Journal of the American Medical Informatics Association, 28(7), 2021, 1591–1599
doi: 10.1093/jamia/ocaa340
Advance Access Publication Date: 26 January 2021
Review

Review

Quality assessment of real-world data repositories across the data life cycle: A literature review

Siaw-Teng Liaw [1], **Jason Guan Nan Guo**[1], **Sameera Ansari**[1], **Jitendra Jonnagaddala** [1], **Myron Anthony Godinho** [1], **Alder Jose Borelli Jr** [1], **Simon de Lusignan** [2], **Daniel Capurro**[3], **Harshana Liyanage**[2], **Navreet Bhattal**[4], **Vicki Bennett**[4], **Jaclyn Chan**[4], **and Michael G. Kahn** [5]

[1]WHO Collaborating Centre on eHealth, School of Population Health, Faculty of Medicine, UNSW Sydney, Sydney, New South Wales, Australia, [2]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom, [3]Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Victoria, Australia, [4]Australian Institute of Health and Welfare, Canberra, Australian Capital Territory, Australia, and [5]Department of Pediatrics (Section of Informatics and Data Sciences), University of Colorado Anschutz Medical Campus, Denver, Colorado, USA

*Corresponding Author: Siaw-Teng Liaw, WHO Collaborating Centre on eHealth, UNSW Sydney, Sydney, New South Wales, Australia; (siaw@unsw.edu.au)

Received 14 July 2020; Editorial Decision 16 December 2020; Accepted 22 December 2020

ABSTRACT

Objective: Data quality (DQ) must be consistently defined in context. The attributes, metadata, and context of longitudinal real-world data (RWD) have not been formalized for quality improvement across the data production and curation life cycle. We sought to complete a literature review on DQ assessment frameworks, indicators and tools for research, public health, service, and quality improvement across the data life cycle.

Materials and Methods: The review followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. Databases from health, physical and social sciences were used: Cinahl, Embase, Scopus, ProQuest, Emcare, PsycINFO, Compendex, and Inspec. Embase was used instead of PubMed (an interface to search MEDLINE) because it includes all MeSH (Medical Subject Headings) terms used and journals in MEDLINE as well as additional unique journals and conference abstracts. A combined data life cycle and quality framework guided the search of published and gray literature for DQ frameworks, indicators, and tools. At least 2 authors independently identified articles for inclusion and extracted and categorized DQ concepts and constructs. All authors discussed findings iteratively until consensus was reached.

Results: The 120 included articles yielded concepts related to contextual (data source, custodian, and user) and technical (interoperability) factors across the data life cycle. Contextual DQ subcategories included relevance, usability, accessibility, timeliness, and trust. Well-tested computable DQ indicators and assessment tools were also found.

Conclusions: A DQ assessment framework that covers intrinsic, technical, and contextual categories across the data life cycle enables assessment and management of RWD repositories to ensure fitness for purpose. Balancing security, privacy, and FAIR principles requires trust and reciprocity, transparent governance, and organizational cultures that value good documentation.

Key words: data quality, DQ measures, DQ indicators, DQ assessment tools, data custodianship, data stewardship, literature review

INTRODUCTION

Background

Globally, the increasing use of electronic health records (EHRs), health information systems, and personalized health monitoring devices has led to repositories of large volumes of complex longitudinal real-world data (RWD). Extracting valid inferences from these RWD repositories requires critical thinking and informatics-based analytic tools. Artificial intelligence (AI) and deep machine learning algorithms are increasingly being harnessed to mine these repositories for research, population health, quality improvement, clinical decision support, and personalized medicine.[1,2] Guiding principles have evolved for data custodians and data stewards to ethically manage these RWD repositories, including Privacy, Ethics, and Data Access frameworks[3] and the FAIR principles of findability, accessibility, interoperability and reusability, for public good and scientific advancement.[4] Improved access to interoperable data will accelerate systematic and innovative research using RWD to generate real-world evidence (RWE).

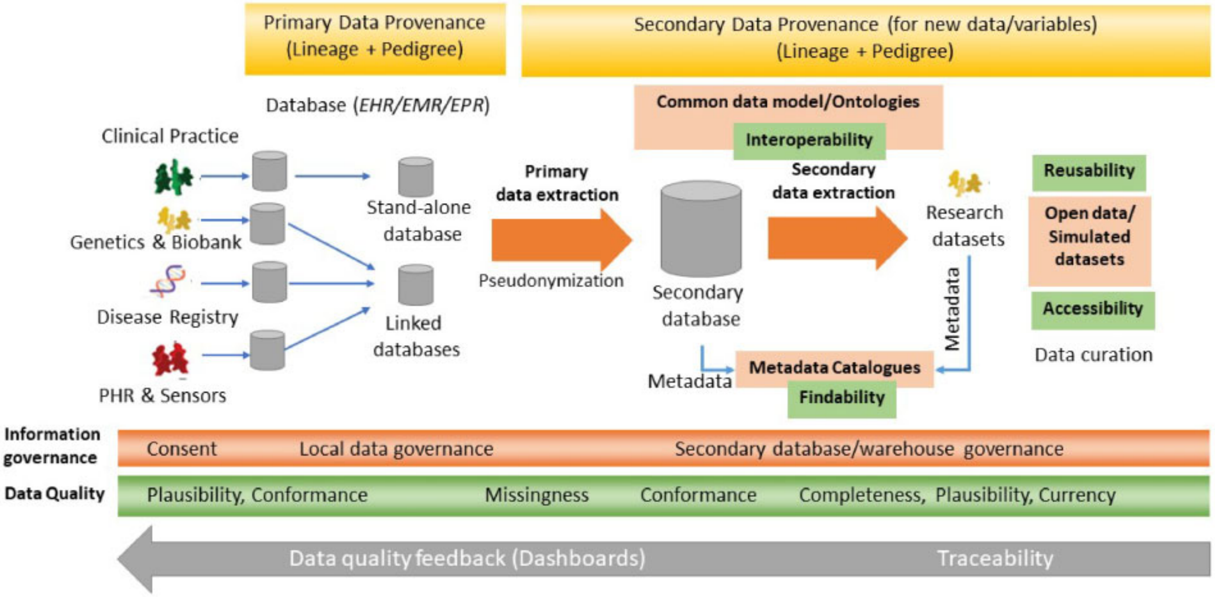
Examples of RWD repositories include PEDsNet in the United States[5] and the My Health Record system[6] in Australia. These datasets usually contain linked patient summary data uploaded from participating health services from diverse primary and secondary care settings, providing a clinician- and patient-centric view of longitudinal health data, collected at point of care in the real world. Relevant regulations enable the release of these RWD for research, public health, and quality improvement purposes. Similar RWD repositories exist globally, with different models of governance and provenance requirements of data custodians to ensure data quality (DQ) and fitness for purpose.

