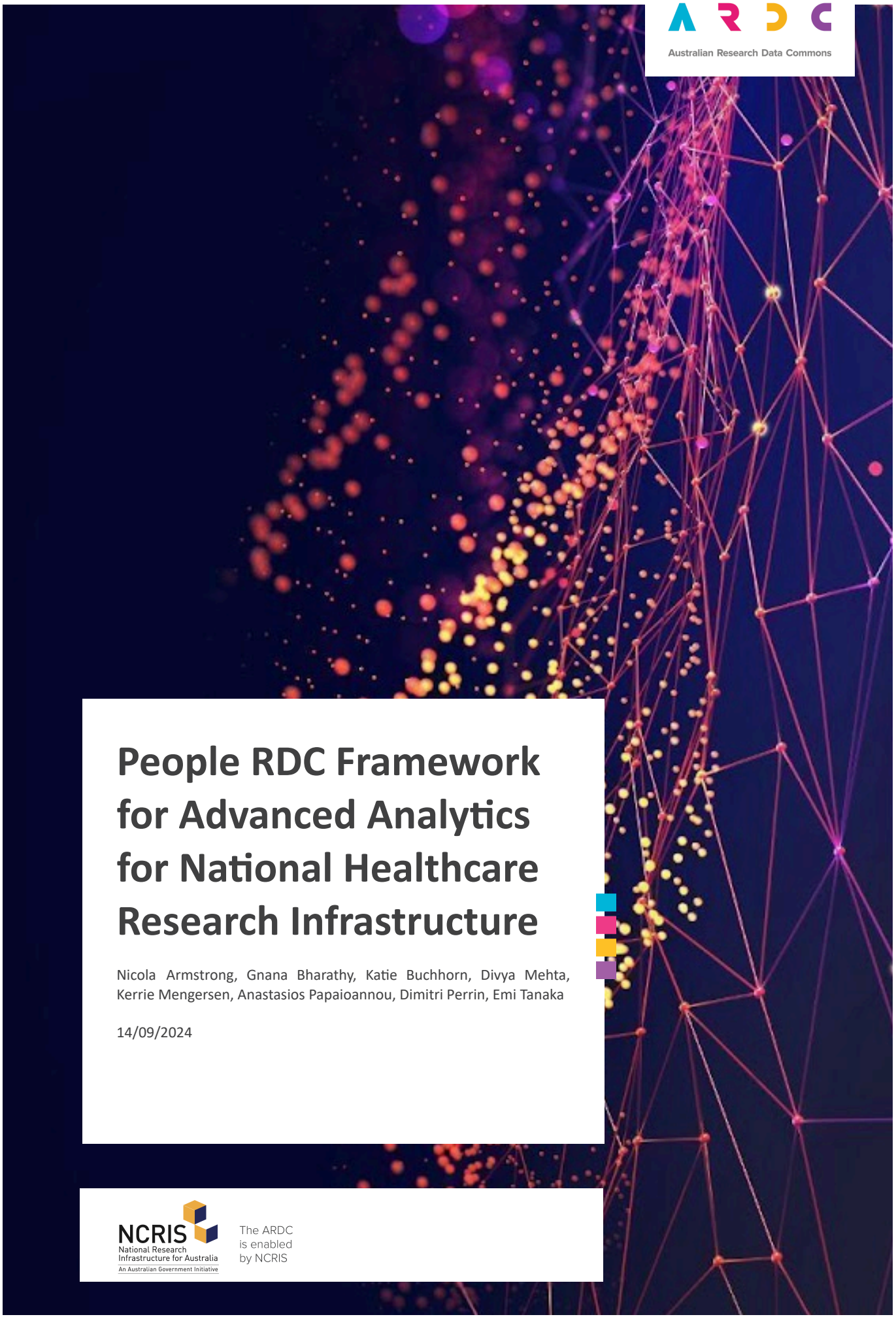


Australian Research Data Commons

People RDC Framework for Advanced Analytics for National Healthcare Research Infrastructure

Nicola Armstrong, Gnana Bharathy, Katie Buchhorn, Divya Mehta, Kerrie Mengersen, Anastasios Papaioannou, Dimitri Perrin, Emi Tanaka

14/09/2024

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

The ARDC is enabled by NCRIS

Cite as:

Armstrong NJ, Bharathy GK, Buchhorn K, Mehta D, Mengersen K, Papaioannou A, Perrin D, Tanaka E, (2024). People RDC Framework for Advanced Analytics for National Healthcare Research Infrastructure. Australian Data Science Network (ADSN). Report to ARDC, Australian Research Data Commons. DOI: 10.5281/zenodo.13831386

Acknowledgement of Country

We acknowledge the traditional custodians throughout Australia and their continuing connection to, and deep knowledge of, the land and waters. We pay our respects to Elders both past and present.

Working Group

Name	Position and Affiliation	ORCID
Prof Nicola J Armstrong	Professor, Curtin University	0000-0002-4477-293X
Dr Gnana K Bharathy	Research Data Specialist (AI/ML & Architecture), Australian Research Data Commons (ARDC) and UTS	0000-0001-8384-9509
Dr Katie Buchhorn	Research Assistant, Queensland University of Technology	
Prof Divya Mehta	Professor, Queensland University of Technology	0000-0002-7359-348X
Dr Anastasios Papaioannou	Data Science Manager, Intersect Australia	0000-0002-8959-4559
Dr Dimitri Perrin	Associate Professor, Queensland University of Technology	0000-0002-4007-5256
Dr Emi Tanaka	Deputy Director, Australian National University	0000-0002-1455-259X

Contributions

AUTHORS: Armstrong NJ, Bharathy G, Buchhorn K, Mehta D, Mengersen K, Papaioannou A, Perrin D, Tanaka E

REVIEWERS: Ward R, Burton A, Bharathy G, Uddin A

WORKSHOP: Armstrong N, Bharathy G, Buchhorn K, Burton A, Channon J, Ferrers R, Liu M, Macuga T, Mehta D, Perrin D, Uddin A, Ward R

COMMUNICATIONS & EDITING: Macuga T, Murphy P, Savill J, Staines C, Vennell A, Yuen J

NCRIS National Research Infrastructure for Australia An Australian Government Initiative
ARDC is enabled by NCRIS

Table of Contents

NCRIS National Research Infrastructure for Australia An Australian Government Initiative
ARDC is enabled by NCRIS

Executive Summary and Recommendations

This document reports the findings of a program of work undertaken to develop a national infrastructure framework for the support of advanced analytics for healthcare research.

The Australian Research Data Commons (ARDC), in collaboration with the Australian Data Science Network (ADSN), consulted with stakeholders to identify the needs, aspirations and challenges of advanced analytics in healthcare.

Based on the synthesis of findings, the Framework for Advanced Analytics National Infrastructure in Healthcare Research was developed, and the following key recommendations were proposed to address identified gaps and needs in Australia's advanced health analytics infrastructure. The recommendations are aimed at creating an Australian Harmonized Approach (AHA) to health data analytics research infrastructure:

1. Enhance Computational Resources and Data Environments:
    a. Increase investment in scalable, secure, high-performance computing resources, including on-demand access to GPUs and TPUs. This includes the development of infrastructure to harness the potential of Generative AI and analytics workloads to meet the dynamic needs of researchers.
    b. Develop and support Trusted Research Environments (TREs) that provide secure, controlled access to sensitive health data.
    c. Establish national federated learning infrastructure to facilitate secure analysis of decentralised datasets while maintaining data privacy.
2. Standardise Data Governance and Curation:
    a. Implement national standards for health data curation, metadata, and access protocols to improve data quality and usability. This includes developing searchable portals for health datasets with comprehensive documentation and metadata.
    b. Establish a national data governance framework to harmonize data privacy regulations and standards across jurisdictions, particularly around key initiatives such as TREs and federated learning. This framework should facilitate data sharing, linkage, and analysis while ensuring compliance with privacy and security requirements.
    c. Develop a centralised ethics approval framework to streamline the process for research involving sensitive health data.
    d. Develop tools for synthetic data generation in appropriate contexts to address data privacy concerns and enhance data accessibility, ensuring that synthetic data accurately reflects real-world complexities and is suitable for various research purposes.

