Data and Services Discovery projects - Transformative Data Collections

Title

What large clinical data sets exist across the member organisations of MACH, Melbourne and how can we collate and curate them to maximise research outcomes?

Approach

This project was undertaken under the auspices of the University of Melbourne Petascale Campus Initiative – a five year technology refresh initiative at the University of Melbourne and the Melbourne Academic Centre for Health (MACH). Under MACH, there is a collaboration agreement between hospital and institute partners and The University under which knowledge discovery projects and collaborations can proceed. Utilising these collaboration foundations, this project aimed to understand datasets of potential research interest held across MACH partners.

Activities: Our initial approach was to undertake interviews with key stakeholders. In month one we determined that organisational permission to run such a survey varied across the institutions and regardless of any collaboration agreement, we would need to follow established processes within each organisation. This combined with a need to find a dataset discovery model that was sustainable led us to develop an administration and survey model that MACH and similar organisations may operate on an on-going basis.
Our activities involved:

1. Survey tool consultation with stakeholders and the generation of a survey tool utilising REDCap operating at two levels:
    a. Initial knowledge discovery: Identification of datasets and the principle contact for each dataset
    b. In-depth metadata collection: Follow-up as required for datasets of strategic value
2. Engagement with ethics and data governance officers at each of the collaborating organisations
3. Collation of the organisational requirements to undertake the survey including requirements for Good Clinical Practice training, recruitment materials, plain English language statements, formal ethics application or quality assurance project applications, identification of organisational sponsors for the survey and any associated fees.
4. Completion of all organisational requirements to gain permission to undertake the survey.
5. Collaborative working with each institution for the survey to be released to clinical and research staff once permissions had been received.

Participants: We engaged through survey with clinicians and researchers across the MACH institutions in central, northern and western Melbourne including: Austin Health, the Bionics Institute, Melbourne Health, Northern Health, The Royal Children's Hospital, St. Vincent's Hospital, The Eye and Ear Hospital, The Women's, Peter Mac and Western Health. Engagement and participation is on-going at some institutions.

Outputs: REDCap databases have been established for each organisation with data collection able to be on-going. Approaches to organisational issues and the mechanisms of engagement form the principle learnings of the project both for MACH and for wider audiences. The processes required for MACH to engage in further surveys as well as examples of the engagement mechanism and documents are now established. The principle findings are outlined below and shall be presented at the 2019 eResearch Conference and the NHMRC Translational research conference in November 2019. An academic paper is also under development. Continuation of this work via the University of Melbourne and MACH is planned with the dataset dictionary being made available to researchers in collaboration with the hospital partners and the ARDC.

FAIR - Findability, Accessibility, Inter-operability and Re-Usability

This project has resulted in us having a baseline knowledge of datasets that was not available previously. We anticipate that this data catalog will have future impact on the findability, accessibility and re-usability of many of these datasets going forward with us anticipating the number of datasets represented growing markedly over time.

In most cases, institutional approval to run the survey has been very recent and hence survey responses at the time of this report is limited. To date and as a result of this project approximately 100 datasets now have a recorded baseline FAIR status. The overall FAIR assessment for the datasets discovered was:

- Findability:
  - A contact person and contact details have been applied to a dataset.
  - In most cases keywords have been applied to assist in findability and relevance analysis
  - Most datasets have a unique ID or a URL was also provided in some cases.
- Accessibility: due to engagement constraints, detailed technical information on accessibility was not captured
- Inter-operability: due to engagement constraints, detailed technical information on inter-operability was not captured
- Re-usability: the survey identified if the dataset was available to researchers and therefore available for re-use.

The FAIR assessment spreadsheet is found in Appendix 1

Collaboration and coverage

The project's scope was MACH institutions in central, northern and western Melbourne. Coverage was achieved across twelve of the eighteen sites in MACH including all of the largest and most research-active hospitals. Further coverage is possible going forward although the outstanding sites include organisations that do not undertake human research and some are small institutes.

Engagement and collaboration with institutional ethics and research offices was good. Although supportive, approval pathways and governance processes varied widely.

Sustainability

Great effort was required to achieve institutional agreements to proceed with this survey. The value of having an on-going dataset repository has been recognised. Now that these agreements have been achieved, the University plans to continue to resource the data collection in the coming months beyond the initial ARDC project funding. The survey will play a role in the strategic planning for Health Data Science at the University of Melbourne and the options to continue the maintenance of the survey on an on-going basis will form a part of these discussions. Further surveys will also be required through MACH on an on-going basis. The documentation and approvals processes required for each organisation are now a core resource that will be made available to MACH on an on-going basis. The challenges faced are also able to form the basis for discussions about how MACH works with the partner hospitals going forward and we hope the learnings will be more widely applicable.

