MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220033

Can Researchers Assess the Suitability of Datasets to Answer Their Research Questions, with Access to Metadata Only?

George Tilston[a,b], Richard Williams[a,c], Emily Griffiths[a], Sarah Al-Adely[b,d], Saskia Lawson-Tovey[b,e], Will Hulme[a], Andrea Short[f], Jim Davies[g,h], James Welch[g,h], and Niels Peek[a,b]

[a] Centre for Health Informatics, Division of Informatics, Imaging, and Data Sciences, School of Health Sciences, The University of Manchester, Manchester, UK
[b] NIHR Manchester Biomedical Research Centre, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK
[c] NIHR Greater Manchester Patient Safety Translational Research Centre, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK
[d] Centre for Epidemiology Versus Arthritis, The University of Manchester, Manchester, UK
[e] Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, Manchester, UK
[f] Manchester Centre for Audiology and Deafness (ManCAD), Division of Human Communication and Hearing, School of Health Sciences, The University of Manchester, Manchester, UK
[g] NIHR Oxford Biomedical Research Centre, Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Old Road Campus, Oxford, UK
[h] Department of Computer Science, University of Oxford, Oxford, UK

Abstract

Health research increasingly requires effective ways to identify existing datasets and assess their suitability for research. We sought to test whether researchers could use an existing metadata catalogue to assess the suitability of datasets for addressing specified research questions. Five datasets were described in the National Institute for Health Research Health Informatics Collaborative metadata catalogue, and for each dataset five associated research questions were formulated, some of which were answerable with the dataset while others were not. Thirteen researchers each assessed whether the ten questions associated with two randomly selected datasets were answerable with the described datasets. After removing instances where participants misunderstood the question or lacked subject matter knowledge to make the assessment, we found that 87 out of 109 assessments (80%) were correct. Participants particularly struggled with one dataset which consisted of EHR data. The most common reason for incorrect assessments was the inability to find the relevant information in the metadata catalogue.

Keywords:

Cataloguing, Data Aggregation, Metadata

Introduction

Health research increasingly requires effective ways to identify existing datasets and assess their suitability for addressing prevailing research questions. However, health data ecosystems tend to prevent data reuse, leading to duplication and inefficiency [1]. Reusability is impeded not by lack of data but by bottlenecks in data management and stewardship [2]. Even if researchers know that a particular dataset exists, they will often not know which information is captured by that dataset precisely, and hence will not know whether it could be used to answer their research question. Often this leads to the decision to prospectively collect data where analysing existing data would have been quicker and less resource-intensive.

Metadata catalogues are increasingly recognised as a possible solution for this problem [2–5]. Metadata catalogues enable researchers to browse and search high-level information about health datasets without having to spend time and resources for accessing these datasets – in a similar way as PubMed provides access to high-level information about scientific publications without having to access the expensive full texts. Ideally, this information can then be used to assess which dataset could be used to answer a given research question, thus creating a fast and efficient process for reusing existing health data. However, metadata are not always perfect, and incorrect assessments of such information could lead to wasted time accessing unsuitable datasets, or missing out on datasets that are suitable for answering the given research question(s).

The Health Informatics Collaborative (HIC) metadata catalogue tool was funded by the UK National Institute for Health Research (NIHR)[6]. It was the basis for the UK's Health Data Finder, the Health Data Research Innovation Gateway [7,8], and the National Health Service Data Model and Dictionary for England [9]. The tool enables data custodians to describe datasets by creating metadata models, which researchers can inspect to assess the suitability of datasets for their research. The objective of this study was to test whether the HIC metadata catalogue can successfully facilitate these assessments. Specifically, we aimed to assess whether independent researchers could assess with high accuracy, via the HIC metadata catalogue, whether given datasets were suitable for answering a range of pre-specified research questions.

