International Journal of Population Data Science

Journal Website: www.ijpds.org

Developing a linked electronic health record derived data platform to support research into healthy ageing

Nadine E. Andrew[1,2,*], Richard Beare[1,2], Tanya Ravipati[2], Emily Parker[2], David Snowdon[1,2], Kim Naude[2], and Velandai Srikanth[1,2,3]

Submission History

Submitted:	28/11/2022
Accepted:	08/04/2023
Published:	12/06/2023

[1]National Centre for Healthy Ageing, Frankston, Victoria, Australia
[2]Department of Medicine, Peninsula Clinical School, Central Clinical School, Monash University, Frankston, Victoria, Australia
[3]Department of Medicine & Geriatric Medicine, Frankston Hospital, Peninsula Health, Melbourne, Australia

Abstract

Introduction

Digitalisation of Electronic Health Record (EHR) data has created unique opportunities for research. However, these data are routinely collected for operational purposes and so are not curated to the standard required for research. Harnessing such routine data at large scale allows efficient and long-term epidemiological and health services research.

Objectives

To describe the establishment a linked EHR derived data platform in the National Centre for Healthy Ageing, Melbourne, Australia, aimed at enabling research targeting national health priority areas in ageing.

Methods

Our approach incorporated: data validation, curation and warehousing to ensure quality and completeness; end-user engagement and consensus on the platform content; implementation of an artificial intelligence (AI) pipeline for extraction of text-based data items; early consumer involvement; and implementation of routine collection of patient reported outcome measures, in a multisite public health service.

Results

Data for a cohort of >800,000 patients collected over a 10-year period have been curated within the platform's research data warehouse. So far 117 items have been identified as suitable for inclusion, from 11 research relevant datasets held within the health service EHR systems. Data access, extraction and release processes, guided by the Five Safes Framework, are being tested through project use-cases. A natural language processing (NLP) pipeline has been implemented and a framework for the routine collection and incorporation of patient reported outcome measures developed.

Conclusions

We highlight the importance of establishing comprehensive processes for the foundations of a data platform utilising routine data not collected for research purposes. These robust foundations will facilitate future expansion through linkages to other datasets for the efficient and cost-effective study of health related to ageing at a large scale.

Keywords

electronic health record; data linkage; big data; ageing; longitudinal cohort

[*]Corresponding Author:
Email Address: nadine.andrew@monash.edu (Nadine E. Andrew)

https://doi.org/10.23889/ijpds.v8i1.2129

Highlights

- Electronic Health Record (EHR) data contains rich clinical, demographic and health service data that can enhance research. However, variations in data quality, completeness and transparency can lead to distrust by researchers.
- The National Centre for Healthy Ageing, Healthy Ageing Data Platform has developed an EHR based research data warehouse containing a core set of high quality, curated data items identified as relevant to healthcare and clinical research.
- Harnessing linked EHR data across multiple healthcare sectors, in a geographic population, has created a naturalistic cohort to understand health journeys across the life-span.
- Processes implemented in the Healthy Ageing Data Platform have the potential to be replicated in other settings, allowing researchers streamlined access to research grade EHR derived data.

Introduction

Globally, the number of people aged 60 years and over has tripled since 1950 and is projected to double over the next 30 years resulting in an ageing population with increasingly complex health needs [1]. In response, there is an impetus for health care systems to transition from traditional single disease-based models of care to more comprehensive interdisciplinary models with a focus on holistic person-centred approaches. Electronic Health Record systems (EHRs) are being rapidly implemented globally, with potential to provide rich clinical data across multiple healthcare sectors at scale [2, 3]. Bringing together sources of EHR data both within and between hospital EHR systems can allow a more comprehensive view of a person's healthcare journey and provide longitudinal data on individuals and populations to monitor disease trends and patient outcomes.

For research, such a resource would provide rich longitudinal data at a much lesser cost than traditional prospectively recruited cohort studies. When applied to a defined region, data on the majority of the population can be captured, reducing biases and generalisability issues inherent in other study designs. Due to the large scale of data there is sufficient statistical power to derive accurate estimates at a population level and the ability to examine rare clinical groups and populations who are traditionally under-represented in recruited studies. The use of routinely collected data within a geographic region can also provide data infrastructure that can be harnessed to support more detailed data from nested cohort studies and provide a framework for healthcare trial evaluation.

