Observational Studies 4 (2018) 171-192 Submitted 7/17; Published 4/18

A Contemporary Conceptual Framework for Initial Data Analysis

Marianne Huebner **huebner@stt.msu.edu**
Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA

Saskia le Cessie **S.le Cessie@lumc.nl**
Department of Clinical Epidemiology and Department of Medical Statistics and Bioinformatics
Leiden University Medical Center
Leiden, The Netherlands

Carsten O. Schmidt **Carsten.schmidt@uni-greifswald.de**
Institute for Community Medicine, SHIP-KEF
University Medicine of Greifswald
Greifswald, Germany

Werner Vach **Werner.vach@usb.ch**
Department of Orthopaedics and Traumatology
University Hospital Basel
Basel, Switzerland

on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies, `http://www.stratos-initiative.org`). Membership of the Topic Group is provided in the Acknowledgments.

Abstract

Initial data analyses (IDA) are often performed as part of studies with primary-data collection, where data are obtained to address a predefined set of research questions, and with a clear plan of the intended statistical analyses. An informal or unstructured approach may have a large and non-transparent impact on results and conclusions presented in publications. Key principles for IDA are to avoid analyses that are part of the research question, and full documentation and transparency.

We develop a framework for IDA from the perspective of a study with primary-data collection and define and discuss six steps of IDA: (1) Metadata setup to properly conduct all following IDA steps, (2) Data cleaning to identify and correct data errors, (3) Data screening that consists of understanding the properties of the data, (4) Initial data reporting that informs all potential collaborators working with the data about insights, (5) Refining and updating the analysis plan to incorporate the relevant findings, (6) Reporting of IDA in research papers to document steps that impact the interpretation of results. We describe basic principles to be applied in each step and illustrate them by example.

Initial data analysis needs to be recognized as an important part and independent element of the research process. Lack of resources or organizational barriers can be obstacles to IDA. Further methodological developments are needed for IDA dealing with multi-purpose studies or increasingly complex data sets.

Keywords: Initial data analysis, data cleaning, data screening, reporting, metadata, research plan, STRATOS Initiative

1. Background

Scientists commonly perform initial data analyses (IDA) as part of studies with primary-data collection, where data are obtained to address a predefined set of research questions, and with a clear plan of the intended statistical analyses. However, there has not been a consensus about the elements of IDA. For some, IDA may be limited to data cleaning, others understand IDA as basic data summaries, and yet others perform analyses or visualization to discern relationships, which are then formalized in statistical models. Although there are articles listing important elements of IDA (6; 16), we lack a more general framework for IDA. How to perform IDA in a structured and strategic way needs further discussion (14). Furthermore, the problem of insufficient reporting of IDA steps is pervasive (6; 13). Insufficient reporting is especially problematic, if IDA has led to a change in the intended analytic strategy or in a rephrasing of the research question.

Of the different IDA elements data cleaning is well established and performed at high standards such as in areas following Good Clinical Practice (GCP) or Good Epidemiologic Practice (GEP) guidelines (20; 15). An example of IDA in high dimensional data analyses is the preprocessing of data (5; 13; 17). However, many approaches or views of IDA may also include hypotheses formulating as part of exploratory data analysis. Since this can have a non-transparent impact on results and conclusions presented in publications it has been pointed out that IDA should refrain from touching upon questions in the research plan (1), or that a pipeline of data analysis with data cleaning, data exploration, and modeling needs more debate (16). This is particularly important, since the conditions and scope for IDA have changed over the last decades. Data sets have grown in size and complexity, often including data from different sources. More and more studies are based on data collected not for research purposes such as administrative data. Electronic health records present additional challenges, since uniform data standards are not widely used (26) and such data are prone to documentation errors. Further methodological developments are needed for IDA addressing this changing landscape. We can harness a vast array of modern tools. Today we can create hundreds of histograms, boxplots and scatterplots in seconds, and many graphs can be viewed on a single screen. More sophisticated tools for data visualization exist (7) that may still be underused in IDA.

In this paper we focus on IDA from the perspective of a study with primary-data collection, collecting data to address a predefined set of research questions, and with a clear plan of the intended statistical analyses. However, this framework may apply to other scenarios. A major obstacle to perform a systematic initial data analysis is that it has not been recognized or appreciated as an important part and independent element of the research process. Thus lack of funding or time or other priorities can lead to an insufficient approach to IDA. The organizational set-up differs at different institutions. Some have a team of statistical analysts who routinely perform initial data analyses, at other places these are done by a single investigator who also handles the main statistical analyses.

Consequently, in this paper we develop a conceptually oriented, contemporary view on IDA. We describe the frame and scope of IDA, the necessary preparations, and basic steps

of IDA. We focus on the logic behind these steps: Why do we do what? We mention typical questions and describe common techniques for these. Additionally, we discuss issues related to the organizational frame of IDA and conclude with a discussion and outlook.

