STRengthening Analytical Thinking for Observational Studies (STRATOS):

Introducing the Initial Data Analysis Topic Group (TG3)

Carsten Oliver Schmidt[1], Werner Vach[2], Saskia le Cessie[3], Marianne Huebner[4] on behalf of TG3

[1]Institute for Community Medicine, SHIP-KEF, University Medicine of Greifswald, Germany; Email: Carsten.schmidt@uni-greifswald.de

[2]Department of Orthopaedics and Traumatology, University Hospital Basel, Basel, Switzerland; Email: Werner.vach@usb.ch

[3]Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands; Email: S.le_Cessie@lumc.nl

[4]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA; Email: huebner@stt.msu.edu

In the previous issues of the Biometric Bulletin, the STRATOS initiative was introduced and the Topic Groups on Missing Data (TG1), and Measurement Error (TG4) described their activities. In this issue, we report on activities of the Topic Group on Initial Data Analysis (TG3). Whereas missing data and measurement error are topics well discussed in literature, this is less so for initial data analysis (IDA) despite IDA being part of the everyday work of many statisticians.

Some aspects of IDA such as data cleaning or specific techniques for exploratory data analysis are discussed in publications. There are numerous online reports or blog posts discussing certain aspects of IDA, sometimes reflecting personal experience or experience in a specific field. However, since the classic paper by Chatfield [1] few systematic activities aimed to further develop IDA and its reporting. Usually, there is only minimal inclusion of elements of IDA in research papers. This absence of a framework and inadequate coverage of IDA was the starting point for Topic Group 3. It became quickly apparent, that initial data analysis is also a challenging part of the research process.

The aim of Topic Group 3 is to improve awareness of IDA as an important part of the research process and to provide guidance on conducting IDA in a systematic and reproducible manner.

The members of this topic group are Marianne Huebner, USA, Saskia le Cessie, Netherlands, Werner Vach, Switzerland (joint chair persons), Dianne Cook, Australia, Heike Hoffman, USA, Lara Lusa, Slovenia, Carsten Oliver Schmidt, Germany, all of whom have experience working with observational studies requiring IDA.

The first task for TG3 was to develop a conceptual framework that emphasizes and clarifies the role of IDA in the research process. This framework was recently published [4]. The main aim of IDA is seen in providing reliable knowledge about the data to enable responsible statistical analyses and interpretation. IDA consists of all steps performed on the data of a study between the end of the data collection/entry and start of those statistical analyses that address research questions. So far, we focused on primary-data collections,

where data are obtained to address a predefined set of research questions, with an elaborated analysis plan. However, IDA is often performed in more complex studies raising additional issues such as an implementation of IDA processes during ongoing data collections to detect data issues while they are potentially remediable.

Our framework distinguished six IDA steps: First, the **Metadata setup** summarizes background information to properly conduct all following IDA steps. Beyond technical metadata such as labels or plausibility limits, this covers conceptual metadata which combines information from the study protocol, secondary information sources and information about the actual study conduct. Second, **Data cleaning** is performed to identify and correct technical data errors. Many errors may not be directly observed and a proper metadata setup is crucial to progress correctly and efficiently in this step. Third, **Data screening** examines data properties to inform decisions about the realizability of the intended analyses. In contrast to the data cleaning step, the focus is on data properties, not technical errors. However, data screening may reveal structural errors that occurred during the data collection process. Fourth, **Initial data reporting** documents all insights obtained from the previous steps to the research body. Fifth comes **Refining and updating the analysis plan** where adaptations of the analysis plan may be made to account for findings from the previous IDA steps. Finally, reporting IDA in research papers is necessary to ensure transparency regarding key findings and actions in the IDA steps that impacted the analysis or interpretation of results. This reporting step is based on the initial data reporting but clearly focused on the specific paper and what has been done, whereas the former provides a general overview of IDA findings and suggestions on ways to handle potential conflicts with the analysis plan.