Despite the potential of EHR data for health research, there are a number of limitations that impact its suitability. Data within EHR systems comprise both structured and unstructured data that have been gathered, captured and assessed during routine clinical care. As such, their primary purpose is that of clinical documentation and business reporting – not research. Due to operational complexities, lack of interoperability between systems and a lack of standardisation of data models, these data are inconsistently recorded both within and between healthcare organisations. When applied to research, these inconsistencies can lead to substantial and cumulative bias resulting in distrust from researchers about the quality and usefulness of EHR data [4]. These constraints, if not addressed, means that the use of these data for healthcare evaluation and other forms of research can be limited.

In 2019, The National Centre for Health Ageing (NCHA) was funded by the Australian Federal Government to develop an EHR derived data platform. The overarching purpose being to make better use of existing health data to support transformative research and translation in health service and aged care innovation, and gain insights into the epidemiology of ageing. The aim of this paper is to provide an overview of the activities undertaken to establish the foundations of the NCHA Healthy Ageing Data Platform.

Methods

Setting

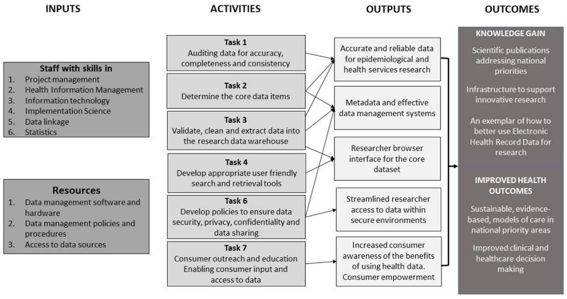
The NCHA is a partnership between Monash University and Peninsula Health and is located in the Frankston/Mornington Peninsula region in Victoria, Australia. The partnership draws upon the strengths of the two organisations. Monash University is a leading international medical research university and is at the forefront of cutting-edge, innovative and translational research involving sensitive data. Monash researchers are supported by significant investment in world-class infrastructure, platforms and services that were used to support the development of the Healthy Ageing Data Platform.

Peninsula Health is the only public health provider for the Frankston/Mornington Peninsula region, services a population of approximately 311,000 people within Australia and is an early adopter of EHR systems. The region contains both metropolitan and regional areas, has wide socio-economic diversity, an ageing population (20% aged >65 years) and one of the highest rates of dementia in Australia [5]. Located on a peninsula, residents access the majority of their healthcare through Peninsula Health, which includes two acute care hospitals, two rehabilitation hospitals and over ten outpatient and community health centres. Unlike most health organisations in Australia, Peninsula Health allocates a single unit record (UR) number to patients across all of the services. These features made Peninsula Health an attractive organisation to partner with in the development of a linked EHR derived cohort for health services research specific to ageing. The activities involved in establishing the Healthy Ageing Data Platform are outlined in Figure 1 and are described in detail below.

Establishment team

The complexity of this undertaking required a multidisciplinary team of Information Technology (IT) professionals, epidemiologists/biostatisticians, implementation scientists, data scientists, legal and governance officers and clinicians.

Figure 1: Program logic model for the development of the Healthy Ageing Data Platform


INPUTS
ACTIVITIES
OUTPUTS
OUTCOMES
Staff with skills in
1. Project management
2. Health Information Management
3. Information technology
4. Implementation Science
5. Data linkage
6. Statistics
Resources
1. Data management software and hardware
2. Data management policies and procedures
3. Access to data sources
Task 1
Auditing data for accuracy, completeness and consistency
Task 2
Determine the core data items
Task 3
Validate, clean and extract data into the research data warehouse
Task 4
Develop appropriate user friendly search and retrieval tools
Task 6
Develop policies to ensure data security, privacy, confidentiality and data sharing
Task 7
Consumer outreach and education Enabling consumer input and access to data
Accurate and reliable data for epidemiological and health services research
Metadata and effective data management systems
Researcher browser interface for the core dataset
Streamlined researcher access to data within secure environments
Increased consumer awareness of the benefits of using health data. Consumer empowerment
KNOWLEDGE GAIN
Scientific publications addressing national priorities
Infrastructure to support innovative research
An exemplar of how to better use Electronic Health Record Data for research
IMPROVED HEALTH OUTCOMES
Sustainable, evidence-based, models of care in national priority areas
Improved clinical and healthcare decision making

Early establishment of the team with a clear and shared vision of the end goal has been critical. To maximise integration within the Healthcare Organisation/University partnership, key technical staff were employed by Peninsula Health, funded through the NCHA, and integrated within existing Peninsula Health operational teams. Academic leads were given appointments with Monash university but also have honorary appointments with Peninsula Health.

