Australian Critical Care xxx (xxxx) xxx

Contents lists available at ScienceDirect

Australian Critical Care

journal homepage: www.elsevier.com/locate/aucc

Methodological series

Data cleaning for clinician researchers: Application and explanation of a data-quality framework

Julia K. Pilowsky, RN, PhD [a, b, c, *], Rosalind Elliott, RN, PhD [b, c, d], Michael A. Roche, RN, PhD [b, e]

[a] Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia; [b] Faculty of Health, University of Technology Sydney, Sydney, NSW, Australia; [c] Royal North Shore Hospital, Northern Sydney Local Health District, Sydney, NSW, Australia; [d] Nursing and Midwifery Directorate, Northern Sydney Local Health District, Sydney, NSW, Australia; [e] University of Canberra and ACT Health Directorate, Canberra, ACT, Australia

ARTICLE INFORMATION

Article history:
Received 11 September 2023
Received in revised form
22 February 2024
Accepted 10 March 2024

Keywords:
Data cleaning
Data management
Initial data analysis
Quantitative methods

ABSTRACT

Background: Data cleaning is the series of procedures performed before a formal statistical analysis, with the aim of reducing the number of error values in a dataset and improving the overall quality of subsequent analyses. Several study-reporting guidelines recommend the inclusion of data-cleaning procedures; however, little practical guidance exists for how to conduct these procedures.

Objectives: This paper aimed to provide practical guidance for how to perform and report rigorous data-cleaning procedures.

Methods: A previously proposed data-quality framework was identified and used to facilitate the description and explanation of data-cleaning procedures. The broader data-cleaning process was broken down into discrete tasks to create a data-cleaning checklist. Examples of the how the various tasks had been undertaken for a previous study using data from the Australia and New Zealand Intensive Care Society Adult Patient Database were also provided.

Results: Data-cleaning tasks were described and grouped according to four data-quality domains described in the framework: data integrity, consistency, completeness, and accuracy. Tasks described include creation of a data dictionary, checking consistency of values across multiple variables, quantifying and managing missing data, and the identification and management of outlier values. The data-cleaning task checklist provides a practical summary of the various aspects of the data-cleaning process and will assist clinician researchers in performing this process in the future.

Conclusions: Data cleaning is an integral part of any statistical analysis and helps ensure that study results are valid and reproducible. Use of the data-cleaning task checklist will facilitate the conduct of rigorous data-cleaning processes, with the aim of improving the quality of future research.

1. Introduction

Data cleaning is the process by which raw data are transformed into data that are of an appropriate quality for formal statistical analysis. It involves the identification and management of incorrect data and is vital for ensuring the validity and reproducibility of research findings. Data cleaning should be performed on all datasets before analysis, regardless of whether they have been collected manually or extracted from an electronic health record or clinical registry as all datasets may contain incorrect or missing data.[1] Reporting guidelines for both observational and predictive modelling studies recommend that data-cleaning processes are reported, particularly aspects relating to missing and outlier data; however, this fails to occur in a significant number of publications.[2–5] This paper aims to provide clinician researchers with guidance to perform and report rigorous data-cleaning procedures according to a previously proposed data-quality framework.[6] Practical examples of how these procedures were undertaken on the Australia and New Zealand Intensive Care Society (ANZICS)

[*] Corresponding author at: Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia.
E-mail addresses: Julia.Pilowsky@sydney.edu.au, Julia.Pilowsky@health.nsw.gov.au (J.K. Pilowsky).

https://doi.org/10.1016/j.aucc.2024.03.004
1036-7314/© 2024 Australian College of Critical Care Nurses Ltd. Published by Elsevier Ltd. All rights reserved.

Adult Patient Database (APD) are also provided to illustrate the theoretical concepts described in the framework.