3. Promote Collaborative and Ethical Research Initiatives:
    a. Foster collaborative research initiatives that bring together academia, industry, and government, promoting international collaborations.
    b. Create a national community of practice to support the implementation of advanced analytics solutions in healthcare, incentivizing researchers based on implementation success.
    c. Establish a national data ethics committee to promote consistent understanding and acceptance of AI technologies across research institutions.
    d. Adopt risk-based data management approaches that balance privacy concerns with data accessibility needs.
4. Support Workforce Development and Practical Implementation:
    a. Invest in education and training programs to build a skilled research workforce capable of utilizing advanced analytics techniques, including data science, advanced statistical methods, and artificial intelligence. This should include specialised training in the use of Generative AI models in health research and responsible AI/ML practice guidelines.
    b. Create a centralised methodological hub with resources specific to the local context to support the translation of research findings into clinical practice.
    c. Establish support systems to assist researchers in using infrastructure technology and navigating data privacy issues.

1. Introduction

In 2023, the Australian Research Data Commons (ARDC) approached the Australian Data Science Network (ADSN) to discuss community needs and aspirations in the context of Healthcare Analytics. The resulting collaboration resulted in the development of an infrastructure framework to support advanced analytics for healthcare research in Australia.

The resultant Framework for Advanced Analytics National Infrastructure aims to create a comprehensive and coordinated approach to health analytics infrastructure, leveraging national and international collaborations to enhance research capabilities, data governance, and technological advancements.

The results and recommendations arising from this project will assist the ARDC in meeting its aims of contributing significantly to the advancement of health research and improvement of healthcare outcomes. Some of the recommendations may go beyond the purview of an infrastructural organization, such as the ARDC, to address directly, however, it could pursue these through its advocacy and influence.

1.1. Background

The ARDC has been instrumental in fostering a robust research infrastructure to support advanced analytics in health data. Recognising the rapid advancements in artificial intelligence (AI) and machine learning (ML), the ARDC initiated the development of a comprehensive framework to enhance the national research infrastructure. This initiative aims to leverage both national and international collaborations to improve research capabilities, data governance, and technological advancements.

The Framework for Advanced Analytics National Infrastructure emerges from the need to address significant gaps in the current health data ecosystem. As the volume and complexity of health data continue to grow, so does the demand for sophisticated tools and platforms that can effectively process, analyse, and interpret these data. The framework aims to identify and prioritise the most valuable components that ARDC should focus on to support advanced health analytics, ensuring that Australia remains at the forefront of health research and innovation.

This initiative is driven by the understanding that a coordinated and well-supported infrastructure is crucial for translating research into tangible healthcare improvements. The framework will guide the development of services and partnership programs, ensuring that researchers have the necessary tools and support to conduct high-impact health research.

1.2. Aims, Scope and Outcomes

1.2.1. Aims & Objectives

The objective is to develop an advanced analytics infrastructure framework, or specifically an Advanced Analytics Implementation Specification for projects conducted by ARDC People Research Data Commons

(People RDC). The framework seeks to respond to the question: “What are the highest value components that the ARDC should prioritise as national research infrastructure to support advanced analytics of health data?”

The purpose of the specification is to:

1. Outline how ARDC will support advanced health analytics with infrastructure;
2. Inform subsequent service development and partnership programs;
3. Ensure researcher perspective input into the program rationale and logic; and
4. Provide opportunities for the research community to work with ARDC in the People RDC journey.

1.2.2. Rationale

The framework serves as a specification for the full-fledged co-investment projects and will:

1. Establish a rapport with the community, fostering support for the program's core philosophy and logic;
2. Encompass national research infrastructure, incorporating pivotal aspects like AI/ML agent based models, systems dynamics, discrete event simulations, advanced statistical models, and indispensable decision support tools suitable for supporting health and health translation;
3. Incorporate ARDC and other key national infrastructure assets, where relevant.

1.2.3. Scope

Stakeholder consultations were conducted around Australian Data Science Network (ADSN)[1] affiliated institutes and other contacts provided by ARDC or ADSN. These different groups were invited together for consultations. The scope of the consultations covered, but were not limited to, the following infrastructure assets:

1. Underpinning Infrastructure: Software (e.g. Nectar[2], MLeRP[3]) and hardware (e.g. graphical processing unit and data storage)
2. National Reference Data Assets: Data curation, vocabularies and analytic reference datasets, synthetic data, Research Data Australia, Research Vocabularies Australia, FAIR model for AI reference data and ML models (Wilkinson, 2016)
3. Socio-Technical Assets for National-Level Coordination or Facilitation:

[1] https://www.australiandatascience.net/
[2] https://ardc.edu.au/services/ardc-nectar-research-cloud/
[3] https://docs.mlerp.cloud.edu.au/

   a. Culture and policy
   b. Communities of practice, skills, capability, or awareness
   c. Training and capacity development
   d. Guidelines for risk management, practice, governance, privacy, responsible AI etc.
4. Tools, Environment Platforms Reference Programs: Library of tools or collaborative infrastructure (analytics tools, including virtual labs, models and codeless environments, hubs, foundational models, other models)

Notes:

- A pathfinder project (targeted proof of concept/ demonstrator) was carried out in parallel to explore and implement specific Federated Learning, as implementation challenges also needed to be explored and a Federated Learning network needed to be built. Both projects were soft aligned through workshops where both working groups participated.
- As an agenda for the national infrastructure capability, the report does not deal with any direct research agendas.
- In the context of advanced analytics, we use the term “AI/ML” to represent the field of artificial intelligence, machine learning and advanced mathematical and statistical methods.

1.3. Approach

The Framework for Advanced Analytics National Infrastructure was designed to capture a wide range of perspectives and insights from the Australian health research community. To develop a comprehensive and effective framework for advanced health analytics infrastructure, a structured and collaborative approach was adopted.

The process began with the formulation of the problem and the establishment of a working group, comprising key members from the Australian Data Science Network (ADSN) and a working representative from the Australian Research Data Commons (ARDC). The working group also obtained support from ARDC experts in facilitating and collating information from public workshops.

This initial phase culminated in a presentation at the 2023 ADSN conference, where the first workshop was conducted to identify high-level issues and set the stage for further investigation.

Following this, the team conducted an extensive environmental scan to understand the current landscape of advanced analytics in healthcare, identifying key gaps and opportunities. This scan provided a solid foundation for the subsequent stages of the project.

The next phase involved a multi-pronged approach to knowledge elicitation, combining surveys, workshops, and interviews to gather insights from a diverse group of stakeholders and experts. Key stakeholders within ADSN were identified, along with representative stakeholders from outside the network.

A survey was designed, pilot tested and deployed to inform an understanding of challenges and issues around health data and advanced analytics within the broader Australian scientific community. The survey questions were grouped into five main themes: data privacy and ethics, version control, data integrity, compute resources, and advanced analytics. The survey was disseminated through Australian Society of Science, the Statistical Society of Australia, Machine Learning Community of Practice for Australia, Australian Bioinformatics and Computational Biology Society, and Data Management Association Australia as well as some Australian federal and state government organizations.

The workshops and focus groups were particularly instrumental in eliciting detailed requirements and fostering discussions on critical issues.

Both the survey and three of the four workshops were openly promoted to ensure a broad and inclusive perspective on the challenges and needs in health analytics infrastructure. The workshops were supplemented by an offer of individual interviews for those unable to attend the group sessions.

