Data Acquisition and Preprocessing in Studies on Humans: What is Not Taught in Statistics Classes?

Yeyi ZHU, Ladia M. HERNANDEZ, Peter MUELLER, Yongquan DONG, and Michele R. FORMAN

The aim of this article is to address issues in research that may be missing from statistics classes and important for (bio-) statistics students. In the context of a case study, we discuss data acquisition and preprocessing steps that fill the gap between research questions posed by subject matter scientists and statistical methodology for formal inference. Issues include participant recruitment, data collection training and standardization, variable coding, data review and verification, data cleaning and editing, and documentation. Despite the critical importance of these details in research, most of these issues are rarely discussed in an applied statistics program. One reason for the lack of more formal training is the difficulty in addressing the many challenges that can possibly arise in the course of a study in a systematic way. This article can help to bridge the gap between research questions and formal statistical inference by using an illustrative case study for a discussion. We hope that reading and discussing this article and practicing data preprocessing exercises will sensitize statistics students to these important issues and achieve optimal conduct, quality control, analysis, and interpretation of a study.

KEY WORDS: Applied statistics courses; Data cleaning; Data code book; Data collection; Data dictionary; Quality control; Statistical education.

1. INTRODUCTION

Statistics classes focus on mathematical, statistical, and computational theories and methods. However, before researchers reach the first step of formal statistical analysis, many data errors and data quality issues may arise of which a researcher needs to be aware. Data errors and problems may include entry errors, missing values, duplicates, outliers, and data inconsistencies and discrepancies, any of which may affect the validity, reproducibility, and thus the quality of studies. In large-scale studies, budgets may be allocated for personnel with distinct roles, including principal investigators, study coordinators, data collectors, database managers, and statisticians. More often than not, researchers may need to play multiple roles in certain study settings, thereby increasing the demands on researchers to oversee quality control over the whole study flow from study design to data acquisition, preprocessing, and analysis. The importance of quality control over data acquisition is well recognized, but is usually not discussed in applied statistics classes. Furthermore, data preprocessing bridges the gap from data acquisition to statistical analysis but has not been championed as a relevant component in statistics curricula.

In this article, we review some critical issues in quality control during data acquisition and preprocessing of which statistics students should be aware but that are typically not taught in statistics courses. The aim of the article is to address these issues in the context of a case study involving human subjects and discuss possible steps to mitigate related problems. We also provide specific recommendations about class discussions and data preprocessing exercises that could be introduced in applied statistics courses. We hope that introducing the concepts and practical approaches of data acquisition and preprocessing in statistics curricula will sensitize statistics students and researchers to these important issues in quality control and encourage more related discussions.

2. WHY DO WE CARE ABOUT DATA ACQUISITION AND PREPROCESSING?

Data errors may appear at any stage of data acquisition and preprocessing, which could affect study results and lead to erroneous statistical interpretation and conclusions. Goldberg, Niemierko, and Turchin (2008) reported error rates of 2.3%–5.2% for demographic data and 10%–26.9% for clinical data in oncology patients, which could be attributed to data entry errors and researchers' misinterpretation of tumor treatment outcomes due to missing and inconsistent data. These data errors could significantly affect the results by increasing the standard errors of the mean and decreasing the statistical power (Day, Fayers, and Harvey 1998). Data errors could also lead to erroneous findings. An erratum to a published paper (Lim et al. 2012) on risk assessment of disease burden reported that an error in the estimates of burden for alcohol use led to incorrect

Yeyi Zhu is Ph.D. Candidate, Department of Nutritional Sciences, The University of Texas at Austin, Austin, TX 78712 (E-mail: *yeyizhu@utexas.edu*). Ladia M. Hernandez is Research Scientist, Department of Nutritional Sciences, The University of Texas at Austin, Austin, TX 78712 (E-mail: *ladia.hernandez@austin.utexas.edu*). Peter Mueller is Professor, Department of Mathematics, The University of Texas at Austin, Austin, TX 78712 (E-mail: *pmueller@math.utexas.edu*). Yongquan Dong is Statistician, Department of Nutritional Sciences, The University of Texas at Austin, Austin, TX 78712 (E-mail: *kanedong@austin.utexas.edu*). Michele R. Forman is Bruton Centennial Professor, Department of Nutritional Sciences, The University of Texas at Austin, Austin, TX 78712 (E-mail: *mforman@austin.utexas.edu*). This work was supported by the National Institute of Child Health and Human Development Grant HHSN275200800020C. The authors thank all the participants and research collaborators across 10 study centers in this study, and the Editor, Associate Editor, and anonymous reviewers for many helpful comments and suggestions.