2. Aims and scope of initial data analysis

The aim of IDA is to provide a data set and reliable findings on this data set which allows researchers to work with this data set in a responsible manner. The latter requires a full awareness of all data properties needed for a correct analysis and interpretation of the data thus minimizing the risk of producing numerical results or interpretations which are misleading or incorrect. Consequently, IDA enriches the already existing "metadata of a study" to improve the basis for sound statistical analyses and interpretations. Examples are information obtained through IDA on the process of the data cleaning, exclusion of cases due to pre-specified criteria, unusual observations, creation of new or transformed variables, distributions, or patterns of missing data.

IDA typically takes place between the end of the data collection or data entry and the start of those statistical analyses that address the research questions. However, some initial data analyses can be performed simultaneously while data are being collected to detect and deal with data issues as early as possible.

Information based on IDA may impact the choice of analysis methods in the later analysis. Planned analyses methods may be recognized as inadequate after IDA due to data properties contradicting the application of the intended method. In some instances IDA may influence even the choice of research questions to be addressed, when available data turned out to be insufficient in number or variation to allow us to consider the model based on scientists' understanding of the field.

Here we assume that IDA has to be performed as part of a single study with a clear research plan and a single data collection or data entry process. IDA can be regarded as a process consisting of six basic steps, each of which is discussed in detail in the following sections.

I. Metadata setup is aimed at systematically setting up all background information required to properly conduct the following IDA steps.

II. Data cleaning is aimed at identifying and correcting data errors.

III. Data screening consists of reviewing and documenting the properties and quality of the data that may affect future analysis and interpretation.

IV. Initial data reporting aims at informing all potential collaborators about all relevant insights obtained from the previous steps. It should provide all necessary information to properly conduct the intended analyses.

V. Refining and updating the analysis plan translates the relevant findings from the previous IDA steps into corresponding adaptations of the analysis plan.

VI. Reporting IDA in research papers is a final step ensuring that all findings and actions from the previous steps that impact the interpretation of results are documented in the paper.

It is important to note that these steps may not necessarily take place in a linear manner. There may be feedback loops. For example, findings from data screening may motivate further data cleaning. These IDA steps interact with other parts of the overall research process part of a single study. Figure 1 illustrates the basic relations among the IDA steps and the key interactions to major external components in the context of IDA with a research plan in mind: the research plan, data collection, the statistical analysis plan, the research publication(s), the data itself, and all metadata. We will explain and comment on these interactions in the following sections.

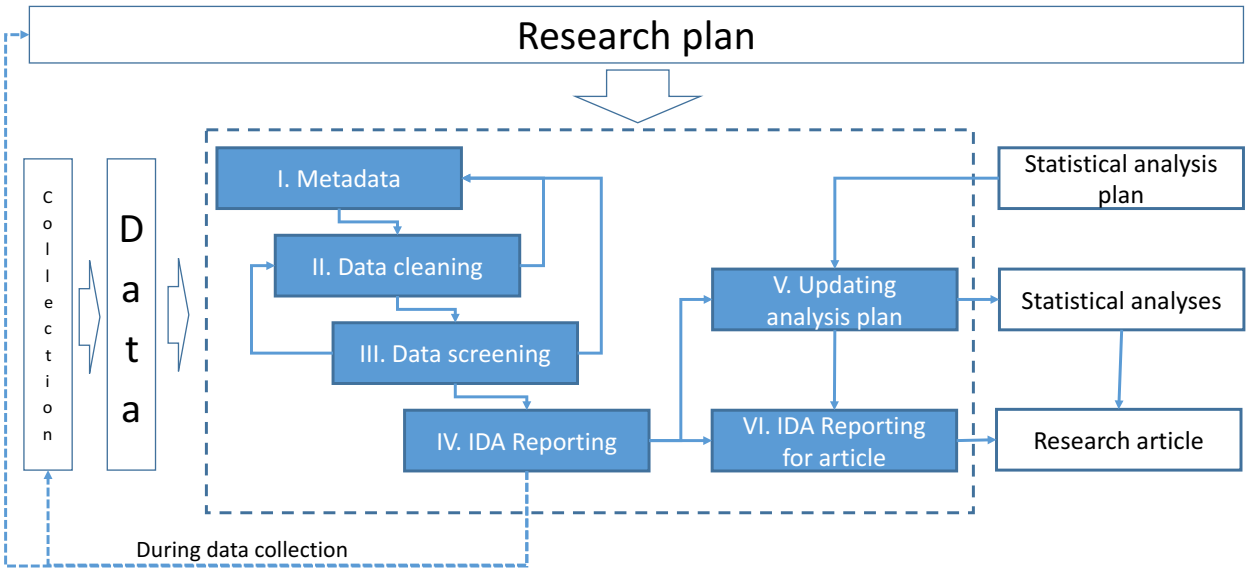

Figure 1: The main connections between the IDA steps and external components

IDA is a productive process, and it is important to have a clear picture about key output elements of the different steps, as summarized in Table 1, and explained further in section 3.

However, the process of IDA has the potential for misconduct . There is a risk that IDA may include analyses addressing the research questions of interest and that knowledge about the results influences the analyses to be performed and published later in a non-transparent manner. This may lead to false positive results. Thus IDA should, as much as possible, refrain from touching the research questions. This requires a good understanding of the research questions and the intended analyses. Further examples are discussed in section 4, the risks of IDA.

