Strengthening Analytical Thinking for Observational Studies (STRATOS): INTRODUCING THE TOWARDS EXCELLENT DATA (TED) PANEL

Carsten Oliver Schmidt 1 , Andreas Klinger 2 , Willi Sauerbrei 3 , Georg Heinze 2 on behalf of the TED panel

1. Institute for Community Medicine, SHIP-KEF University Medicine of Greifswald, Greifswald, Germany,
2. Medical University of Vienna, Center for Medical Data Science, Institute of Clinical Biometrics, Vienna, Austria
3. Faculty of Medicine and Medical Center, University of Freiburg, Germany

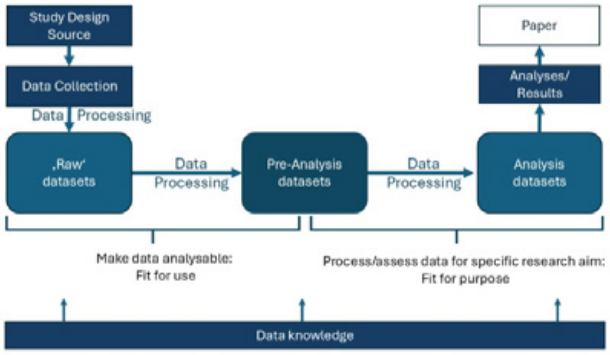
The abstract of the first STRATOS paper (1) states 'The validity and practical utility of observational medical research depends critically on good study design, excellent data quality, appropriate statistical methods and accurate interpretation of results', a statement that certainly achieves the widest possible consensus. As general requirements for STRATOS publications, papers should be made open access, and accompanied by openly available sample data and software. To support this commitment, from the beginning, STRATOS established a dataset panel. At the recent STRATOS workshop in Leiden (see Biometric Bulletin 4/24 for a summary of the key theme 'Estimands') it was decided to relaunch the panel under the new name 'Towards Excellent Data' (TED) with the aim of more directly addressing the recurring shortcoming that analysts frequently receive datasets lacking sufficient information about their provenance, transformations, or quality. This absence of data knowledge can lead to undetected biases or artefacts being carried into the modelling stage. Compounding the problem, readers of scientific articles are rarely provided with a full account of the data properties or the methodological procedures that shaped an analysis dataset, as documented in Topic Group 3's review of Initial Data Analysis (IDA) reporting in high-impact journals (2) . This multifaceted deficit in data knowledge, affecting both analysts and readers, undermines credible science.

The relaunched TED panel pursues two primary aims. The first aim is to guide members of the STRATOS initiative and other researchers in establishing principles and applying tools that support the acquisition, documentation, and responsible sharing of research data, aligned with existing works, e.g., (3, 4) . The second aim is to create a repository of datasets suitable for methodological research, focusing on those that have been used in publications of the STRATOS initiative. For some of the datasets we will also add a 'complete case version' which replaces missing values by a value created with a suitable single imputation method. Such a modified dataset is intended for the use in methodological comparisons that do not focus on missing data handling. The TED panel has set up a new website, accessible at https://stratostedp.github.io on which activities of the TED panel are described.

Figure 1 illustrates the above-mentioned challenges. To conduct valid analyses, analysts must understand the origins and properties of their data. Yet, because they are often not involved in data collection or preprocessing, they must rely on the information provided to them. Despite this, analysts remain accountable for the integrity of their analyses - a responsibility that depends both on how they use the data and on their understanding of it. Moreover, analysts bear responsibility for the stewardship of data after analysis, including proper documentation,

curation, and preservation. This ensures that others can reproduce their findings and evaluate the data's suitability for future use. The TED panel will focus on the descriptive metadata and documentation needed to assess datasets prior to use. In this sense, it operationalises the "Fit for Use" perspective.

Figure 1: Data knowledge and responsible use of data in analyses as the pillars of data excellence



As a first activity in support of the second aim of the TED panel, we have recently collected information about all datasets that have been used to date by STRATOS publications, and summarized this information in a directory, available on the new website. The directory includes the bibliographic references of the papers, keywords (to facilitate searches), and the download links to the datasets or details on how to access the datasets if they are not openly available. The directory will be updated periodically to include future articles that are published on behalf of topic groups and panels of the STRATOS initiative.

We will propose an amendment to STRATOS' publication principles requiring that future authors of articles written on behalf of the STRATOS initiative provide comprehensive information about any datasets used in their work at the time of submission to the STRATOS Publication panel for review. This will enable us to incorporate the relevant information when such a paper is finally published. Potential future extensions concern datasets that have not yet been used by STRATOS, but may be useful for methodological research or comparison studies.

Our aims resonate with international efforts to improve data quality and transparency - which require data to be Findable, Accessible, Interoperable, and Reusable (FAIR) (5) - highlighting the need for structured metadata and long-term usability. The FAIR principles form an important basis for data sharing (4). Similarly, policy efforts, for example within the European Health Data Space seek to define quality criteria and data labelling mechanisms. While STRATOS and TED do not replicate these efforts, they contribute methodological aspects and assess the analytical implications of poorly documented or insufficiently understood datasets.

In summary, STRATOS with its TED panel promotes a vision of research in which dataset understanding and data knowledge is neither optional nor informal but a core pillar of credible science.

This vision includes:

- Datasets accompanied by structured meta-data and contextual information that provide a transparent account of data generation and processing;
- Data quality assessment results reported alongside modelling results in scientific papers, for example as supplementary material, enabling readers to better evaluate the base of reported findings;
- Metadata recognized as an essential research output, not merely a technical add-on, ensuring reuse and interpretability of datasets over time.

REFERENCES

1. Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J, on behalf of the STRATOS initiative. STRengthening analytical thinking for observational studies: the STRATOS initiative. Stat Med. 2014;33(30):5413-32.
2. Huebner M, Vach W, le Cessie S, Schmidt CO, Lusa L on behalf of the Topic Group 'Initial Data Analysis' of the STRATOS Initiative. Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Methodol.2020;20:1-10.
3. Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, Giaretta D, et al. The TRUST Principles for digital repositories. Sci Data. 2020;7(1):144.
4. Pellen C, Munung NS, Armond AC, Kulp D, Mansmann U, Siebert M, et al. Data management and sharing. J Clin Epidemiol. 2025;180:111680.
5. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.