The information gathered through these mixed methods was then collated and analyzed, providing a comprehensive picture of the current challenges in advanced analytics in healthcare and potential recommendations for developing the research infrastructure. This analysis formed the basis for the development of a documented framework, which synthesised the findings into key themes, providing guidance and examples for each, along with recommendations and limitations.

The final step in this process involved the validation of the draft framework through exposure to stakeholder and subject matter expert networks, ensuring its robustness and applicability to the Australian health research context.

The process involved the following key steps:

1. Formulation of the Problem and Working Group
    a. The key working group members within ADSN, as well as a representative from ARDC, were identified.
    b. The problem was presented at the 2023 ADSN conference, and the first workshop was conducted to identify key issues at a high level.
2. Carrying out the Environmental Scan
    a. Environmental scans were carried out to understand the landscape.
    b. Key gaps were identified.
3. Knowledge Elicitation from broader Research Community via Survey
    a. Survey was designed, pilot tested and disseminated to broader stakeholder communities for a 2-month period
    b. Survey responses were analyzed and used to identify the key challenges
4. Knowledge Elicitation from Stakeholders and Expert Groups via Workshops

a. Key stakeholders within ADSN were identified, as well as a representative sample outside ADSN in collaboration with ARDC.
b. Consultations with key stakeholders were conducted through workshops and focus groups to elicit key needs and requirements. This was supplemented by individual interviews as needed.

5. Development of a Documented Framework for Advanced Analytics Projects
    a. The findings were consolidated to document the key infrastructure needs for health analytics and existing and reusable resources.
    b. Evidence was synthesised, organised into themes, and guidance and examples have been provided around each theme with recommendations and limitations.
6. Validation of Framework with Subject Matter Expert and Stakeholder Communities
    a. After developing the framework, a draft will be made available to stakeholder and subject matter expert (SME) networks such as ADSN, ensuring the framework stands up to scrutiny.

The overall approach to framework development is summarised in Figure 1.1.

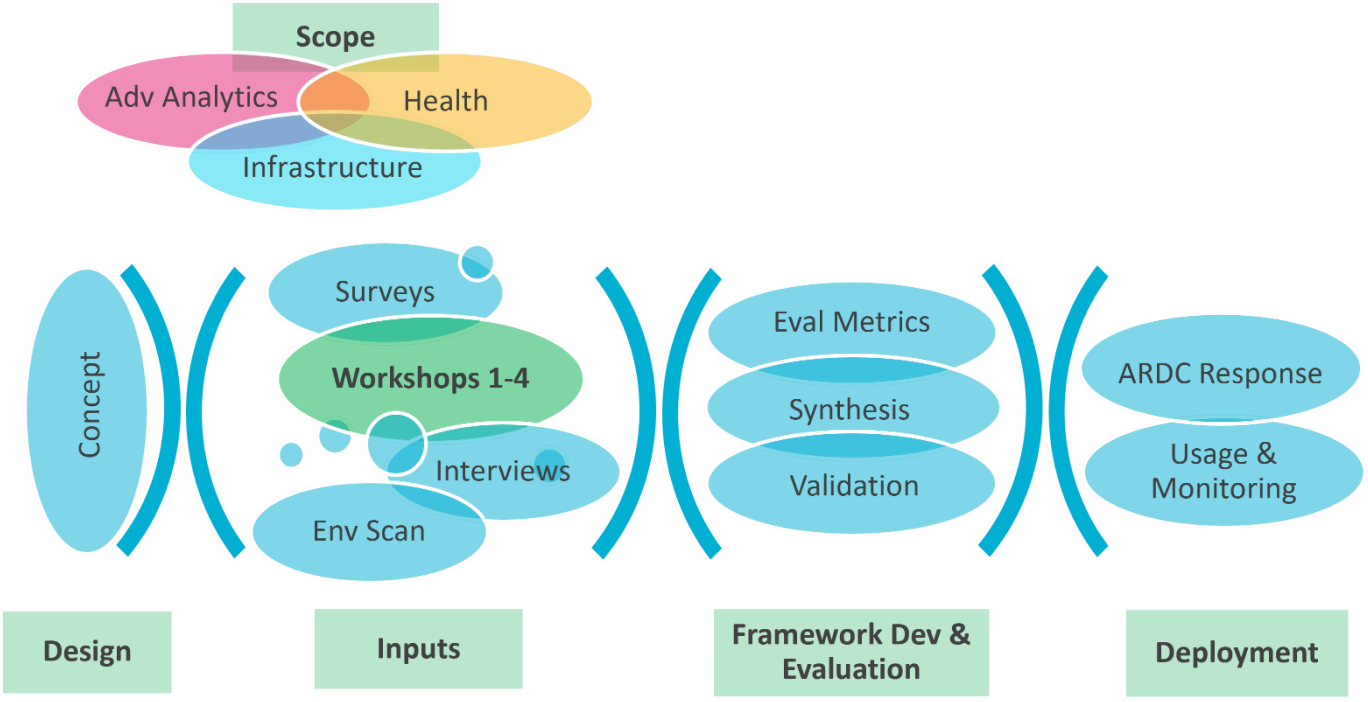


Figure 1.1: Project Methodology

2. Environmental Scan

An environmental scan was conducted to determine the current landscape of advanced analytics in healthcare, nationally and internationally. It aimed to assess healthcare analytics infrastructure by identifying significant gaps and assets, including infrastructure, tools, platforms, socio-technical systems, and national reference data assets.

This approach mapped the existing landscape in a structured manner. Data was gathered from academic literature, policy reports, and technological assessments, as well as real-world data from national initiatives and international benchmarks. The scan systematically reviewed platform capabilities, technological trends, and existing national infrastructures to identify tools, frameworks, and systems that support advanced analytics in healthcare. It also incorporated information from national initiatives and international benchmarks (e.g., the European Open Science Cloud, NIH's All of Us Program, and BioData Catalyst).

A comparative analysis of identified tools, platforms, and infrastructures was conducted to identify strengths and limitations, providing insights into how they can address challenges related to data privacy, integration, and governance. The environmental scan synthesized the data into themes and patterns, identifying gaps in capabilities, infrastructure, and governance, providing a structured overview of current challenges, emerging trends, and future opportunities.

The methodology employed closely mirrored a literature review process but extended it to include both **technological and socio-technical assessments** of the current state of selected assets.

The findings of the environmental scan are summarised in Table 2.1 and presented in more detail in the following sections.

Table 2.1 Summary of the findings of the environmental scan

Underpinning Infrastructure	Countries are investing significantly in national research infrastructures to support advanced analytics of health data. Notable examples include the National Institutes of Health's All of Us Research Program and the National Science Foundation's Extreme Science and Engineering Discovery Environment in the U.S., the European Open Science Cloud and ELIXIR in the EU, UK Research and Innovation and Jisc in the UK, the Digital Research Alliance in Canada, China's supercomputing centers and National Genebank, and Singapore's Agency for Science, Technology and Research. The private sector is increasingly playing a crucial role in research infrastructure by offering advanced computing resources, data analytics platforms, and collaborative tools. The Kaggle Data Science & ML Surveys highlight the popularity of Amazon Web Services as a cloud computing