 DOI: 10.1080/00031305.2013.842498

estimates of mortality and morbidity from ischemic heart disease attributable to alcohol use, which required corrections in the Summary, Results, and Discussion sections, three tables, all figures, and the Appendix. This type of data error could be due to: conversion errors from ounces to grams as the measure of alcohol use in different countries/regions; miscoding of alcohol use as some other risk factors of interest in the study; or errors when data were merged from different sources. With proper data preprocessing, these errors could be avoided before the formal analysis and reporting. Moreover, data errors could result in opposite conclusions. In a clinical study comparing the effect of two treatment protocols on patients with Hodgkin's disease, Levitt et al. (1993) demonstrated that omission of one select patient from one treatment group ($n = 37$) changed the comparison results from statistical insignificance to significance.

Despite recognition of these data issues and adverse consequences, data preprocessing has received relatively little attention in instructional environments, compared to the emphasis on optimal study design and adherence to research protocols. Students or researchers who are new but want to perform data preprocessing may be challenged by limited and difficult-to-access resources. Although the Ethical Guidelines for Statistical Practice by the American Statistical Association state that researchers should report "the data cleaning and screening procedures used, including any imputation" in publications (American Statistical Association 1999), it is uncommon to see all information reported in publications. Some universities and institutes do provide online information about data acquisition and preprocessing, but they are usually request-based services. Indeed, it is difficult for students to find comprehensive manuals or guidelines regarding these issues. Therefore, given the significance of data acquisition and preprocessing in relation to data quality control, it is important that applied statistics curricula provide students a platform to learn, discuss, and practice data acquisition and preprocessing skills in a systematic and planned way.

3. STEPS TO DATA ACQUISITION AND PREPROCESSING USING A CASE STUDY AS AN ILLUSTRATION

The data acquisition process in studies involving human subjects typically includes participant recruitment, screening, consent, and data collection. Data preprocessing usually has five steps: data review; entry and verification; cleaning; editing; and documentation (Maletic and Marcus 2000). Given the large variability in data issues within study-specific contexts, it is impossible to enumerate all possible data errors and corresponding preprocessing strategies. Instead, in this article we use an epidemiological study of infant feeding practices and childhood growth to illustrate common approaches to data acquisition and preprocessing to improve data quality and integrity.

3.1 Background: Study Description

The case study based on the National Children's Study (NCS) Formative Research in Physical Measurements is a cross-sectional study involving 1634 mother–offspring dyads across 10 study sites in the U.S. Mothers were administered a questionnaire on socio-demographic, reproductive, and child feeding factors. Children aged < 6 years were measured for standard anthropometrics (length, height, and weight) and ulnar length by different tools (caliper, ruler, and measuring paper grid). The primary objective of this study was to evaluate ulnar lengths measured by different tools as surrogate measures of body length and height by age, sex, and ethnicity in infants and children aged 0–6 years. A secondary goal was to examine the associations between infant feeding practices and childhood linear growth parameters in this sample. The following discussion is based on data acquisition and preprocessing procedures used in this study.

3.2 Quality Control in Data Acquisition

Participant Recruitment, Screening, and Consent. Statistical classes usually do not discuss participant recruitment, screening, or consent. However, it is important that statisticians are aware of these procedures so that they can provide feedback for study design before actual data collection begins and select appropriate analytic approaches. For example, this NCS formative research involved multisite data collection. After participant recruitment, we examined subject characteristics in each site to determine whether a study site effect was present and whether a linear mixed-effect regression model with study site as a random effect was appropriate. In addition, according to subject eligibility criteria, we created filters to exclude ineligible participants although prescreened and observed in the dataset. It is also worth noting that any study that depends on volunteer participants is subject to a possible selection bias. It is important that the statistical collaborators are aware of these challenges and can sensitize study investigators to these problems in the study design phase. In our case, to reduce the impact of possible selection bias inherent to this convenience sample, we collected relevant covariates of childhood growth and infant feeding to compare participants by selective characteristics that might bias results, for example, ethnicity, maternal prepregnancy body mass index, perinatal morbidity, and child's birth weight.