2. Methods

2.1. Data-quality framework

The data-quality framework used to facilitate the data-cleaning procedures described in this paper consists of four domains: integrity, consistency, completeness, and accuracy.[6] Data integrity refers to defining structural and technical aspects of the dataset and involves tasks such as ensuring all data are in the correct format and all variables required for the analysis are present. Data consistency refers to how correct the data are, which is evaluated by identifying impossible values in individual variables, and by identifying contradictions when assessing several variables together. Data completeness refers to understanding the amount of data missing in the dataset. Data accuracy involves assessing variables to identify potential errors or unexpected values through statistical testing rather than through comparison with other variables as was performed in the data consistency checks. An overview of the four domains and the tasks that are performed in each is provided in Table 1 and can be used as a checklist by clinician researchers to facilitate their own data-cleaning processes. Items in the checklist are referred to throughout the article, and examples are provided to illustrate how the task associated with each item could be completed. The items in the checklist are intended to be performed in the order they are described as some items are contingent on others, for example, ensuring all data are formatted correctly (Item 1.2) before undertaking any filtering procedures (Item 1.4). It should also be noted that this checklist is intended to be used when analysing data that have already been obtained or collected for a study and not when setting up a study that may involve prospective data collection. However, it may be useful to refer to the checklist during this initial preparation phase as many of the suggested items can be addressed while collecting data by building various checks into the data collection software that will prevent the entry of incorrect values.

2.2. Data source

The examples described in this paper illustrate the data-cleaning processes performed on data from the ANZICS APD obtained for a cohort study examining the effect of pre-existing mental health disorders in patients admitted to the intensive care unit (ICU).[7] The study linked the APD to data from the electronic medical record (eMR) to determine the prevalence of pre-existing mental health disorders in the cohort and then performed several multivariable analyses to investigate potential associations between the presence of a pre-existing mental health disorder and outcomes such as mortality and the need for mechanical ventilation. The ANZICS APD is a clinical registry maintained by ANZICS

Table 1
Data-cleaning task checklist.

Item name	Task description	✓
Domain 1: Data integrity		
1.1 Create data dictionary	Create a document that lists all variables present in the datasets along with relevant information including variable name, variable definition, value labels, and variable data type (e.g., continuous, categorial).	□
1.2 Format variables	Ensure all variables are formatted according to the data dictionary and have been imported into the analysis program correctly.	□
1.3 Define research question	Clarify the research question being asked and consider how the various aspects of the question will be defined in the data (e.g., patient population, outcome measures).	□
1.4 Remove inappropriate data	Apply study inclusion/exclusion criteria to the dataset and remove any inappropriate entries (e.g., readmissions, duplicate entries).	□
1.5 Assess variables	Determine whether variables are appropriate for answering the research question in their current form or if they require transformation, including the creation of new variables (e.g., convert postcode to a measure of socioeconomic disadvantage, aggregate admitting diagnoses into broad categories).	□
1.6 Assess overall dataset	Based on the findings of the aforementioned items, consider whether the dataset is appropriate for use in the proposed study (e.g., relevant variables are available, dataset is sufficiently representative of population of interest).	□
Domain 2: Data consistency		
2.1 Identify impossible values	Assess each variable for impossible values such as those considered to be biologically implausible or categories not listed in the dataset data dictionary.	□
2.2 Identify inconsistent values	Identify variables that are related to each other and ensure values are consistent (e.g., ventilation duration [hours] and ventilation indicator [yes/no] or sex [male/female] and pregnancy status [yes/no]).	□
Domain 3: Data completeness		
3.1 Identify missing data	Ensure missing data are coded consistently across the dataset and are coded differently to not-applicable data (e.g., ventilation duration when ventilation indicator is no).	□
3.2 Quantify missing data	Calculate the percentage of missing data in each variable.	□
3.3 Assess missing data	Assess the missingness pattern and determine whether data are missing completely at random, at random, or not at random.	□
3.4 Manage missing data	Determine how missing data will be handled in the study analysis and report this as part of the study methods.	□
Domain 4: Data accuracy		
4.1 Assess continuous variables	Visualise continuous variables to determine if outliers are present and how the data are distributed (e.g., normally, skewed)	□
4.2 Identify outlier data	Use an appropriate statistical method to identify outliers in the dataset.	□
4.3 Evaluate outlier data	Assess datapoints identified as outliers and correct any errors found.	□
4.4 Manage outlier data	Choose an appropriate method for handling outlier data and report this as part of the study methods.	□
4.5 Sensitivity analysis	If removing outliers from a dataset, perform a sensitivity analysis to quantify the impact this has on the study results.	□

Centre for Outcome and Resource Evaluation, and while its primary use is for quality assurance, use for research purposes is also encouraged.[8] The APD contains a range of variables relating to a patient's ICU admission, including severity of illness scores, reason for admission to the ICU, interventions within the ICU such as mechanical ventilation and renal replacement therapy, and outcomes such as ICU mortality and length of stay.[9] Ethical approval for the cohort study was obtained from the Northern Sydney Local Health District Human Research Ethics Committee (2020/ETH02376). All analyses were conducted in R version 4.0.4.

