical tools * Specify all measures used to evaluate model performance, if applicable * If missing data are expected, describe how this will be handled with details about any imputation or augmentation method * If measurement errors are expected, describe how they will be addressed. Specify planned sensitivity analyses, if applicable * Specify planned sensitivity analyses, if applicable * Describe potential limitations of the analyses |
| Sample size | 5.4 | **Describe how the sample size was determined, including all assumptions supporting the sample size calculation**   * Describe available information about sample size, number of outcome events, or follow-up time, if applicable * If the provided numbers are based on expectations provide the rationale * Provide all details and assumptions needed to replicate the sample size or power calculation independently |
| Software | 5.5 | **Describe software used for all analyses, visualizations, data management, data archiving, or backups** |
| **METHODS: INITIAL DATA ANALYSIS (IDA)** | | |
| Data preprocessing | 6.1 | **Describe any methods for preparing data for the analyses**   * Describe how consistency of dates is checked * Describe how new variables needed for the main data analysis are computed * Describe any preparation or formatting of data needed for the analyses, e.g., multistate or wide and long format |
| Unit missingness | 6.2 | **Describe any methods to identify the extent of unit missingness**   * Describe methods to explore the lack of data availability for an observation unit * Describe methods to compare characteristics of those with and without unit missingness, where possible. For example, specify which characteristics of participants will be compared to those from census data, or specify what baseline data will be compared for those who did and did not drop out before the study sample was recruited. |
| Unit profile | 6.3 | **Describe any methods to summarize the temporal or structural pattern of observations for each observation unit**   * Unit profile refers to such patterns of observations, for example the number of units enrolled over time, over geographical locations, or other relevant structures |
| Item missingness | 6.4 | **Describe the methods to examine item missingness (i.e., missing values in variables)**   * Describe methods to explore the number and proportion of missing values for each variable and reasons for missingness, if known * Describe methods to evaluate the number of complete observations for each analysis, if applicable * Describe methods to examine patterns of missingness and relate these to completely observed variables, for example to outcome and exposure variables or any important structural variables that are completely observed. |
| Univariable descriptions | 6.5 | **Describe the methods to summarize the distribution of each variable used in the analyses with numerical or graphical summaries** |
| Multivariable descriptions | 6.6 | **Describe any multivariable descriptive statistics and graphical summaries**   * Specify (“structural”) variables which will be used to help structure the results of multivariable analyses * Specify which associations between covariates with these structural variables or other variables will be evaluated and the numerical measures or graphical summaries that will be used * This does not include evaluating associations with outcome variables |
| **EVALUATION AND UPDATES** | | |
| Evaluating the SAPI | 7.1 | **Indicate if an update of the SAPI is needed after IDA**   * This information is provided after completion of IDA * If there is no need for an update, this should be stated |
| Updating the SAPI, if applicable | 7.2 | **List all SAPI sections affected by updates and state the reasons for the updates**   * This may include changes to inclusion or exclusion criteria for observation units or updated information about the variables, for example ranges, categories, transformations * Updates to the main data analysis plan may also be needed |
| **SUPPLEMENT** | | |
| Key references | 8.1 | **Include references related to statistical methods, research aims, or other background information**   * Specify reporting guidelines to be used for the manuscript (EQUATOR Network) |
| Sustainable handling of analysis outputs | 8.2 | **Describe the measures to ensure sustainable handling of analysis outputs and FAIR principles (e.g., archiving analysis reports, code files, updated datasets, documentation for interoperability and reuse)**   * List the analysis documents, for example final SAPI, IDA report, MDA report, data quality report, code files, updated data files * Describe data formats and documentation that will be used to ensure interoperability and reuse by the research team or third parties * Indicate where and how these outputs will be archived or who they will be shared with |

Abbreviations: SAPI – statistical analysis plan with initial data analysis, MDA – main data analysis, IDA- initial data analysis, FAIR – findable accessible, interoperable, and reusable