Staff Training and Data Collection Standardization. To implement effective quality control, a set of procedures should be established prior to data collection to ensure the staff adheres to the defined set of quality control criteria. In this NCS study, researchers at each study center were *initially trained* by experienced principal investigators with hands-on practice of measurements on young children volunteers. *A manual of procedures for anthropometric measurements* was provided to the staff with detailed steps for conducting the standardized measurements. *A training video* was provided to each study center for subsequent re-training in anthropometrics to standardize collection procedures. *Webinars* describing interview procedures were held by principal investigators to all study sites. *Weekly conference calls* were held to discuss and share field experiences regarding participant recruitment, data collection, interaction with participants, and field conditions. *Daily calibration of equipment* was required and recorded on forms.

Actual data collection required careful attention to the administration of the study questionnaire(s) by interviewers and

Table 1. An example of variable coding documented in a data code book

Survey questions *[responses]*	Variable definitions	Variable name	Label	Codes	Type
18. Was XXX ever fed breast milk? *[Yes/No]* 19. How old was XXX when s/he completely stopped being fed breast milk? *[Age A]* 20. How old was XXX when s/he was first fed something other than breast milk or water? *[Age B]* 21. Was XXX ever fed formula? *[Yes/No]*	a) Exclusive breastfeeding (XBR): *Yes* to Q18 & *No* to Q21; or *Yes* to Q18 and Q21 & *Age A* < *Age B* b) Breast–bottle feeding (BrBot): *Yes* to Q18 and Q21 & *Age A* ≥ *Age B* c) Exclusive bottle-feeding (XBot): *No* to Q18 & *Yes* to Q21	IF_practices	Infant feeding practices	1 = XBR 2 = BrBot 3 = XBot 97 = Refused 98 = Do not know 99 = Missing	Nominal

NOTES. Age A: age stopped breastfeeding, Age B: age started feeding something other than breast milk or water, IF: infant feeding.

completion by participants. In our study, despite the preferred mode of administration by in-person interview, approximately 11% of the mothers self-completed the questionnaire at one site due to logistical issues. Self-administration compared to interviewer-guided administration could potentially affect data quality due to participants' misinterpretation of questions, thus requiring special attention during data preprocessing. In addition, comments regarding logistical conditions during measurement, participants' compliance to the measurement protocol, and reasons for measurement interruption or failure were documented on the anthropometric form. If statisticians identify invalid or implausible values during the data cleaning stage, these comments may help data evaluation and interpretation.

Variable Coding and a Data Code Book. Variable coding is a process that distills and aggregates useful information from the original data and assigns codes to make data analyzable. A data code book describes the content of the dataset and typically has information including original survey questions and skip patterns; variable definitions; variable name, type, label, and values; code for missing data; and other characteristics of each variable. In our study, we collected a series of open-ended and multiple choice questions about whether the mother has ever fed the child in the study on breast milk and/or formula, age started and stopped feeding breast milk and/or formula, and age at introduction of solid foods. As illustrated in Table 1, a data code book documents variable definitions, derivations from the original data, and variable coding. Besides the use of a word document or a spreadsheet, there are software programs available for recording data and developing a data code book, such as IBM SPSS, SAS, and STATA.

3.3 Quality Control in Data Preprocessing

Data errors and issues are prone to arise at any stage even in carefully planned studies. Thus, a systematic and thorough preprocessing approach developed before data analysis is critical to enhance overall data quality and integrity. We delineate data preprocessing into a series of five steps as follows.

Step 1: Data Review. As a front-end process, data review of forms and questionnaires by trained researchers is critical to reduce errors and evaluate data integrity. In this study, *a data review guidebook* was developed to identify and describe the errors and solutions. Detectable errors at the data review stage could be due to, but not limited to: misinterpretation of survey questions by interviewers and/or interviewees; conversion errors due to the use of different metric units in measurements; transcription errors from measurement equipment to forms; and correct values recorded in wrong boxes. Error screening methods at this phase are not necessarily restricted to the statistical. These errors can be detected usually based on study-specific expected ranges; researchers' knowledge of the subject; and common sense. For example, data collection of the study occurred between June 2011 and August 2012; thus dates out of this range must be errors. If not corrected before data entry, these errors could be very difficult to detect and could subsequently result in errors for age calculated as the difference between study date and birth date. Another example is that a child aged 22 months was measured for recumbent length, but the value was recorded in the box for standing height while only measurement for the former was expected. In some circumstances, a data reviewer may be able to requery participants or research collaborators to correct errors; however, considering the increasing difficulty of requerying as time goes by and the relevant person becomes unreachable, it is highly recommended to initiate the data review process as early as possible, ideally soon after data collection begins.

