BJR

https://doi.org/10.1259/bjr.20190574

Received: 28 June 2019 | Revised: 14 January 2020 | Accepted: 15 January 2020

Cite this article as:
Smith NAS, Sinden D, Thomas SA, Romanchikova M, Talbott JE, Adeogun M. Building confidence in digital health through metrology. *Br J Radiol* 2020; **93**: 20190574.

COMMENTARY

Building confidence in digital health through metrology

NADIA A. S. SMITH, PhD, DAVID SINDEN, PhD, SPENCER A. THOMAS, PhD, MARINA ROMANCHIKOVA, PhD, JESSICA E. TALBOTT, BSc and MICHAEL ADEOGUN, PhD

National Physical Laboratory, Hampton Road, Teddington, United Kingdom

Address correspondence to: Dr Nadia A. S. Smith
E-mail: *nadia.smith@npl.co.uk*

ABSTRACT

Healthcare is increasingly and routinely generating large volumes of data from different sources, which are difficult to handle and integrate. Confidence in data can be established through the knowledge that the data are validated, well-curated and with minimal bias or errors. As the National Measurement Institute of the UK, the National Physical Laboratory (NPL) is running an interdisciplinary project on digital health data curation. The project addresses one of the key challenges of the UK's Measurement Strategy, to provide confidence in the intelligent and effective use of data. A workshop was organised by NPL in which important stakeholders from NHS, industry and academia outlined the current and future challenges in healthcare data curation. This paper summarises the findings of the workshop and outlines NPL's views on how a metrological approach to the curation of healthcare data sets could help solve some of the important and emerging challenges of utilising healthcare data.

THE NEED FOR WELL-CURATED DATA

The digital revolution and new measurement modalities within healthcare are creating vast amounts of high-dimensional data from disparate sources.[1] These data may include genomic, imaging, blood tests, electronic healthcare records, and data from wearable devices. Accessing and linking different sources of medical data is a challenge due to their multimodal and confidential nature. Meanwhile, healthcare is increasingly reliant on the integration of large data sets as well as trusted and robust analysis methods.[1] The data curation process in healthcare includes extraction, de-identification and annotation of data sets with metadata, as well as data fusion and linkage. Therefore, future-proof scalable curation methods that handle rapidly growing data volumes are needed.

Curated data are essential for reproducible science, as it makes it possible to correctly group or distinguish data sets based on their provenance. This enables meta-analysis to determine whether a difference in the data is real or arises from inappropriate comparisons. For measurements affected by large variations, capture of the experimental setup in the metadata may add information about uncertainties associated with the data set allowing metrological assessment of the confidence in the data. There is also a need to define quality metrics or readiness levels for various types of data. While the idea of technology readiness levels is well-established, the definition of data readiness levels,[2] by which data sets can be located, linked and combined is still in its infancy.

Metrology for data curation

Recent publications have recognised the pivotal role of metrology in increasing confidence in research results and ensuring reproducibility, saving resources and accelerating bench-to-bedside transition.[3–5] Metrology offers multiple tools to support data curation, increase re-use, improve linkage and guarantee consistent quality with growing data volume and veracity. These tools include intercomparisons, reference datasets and standards with an emphasis on traceability and measures of uncertainty.[6,7] Table 1 provides definitions of metrological and other relevant terms used throughout this paper.

Applying metrology tools to data curation will enable objective assessment of data quality, extend data life-time, increase its utility and ensure its compliance with FAIR principles.[8] Implementation of best practices in data curation will support advancement of new analysis tools and help identify predictors of clinical outcomes.

Improving data interoperability and comparability

Data sets with clearly defined quality indicators such as uncertainty estimates, calibration information and interoperability metrics, will provide confidence in analysis of data and thus improve the quality of healthcare at population and patient levels. For example, in multicentre medical imaging studies,

Table 1. Definitions of the most relevant terms used throughout the paper

Data curation	Organisation and integration of data collected from various sources, annotation of the data, and publication and presentation of the data such that the value of the data is maintained over time, and the data remain available for reuse and preservation.
Data interoperability	Addresses the ability of systems and services that create, exchange and consume data to have clear, shared expectations for the contents, context and meaning of that data.
Data integration	Aggregation of data sets from heterogeneous sources by linking, combination or fusion.
Data provenance	A history of the data set that captures its origin, purpose and all modifications it underwent since its creation.
FAIR principles	A set of guiding principles to make data Findable, Accessible, Interoperable and Re-usable.
Measurement	The process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity.
Metadata	A set of descriptors of the data set that make the data set easier to locate, understand and use. Metadata can include information about the purpose of the data set creation, the experimental setup or the data formats. Depending on the nature of the investigation, metadata can be data in its own right.
Metrology	The science of measurement and its application.
Ontology	A formal representation of a knowledge domain that comprises a vocabulary of terms or concepts as well their inter-relationships within the specific domain.
Reproducibility	Condition of measurement, out of a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects.
Traceability	Property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty.