3. Results and discussion

3.1. Data integrity

Before commencing any data-cleaning processes, it is important to define the structural and technical aspects of the dataset, also known as the metadata.[10] The study data dictionary (Item 1.1) should summarise all relevant metadata including a brief definition for each variable, the type of data for each variable, and how this data type should appear when imported into the statistical analysis program.[11] Confirming the data type for each variable ensures that appropriate statistical tests are selected, and checking that all variables are formatted correctly avoids errors such as when a categorical variable with multiple levels is considered to be a continuous variable by the analysis program (Item 1.2). A detailed data dictionary is an important tool to refer to throughout the conduct of a study, particularly if there are multiple collaborators performing the data collection or analysis.[1,12] Some datasets may also have been collected over a long period of time, and the types or definitions of included variables may have changed over this period. Researchers should refer to documentation for the dataset, which is contemporaneous to the period of data being analysed to ensure that the definitions being used to guide the analysis match those used when the data was originally collected.

Once the researcher has a clear understanding of the study dataset, they should clarify their research question (Item 1.3). The patient population for the study should be defined and a set of inclusion and exclusion criteria developed to capture this group. These criteria can then be applied to the dataset, and any inappropriate entries should be removed (Item 1.4). For example, a study may have an exclusion criterion related to patient age and so should remove any entries in the dataset accordingly. If the researcher chooses to exclude any cases at this stage due to missing data, for example, in the primary outcome variable, then the potential bias this introduces into the study should also be considered (see “Data Completeness” section for a more detailed discussion of missing data). Any duplicate entries should also be removed at this stage. Duplicates can occur when data about a single case are accidentally recorded more than once in a dataset, resulting in multiple entries being present about the same case. Duplicate entries usually have the same values in all variables, although some duplicates may contain some minor differences. Finally, the researcher should determine whether the dataset contains all variables required to answer their question or if new variables will need to be created to facilitate this (Item 1.5). For example, if the aim of a study is to examine the impact of an intervention on ICU length of stay but the dataset only contains variables for ICU admission and discharge date/time, then a new variable will need to be created that calculates the time difference between these two variables.

Once these initial tasks have been completed, the researcher should consider whether the structural and technical aspects of the dataset are adequate to complete their proposed analysis (Item 1.6). For example, if the dataset being analysed was originally collected for another purpose, the researcher should consider whether it contains variables relevant to the phenomenon of interest and whether it is sufficiently representative of the population they intend to apply their findings to. For datasets collected over a long period of time, the researcher will also need to consider if and how the dataset has changed, and whether this will affect the analysis. For example, in addition to changes in the types of variables collected, the scope of a dataset may have also changed over time to include a greater or fewer number of participants, which may influence how the study's findings should be interpreted. Any limitations identified at this point should be noted and reported along with the overall results of the study.

3.1.1. Cohort study example—data integrity

For the cohort study, the data dictionary for the ANZICS APD was used to obtain the initial metadata and create the first version of the study data dictionary.[9] Information relating to the data type, and how that type should appear when imported into R, was also added (Item 1.1, Table 2). The only exclusion criterion for the cohort study was readmissions to ICU within the same hospitalisation, and so episodes of care flagged as being a readmission were removed from the dataset (Item 1.4). Next, each variable in the APD dataset was assessed to determine if it required transformation before performing any further analyses to answer the study question (Item 1.5). Several variables were transformed, including conversion of

Table 2
Example data dictionary.

Variable name	Variable description	Value labels	Variable data type	Variable data type (R)
Age	Age at hospital admission (years)	N/A	Continuous	Numeric
ICU_SRCE	ICU admission source	1 = OT or PACU 2 = ED 3 = Ward 4 = ICU same hospital 5 = Other hospital 6 = ICU other hospital 9 = Home	Categorical	Factor
SOCEC_SUM	Index of Relative Socio-Economic Disadvantage—summary measure derived from postcode variable	1 = Most disadvantage 2 = Medium disadvantage 3 = Least disadvantage	Ordinal	Ordered
INV_IND	Any type of invasive ventilation during ICU stay	1 = Yes 2 = No	Categorical	Factor

Abbreviations: ED = emergency department; ICU = intensive care unit; PACU = post-anesthaesia care unit.

the postcode variable to indices of relative socioeconomic disadvantage and remoteness and aggregation of the Acute Physiology and Chronic Health Evaluation III-J Diagnosis codes into broader diagnostic categories. The study data dictionary was updated to reflect these changes, an excerpt of which is presented in Table 2.