RWD is only as good as its quality. High quality data should be "intrinsically good, contextually appropriate, clearly represented and accessible to data consumers."[7] However, the DQ domain has lacked a commonly agreed terminology and clear conceptualization of the intrinsic and contextual determinants of DQ.[8–10] This diversity reflects the semantic and syntactic heterogeneity in the data, metadata, databases, and diverse needs of the creators, custodians, informaticians, researchers, and users of data across the data production and curation life cycle (Figure 1).[2,11–13]

Figure 1 summarizes the data life cycle along with the required primary and secondary data provenance, governance and DQ assessment at various points in the life cycle. DQ assessment must be conducted systematically at all stages and at various points in the life cycle, including data creation and collection to the extract, transform, and load process; data processing and annotation with metadata; and curation, visualization, and sharing.[2] Traceability is important to ensure that DQ feedback can be provided to data stewards and custodians at each point in the life cycle to address and improve DQ.[14] Some categories of DQ may be more important or easier to assess and manage at different points of the life cycle. The distribution of data and its quality attributes along the life cycle (green bar in the bottom third of Figure 1) illustrates how context and plausibility may be more important to know and easier to manage at the data source.[2,15]

In 2016, an international DQ research collaboration reviewed the literature and developed and published a harmonized framework and terminology to assess the intrinsic quality of a dataset.[15] This framework defined 3 DQ categories—conformance, completeness, and plausibility—and 2 DQ assessment contexts—verification and validation. Conformance and plausibility categories are further divided into subcategories. Data may be verified with organizational data or validated against an accepted gold standard, depending on proposed context and uses. The coverage of this harmonized intrinsic DQ framework (HIDQF) and terminology was validated by successfully aligning to multiple published DQ terminologies.[15]

The use and evolution of the HIDQF since 2016 has included the development of indicators and tools to support the intrinsic DQ assessment of RWD repositories for research, public health, and qual-



NOTE: PHR = personal health record; EHR = electronic health record; EMR = electronic medical record; EPR = electronic patient record.

Figure 1. The data production and curation life cycle with associated provenance, governance, and data quality assessment. EHR: electronic health record; EMR: electronic medical record; EPR: electronic patient record; PHR: personal health record.

ity improvement purposes.[11,16–19] These developments have neither explicitly nor systematically addressed the contextual and process categories that are vital to DQ assessment and management in RWD production and curation life cycles. The senior authors (S.-T.L., M.G.K., S.d.L.) therefore guided the integration of the HIDQF and data life cycle framework as the conceptual starting point for this literature review to identify practical and potential gaps in the assessment and management of DQ.

Objective

We sought to conduct a literature review on DQ assessment frameworks, indicators, and tools for research, public health, and quality improvement, incorporating the additional perspective of the full data production and curation life cycle.

MATERIALS AND METHODS

Developing the conceptual framework for the review and synthesis

Figure 1 illustrate some of the included intrinsic DQ categories and subcategories from the HIDQF, along with the FAIR guiding principles for the secondary use of data, and their relevance and importance at various points in the data life cycle. The Privacy, Ethics, and FAIR principles are implicit in and central to the governance and provenance concepts in Figure 1. The elements of this conceptual framework guided the search strategy and criteria for inclusion of articles and DQ concepts.

Conducting the literature review

Guided by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, the published literature—including reviews, conference articles and original research articles—from 2009 onward was examined for research on DQ frameworks, indicators ,and tools. While the focus is health, we were also interested in similar work reported in the social and physical sciences. As such, the databases used were Cinahl, Embase, Scopus, ProQuest, Emcare, PsycINFO, Compendex, and Inspec. Embase was used instead of PubMed, which is an interface to search MEDLINE, because it covers all the journals that are in MEDLINE, as well as covers 2900 additional unique journals and conference abstracts that are not in MEDLINE. Embase is indexed with the Emtree terminology, which includes all the MeSH (Medical Subject Headings) terms used in MEDLINE. Gray literature sources included:

1. Australian Institute of Health and Welfare (https://www.aihw.gov.au/)
2. Australian Digital Health Agency (https://www.digitalhealth.gov.au/)
3. Australian Bureau of Statistics (ABS) (https://www.aihw.gov.au/)
4. Australian Department of Health (https://www.health.gov.au/)
5. Irish Health Information Quality Authority (https://www.hiqa.ie/)
6. World Health Organization Institutional Repository for Information Sharing (https://apps.who.int/iris/browse).

Search strategy

This study built on our previous DQ literature reviews.[8,15] The full search strategy is available as Supplementary Appendix 1. The search syntax in Scopus is presented as an example (Box 1):

Consensus process

Two reviewers independently screened the titles and abstracts for studies of DQ frameworks, indicators, or assessment tools. A third reviewer was involved if consensus was not reached. Following consensus, full-text appraisal defined the final set of included reviews, conference articles, and original research articles.

Data were extracted by combinations of 2 authors, including details of the following:

- DQ framework, categories, subcategories, indicators and assessment tools;
- setting and purpose (research, population/public health, clinical/managerial quality improvement); and
- methodology (machine learning, statistics, models, ontologies), including innovative concepts and processes.