Step 2: Data Entry and Verification. Several approaches could be used for data entry, for example, manual entry of paper records, data transfer from handheld computers used for data collection, and optical scanning (Roberts et al. 1997). We manually entered values from measurement forms and questionnaires into an informatics platform. When additional data errors were discovered at this stage, researchers implemented rules in the *data review guidebook* and documented the changes. Despite careful entry by well-trained staff, data entry is inevitably prone to errors. Based on study-specific features of data, investigators might choose different data verification methods including double data entry or visual comparison (Blumenstein 1993; Kawado et al. 2003). In this study, due to the unavailability of double entry in the informatics platform, a different person from that of the data enterer verified entries by visual comparison, following specific guidelines for error detection, correction, and documentation in the *data review guidebook*.

Step 3: Data Cleaning. Although many data errors can be detected by initial review, as a good practice, it would be important

for researchers to pre-establish or define rules for data cleaning and editing. In this study, we cleaned demographic, anthropometric, and infant feeding data using the following checks for logic and consistency, outliers, and missing data.

(a) Logic and Consistency Checks. Compared to outliers, erroneous data within the expected range are more difficult to distinguish from valid data (Winkler 1998). In this study, given the logic behind maternal responses to infant feeding practices, two types of checks were performed as below. First, the authors performed cross-checks on the same variable measured on repeated occasions using different questions. For example, alerts were triggered when the child was reported as being currently formula-fed but not fed formula in the past 7 days based on responses to two separate questions. The second type was pairwise and multivariable cross-checks among variables that should be internally related. An example for pairwise cross-checks was to compare the ages when formula feeding started and stopped. Flags were created if the age started was later than the age ended. An example for multivariable cross-checks was to examine whether those breast–bottle fed children (categorized using multiple questions shown in Table 1) were reported as being fed formula only or both formula and breast milk during the mixed feeding period. Obviously, maternal report of feeding formula only was inconsistent with the practice of mixed feeding.

In short, the logic and consistency checks are essential to data cleaning and facilitate the identification of suspected erroneous data which otherwise would be difficult to locate using regular statistical methods for outlier checks as discussed below. In addition, data cleaning is highly recommended in the early stage of a study to provide feedback about data error sources to investigators and develop study-tailored quality control strategies for data collection.

(b) Outlier Detection. Screening approaches for outlier detection can be statistical and/or empirical. Statistical packages usually have functions to perform univariate outlier detection, such as frequency checks using histograms or frequency distribution tables; range checks using box plots and stem-and-leaf plots; and central tendency and dispersion checks calculating the mean, median, and standard deviation for example. Also, multivariate outlier detection with the assistance of graphics can assist outlier detection via data visualization. For example, a scatterplot stratified by child's age group (Figure 1) and Bland–Altman plot (Figure 2) of two related measures visually identified potential outliers which could be difficult to detect using univariate statistical methods.

Outlier screening can also be based on researchers' knowledge and experience. One example in our study is outlier checks on maternal-reported birth weight of the child. The quality of maternal-reported birth data could be compromised by recall bias and conversion errors due to different metric units used by mothers of different ethnicities. For example, a birth weight of 5000 g would be about 3.5 standard deviations above the reported national mean (mean = 3389 g, standard deviation = 466 g) (Donahue et al. 2010) and hence would be flagged as an unusual observation. To screen further, it was necessary to take relevant maternal characteristics into consideration. If the infant mentioned above was born to a mother who had severe anemia and smoked frequently during pregnancy, it would trigger

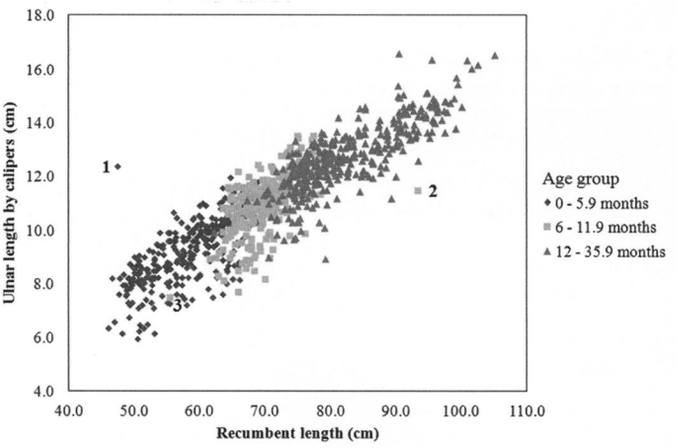

Figure 1. An example of multivariate outlier detection using a scatterplot of ulnar length versus recumbent length by age group. Potential outliers are flagged with numbers. The online version of this figure is in color.

an alert because these maternal complications are risk factors for low birth weight (Thompson et al. 2001). In addition, if other data such as newborn's birth length were available, further reviews of the data would provide helpful information for consideration as birth weight and length are positively correlated. For example, an infant boy with a high birth weight of 5000 g and a short birth length of 45 cm would be a potential outlier based on the weight-for-length growth chart (Centers for Disease Control and Prevention 2000).

