The European Journal of Social and Behavioural Sciences
EJSBS
ISSN: 2301-2218 (online)

The European Journal of Social and Behavioural Sciences
EJSBS Volume XIII, Issue II (e-ISSN: 2301-2218)

A FRAMEWORK TO ASSESS HEALTHCARE DATA QUALITY

William Warwick[a], Sophie Johnson[a], Judith Bond[a], Geraldine Fletcher[a], Pavlo Kanellakis[a]*

[a]*Health Education West Midlands, UK*

Abstract

Assessing data quality is a fundamental task during the research process. Information derived from data of inadequate quality may lead to invalid conclusions and misinformed management decisions for healthcare organisations. To minimise such risk a data quality framework can be utilised to ensure suitability and to quality assure datasets. This current research involved a review of existing frameworks and the formation of a new framework which combines quality criteria derived from different research disciplines. The current framework is robust, it can effectively assess data quality across a range of criteria and supports researchers to formulate a decision on whether to use the dataset. Further development of the quality framework would include an emphasis on the interdependencies of quality criteria.

Keywords: Healthcare, healthcare organisations, data quality, DatQAM

*Corresponding author.
E-mail address: patopavlokan@gmail.com

doi: 10.15405/ejsbs.156

https://doi.org/10.15405/ejsbs.156
eISSN: 2301-2218 / Corresponding Author: Pavlo Kanellakis
Selection & Peer-review under responsibility of the Editors

1. Introduction

Data quality assessment is a fundamental task when undertaking research. A wealth of healthcare data provided from the National Health Service is available and can be easily accessed and utilised for research. Even though health related datasets are obtained from authoritative sources, issues within the quality of data may be apparent. Data quality issues can lead to an array of errors within research findings including incorrect demographical information and exaggeration of disorder prevalence. Moreover, the consequences of decisions made from inaccurate results can be damaging to organisations within the healthcare sector (Goodchild, 1993).

It is therefore important to use a framework to assess the quality of data obtained from the data mining process. This will help determine whether it can be used to test hypotheses, and increase confidence of validity.

The motivation for this current research was to investigate data quality issues encountered for a research report undertaken by the NHS Coventry and Warwickshire Partnership Trust entitled 'Up Skilling the Adult Mental Health Workforce in Psychological Practice Skills'. As researchers had access to a wealth of data from several sources, it was important to examine the data available to the research and what data quality criteria would be necessary to draw conclusions on its suitability. As many of the available datasets had not been collected with a specific research question, the selection quality and methods were not under control of the research and therefore, were difficult to validate (Sorensen, Sabroe & Olsen, 1996). From this, there was a need to construct a robust framework to assess the quality of data. This led to a review of existing frameworks and the formation of a new framework specific for this research.

2. Review of Quality Framework

The existing literature instigated that the criteria for a quality framework must be general, applicable across application domains and data types and clearly defined, Price & Shanks, (2004). Eppler (2001), put forward that quality frameworks should show interdependencies between different quality criteria, to allow researchers to become familiar with how data quality issues impact other criteria.

The Data Quality Assessment Methods and Tools (DatQAM) provides a systematic implementation of data quality assessment which includes a range of quality measures which considers the strengths of official statistics. It is concerned with user satisfaction concerning relevance, sampling and non-sampling errors, production dates concerning timeliness,

https://doi.org/10.15405/ejsbs.156
eISSN: 2301-2218 / Corresponding Author: Pavlo Kanellakis
Selection & Peer-review under responsibility of the Editors

availability of metadata and forms for dissemination, changes over time and geographical differences and coherence (Eurostat, 2007).

The Quality Assurance Framework (QAF) developed by Statistics Canada (2010) includes a number of quality measures for assessing data quality including measures for timeliness, relevance, interpretability (completeness of metadata), accuracy (coefficient of variance, imputation rates), coherence and accessibility. These two data quality (DQ) frameworks are similar in the way that they consider measures for data quality and for the data quality criteria themselves. They are also widely used, an example of this is that the HSCIC uses the DatQAM for data quality assessments (HSCIC, 2013).

