STRengthening Analytical Thinking for Observational Studies (STRATOS): On the importance of Data Quality Assessments and Initial Data Analysis

C. O. Schmidt[1], G Heinze[2], L Lusa[3], M Huebner[4]

[1]Institute for Community Medicine, SHIP-Clinical Epidemiological Research, University Medicine of Greifswald, Greifswald, Germany

[2]Section for Clinical Biometrics; Center for Medical sTatistics, Informatics and Intelligent Systems; Medical University of Vienna, Vienna, Austria

[3]Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technology, University of Primorska, Koper, Slovenia.

[4]Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

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 which certainly achieves the widest possible consensus. STRATOS has no panel on data quality assessments (DQA), but related issues are discussed in the topic group on initial data analysis (IDA, TG3) and some of the STRATOS members are also active in DQA. For example, the STRATOS topic group TG3 on IDA has developed a guidance paper which outlines an IDA framework.[2] The framework consists of six steps: (I) metadata setup to systematically represent all background information or relevance, (II) data cleaning to identify and correct data errors, (III) data screening to review and document data properties, (IV) initial IDA reporting, (V) refining or revising the analysis plan, and (VI) IDA reporting in research papers. It is obvious that particularly the data cleaning and data-screening step closely relate to DQA. A subsequent STRATOS review of selected articles published in major medical journals revealed considerable shortcomings regarding IDA reporting.[3] Factors contributing to the partial neglect of IDA in scientific practice include lack of funding for such activities, ad hoc approaches for different research projects, as well as the extent and complexity of such initial data analyses despite their seeming simplicity. Analysts are faced with a myriad of choices of what to assess, which tools, routines, and visualizations to use, and how to structure a report of the potentially extensive output

in a digestible fashion when conducting a full scope IDA. Therefore, TG3 has expanded its work on guidance at different levels. The paper "Ten simple rules for initial data analysis" [4], compiles a brief overview of the aspects of IDA and the benefits of integrating IDA in the research practice. It should be emphasized that IDA must not be confused with exploratory data analysis (EDA) as it does not aim at creating any scientific findings. Current activities of TG3 are the development of guidance when using cases with public data. This includes the project "Regression without Regrets" on IDA for multivariable regression models with continuous or binary outcomes and IDA for longitudinal data including complex survey data.

DQA and IDA are closely related as illustrated in a cooperative work with several STRATOS members, most of them from TG3, in a project funded by the German Research Council (DFG). This includes a data quality framework for observational studies along with the dataquieR R package, to facilitate DQAs.[5,6] The framework describes four data quality dimensions (integrity, completeness, consistency, and accuracy) with a total of 34 indicators. There is a considerable overlap between IDA and DQA in addressing data errors. However, while IDA assesses the match between the data at hand and its suitability for the intended statistical analyses according to the underlying research questions and analysis plan, DQA commonly assesses deviations from requirements on the data not tied to one specific research project as it works mostly from the data generation perspective.

The strong interest in a DQA/IDA workshop held in Berlin on November 17th - 18th, 2022, by over 130 participants demonstrated the growing interest and demand for guidance on DQA and IDA. The workshop was organized by several statistical and methodological societies from Germany, including GR-IBS, the German Association for Medical Informatics, Biometry and Epidemiology (GMDS), the German Society for Epidemiology (DGEpi), the Technologie- und Methodenplattform für die vernetzte medizinische Forschung (TMF), STRATOS, and funded by the National Research Data Infrastructure for Personal Health Data (NFDI4Health). In addition to presentations related to the TG3 activities mentioned above, other presentations discussed programming tools of relevance to IDA and DQA. This comprises R packages[7] and a range of open source as well as commercial stand-alone software.8 The workshop highlighted that IDA and DQA are more than a statistical task in being an infrastructure challenge that must consider the broader context of the entire scientific data lifecycle, from data generation to downstream use of data sets. For example, the workshop stressed the importance of using common data models (CDMs) and syntactic standards such as the Observational Medical Outcomes Partnership Common Data Model (OHDSI-OMOP), Clinical Data Interchange Standards Consortium (CDISC), Fast Healthcare Interoperability Resources (FHIR), or others, to provide interoperable data bodies with rich, well-curated metadata; this is desperately needed for transparency, reproducibility, and accountability to overcome shortcomings of the frequently employed practice of using spreadsheets.

As an outlook, future STRATOS work will increasingly combine statistical demands with recent developments from information technologies to facilitate efficient IDA and DQA work flows. The appropriate conduct of IDA and DQA may prove an important cornerstone in advancing FAIR and reproducible science.[9]

Further overviews of current and past activities of TG3 can be found at https://www.stratosida.org and on the website of the STRATOS initiative http://www.stratos-initiative.org/.

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;33:5413-32.

2. Huebner M, Le Cessie S, Schmidt CO, Vach W. A contemporary conceptual framework for initial data analysis. Observational Studies 2018;4:71-192.

3. Huebner M, Vach W, le Cessie S, Schmidt CO, Lusa L, Topic Group "Initial Data Analysis" of the SI. Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Methodol 2020;20:61.

4. Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M, Topic Group "Initial Data Analysis" of the SI. Ten simple rules for initial data analysis. PLoS Comput Biol 2022;18:e1009819.

5. Schmidt CO, Struckmann S, Enzenbach C, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021;21.

6. Richter A, Schmidt CO, Krüger M, Struckmann S. dataquieR: assessment of data quality in epidemiological research. JOSS 2021;6:3039.

7. Marino J, Kasbohm E, Struckmann S, Kapsner LA, Schmidt CO. R Packages for Data Quality Assessments and Data Monitoring: A Software Scoping Review with Recommendations for Future Developments. Applied Science 2022;12.

8. Ehrlinger L, Wöß W. Survey of Data Quality Measurement and Monitoring Tools. Front Big Data 2022.

9. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018.