	The six IDA steps					
Type of output	I. Metadata Setup	II. Data cleaning	III. Data screening	IV. Initial data reporting	V. Refining & updating the analysis plan	VI. Reporting in research papers
Analysis plan					Updated analysis plan	
Dataset		Cleaned dataset(s) with all original and derived variables			Analysis dataset to be used in final analyses	
Technical metadata / Data documentation	Comprehensive metadata; Comprehensive data dictionary	Updated data dictionary			Updated data dictionary	
Documentation	Documentation of all metadata aspects for IDA including technical and contextual metadata.	Document/ code of all data manipulations and conducted data cleaning activities	Document/ code of all conducted data screening activities	Summary of all findings from I-III with regards to their importance for subsequent analyses	Document/ code on suggested, discussed and accepted changes in the analysis plan	IDA findings influencing the interpretation of results included in Methods, Results, or Discussion of manuscript

Table 1: Key output elements of the six IDA steps

IDA presented so far can be too narrow. For example, in multi-purposes studies with no clear research questions and analysis plan, studies reusing existing data sets, or continuously growing administrative data sets, these situations will be discussed in section 5.

3. IDA Steps

3.1 Metadata setup

A data set alone is not sufficient to perform meaningful analyses. We need background information to guide IDA activities. Often, this information is referred to as metadata. The National Information Standards Organization defines metadata as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (10). Metadata of relevance should be set up in a structured systematic way to make sure that all elements of interest have been covered. In general, one should distinguish between the two types of metadata as follows.

3.1.1 Technical metadata

Technical metadata comprises information such as variable and value labels, data type, measurement unit, plausibility limits, codes for missing values or permitted jumps. It is typically associated with single study variables. For example, the original phrasing of a question in a questionnaire carries information not covered by the variable name and might be stored as a variable label. Technical metadata may also relate to examiners, devices or examination centers. This type of metadata is essential to perform IDA aspects related to data cleaning and data screening and should be stored electronically as part of data dictionaries.

3.1.2 Conceptual metadata

Metadata may also be related to the background of a data collection and its implementation such as information contained within the study protocol related to design (e.g., data sources, data collection methods), or recruitment (e.g., target population, inclusion and exclusion criteria, sampling methods, time of data collection). In addition, background information regarding the study (e.g., validation status for instruments, training of examiners) is of importance. Beyond technical documentations, feedback from colleagues with experience in the data collection or handling process may be important to consider. For example, when using routine data, feedback from clinical staff may be crucial to understand certain aspects of disease codings in a given outpatient setting. Knowing the original aim for collecting a variable and the actual mechanism of collecting data is relevant for data cleaning and data screening steps. Conceptual metadata is not only associated with single variables but with the entire examination of the dataset or the entire study.

Preferably, all relevant metadata should be set up in electronic form before the start of a study. This is crucial in order to support IDA processes and to control the entire data collection. For technical metadata, tools like REDCap (Research Electronic Data Capture) (12) are available. This part of metadata should be set up in a structured electronic form before the start of IDA processes. Tools like the memisc package (9) in the statistical software R (22) support this. However, the full potential of metadata may be hard to handle outside appropriate databases such as MySQL. Metadata, which cannot be handled as part of a data dictionary in a given study environment should be set up in an electronic document to ensure a systematic overview on metadata of relevance for IDA.

3.2 Data cleaning

In this step the raw data sources are processed to arrive at a cleaned dataset in the format needed for the analysis. This may include the definition of new variables, changing between wide and long formats, and merging or splitting of data sets. Data cleaning should be seen as a systematic attempt to find data errors and, if possible, to correct them. Data cleaning processes may discover errors highlighting impossible values or technically deficient data representations. The concept of an error implies that we imagine a virtual second data set including the "true" values, such that detecting a difference should force us to correct our data. In Table 2 we point to some typical sources for error detection.

It is obvious from the examples in Table 2 that the level of evidence for an error may vary. The fundamental problem with many data errors in a data set is the lack of direct

Source for error detection	Examples
Inconsistencies between observed data values and the formal frame given by the data structure or metadata	• Impossible codes for categorical variables • Values outside the allowed range for continuous variables • Date outside the time frame of the study • Unexpected missing values or missing patterns • Number of repeated observations for a subject differs from planned number • Consecutive dates which are out of order • Unreasonable time differences
Inconsistencies in the reporting of values for a single variable	• Large gaps in the distribution of a continuous variable indicating a technical error, e.g. change of the measurement unit in the data collection process or a use of different systems of measurements (e.g. metric vs. English, or a different lab assay) • Outliers explainable by entry errors (e.g. switches in labels) • Inconsistent use of date formats • Inconsistent use of upper and lower case letters, or typographical variants of the same term
Logical inconsistencies between variables	• Incompatible pairs of values (e.g. pregnant men) • Inconsistency between detailed information and overall information (e.g. the sum of reported durations of disjoint time periods should not exceed a reported total) • Incompatibility of a series of reported values (e.g. the sum of reported durations of daily activities should not exceed 24 hours)
Indications of typical mistakes in the data collection or entry	• Duplicate records (indicating double entry) • Partial duplicates (indicating unintended copy and paste) • A reversal of digits in non-matching key variables.