(c) Missing Data Preprocessing. Missing data are a common issue for most studies. Before statisticians apply approaches to address this issue in the analysis phase, it is important to understand why data are missing and be aware of the approaches to avoid missingness. Besides participants' failure to provide responses, missing values could be due to incomplete data forms sent by study coordinators, data entry errors, or interruptions

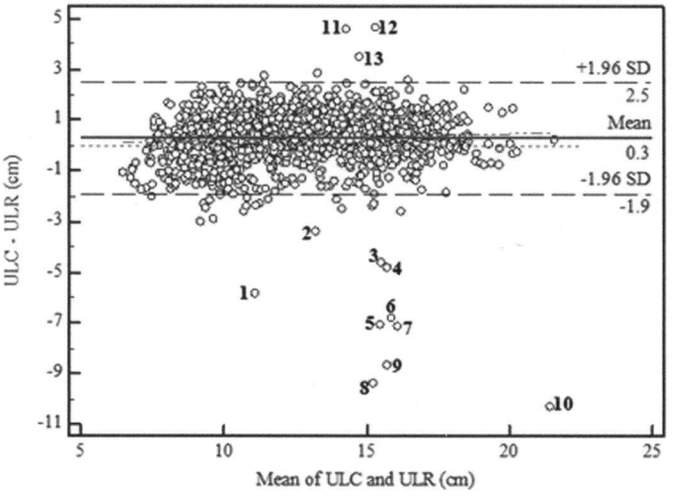

Figure 2. An example of multivariate outlier detection using a Bland–Altman plot of ulnar length measured by calipers versus ulnar length by rulers. Potential outliers are flagged with numbers. The online version of this figure is in color.

of data transmission from informatics to personal computers, which could be avoided by resending forms, correcting entry errors, and careful data transmission, respectively. In addition, missing data may be avoided by abstracting the same or relevant information collected from different questions on other forms. For instance in this study, child's birth date should be reported on both anthropometric measurement form and maternal questionnaire. In case one was missing, the same information reported in the other source could be used to fill in the blank.

Step 4: Data Editing. Following data cleaning, researchers need to clarify and determine whether the suspected data errors or issues detected at Step 3 are real errors, true extremes, or unable to be verified. In this study, authors applied diagnostic procedures as follows. First, we checked for data entry accuracy and corrected entry errors. Second, for data errors that did not pass checks for entry accuracy (which could be due to errors in data collection), we contacted the corresponding study site coordinators for subject requerying. If no further confirmation or requery was obtained, the suspected erroneous variables were recoded as missing values. Finally, data editing rules were established according to the consensus reached in laboratory meetings based on the investigators' research experience and knowledge. All the comments, flags, and corrections were documented accordingly.

Step 5: Documentation in a Data Dictionary. Clear and detailed documentation for data preprocessing is important for data integrity and serves as a useful tool for statistical analysis and interpretation. Different from a data code book described in Section 3.2, a data dictionary has detailed documentation on suspected errors including: diagnostic strategies for data cleaning; justification and rules for data editing; decisions for error treatment; and information on dates and personnel involved with specific data preprocessing steps. Such information from the data dictionary should also be reported in the final publication as an essential component of quality and validity assessment as suggested by the American Statistical Association (1999). Furthermore, the data dictionary provides important feedback about data error sources to study investigators, which could assist researchers in improving research protocols and training procedures and in harmonizing data across studies to improve data quality control.

4. RECOMMENDATIONS TO TEACH DATA ACQUISITION AND PREPROCESSING IN STATISTICS CLASSES

Although the importance of data acquisition and preprocessing in terms of quality control and assessment is well recognized, these concepts are rarely introduced in statistics classes. We provide some practical recommendations as a beginners' guide for those who are interested in adding these concepts in a statistics curriculum.

