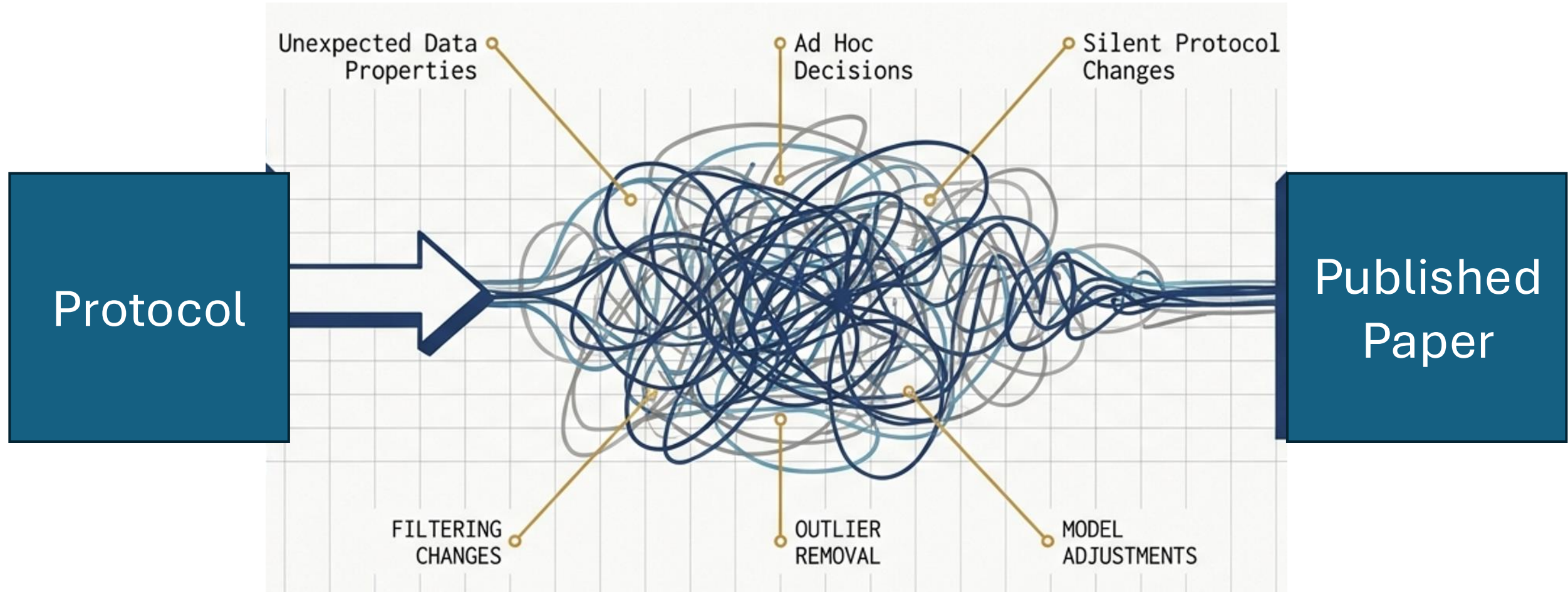


Statistical analysis plan with initial data analysis (SAPI): Validating the SAPI checklist in analysis projects

Marianne Huebner (Michigan State University, USA),
Carsten O Schmidt (University of Greifswald, Germany),
Lara Lusa (University of Primorska, Slovenia),
Georg Heinze (Medical University of Vienna, Austria)

Without a systematic plan to handle realities of observational data:



SAPI Statement

SAPI

A systematic framework developed via a Delphi consensus process that bridges the gap between a protocol and the final analysis by explicitly accounting for realities of observational data.

