THE FUTURE OF INFORMATION SCIENCES

INFUTURE2019

KNOWLEDGE IN THE DIGITAL AGE

Edited by

Petra Bago, Ivana Hebrang Grgić, Tomislav Ivanjko,
Vedran Juričić, Željka Mikulošević and Helena Stublić

Zagreb, November 2019

https://doi.org/10.17234/INFUTURE.2019.12 Preliminary communication

Data Quality in the Context of Longitudinal Research Studies

Tonko Carić
Institute for Anthropological Research, Zagreb, Croatia
tcaric@inantro.hr

Kristina Kocijan
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
krkocijan@ffzg.hr

Summary
This paper discusses the concept of data quality in the context of longitudinal research. By deconstructing quality assurance process and data collection strategies through a case study of the "Croatian Birth Cohort Study", we try to define causes and sources of poor data quality in the context of longitudinal studies. Besides the problems discussed throughout the known literature (panel conditioning, sample attrition, recall bias, temporal and financial demands), we introduce single-source problems, multi-source problems, security problems, design questionnaire problems and QA workflow problems as important aspects in the domain of the possible sources of errors. Additionaly we propose models for eliminating the errors through prevention and detection in order to improve data quality

Key words: data quality, quality assurance, data collection, research data, longitudinal study

Introduction

Data may be defined as a representation of facts or concepts or instructions in a formalized manner, suitable for communication, interpretation or processing by manual or electronic means. Tayi and Ballou (1998) define data as a "raw material for the information age". An element of data is an item, idea, concept or raw fact (Abdelhak et al., 1996 as cited in World Health Organization, 2003).
Information is "useful data" that is processed by the end-user in such a way that the information received is manifested as "knowledge (McFadden et al., 1998). In the literature about research data, the term "information" is often used interchangeably with the term "data". In the context of research studies, data can also be referred as „population-based data" (Chen et al., 2014) and such data go through the processes of collection, storage, processing and compilation.
A longitudinal study is a research design that involves repeated observations of the same variables (e.g., people) over short or long periods of time (Young et al., 2007). Longitudinal studies share many similarities with transversal studies, while differences do exist. Key benefits of collecting data through longitudinal studies include analysis advantages and measures of stability or instability. Moreover, longitudinal surveys can help understand causality - only the longitudinal survey can provide information about cumulative phenomena, following changes over time in particular individuals within the cohort (Young et al., 2007).
This paper intends to define possible causes and sources of poor data quality particularly in the context of longitudinal studies. The structure of the paper will flow from the introduction to quantitative data collection, total survey error and data collection to the CRIBS project that was used as our case study, including the steps for its data collection and quality assurance, detecting sources of errors and eliminating the errors. The paper will conclude with some main discussion points.

Quantitative data collection

Quantitative data collection methods rely on random sampling and structured data collection instruments that fit diverse experiences into predetermined response categories. Primary longitudinal data can be collected by direct observation (e.g., interviews, field observation), survey (e.g., personal structured interview, mail questionnaires, telephone surveys, diaries), tests and instruments or retrospective measures (e.g., investigation of archived documentation, interviews) (Leedy, Ormrod,

2001). This paper puts focus on the usage of primary longitudinal data for longitudinal data collected through a survey.

There are several ways in which longitudinal surveys provide benefits in terms of data collection. These are mostly connected to either the quantity or quality of data that can be collected compared to alternatives such as the use of retrospective recall (Leedy, Ormrod, 2001). Longitudinal research can utilize either primary data or secondary data. With primary data collection, the principal investigator designs the measures and methods of data collection and supervises the data collection effort.

Total Survey Error perspective

Surveys are a common method in academic research to collect data. The Total Survey Error (hereafter TSE) approach has been established as a systematic framework to understand the various sources of error that are associated with each of these steps (Biemer, Lyberg, 2003; Jedinger et al., 2018). The term survey error refers to the deviation of an estimator from the true value in a population (Biemer, Lyberg, 2003).

According to Weisberg (2009), these potential errors can be divided into three categories of respondent selection (e.g., coverage error), response accuracy (e.g., item nonresponse error) and survey administration (e.g., mode effects). Current research that relies on the TSE approach, however, focuses on a narrow concept of survey data quality that involves errors that are induced by sampling, measurement and non-responses, but does not include other factors present while working with survey data.

Data quality – beyond the TSE

There is a need to analyse the quality of data (hereafter DQ) outside the TSE approach. However, there is no single definition of the quality in the context of research data accepted by researchers and those working in the discipline.

