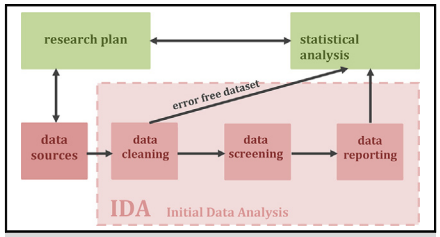
A systematic approach to initial data analysis is good research practice

Marianne Huebner, PhD,[a,b] Werner Vach, Dr rer. nat.,[b] and Saskia le Cessie, PhD[c]

ABSTRACT

Initial data analysis is conducted independently of the analysis needed to address the research questions. Shortcomings in these first steps may result in inappropriate statistical methods or incorrect conclusions. We outline a framework for initial data analysis and illustrate the impact of initial data analysis on research studies. Examples of reporting of initial data analysis in publications are given. A systematic and careful approach to initial data analysis is needed as good research practice. (J Thorac Cardiovasc Surg 2016;151:25-7)



The framework for initial data analysis as an integral part of the research process.

Central Message

We provide a conceptual framework for initial data analysis and good practice recommendations in surgical outcomes research.

Perspective

Initial data analysis is the process of data inspection steps to be carried out after the research plan and data collection have been finished but before formal statistical analyses. The purpose is to minimize the risk of incorrect or misleading results. Methods used and decisions made in the initial data analysis should be included in medical scientific reports.

Initial data analysis (IDA) has been described as part of "the rest of the iceberg that may sink science" in contrast with the much more often discussed *P* value as the "tip of the iceberg."[1] As much as 80% of the time allocated to the statistical analysis process is spent on data cleaning and preparation[2,3]; however, these first steps in the analysis of data are often neglected or disorganized,[4] and decisions made during these steps, such as changing the methods used or the outcomes measured, are often unreported.[5,6] In a study that compared protocols with published articles for randomized trials, at least 1 primary outcome was changed, introduced, or omitted in 62% of the studies examined.[6] Some of these changes may be due to the discovery of data properties that do not agree with the expectations or requirements of the analysis plan. To prevent false-positive results, a systematic approach to initial data analysis and its reporting is needed.

CONCEPTUAL FRAMEWORK FOR IDA

IDA is the process of data inspection and the screening steps in the analysis to be carried out after the research plan and data collection have been finished but before performing the formal statistical analyses. IDA is conducted independently of the analysis needed to address the research questions and does not include analyses that touch directly the research aims; this restriction prevents hypothesis generation leading to false-positive results. The purposes of IDA are to ensure that the later statistical analysis can be performed efficiently and to minimize the risk of incorrect or misleading results.

IDA can be divided into 3 main steps:

1. Data cleaning is the identification of inconsistencies in the data and the resolution of any such issues.
2. Data screening is the description of the data properties.
3. Documentation and reporting preserve the information for the later statistical analysis and models.

From the [a]Department of Statistics and Probability, Michigan State University, East Lansing, Mich; the [b]Institute for Medical Biometry and Statistics, Medical Center, University of Freiburg, Freiburg, Germany; and the [c]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands.

Received for publication Sept 5, 2015; accepted for publication Sept 22, 2015; available ahead of print Oct 23, 2015.

Address for reprints: Marianne Huebner, PhD, Department of Statistics and Probability, Michigan State University, 619 Red Cedar Rd, East Lansing, MI 48824 (E-mail: huebner@stt.msu.edu).

0022-5223/$36.00

http://dx.doi.org/10.1016/j.jtcvs.2015.09.085

Abbreviation and Acronym
IDA = initial data analysis

These steps should be seen as part of the overall research work flow (Figure 1) in addition to a well thought-out statistical analysis plan.