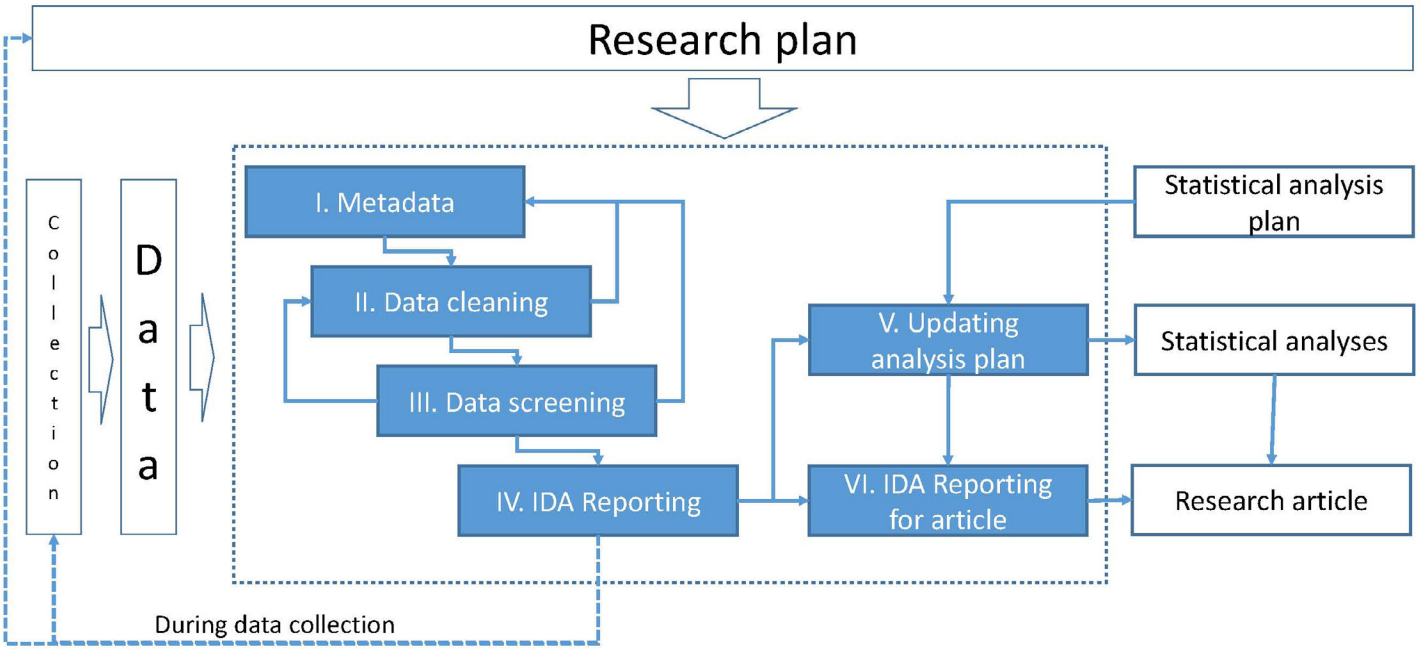
Figure 1 *(below)* illustrates the relationships among the IDA steps and external components. IDA steps may not necessarily take place in a linear manner and feedback loops may occur. Examples for each step are discussed in [4]. Note that conduct and handling of IDA steps 3 and 5 provide direct links to activities in the other STRATOS TGs.

There are risks associated with IDA, namely when analyses touch upon research questions of interest. These may influence the final analyses and conclusions in a non-transparent manner, and may increase the risk of false positive results [5]. Key principles for IDA are therefore to avoid touching the research question, and to provide full documentation of the process.

IDA requires an organizational framework, the proper assignment of responsibilities, identification of resources, as is the case for established work like data collection, statistical analysis, or writing. There is a need to agree on concepts and techniques in order to facilitate IDA and its reporting, including options for an automatization of IDA processes and recommendations for reporting IDA.

We aim to provide worked examples guiding through the steps of initial data analyses with selected datasets, reporting guidance for initial data analyses in manuscripts, and to examine datasets that may need more specialized strategies. We hope to stimulate discussions about statistical methods for IDA [3], software to facilitate IDA processes in complex studies [6], or modern graphical techniques [2], new methods addressing extensions of the IDA

Figure 1



framework, and are looking forward to interactions with other researchers interested in these topics.

Conceptual links exist to all other STRATOS Topic Groups, particularly to groups dealing with data deficiencies like TG1 (Missing Data), TG4 (Measurement error and misclassification), and TG2 (Selection of variables and functional forms in multivariable analysis), or TG9 when addressing large administrative data sets.

References:

[1] C. Chatfield, The Initial Examination of Data, *J R Stat Soc Ser Gen* **148** (1985), 214-253.

[2] D. Cook and D.F. Swayne, *Interactive and Dynamic Graphics for Data Analysis*., Springer, New York, 2007.

[3] T. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*, Wiley-Interscience, New York, 2003.

[4] M. Huebner, S. le Cessie, C.O. Schmidt, and W. Vach, A Contemporary Conceptual Framework for Initial Data Analysis, *Observational Studies* **4** (2018), 171-192.

[5] J.T. Leek and R.D. Peng, Statistics: P values are just the tip of the iceberg, *Nature* **520** (2015), 612.

[6] C.O. Schmidt, C. Krabbe, J. Schossow, M. Albers, D. Radke, and J. Henke, Square(2) - A Web Application for Data Monitoring in Epidemiological and Clinical Studies, *Studies in Health Technology and Informatics* **235** (2017), 549-553.

Solution to the Mathematical Riddle of Vol 35 1st Issue

The solution to the last issue's mathematical riddle was: 46. Note that in the last row the women do not hold a bag.

Only **Moshe Kelner** (Research Unit, Cellcom Israel) solved it correctly.

For Genius