Table 2: Typical sources of errors in data

observability. They can be detected only indirectly by making adequate use of available metadata. The key to such indirect detections are inconsistencies, typically a deviation of an observed value (or a set of values) from some expectation based on metadata (e.g., the value range of a variable) or other observed data. In some cases it is possible to use external

information (e.g., original paper forms) to substantiate the level of evidence for an error. In other cases, the situation is more difficult. For example, a subject may tell the interviewer to have a paid job but no income. Any decision on how to deal with this inconsistency is likely to contain some element of uncertainty.

Technically, data cleaning is performed by a careful inspection of the data using simple techniques like tabulations, cross tabulations, histograms, or scatter plots. Many of the inconsistencies can be checked automatically at the time of data entry with a careful database design. Using well established coding schemes or formats like the ISO 8601 date and time formats supports consistency across multiple information sources. The more effort is invested on this, the faster will be the data cleaning performed as part of IDA. Ideally there should be a body separate from the person(s) conducting IDA to decide on changes in the data. Changing data is likely to be an iterative process, if information sources are accessed or additional enquiries are made.

The output of the data cleaning process is a "cleaned" dataset free of errors according to the scope of the data cleaning activities, supplemented with a record of all decision steps and all subsequent data corrections. All error flags which could not be clarified should be documented, as this may have an impact on the data screening discussed in the next section as well as the final interpretation of results. The original raw data must never be changed and the corrections, preferable incorporated in a programming code, should only be realized in a new "cleaned" dataset so that any changes are reproducible and reversible (1).

The data cleaning step will not only focus on the originally measured variables but also on all additionally derived variables which are of potential importance for later data screening and scientific analyses. Examples are the calculation of the follow-up time from a baseline and follow-up measurement, or the calculation of a body mass index from height and weight measurements.

3.3 Data screening

Data screening consists of all attempts to understand those properties of the data that may affect future analysis and interpretation. It involves analyses without touching the research hypothesis, and is needed to check whether explicit or implicit expectations about certain properties of the data are met such that the intended analysis of the study is applicable and can yield convincing results. For example, in a study investigating the effect of some type of risky behavior we typically aim to include subjects with low, medium, and high risk. If the number of subjects in one of the categories is very low the study may allow only limited conclusions. The reason for such a deviation may be related to the recruitment to the study, or to the way the risky behavior is measured, or to sources and algorithms used to collect the data. Classical examples consist of the check of randomization by comparing baseline characteristics between intervention groups or the comparison of characteristics of the study population with those for a background population. More generally, this is an attempt to detect "structural errors" in the data which point to problems in recruitment process, choice of instruments, or other aspects of study planning and conduct, that may result in a deviation between expected and observed properties of the data. Furthermore, data screening is needed to check whether certain properties of the data may be potential threats to correct application of statistical methods or adequate interpretation of results.

A typical example is the distribution of a variable. If a variable has a skewed distribution statistical methods not relying on the assumption of a normal distribution may be preferred and observed associations to other variables may depend on few observations, and should be interpreted with care. If missing values are observed in a variable, conclusions of an analysis restricted to subjects with complete data may not be valid for the whole study population. Moreover, the occurrence of missing data could indicate the need to use specific statistical methods to handle this incompleteness. The inspection of the distribution of all single variables is a basic step of data screening, but many more aspects of the data are important. In Table 3 we comment on relevant aspects, the aims of data screening when considering these aspects and points to be considered or techniques to be used.

Aim	Points to be addressed/methods
Distribution of single variables	
• Identifying qualitative properties such as outliers, bimodality or skewness • Comparing quantitative characteristics like sample means, percentiles or fractions below or above selected thresholds with expectations • Informed choice of methods to handle missing data • Insights into participants' behavior and attitudes • Insights about the data collection process	• Visual inspection by boxplots, histograms, bar charts, probability plots etc. • Computation of sample characteristics
Missing Data	
• Informed choice of methods to handle missing data • Insights into participants' behavior and attitudes • Insights about the data collection process	• Types of missing values (e.g. active non-response, not applicable items, technical difficulties, loss of existing data) • Level of occurrence (e.g. within a single variable, for all variables within an individual, or for all individuals within certain strata) • Missing frequencies per variable, missing patterns, associations across variables • Association of the occurrence with individual characteristics, higher level (e.g. center) characteristics, or with previous measurements (in the case of longitudinal data)
Association between variables	

Aim	Points to be addressed/methods
• Comparison with expectations: – Higher association than expected may hinder the intended adjustment for confounding – Lower association than expected may hint to factors measured with insufficient precision • Developing ideas for summary variables to reduce dimensionality	• Visualization and quantification of pairwise associations • Multivariate methods like factor analyses, latent class analyses, redundancy analyses, or log-linear models
Individual trajectories in longitudinal data	
• Understanding variability of individual development • Checking the adequateness of summary measures (e.g. slope) to address research questions • Informing modeling of the data	• Visual inspection of individual trajectories • Exploratory modeling of trajectories and drop out
Levels of measurements	
• Understanding deviations from uniform measurement conditions, e.g. related to centers, observers, treatment providers, place of residence, time of day or day of week • Informing interpretation and model choice	• Description of corresponding differences • Quantification via random effect models
Measurement error	

Aim	Points to be addressed/methods
• Assessing magnitude in order to inform about the well understood impact on interpretation of results (Buonaccorsi, 2010; Keogh & White, 2014)	• Often hard to assess due to lack of exact repeated measurements • Indirect insights possible due to unexpected low correlations or frequent inconsistencies • Use of surrogate repetitions, e.g. pre- and post-intervention assessments of outcomes in a placebo or usual care control group

Table 3: Data screening: topics, aims, and methods.