IMPACT OF IDA

It is tempting to start immediately with the statistical analyses, as described in the statistical analysis plan. Neglecting the initial data analysis, however, can result in lack of validity of a study.[7] When data errors or data properties inconsistent with the planned statistical models are discovered later, at the analysis stage or the manuscript writing stage, this leads to time-consuming and cost-inefficient rechecking of data, redoing analyses, and rewriting of the manuscript. The following is a list of issues that IDA may detect that show the possible importance of such detection:

- Duplicate records need to be eliminated to ensure correctness of the results.
- Coding 0 or 1 may be reversed for 2 related dichotomous variables, which would produce opposite associations than expected.
- Inconsistencies in date and time stamps for sequential measurements bear the risk of errors in time variables derived later as part of the analysis.
- Bimodal distributions may indicate an inconsistent use of measurement units.
- Skewed distribution of a variable may forbid the use of statistical methods that assume a symmetric distribution.
- Ceiling and floor effects have to be taken into account in interpreting effects different from those expected.
- Statistical models may be invalid and may yield misleading results in the presence of outliers.
- The pattern of missing data at different levels (eg, for a record, a variable, or a group) may indicate a systematic error in the data collection.
- Error rates and types for the main variables may indicate a problem in the data collection process and can inform future studies.

IDA cannot and should not take the place of error prevention[8]; however, some of the issues mentioned here can be avoided by a careful design of the database.

REPORTING IDA IN PUBLICATIONS

The main results of IDA should be reported if they have an impact on the analyses or the interpretation of the results. Some examples are given in the REMARK Guidelines.[9]

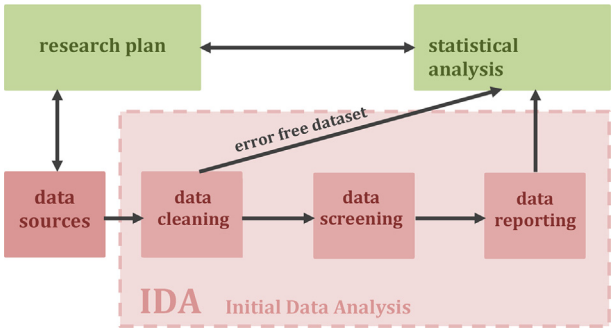


FIGURE 1. The framework for initial data analysis as an integral part of the research process. *IDA*, Initial data analysis.

1. Any changes in the original analysis plan that are based on insights obtained by IDA should be documented.
2. Any deletion of records or other changes in the data as a result of insights obtained by IDA should be reported.
3. Information on location, variation, and shape of the distribution of variables that are part of the research plan should be given.
4. The frequency of missing values in single variables used in the statistical analysis and the patterns of missing values in related variables should be described.
5. The criteria used to detect extreme values and the handling of these observations need to be explained.

Such checklists can also help investigators to decide on whether their data is ready for the statistical analyses planned in the research protocol or whether decisions about a modification of the planned statistical models have to be made. More detailed guidance for IDA is currently being developed as part of the Strengthening Analytical Thinking for Observational Studies (STRATOS) initiative.[10]

CONCLUSIONS

Careful IDA facilitates fast and smooth statistical analyses and investigation and reduces the risk of publishing incorrect results. Due diligence in data management and carrying out initial data analysis is the responsibility of each investigator. A research protocol should include a systematic plan and a budget for the initial data analysis. Methods used for the initial data analysis and the decisions made in and on the basis of the initial data analysis should be included in medical scientific reports.[11]

Conflict of Interest Statement

Authors have nothing to disclose with regard to commercial support.

References

1. Leek JT, Peng R. Statistics: *P* values are just the tip of the iceberg. *Nature*. 2015; 520:612.

2. Dasu T, Johnson T. *Exploratory data mining and data cleaning*. Hoboken (NJ): John Wiley & Sons; 2003.
3. Lohr S. For big-data scientists, 'janitor work' is key hurdle to insights. *The New York Times*. August 17, 2014. Available at: http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0. Accessed August 10, 2015.
4. Chatfield C. The initial examination of data. *J R Stat Soc Series A*. 1985;148:214-53.
5. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383:267-76.
6. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004;291:2457-65.
7. Osborne JW. *Best practices in data cleaning*. Thousand Oaks (CA): SAGE Publications; 2012.
8. Huber PJ. *Data analysis: what can be learned from the past 50 years*. Hoboken (NJ): John Wiley & Sons; 2011.
9. Altman DG, McShane LM, Sauerbrei W, Taube SE. Reporting recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. *PLoS Med*. 2012;9:e1001216.
10. Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J, STRATOS Initiative. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Stat Med*. 2014;33:5413-32.
11. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Med*. 2005;2:e267.

Key Words: initial data analysis, data cleaning, data screening