3.2. Data consistency

The consistency, or correctness, of a dataset should be assessed in two stages. First, each variable should be considered in isolation to identify any impossible values (Item 2.1). For continuous variables, biologically implausible values should be defined and either corrected, if possible, or removed from the dataset. For categorical variables, any value labels not present in the data dictionary should also be either corrected or removed. For example, if a categorical variable has only two value labels listed in the data dictionary, 1 and 2, then any value other than these should be considered an error and either be marked as missing or replaced with the correct value if available. Wherever possible, correct values should be obtained through verification against the source from which the data was derived, commonly the patient's medical record. In some studies involving multiple datasets, it may also be possible to determine correct values by cross checking which values are recorded in these other datasets. It should also be noted that many types of consistency checks can be built into data collection software, including commonly used platforms such as Research Electronic Data Capture, which can reduce the number of errors made during the data collection process.[13]

Next, potentially contradictory values should be assessed by considering multiple variables together (Item 2.2). For example, a variable relating to pregnancy status could be considered together with the sex variable to identify any males with a positive pregnancy status. Again, these contradictory values should be checked and replaced with the correct value wherever possible, either through source data verification or by cross referencing other values in the available datasets. If the researcher is unable to access the source from which the data were derived, then they may potentially choose to make an assumption about which of two contradictory variables is likely to be more correct. Assumptions require a thorough understanding of the context in which the data is collected to be able to make an appropriately informed judgement which contradictory variable is likely to be correct. It should also be noted that making assumptions can lead to inaccuracies and biases being introduced into a study and so should be made with caution and be clearly recorded as part of the analysis process. In most cases, where the researcher cannot verify a contradictory or implausible value against source data, or through cross-checking against other values in the dataset, they should simply consider the values as errors and mark them as missing.

3.2.1. Cohort study example—data consistency

For the cohort study, biologically implausible values were defined for each continuous variable in the APD. For example, for the age variable, an upper limit of 110 years was chosen, and all recorded ages greater than this were removed and marked as missing. To check for impossible values in the categorical variables, the APD data dictionary was consulted to identify permissible value labels.[9] All value labels in the dataset were listed in the data dictionary, and so no values were removed. A series of multivariate consistency checks were also performed on the APD data (Item 2.2, Table 3). When performing these checks, it was assumed that variables relating to the ICU admission were more correct than those relating to the hospital admission, given that the data collectors for the APD are based in the ICU. Many of these checks are also built into the software used to collect and submit data to the APD meaning that very few of these errors were identified (Table 3).

3.3. Data completeness

Assessments in the data completeness domain relate to identifying and managing missing data in the dataset. A separate code should be used to differentiate missing data from data that is not applicable to that patient, for example, pregnancy status in a male patient (Item 3.1). This will enable researchers to accurately calculate the amount of missing data across their study dataset (Item 3.2), which should be reported as the percentage of data missing for each variable. Understanding patterns of missingness in a dataset not only informs which type of method should be used to manage missing data in the study analyses but can also provide insights into the study cohort itself such as identifying which patient characteristics are associated with higher levels of missing data.[10] Frameworks for how to manage missing data specifically, rather than as part of the overall data-cleaning procedures for a study, have also been proposed, which can be used in conjunction with the brief guidance provided here.[14,15]

Missing data can be classified either as missing completely at random, missing at random, or missing not at random, depending on whether the patterns of missingness are related to or dependent upon other variables in the dataset associated with the research question.[6,16] For example, if follow-up data for patients in the intervention arm of a clinical trial are more likely to be missing than for patients in the control arm, then clearly there is a pattern to the missingness of data, and failing to account for this will lead to a biased result. It is equally important to understand the amount and

Table 3
Exemplar consistency checks performed (Item 2.2) and updates made to APD data.