Data management

Most EHR systems, including Peninsula Health, are examples of relational databases that require internal linkages. As the Peninsula Health system contains multiple systems, e.g. Cerner (inpatient data), Mastercare (community data), iPharmacy (pharmacy data) etc, linkages also need to occur across systems. This can create challenges from both a technical and internal governance perspective, especially as these data are primarily collected stored and managed for operational and not research purposes. Consequently, having a robust internal data governance framework that can maximise the value of the data for research whilst managing data-related risks is paramount [6].

For security purposes all operational activities for the Healthy Ageing Data Platform are undertaken inside the Peninsula Health's data centre and so are governed by the Peninsula Health cyber security policies. This approach ensures compliance with data security related to the operations associated with the transmission and storage of data. To avoid negative impact on operational systems key tables are copied, with weekly incremental updates, into staging areas within the NCHA data warehouse where business rules and cleaning operations are applied to merge the data with the core dataset. Once the warehouse is fully established, audit trails will be automated within the Healthy Ageing Data Platform warehouse to monitor who is accessing the data, movement of data between systems and changes in data use and structure. Audit requirements will be regularly reviewed to ensure that they are responsive to regulatory requirements.

Internal data access is guided by the separation principle whereby identifiable data are kept separate to clinical data and staff only have access to the information needed for them to perform their role. For example, staff involved in managing the data are employed by Peninsula Health, have limited access rights to the data and are not given access to patients' full records unless required. Person level data are assigned a platform ID within the Research Data Warehouse which can only be linked back to the hospital UR number using a linkage map held within Peninsula Health but external to the warehouse.

Data quality and completeness is of the utmost importance and is critical to determining the suitability of items for incorporation into the research data warehouse. The World Health Organisation's Data Quality Review (DQR) Framework [7] is being used to guide the assessment of data quality. The DQR dimensions include: (1) completeness i.e. the proportion not containing null or blank values; (2) specificity, calculated by excluding unknown, not specified, not stated, or 'other'; (3) timeliness, validated against data recording time; (4) internal consistency, measured as consistency between similar variables recorded within different datasets; and (5) external consistency, validated against published literature and data collection points.

Determining the content

To address the excessively large amount of data of varying quality within the Peninsula Health EHR systems, a decision was made to focus on high value, high quality data items that are frequently used in healthcare research. An existing, published review of the international literature was used as one of the sources to identify EHR items commonly used in research [8]. Although the overarching purpose of the NCHA is to support transformative research and translation in the area of healthy ageing, the core data items did not need to have a specific focus on ageing research. This means that the Healthy Ageing Data Platform can be used for any type of investigator led research. The Peninsula Health EHR systems were then audited to determine item availability and information on data quality was derived from assessment against the DQR. Items that were not routinely available or performed poorly on multiple aspects of the DQR were removed from the list.

To further refine the content, researchers from Monash University and clinicians from Peninsula Health, from a broad range of areas, were invited to take part in a modified Delphi process, a structured communication method to reach a decision by surveying a panel of experts [9, 10]. The majority of voting is performed online using live online survey/polling tools. Each workshop is focussed on a particular set of data items. For example, workshop 1 presented items from the demographic, inpatient and theatre/surgical datasets. Prior to each online workshop participants are sent information about the items that will be polled at that particular workshop. During the online workshop but prior to polling, each data item is presented with information on use and quality. Participants are encouraged to ask questions about the data items and subject matter experts are consulted prior to providing advice to the group on specialised datasets such as radiology or costing data. Group members who are unable to attend the online workshop are given the option of completing an online survey instead. They are also provided with an opportunity to meet with the Data Operations Manager to discuss the items prior to completing the survey. Some items e.g. age, gender, admission/contact dates were considered mandatory and were not polled but were presented to the group for comment. Participants rated the items for inclusion using a 6-point Likert scale and consensus was set at >70%. Results are collated and provided to the participants for review and comment. If consensus is not gained on all items a second round of online polling is performed and after this second round items that remain under 70% are excluded.