In order to build a framework which considers measures for DQ we can consider these two frameworks and how the criteria are measured within them in order to gain a comprehensive framework that can be applied to data that we use within our research. These measures have been adapted from the DatQAM and QAF frameworks in order to quantify our data quality assessments.

Furthermore, the World Health Organisation's (WHO) 'quality criteria' was utilised in order to categorise the quality measurements. The Data Quality Audit Tool (DQAT) is utilised by the WHO and Global Fund. After cross referencing it was decided that a 'confidentiality' criteria be added to the framework which was adapted from the DQAT (2008).

Data quality criteria often influence the execution of data cleansing methods on raw data (Muller & Freytag, 2005) Considering that the data that is used in a project such as this often comes from authoritative sources, it is likely that data cleansing methods have already been undertaken. As a consequence, a criteria was added to the current framework to assess whether data cleansing methods were already implemented.

The current framework highlights whether sufficient practice was carried out during data collection, these criteria were adapted from a Quality Category Information Framework by Price and Shanks (2004). The researchers favoured an integrated quality framework using intuitive, empirical and theoretical approaches to ascertain rigour and scope. Aspects of Price and Shanks (2004) framework criteria regarded the objectivity of the dataset, so whether the dataset is completely independent of user or use. The current research implements this measure, to prompt the researcher to examine objectivity.

Previous work by Eppler (2001) Price and Shanks (2004) outlines accessibility criteria. For example, prompting users to examine whether access to the data needs to be authorised, and question whether the dataset has been protected from bias and cannot be

https://doi.org/10.15405/ejsbs.156
eISSN: 2301-2218 / Corresponding Author: Pavlo Kanellakis
Selection & Peer-review under responsibility of the Editors

misused. The current research supports the notion of covering a broad range of research and data management processes to ensure efficient practice.

3. Final Data Quality Framework

The current framework aims to highlight good data management practice and data issues and inaccuracies. The user will be prompted to record any possible resolutions for data inaccuracies, for example, requesting missing or incomplete data. A validity criteria has also been added so that contradictions between datasets from different sources are highlighted. Data sets with negative reports cannot pass the validation criteria however may not fail if engagement with researchers is evident and/or the data is subject to change or indeed inaccuracies have been corrected. The table below presents the criteria, measurements and definitions of the current quality framework (see table 1).

Table 1. Final Data Quality Framework with Definitions

Criteria	**Measurement**	**Definition**
Accessibility	Assess which researchers need access to the data, and does access need to be authorised?	*To ensure only those who need to use the dataset have access to the file*
	Has the data been protected from deliberate bias?	*Can the process of acquiring the dataset be traced?*
	Will the appropriate steps be undertaken to ensure the dataset cannot be damaged or misused?	*Ensure the dataset is saved in a secure file for analysis*
Relevance	Are the concepts in the dataset needed for the current user?	*Refer to hypotheses and evaluate whether the dataset is relevant*
	Are the produced statistics needed by the user?	*Investigate whether statistics have been formulated and whether these could be used in the present research*
Accuracy	Is the coefficient of variation available?	*Compare the degree of variation from one data series to another*
	What is the response rate?	*Reported as a percentage of how many participants returned the data collection*
	Does the data represent a complete list of eligible persons or units? and not just a fraction of the list	*Review the response rate and determine whether datasets were not submitted or incomplete. Depending on the severity of this issue, contact the data source or consider using statistical tests to account for missing values.*
	Is the imputation rate available?	*How many fields have been inserted to account for missing data*

https://doi.org/10.15405/ejsbs.156
eISSN: 2301-2218 / Corresponding Author: Pavlo Kanellakis
Selection & Peer-review under responsibility of the Editors