Variable name	Consistency check
HOSP_SRCE ICU_SRCE	For all ICU admissions with a source ‘Other hospital’ or ‘Other hospital ICU’ the hospital source should also be ‘Other hospital’. HOSP_SRCE was updated to reflect this (12 values, 0.1% of available data)
HOSP_OUTCM ICU_OUTCM DIED DIED_ICU DIED_HOSP ICU Discharge Date/Time Hospital Discharge Date/Time	If ICU_OUTCM is died and ICU and hospital discharge dates are the same, then HOSP_OUTCM should also be died. The variables DIED_ICU, DIED_HOSP, and DIED should also all be ‘yes’ and were updated to reflect this (5 values, 0.03% of available data). If HOSP_OUTCM is survived, then ICU_OUTCM should also be survived, and so was updated to reflect this (5 values, 0.03% of available data).
ICU Admission Date/Time Hospital Admission Date/Time	Hospital admission date/time should be less than or equal to ICU admission date/time. Any hospital admission date/time values that occurred after the ICU admission date/time were updated to equal the ICU admission date/time (65 values, 0.4% of available data).

Abbreviations: APD = Adult Patient Database; ICU = intensive care unit.

Please cite this article as: Pilowsky JK et al., Data cleaning for clinician researchers: Application and explanation of a data-quality framework, Australian Critical Care, https://doi.org/10.1016/j.aucc.2024.03.004

type of missing data when conducting observational research, particularly when using administrative datasets. Selection bias can be introduced into observational studies if they depend on certain variables being recorded in the eMR for use as outcome data, such as attendance at follow-up appointments.[17] Attendance at these appointments may correlate with other variables such as age or chronic health conditions, meaning particular subgroups may be underrepresented in the study population. This again highlights the importance of understanding the clinical context in which administrative data are collected in order to make informed judgements regarding why a particular variable might have contained missing data, and whether those data are missing at random or not. For example, it might be routine for elective surgical patients to have their height and weight recorded in the eMR during a visit to a preadmission clinic, and therefore, these patients will be less likely to have missing data in these variables than emergency medical patients.

Once the amount and type of missing data has been determined, a method for handling the missing values needs to be selected. Multiple methods for handling missing data exist, although they can be grouped into two main categories: those which impute the missing data and those which exclude it.[18] One common imputation method simply replaces missing values with the median, mean, or most frequent value in the dataset. This method creates bias by reducing the variability in the dataset, thereby resulting in falsely precise estimates, and is not recommended.[19] A potentially less biased method for dealing with missing data is multiple imputation, although when performing a type of multiple imputation, it is recommended that expert statistical advice is obtained as these procedures can be complex and require multiple assumptions about the data to be met.[20] Alternatively, it may be appropriate to exclude variables with large amounts of missing data from the analysis. Excluding patients with missing outcome data is known as a complete case analysis, and while this may be appropriate in large datasets with low levels of missing data, it can also lead to bias if a particular subgroup is more likely to have missing outcome data than the rest of the cohort.[21]

3.3.1. Cohort study example—data completeness

When assessing missing data in the APD for the cohort study, levels of missing data were found to vary widely and were largely dependent on whether the variable was deemed mandatory for