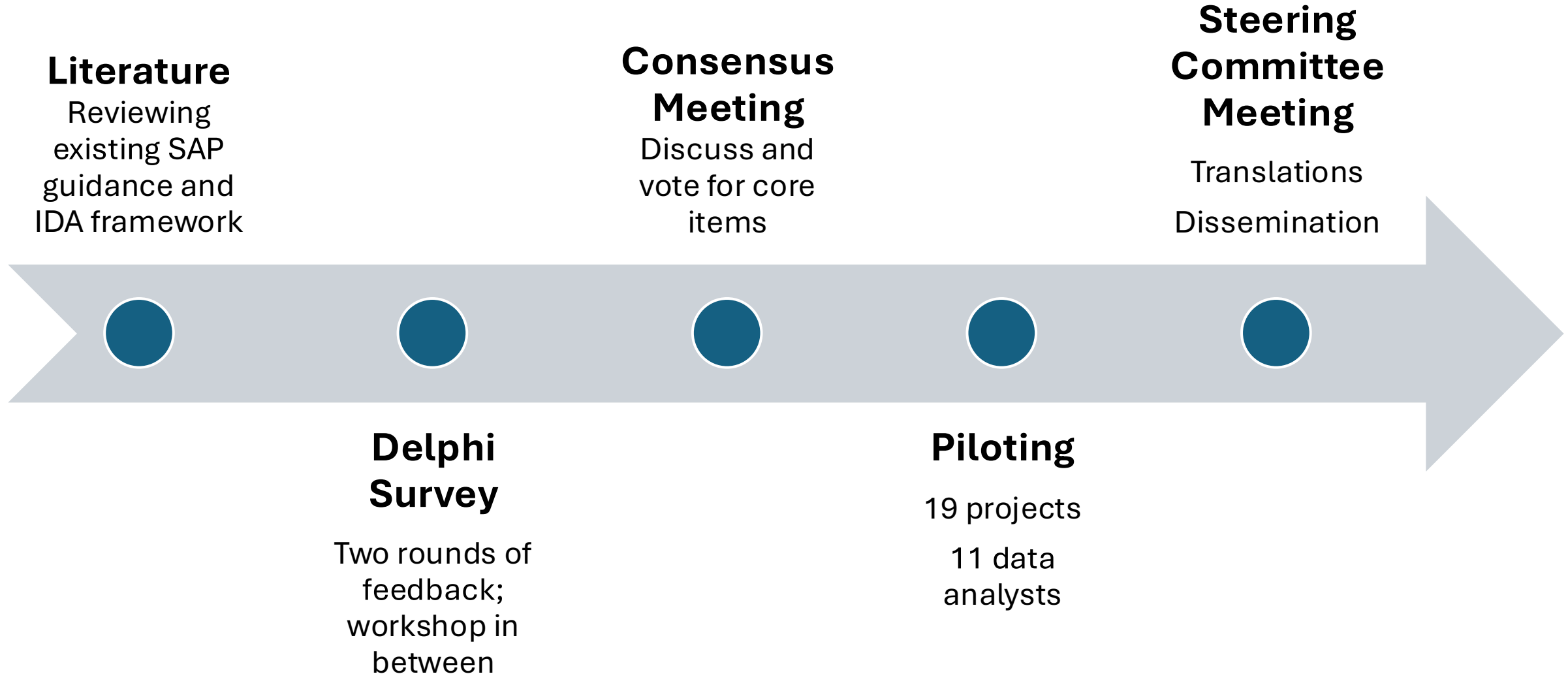
Applicable across a wide range of observational study designs and research objectives (descriptive, predictive, causal)