	platform, Amazon S3 for data storage, Amazon SageMaker for ML, and Google Cloud AutoML, while also emphasising the dominance of Python and SQL, the widespread use of VSCode, and the growing traction of specialised hardware like Tensor Processing Units (TPUs) among data scientists. The Netherlands eScience Center's Technology (2024) forecasts in AI include hardware specialisation and exotic architectures like quantum computing potentially leading to a shortage of experts, the increasing need for portable libraries to avoid HPC vendor lock-in, and the rise of augmented and edge analytics for accessibility and real-time insights.
National Reference Data Assets	Synthetic data generation is a transformative tool that enhances data privacy and accessibility in health research, enabling the development of AI models without compromising confidentiality and supporting collaborative efforts by providing high-quality artificial datasets. However, it may not capture the complexity of real-world data, requires extensive validation, can introduce bias, and requires advanced technical resources that may not be available. Federated learning offers data privacy benefits by allowing models to be trained on decentralised datasets without sharing sensitive information. However, this approach presents challenges in data ethics, accuracy, and bias, requiring improved interoperability and collaborative governance efforts. The FAIR (findability, accessibility, interoperability, and reusability) principles are crucial for reproducibility, but their application to code, software, workflows, documentation, methodology, and technology must be clarified. Achieving reproducible research requires a cultural change, supported by tools, training, incentives, policies, as well as national networks, coordination efforts, and university consortia (Knowledge Exchange).
Tools & Environment Platforms Reference Programs	Robust infrastructure is essential for AI and analytics, providing the computational power, data storage, and connectivity needed for complex algorithms and large datasets, impacting healthcare by enabling efficient data processing, model training, and real-time insights. In Australia, national initiatives like the National Computational Infrastructure (NCI), Pawsey Supercomputing Centre, and ARDC Nectar Research Cloud support large-scale AI projects across various sectors, driving innovation in areas such as genomic research, climate modeling, personalised medicine, and smart city initiatives. AI in medical imaging, drug discovery, predictive analytics, personalised medicine, and clinical decision support systems revolutionise diagnostics,

	patient care, and treatment development, significantly improving efficiency, accuracy, and outcomes in health and medical research. Generative AI and foundation models, such as GPT-4 and BERT, enhance health research by analysing medical literature, generating clinical notes, and aiding in patient communication, while also supporting predictive analytics and personalised treatment plans through their ability to process diverse data types.
Socio-Technical Assets for National-Level Coordination	*Skills & workforce:* Researchers in statistical agencies, government, academia, and beyond increasingly depend on the professional skills of the Research computing and data (RCD) workforce to facilitate the use of vast and ever-evolving technical resources (Schmitz, 2021). However, challenges are recognised in recruiting and retaining RCD professionals due to competition with the private sector and a lack of standardised job titles and career paths (Chaudhry, 2022). *Culture & Policy:* Many organisations handle digital transformation investments ad-hoc rather than strategically, with research showing one-third of healthcare providers lacking a long-term strategy; organisations should align on a vision, identify service gaps, and integrate transformation efforts into comprehensive master plans (Srivastava, 2024). *Guidelines for Risk Management:* Any data release, whether blended or non-blended, introduces disclosure risks that must be balanced with data usefulness. Informed consent issues are amplified in blended data. New research and development efforts are needed to enhance privacy-protecting technologies (National Academies of Sciences, Engineering, and Medicine, 2024a). *Governance:* The Organisation for Economic Co-operation and Development (OECD)[4] Council recommendations for data governance have focused on data sharing, accessibility, quality, interoperability, and privacy protections, with an ongoing focus to enhance cybersecurity and global health data interoperability. *Training & Capacity Development:* Exemplary platforms include BioData Catalyst[5], with extensive support resources such as documentation, video tutorials, and community forums to enhance researchers' data integration and analysis skills. Similarly, Alteryx offers a community hub with micro-certifications, digital learning opportunities, and interactive resources

[4] https://www.oecd.org/
[5] https://biodatacatalyst.nhlbi.nih.gov/

	such as video tutorials and e-books for various skill levels. However, the National Research Infrastructure Roadmap (Australian Government, 2021) notes “Rapid advances in computing techniques and analysis, and management of large and complex datasets, have resulted in researchers no longer having sufficient expertise in data management, computational and analysis techniques.” Therefore, there is a knowledge gap and an increasing need for upskilling the research workforce in data science and advanced analytics skills.
Existing platforms	Existing platforms (both commercial and public) fall short in key areas such as advanced machine learning capabilities, no-code solutions, collaboration tools, interoperability, security, and comprehensive support, data cataloguing and data integration.

2.1. Technological Landscape

The adoption of digital technologies like online self-scheduling, telehealth, and wearables have transformed patient experiences. At the same time, advancements in predictive analytics, internet-connected medical devices, and artificial intelligence offer the potential for increased efficiency through more integrated and personalised treatment. The recent pandemic accelerated investments in digital health technologies as organisations prioritised spending to address immediate needs in a digital transformation (Knawy, 2022; Gunasekeran, 2021). The market for digital health technologies has been growing at approximately 18% annually, driven by expenditures on digital health-related software and hardware, as well as the increasing demand for mobile services from patients. According to Precedence Research (2022), the market is projected to exceed $152 billion in North America by 2027. However, many organisations handle digital transformation investments on an ad-hoc basis rather than as part of a sustained, long-term strategy. Research by Huron Consulting found that one-third of healthcare providers lack an organisational-level strategy and are uncertain about the benefits of large-scale efforts. To achieve the best long-term results, organisations should align on a vision, identify service gaps, and integrate transformation efforts into their comprehensive plans (Srivastava, 2024).

Countries Investing in Research Infrastructure

Other countries are investing significantly in national research infrastructures to support advanced analytics of health data. These initiatives reflect a strong commitment to fostering scientific research and technological innovation across the region at a national level.

In the United States, the National Institutes of Health (NIH) supports major initiatives such as the All of Us Research Program[6], which aims to gather data from one million or more people to accelerate

[6] https://allofus.nih.gov/

research and improve health. The National Science Foundation (NSF) contributes with XSEDE (Extreme Science and Engineering Discovery Environment)[7], providing advanced digital resources and services to researchers.

In the European Union, the European Open Science Cloud (EOSC)[8] aims to provide researchers with a virtual environment for storing, managing, analysing, and reusing research data across borders and disciplines, offering services for compute, storage, and data management. ELIXIR[9], an intergovernmental organisation, brings together life science resources from across Europe, including databases, software tools, training materials, and cloud computing resources.

The United Kingdom's UK Research and Innovation (UKRI)[10] provides funding and support for research and innovation across various fields, while Jisc[11] offers digital solutions for UK education and research, including cloud services, cybersecurity, and open research resources.

In Canada, the Digital Research Alliance of Canada[12] focuses on advanced research computing, data management, and research software. China is expanding its research infrastructure with notable investments in advanced computing and genomics. The National Supercomputing Centers[13] host some of the fastest supercomputers globally. Additionally, China's National Genebank[14] supports cutting-edge genomics research. While the Agency for Science, Technology and Research (A*STAR)[15] in Singapore supports diverse research activities, including AI and data analytics.

2.1.1. Is the future commercial?

The commercial sector is increasingly becoming a significant player in the landscape of advanced analytics and research infrastructure. There are several factors and trends driving this shift:

Big Data and Cloud Computing Services

Companies like Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure offer robust cloud computing services tailored for big data analytics and machine learning. These platforms provide scalable compute power, storage solutions, and advanced analytics tools that are becoming essential for research.

[7] https://www.xsede.org/
[8] https://eosc.eu/
[9] https://elixir-europe.org/about-us
[10] https://www.ukri.org/
[11] https://www.jisc.ac.uk/
[12] https://alliancecan.ca/en
[13] https://en.wikipedia.org/wiki/Supercomputing_in_China
[14] https://db.cngb.org/
[15] https://www.a-star.edu.sg/

AI and Machine Learning Platforms

Firms such as IBM (with Watson), NVIDIA, Databricks, Snowflake, and other AI companies and startups provide advanced machine learning platforms and tools. These companies are at the forefront of developing and commercialising AI technologies that can be leveraged for research purposes.

Kaggle (Kaggle.com), a large online community of data scientists and machine learning engineers, conducted surveys in 2021 and 2022 to identify the tools and technologies favoured by data scientists. Summaries of the survey findings are shown in Table 2.2.