Table 1 - Datasets that were described in the metadata catalogue and tested in this study

Dataset	Inclusion Criteria	Population size	Data fields	Study period
Lupus Extended Autoimmune Phenotype cohort study [10]	Patients with undifferentiated connective tissue disease from Greater Manchester	500	Patient demographics, medical history, diagnosis date, manifestations of the disease, questionnaire answers: disease manifestations and general wellbeing.	Data collected in 2013
Salford Integrated Record [11]	Salford residents registered at general practices	>250,000	Primary care electronic health record (EHR): diagnoses, laboratory test results, medication prescriptions, symptoms, procedures, referrals.	2008 onwards
Salford Kidney Study [12]	Patients referred to a tertiary renal centre with eGFR< 60 mL/min/1.73m^2 and not requiring immediate renal replacement therapy	3,060	Demographics, comorbidities, laboratory test results, self-reported cerebrovascular and cardiovascular events, and mortality.	2002 to 2018
Imaging studies [13,14]	Participants of studies "STOpFrac: Software Tool for Opportunistic diagnosis of vertebral Fractures" and "An Automated Tool to Identify Vertebral Fractures in Various Imaging Modalities"	25	Information relating to clinical images of the spine: technical information about the device used and the pixels within the image, as well as patient demographics.	2013 onwards
Tissue & blood sample	Fictitious dataset representing typical blood/tissue samples for a musculoskeletal clinical research study.	N/A	Information about the handling, storage, and analysis of nine sample types, including cells, DNA, serum, and tissue.	N/A

Methods

Study design

A prospective, laboratory-based study was conducted at the University of Manchester.

Creating and entering metadata into the catalogue

This study used five patient-level datasets (described in Table 1), which cover a range of health research scenarios. Two datasets (the Lupus Extended Autoimmune Phenotype cohort study [10] and the Salford Kidney Study [12]) stem from clinical cohort studies where a set of predefined clinical measures was collected at specified follow-up times. One dataset, the Salford Integrated Record [11], contained transactional data from electronic health records. The final two datasets were obtained through clinical imaging or by analysing samples from tissue and blood.

For each dataset, we worked with the data custodian who was involved in the data collection, to create metadata and populate the catalogue. To create a uniform set of metadata models, we subsequently reviewed all five metadata models, documenting major differences, and agreeing on actions to resolve them. The metadata models describe how and why each dataset was collected, along with data types and descriptive information (such as value range) for each data field.

Designing research questions

We formulated at least five research questions for each dataset. Working with the relevant data custodian for each dataset, we established whether each question was answerable (yes or no) and ensured that this was unambiguous. In making this assessment, the data custodian used their knowledge and experience of the dataset, and not just the information captured in the metadata catalogue.

The clarity of each question was assessed by a member of the study team that was not involved in the question formulation, and their feedback was used to re-word the questions. The re-worded questions then had their complexity assessed by three independent University of Manchester researchers with experience in health research, who were not part of the study team. Any questions that were considered ambiguous, too complex, or too simple compared to the other questions, were removed. The result was a set of five questions per dataset.

Participant recruitment

Participants were recruited from the Faculty of Biology, Medicine and Health at the University of Manchester. We recruited participants by placing online advertisements in Faculty newsletters, inviting PhD students, postdocs, research fellows, and academic staff with at least three months of experience of working with health research datasets to participate in the study. To take part in the study, participants visited our lab.

Study process

Two datasets were randomly assigned to each participant. We ensured that the participant had not worked with those datasets in their research. Using a laptop to browse the metadata catalogue, the participant assessed five research questions for both datasets, deciding which questions they thought were answerable.

If the participant answered 'Yes', they were asked to provide the necessary data fields to answer the research question. If they answered 'No', they were asked for a justification. A correct justification, or the correct data fields, was required alongside a correct answer. This reduced the chance of achieving a correct answer by guessing. Finally, the participant provided feedback on their experience using the catalogue.

To understand the likely reason behind each incorrect assessment, six categories were developed. Two authors (GT and RW) independently assigned each incorrect assessment to a category, based on the answer and justification provided. Most assessments were assigned to the same category by both authors. Where disagreements occurred, another author (NP) adjudicated and decided the most suitable category (Table 2).

Outcome measures

Our outcome measure was the proportion of correct assessments out of all assessments made. This was established by comparing each assessment against the correct answer determined by the study team together with the data custodians.

Sample size

Metadata catalogues can describe tens or hundreds of different datasets. We therefore aimed to assess whether our metadata catalogue provided sufficient information to assess datasets with high accuracy, corresponding to a success rate of at least 90%. Hence the study was powered to detect a significant non-inferiority from a 90% success rate with the 95% confidence interval having width no more than 10%. Exact confidence intervals were used. For a superiority approach (is the success rate over 90% at a 95% confidence level?) this yields a minimum sample size of 36 (success rate = 1, width = 0.097, 95% lower bound = 0.903, alpha = 0.05) per question.