The relevance and importance of all the DQ concepts extracted were qualitatively assessed and, using an iterative consensus process, categorized according the HIDQF and data life cycle conceptual frameworks. Final inclusion was determined following discussion,

Box 1. Search Syntax in Scopus

SYNTAX 1

ALL ((((data PRE/0 (assessment OR accuracy OR quality OR completeness OR conformance OR validation OR verification OR plausibility)) PRE/5 (tool* OR framework OR software)) AND ("electronic medical record*" OR "electronic health record*" OR "my health record" OR "electronic patient record*" OR "personal health record*"))) AND PUBYEAR > 2009 AND (LIMIT-TO (SUBJAREA,"MEDI") OR LIMIT-TO (SUBJAREA,"COMP") OR LIMIT-TO (SUBJAREA,"ENGI") OR LIMIT-TO (SUBJAREA,"NURS"
) OR LIMIT-TO (SUBJAREA,"BUSI") OR LIMIT-TO (SUBJAREA,"DECI")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))

SYNTAX 2

ALL ((((tool* OR framework OR software) PRE/5 (data PRE/0 (assessment OR accuracy OR quality OR completeness OR conformance OR validation OR verification OR plausibility))) AND ("electronic medical record*" OR "electronic health record*" OR "my health record" OR "electronic patient record*" OR "personal health record*"))) AND PUBYEAR > 2009 AND (LIMIT-TO (SUBJAREA,"MEDI") OR LIMIT-TO (SUBJAREA,"COMP") OR LIMIT-TO (SUBJAREA,"ENGI") OR LIMIT-TO (SUBJAREA,"NURS"
) OR LIMIT-TO (SUBJAREA,"BUSI") OR LIMIT-TO (SUBJAREA,"DECI")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))

Note: Search was run in 2 syntax to cover tool/framework group occurring in front of group 1.

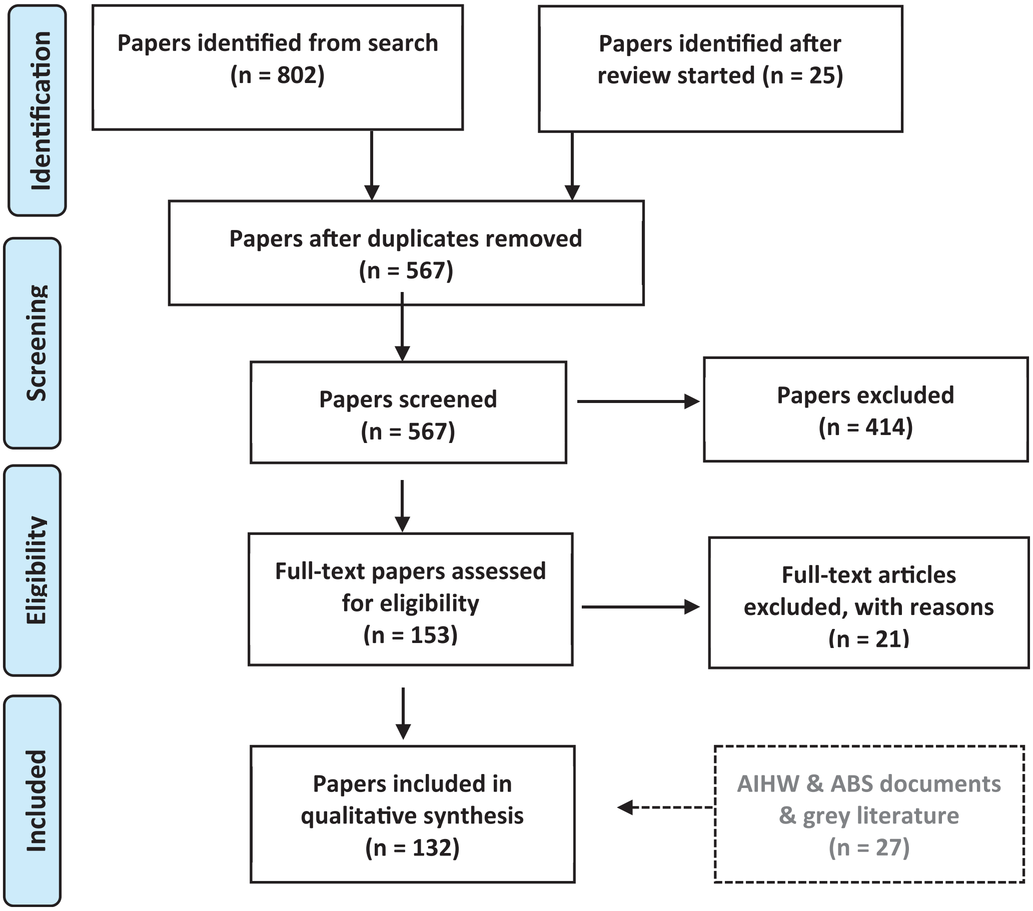

Identification
Screening
Eligibility
Included
Papers identified from search (n = 802)
Papers identified after review started (n = 25)
Papers after duplicates removed (n = 567)
Papers screened (n = 567)
Papers excluded (n = 414)
Full-text papers assessed for eligibility (n = 153)
Full-text articles excluded, with reasons (n = 21)
Papers included in qualitative synthesis (n = 132)
AIHW & ABS documents & grey literature (n = 27)

Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram for data quality (DQ) literature review.

including consideration of the frequency of separate articles that addressed them, and consensus by the UNSW team and, subsequently, all authors. The full data extraction matrix is available as Supplementary Appendix 2.

The synthesis aimed to map the different terminologies used to describe the same or similar DQ concepts. Commonalities in the key features and purpose of DQ assessment across the data life cycle, including the indicators and tools used, were identified and agreed. The implications for the HIDQF specifically and DQ assessment generally were examined.

RESULTS

Figure 2 summarizes the literature review process and articles included and excluded at all stages of the review, using the PRISMA guidelines.

Articles that were published after the review began included the Computer Methods and Programs in Biomedicine Special Issue on Data Quality.[20] Relevant documents from the gray literature were also included as they became available and eligible.