Table 2.2 Summaries of findings of Kaggle Data Science and Machine Learning surveys 2021 and 2022

2021 Survey
Respondents: 25,000+, focusing on 14% who are data scientists. **Cloud Computing Platforms:** Amazon Web Services (AWS) is the most popular, closely followed by Google Cloud Platform (GCP) and to a lesser extent by Microsoft Azure . **Fully Managed ML Services:** Amazon SageMaker leads; Databricks, Azure ML Studio, and Google Cloud Vertex AI notable. **Data Storage Products:** Top databases are MySQL, PostgreSQL, and Microsoft SQL Server. **Automated ML:** Google Cloud AutoML leads; Azure Automated ML; Amazon Sagemaker Autopilot; Databricks AutoML.
2022 Survey
Programming Trends: Python and SQL remain top skills; VSCode used by over 50% of data scientists. **Cloud Computing Trends:** TPUs gaining traction.
For detailed results, see Kaggle Survey 2021 and Kaggle Survey 2022.

Data Marketplaces

Commercial entities are creating data marketplaces where datasets can be bought and sold. Companies like DataRobot and Snowflake offer platforms that integrate data from various sources, providing researchers with access to vast amounts of data.

Collaborative Research Initiatives

Tech giants are increasingly partnering with academic institutions and research organisations. For example, Google's AI for Social Good[16] program and Microsoft's AI for Health[17] initiative provides

[16] https://ai.google/responsibility/social-good/
[17] https://www.microsoft.com/en-us/research/project/ai-for-health/

funding, tools, and expertise to researchers. Another example, AlphaFold[18], is an initiative between Google DeepMind and EMBL's European Bioinformatics Institute to predict protein structures that are important for research and global health (Jumper, 2021; Abramson 2024). NVIDIA also partnered with King's College to develop an open-source framework (MONAI[19]) for AI in healthcare imaging (Cardoso, 2022).

In January 2024, the U.S. National Science Foundation and collaborating agencies launched the National Artificial Intelligence Research Resource (NAIRR)[20] pilot, a first step towards a shared research infrastructure that will strengthen and democratise access to critical resources for responsible AI discovery and innovation. The pilot, partnering with 10 federal agencies and 25 private sector, nonprofit, and philanthropic organisations, will provide access to advanced computing, datasets, models, software, training, and user support to U.S.-based researchers and educators. The NAIRR pilot aims to set the foundation for a shared research infrastructure, strengthening and democratising access to critical resources needed for responsible AI discovery and innovation. The partnership highlights the pivotal role of investment in the nation's continued leadership in AI.

The NAIRR pilot aims to support AI research in healthcare, environmental sustainability, and infrastructure. It will focus on four areas: NAIRR Open, NAIRR Secure, NAIRR Software, and NAIRR Classroom. Researchers can access initial NAIRR resources through the NAIRR pilot portal, and a call for proposals will be released in Spring, 2024. The pilot aims to democratise access to AI innovation and support trustworthy AI development.

As this pilot demonstrates, the future of advanced analytics in research may be shaped by a combination of public and commercial efforts. The commercial sector's involvement brings innovation, scalability, and expertise, but it is essential to address challenges related to data privacy, ethical considerations, and access inequality. A balanced approach that leverages the strengths of both sectors can create a robust and inclusive research ecosystem.

2.1.2. Forecasting trends in AI

The Netherlands eScience Center's Technology Forecast (2024) highlights transformative trends in artificial intelligence (AI), computing, data processing and data analytics. The following trends were identified by technology experts based on survey feedback covering a wide range of expertise.

1. AI-powered innovations: AI advancements in natural language processing (NLP) and computer vision promise to revolutionise data analytics, with efficiency gains and cost reductions expected.
2. Computing challenges: addressing limitations through hardware specialisation and exotic architectures like quantum computing poses programming challenges, potentially leading to a shortage of experts.

[18] https://alphafold.ebi.ac.uk/
[19] https://monai.io/
[20] https://nairrpilot.org/about

3. Cloud solutions: cloud-based access to high-performance computing (HPC) hardware is increasing, emphasising the need for portable libraries to prevent vendor lock-in.
4. Data explosion and edge computing: improved sensors and Internet of things (IoT) devices result in a data explosion, leading to efforts to process data closer to sources for cost reduction and enhanced privacy.
5. FAIR principles: adherence to FAIR data principles is crucial for a highly distributed infrastructure, promoting findability, accessibility, interoperability and reusability.
6. Data analytics: to meet challenges in extracting value from massive amounts of unstructured data, recent developments focus on techniques like dimensionality reduction and incremental learning.
7. Augmented analytics and edge insights: augmented analytics targets accessibility for non-experts, while edge analytics ensures real-time insights closer to data sources.

2.1.3. Emerging tools

Emerging tools and technologies in the AI health and medical field are rapidly transforming the landscape of healthcare delivery, research, and patient management. Some of the most promising and innovative tools and methodologies are described below.

Synthetic Data Generation

Synthetic data generation is a transformative tool addressing critical issues of data privacy and accessibility, especially in the health and medical field. By creating artificial datasets that closely mimic real data, synthetic data generation enables researchers to develop and validate AI models without compromising confidentiality (Yoon, 2020).

Synthetic data, such as synthetic PET scans from CT images, are utilised in medical imaging to improve diagnosis and prognosis in fields like oncology (Levine, 2020). They are also used in machine learning and AI training to create robust models by providing diverse and comprehensive datasets. Synthetic data bridges the gap between different imaging modalities, enhancing diagnostic accuracy and reducing the need for multiple imaging sessions. It also augments existing datasets, providing variations that help models generalise better. Synthetic data fills data gaps, is cost-effective, enhances privacy, and improves model performance by providing a more comprehensive training set, particularly in image recognition and natural language processing tasks (Lu, 2023).

MD Anderson Cancer Center used generative models to create synthetic PET images from diagnostic CT scans, a crucial tool for lung cancer diagnosis, staging, and prognosis (Salehjahromi, 2024). Involving 1,300 lung cancer scans, the study confirmed the comparable imaging quality and tumor contrast between synthetic and actual PET scans. Radiologists verified this comparability, and the cancer radiogenomics were consistent across both modalities. This consistency highlights the potential of synthetic data to enhance diagnostic processes.

There are several tools that can generate high-quality synthetic data, preserving the statistical properties of the original datasets while ensuring privacy. This technology facilitates robust health research and analysis by providing a large amount of data for training machine learning algorithms. It also supports collaborative research efforts across institutions by enabling data sharing without the risk of exposing sensitive information. As synthetic data generation continues to advance, it holds the promise of accelerating innovation in medical and health research, promoting the development of predictive models, personalised medicine, and improved patient outcomes.

Synthetic data provides accessibility, scalability, and innovation in data-driven techniques, especially in resource-limited settings. However, it may not capture the full complexity of real-world data, requires extensive validation, can introduce bias, and requires advanced technical skills and resources that may not be available in all settings.

Federated Learning

Federated Learning (FL) is a promising approach that allows machine learning models to be trained across multiple decentralised datasets while addressing privacy concerns. However, FL is a tool that requires attention to a wide range of technical considerations that have both ethical and practical implications, such as increased risks of re-identification and responsibility diffusion among multiple partners (Bak, 2024). FL models can also suffer from accuracy issues due to data quality and potential adversarial attacks, as well as biases due to under-representation of minority populations and uneven contributions from participating entities. Given the heterogeneity of clinical practices and electronic health records systems, data curation and meaningful dataset interoperability in FL remain a challenge often requiring additional and costly steps (Crowson, 2022). To address these challenges, a European initiative (JA-InfAct) is being set up that aims to improve interoperability for FL (González-García, 2021).

Generative AI and Foundation Models

Generative AI (GenAI) and foundation models are very popular, offering transformative capabilities for health and medical research. These models, such as OpenAI's GPT-4o (Achiam, 2023), Google's BERT (Devlin, 2018), and more, leverage vast amounts of data to understand and generate human-like text, enabling advanced natural language processing tasks. In the health and medical field, generative AI can analyse medical literature, generate clinical notes, and assist in patient communication, significantly enhancing efficiency and accuracy. Foundation models are also instrumental in developing predictive analytics and personalised treatment plans by processing and interpreting complex medical and health data. Their ability to handle diverse data types, from genomic sequences to radiology images, makes them invaluable in multidisciplinary research. These models continue to evolve rapidly and can enable researchers and scientists to accelerate and advance their work on more precise diagnostics, drug discovery, and tailored treatment strategies.