The World Health Organization (2003) defines research data quality as the ability to achieve desirable goals. Quality data represent what is intended or defined by their official source, are objective, unbiased and comply with known standards (Abdelhak et al., 1996 as cited by World Health Organization, 2003). Following on from this, World Health Organization (2017) has created a DQ review framework that, in addition to the known parameters from the TSE perspective, has introduced the following data quality indicators: bias and human errors in data entry and computation.

They also used the term “data quality dimension”. Why “data quality dimension”? According to the available literature, data quality in scientific studies is a multifaceted concept for which there is no precise or unique definition. One way of explaining DQ is through the concept of dimensions. Dimensions deconstruct data quality into practical, definable and measurable constructs (Tayi, Ballou, 1998; Bai et al., 2018).

Whitney et al. (1998) discuss data quality in longitudinal studies and emphasize the need for quality assurance and quality control procedures beyond the TSE approach. Quality assurance (hereafter QA) consists of activities undertaken before a data collection to ensure that the data are of the highest possible quality at the time of collection. Quality control takes place during and after data collection (Whitney et al., 1998).

QA is a process used to prevent problems in the data collection process and to support subsequent data quality. It plays an important role in the conduct of a research study by helping to ensure findings and conclusions are correct and justifiable (Yamanaka et al., 2016). According to the Szklo and Nieto (2014), QA activities before data collection aim to prevent or at least minimize systematic or random errors in collecting and analysing data. Traditionally, these activities have consisted of detailed protocol preparation, development of data collection instruments and procedures and their manuals of operation, and training and certification of staff. The development of manuals specifying quality control activities can also be considered as a quality assurance activity. QA therefore includes methods and procedures for preventing and correcting problems that may affect the quality of survey data (Biemer, Lyberg, 2003)

The available literature on QA focuses mostly on standardizing the protocols and personnel training (Sáez et al., 2012; Chen et al., 2014; Szklo, Nieto, 2014; Yamanaka et al., 2016). QA steps mentioned in research papers can be summarized into three steps: (1) developing a procedure manual for data

collection, (2) developing a detailed recruitment and training plan to enforce the value of collecting accurate data and (3) monitoring and evaluating the process in the field and identifying areas of improvement to strengthen the study's protocol. However, the abovementioned steps lacked information science and computer science perspectives on data-related issues.

CRIBS case study

The project "Croatian Birth Cohort Study" of the Institute for Anthropology (hereinafter referred to as "CRIBS") is a pilot of a longitudinal study aimed at the Croatian populations of the eastern Adriatic islands and the neighbouring mainland, in particular the population of pregnant women and their born children. It is a public health longitudinal study in which a sample of 500 pairs of mothers and their children will be examined, namely mothers' lifestyle, diet and health before and during the pregnancy, and growth and development of their children.

In the context of the study, 6 surveys are being collected: 3 surveys before pregnancy and 3 surveys from pregnancy to the child's first year. The study expands over time and adds new data sources and collection methods such as allergy tests and additional surveys.

CRIBS quality assurance

Due to the unique characteristics of the longitudinal study, quality assurance is seen as an iterative process within the CRIBS study, where the characteristics of data collection processes and data handling are evaluated at each time point and analysed to improve the QA process for the following time point. In this way, methods and procedures for preventing and correcting problems that can affect the quality of the survey data are constantly being upgraded over time.

At the beginning of the study, the QA consisted of the following steps:

1. prevention - standard procedures were used to ensure accurate and consistent measurements throughout the study. Standardized training manuals were developed to document measurement protocols, detail procedures, and minimize errors. Data collection procedures for each registry were clearly defined and described. Manuals were presented in paper form;
2. detection - exploratory data analysis prior to data analysis was used in different software packages, depending on the researcher's preference (R, IBM SPSS, MS Excel);
3. correction - QA process concluded with a team debriefing of measurement activity to review results, discuss corrections and provide clarifications. The aim was to establish a continuous feedback mechanism between data sources and the research team to ensure consistency of data types, quantity, quality and origin.

CRIBS data collection

In the first wave of survey data collection, data were collected primarily through web surveys. CRIBS surveys do not contain HIPAA identifiers, and respondents are identified by a unique code. The advantages of web surveys are that they are very cost-effective. Relying on web surveys also has its drawbacks, such as excluding those participants who do not have a computer or are unable to access a computer.

For the purposes of conducting the CRIBS study web survey, we have opted for Google Forms as a commonly used survey data collection tool. Call for such surveys are sent by email. Respondents who did not have an email received hard copies of the surveys at their postal address. Upon receiving them, the researchers would manually enter such copies through the web form into the database. The collected data were then reviewed in the software package according to the preferences of the researchers, mainly IBM SPSS and MS Excel, in which Exploratory Data Analysis was performed to find possible errors.