Data access governance

The five-safes framework is our overarching principle for guiding data governance related to data access and release (Figure 2). The framework was developed by the United Kingdom (UK) office for National Statistics in 2003. It has since expanded to a number of other countries and contexts including government and health organisations in countries such as the United Kingdom, Germany, Australia and New Zealand. The principles of the five-safes [11] ensure oversight with reference to safe: (i) people - user trust with regards to the knowledge and data management skills; (ii) projects – appropriate use of the data from a legal, moral and ethical perspective, (iii) systems - authorisation of use controls over data access (e.g. use of secure analysis environments) (iv) data - potential for identification of individuals within the data, (v) outputs - ensuring outputs/publications maintain confidentiality.

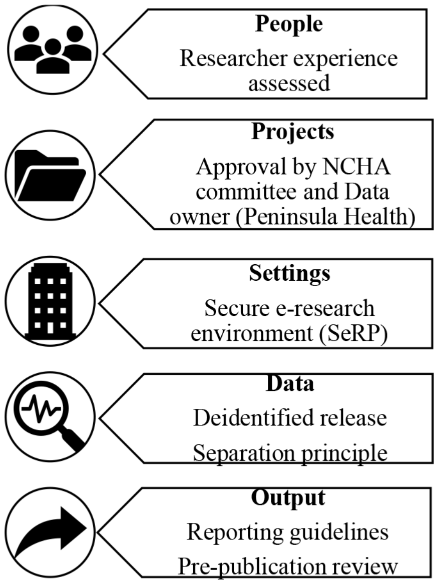
A working group, consisting of senior legal, research, ethics and operation managers from Peninsula Health and Monash University was established. The groups purpose was to develop a data access governance framework that would support: streamlined access to data for research (including quality improvement, clinical and epidemiological research); transfer of data into secure research environments for analysis; linkage with a range of external data sources; and incorporation of a variety of data types from current and future sources, including industry, with varying levels of identification and aggregation.

To guide the working group's activities and avoid duplication of already established processes, a review of existing national, international and local data governance frameworks was undertaken. Stakeholder engagement and early piloting through use cases helped establish the needs of a range of end-users and enabled a better understanding of current processes to ensure compliance and streamlining of future processes. So far, these activities have been focussed on governing data access by members of the two partner organisations. However, the foundations of the governance framework are designed to allow future extension to data access by other external researchers and data users such as researchers from other academic institutions or industry and other data types e.g. linkages with administrative data or personal device data.

Implementing natural language processing

Electronic Health Records include many forms of narrative text, ranging from short progress notes to much longer discharge summaries and specialist reports. These documents are *unstructured* data that cannot be used directly for research or operational purposes unless they are reviewed and coded by, for example, Health Information Managers, into classification systems such as the International Classification of Diseases (ICD) codes. However, there are limitations to these codes for use in research whereby some conditions are not well coded and have been shown to have low specificity when validated against clinical cohorts [12]. Narrative text may be manually coded to answer specific research

Figure 2: Application of the Five Safes framework


People
Researcher experience assessed
Projects
Approval by NCHA committee and Data owner (Peninsula Health)
Settings
Secure e-research environment (SeRP)
Data
Deidentified release
Separation principle
Output
Reporting guidelines
Pre-publication review

NCHA: National Centre for Healthy Ageing, SeRP: Secure e-research Platform.

questions, but the scale is typically small, and the cost is significant.

Natural Language Processing (NLP) is a family of computational tools for automated analysis of narrative or unstructured text. Having an NLP framework will allow researchers to access information that may not be otherwise accessible, especially information that is either poorly coded in standard datasets or not coded at all.