secure data sharing between centres must include a pre-agreed definition of mandatory metadata to capture data provenance and acquisition protocols. These annotations result in clearer interpretability and comparability of data sets, enable uncertainty estimation and future analysis. It must be noted data fusion and variations in data quality of individual data sets present important challenges to be addressed in the future.

Reducing uncertainty in multimodal data

Conventional data handling practices often ignore the fact that data and metadata are multimodal, can originate from different sources, often stored in different formats and have a variety of attributes. Multimodal metadata can be described by several categories, each of which may have multiple sources: information related to the sample and its provenance (*e.g.* patient information); metadata from the experiment from the instrument (*e.g.* protocols and settings); and any operator-related steps (*e.g.* biopsies); information around data processing and analysis. Together, these provide a well-curated data set.

Data provenance, and therefore curation, are predicated on a complete record of all multimodal sources of metadata; deviations from any of these sources can impact the data and downstream analysis. For example, a quantitative assessment of contrast magnetic resonance images taken on different modalities or using different imaging protocols can be extremely challenging. In such cases, traceable and complete image metadata enables the distinction between real observed effects and measurement artefacts or data errors. This is vital for the reproducibility of scientific data[9] and audit trails for detecting errors or scientific fraud.[10]

Integrating large volumes of healthcare data from multiple sources presents further metrological challenges. For example, the combination of images acquired under different conditions may require specific pre-processing or normalisation steps. Curating this combined data requires details of the data integration process, including any processing of the data and linkage of their respective metadata, which may include redundant or conflicting attributes. A well-defined set of rules to combine these multimodal metadata is required for long-term storage, otherwise vital information may be missing for future analysis or attempts to reproduce data.[2,10]

Use of metadata and ontologies to unlock the long-term data value

Due to the lack of structure, standardisation, provenance or knowledge of the experiment setup, a lot of metadata critical for correct data interpretation is lost. Frequently, the only metadata available to the user are file size, file name and last modification date. This reduces the data usability, decreases the data quality and slows down clinical research.

In absence of adequate curation, knowledge about the data decays over time. For example, a large data set of equal numbers of healthy, early-stage and late-stage disease can provide valuable insight into early indicators, changes in heterogeneities, differentiators between stages etc. This insight becomes impossible if the disease staging information is lost, decreasing the dataset value. Well-designed curation mechanisms preserve the long-term data value for future information retrieval,[8] while providing a measure of data relevance and usefulness to the future research context.

The lack of healthcare domain ontologies often makes medical data less usable and their access less systematic and non-replicable. An ontology can provide a systematic replicable method for clinical case definition by supplying domain researchers with a common vocabulary, thus aiding data retrieval, interoperability and curation. Ontological approaches have been successfully developed to identify cases of particular diseases from routine primary care data.[11]

Some efforts to ensure that the data curation complies with FAIR principles have been implemented,[8,12] but there are still necessary future developments when applying these principles to medical datasets: issues around data accessibility and data privacy make the findable and accessible principles difficult to comply with; the lack

of use of common vocabulary, standards and information models such as HL7 FHIR[13] (Fast Healthcare Interoperability Resources) make healthcare data less interoperable, and previously stated data provenance issues mean data can rarely be re-used.

ENGAGING DIGITAL HEALTH PROFESSIONALS TO APPLY METROLOGY METHODS TO DATA CURATION

As part of the National Measurement System, NPL's mission is to support UK prosperity and quality of life. Providing confidence in the intelligent and effective use of data is at the heart of the UK's Measurement Strategy.[14] To address the data curation challenges, NPL is running an interdisciplinary project themed around Digital Health data curation in three areas in line with government priorities and investments[15]: cancerous, cardiovascular and neurodegenerative diseases.