Sources of errors

Through a case study and semi-structured interview with members of the research team, the following problems were identified in the research workflow. We detected expected data collection problems but also some less often discussed problems. Problems detected within the CRIBS study corresponding to problems discussed in other research papers (Yamanaka et al., 2016; Read et al., 2017; Young et al., 2007) were:

1. panel conditioning - the response may have been conditioned by previous experience of taking part in the survey;
2. sample attrition - continued loss of respondents from the sample due to nonresponse at each wave of a longitudinal survey;
3. recall bias;
4. generally-increased temporal and financial demands associated with these longitudinal studies.

Issues that were not found in relevant research papers were classified into the following categories: single-source problems, multi-source problems, security problems, design questionnaire problems, and QA workflow problems.

Single source problems are related to inconsistency and inaccuracy of collected data point and they do not reflect the quality of the database. Examples include *errors in data entry* (errors because of the interpretation of questions by the participant, unintentional errors such as misspellings, intentional distortion of data), *missing values*, *embedded values* (multiple values entered in only one field), *misplaced values* (values entered in the incorrect field), *duplicate entries* and *contradictory entries*.

Multi-source problems occur when multiple data sources, i.e. multiple surveys, have to be merged into a warehouse or aggregate database. Data sources often contain the same data but in different representations, which are often contradictory to one another. Such problems are a reflection of faulty survey design and are characteristic of longitudinal studies given the incidence of recurring questions. An example of a multi-source problems occurr when multiple data sources, i.e. surveys are designed with different names for the same variable which creates structural conflicts (e.g., survey "A" uses the term "customer" and source "B" uses the term "client"). Second example refers to a different representation of the same values when the variable (i.e. column in a tabular database) is called the same (e.g., survey "A" for a dichotomous "gender" variable uses "0/1" labels, while survey "B" for a variable of the same dichotomous "gender" variable uses different value labels such as "M/F").

Security problems. We recorded a case where an anonymous employee changed the survey content. The data was recovered because we connected Google Forms survey data with Python script for backing up the data that was called twice a day via cron job at the beginning and the end of working time. However, data recovery was possible only because including data backup in our QA plan and making a custom backup script, since Google Forms do not have advanced backup features.

Problems related to the design of the questionnaire arise from the general design of the questionnaire and the chosen data collection tool, Google Forms. The general design of the questionnaire refers to the structuring of the question. Some of our surveys had a certain number of so-called "free text" questions that lead to single-source issues such as values entered in the wrong field. Abstracting data from free text is often a tedious process and it usually requires a human reader. The next decision in the questionnaire design concerns the decision of which questions to ask as mandatory. Specifically, after making questions mandatory, we have noticed a "trade-off" between missing values and the number of errors in data entry. In the case of mandatory questions, the number of missing values was kept to a minimum, but the number of single-source problems increased towards the end of the survey. In the case of optional questions, a noticeably smaller number of single-source problems was observed, but the number of skipped questions, which generated missing values, was increased. Furthermore, Google Forms does not contain the "save progress" feature, which is why the validity of such surveys may be in question as people might be in a hurry to complete it and so might not give accurate responses. Google Form also lacks advanced validation features for data input, making the "data cleaning" process extremely demanding.

Quality assurance workflow problems. A sustainable workflow model needs to be made. We detected some parts of our QA workflow lacking a reproducibility feature. For example, multiple software packages such as MS Excel, IBM SPSS and Statistica have been used for the same purpose. Since each of the following programs works with its proprietary file, as a result, a large number of heterogeneous files were created for the same data set which are not fully compatible with each other. That led to problems in creating a consistent workflow for working with the data, namely prevention and detection. Also, QA workflow required more comprehensive documentation of procedures in a more detailed manual.

Eliminating the errors

After a semi-structured interview, which examined the research team's attitudes towards the sources of errors found in the case study and discussed proposals to address them, a focus group was organized with the same members of the research team to provide a more thorough argumentation of the same topics and to obtain a wider range of information. The two sets of interventions (prevention and detection) were made according to the specified sources of error according to the steps of the QA process and will be discussed in the following sections. The third type of intervention (correction) will not be discussed at this time due to the length and complexity of the steps involved.

Prevention

Interventions that can be classified as a prevention step, can further be subdivided into four distinct models.