In response we have deployed the CogStack framework [13, 14] within the Healthy Ageing Data Platform. CogStack is an open-source project developed by Kings College, London that includes modules for extraction of documents from source systems, conversion formatted documents (such as rich text format) to standard text and creation of searchable document indexes. These modules provide an operational environment that supports large scale, such as an entire health organisation, processing of documents. The core NLP component provided by CogStack is MedCAT, although other approaches can be developed and installed in the CogStack framework. MedCAT provides an annotation service which links tokens in text to concepts in an ontology such as the Unified Medical Language System (UMLS) or SNOMED-CT. The concepts are useful for subsequent analysis via statistical classifiers or machine learning algorithms.

Consumer engagement

Due to the focus of our Research Data Warehouse on data from a defined geographic region, we identified that a critical component to our success would be early and meaningful engagement with the local community. To guide our approach, facilitated workshops were conducted with consumer representatives recruited through the Peninsula Health Consumer Advisory Groups. The aim of the workshops was to co-design with consumers, the consumer engagement strategy for the NCHA Healthy Ageing Data Platform. Workshops were structured to facilitate discussion about the risks and benefits of using EHR data for research and ways that consumers could be involved in the use of their data. Participants were asked to brainstorm how best to communicate the Healthy Ageing Data Platform to the broader local community taking into consideration identified risks and benefits, language used and target groups. Information from these workshops will be used to guide our overarching consumer engagement strategy.

Relevant to consumer involvement is the incorporation of a related program of work, implementing routine collection of Patient Reported Outcome Measures (PROMS) into Peninsula Health for incorporation into the Healthy Ageing Data Platform [15]. This activity will ensure that data obtained from consumer, and not just clinician perspectives are available within the Platform.

Results

Approvals for 9 out of the 11 identified source systems have been obtained from the different dataset owners for access to research relevant datasets held within the Peninsula Health network environment. Core data for >800,000 patients over a 10-year period have been curated within the Healthy Ageing Data Platform's research data warehouse. Migration of these data into the warehouse to allow automated data extraction, is in progress. So far, 129 core data items have been through the Delphi process with 23 individuals participating in one or more of the workshops (13 clinicians from 11 areas, 10 researchers from 8 departments). Of the 129 items polled, 113 items reached consensus after the first round of polling and 117 reached consensus during the Delphi process and have been included in the core dataset (see Table 1 for examples). Examples of items that did not reach consensus include: Religion (Demographic dataset), Previous drug Overdose (Inpatient dataset), and Medication Order Status (Pharmacy dataset). A data dictionary, based on the National Health Data Dictionary (NHDD) is under development to catalogue the data attributes of these core data items, providing detailed meta-data for researchers wanting to use data from the Healthy Ageing Data Platform.

The Cogstack NLP pipeline has been implemented within the Platform, deployed on a Linux server and configured to retrieve documents from the staging are of the data warehouse via SQL queries. So far it has been used to identify dementia cases through extraction of dementia related concepts from clinical notes. Future uses may include the extraction of concepts not available in the structured data e.g. social concepts, and improving the quality of core data items for diagnoses that are known to be poorly recorded in the structured data, for incorporation into the research data warehouse.

Data access governance frameworks and policies from government held administrative datasets, clinical quality registries and large clinical trials ($n = 7$) were reviewed by the data access working group (Table 2). Following the review, a template based on existing stakeholder collaborations and institutional agreements between the two NCHA partners was chosen to use as a basis. A streamlined approach for access to core data items by researchers from the partner organisations has been agreed upon and is currently being formalised. The committee identified that additional frameworks will be needed to accommodate future data sharing and linkage activities with industry, government and external academic institutions.

Decisions were also made with regards to governing data release. Person level data released external to Peninsula Health will be released into Monash SeRP - an e-research environment. Similar to other e-research environments SeRP ensures secure data transfer and provides an analysis environment with audit trails via a curated gateway with control over which researchers can access the data. The exception to SeRP is that local Peninsula Health projects will have the option of release of research data to a secure folder or internal secure environment within the Peninsula Health network in alignment with current practice.