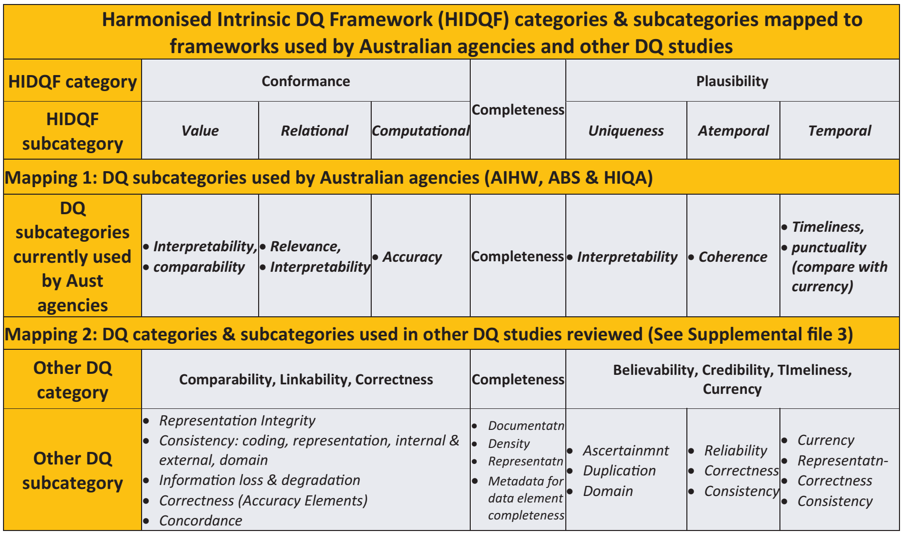
The complete list of articles reviewed is available as Supplementary Appendix 3.

DQ frameworks

Figure 3 summarizes the existing HIDQF and mapped its categories and subcategories to those used by the Australian Institute of Health and Welfare, Australian Bureau of Statistics, and Irish Health Information Quality Authority (Mapping 1), and other published DQ studies as listed in Supplementary Appendix 3 (Mapping 2). It was not possible to map all the diverse DQ concepts used by Australian agencies and reported in the included DQ studies to the HIDQF subcategories. The ABS DQ Framework lists 7 dimensions, equivalent to categories in HIDQF: institutional environment, relevance, timeliness, accuracy, coherence, interpretability and accessibility.[21] This framework underpins the ABS DQ declaration or statement.[22]

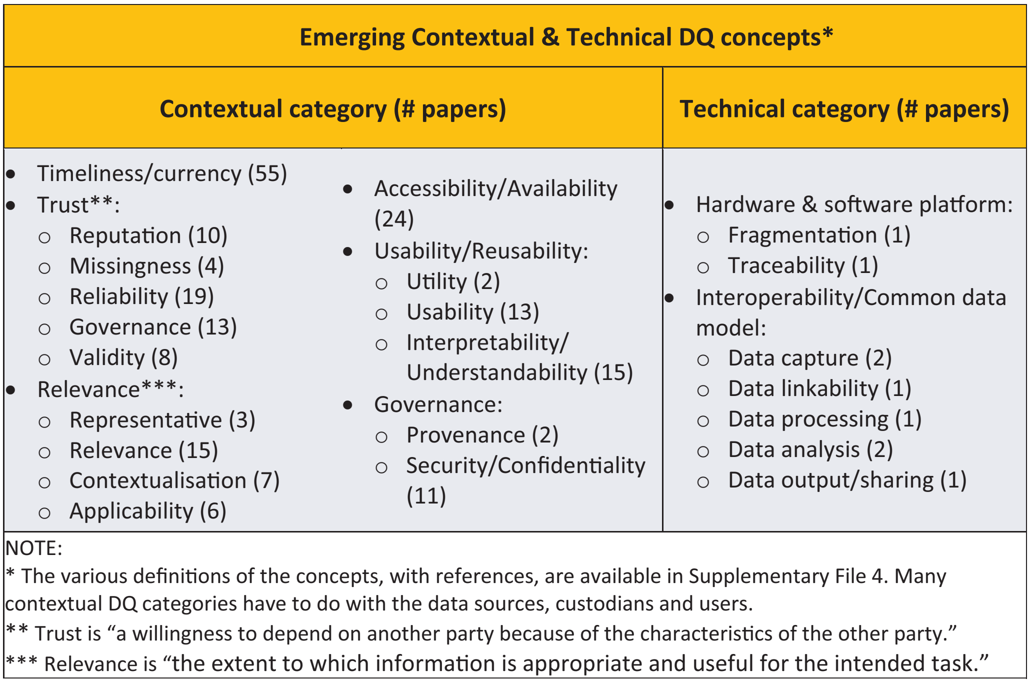
A number of concepts that may determine the DQ of a RWD repository were identified from the 120 included articles (Figure 4), including timeliness or currency (55 articles), accessibility (18 articles), relevance or irrelevance (15 articles), reliability (13 articles), usability (13 articles), interpretability (10 articles), reputation/believability (10 articles) articles, validity (8 articles), contextualization (7 articles), and applicability (6 articles). The definitions of these concepts, with references, are available in Supplementary Appendix 4.

DQ concepts also emerged from a number of articles reporting on DQ research and development associated with interoperability and common data models.[23] The extrinsic DQ concepts were categorized as organizational (eg, reputation and governance), functional (eg, timeliness and usability), and technical (eg, interoperability). The "trust" construct[24] is a good example of the utility of the DQ across the life cycle framework: in addition to intrinsic DQ concepts, contextual concepts such as reputation (trusted source), missingness (absence of data or fields), usability, accessibility, and relevance (to intended task) are important contextual determinants of DQ (Figure 4).



Harmonised Intrinsic DQ Framework (HIDQF) categories & subcategories mapped to frameworks used by Australian agencies and other DQ studies							
HIDQF category	Conformance			Completeness	Plausibility		
HIDQF subcategory	*Value*	*Relational*	*Computational*		*Uniqueness*	*Atemporal*	*Temporal*
Mapping 1: DQ subcategories used by Australian agencies (AIHW, ABS & HIQA)							
DQ subcategories currently used by Aust agencies	• *Interpretability,* • *comparability*	• *Relevance,* • *Interpretability*	• *Accuracy*	Completeness	• *Interpretability*	• *Coherence*	• *Timeliness,* • *punctuality (compare with currency)*
Mapping 2: DQ categories & subcategories used in other DQ studies reviewed (See Supplemental file 3)							
Other DQ category	Comparability, Linkability, Correctness			Completeness	Believability, Credibility, TImeliness, Currency		
Other DQ subcategory	• *Representation Integrity* • *Consistency: coding, representation, internal & external, domain* • *Information loss & degradation* • *Correctness (Accuracy Elements)* • *Concordance*			• *Documentatn* • *Density* • *Representatn* • *Metadata for data element completeness*	• *Ascertainmnt* • *Duplication* • *Domain*	• *Reliability* • *Correctness* • *Consistency*	• *Currency* • *Representatn-* • *Correctness* • *Consistency*

Figure 3. Existing harmonized intrinsic data quality (DQ) framework (HIDQF) categories and subcategories mapped to frameworks used by Australia agencies and other DQ studies. ABS: Australian Bureau of Statistics; AIHW: Australian Institute of Health and Welfare; HIQA: Irish Health Information Quality Authority.