SPIRIT

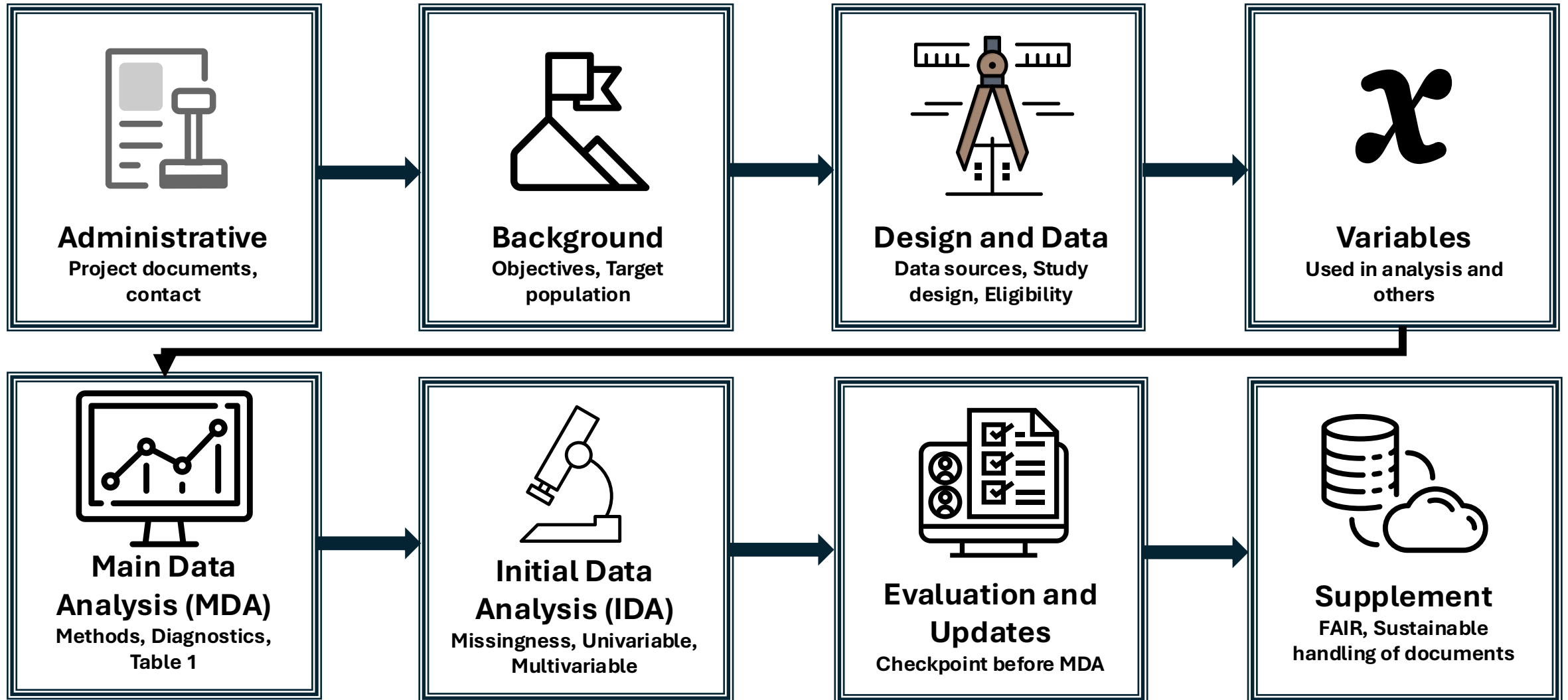
DEBATE

STROBE

Building Consensus: The DELPHI process



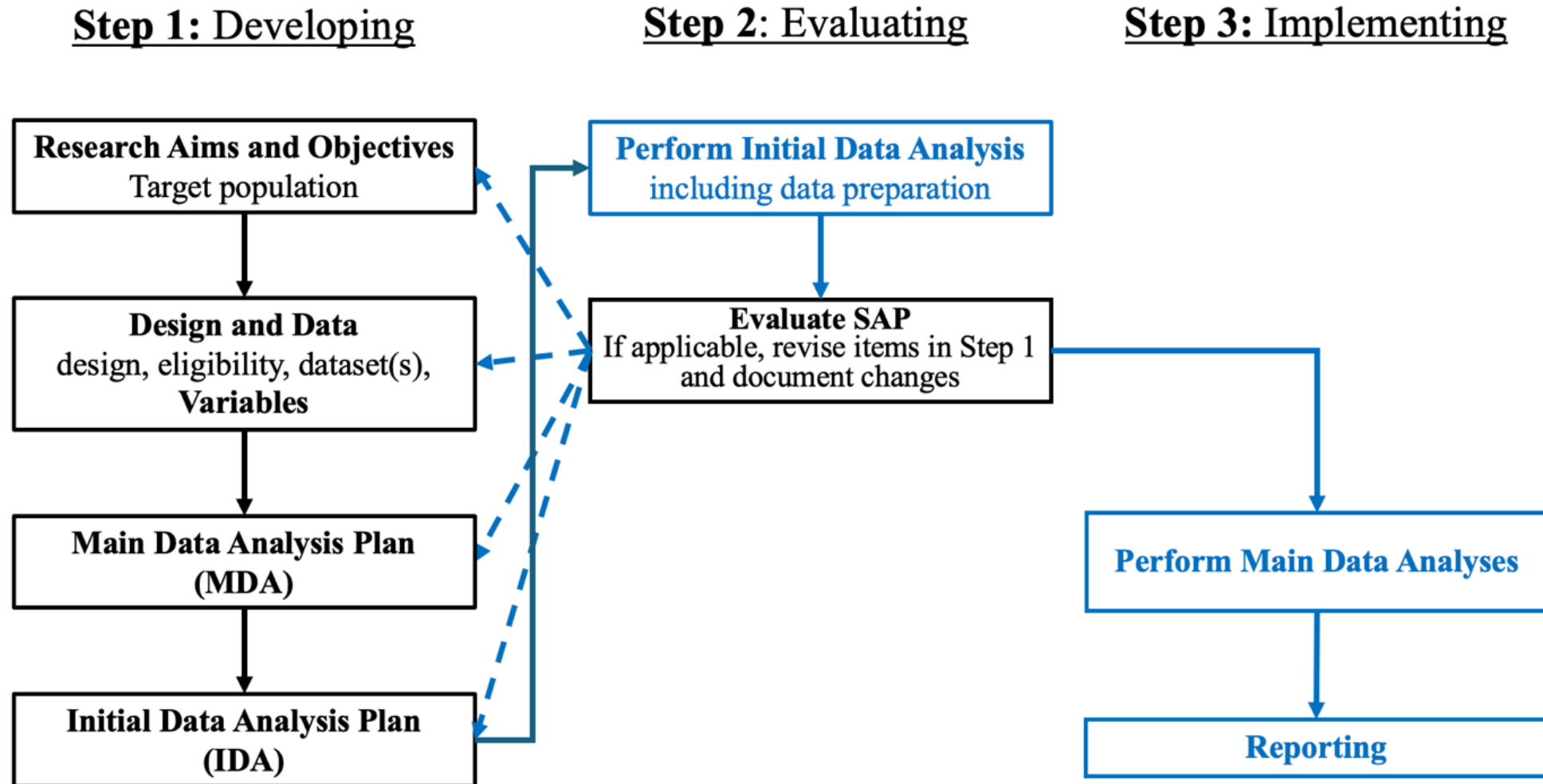
Statistical Analysis Plan with Initial data analysis (SAPI)