	Has the dataset been revised?	*Check for number of revisions and ensure the researchers access the latest version*
	Were data cleansing methods used?	*Investigate the responsible statistician, and review the cleansing methods*
Reliability	Is the data generated based on protocols and procedures that do not change according to who is using them? So, is the data completely objective, independent of user or use?	*Search for published guidelines for data collection, and examine the process.*
	Are variables defined, and are these definitions standardised and based on a referenced source?	*Determine whether definitions of variables are available*
Timeliness	Can the amount of time between the dataset and reference point be calculated?	*Important when planning further research and comparisons.*
Clarity	Is the metadata completed?	*Imperative to assess data quality. Contact the source if metadata is not available*
Comparability	What is the length of the time-series?	*The occurrence of the publication of the dataset*
	Which geographical areas are used? And, can these be transformed into larger geographies?	*List of geographical granularity, for example, County and District*
	Can the data be easily manipulated and presented as needed?	*Can the dataset be modified to suit the researcher's needs, for example, can units be converted?*
Coherence	Taking the above questions into account, can the current data be compared to other datasets?	*Prompts the researcher to reflect on the information*
Validity	Is engagement with researchers evident?	*During the data collection process, and publication of the dataset were relevant researchers liaised with?*
	Are the reports provisional and subject to change or have inaccuracies been reported separately?	*Find out whether the report is provisional and/or search for documentation of inaccuracies*
	Is there evidence of positive reports and no negative reports on the findings?	*Review the data source. Negative reports will be those that suggest that there are contradictions between different data sources for the same data.*
	Overall, does the dataset meet validation criteria?	*Dependent on the aforementioned. Mark the dataset as Pass, Borderline or Fail.*
Confidentiality	Does the dataset meet the BPS code of conduct for confidentiality?	*Check the data contains no identifiable information*

4. Discussion

Researchers wishing to use the current framework will benefit from the range of data quality criteria to assess suitability of datasets for specific research questions. The criteria and measures were adapted from an array of multidiscipline sources, the reason for this being to ensure the framework is robust and will effectively assist researchers. Another benefit of the current framework is that it is not sequential in nature. We anticipate the framework is shared across researchers and colleagues so individuals are able to gain a better understanding of their data collectively. The distribution of completed frameworks also promotes quality assurance through inter-observer reliability.

A further modification of the current framework would be to include a stronger emphasis on the interdependencies of quality criteria (Price & Shanks, 2004; Singh & Singh, 2010), as inter- dependencies may affect the analytical methods used in evaluation.

The testing phase of this quality framework (see appendix 1) concluded it is suitable for analysing and assessing datasets, and it will prompt researchers to formulate decisions on whether the dataset should be used in their research and with what level of confidence. If possible, this decision could be communicated back to the data source to raise issues and initiate solutions.

Acknowledgements

The author(s) declare that there is no conflict of interest.

References

Eppler, M. J. (2001). The Concept of Information Quality: An Interdisciplinary Evaluation of Recent Information Quality Frameworks, *Studies in Communication Sciences, 1*(2), 167-182.

Eurostat (2007): Handbook on Data Quality Assessment Methods and Tools. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/HANDBOOK%20ON%20DATA%20QUALITY%20ASSESSMENT%20METHODS%20AND%20TOOLS%20%20I.pdf

Goodchild, M. F. (1993). Data models and data quality: problems and prospects. *Environmental modeling with GIS*, 94-103.

Health and Social Care Information Centre (2013): Routine Quarterly Improving Access to Psychological Therapies Dataset Report: Final Q4 2012/13 summary statistics and related information, England, experimental statistics. Available at: http://www.hscic.gov.uk/catalogue/PUB11206

Muller, H., & Freytag, JC. (2005). Problems, Methods and Challenges in Comprehensive Data Cleansing. Available at: http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf

Price, R., & Shanks, G. (2004). A semiotic information quality framework. In Proceedings of the International Conference on Decision Support Systems DSS04 (pp. 658-672).

Singh, R., & Singh, K. (2010). A descriptive classification of causes of data quality problems in data warehousing. *International Journal of Computer Science Issues, 7*(3), 41-50.

Sorensen, H. T., Sabroe, S., & Olsen, J. (1996). A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology, 25*(2), 435- 442. https://doi.org/10.1093/ije/25.2.435

Statistics Canada. (2010). Quality Indicators and the Role of Metadata Repositories. Paper presented at the 3rd session of the Work Session on Statistical Metadata (METIS), Geneva, Switzerland, 10-12 March 2010.

Tran Ba Huy et al (2008). Data Quality Audit Tool: Guidelines For Implementation. Available at http://www.cpc.unc.edu/measure