Emerging Contextual & Technical DQ concepts*		
Contextual category (# papers)		**Technical category (# papers)**
• Timeliness/currency (55) • Trust**: ○ Reputation (10) ○ Missingness (4) ○ Reliability (19) ○ Governance (13) ○ Validity (8) • Relevance***: ○ Representative (3) ○ Relevance (15) ○ Contextualisation (7) ○ Applicability (6)	• Accessibility/Availability (24) • Usability/Reusability: ○ Utility (2) ○ Usability (13) ○ Interpretability/ Understandability (15) • Governance: ○ Provenance (2) ○ Security/Confidentiality (11)	• Hardware & software platform: ○ Fragmentation (1) ○ Traceability (1) • Interoperability/Common data model: ○ Data capture (2) ○ Data linkability (1) ○ Data processing (1) ○ Data analysis (2) ○ Data output/sharing (1)

NOTE:
* The various definitions of the concepts, with references, are available in Supplementary File 4. Many contextual DQ categories have to do with the data sources, custodians and users.
** Trust is "a willingness to depend on another party because of the characteristics of the other party."
*** Relevance is "the extent to which information is appropriate and useful for the intended task."

Figure 4. Emerging contextual and technical data quality (DQ) concepts.

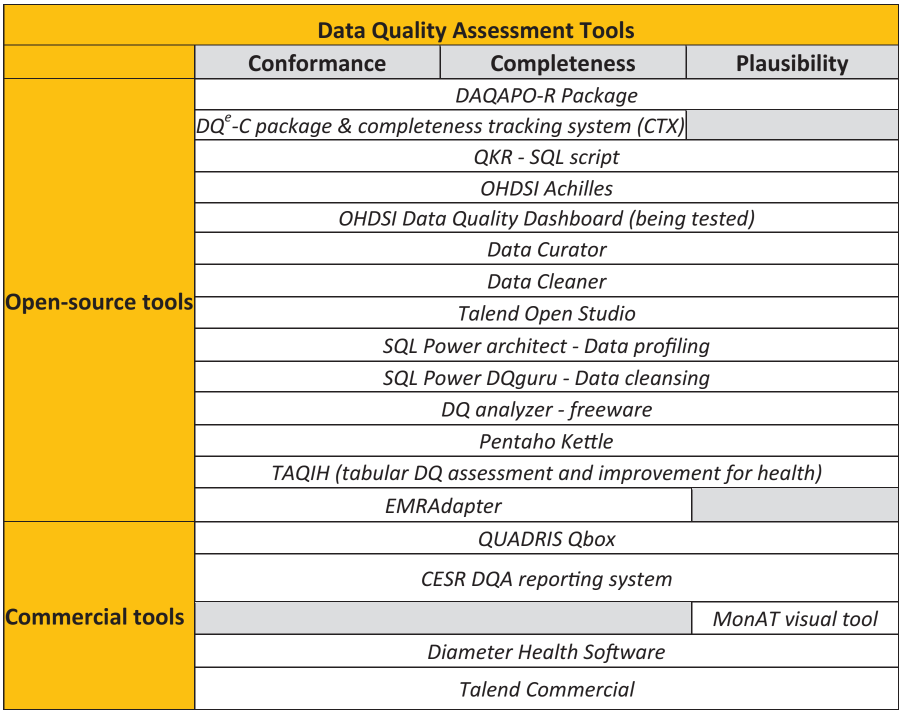
A use case of this additional contextual and technical categories is when RWD become dated or fall outside the regulatory requirements for storage, as in data sunsetting. The "sunset data" can be identified and addressed through a combination of temporal plausibility (intrinsic DQ) and provenance and governance provisions (contextual DQ) and computational (technical DQ) aspects in the extended HIDQF.

Indicators and measures of DQ

Many published DQ indicators and measures found were relevant and mappable to indicators for the HIDQF categories and subcategories. The DQ indicators and assessment tools appeared to be robust, with testing and validation in the open source domain, including those developed by the Observational Health Data Science and Informatics (OHDSI) collaborative. [25]

Conformance of data was measured by a value for homogeneity and compliance to metadata, data model, specifications, and standards. Examples of conformance measures would be a ratio of known and unknown data types or calculated as *[Total Semantically Consistent Rows] divided by [Total Rows]*.

Plausibility was measured as a believable unique value, range, or pattern at one time point or over many time points. Examples of temporal plausibility would be measures of single or multisite temporal patterns and variability, including probabilistic variability, over time.



<table>
<tr><th colspan="4">Data Quality Assessment Tools</th></tr>
<tr><th></th><th>Conformance</th><th>Completeness</th><th>Plausibility</th></tr>
<tr><td rowspan="14">Open-source tools</td><td colspan="3">DAQAPO-R Package</td></tr>
<tr><td colspan="2">DQ^e-C package & completeness tracking system (CTX)</td><td></td></tr>
<tr><td colspan="3">QKR - SQL script</td></tr>
<tr><td colspan="3">OHDSI Achilles</td></tr>
<tr><td colspan="3">OHDSI Data Quality Dashboard (being tested)</td></tr>
<tr><td colspan="3">Data Curator</td></tr>
<tr><td colspan="3">Data Cleaner</td></tr>
<tr><td colspan="3">Talend Open Studio</td></tr>
<tr><td colspan="3">SQL Power architect - Data profiling</td></tr>
<tr><td colspan="3">SQL Power DQguru - Data cleansing</td></tr>
<tr><td colspan="3">DQ analyzer - freeware</td></tr>
<tr><td colspan="3">Pentaho Kettle</td></tr>
<tr><td colspan="3">TAQIH (tabular DQ assessment and improvement for health)</td></tr>
<tr><td colspan="2">EMRAdapter</td><td></td></tr>
<tr><td rowspan="5">Commercial tools</td><td colspan="3">QUADRIS Qbox</td></tr>
<tr><td colspan="3">CESR DQA reporting system</td></tr>
<tr><td colspan="2"></td><td>MonAT visual tool</td></tr>
<tr><td colspan="3">Diameter Health Software</td></tr>
<tr><td colspan="3">Talend Commercial</td></tr>
</table>

Figure 5. Data quality assessment tools.