Why SAPI?

Key challenges in observational studies	How SAPI addresses it
Unknown data quality and data properties	Planned initial data analysis (IDA)
Analysis plan revisions	Deviations documented prior to main data analysis (MDA); analysis decisions recorded
Multiple objectives in the same study (e.g. descriptive, predictive, causal)	Objectives → Variables → MDA/IDA workflow
Poor reproducibility	Systematic, aligned with MDA; augments metadata
Limited reuse of analysis outputs	FAIR, access, archiving, reuse

Workflow



Initial data analysis

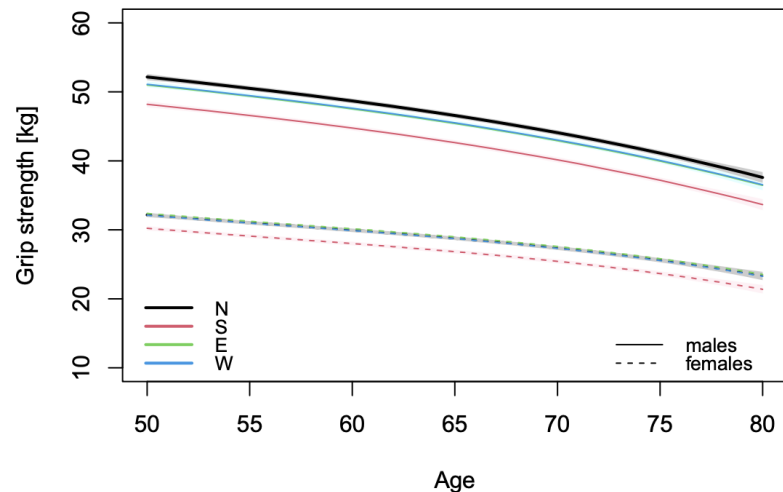
Provide reliable knowledge about the data to determine the suitability of the data for the main data analysis. No hypotheses generating activities.