3.4 Documentation and initial data reporting

All previous steps are parts of an elaborate and comprehensive process. Each step may consist of many single analyses and decisions, reflecting a long process of working with the data. We may generate a huge amount of information, and consequently there is some danger of losing pertinent facts or important insights. In Section 2 we mention that full documentation of all activities in the first three steps and updating the data dictionary are key output elements of IDA. However, it is possible to get lost in the full documentation. Hence a specific initial data report is useful and necessary to allow researchers to work with this data set in a responsible manner. We suggest the following six elements to be included in such a report:

1. A short summary of the meta data set up step. This should focus on the sources used to build up the meta data and a summary of the research questions and the study design.

2. A detailed flow chart of the study. This starts with all criteria defining the study population and/or the recruitment process, reporting the number of subjects originally approached and the numbers remained after subjects are excluded due to inclusion/exclusion criteria, impossible values, missing data or other reasons. Often these steps can be visualized in a flow diagram.

3. A short summary of the data processing and cleaning step. This should start with a verbal summary of the main issues identified. In addition, a list of all explicitly applied correction rules should be provided as well as a summary making the impact of suggested/conducted changes transparent (e.g. proportion of data elements /observations etc. affected by some type of error and changed/unchanged)

4. A short summary of the data screening step. In addition to a short summary of the main issues identified, a basic step is a description of the distribution of all variables for the study population and important subpopulations. Then it should be feasible

to evaluate whether the study population resembles the intended source populations. If external information on the source populations is known, it is helpful to add this information to the description of the study population. Depending on the issues identified, further tables with summary information can be useful, e.g., about missing frequencies, about study design factors (observers, devices, centers) or multivariable descriptions like average trajectories.

5. A summary of all insights which may influence the interpretation of the results. This would include findings showing a deviation between expected and observed properties of the data, e.g. structural errors in the data that can point to a problem in the recruitment process.

6. A summary of all insights which may influence the further statistical analysis. These are data properties not in accordance with requirements of the intended statistical methods of the study, e.g. unexpected sizes of subgroups, or unexpected variable distributions.

Each short summary should be accompanied by an overview about all steps performed and an instruction how to read the full documentation of the step. Some elements such as the study flow chart, description of missing values, etc., may be used directly in the research papers. The insights which may influence the further statistical analysis should be included in the discussion of the research papers, in particular with respect to limitations of the study. It is crucial, as it prepares the interaction with the final statistical analysis. This step is discussed further in the next section.

3.5 Refining and updating the statistical analysis plan

One important aspect of IDA is to avoid misleading statistical analyses due to ignorance about properties of the data. Therefore, in the process of IDA analysts should be aware of the statistical analysis plan to ensure that all variables used in the projected analysis are thoroughly checked for errors in the cleaning phase and are reported in the screening phase. In the initial data reporting step results of the cleaning and screening steps have already been reviewed in light of the intended statistical analysis. Subsequently, it may be required that the statistical analysis plan is refined, updated, or extended. Table 4 outlines typical topics which may appear in this context.

Topic	Examples	Possible influence on statistical analysis plan
Suspicious values	• Suspicious outliers • Inconsistent follow-up dates • Different sources yield inconsistent information about a diagnosis	Decision on a rule whether and how to use this inconsistent or suspicious information in the statistical analysis
Suspicious subjects/rows	• Doubts whether subject fulfills inclusion criteria of the study. • Data of a subject looks so corrupt that it may add more noise than information	Decision on exclusion of subjects - balancing the potential benefit of having a "cleaner" data set against a potential selection bias
Unexpected heterogeneity of the study population	• Confusing information about the education of immigrants due to incompatibility of education systems • Heterogeneity between study centers suggesting a misunderstanding of the measurement instructions	Decision on exclusion of subjects - balancing the potential benefit of having a "cleaner" data set against a potential selection bias
Distribution of a variable	• Covariates with (unexpected) skewed distributions in regression analysis • Bimodal distribution for continuous variables	Transformation of variables; splitting bimodal variables into a binary variable to improve interpretability of results; adaption of statistical methods
Data properties not in accordance with requirements of intended statistical methods	• Unexpected skewness in the distribution of an outcome variable • Some subgroups are much smaller than expected and asymptotic tests may not be appropriate	Refinement, extensions or reduction of models

Table 4: Topics which may influence the statistical analysis plan

The examples presented in Table 4 illustrate that there can be reasons for changing the statistical analysis plan after a careful and comprehensive IDA. However, this should not mask the fact that changing the analysis plan is problematic. It bears the unintentional risk

to "optimize" the findings and reviewers and readers will be sceptic of, possibly extensive, changes. Transparency is key.