Preliminary data for twelve use case projects have been released to researchers using manual extraction. Once the warehousing is complete the extraction process will be automated for the core data items, allowing scale up of data release processes. Linkages between data from the core dataset and other administrative datasets have also been undertaken for individual projects. This includes linkages with Commonwealth and State administrative datasets (e.g. hospital, Medicare pharmaceutical dispensing, aged care and

Table 1: Overview of the datasets and examples of core data items that achieved consensus during the Delphi process and are included in the Healthy Ageing Data Platform

Dataset	**Example of variables within datasets**
Demographics	month/year of birth, gender, postcode, country of birth, marital status, interpreter requirement, living arrangement, usual accommodation
Admission	admission/discharge/onset dates, diagnosis codes, discharge destination, basic vital signs, procedures/interventions, death date, admission/discharge specialty, admission care type, admission type, ICU/CCU stay duration, discharge destination, transfer reason, tobacco/alcohol misuse, aged care assessment status
Emergency Department	attendance dates, triage category, time to triage, diagnosis codes, time to departure, arrival mode, injury cause/intent, visit type, departure destination
Theatre/surgical	case specialty, primary procedure code id, operation type, surgery start date, operation outcome, ASA score, anaesthetic type, unplanned return to theatre
Outpatient	episodes dates, contact dates, episode/contact program, contact purpose, contact delivery setting, contact professional group, episode health conditions, other factors affecting health
Pharmacy	medication dates, medication order type, drug type, medication name, medication dosage, medication route, clinician specialty
Pathology	pathology order/collection dates, test name, reason for test, specimen type, collection priority, test result, result units/limits
Costing	encounter type, cost buckets: allied health, emergency, ward medical, ward nursing, ward supplies, critical care unit, operating room, imaging, pathology, pharmacy, prosthesis, and non-clinical
Mental Health	Episode/service contact dates, diagnosis codes, program, phase of care, triage outcome
Radiology	In process
Community Health	In Process

ICU: Intensive Care Unit, CCU: Coronary Care Unit, ASA: American Society of Anesthesiologists.

Table 2: Data access governance frameworks and policies used to inform the National Centre for Healthy Ageing data access policy

Health data sharing and linkage centres
Australian Institute of Health and Welfare's data governance framework
Rochester Epidemiology Project, Mayo Clinic's Department of Quantitative Health Sciences
Monash Partners Data Sharing Agreement and Principles
International Organisation for Standardization and the International Electrotechnical Commission; Application of ISO/IEC 38500 to the governance of data; guiding principles ISO/IEC 385051:2017
Victorian Department of Health Datasets; Centre for Evaluation and Research external research application process
National clinical registries
Australian Stroke Clinical Registry
Clinical Trials
ASPirin in Reducing Events in the Elderly (ASPREE) Clinical Trial; Working with ASPREE, Version 1.2: 31 July 2018

ISO: International Organization for Standardization, IEC: International Electrotechnical Commission.

death data) for all people living aged $\geq$60 years in the region and linkages with local primary care health record data through Outcome Health, a not for profit organisation that that acts as a data custodian for a large number of Primary Care practices in the region.

Based on findings from our two consumer engagement workshops (workshop 1=7 participants; workshop 2=9 participants) the community engagement strategy for the Healthy Ageing Data Platform will incorporate the following: (i) understanding who the target community is; for example, age, culture, and health/digital literacy, (ii) conveying information that is simple, clear and devoid of academic and/or technical jargon, (iii) respect of consumers' genuine fear of who can access their data and concerns around whether their data is identifiable, and (iv) transparency in how data are being used and managed.

Discussion

In this study we provide a detailed account of the multifaceted approach taken to develop the NCHA Healthy Ageing Data Platform utilising EHR data across multiple sectors of a healthcare organisation within a geographic region. Our progress to date has demonstrated early evidence of feasibility of our approach, providing a solid foundation for future expansion of this work.

Although a number of publicly accessible EHR derived datasets exist, they are generally restricted to a single condition, disease or specialty type [16, 17] a single unit or hospital. Other models have involved the bringing together of state or national government held administrative data through a central agency or databank to establish bespoke project specific research cohorts [18, 19] or to support large prospectively established national cohorts or data banks [20]. Governance frameworks differ between these programs with different consent and agreements required. However, similar to us, most are working to streamline access to maximise the use of their data for research purposes with most programs using some form of e-research environment or data enclave for data access and analysis [17, 19–21]. The Rochester Epidemiology Project (REP) is the model most similar to ours. The REP, associated with the Mayo Clinic in the United States, is the longest standing example of a population medical record linkage system, with data dating back over 50 years [21]. REP data became available for electronic linkage and storage from 2010 onwards allowing expansion and streamlining of the system [22]. A more mature system than ours, the REP has supported over 2,000 peer-reviewed publications covering a diverse range of clinical conditions and populations.