Assessment	Observation units	Predictors Exposures	Covariates	Outcomes	Structural variables
Missingness (proportion, pattern, profiles)					
Univariable distributions (scales, shape, outliers)					
Multivariable distributions/ associations/groups					

Case Study 1: Grip strength, aging, and physical activity

Research aim

to assess the age-associated decline of hand grip strength by sex



Variables

Outcome: grip strength

Time metric: age

Time-fixed: sex, education, height

Time-varying: weight, smoking, physical activity

<<Data dictionary>>

Case Study 1: Grip strength, aging, and physical activity

Main data analysis

Linear mixed model for maximum grip strength

Primary estimand: Age $\times$ sex interaction

Random intercept for age

Longitudinal repeated measurements

Initial data analysis

Unit missingness (non-enrollment, drop-out, intermittent missingness)

Item missingness

Univariable descriptions:
baseline and at different time points

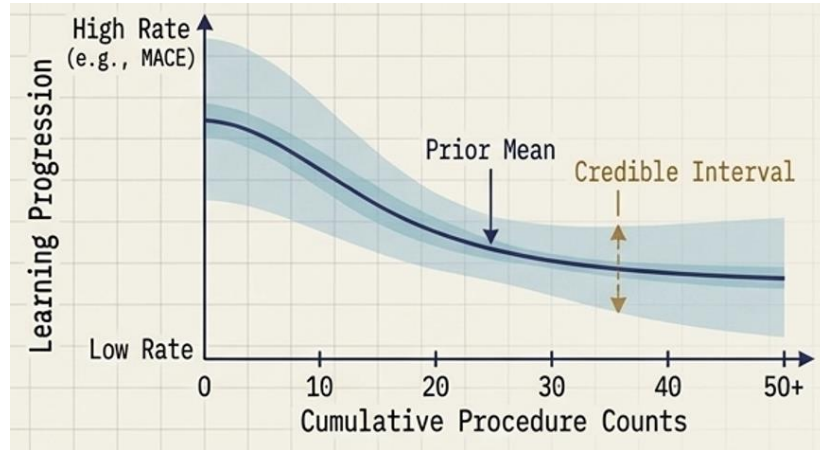
Multivariable descriptions

Longitudinal aspects (trends of time-varying variables, individual profiles across time metric)

SAP v1 versus SAP v2

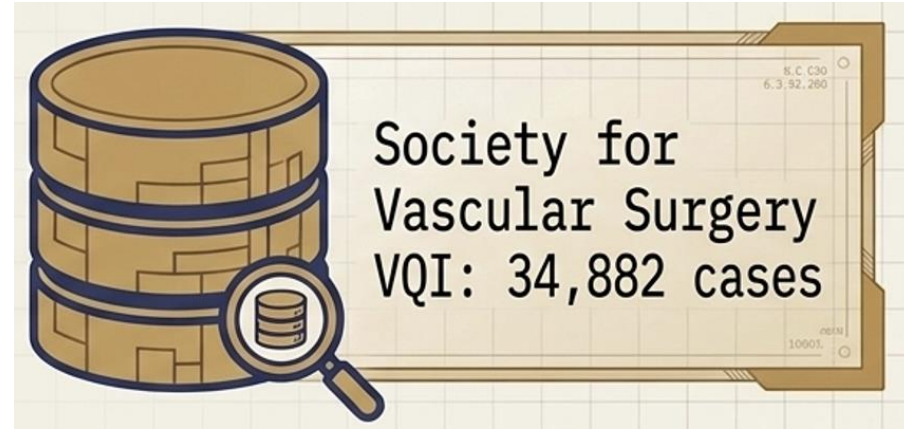
	SAP v1 before IDA	SAP v2 after IDA
Variables	Smoking: time-varying variables	Smoking: baseline only (missing by design later)
Eligibility	Ages 50+ at first interview, with known birth year at	Study population: Analyses restricted to age ≤ 90 years (limited data at older ages)
Main data analysis (missingness)	Did not account for missing by design in waves; unexpected attrition in outcome measurements	Missing data: Birth cohort adjustment and multiple imputation added; attrition explicitly evaluated
Main data analysis	Assumed linear decline in age	Age: Flexible functional forms (splines/fractional polynomials) evaluated
Sensitivity analyses	none	Sensitivity analyses: Informative mortality, dropout, correlation structure, age range, and measurement error

Case Study 2: Surgical learning curves



Research aim

Estimate the learning curve for transfemoral carotid artery stenting for adverse events and operative time based on procedure count



Variables

Outcome: MACE or OP time

Key: Procedure count by physician

Explanatory: Age, BMI, sex, comorbidities, emergency, insurance

Case study 2: Surgical learning curves

Main data analysis

Multivariable Bayesian regression models with Gaussian process priors to map progression over procedure counts

