Methods to assess the quality of data for public health: a critical review

Authors: Hong CHEN\textsuperscript{a, b}, Ping YU\textsuperscript{a}, David HAILEY\textsuperscript{a} and Ning WANG\textsuperscript{c}

\textsuperscript{a} School of Information Systems and Technology, Faculty of Engineering and Information Sciences, University of Wollongong, Australia

\textsuperscript{b} Jiangxi Provincial Centre for Disease Control and Prevention, China

\textsuperscript{c} NCAIDS, China CDC, Beijing, China

Corresponding Author: Ping YU, E-mail: ping@uow.edu.au
Outline

- Background
- Aims
- Methods
- Findings
- Conclusion
Fig. 1. Conceptual framework of data quality assessment for public health
Two key components of data quality assessment

- Attributes of data quality
  - Precisely defined.
  - Measurable.
  - Complete.

- The measurement method
Criteria to judge the quality of measurement methods

**Reliability**
The extent to which a measurement gives results that are consistent and repeatable.

**Validity**
The degree to which an assessment measures what it is supposed to measure.
There has been no systematic review of the methods to assess the quality of data in public health.
Research aim

To understand
- which attributes of data quality were assessed,
- how the assessment was conducted, and
- the reliability and validity of the methods.
Literature search strategy

- **Databases**
  - Scopus, IEEE Xplore, Web of Science, ScienceDirect, PubMed, etc.

- **Terms individually or in combination**

- **Additional search**
  - Well-known institutional websites: WHO, US CDC
  - Hand-search the reference lists

- **Inclusion criteria**
  - Empirical study
  - Peer reviewed or institutional reports during 2001 and 2013.

- **Exclusion criteria**
  - Narrative reviews, expert opinion, correspondence, and commentaries.
202 publications identified

73 duplicates and 51 irrelevant

78 potentially relevant publications for scrutiny

9 publications identified from the reference list

48 publications excluded

39 publications included

35 on assessment of data quality

4 on assessment of data use or data collection process

Figure 2. Flow diagram of selection of publications
Methods: qualitative content analysis

- Attributes of the quality of data
- Study design
- Measures
- Reliability and validity
### Attributes and the frequency to be assessed

<table>
<thead>
<tr>
<th>Item (No.=49)</th>
<th>Attribute (frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High data quality (38)</strong></td>
<td>Completeness (24), Accuracy (21), Timeliness (9) or up-datedness or currency, validity, periodicity, relevance, reliability, precision, integrity, confidentiality or data security, comparability, consistency or internal consistency or external consistency, concordance, granularity, repetability, readily useableness or usability or utility, objectivity, ease with understanding, importance, reflecting actual sample, meeting data standards, use of standards, accessibility, transparency, representativeness, disaggregation, data collection method or adjustment methods or data management process or data management</td>
</tr>
<tr>
<td><strong>Poor data quality (11)</strong></td>
<td>Missing data, under-reporting, inconsistencies, data errors or calculation errors or errors in report forms or errors resulted from data entry, invalid data, illegible hand writing, non-standardization of vocabulary, and inappropriate fields</td>
</tr>
</tbody>
</table>

The Attributes without brackets were assessed less than 3 times.

22 /35 studies did not explain how the attributes were selected and why.
No consistent definition of attributes

- The same attribute was given different meaning
  - Completeness
    - Defined as “standard data element entry” in the reporting forms; or
    - Defined as either “the correctness of data collection methods” or “the proportion of facilities reporting in an administrative area”.

- Different attributes were given the same meaning
  - Accuracy and completeness
    - Both stated as “no missing data elements”.
  - Accuracy and validity
    - Both defined as measurement validity.
Types of study subjects

- Data values: 31 studies
- Data users and documentation process: 13 studies
- Of 35 studies, 9 used both methods.
Study design for data quality assessment

- Data values: 31 studies
  - Desk review of the datasets
  - Field verification
  - Cross-sectional survey
  - Record-linkage and cross-check
  - No explicit description of methods.
Study design for assessment of data users or documentation process

- Data users and documentation process: 13 studies
  - Open, semi-structured or structured interviews with data users: 12 studies.
  - Reviewing the quality of documentation/guideline/instruction: 5 studies.
# Sampling methods

<table>
<thead>
<tr>
<th>Study subjects</th>
<th>Data values (n=31)</th>
<th>Data user &amp; documentation (n=13)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sampling methods</strong>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probability</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Non-probability</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>No information</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td><strong>Sample size</strong></td>
<td>72 to 7.5 million</td>
<td>6-178</td>
</tr>
<tr>
<td><strong>Power calculation</strong></td>
<td>1</td>
<td>No information</td>
</tr>
</tbody>
</table>

Note: * The sum is greater than 31 because of multi-stage sampling methods.
<table>
<thead>
<tr>
<th>Items</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td></td>
</tr>
<tr>
<td>*Reporting reliability value</td>
<td>2</td>
</tr>
<tr>
<td>Repeatability</td>
<td>16</td>
</tr>
<tr>
<td>Validity</td>
<td>4</td>
</tr>
</tbody>
</table>

*One paper reported the reliability in kappa statistics.
One paper reported the Cronbach’s alpha value.
Five deficiencies

1. A lack of consensus on which attributes should be measured for data quality in public health.
2. A lack of clear, consistent definition of data attributes and measures.
3. A lack of systematic procedures. Rarely applied mixed methods.
4. Little field verification of the quality of data.
5. The quality of the assessment methods need to be improved.
   - The validity of assessment was rarely examined.
   - Use of non-probability sampling methods is sub-optimal for the representativeness and external validity.
Conclusions

Contribution of this study:

- Systematic review of extant literature on data quality assessment for public health.
- Identifying the major attributes that were measured.
- The measurement methods used.
- Areas for further improvement.
To date, 49 attributes were measured.

- 38 positive attributes – high data quality
- 11 negative attributes – low data quality

The three most measured attributes:

- Accuracy
- Completeness
- Timeliness
The Measurement Methods

- Diverse.
- Could benefit from mixed methods approach.
- Need to consider reliability and validity of measurement.
- Field verification of the quality of data.
Areas for further development

Improving the definition of attributes of data quality

Improving reliability and validity of data quality assessment.

Triangulation of mixed methods.
• A. Alkon, Understanding Research: An Overview for Health Professionals. 2014, www.coursera.org
• J.A. Pereira, et al., Exploring the feasibility of integrating barcode scanning technology into vaccine inventory recording in seasonal influenza vaccination clinics. *Vaccine*, **30** (2012), 794-802.
• K.S. Mate, et al., Challenges for routine health system data management in a large public programme to prevent mother-to-child HIV transmission in South Africa. *PLoS ONE*, **4** (2009), e5483.


• D. Freestone, D. Williamson, and D. Wollersheim, Geocoding coronial data: Tools and techniques to improve data quality. *Health Information Management Journal*, **41** (2012), 4-12.


D. Lowrance, et al., Assessment of a national monitoring and evaluation system for rapid expansion of antiretroviral treatment in Malawi. *Tropical Medicine & International Health*, **12** (2007), 377-381