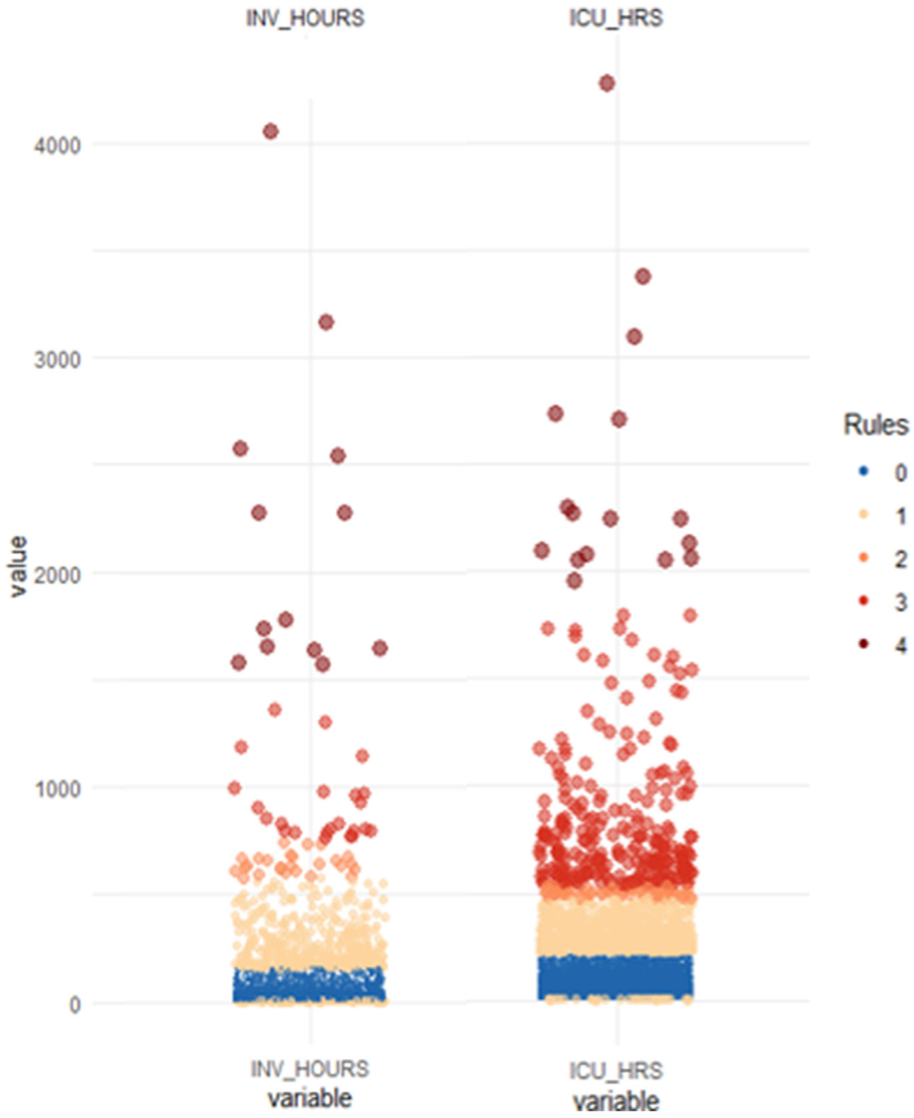

Fig. 1. Scatterplots of the ICU length of stay (ICU_HRS) and hours of invasive ventilation (INV_HOURS) variables with the number of statistical rules (Rules) classifying each datapoint as an outlier (Item 4.2).

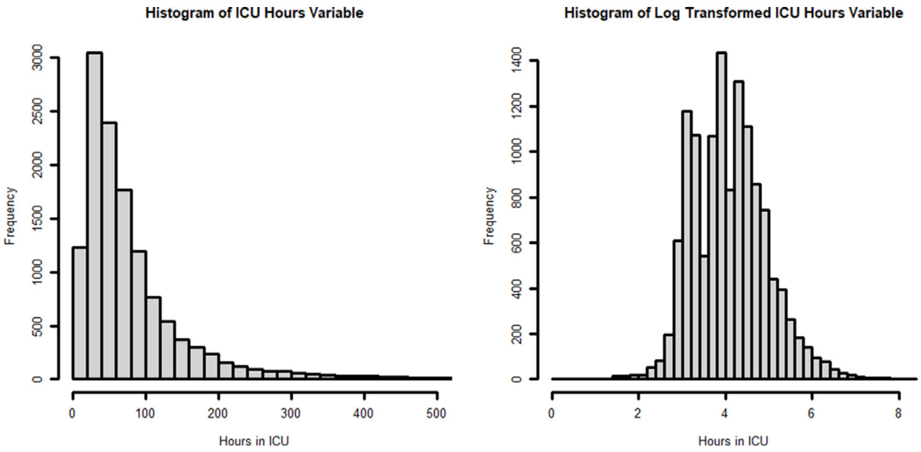

Fig. 2. ICU length of stay variable before and after log transformation (items 4.1 and 4.4).

collection by ANZICS or not.[9] Overall, the final amount of missing data present in the cohort study dataset was low, with over half the variables having no missing data, and only seven (9.7%) having more than 10% missing. Given the low levels of missing data, and the large size of the overall study cohort (n = 16,228), a complete case analysis was used for multivariable analyses conducted in the study.