Learnings

The challenges of institutional engagement were at the center of this project. Below are some of the key learnings:

Willingness to participate: Most organisations approached were willing and in some instances keen to discover what datasets were held in their institution. Dataset discovery as a concept was welcomed by leads, ethics departments and most governance approvers. It was recognised however that the survey uptake would be poor as dataset discovery is a low priority for potential respondents. There was also no internal capacity to do more than send recruitment emails with a survey link to department heads, and then follow up with a reminder email. There was no capacity in any organisation to follow up and promote survey completion at departmental or unit level in the time available.

Inconsistent response to survey request: Institutional ethics and governance responses to the survey request were inconsistent. Each institution evaluated and then managed the project's ethics and governance risk differently. Many of the larger institutions are regularly approached for access to patient/participant records and are highly conscious and protective of these records.

Ethics requirements for the survey differed from 'Low to Negligible Risk' approval to no ethics approval. The same inconsistency applied to governance or institutional support requirements which varied from an email exchange to signing of the MACH collaboration agreement. This inconsistent response was escalated to the MACH executive and may result in a more standardised process for engagement for non-research project surveys that are not accessing patient or participant data going forward.

Regardless of wording in the documents used to engage with the organisations, repeated reassurances were often required in the following areas:

- Assurances that no patient/participant records were to be accessed
- Clarity that the project was not a research project.
- Clarity that the project was not just interested in research data, but also clinical data.

Legacy survey materials and outcome: Each institution which took the survey will receive the results of their own survey and that of other willing MACH institutions. The project has no insight into what the institution does with this information. Each institution has however been provided with the data dictionary and recruitment templates used in the survey for their own use. One institution chose to use the data dictionary as the basis for their own survey independent of the ARDC survey.

Process Learnings:

- Providing funding at institutional level for engagement and dataset discovery would help.
- Start engagement via the ethics and governance departments. They have experience in evaluating the perceived risk of projects and provide guidance on approvals. This project spent time contacting leads when ultimately the ethics and governance teams were the units who could effect the survey.
- Recognise that dataset discovery by an external organisation is voluntary and is a low to no priority for a busy researcher or clinician. Options other than voluntary online survey could include: survey plus in-house promotion to complete which requires in-house resourcing or inducements to complete such as movie tickets
- Recognise that the target organisation has no staff to closely support discovery. Institutional staff can do little more than seek approvals and prepare and send emails. A well-resourced in-house project with dedicated staff to follow up on survey completion would greatly improve discovery
- Limit the number of institutions engaged in the first instance. The scope of the project was too wide starting analysis with eighteen MACH institutions. This approach was necessary given a three-month project window for completion. Given more time, the engagement would have been piloted and refined before a wider implementation.
- Recognise the diversity of ethics and governance requirements. Each organisation will have its own approach to ethics approval based on the institution's purpose and history. Each organisation will have its own governance support. Work with the organisation to identify governance and ethics approval processes and carefully follow them. Don't miss the ethics approval cycle.

For the ARDC, the learnings here are directly applicable and can help any similar national dataset discovery initiative at the design stage to achieve better and faster outcomes.

Impact

The project is intended as a first step towards collating clinical and research health data from hospitals, clinical practices and research institutes that fall under the collaborative umbrella of the MACH into a curated data commons for research. The project has successfully engaged with the majority of MACH members and has collected dataset metadata.

The immediate benefit of the project was somewhat contrarily to a) highlight to the MACH how administratively difficult dataset discovery is across the diverse membership of MACH, and b) demonstrate how keen most institutions we engaged with are to discover their datasets.

As a result of the project, participating institutions have a set of email templates and a survey model to reuse for their own dataset discovery, a list of their own discovered datasets and a list of other MACH members datasets for analysis for collaboration. Three institutions have indicated that they will reuse the survey and engagement material for ongoing dataset discovery.

A research paper is in development and funding is being sought to continue the data collection. The collaboration and engagement will play a key role in helping formulate the University of Melbourne and MACH's health data strategy going forward and hence this exercise is a key contribution in this space. Through this knowledge discovery and contribution to The University and MACH strategy, it is likely that the survey will have a direct impact in the utilisation of datasets identified by this initiative for research going forward.

Broader impact: Once fully collated and released, this survey will provide opportunity to the Victorian research community as well as potential for wider government and industry engagement. Impact generated is likely to be through the new research and development opportunities found through the innovative linkage of datasets that were previously unknown or not well known. This research and innovation and productivity benefit is on-going. If the dataset discovery continues, this will help avoid data silos going forward and maximise data use. The project methods, results and learnings will be more extensively reported in a research paper that is currently under development. This will be actively shared across the Australian Health Research Translation Centers via the Australian Health Research Alliance. The findings will also be made available to the ARDC for wider distribution and inclusion in any national dataset collection initiative.

Report prepared by: Assoc Prof Douglas Boyle and Ursula Soulsby, The University of Melbourne
Date: 08/10/2019