The coverage of these mapped indicators, with some examples, are shown in Supplementary Appendix 5. A high-level summary of these indicators is also included in Figure 6.

The DQ indicators not completely covered by the HIDQF were categorized as potentially intrinsic (eg, conciseness or objectivity), contextual (eg, applicability or understandability), or technical (eg, Euclidian distance or correct location or format after migration of records). Our approach to "context" was realist and includes places, people, time, institutions, resources available (or not), social relationships, rules, norms, and expectations that constitute them.[26]

These indicators are summarized in Supplementary Appendix 6.

DQ assessment tools

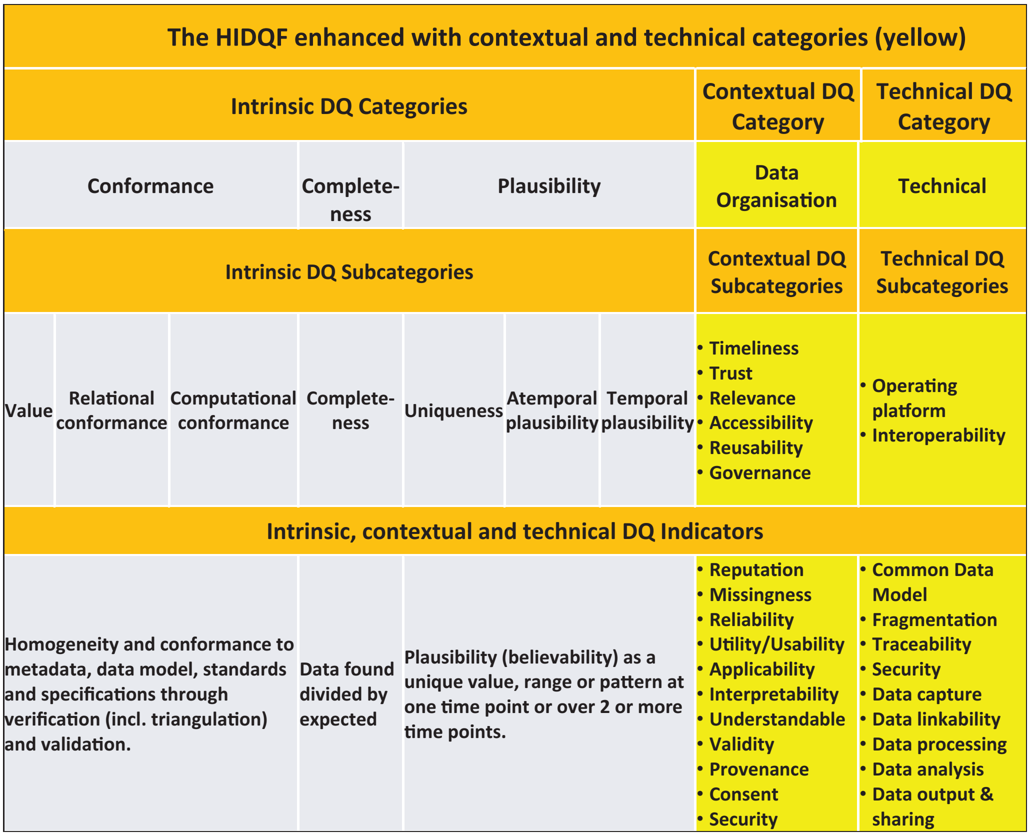
Figure 5 lists the currently available open source and commercial DQ assessment tools with an indication of the HIDQF category they addressed. This review included only open source tools, based on a qualitative examination of the logic and innovativeness of the approach and whether they have been field-tested (not many have been systematically evaluated for processes and outcomes). The included tools mainly addressed intrinsic DQ indicators because they are relatively easy to compute as DQ rules. The Quality Knowledge Repository was developed to store DQ Concepts and their methods of computation across information quality (IQ) domains with relationships across domains determined by a DQ meta-model. This knowledge repository has been leveraged into a service-oriented architecture to perform as a scalable and reproducible framework for DQ assessments of disparate data sources.[18] Web-based tools such as the TAQIH (tabular DQ assessment and improvement for health) have been developed to conduct exploratory data analyses to improve completeness, accuracy, redundancy, and readability.[27]

The ACHILLES (Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems) tool was developed and maintained by the OHDSI collaborative to conduct DQ checks on databases that are based on or mapped to the Observational Medical Outcomes Partnership Common Data Model (CDM).[25] The Achilles DQ rules have been well tested, including a published application to 24 large healthcare datasets across 7 different organizations in America, Europe, and Asia. This highlighted at least 12 DQ rule violations in 10 of the 24 datasets and violations of the full set of 71 DQ rules in at least 1 dataset.[23] This CDM-enabled rapid comparisons of the DQ of multiple international datasets reduced the data model variations and increased the robustness and generalizability of the methods and outputs. These findings also highlighted the importance of interoperability, a FAIR guiding principle,[4] as a core component of DQ assessment of RWD repositories that integrate data from multiple EHRs and health information systems. The OHDSI Data Quality Dashboard (https://ohdsi.github.io/DataQualityDashboard/) has been developed based on the HIDQF and Achilles. It is a potential mechanism to standardize DQ assessment of RWD within a CDM.

These tools do not appear to be used widely as only a small number of included studies reported on the quality of the underpinning dataset. There were no DQ statements that met, for example, the criteria promoted in the ABS DQ declaration checklist.[21]

DISCUSSION

We found a diverse range of DQ assessment frameworks, indicators, measures, tools, and use cases across the informatics, health services, and population health domains. These use cases highlighted information flow, business processes, and data elements; relevance to user and system requirements; and uses and challenges to DQ across the data life cycle (Figure 1). The diversity of frameworks, indicators, and tools presented different data and DQ needs depending on whether the purpose was for patient-level prediction or population-level estimation, or to answer questions of varying complexity in terms of time, place, and target population.