3.4. Data accuracy

Similar to data consistency, assessments of data accuracy are concerned with how correct the data are. However, unlike data consistency, which determines correctness by comparing variables against each other, measures of data accuracy are derived by considering each variable in isolation and by using a variety of statistical techniques to assess if the value of a datapoint may be incorrect or unexpected. Datapoints that lie outside the expected data distribution are known as outliers.[22] Three broad categories have been proposed for classifying outliers, namely error, interesting, and random outliers.[23,24] Error outliers are incorrect data, for example, due to data-entry error. Interesting outliers are true datapoints that when considered individually as case studies, or together as a separate subgroup, have the potential for generating new knowledge. For example, by investigating the causes behind an outlier ICU with an unusually high mortality rate, the negative impact of low staffing levels and high rates of after-hours discharge were able to be demonstrated.[25] Finally, random outliers are true datapoints that fall outside the expected distribution due to chance rather than underlying reasons worthy of investigation. Outliers can also be classified either as univariate, when outliers in a particular variable are considered in isolation, or multivariate when a datapoint only becomes an outlier when additional dimensions are added to the variable. For example, a patient's weight might be considered to be within the normal range; however, their body mass iIndex, which includes both height and weight dimensions, may be assessed as an outlier.

There is no consensus definition for how to identify an outlier.[23] Multiple methods for detecting these values are available, especially when assessing outliers in non-normally distributed or heavily skewed data, such as length of stay.[24,26] Similarly, there is no consensus on how outliers should be handled, beyond correcting or removing any error outliers and clearly reporting how all outliers were handled.[23] If considering removing outliers from the dataset, the researcher should also perform a sensitivity analysis to quantify the difference observed when outliers are excluded from the dataset and the analysis. An alternate method for handling outliers is to winsorize the variable, which replaces all values above or below a specified percentile with values at that percentile of the distribution.[27] For variables with a highly skewed distribution, other types of transformation can be considered such as a logarithmic transformation, which reduces the influence of outlier datapoints.[28] A detailed discussion of how to identify and manage outliers can be found in the methodology paper by Mowbray et al.[22]

3.4.1. Cohort study example—data accuracy

When assessing univariate outliers in continuous variables in the APD data for the cohort study (Item 4.1), four different statistical methods were used (Tukey, 6-Sigma, Hubert and Sigma-gap) as implemented in the R *dataquieR* package.[29–32] Fig. 1 shows scatterplots for two of the assessed variables (ICU length of stay and hours of invasive ventilation) with the different colours highlighting how many statistical methods or rules classified each datapoint as an outlier. Outliers in the ICU length of stay variable were left unchanged for the descriptive analyses in the study; however, the variable was log transformed before use in the multivariable analyses (Item 4.4). Histograms of the ICU length of stay variable before and after log transformation are shown in Fig. 2.

4. Conclusion

This article has provided an overview of how a previously proposed data-cleaning framework can be applied to the ANZICS APD. The four domains, namely data integrity, consistency, completeness, and accuracy were explained, along with practical examples of how the various domain assessments should be conducted. A summary of tasks for the four domains is provided in the form of a data-cleaning checklist (Table 3) to enable clinician researchers to conduct comprehensive data-cleaning assessments and improve the quality of future clinical research.

Funding

Dr Pilowsky was supported by an Australian Government Research Training Program Scholarship and a UTS Research Excellence Scholarship. The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources.

CRediT authorship contribution statement

Julia Pilowsky: Conceptualisation (lead), Data Curation, Software and Formal Analysis (lead), Funding Acquisition (lead),

Please cite this article as: Pilowsky JK et al., Data cleaning for clinician researchers: Application and explanation of a data-quality framework, Australian Critical Care, https://doi.org/10.1016/j.aucc.2024.03.004

Writing—Original Draft Preparation. **Rosalind Elliott**: Conceptualisation (supporting), Data Curation, Software and Formal Analysis (supporting), Resources (lead), Funding Acquisition (supporting), Writing—Review and Editing (equal), Supervision (supporting). **Michael Roche**: Conceptualisation (supporting), Data Curation, Software and Formal Analysis (supporting), Resources (supporting), Funding Acquisition (supporting), Writing—Review and Editing (equal), Supervision (lead). All authors read and approved the final manuscript.

Conflict of interest

The authors declare they have no conflict of interest.

Data availability statement

Data used for the illustrative examples described in this paper are not publicly available, and any requests should be made to the ANZICS CORE committee.

Acknowledgements

The authors would like to thank the Australian and New Zealand Intensive Care Society Centre for Outcome and Resource Evaluation for providing the data for the study, from which the illustrative examples described in this paper were derived. The authors would also like to thank the clinicians, data collectors, and data managers at the sites who contributed data to the Australian and New Zealand Intensive Care Society Adult Patient Database.