Since we asked 5 questions per dataset (reducing sample size by a factor of 5), there are 5 datasets in total (increasing sample size by a factor of 5), and each participant was asked to consider at least two datasets (reducing sample size by a factor of 2), the target sample size was 36*(5*(1/5)*(1/2)) = 18 participants.

Ethics

In line with University guidelines, this study was reviewed by the Research Ethics Signatory for the Division of Informatics, Imaging and Data Sciences. It was deemed exempt from requiring an ethical review because it was service evaluation by professionals in their professional capacity [15].

Results

In total, 13 participants were recruited, with roles varying from PhD students to clinical lecturers. 11 participants (85%) held a PhD or MSc Degree as their highest qualification. Seven participants (54%) had at least five years, and 11 (85%) had at least two years, of experience in health research.

Table 2 - Incorrect assessment categories

Category	N
Unable to find information	13
Correct reason but incorrect answer	5
Incorrect field chosen	4
N/A: Not enough information provided	4
Misunderstood research question	12
Lack of knowledge	5

From a total of 130 assessments, 87 were correct, meaning that the participant answered yes or no correctly and named the correct data fields or an acceptable justification for this assessment. The 43 incorrect assessments are categorised in Table 2.

Three of the categories ('Misunderstood', 'Lack of Knowledge', and 'N/A') were excluded from analysis for failing to replicate a real-life research scenario. In these cases, the incorrect assessments seemed to be due to the individual's lack of subject matter knowledge or misunderstanding of the question, rather than due to interaction with the catalogue. This resulted in 87 correct assessments out of 109.

Table 3 contains a breakdown of these assessment outcomes across the five datasets. Inaccurate assessments were balanced between false positives (believing that a research question was answerable where in fact it was not; 11 instances) and false negatives (believing that a research question was not answerable where in fact it was; also 11 instances).

Table 3 - Assessment outcomes across the five datasets. A false positive represents an inaccurate assessment of a research question as answerable. A false negative represents an inaccurate assessment of a question as unanswerable.

	Correct	False positive	False negative
Lupus Extended Autoimmune Phenotype cohort study	20	0	3
Salford Integrated Record	12	5	3
Salford Kidney Study	15	2	0
Imaging studies	22	1	2
Tissue & blood sample	18	3	3

Dividing the number of correct assessments by the total number of assessments after the exclusions yields a success rate of 80% (95% confidence interval: 72% to 87%). Given that the desired result (90%) was not within the 95% confidence range, it was assumed that the success rate would not be met by recruiting five more participants to reach the target of 18 participants.

The average time taken to assess a dataset's suitability for five research questions was 15.3 minutes. The variation in assessment accuracy between the five datasets was found to be insignificant. However, a post-hoc test comparing the Salford Integrated Record against the other four datasets found a significant difference ($X^2 = 4.56$, df = 4, $N = 109$, $p = .03$). The variation between individual participants was also significant ($X^2 = 27.20$, df = 12, N = 109, p = .0.01) Number of years' experience in health research was not associated with assessment accuracy.

In a sensitivity analysis we did not exclude the assessments categorised as 'N/A', 'Misunderstood' or 'Lack of knowledge', and found a success rate of 67% (95% CI 59% to 75%).

Feedback

At the end of each study session, participants were asked to provide general feedback on the catalogue. The most common suggestion was to add extra detail in field names and descriptions. Secondly, participants highlighted the inefficient search function, suggesting improvements to its predictive capability. Thirdly, the catalogue was difficult to use initially but became easier to use over time. Participants suggested a user manual for first-time users. Positive comments included the tool's ease of navigation and intuitive folder structures.

Discussion

This study investigated whether researchers could use a metadata catalogue to assess the suitability of given datasets to answer given research questions. Participants made correct

assessments in 80% of cases, which failed to meet our pre-defined success rate (90%).

To the best of our knowledge, this is the first study that tests whether researchers can make accurate assessments on the utility of health datasets using a metadata catalogue. Dixit and colleagues have previously conducted a usability assessment of DataMed, a catalogue that collates machine-generated metadata [16]. Similar to what was found in that study, several of our participants mentioned the inefficiency of searching for specific terms within the catalogue [16]. This highlights the need for metadata catalogues to implement good search functionalities.