A workshop to engage external stakeholders in outlining the current and future healthcare data challenges was organised by NPL in 2019. During the workshop, participants from NHS, academia and industry helped identify the key challenges in handling and analysing healthcare: a lack of data and metadata standards; data linkage; multimodal sources of data; and data trustworthiness including traceability, quality assurance and uncertainty quantification. Four pilot case studies on curation of healthcare data and metadata were proposed: (1) integration of equipment calibration data with medical images; (2) ontology-based linkage of primary and secondary care data sets for prostate cancer patients; (3) development of data-driven models to identify key prognostic markers in healthcare records; and (4) generation of synthetic data sets with associated data quality metrics. The unique position of NPL, as the UK's National Measurement Institute, is instrumental in facilitating reproducible research and strengthening links between academia, industry, and NHS, and can help address some of the challenges listed above by providing: metrology training to clinical professionals; reference data sets as exemplars of well-curated data with defined quality metrics; validated models and methods, with uncertainty quantification; and standards to capture important clinical data and metadata.

The long-term value of healthcare data and its compliance with the FAIR principles require well-crafted data curation mechanisms. These mechanisms can be implemented through traceable metadata and well-annotated data analysis pipelines that account for uncertainty propagation. Curation of healthcare data will increase long-term data value, improve clinical decision-making and uncover new insights into healthcare problems through secondary or meta-analysis. Curated data sets will form a basis for unsupervised data-driven analysis and advance scientific research through data-assisted-hypothesis generation. Metrology methods and data standards will provide foundation for data quality assessment and increase confidence in clinical data. Furthermore, well curated healthcare data sets could be combined with other sources of non-medical data, opening potentially novel diagnostic pathways that have not been explored so far.

ACKNOWLEDGMENT

The authors would like to thank all the presenters and participants at the workshop.

FUNDING

This work was funded by the department of Business, Engineering and Industrial Strategy (BEIS) through the cross-theme national measurement strategy under the Life Sciences & Healthcare theme (Digital Health – Data Curation).

REFERENCES

1. Topol E. The Topol Review: Preparing the healthcare workforce to deliver the digital future. NHS Health Education England. Available from: https://topol.hee.nhs.uk/ [February 2019].
2. Laurence ND. Data Readiness Levels. ArXiv, 1705.02245v1, 2017. Available from: https://arxiv.org/pdf/1705.02245.pdf.
3. Sené M, Gilmore I, Janssen JT. Metrology is key to reproducing results. *Nature* 2017; **547**: 397–9. doi: https://doi.org/10.1038/547397a
4. Plant AL, Becker CA, Hanisch RJ, Boisvert RF, Possolo AM, Elliott JT. How measurement science can improve confidence in research results. *PLoS Biol* 2018; **16**: e2004299. doi: https://doi.org/10.1371/journal.pbio.2004299
5. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. *PLoS Biol* 2015; **13**: e1002165. doi: https://doi.org/10.1371/journal.pbio.1002165
6. Joint Committee for Guides in Metrology JCGM 200: International vocabulary of metrology – Basic and general concepts and associated terms (VIM). 2012;.
7. Joint Committee for Guides in Metrology JCGM 100: Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement. 2008;.
8. Wilkinson MD, Dumontier M, Jan Aalbersberg I, Appleton G, Axton M, Baak A, et al. Addendum: the fair guiding principles for scientific data management and stewardship. *Sci Data* 2019; **6**: 6. doi: https://doi.org/10.1038/s41597-019-0009-6
9. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016; **533**: 452–4. doi: https://doi.org/10.1038/533452a
10. George SL, Buyse M. Data fraud in clinical trials. *Clin Investig* 2015; **5**: 161–73. doi: https://doi.org/10.4155/cli.14.116
11. Cole NI, Liyanage H, Suckling RJ, Swift PA, Gallagher H, Byford R, et al. An ontological approach to identifying cases of chronic kidney disease from routine primary care data: a cross-sectional study. *BMC Nephrol* 2018; **19**: 85. doi: https://doi.org/10.1186/s12882-018-0882-9
12. Sun C, Ippel L, Wouters B, Soest JV, Malic A, Adekunle O, et al. Analyzing partitioned fair health data Responsibly. *arXiv* 2018;: 1–6.
13. https://www.hl7.org/fhir/overview.html.
14. HM Government. Industrial Strategy: Life Sciences Sector Deal 2, 2018. Available from: https://www.gov.uk/government/publications/life-sciences-sector-deal.
15. NHS report. 'Next Steps on the NHS Five Year Forward View'; March 2017. Available from: https://www.england.nhs.uk/publication/next-steps-on-the-nhs-five-year-forward-view/.