Table 2. Practical approaches and guidelines to implement data preprocessing using real, unprocessed data

Steps	Objectives	Data preprocessing guidelines
1. Get to know the study	(1) Assess the quality and integrity of collected data by looking into data acquisition process (2) Get a sense of potential bias or data issues in the dataset	(1) Learn details about the study: • What is the research question and study design? • What are the subject recruitment criteria? • How and what data are collected? (2) Check whether subjects in the dataset meet the eligibility criteria. If not, exclude them and document the changes in the *data dictionary*.
2. Assess the validity of variable coding	Ensure the variables of interest are coded in a meaningful and clear language	(1) Check how variables are coded and assess if the coding is appropriate according to the sampling distribution, specific research question, and coding methods used in related literature. (2) If the current coding is inappropriate, recode the variable and document justifications for recoding in the *data code book*.
3. Assess data entry accuracy	Make sure information in the dataset is valid and accurate	(1) If original data are accessible, review and verify data entry (assuming data are already entered and electronically available). (2) If data entry errors are detected, correct the invalid entries and document the changes or comments in the *data dictionary*.
4. Perform data cleaning	Detect suspected data errors	(1) Perform logic and consistency checks: • Cross-check the same variable collected and measured using different questions. • Conduct pairwise and multivariable cross-checks among variables that are internally related. (2) Review for outliers: • Statistical methods: univariate (frequency, range, and central tendency and dispersion) and multivariate checks (graphics plotting multiple related measures). • Empirical methods based on related knowledge or experience. (3) Look for missing data and assess if missingness can be avoided. (4) Document suspected data errors in the *data dictionary*.
5. Edit identified data errors	Improve data quality by addressing data errors	(1) Recheck data entry accuracy and correct errors if necessary. (2) Requery study coordinators or participants for problematic data. (3) Edit data for suspected errors: deletion, correction, or no change. (4) Document data editing rules and decisions in the *data dictionary*.

4.1 Courses to be Involved

We recommend teaching data preprocessing concepts and skills in applied statistics classes. Other classes involving applied statistics in discipline-specific subjects are also platforms to implement the concepts. In addition, research teams in need of members' ability to manage data quality control may also benefit by providing training sessions introducing data acquisition and preprocessing procedures.

4.2 Conduct a Case Study

Students new to the topic of data acquisition and preprocessing might be challenged by the question "Where and what to start with?" Given data preprocessing issues vary considerably by subject-specific research area, we advocate that instructors begin by selecting research topics and associated datasets of interest and conduct a case study in class. For example, the NCS formative research discussed in this article could be an example of an in-class case study in epidemiological research. Instructors can either lecture about the case study or conduct interactive class discussions with students. Outside speakers with access to an actual dataset and experiences in data acquisition and preprocessing are recommended to be invited for an in-class talk as well.

4.3 Assign a Data Preprocessing Exercise

To better motivate and involve students in real data preprocessing practices, we recommend that a class project or assignment for data preprocessing be supplemented with case study discussions. Real case exercises give students an opportunity to better understand the components of preprocessing through examples of data collection. Within the context of a specific research question, students may need to design and implement tailored preprocessing approaches that might have not been discussed in this article. For purposes of practicing data preprocessing using a real, unprocessed dataset, a series of five key steps can be provided to students as guidelines (Table 2). It is important for students to realize that each step alone is insufficient, while the entirety of steps creates a platform for data quality control via optimal data acquisition and systematic preprocessing approaches.

5. CONCLUDING REMARKS

In a typical statistics curriculum, the primary focus is usually on study design, statistical theories, and methods, and the use of statistical packages with either a brief or even absent description of data acquisition and preprocessing. This article uses a real-life example to illustrate the process of data acquisition with respect to quality control, and the need for and approaches to data preprocessing in an epidemiological study involving human subjects. It also demonstrates data preprocessing as a pivotal component in the interactive feedback system among study design, data acquisition, statistical analysis, and publication (Figure 3). Each component in the system has a unique contribution in terms of improving overall data quality control and cannot be removed. This real-case scenario may