On average, three minutes were spent per assessment. Participants spent significantly less time on their second attempt compared to their first. This may be due to the different datasets involved, but the fact that the trend is seen across the participants suggests an increased familiarity with the catalogue after their first attempt. As suggested in the feedback, a manual for first time users could further reduce the assessment time, and potentially improve researchers' interest in the catalogue [5].

Our analysis revealed a significantly lower assessment accuracy for metadata related to EHR data. This kind of data, collected for direct care rather than research, has a more complex structure with many more tables and data fields than typical research cohort datasets.

Various studies have mapped disparate sets of metadata to a common metadata model [3,4,16–18]. However, our study revealed that more complex datasets (such as the EHR data) were more difficult to describe, thus highlighting the fact that enforcing consistency between metadata that differs in structure, detail, and complexity is not trivial.

Limitations

Although each of them assessed ten research questions, the small number of participants was a limitation of this study.

For pragmatic reasons, we only used one type of metadata catalogue software in this study. The findings might have been different if other software had been used. But this seems unlikely as all these tools operate in the same way.

The metadata created for each dataset provided information on data schema and data types but lacked information on data quality and completeness. In a real-world setting, data quality and completeness would also influence the feasibility of answering specific research questions, but this was not considered in our study.

Participants and datasets were re-used several times, which may have led to clustering of results. Re-using datasets and participants in this way was essential given resource limitations. Clustering was considered by analysing the variance between datasets and participants, which was detectable in both cases but not enough to alter conclusions.

Unanswered questions and future research

Research is often carried out in a multi-disciplinary team, so future research should test the catalogue tool with groups of researchers rather than individuals.

The metadata were created manually. Future studies could compare participant assessments using manual versus automatically generated metadata.

Finally, if the metadata were developed to include more specific information about data fields, such as completeness, it would be useful to test how this affects assessment accuracy.

Conclusion

A metadata catalogue of health research datasets, that was constructed manually in collaboration with data custodians, provided insufficient information for experienced health researchers to assess with high accuracy whether these datasets could be used to answer given research questions. Participants made correct assessments on the feasibility of answering research questions in 67% to 80% of cases. They particularly struggled to assess the suitability of EHR data for answering given research questions. The most common reason for incorrect assessments was the inability to find the relevant information in the metadata catalogue.

Data Sharing

The research questions and participant results are available via figshare [19].

Acknowledgements

We thank the participants for their time and feedback, Ms Caige Huang for recruiting participants, and Dr Paul Bromiley and Dr Nisha Nair for helping us to develop metadata and related questions, for the Imaging and Tissue datasets, respectively.

This research was funded by the National Institute for Health Research (NIHR) Manchester Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

This work was facilitated by infrastructure development at Oxford Biomedical Research Centre funded by the NIHR Health Informatics Collaborative (HIC).

References

[1] T. Souza, R. Kush, and J.P. Evans, Global clinical data interchange standards are here!, *Drug Discov. Today*. **12** (2007) 174–181. doi:10.1016/j.drudis.2006.12.012.

[2] X. Chen, A.E. Gururaj, B. Ozyurt, R. Liu, E. Soysal, T. Cohen, F. Tiryaki, Y. Li, N. Zong, M. Jiang, D. Rogith, M. Salimi, H. eui Kim, P. Rocca-Serra, A. Gonzalez-Beltran, C. Farcas, T. Johnson, R. Margolis, G. Alter, S.A. Sansone, I.M. Fore, L. Ohno-Machado, J.S. Grethe, and H. Xu, DataMed - an open source discovery index for finding biomedical datasets, *J. Am. Med. Informatics Assoc.* **25** (2018) 300–308. doi:10.1093/jamia/ocx121.

[3] M.A. Musen, C.A. Bean, K.H. Cheung, M. Dumontier, K.A. Durante, O. Gevaert, A. Gonzalez-Beltran, P. Khatri, S.H. Kleinstein, M.J. O'Connor, Y. Pouliot, P. Rocca-Serra, S.A. Sansone, and J.A. Wiser, The center for expanded data annotation and retrieval, *J. Am. Med. Informatics Assoc.* **22** (2015) 1148–1152. doi:10.1093/jamia/ocv048.

[4] J.L. Oliveira, A. Trifan, and L.A. Bastião Silva, EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data, *Int. J. Med. Inform.* **126** (2019) 35–45. doi:10.1016/j.ijmedinf.2019.02.006.