Examples of GenAI reported in the literature include, but is not limited to:

1. The potential role of GenAI in optimising healthcare supply chains ethically (Ijiga, 2024), emphasising the importance of responsible implementation;
2. Transformative insights that GenAI could bring to health system management, particularly in enhancing data processing, diagnostics, and patient care (Malsia, 2024);
3. The emergence of foundation models as powerful tools in various healthcare applications (Briganti, 2023);
4. Application of GenAI in specific data management and analysis tasks such as FHIR (Fast Healthcare Interoperability Resources) and EMHR (Electronic Medical Health Records). Taxonomy and architectures of foundation models trained on non-imaging Electronic Medical Record (EMR) data, highlighting their potential use cases and training data (Wornow, 2023). Potential for GenAI such as FHIR-GPT has potential to support Fast Healthcare Interoperability Resources (FHIR) standard (Brat, 2024); and
5. Biomedical text generation and mining (Luo, 2022).

2.1.4. AI in Health Fields

AI in Medical Imaging

AI in medical imaging can revolutionise diagnostics and patient care. Advanced AI algorithms enhance the accuracy and speed of interpreting medical images, from X-rays to MRIs. These tools utilise deep learning algorithms to detect anomalies and patterns that may be imperceptible to the human eye, leading to earlier and more accurate diagnoses. AI-driven imaging solutions also streamline workflows by automating routine tasks and highlighting areas of concern, thereby improving efficiency and effectiveness. Moreover, these tools and technologies facilitate personalised treatment plans by integrating imaging data with other patient information, enabling more precise and tailored medical interventions. AI in medical imaging promises to significantly improve diagnostic capabilities, patient outcomes and the overall efficiency of health and medical research and analytics.

AI in Drug Discovery and Development

AI in drug discovery and development is a transformative element significantly accelerating the creation of new therapeutics. Advanced AI tools and platforms utilise machine learning to predict molecular interactions and identify promising drug candidates. These tools and technologies can analyse large datasets of chemical compounds and biological data, uncovering potential treatments that might be overlooked by traditional methods. AI also optimises drug design by simulating how drugs interact with targets, thereby reducing the time and cost associated with experiments. Additionally, AI-driven models can streamline clinical trials by identifying optimal patient cohorts and predicting trial outcomes. This accelerates the drug development pipeline, bringing effective treatments to market faster.

Predictive Analytics and Personalised Medicine

Predictive analytics and personalised medicine are offering unprecedented opportunities for tailored healthcare. Using AI-driven tools, researchers and scientists can analyse large amounts of patient data to identify patterns and predict future health outcomes. These tools enable the early detection of diseases, personalised treatment plans, and proactive management of chronic conditions. By integrating data from various sources, including genetic information, medical history, and lifestyle factors, predictive analytics provide a comprehensive view of a patient's health. Personalised medicine, supported by these analytics, ensures that treatments are specifically tailored to the individual, improving efficacy and reducing adverse effects.

Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) can enhance the decision-making capabilities of healthcare providers. Advanced CDSS tools leverage AI to analyse patient data and provide evidence-based recommendations. These tools integrate seamlessly with electronic health records, offering real-time insights and alerts that assist clinicians in diagnosing conditions, prescribing treatments, and managing patient care. By using machine learning algorithms to process complex datasets, CDSS can identify patterns and potential issues that may not be immediately apparent. This leads to more accurate diagnoses, improved treatment outcomes, and enhanced patient safety. Moreover, CDSS tools support personalised medicine by tailoring recommendations to the individual patient's unique health profile.

2.1.5. Infrastructure for AI

Infrastructure is the backbone of any advanced AI and analytics initiative, providing the necessary computational power, data storage, and connectivity required to support complex algorithms and very large datasets. In the realm of AI, stable, scalable and sustainable infrastructure ensures that data can be processed efficiently, models can be trained effectively, and insights can be generated in real-time. This is particularly critical in healthcare, where timely and accurate data processing can significantly impact scientific outcomes and accelerate discoveries.

In Australia, national initiatives have leveraged AI within computing infrastructure to drive innovation across various sectors. For instance, the National Computational Infrastructure (NCI)[21] and the Pawsey Supercomputing Centre[22] provide the high-performance computing resources essential for large-scale AI projects. ARDC Nectar Research Cloud[23] also provides a research-focused Cloud Computing infrastructure to researchers across Australia. These National Collaborative Research Infrastructure Strategy (NCRIS) facilities support a wide range of applications from genomic research to climate modeling, showcasing the versatility and importance of AI in different fields.

[21] https://nci.org.au/
[22] https://pawsey.org.au/
[23] https://ardc.edu.au/services/ardc-nectar-research-cloud/

Other initiatives also try to advance research in specific areas by building robust infrastructure and providing access to specialised platforms. The National Imaging Facility (NIF) in Australia provides state-of-the-art research infrastructure and cutting-edge tools to advance research in medical imaging. Utilising the Neurodesk platform[24], NIF offers researchers a comprehensive and flexible environment for neuroimaging analysis. Neurodesk integrates a wide range of neuroimaging software and tools within a cloud-based platform, enabling seamless access and efficient data processing. This platform supports advanced imaging techniques and facilitates collaborative research efforts, driving innovations in medical imaging. By leveraging Neurodesk, NIF empowers researchers to conduct high-quality, reproducible research, ultimately contributing to significant advancements in medical science and healthcare.

Another example is the E-Research Institutional Cloud Architecture (ERICA)[25] that provides essential research infrastructure and tools designed to advance research, particularly for those handling sensitive, large-scale data. ERICA offers a secure, scalable cloud environment that facilitates the storage, management, and analysis of vast datasets while ensuring compliance with stringent data protection regulations. By leveraging ERICA, researchers gain access to high-performance computing resources and advanced analytical tools, enabling them to conduct complex data-driven studies with ease and efficiency. This robust platform not only enhances data security and accessibility but also promotes collaborative research efforts across institutions, driving innovation and discovery in various scientific fields.

AI is revolutionising health and medical research. Universities, government departments and medical research institutes utilise AI for predictive analytics, personalised medicine, drug discovery, and automated diagnostics, enhancing the quality and efficiency of research outcomes and discoveries. Beyond the health and medical field, AI is making substantial impacts in adjacent fields such as environmental management, where machine learning models predict and mitigate natural disasters, and in smart city initiatives, where AI-driven systems improve traffic, reduce energy consumption, and enhance public safety. The broad applicability underscores the critical role that robust infrastructure plays in harnessing the full potential of AI across diverse domains.

2.2. Overview of Selected Platforms

This section provides a comparative analysis of several key platforms in healthcare research, each offering different capabilities and resources to support advanced research. These platforms include diverse data ecosystems and software solutions that cater to various research and business intelligence needs through comprehensive data access, analytical tools, and robust computational resources. The list of key platforms, and the target audience and analytics capabilities for each are as follows:

1. BioData Catalyst: Cloud-based ecosystem for biomedical research, providing access to data, analytical tools, and computational resources. Target Audience: Researchers and data scientists

[24] https://www.neurodesk.org/docs/overview/
[25] https://research.unsw.edu.au/erica

focused on biomedical research, particularly in areas related to heart, lung, blood, and sleep disorders. Capability: highly capable for advanced analytics. Integrates with RStudio and Jupyter Notebooks, enabling the use of external libraries.