References

[1] Arts DGT, de Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inf Assoc 2002;9(6):600–11.

[2] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335(7624):806–8.

[3] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol 2015;68(2):112–21.

[4] Huebner M, Vach W, le Cessie S, Schmidt CO, Lusa L, Cook D, et al. Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Methodol 2020;20(1):61.

[5] Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015;12(10):e1001885.

[6] Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021;21(1):63.

[7] Pilowsky JK, Elliott R, Roche MA. Association between preexisting mental health disorders and adverse outcomes in adult intensive care patients: a data linkage study. Crit Care Med 2023;51(4):513–24.

[8] Secombe P, Millar J, Litton E, Chavan S, Hensman T, Hart GK, et al. Thirty years of ANZICS CORE: a clinical quality success story. Crit Care Resusc 2023;25(1):43–6.

[9] Australia and New Zealand Intensive Care Society. APD data dictionary version 5.10. 2020.

[10] Huebner M, le Cessie S, Schmidt CO, Vach W. A contemporary conceptual framework for initial data analysis. Observ Stud 2018;4(1):171–92.

[11] Richter A, Schössow J, Werner A, Schauer B, Radke D, Henke J, et al. Data quality monitoring in clinical and observational epidemiologic studies: the role of metadata and process information. GMS Med Inform Biom Epidemiol 2019;15.

[12] Buchanan EM, Crain SE, Cunningham AL, Johnson HR, Stash H, Papadatou-Pastou M, et al. Getting started creating data dictionaries: how to create a shareable data set. Adv Methods Pract Psychol Sci 2021;4(1): 2515245920928007.

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

[14] Carpenter JR, Smuk M. Missing data: a statistical framework for practice. Biom J 2021;63(5):915–47.

[15] Lee KJ, Tilling KM, Cornish RP, Little RJ, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: the treatment and reporting of missing data in observational studies framework. J Clin Epidemiol 2021;134:79–88.

[16] Rubin DB. Inference and missing data. Biometrika 1976;63(3):581–92.

[17] Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. JAMA Netw Open 2021;4(2):e210184.

[18] Bell ML, Fiero M, Horton NJ, Hsu C-H. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol 2014;14(1):118.

[19] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338.

[20] Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Med Res Methodol 2017;17(1):162.

[21] Altman DG, Bland JM. Missing data. BMJ 2007;334(7590):424.

[22] Mowbray FI, Fox-Wasylyshyn SM, El-Masri MM. Univariate outliers: a conceptual overview for the nurse researcher. Can J Nurs Res 2019;51(1):31–7.

[23] Leys C, Delacre M, Mora YL, Lakens D, Ley C. How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. Int Rev Soc Psychol 2019;32(1):1–10.

[24] Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organ Res Methods 2013;16(2):270–301.

[25] McClean K, Mullany D, Huckson S, van Lint A, Chavan S, Hicks P, et al. Identification and assessment of potentially high-mortality intensive care units using the ANZICS Centre for Outcome and Resource Evaluation clinical registry. Crit Care Resusc 2017;19(3):230–8.

[26] Verardi V, Vermandele C. Univariate and multivariate outlier identification for skewed or heavy-tailed distributions. STATA J 2018;18(3):517–32.

[27] Steyerberg EW. Coding of categorical and continuous predictors. In: Steyerberg EW, editor. Clinical prediction models: a practical approach to development, validation, and updating. 2nd ed. Springer International Publishing AG; 2019. p. 175–90.

[28] Bland JM, Altman DG. Statistics notes: transformations, means, and confidence intervals. BMJ 1996;312(7038):1079.

[29] University Medicine Greifswald, Richter A, Schmidt CO, Struckmann S. acc_univariate_outlier: function to identify univariate outliers by four different.... 2021. Available from: https://rdrr.io/cran/dataquieR/man/acc_univariate_outlier.html.

[30] Bakar ZA, Mohemad R, Ahmad A, Deris MM. A comparative study for outlier detection techniques in data mining. In: Proceedings of the 2006 IEEE conference on cybernetics and intelligent systems; 2006. p. 1–6.

[31] Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. Comput Stat Data Anal 2008;52(12):5186–201.

[32] Tukey JW. Exploratory data analysis. Reading, Mass. 1977.