[5] F. Wang, C. Vergara-Niedermayr, and P. Liu, Metadata based management and sharing of distributed biomedical data, *Int. J. Metadata, Semant. Ontol.* **9** (2014) 42–57. doi:10.1504/IJMSO.2014.059126.

[6] National Institute for Health Research, Metadata – Health Informatics Collaborative, (n.d.). https://hic.nihr.ac.uk/metadata (accessed May 17, 2021).
[7] National Institute for Health Research, Health Data Finder for Research, (2020). http://hdf.nihr.ac.uk/ (accessed August 26, 2020).
[8] Health Data Research UK, HDRUK Innovation Gateway, (2020). https://www.healthdatagateway.org/ (accessed September 21, 2020).
[9] National Health Service, NHS Data Model and Dictionary, (2020). https://datadictionary.nhs.uk/index.html (accessed November 4, 2020).
[10] Health Research Authority, Lupus Extended Autoimmune Phenotype Study (LEAP), *Lupus Ext. Autoimmune Phenotype Study (LEAP)*. (n.d.). https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/lupus-extended-autoimmune-phenotype-study-leap/ (accessed August 26, 2020).
[11] J. Tollitt, A. Odudu, E. Flanagan, R. Chinnadurai, C. Smith, and P.A. Kalra, Impact of prior stroke on major clinical outcome in chronic kidney disease: The Salford kidney cohort study, *BMC Nephrol.* **20** (2019) 432. doi:10.1186/s12882-019-1614-5.
[12] J.P. New, D. Leather, N.D. Bakerly, J. McCrae, and J.M. Gibson, Putting patients in control of data from electronic health records, *BMJ.* **360** (2018) j5554. doi:10.1136/bmj.j5554.
[13] B. PA, STOpFrac: Software Tool for Opportunistic diagnosis of vertebral Fractures, (n.d.). https://personalpages.manchester.ac.uk/staff/paul.a.bromiley/stopfrac.html (accessed August 26, 2020).
[14] Cootes TF, Project: An automated tool to identify vertebral fractures in various imaging modalities, (n.d.). https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/Projects/HICF_VertFrac/hicf_vert_frac.html (accessed August 26, 2020).
[15] The University of Manchester, Does Your Research Require Ethical Approval? | StaffNet | The University of Manchester, (2020). http://www.staffnet.manchester.ac.uk/services/rbess/governance/ethics/does-your-research-require-ethical-approval/ (accessed March 4, 2021).
[16] R. Dixit, D. Rogith, V. Narayana, M. Salimi, A. Gururaj, L. Ohno-Machado, H. Xu, and T.R. Johnson, User needs analysis and usability assessment of DataMed - a biomedical data discovery index, *J. Am. Med. Informatics Assoc.* **25** (2018) 337–344. doi:10.1093/jamia/ocx134.
[17] S.A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.A. Coleman, J. Copeland, S. Das, A. De Daruvar, P. De Matos, I. Dix, S. Edmunds, C.T. Evelo, M.J. Forster, P. Gaudet, J. Gilbert, C. Goble, J.L. Griffin, D. Jacob, J. Kleinjans, L. Harland, K. Haug, H. Hermjakob, S.J.H. Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C.E. Shamu, C.A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, and W. Hide, Toward interoperable bioscience data, *Nat. Genet.* **44** (2012) 121–126. doi:10.1038/ng.1054.
[18] S.A. Sansone, A. Gonzalez-Beltran, P. Rocca-Serra, G. Alter, J.S. Grethe, H. Xu, I.M. Fore, J. Lyle, A.E. Gururaj, X. Chen, H.E. Kim, N. Zong, Y. Li, R. Liu, I.B. Ozyurt, and L. Ohno-Machado, DATS, the data tag suite to enable discoverability of datasets, *Sci. Data*. **4** (2017) 170059–170059. doi:10.1038/sdata.2017.59.
[19] G. Tilston, Testing the usefulness of a metadata catalogue for researchers to assess the suitability of biomedical datasets: results, (n.d.). https://figshare.com/articles/dataset/Testing_the_usefulness_of_a_metadata_catalogue_for_researchers_to_assess_the_suitability_of_biomedical_datasets/13516730 (accessed January 4, 2021).

Address for correspondence

Professor Niels Peek, Centre for Health Informatics, Division of Informatics, Imaging, and Data Science, The University of Manchester, Vaughan House, Portsmouth Street, Manchester, M13 9GB. Niels.peek@manchester.ac.uk