2. Datapine: Business intelligence (BI) and data visualisation. Target Audience: Business users, executives, and departmental managers who need easy-to-use tools for data visualisation, reporting, and basic analytics. Capability: user-friendly analytics but lacks deeper machine learning and AI capabilities.
3. Alteryx: Advanced data analytics, data preparation, and predictive modeling. Target Audience: Data analysts, data scientists, and business analysts who require robust analytics capabilities without needing to write extensive code. Capability: highly capable for advanced analytics; integrates with popular data science libraries in Python and R.
4. UK Biobank: Comprehensive biomedical database and research resource with extensive genetic, lifestyle, and health data. Target Audience: Researchers, scientists, and healthcare professionals from academia, industry, and charitable organisations involved in health-related research and data analysis. Capability: supports ML but not specialised in advanced analytics.
5. dbGaP: Repository for archiving and distributing data from studies investigating the interaction of genotype and phenotype in humans. Target Audience: Researchers and scientists, primarily serving the need to submit or access large datasets for comprehensive genomic studies. Capability: lacks integrated advanced analytics tools.

The following tables provide a comparison of the platform features in terms of advanced analytics and the underlying platform system.

2.2.1. Advanced Analytics

Table 2.3 Feature description for advanced analytics.

Visualisations	Interactive dashboards; charting tools; mapping.
Statistical Summary	Automated summary statistics; data aggregation; report generation.
No Code	Drag-and-drop interface; access to a library of pre-built analytical models and algorithms.
Collaboration	Shared workspaces; real-time collaboration; version control; data sharing; task management; granular permissions.
Workflows	Setting up automated data workflows and pipelines.
Machine Learning (ML) Integration	Integration with machine learning frameworks and libraries; in-built models.

Table 2.4 Platform summary for advanced analytics

	Visualisations	Summaries	No Code	Collaboration	Workflows	ML Integration
BioData Catalyst	+	+	+	++	++	++
Datapine	++	++	++	+	+	+
Alteryx	++	++	++	+	+	++
UK biobank	+	+	-	+	++	+
dbGaP	-	-	-	-	-	-

.

2.2.2. Platform System

Table 2.5 Feature description for platform system.

Accessibility	User-friendly interface; support for multiple languages; integration with other systems.
Interoperability	API integration with other systems and applications; common data exchange standards.
Data integration	Extract, Transform, Load (ETL) tools to import data from various sources; data warehousing in centralised storage; data mapping and transformation.
Security	Robust mechanisms to verify user identity; role-based access; audit trails; data encryption both at rest and in transit.
Privacy	Data de-identification; adherence to privacy regulations (GDPR, HIPAA, CCPA)
Scalability	Ability to process and analyse large volumes of data efficiently; distributed computing; load balancing; performance monitoring of platform.
Data Catalogue	Searchable catalogue of available datasets and metadata.
Support	Technical support channels; continuous improvement and updates to the platform.
Training and Documentation	Access to (and quality of) training materials, documentation, and user guides.

Table 2.6 Platform summary for platform system.

	Accessibility	Interoperability	Data integration	Security	Privacy	Scalability	Data Catalogue	Support	Training and Documentation
BioData Catalyst	+	++	+	++	++	++	+	+	++
Datapine	++	+	+	++	+	+	-	+	+

Alteryx	++	++	++	+	+	++	-	+	++
UK biobank	+	+	+	++	++	++	++	+	+
dbGaP	-	-	-	+	+	-	-	-	-

See Appendix A.2 for a detailed analysis of these selected platforms, including a discussion on the features and capabilities, access and cost, security and compliance, and user support and community.

2.2.3. Key gaps in the market

There are areas where existing platforms fall short in meeting the needs of researchers and healthcare professionals. The following highlights some of the key gaps:

Advanced analytics (Table 2.1)

- Platforms like the UK Biobank and Datapine have limited machine learning (ML) capabilities, which is a significant shortfall given the increasing importance of ML in data-driven healthcare research.
- While platforms like Alteryx and Datapine offer no-code solutions, others lack this feature, potentially limiting accessibility for researchers without advanced programming skills.
- Collaboration features are not consistently offered across platforms. This highlights a need for improved collaboration tools, such as real-time data sharing and version control.

Platform system (Table 2.4)

- There is a need for a platform that offers more robust interoperability with standard APIs and data exchange formats to existing healthcare systems.
- There are inconsistencies in security and privacy features. Platforms need to enhance data encryption and compliance with privacy regulations like GDPR[26] and HIPAA[27].
- The level of support is generally low, and training varies significantly. While platforms like BioData Catalyst offer comprehensive support, others like dbGaP lack adequate resources. Providing extensive documentation, tutorials, and responsive support can help bridge this gap.
- Efficient data management, including the ability to easily import, curate, and catalog data, is often lacking. Platforms that offer comprehensive data governance tools can help ensure data quality and usability.

[26] https://gdpr-info.eu/
[27] https://www.hhs.gov/hipaa/index.html

2.3. Broader Perspectives

In the last decades, the number of researchers in higher education has surged, rising from 4 million in 1980 to around 15 million today, resulting in a fivefold increase in research papers. Universities are expected to generate breakthroughs that benefit businesses, governments, and the public, theoretically boosting productivity and economic growth. However, research suggests that productivity gains were greater under the old corporate-led model of science compared to the current university-led approach (Arora, 2023). The current boom in AI innovation is driven by corporate researchers, rather than universities. This success highlights the potential for improved collaboration between universities and the corporate sector, and tighter competition policies could encourage businesses to reinvest in internal research (The Economist, 2024).

2.3.1. Skills and Workforce

Research computing and data (RCD) includes the broad range of infrastructure and services - and all the people - needed to support research, such as computing, storage, networking, virtualisation, software, cybersecurity, etc. Researchers face challenges in adopting new computational and data-intensive methods due to a lack of training and tradition in using advanced technologies, creating a significant skills gap (Schmitz, 2021). As such, research is increasingly dependent upon RCD skilled professionals who can help facilitate researchers' use of technical resources. However, the roles of RCD professionals are poorly understood (e.g., relative to traditional/enterprise IT), and recruitment and retention of RCD professionals is challenging, in part due to a lack of clear career paths. Schmitz et al (2021) suggest that overcoming these challenges requires establishing a professional association, coordinating community efforts, and expanding tools and services for RCD professionals.

In 2016–2017, a series of National Science Foundation (NSF)-funded Research Coordination Network (RCN) meetings brought together over 30 leaders in RCD and organisational scientists to discuss the evolution of RCD as a community of practice into a professional field (Berente, 2018). The conclusions of the participants led to the formation of the Campus Research Computing Consortium (CaRCC, carcc.org) - an organisation of dedicated professionals developing, advocating for, and advancing campus RCD and associated professions.

Furthermore, the 2021 National Research Infrastructure Roadmap[28] notes "*Rapid advances in computing techniques and analysis, and management of large and complex datasets, have resulted in researchers no longer having sufficient expertise in data management, computational and analysis techniques*". A NSF study (Barone, 2017) further underscores this issue, where 704 Biological Sciences Principal Investigators in this survey said "*The most pressing unmet needs are training in data integration, data management, and scaling analyses for HPC—acknowledging that data science skills will be required to*

[28] https://www.education.gov.au/national-research-infrastructure/2021-national-research-infrastructure-roadmap

build a deeper understanding of life". Environmental and other scientists face the same issues and unmet need for training and mentorship in computational skills (Baker 2017, Hampton 2017).

2.3.2. FAIR Data and Software

The FAIR principles (findability, accessibility, interoperability, and reusability) play an important role in making research data reproducible. But how should FAIR be applied to the complexities around code and software, workflows and documentation, methodology and technology? What do researchers need and how can institutional managers help? The FAIR Data and Software supporting Reproducible Research (FDSR) initiative by the Knowledge Exchange (KE) aims to answer these questions.