If questions about adequate analysis for the research questions remain after the update of the statistical analysis plan based on data, it might be necessary to apply different strategies. However, such analyses are beyond IDA and should be reported accordingly as part of the main analyses, for example as a sensitivity analysis.

3.6 Reporting IDA in research papers

It is important that all relevant findings from IDA and all relevant decisions made based on IDA are reported in a research paper. Reporting guidelines exist for different types of studies, such as the CONSORT statement for randomized trials (18), the STROBE statement for observational studies (28), or the REMARK recommendations for prognostic studies (3). While the initial data report should include a list of all activities and findings, reporting IDA in research papers focuses on findings of relevance for the interpretation of results. Actual changes made in the analysis plan with respect to the analyses presented should be mentioned briefly in the main part of the paper, as well as any changes in the choice or definitions of outcome variables or variables used in the main analysis. Motivation for these changes should be given. The section discussing limitations of a study is often a natural place for this. More extensive IDA reports could be included in online supplements. It might be appropriate to classify the findings with respect to their degree of relevance for the interpretation of the results presented in the current paper or in the supplementary material.

4. Risks of IDA

Because IDA produces new information, it includes elements with a potential for misconduct. IDA may lead to unjustified removal of "disturbing" observations, to data driven hypotheses, to nontransparent changes in the statistical analysis plan, or to "optimizing" of analysis strategies similar to trial and error approaches. "Data dredging" and "Data snooping" are names for such behavior we have to avoid in IDA (27). IDA should not be misused in this step to sanctify poor research practice. Potential examples of misuse can be the optimization of cut-points or the categorization of continuous variables without any need (2; 19).

Data cleaning may result in unjustified changes, in particular masking poor data quality. To address this, the data cleaning step (step II) should clearly distinguish between a phase where potential errors are flagged and suggestions for changes are generated, and a second phase with decisions on handling these proposals. It should be clear who had the mandate to accept or perform such changes, preferable an independent body. Secondly, there is a risk that IDA may include analyses addressing the research questions of interest and results of IDA influence the analyses to be performed and published later in a non-transparent manner. This may lead to biased results and interpretations. To address this issue a good practice would be to explicitly exclude any analysis that touches the research question of interest as much as possible. However, not touching the research question may occasionally be difficult. For example, if the research question is about the distribution of a time to a

certain event, it is hard to identify suspicious outliers without looking at the shape of the event time distribution.

One important aim of IDA is to inform the later statistical analyses about unexpected and undesirable data properties, which may suggest to change the analysis plan. Allowing such changes may be misused. Hence transparency and complete documentation are essential principles to aim for. To achieve this, IDA should be seen as a structured process with clear rules for allowed and intended actions and their documentation. In conclusion, we have to be conservative in defining the scope of the core IDA. To assure this we suggest to explicitly exclude certain analyses a priori to avoid biased scientific analyses.

5. Extensions of the IDA framework

We have focused on IDA from the perspective of a study with primary-data collection, collecting data to address a predefined set of research questions, and with a clear plan of the intended statistical analyses. IDA may apply to other scenarios with the same steps outlined earlier. However some aspects of IDA may have to be revised or adapted to these more complex situations.

5.1 Multi-purpose studies

More and more with studies lacking a clear frame of predefined research questions are appearing. Large-scale cohort studies, local clinical, or nationwide/global epidemiological registries are based on planned data collection, but intended to be used for a large variety of research questions which may not have been specified at the time of data collection. This interferes with the principle of IDA of not touching the research question. It may be possible to define potential types of future research questions at the time of database set-up. Another approach might be to keep the IDA report confidential and to make it accessible only in part whenever a concrete project starts to avoid potential bias. An advantage of large scale multi-purpose studies is the likely separation of personnel responsible for IDA versus for scientific analyses. This separation reduces the risk of tailoring IDA to meet intended scientific outcomes.

5.2 Reusing existing data

Instead of collecting all data, a study may use existing data sources, for example data from previous research studies, or data collected for administrative purposes, like medical records, pharmacy, laboratory, or clinical notes. "Data-repurposing" (BD2K Home Page — Data Science at NIH) (NIH: Big Data to Knowledge program) refers to the reuse of such data to solve research questions. These data may have already been subjected to some type of IDA activities, and we may rely partially on the results of these activities and include them in the meta information.

5.3 Continuously growing data

Clinical or epidemiological registers are examples of data sets which are continuously growing and continuously used for research purposes. Consequently, it becomes impossible to wait with IDA until the data collection is finished; it has to be performed continuously.

This implies that we have to redo at least step II to step IV in fixed intervals, preferably reusing the procedures and knowledge from the previous analyses. Appropriate techniques and procedures have to be defined to ensure this in an efficient and reliable manner. For example, a process could be such that incrementally checking only new observations for data cleaning, while the screening approach is carried out on all observations.