Sensitivity analysis with other priors

Initial data analysis

Item missingness (n, %)

Univariable (quantiles, histogram, distributions of transformed variables for skewed variables)

Multivariable (patient characteristics stratified by year of surgery and by bins of procedure count)

Case study 2: Supplement

Data dictionary, including
derived variables

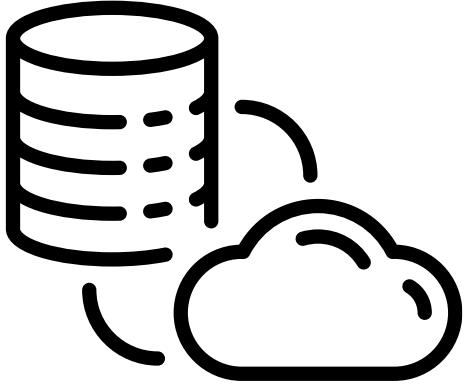
Variable	Variable type	Values
MACE	Categorical	0 = No incident 1 = Incident
Operating Time	Continuous	Minutes, ≥ 27 , <1200
Physician ID	Nominal	1 - 2299
Patient ID	Nominal	1 - 37579
<i>Procedure count</i>	<i>Count</i>	<i>1-453</i>
<i>Procedure count</i>	<i>Categorical</i>	<i><u>1, 2-10, 11-50, 51-100, 101+</u></i>
Age	Continuous	Years, ≥ 40
Sex	Binary	1 = Male 2 = Female
Emergency type	Categorical	1 = Not emergent 2 = Urgent 3 = Emergent

Reporting guideline:
STROBE

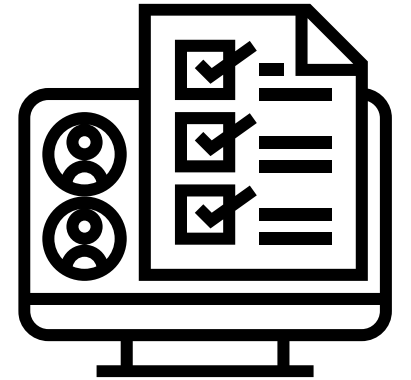
Research team:
PIs: SAP, IDA+MDA reports
Statisticians: Data, Code

Open science:
Code, SAP, IDA+MDA
reports on Zenodo

Sustainable handling of analysis outputs



Archive original and updated SAP
Preserve IDA and MDA reports
Publish code
Share data, if possible
Ensure full analytical lifecycle is
reproducible



Lessons learned

SAPI is practical and feasible to implement in real analysis projects

The workflow is logical and efficient, without repetition of information across sections, limits duplication from protocol items

Roles, responsibilities, and access of research team are clearly defined, improving team coordination and transparency/awareness

The framework encourages reproducibility, reporting, and long-term archiving of analysis decisions

SAP v2 documents analysis updates transparently, providing a clear rationale for changes after IDA

Note: A meeting of the research team should take place to discuss IDA results before SAP v2. Changes may also be documented in IDA/MDA reports.

- * Explicit planning of IDA
- * Checkpoint after IDA
- * Transparent updates
- * Integration of FAIR principles



Scan for the SAPI OSF repository



STRATOS
INITIATIVE

← Initial Data Analysis Topic Group, SAPI checklist, Resources

Abstract

The SAPI statement (Statistical Analysis Plan for observational studies with Initial data analysis) provides recommendations for developing analysis plans for observational studies. It promotes a systematic process to improve completeness and transparency, and ultimately to better interpretability of research results for evidence-based health decision making. It can be used for descriptive, predictive, or causal research objectives. The SAPI checklist consists of 27 items comprising administrative information, project background, design and data, variables, main data analysis (MDA), initial data analysis (IDA), evaluation and updates, and supplement with plans for disseminating and archiving research outputs.

We demonstrate these items in several examples of analysis plans and their updates following IDA including investigating functional forms in regression models or estimating learning curves.

In each of these examples, updates of the SAP were implemented after IDA. These examples serve to demonstrate that a full pre-specification of an SAP is rarely possible without data examinations. The SAPI checklist and IDA framework guarantee that these looks are performed while strictly ignoring outcome-predictor associations and that their impact on the final analysis specification is transparently reported.