Our approach differs from these examples in our creation of a core set of data items that are highly curated inside our research data warehouse prior to release to researchers. This not only to maximises data quality and completeness but saves time and enhances consistency between researcher projects. Weekly updates of our warehouse ensure that our data are contemporary, making it ideal for healthcare evaluation and quality improvement. Our core also provides a standardised set of data that can be linked to other data sources such as government held administrative datasets. Our routine collection of PROMS across a whole healthcare organisation for incorporation into our data platform is also novel.

Despite the many strengths of our approach there are a number of weaknesses and challenges that should be acknowledged. Our core dataset is limited to residents that attend services within Peninsula Health. As Australian residents are covered by universal healthcare they are entitled to seek care at any public hospital and those with private health insurance may attend private hospitals - though in our region this is predominantly for elective surgery. To better understand this gap, we have recently linked our core dataset with government held administrative data, providing information on all hospital admissions, emergency department presentations and outpatient visits for all public and private hospitals in Victoria. Although our region has socio-economic diversity, and both metropolitan and regional areas, we have limited cultural diversity and so our data may not be suitable for answering questions related to specific cultural groups. There are also a number or data items that may be important for research but are not part of the core dataset due to poor quality and/or completeness. To address this gap, we plan to work with clinicians and the Peninsula Health, health information management team to maximise data quality at the recording interface. Future NLP activities will also focus on improving data quality. Finally, core data items are limited to those routinely collected in clinical practice, highlighting the additional benefit of using our Healthy Ageing Data Platform to support prospectively recruited sub-cohorts involving collection of specific research relevant data.

Conclusion

We have provided an overview of the methods and activities undertaken to establish a curated research data platform utilising EHR data from multiple systems within a comprehensive healthcare organisation. Having an established EHR research dataset will improve the feasibility of data linkage and potential for future expansion of linkages within the NCHA. Once fully established the Healthy Ageing Data Platform will provide a national asset and international exemplar for better understanding the complexity of ageing and the role that healthcare access and utilisation plays, within a defined geographic cohort.

Acknowledgments

We wish to acknowledge the members of the National Centre for Healthy Ageing Data Platform Working Group and Data Access Working Group. The National Centre for Healthy Ageing acknowledges the financial support of the Australian Government Department of Health to support the Centre establishment.

Statement of conflicts of interest

None declared.

Ethics statement

Ethics approval for this work was obtained from the Peninsula Health Human Research Ethics Committee (HREC) [Project number 82353, Approval numbers QA/82353/PH-2022-296956(v1); LNR/66113/PH-2020; LNR/78268/PH-2021-274394; LNR/73731/PH-2021; AM/73610/PH-2021-281834(v1); QA/69403/PH-2020-237290(v2). Workshops attendees were provided with a Participant Information Sheet and informed consent was implied by attendance at the workshop.

References

1. World Health Organisation. World report on ageing and health. Geneva, Switzerland 2015.

2. Atasoy H, Greenwood B, McCullough J. The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. Annual Review of Public Health. 2019;40:487–500. https://doi.org/10.1146/annurev-publhealth-040218-044206

3. Hemingway H, Asselbergs F, Danesh J, Dobson R, Maniadakis N, Maggioni A, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. European Heart Journal. 2018;39:1481–95. https://doi.org/10.1093/eurheartj/ehx487

4. Savitz S, Savitz L, Fleming N, Shah N, Go A. How much can we trust electronic health record data? Healthcare. 2020;8:100444. https://doi.org/10.1016/j.hjdsi.2020.100444

5. Department of Health. Victorian Population Health Survey 2020 2022 [Available from: https://www.health.vic.gov.au/population-health-systems/victorian-population-health-survey-2020#.

6. Abraham R, Schneider J, vom Brocke J. Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management. 2019;49:424–38. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

7. World Health Organisation. Discrete desk review of data quality 2020 [Available from: https://cdn.who.int/media/docs/default-source/data-quality-pages/2021_-dqa_module-2_desk-review-of-data-quality.pdf?sfvrsn=7a0999e_9.