The HIDQF enhanced with contextual and technical categories (yellow)								
Intrinsic DQ Categories							Contextual DQ Category	Technical DQ Category
Conformance			Completeness	Plausibility			Data Organisation	Technical
Intrinsic DQ Subcategories							Contextual DQ Subcategories	Technical DQ Subcategories
Value	Relational conformance	Computational conformance	Completeness	Uniqueness	Atemporal plausibility	Temporal plausibility	• Timeliness • Trust • Relevance • Accessibility • Reusability • Governance	• Operating platform • Interoperability
Intrinsic, contextual and technical DQ Indicators								
Homogeneity and conformance to metadata, data model, standards and specifications through verification (incl. triangulation) and validation.			Data found divided by expected	Plausibility (believability) as a unique value, range or pattern at one time point or over 2 or more time points.			• Reputation • Missingness • Reliability • Utility/Usability • Applicability • Interpretability • Understandable • Validity • Provenance • Consent • Security	• Common Data Model • Fragmentation • Traceability • Security • Data capture • Data linkability • Data processing • Data analysis • Data output & sharing

Figure 6. The harmonized intrinsic data quality (DQ) framework (HIDQF) enhanced with contextual and technical categories.

The majority of the concepts identified could be mapped to the HIDQF categories and subcategories. It was difficult to map some extracted concepts, such as conciseness, objectivity, ease of manipulation, status, content, and presentation to the HIDQF. This was mainly because the measures were mostly qualitative and more contextual than intrinsic. The common practice has been to use contextual DQ concepts qualitatively on a case-by-case basis as, for example, in the case of the ABS DQ statement and declaration.[22]

Much of the gray literature, especially those published by government agencies such as the ABS (Figure 3) and the WHO, [13] continue to describe DQ with a mix of intrinsic and contextual categories in a nonstandardized or systematic manner. For example, timeliness can refer to when data were collected, made findable, or accessible; or missingness of data or data fields can occur during collection, extraction, or visualization. A fundamental challenge is trust, which we have developed as a DQ construct (Figure 4).[24] While contextual DQ dimensions are not easily quantifiable or measurable, they are important to meaningfully assess and manage DQ across the data life cycle.

The diversity found emphasizes the need for an extended approach to DQ that standardizes conventions to describe and assesses contextual and technical as well as intrinsic DQ categories in RWD repositories across the data life cycle. Because the repositories are ever-changing in content and purpose, and ever-extending in duration, DQ assessment models will need to consider continuous DQ assessment for periods longer than years or decades. These contextual elements may potentially be addressed with a common metadata model or an enhanced CDM or a combination of both.

The literature reflected an increasing recognition that a CDM is useful to constrain and, over time, overcome the inherent interoperability challenge with aggregating and integrating different RWD repositories. At the very least, it will improve the reuse of the data in one by another RWD repository. Interoperability standards are important to ensure that RWD collected as part of clinical care are captured, represented, curated, and shared appropriately and accurately among users in the primary and secondary health services in an integrated health neighborhood.[28] The benefits to integrated care will be further enhanced if and when research datasets created from the secondary use of RWD are also interoperable across domains and disciplines. Transparent, comprehensive reporting of DQ features directly aligns with the Privacy, Ethics, and FAIR guiding principles by engendering trust among stakeholders of RWD repositories.[2,29,30]

The CDM-based ACHILLES[23] approach to DQ assessment is being enhanced through a service-oriented architecture-based Open Quality and Analytics Framework.[18] The Open Quality and Analytics Framework includes a DQ metamodel, a federated data integration platform to support semantically consistent metadata-centric querying of heterogeneous data sources, and a visualization metaframework to store visualizations for different DQ concepts, indicators, and measures. This supports the inclusion of a technical category in the HIDQF.

This review suggests that RWD repositories require a DQ assessment framework that includes new or expanded intrinsic, technical, and contextual categories. The context needs to address best practice across micro-, meso-, and macro-organizations as well as the health system across the data life cycle. To support this role, we recommend that the current HIDQF be enhanced with 2 categories—technical and contextual. Regulatory, organizational, and with increasing use, other subcategories can be included in the contextual category.

Figure 6 summarizes the enhanced HIDQF with proposed categories and indicators. The HIDQF needs to be enhanced because DQ assessment is highly context sensitive and dependent on the purpose. The needs and requirements of a range of diverse stakeholders across the RWD production and curation life cycle are central to DQ assessment and management of the data errors and problems found.

There is sufficient research and development of intrinsic DQ indicators to provide a library from which data custodians can choose to adopt or adapt to meet their needs. However, for the enhanced HIDQF, novel quantitative and qualitative DQ measures and metadata will need to be developed to meaningfully characterize trust, relevance, accessibility, reusability, and governance as well as the operating platforms and interoperability.

CONCLUSION

An enhanced HIDQF, combined with continuous quality improvement protocols across the RWD life cycle, is recommended to ensure that RWD can be assessed and managed, including consideration of the context, to be fit for purpose. It can lead to more meaningful and useful standardized DQ statements about RWD repositories or specific datasets. Data custodians and researchers must routinely report the DQ—a DQ statement—as well as the actual and potential impacts of the results of data-driven research and development. This is especially important with AI and deep machine learning being the inevitable future. The enhanced DQ assessment framework will determine fitness for purpose through assessment of intrinsic DQ, contextual DQ (timeliness, trustworthiness, relevance, accessibility, reusability, governance), and technical DQ (traceability and interoperability).

The CDM-based ACHILLES DQ assessment tool highlighted the importance of CDMs as a strategy to reduce variations in DQ assessments of distributed RWD datasets. Interoperability (eg, conformance to a semantic and syntactic standard) is an integral concept in the DQ framework. More research is required in this space.

Trust and willingness to share data for public good and scientific advancement is a core requirement. DQ assessment, management, visualization, and sharing requires an optimal balance of privacy and security arrangements with adherence to FAIR principles. An ethical and secure framework based on public good is essential to produce a data asset that is fit for purpose.

Comprehensive DQ assessment requires a culture of reciprocity, transparency, and interoperability across the data production and curation life cycle. Effective DQ assessment is underpinned by rigorous documentation at point of care, good management, and appropriate governance across the RWD production and curation life cycle.

