**RESEARCH AREA: Statistical Methods in Biomedical Research**

**Hidden analyses: a systematic review of reporting practice and recommendations for more transparent reporting of initial data analyses**

**Keywords:** initial data analysis, reporting, observational studies, STRATOS Initiative

**Background**

In the data pipeline from the data collection process to the planned statistical analyses, initial data analysis (IDA) typically takes place between the end of the data collection and do not touch the research questions. A systematic process for IDA and clear reporting of the findings would help to understand the potential shortcomings of a dataset, such as missing values, or subgroups with small sample sizes, or shortcomings in the collection process, and to evaluate the impact of these shortcomings on the research results. A clear reporting of findings is also relevant when making datasets available to other researchers. Initial data analyses can provide valuable insights into the suitability of a data set for a future research study.

**Aim**

Our aim is to describe the practice of reporting of initial data analyses in observational studies in five highly ranked medical journals with focus on data cleaning, screening, and reporting of findings which led to a potential change in the analysis plan.

**Methods**

Reporting adheres to the PRISMA guidelines. To aid transparency, the PubMed search strategy and data collection form will be available on the STRATOS TG3 website (https://www.stratosida.org/ activities/project-systematic-review-of-ida-reporting).

*Sampling frame*

Papers will be selected from five highly ranked medical journals (The New England Journal of Medicine (NEJM), Lancet, Journal of Clinical Oncology (JCO), Circulation (CIRC), Journal of the American Medical Association (JAMA)). All papers published in a six-month window from January 1, 2018 to July 15, 2018 meeting the inclusion criteria were included. The primary reviewer [MH] screens the titles and abstracts against the inclusion criteria. Full reports will be obtained of all articles which appear to meet the inclusion criteria below. Each statement in a selected paper needs to be carefully evaluated regarding its relation to initial data analysis. For an equal representation across journals five papers from each journal will be randomly selected and reviewed by two reviewers. The random sampling protects against unforeseen selection bias. For each journal selected papers will be ordered, then the order will be permuted using the statistical software R, and the first 5 papers on the list will be selected, to retain t equal representation across journals. If, upon examination, an article does not meet the inclusion criteria, it will be replaced by the next paper on the list from the target journal.

**Sample size considerations:** The sample size of 25 papers is not based on a formal sample size criterion, but is perceived as sufficient to gain general insights on IDA reporting.

**Inclusion criteria**

* Observational study, original research articles
* Published in one of the selected journals and available between January 2018 and July 15th, 2018 (including Epubs ahead of print).

**Exclusion criteria**

* Clinical trials, randomized experiments, laboratory studies, genetics or genomics studies, letters, editorials, reviews, guidelines, comments
* Fewer than 50 participants
* Simulation studies, imaging studies, cost studies
* Studies published only in abstract form
* No clear research aim stated

*Data extraction*

Data will be extracted from the selected papers using a standardized data extraction form developed for this review. An online submission form will be prepared, piloted and refined prior to use by two authors LL and MH. The form includes data on study background (author, country, sample size, data source), elements of IDA framework reported (data cleaning, screening, change in the analysis plan). Each aspect was classified by the location in the paper where the respective aspect was targeted and ranked by sufficiency of information (not mentioned, mentioned, mentioned with sufficient detail or not applicable). The full articles will be reviewed, and the location of IDA statements will be noted as Introduction, Methods, Results, Discussion, and Supplement. If topics are mentioned in more than one section, the main selections are reported and therefore the sum of reported locations can exceed the sample size of 25 articles. We report consensus between two reviewers.

**Statistical methods**

Both quantitative summaries and qualitative evaluation of text excerpts will be employed. Each extracted item will be summarized overall and by location in the article. A summary stratified by journal will not be attempted due to the small number of articles from each journal.

**Strengths and limitations**

This review is limited to 25 papers in medical journals. However, the aim is to get a general impression of IDA reporting with examples across five medical journals and a discussion on how reporting might be improved.

IDA in studies based on disease registries, large electronic health record data bases, or population cohorts may have been performed prior to the study leading to less IDA reporting.

**References**

1. Huebner M, le Cessie S, Schmidt C, Vach W. A contemporary conceptual framework for initial data analysis. Obs Stud. 2018;4:171–92.
2. Ioannidis JPA. Why Most published research findings are false. PLoS Med. 2005;2:e124.

**Abbreviations**

IDA – Initial data analysis

**Principal Investigators**

Marianne Huebner, Michigan State University

Lara Lusa, Institute of Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia.

**Co-Investigators**

Werner Vach, Department of Orthopaedics and Traumatology, University Hospital Basel, Basel, Switzerland.

Saskia leCessie, Department of Clinical Epidemiology and Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Carsten Schmidt, Institute for Community Medicine, SHIP-KEF University Medicine of Greifswald, Greifswald, Germany.