8. Bruland P, McGilchrist M, Zapletal E, Acosta D, Proeve J, Askin S, et al. Common data elements for secondary use of electronic health record data for clinical trial execution and serious adverse event reporting. BMC Med Res Methodol. 2016;16(1):159. https://doi.org/10.1093/eurheartj/ehx487

9. Dalkey N. An experimental study of group opinion: The Delphi method. Futures. 1969;1(5):408–26. https://doi.org/10.1016/S0016-3287(69)80025-X

10. Delbecq A, Van de Ven A, Gustafson D. Group Techniques for Program Planning; a guide to nominal group and Delphi processes. Group and Organization Studies. 1975;1(2):256–. https://doi.org/10.1177/105960117600100220

11. Ritchie F. The ’Five Safes’: a framework for planning, designing and evaluating data access solutions. Data for Policy 2017; London2017.

12. Kilkenny M, Dalli L, Sanders A, Olaiya M, Kim J, Ung D, et al. Comparison of comorbidities of stroke collected in administrative data, surveys, clinical trials and cohort studies. Health Information Management Journal. In press;Epub ahead of print. https://doi.org/10.1177/18333583221124371

13. Jackson Rea. CogStack – experiences of deploying integrated information retrieval and extraction services in a large National Health Service Foundation Trust hospital. BMC Medical Informatics and Decision Making. 2018;18(1):47. https://doi.org/10.1186/s12911-018-0623-9

14. Wu H. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. Journal of the American Medical Informatics Association. 2018;25(5):530–7. https://doi.org/10.1093/jamia/ocx160

15. Snowdon D, Srikanth V, Beare R, Marsh L, Parker E, Naude K. Landscape assessment of the use of Patient Reported Outcome Measures in research, quality improvement and clinical care across a healthcare organistaion. BMC Health Services Research. 2023;23. https://doi.org/10.1186/s12913-023-09050-1

16. Fleuren L, Dam T, Tonutti M. The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients. Critical Care. 2021;25:304. https://doi.org/10.1186/s13054-021-03733-z

17. Pfaff E, Girvin A, Gabriel D, Kostka K, Morris M, Palchuk M, et al. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. J Am Med Inform Assoc. 2022;21(4):609–18. https://doi.org/10.1093/jamia/ocab217

18. Andrew NE, Kim J, Cadilhac DA, Sundararajan V, Thrift AG, Churilov L, et al. Protocol for evaluation of enhanced models of primary care in the management of stroke and other chronic disease (PRECISE): A data linkage healthcare evaluation study. Int J Popul Data Sci. 2019;4(1):1097. https://doi.org/10.23889/ijpds.v4i1.1097

19. Lyons J, Akbari A, Agrawal U, Harper G, Azcoaga-Lorenzo A, Bailey R, et al. Protocol for the development of the Wales Multimorbidity e-Cohort (WMC): data sources and methods to construct a population-based research platform to investigate multimorbidity. BMJ Open. 2021;11(1):e047101. https://doi.org/10.1136/bmjopen-2020-047101

20. All of Us Research Program Investigators, Denny J, Rutter J, Goldstein D, Philippakis A, Smoller J, et al. The "All of Us" Research Program. N Engl J Med. 2019;38(17):668–76. https://doi.org/10.1056/nejmsr1809937

21. St Sauver J, Grossardt B, Yawn B, Melton Lr, Pankratz J, Brue S, et al. Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system. Int J Epidemiol. 2012;41:1614–24. https://doi.org/10.1093/ije/dys195

22. Rocca W, Grossardt B, Brue S, Bock-Goodner C, Chamberlain A, Wilson P, et al. Data Resource

Profile: Expansion of the Rochester Epidemiology Project medical records-linkage system (E-REP). Int J Epidemiol. 2018;47(2):368. https://doi.org/10.1093/ije/dyx268

Abbreviations

DQR:	Data Quality Review
EHR:	Electronic Health Record
IT:	Information Technology
NCHA:	National Centre for Healthy Ageing
NLP:	Natural Language Processing
SeRP:	Secure eResearch Platform
UR:	Unit record