FUNDING

This work was supported by the Australian Institute of Health and Welfare.

AUTHOR CONTRIBUTIONS

S-TL, MK, and SdL conceptualized the combined HIDQF and life cycle framework and guided the literature review. All authors contributed substantially to the review process; screening of abstracts and full texts, data extraction, synthesis, analysis and interpretation; and iterative enhancement of conceptual framework. S-TL initiated and completed the manuscript. All authors reviewed and contributed to the intellectual content, approved the final draft and agreed to be accountable for the accuracy and integrity of the work.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

DATA AVAILABILITY STATEMENT

All the data has been made available through the article and online Supplementary Appendices.

ACKNOWLEDGMENTS

We thank UNSW Sydney research librarian Ms Monica O'Brien for guidance on the search strategy.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

REFERENCES

1. Liyanage H, Liaw S-T, Jonnagaddala J, *et al.* Artificial intelligence in primary health care: perceptions, issues, and challenges. *Yearb Med Inform* 2019; 28 (1): 41–6.
2. Liaw S, Liyanage H, Kuziemsky C, *et al.* Ethical use of electronic health record data and artificial intelligence: recommendations of the primary care informatics working group of the international medical informatics association. *Yearb Med Inform* 2020; 29 (1): 51–7.
3. Liyanage H, Liaw ST, Di Iorio CT, *et al.* Building a privacy, ethics, and data access framework for real world computerised medical record system data: a Delphi study. Contribution of the Primary Health Care Informatics Working Group. *Yearb Med Inform* 2016; 25 (1): 138–45.
4. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016; 3 (1): 160018.
5. Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating foundational data quality in the national patientcentered clinical research network (PCORnet®). *EGEMS (Wash DC)* 2018; 6 (1): 3–9.
6. Australian Digital Health Agency. Commonwealth of Australia. My Health Record - Practice Incentive Program Digital Health Incentive; 2016. https://www.myhealthrecord.gov.au/for-healthcare-professionals/what-is-my-health-record. Accessed January 17, 2020.
7. Wang R, Strong D, Guarascio L. Beyond accuracy: what data quality means to data consumers. *J Manage Inf Syst* 1996; 12 (4): 5–33.
8. Liaw S, Rahimi A, Ray P, *et al.* Towards an ontology for data quality in integrated chronic disease: a realist review of the literature. *Int J Med Inform* 2013; 82 (1): 10–24.
9. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013; 20 (1): 144–51.
10. Saez C, Martinez-Miranda J, Robles M, *et al.* Organizing data quality assessment of shifting biomedical data. *Stud Health Technol Inform* 2012; 180: 721–5.
11. Huser V, Kahn MG, Brown JS, *et al.* Methods for examining data quality in healthcare integrated data repositories. *Biocomputing* 2018; 23: 628–33.
12. World Health Organization. *Data Quality Review: A Toolkit for Facility Data Quality Assessment. Module 1. Framework and Metrics*. Geneva, Switzerland: WHO Press; 2017.

13. World Health Organization Western Pacific Regional Office. *Improving Data Quality: A Guide for Developing Countries*. Manila, the Phillippines: WHO Press; 2003.
14. Taggart J, Liaw S-T, Yu H. Structured data quality reports to improve EHR data quality. *Int J Med Inform* 2015; 84 (12): 1094–8.
15. Kahn MG, Callahan TJ, Barnard J, *et al.* A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)* 2016; 4 (1): 18.
16. Huser V, Li X, Zhang Z, *et al.* Extending Achilles heel data quality tool with new rules informed by multi-site data quality comparison. *Stud Health Technol Inform* 2019; 264: 1488–9.
17. Khare R, Utidjian L, Ruth BJ, *et al.* A longitudinal analysis of data quality in a large pediatric data research network. *J Am Med Inform Assoc* 2017; 24 (6): 1072–9.
18. Rajan NS, Gouripeddi R, Mo P, *et al.* Towards a content agnostic computable knowledge repository for data quality assessment. *Comput Methods Progr Biomed* 2019; 177: 193–201.
19. Henley-Smith S, Boyle D, Gray K. Improving a secondary use health data warehouse: proposing a multi-level data quality framework. *EGEMS (Wash DC)* 2019; 7 (1): 38.
20. Sáez C, Liaw S-T, Kimura E, Coorevits P, Garcia-Gomez JM. Guest editorial: Special issue in biomedical data quality assessment methods. *Comput Methods Programs Biomed* 2019; 181: 104954.
21. Australian Bureau of Statistics. The ABS Data Quality Framework; 2015. https://www.abs.gov.au/websitedbs/D3310114.nsf//home/ABS+Data+Quality+Statement+Checklist#:~:text=Checklist%20of%20questions%20to%20generate,quality%20of%20the%20specific%20data. Accessed August 30, 2020.
22. Lee G, Allen B. Educated use of information about data quality. In: *Proceedings of the International Statistical Institute 53rd Session*; 2001.
23. Huser V, DeFalco FJ, Schuemie M, *et al.* Multisite evaluation of a data quality tool for patient-level clinical data sets. *EGEMS (Wash DC)* 2016; 4 (1): 1239.
24. McKnight DH, Carter M, Thatcher JB, *et al.* Trust in a specific technology: An investigation of its components and measures. *ACM Trans Manage Inf Syst* 2011; 2 (2): 1–25.
25. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
26. RAMESES II project team. What realists mean by context; or, Why nothing works everywhere or for everyone. http://www.ramesesproject.org/media/RAMESES_II_Context.pdf. Accessed September 10, 2020.
27. Álvarez Sánchez R, Beristain Iraola A, Epelde Unanue G, *et al.* TAQIH, a tool for tabular data quality assessment and improvement in the context of health data. *Comput Methods Programs Biomed* 2019; 181: 104824.
28. Liaw S, de Lusignan S. An 'integrated health neighbourhood' framework to optimise the use of EHR data. *J Innov Health Inform* 2016; 23 (3): 547–54.
29. Kahn MG, Brown JS, Chun AT, *et al.* Transparent reporting of data quality in distributed data networks. *EGEMS (Wash DC)* 2015; 3 (1): 7–1052.
30. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. *Med Care* 2013; 51: S22–9.