5.4 IDA as part of data quality monitoring

Any study with primary-data collection should aim to detect data issues as early as possible during data collection to reduce or avoid data problems instead of optimizing ways to deal with them during the final statistical analyses (24). Therefore, it would be reasonable to start IDA already during data collection, as the output of IDA can then serve as a feedback to improve the ongoing data collection and the ongoing research process. In particular, large scale epidemiological studies use IDA as part of a data quality monitoring process. Despite the obvious importance of such approaches, a comprehensive data monitoring commonly does not receive the attention it deserves (11). The first implication for IDA in this context is that it has to start as early as possible and should continue with the growth of the study. This may result in subtle changes in the focus of IDA, since optimizing the data collection is now the main interest. For example, traditionally IDA informs researchers only about the size of measurement errors in the data after data collection had stopped, here we can define acceptable magnitudes and initiate actions to improve data quality. IDA may now also initiate specific sub-studies to clarify data issues, e.g. validation studies for measurement procedures.

6. Organizational frame of IDA and related issues

6.1 IDA team

IDA requires different types of expertise. Statistical competence is necessary to select adequate tools for the different steps, to assist in the interpretation and to judge the potential impact on statistical issues of the later analysis. Domain knowledge is necessary to formulate expectations (which preferably should be justified by reference to or generation of systematic reviews) and to judge the potential impact on conceptual issues in later analyses. Familiarity with the data collection procedure is required in order to relate IDA findings to data collection. We can rarely expect to find all this expertise in one person, so it is natural to regard IDA as a team task. In larger projects each of the three types of expertise may involve several people.

At the technical level most IDA tasks do not require highly sophisticated skills, so it is natural to organize such a team in the way that one person is actually performing all analyses, whereas all other team members contribute by planning, supervising and discussing the work of this person.

With respect to data cleaning, we suggest two distinct phases, where the IDA team suggests data changes to a body which has the mandate to accept or perform changes. In practice, the IDA team may overlap or be identical to this body, which will add responsibilities for the team. Similar, there may be an overlap between the IDA team and the

scientists responsible for the statistical analysis. This requires increased care to draw the line between IDA and the main analysis when refining and updating the analysis plan.

6.2 The limits of manual inspection

We have described IDA as a process where the different steps of inspection and reporting can be performed manually. However, the advances in automated or semi-automated data collection, multiple examinations from large scale studies, the increasing possibilities of data linkage allowing to incorporate existing data from different sources, and the many sources of high dimensional data in the biomedical sciences characterized by large number of observations or variables per subject in a single project today are more the rule than the exception. Then the paradigm of IDA as a manual process breaks down. Already in the case of 100 variables an analysis of all pairwise associations would require to inspect 4950 scatter plots. In the future automated or semi-automated procedures for data inspection are desirable. Pipelines for preprocessing high dimensional data are one example of such approaches. For an example, see (4).

6.3 IDA costs

If we take IDA serious, it requires a lot of time, effort, and resources. Some estimate that it can take up to 80-90% of the total analysis time (8; 16). It can be expensive, and it may be more expensive than the main analysis. Consequently, the performance of IDA should be a major part of the budget in any research proposal, and funding agencies should be prepared to accept such a budget plan.

7. Discussion and outlook

There are numerous online reports or blog posts discussing certain aspects of IDA, reflecting personal experience or experience in a specific field. These may get reinvented as others are faced with the task of IDA. However, these reports do not give a comprehensive view of IDA and its role in the research process, and lack a clear definition of the aim of IDA. In this paper we suggest a framework for IDA that encompasses six conceptually different steps. Overlap in the analysis tasks may be a matter of debate:

- Data cleaning and data screening, based on a systematic inspection of data, can refer to similar analytical steps. Some tasks will be performed only once, but their interpretation can be part of data cleaning, for example record verification, as well as part of data screening when describing properties of variables.

- The initial data reporting is a summary of the metadata, data cleaning and data screening steps. However, a systematic reporting from these steps is often ignored (Chatfield, 1985; Huber, 2011), and this constitutes a source for producing waste or possibly misconduct. The summary of the insights serves as a basis to reach full transparency in the refining and updating of the analysis plan (step V) and the reporting of IDA (step VI). Hence it is reasonable to regard initial data reporting as a step on its own.

- "Refining and updating the analysis plan" implies crucial decisions about where IDA actually ends. Current widespread practice is that IDA stops with the initial data reporting and any further decisions are made by the scientists performing the statistical analyses. However, informal decision making can lead to negative consequences regarding the reliability and validity of research results. Hence we feel it is an advantage to include this step explicitly in IDA.

We have touched the question of concrete statistical methods to be used in IDA only briefly. However, it is clear that the large tool box of exploratory data analysis (EDA) (8), also in form of modern graphical techniques (7), will play a major role. EDA is often used as an explorative, hypothesis generating approach, whereas IDA is aiming to clarify the conditions for answering pre-specified research questions. It is beyond the scope of this paper to discuss the available tools in detail and provide guidance and checklists for the separate steps. It is the aim of the STRATOS Initiative (Strengthening Analytical Thinking for Observational Studies) to develop such guidance for IDA in the future (23).

In this paper we developed a contemporary view on IDA by focusing on practical challenges that are part of analysts' work today. The rapidly growing sizes of datasets shows the need to use automated and semi-automated techniques in a structured way to allow performing IDA in an efficient and fast manner, often as part of a routine to ensure validity of growing data sets (25). IDA may also deal with tasks to handle not only numerical or categorical variables and time stamps, but also more complex structures, requiring the use of natural language processing techniques or use of methods for knowledge representation and processing. IDA may in future also have an impact on the design of standard data base system, for example by allowing to attach additional information to single observations about their trustworthiness. Due to their experience and skills, statisticians will often have a leading role ensuring that IDA contributes productively to the research process, and will contribute to the development of new methodology to support IDA.