Managing attrition rate. Methods of email campaigns were used when sending web surveys to respondents for a more detailed insight into participants' behaviour. Of the last 353 web surveys submitted, we had a click rate of 71%, that is, 29% of respondents did not open an email within the span of 3 weeks. Of the 71% open emails, 78% of surveys were completed within three weeks. By stratifying respondents by their behaviour, we could elaborate campaigns tailored to a specific group of respondents to reduce sample attrition. A smaller group of respondents with a smaller click rate is devoted more time and is contacted by telephone.

Particular attention should be paid to sample attrition as a source of data quality problems. The problem of study attrition is unique to longitudinal designs and must be accounted for while presenting study results. From an analysis perspective, sample attrition is information about sample behaviour and can thus provide additional insight into the results. Still, sample attrition is not often talked about in the context of research (longitudinal) studies.

Project tailored tool. Downsides of a general survey tool have been revised that led to a decision to implement a new web survey collection tool, REDCAP (Harris et al., 2009). REDCAP is a fully compliant data collection tool with DPA & GDPR. User privileges and rights can be controlled and all interactions are logged and auditable. It has an advanced form with data validation features that eliminates certain single-source problems such as errors in data entry, values entered in the wrong field, and duplicated values. Moreover, it has the feature of saving data entry progress and resuming later. This reduces the trade-off effect between missing values and the number of single-source issues, but also attrition rates.

Live chat service. Demo live chat service is underway where the respondent can contact a research team member in real-time. In the demo version, such a feature proved to be extremely useful for reducing single-source problems such as errors in the interpretation of a question by the participant. However, maintaining the real-time help desk service is extremely challenging and time-consuming for a small research team.

Data documentation design. Documentation of procedures in a more detailed manual is under development. The manual now contains new information such as *a priori* specification of potential confounding variables. Documentation and manuals are in paper form and also in the form of self-hosted wiki, as wiki form turned out to be a great way to set up an in-house knowledge base. Wiki is being constantly updated. In addition to the manual, an interactive codebook following DDI Alliance instructions is under construction. The Data Documentation Initiative (DDI) (Data Documentation Initiative, n.d.) is an international standard for describing data produced by surveys and other observational methods for research data. Codebook also eliminates most multi-source problems, i.e. problems that occur when multiple data sources, i.e. multiple surveys, must be merged into an aggregate database or a data warehouse. Finally, the codebook can integrate within the new workflow written in the R programming language which makes it easier to export data to other analysis programs preferred by other members of the research team.

Detection

In addition to prevention models, we distinguish two observable cases of detection models.

Exclusive data analysis tool. Using multiple versions of software packages (R, IBM SPSS, MS Excel) for the same purpose (namely exploratory data analysis) is no longer possible. The ETL process within the R programming language is being made. R is not really designed for ETL - R by

design loads data into memory, so it is limited by the amount of memory the user has available in his system. However, since the size of the data used is usually one of the major determinants of viable ETL, R's "Tidyverse" package has shown to be a good choice for narrow scope ETL – in our case, for small-scaled survey data. It should be noted, however, that R lacks high-level ETL process support and lacks features such as staging objects, manual logging and visualizing data pipes.

Traceability. An "R Markdown" in HTML format is created within the ETL process for an interactive report where each member of the research team can see the status of each ETL step and provide their feedback on the issues. The next step is to implement a more comprehensive data quality report.

Discussion and conclusion

The aim of this paper was to promote transparency and to share the insights about the errors inherent in most studies containing survey data. Those errors affect data quality, and data quality should be a key priority when planning a longitudinal study to guarantee appropriate results and conclusions from survey data. In practice, the commonly used TSE approach has not proven to be sufficient when working with survey data and analysing their quality. Biemer and Lyberg (2003) criticize TSE and say that it lacks a user perspective and should be complemented by a more modern quality paradigm.

Survey data quality is currently a vague concept with multiple definitions and sources, and according to Houston (2018), only a small body of academic research has described the use of data quality in research. Papers on quality assurance in research (longitudinal) studies predominantly talk about structures, processes, and policies that need to be a place to ascertain the quality of the data collected, but in-depth insights of information science and computer science approaches are rarely seen. Many of the data-centric topics such as data cleaning and transforming research data, building data pipelines, detection of errors in the data collection process and database-related challenges are rarely discussed - especially in social science and humanities.