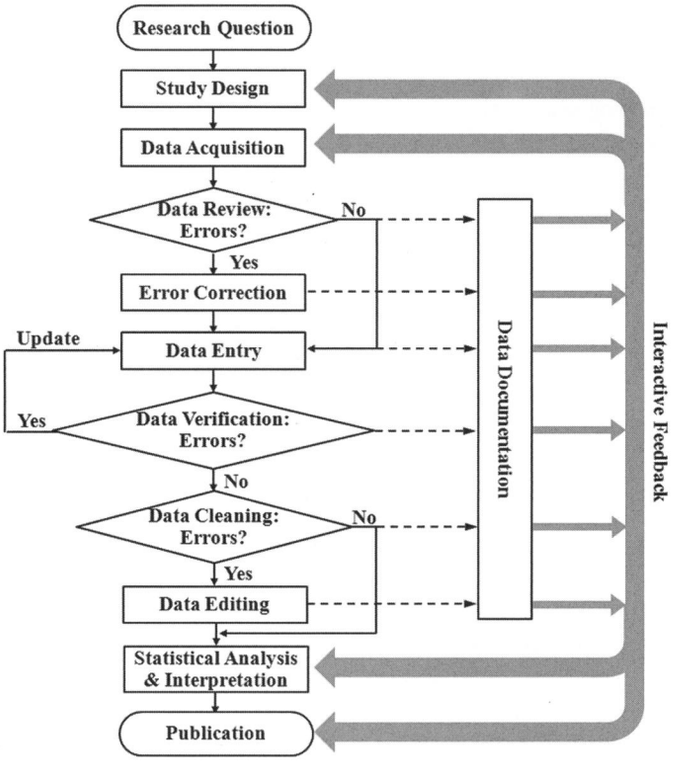

Figure 3. The interactive feedback system including study design, data acquisition, preprocessing, statistical analysis, and publication. The online version of this figure is in color.

provide insights into data quality control via optimal data acquisition and systematic data preprocessing for teachers and students in applied statistics related disciplines. We advocate that instructors incorporate data acquisition and preprocessing components into current statistics curricula by introducing a case-study discussion or lecture and supplement it with a real case data preprocessing exercise. We hope that data acquisition and preprocessing can serve to more closely bridge the phase of study design to the phase of statistical analysis, and ultimately help both investigators and statisticians to achieve optimal conduct, quality control, data analysis, and interpretation of a study.

[Received December 2012. Revised August 2013.]

REFERENCES

American Statistical Association (1999), "Ethical Guidelines for Statistical Practice." Available at *http://www.amstat.org/about/ethicalguidelines.cfm* [236,239]

Blumenstein, B. A. (1993), "Verifying Keyed Medical Research Data," *Statistics in Medicine*, 12, 1535–1542. [237]

Centers for Disease Control and Prevention (2000), "CDC Growth Charts: United States." Available at *http://www.cdc.gov/growthcharts/data/set2/chart-11.pdf* [238]

Day, S., Fayers, P., and Harvey, D. (1998), "Double Data Entry: What Value, What Price?," *Controlled Clinical Trials*, 19, 15–24. [235]

Donahue, S. M., Kleinman, K. P., Gillman, M. W., and Oken, E. (2010), "Trends in Birth Weight and Gestational Length Among Singleton Term Births in

the United States: 1990–2005," *Obstetrics and Gynaecology*, 115, 357–364. [238]

Goldberg, S. I., Niemierko, A., and Turchin, A. (2008), "Analysis of Data Errors in Clinical Research Databases," in *Proceedings of the American Medical Informatics Association Annual Symposium*, pp. 242–246. [235]

Kawado, M., Hinotsu, S., Matsuyama, Y., Yamaguchi, T., Hashimoto, S., and Ohashi, Y. (2003), "A Comparison of Error Detection Rates Between the Reading Aloud Method and the Double Data Entry Method," *Controlled Clinical Trials*, 24, 560–569. [237]

Levitt, S. H., Aeppli, D. M., Potish, R. A., Lee, C. K., and Nierengarten, M. E. (1993), "Influences on Inferences. Effect of Errors in Data on Statistical Evaluation," *Cancer*, 72, 2075–2082. [236]