Acknowledgments

This work was developed as part of the international initiative of Strengthening Analytical Thinking for Observational Studies (STRATOS). The objective of STRATOS is to provide accessible and accurate guidance in the design and analysis of observational studies (`http://stratos-initiative.org/`). Members of the Topic Group *Initial Data Analysis* of the STRATOS Initiative are Dianne Cook (Australia), Heike Hoffman (USA), Marianne Huebner (USA), Saskia le Cessie (Netherlands), Lara Lusa (Slovenia), Carsten Oliver Schmidt (Germany), Werner Vach (Switzerland).

Work on this paper was supported by the German Research Foundation (DFG, SCHM 2744/3-1)

References

[1] Ader, H. and Mellenbergh, G. (2008). *Advising on research methods: a consultant's companion*. Johannes van Kessel Publishing, 1st edition.

[2] Altman, D. G., Lausen, B., Sauerbrei, W., and Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86(11):829–835.

[3] Altman, D. G., McShane, L. M., Sauerbrei, W., and Taube, S. E. (2012). Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *PLoS Medicine*.

[4] Altmann, A., Weber, P., Bader, D., Preuss, M., Binder, E. B., and Müller-Myhsok, B. (2012). A beginners guide to snp calling from high-throughput dna-sequencing data. *Human Genetics*, 131(10):1541–1554.

[5] Baggerly, K. A. and Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Stat.*, 4(1309–1334).

[6] Chatfield, C. (1985). The initial examination of data. *Journal of the Royal Statistical Society. Series A (General)*, 148(3).

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

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

[9] Elff, M. (2017). memisc: Management of survey data and presentation of analysis results. Available from: `https://CRAN.R-project.org/package=memisc`.

[10] Guenther, R. and Radebaugh, J. (2004). *Understanding metadata*. National Information Standards Organization, NISO Press, Bethesda, MD.

[11] Harel, O., Schisterman, E. F., Vexler, A., and Ruopp, M. D. (2008). Monitoring quality control: can we get better data? *Epidemiology*, 19(4):621–627.

[12] Harris, P., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. (2009). Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*, 42(2):377–81.

[13] Huber, P. (2011). *Data Analysis: What Can Be Learned From the Past 50 Years*. Wiley, Hoboken, N.J.

[14] Huebner, M., Vach, W., and le Cessie, S. (2016). A systematic approach to initial data analysis is good research practice. *The Journal of Thoracic and Cardiovascular Surgery*, 151(1):25–27.

[15] International Epidemiological Association (IEA) (2007). Good Epidemiological Practice (GEP). Available from: http://ieaweb.org/good-epidemiological-practice-gep/ [cited October 24, 2017].

[16] Leek, J. T. and Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature*, 520(7549):612.

[17] McShane, L. M., Cavenagh, M. M., Lively, T. G., Eberhard, D. A., Bigbee, W. L., Williams, P. M., J.P., M., Polley, M.-Y. C., Kim, K. Y., Tricoli, J., Taylor, J. M. G., Shuman, D. J., Simon, R. M., Doroshow, J. H., and Conley, B. A. (2013). Criteria for the use of omics-based predictors in clinical trials. *Nature*, 502(7471):317–320.

[18] Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., and Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340:c869.

[19] Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., and Altman, D. G. (2011). Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, 32(3):437–440.

[20] National Institute on Drug Abuse (NIDA), Center for Clinical Trials (CCTN), and Clinical Trials Network (CTN) (2005). Good clinical practice. Available from: https://gcp.nidatraining.org [cited October 24, 12017].

[NIH: Big Data to Knowledge program] NIH: Big Data to Knowledge program. BD2K Home Page — Data Science at NIH [online]. Available from: https://datascience.nih.gov/bd2k.

[22] R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

[23] Sauerbrei, W., Abrahamowicz, M., Altman, D. G., le Cessie, S., Carpenter, J., and initiative, S. (2014). Strengthening analytical thinking for observational studies: the stratos initiative. *Statistics in Medicine*, 33(30):5413–5432.

[24] Schmidt, C. (2014). *Implementation recommendations [for data quality in cohorts]*, pages 117–127. Data quality in the medical sciences: Guideline for the adaptive management in cohorts and registries. TMF e.V., Berlin.

[25] Schmidt, C. O., Krabbe, C., Schoessow, J., Albers, M., Radke, D., and Henke, J. (2017). Square2 - a web application for data monitoring in epidemiological and clinical studies - semantic scholar. *Technology and Informatics*, 235:549–553.

[26] Schneeweiss, S. (2014). Learning from big health care data. *New England Journal of Medicine*, 370(23):2161–2163.

[27] Smith, G. D. and Ebrahim, S. (2002). Data dredging, bias, or confounding. *BMJ*, 325(7378):1437–1438.

[28] Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J., Egger, M., and Initiative, S. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Epidemiology*, 18(6):805–835.