Hence, it is necessary to reopen the methodological discussion about data quality and data in general research studies - a space where information experts can certainly find their place - especially when looking for new challenges on the horizon. For instance, setting up a causal frame according to observational data from the field was always a challenge. Consequently, researchers started to consider other data sources and integrating them into their survey-based analyses in order to work with innovative research questions (Spjuth et al., 2016). This entails challenges such as data linkage, i.e. merging survey data with other sources. Along with data linkage techniques, we can see the rise of harmonizing research data that refers to linking multiple different studies into one unified data warehouse (Spjuth et al., 2016). Finally, one cannot ignore the importance of FAIR guiding principles for research data management and stewardship which emphasises the capacity of computational systems to find, access, interoperate, and reuse data (European Commission, 2016). For this reason only, it is advisable to create a reproducible analytical pipeline - which opens a myriad of new challenges known in the IT sector but is rarely mentioned in the context of scientific research, such as usage of version control, dependency management, the need for good schema design and choosing the right tools in general.

Information scientist and computer experts (or as the trends suggest, "data scientists") should play a more prominent role within the work of research (longitudinal) studies and be more open about their techniques and their advantages and disadvantages used for dealing with the research data and aim to integrate those insights into quality assurance as well as data management plans.

References

Bai, L., Meredith, R., Burstein, F. (2018). A Data Quality Framework, Method and Tools for Managing Data Quality in a Health Care Setting: An Action Case Study. // Journal of Decision Systems 27, 144-154

Biemer, P. P., Lyberg, L. E. (2003). Introduction to Survey Quality. John Wiley & Sons

Chen, H., Hailey, D., Wang, N., Yu, P. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. // International Journal of Environmental Research and Public Health 11, 5, 5170-5207

Data Documentation Initiative. Welcome to the Data Documentation Initiative. (2019). https://ddialliance.org/ (25.8.2019)

European Commission. (2016). H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 v3.0. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (1.9.2019)

Harris, P. A., Taylor, R., Thielke, R. (2009). Research Electronic Data Capture (REDCap) - A Metadata-Driven Methodology and Workflow Process for Providing Translational Research Informatics Support. Journal of Biomedical Informatics 42, 2, 377-381

Houston, L., Yu, P., Martin, A., Probst, Y. (2018). Defining and Developing a Generic Framework for Monitoring Data Quality in Clinical Research. // AMIA Annual Symposium Proceedings, 1300-1309

Jedinger, A., Watteler, O., Förster, A. (2018). Improving the Quality of Survey Data Documentation: A Total Survey Error Perspective. Data 3, 4

Leedy, P. D., Ormrod, J. E. (2001). Practical Research: Planning and Design. Upper Saddle River, N. J.: Merrill Prentice Hall

McFadden, F. R., Prescott, M. B., Hoffer, J. A. (1998). Modern Database Management. 5th ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co.

Read, K. B., LaPolla, F. W., Tolea, M. I., Galvin, J. E., Surkis, A. (2017). Improving Data Collection, Documentation, and Workflow in a Dementia Screening Study. // Journal of the Medical Library Association 105, 2, 160-166

Sáez, C., Martínez-Miranda, J., Robles, M., García-Gómez, J. M. (2012). Organizing Data Quality Assessment of Shifting Biomedical Data. Studies in Health Technology and Informatics 180, 721-725

Spjuth, O., Krestyaninova, M., Hastings, J. (2016). Harmonising and Linking Biomedical and Clinical Data across Disparate Data Archives to Enable Integrative Cross-Biobank Research. // European Journal of Human Genetics 24, 521-528

Szklo, M., Nieto, F. J. (2014). Epidemiology: Beyond the Basics. Jones & Bartlett Publishers

Tayi, G. K., Ballou, D. P. (1998). Examining Data Quality. // Communications of the ACM 41, 2, 54-57

Weisberg, H. F. (2009). The Total Survey Error Approach: A Guide to the New Science of Survey Research. University of Chicago Press

Whitney, C. W., Lind, B. K., Wahl, P. W. (1998). Quality Assurance and Quality Control in Longitudinal Studies. Epidemiologic Reviews 20, 1, 71-80

World Health Organization. (2003). Improving Data Quality: A Guide for Developing Countries. Manila: WHO Regional Office for the Western Pacific. https://apps.who.int/iris/handle/10665/206974 (25.7.2019)

World Health Organization (2017). Data Quality Review: Module 3: Data Verification and System Assessment. https://apps.who.int/iris/handle/10665/259226 (25.7.2019)

Yamanaka, A., Fialkowski, M. K., Wilkens, L. (2016). Quality Assurance of Data Collection in the Multi-Site Community Randomized Trial and Prevalence Survey of the Children's Healthy Living Program. // BMC Research Notes 9, 1, 432

Young, A., Powers, J., Wheway, V. (2007). Working with Longitudinal Data: Attrition and Retention, Data Quality, Measures of Change and Other Analytical Issues. // International Journal of Multiple Research Approaches 1, 2, 175-186