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., Amann, M., Anderson, H. R., Andrews, K. G., Aryee, M., Atkinson, C., Bacchus, L. J., Bahalim, A. N., Balakrishnan, K., Balmes, J., Barker-Collo, S., Baxter, A., Bell, M. L., Blore, J. D., Blyth, F., Bonner, C., Borges, G., Bourne, R., Boussinesq, M., Brauer, M., Brooks, P., Bruce, N. G., Brunekreef, B., Bryan-Hancock, C., Bucello, C., Buchbinder, R., Bull, F., Burnett, R. T., Byers, T. E., Calabria, B., Carapetis, J., Carnahan, E., Chafe, Z., Charlson, F., Chen, H., Chen, J. S., Cheng, A. T., Child, J. C., Cohen, A., Colson, K. E., Cowie, B. C., Darby, S., Darling, S., Davis, A., Degenhardt, L., Dentener, F., Des Jarlais, D. C., Devries, K., Dherani, M., Ding, E. L., Dorsey, E. R., Driscoll, T., Edmond, K., Ali, S. E., Engell, R. E., Erwin, P. J., Fahimi, S., Falder, G., Farzadfar, F., Ferrari, A., Finucane, M. M., Flaxman, S., Fowkes, F. G., Freedman, G., Freeman, M. K., Gakidou, E., Ghosh, S., Giovannucci, E., Gmel, G., Graham, K., Grainger, R., Grant, B., Gunnell, D., Gutierrez, H. R., Hall, W., Hoek, H. W., Hogan, A., Hosgood, H. D., III, Hoy, D., Hu, H., Hubbell, B. J., Hutchings, S. J., Ibeanusi, S. E., Jacklyn, G. L., Jasrasaria, R., Jonas, J. B., Kan, H., Kanis, J. A., Kassebaum, N., Kawakami, N., Khang, Y. H., Khatibzadeh, S., Khoo, J. P., Kok, C., Laden, F., Lalloo, R., Lan, Q., Lathlean, T., Leasher, J. L., Leigh, J., Li, Y., Lin, J. K., Lipshultz, S. E., London, S., Lozano, R., Lu, Y., Mak, J., Malekzadeh, R., Mallinger, L., Marcenes, W., March, L., Marks, R., Martin, R., McGale, P., McGrath, J., Mehta, S., Mensah, G. A., Merriman, T. R., Micha, R., Michaud, C., Mishra, V., Hanafiah, K. M., Mokdad, A. A., Morawska, L., Mozaffarian, D., Murphy, T., Naghavi, M., Neal, B., Nelson, P. K., Nolla, J. M., Norman, R., Olives, C., Omer, S. B., Orchard, J., Osborne, R., Ostro, B., Page, A., Pandey, K. D., Parry, C. D., Passmore, E., Patra, J., Pearce, N., Pelizzari, P. M., Petzold, M., Phillips, M. R., Pope, D., Pope, C. A., III, Powles, J., Rao, M., Razavi, H., Rehfuess, E. A., Rehm, J. T., Ritz, B., Rivara, F. P., Roberts, T., Robinson, C., Rodriguez-Portales, J. A., Romieu, I., Room, R., Rosenfeld, L. C., Roy, A., Rushton, L., Salomon, J. A., Sampson, U., Sanchez-Riera, L., Sanman, E., Sapkota, A., Seedat, S., Shi, P., Shield, K., Shivakoti, R., Singh, G. M., Sleet, D. A., Smith, E., Smith, K. R., Stapelberg, N. J., Steenland, K., Stockl, H., Stovner, L. J., Straif, K., Straney, L., Thurston, G. D., Tran, J. H., Van Dingenen, R., van Donkelaar, A., Veerman, J. L., Vijayakumar, L., Weintraub, R., Weissman, M. M., White, R. A., Whiteford, H., Wiersma, S. T., Wilkinson, J. D., Williams, H. C., Williams, W., Wilson, N., Woolf, A. D., Yip, P., Zielinski, J. M., Lopez, A. D., Murray, C. J., Ezzati, M., AlMazroa, M. A., and Memish, Z. A. (2012), "A Comparative Risk Assessment of Burden of Disease and Injury Attributable to 67 Risk Factors and Risk Factor Clusters in 21 Regions, 1990–2010: A Systematic Analysis for the Global Burden of Disease Study 2010," *Lancet*, 380, 2224–2260. [235]

Maletic, J. I., and Marcus, A. (2000), "Data Cleansing: Beyond Integrity Analysis," in *Proceedings of the Conference on Information Quality*. Available at *http://www.sdml.info/papers/IQ2000.pdf* [236]

Roberts, B. L., Anthony, M. K., Madigan, E. A., and Chen, Y. (1997), "Data Management: Cleaning and Checking," *Nursing Research*, 46, 350–352. [237]

Thompson, J. M., Clark, P. M., Robinson, E., Becroft, D. M., Pattison, N. S., Glavish, N., Pryor, J. E., Wild, C. J., Rees, K., and Mitchell, E. A. (2001), "Risk Factors for Small-for-Gestational-Age Babies: The Auckland Birthweight Collaborative Study," *Journal of Paediatrics Child Health*, 37, 369–375. [238]

Winkler, W. E. (1998), "Problems With Inliers," *Bureau of the Census Research Reports Series RR98/05*. Available at *https://www.census.gov/srd/papers/pdf/rr9805.pdf* [238]