KE partners, composed of key national organisations within Europe, are developing infrastructure and services to leverage digital technologies to improve higher education and research. KE commissioned a report called "Approaches to scaling up reproducibility in research organisations" which provides a framework to scale up reproducible research practices within organisations by focusing on meso-level factors and identifying enablers such as tools, training, incentives, and policies. The status of reproducibility differs widely across research ecosystems and within organisations, teams, and faculties. National reproducibility networks and coordination efforts, as well as university consortia, offer potential support avenues.

The challenges for researchers to successfully carry out research in a reproducible way go far beyond technology and principles, it is a cultural change. Knowledge Exchange (KE) is currently investigating the minimal requirements for researchers to develop reproducible research practices as the norm, specifically within the scope of Artificial Intelligence (AI) methodology. This investigation will explore the ambitions, perspectives, and considerations of institutional management in stimulating reproducible research.

2.3.3. Data Risk Management

A data release is the process of making data publicly available or accessible to specific users or organisations, involving compiling, formatting, and documenting the data to ensure it is usable and understandable. However, any data-release method that produces useful data inherently comes with some nonzero risk to privacy and confidentiality (National Academies of Sciences, Engineering, and Medicine, 2024a). Blended data, which combines information from multiple sources, can pose significant disclosure risks, potentially allowing adversaries to learn sensitive health information. Clear communication of disclosure risk to the public is essential but currently challenging. Informed consent for data is complex and needs improvement to adequately convey risks and benefits to data subjects. Informed consent issues are amplified in blended data (National Academies of Sciences, Engineering, and Medicine, 2023b).

Effective risk management strategies should include both technical and policy approaches, be dynamic, involve stakeholder input, and follow best practices (National Academies of Sciences, Engineering, and

Medicine, 2024a). Coordination among data holders is necessary to manage these risks effectively, as lack of coordination can increase them. However, agencies often struggle to obtain stakeholder feedback, crucial for making informed decisions about disclosure risk/usefulness trade-offs. As such, there is a significant need for a robust Research Computing and Data (RCD) workforce to support the technical aspects of data blending and confidentiality protections. Additionally, new research and development efforts are needed to enhance privacy-protecting technologies.

2.3.4. Data Governance and Privacy

Data governance encompasses a wide range of technical, policy, regulatory, and institutional arrangements that manage the entire data lifecycle: creation, collection, storage, use, protection, access, sharing, and deletion. Some studies show public and private-sector data are estimated to generate social and economic benefits worth between 1% and 2.5% of GDP but have not achieved their potential due to challenges such as lack of trust, and conflicting interests of different stakeholders (OECD, 2019).

The OECD's Recommendation on Health Data Governance[29] shows that effective data governance balances the use of health data to improve healthcare quality, surveillance, system management, and research with the need to protect privacy and security. This requires comprehensive national frameworks based on high-level principles and developed through multi-stakeholder consultations. The OECD emphasises the importance of making health data widely available and processable for public policy while minimising risks. Specific undertakings include encouraging countries to develop and implement national health data governance frameworks and facilitating international harmonisation. Regular assessments are crucial for leveraging data-driven innovation and maintaining robust data governance frameworks.

In Japan, the Research Center for Open Science and Data Platform (RCOS) at the National Institute of Informatics[30] demonstrate that effective data governance involves a structured approach to expanding data access, regular quality checks, and the active use of Data Management Plans (DMPs).

UNESCO's guidelines on digital platform governance[31], while primarily focused on social media platforms, also offer insights in data governance for sensitive health data. Transparency and accountability, as highlighted by UNESCO, can be enhanced by establishing clear data usage policies and independent oversight mechanisms. Risk mitigation strategies for misinformation and data misuse ensure responsible and secure data handling. Promoting cultural diversity and conducting regular reviews to update governance practices in response to technological advancements are crucial for maintaining effective and relevant governance frameworks.

[29] https://www.oecd.org/health/health-data-governance.htm
[30] https://rcos.nii.ac.jp/en/service/dmp/
[31] https://unesdoc.unesco.org/ark:/48223/pf0000387339

2.3.5. Socio-Tech Systems and AI

In terms of artificial intelligence in research data infrastructure, there are several key socio-technical needs and considerations. There is a recognised need for AI/ML resources, including guidelines, checklists, and frameworks (Achten et al., 2020). Sensitive data management is crucial, specifically secure platforms, privacy-enhancing technologies (PETS), and synthetic data (Achten et al., 2020). Managing AI risk, ethics, explanation, trust, translation, and adoption (REETTA) is also necessary. Integrating AI and data quality management through information extraction and translation is also an important consideration (Achten et al., 2020). An AI infrastructure stack for Research Data Centers (RDCs) should encompass computational, spatial, and graph capabilities, along with data integration, linkage, and big data management. These findings are aligned with previous research in various domains, including HR management (Sivathanu & Pillai, 2018), telehealth and remote patient monitoring (Leung, 2023), and disaster risk management (Velev, 2023).

Comprehensive guidelines are crucial for ensuring best practices, ethical considerations, risk assessment, and transparency in AI/ML implementation in research and research infrastructure. These guidelines are essential across various domains to ensure effective and responsible use of AI. For instance, Olczak et al. (2021), guidelines are essential for reporting AI research in the medical field and facilitating consensus among stakeholders. Similarly, Aung et al. (2021) emphasise the need for regulatory guidelines to ensure the safe implementation and assessment of AI technology in healthcare. In talent management systems, Gonzalez et al. (2019) discuss the importance of addressing fairness and potential adverse reactions from job applicants during selection procedures. In higher education institutions, Kuleto et al. (2021) emphasise the need for best practices and understanding students' knowledge and attitudes towards AI/ML. Ethical and logistical considerations associated with AI implementation are discussed by Pethani (2021), while Zhang et al. (2021) highlight the significance of comprehensively understanding AI applications and opportunities for ethical and sustainable implementation. Additionally, Kourou et al. (2021) stress the importance of explainability and transparency in predictive models in cancer research. Together, these perspectives illustrate the overarching necessity of comprehensive guidelines to navigate the complexities and ethical consideration of AI/ML implementation.

The need for managing AI risk in research infrastructure is crucial due to several ethical and practical considerations. Ethical challenges arise when deploying AI infrastructures, particularly in fields like pathology and healthcare (Mckay et al., 2022; Eppler, 2023). These challenges include privacy, choice, equity, and trust (Mckay et al., 2022). To mitigate ethical concerns related to data-driven AI, key considerations such as auditing, benchmarking, confidence and trust, and explainability and interpretability should be reflected upon (Baird & Schuller, 2020). Additionally, standardisation and formalisation of the lifecycle of AI development and use are needed to ensure secure AI systems (Neretin & S, 2022). The implementation of AI infrastructure in research organisations, such as the Council for Scientific and Industrial Research[32], requires careful planning and coordination (Modiba et al., 2023).

[32] https://www.csir.res.in/

Building cybersecurity capacity and human capital is also necessary to secure AI infrastructure and enable knowledge translation, verification, and actionable decision-making (Ramim & Hueca, 2021). While AI has the potential to improve efficiency and effectiveness in human resource management, ethical challenges and governance issues need to be addressed to build trust in AI systems (Agustono, 2023; Guan, 2019). Establishing infrastructures that provide legal and regulatory guidance, clinician competencies, and learner-centric resources can support the safe and effective use of AI-based tools in healthcare (Novak, 2023).

2.4. Detailed Analysis of Selected Organisations

2.4.1. Organisation for Economic Co-operation and Development (OECD)

Introduction

The Organisation for Economic Co-operation and Development (OECD)[33] is an intergovernmental organisation with 38 member countries, headquartered in France. The OECD is working with countries as they develop policies to strengthen their health data infrastructure, with the objective of creating safer, better and more efficient health systems and healthier populations.

Countries need the right data infrastructure in place for producing health statistics and measuring healthcare quality and outcomes, leveraging data and extracting information from registries, administrative data, electronic health records, and referencing them with other sources often beyond the health system.

Key Programs and Initiatives

Health Information Review: The OECD assists countries in evaluating and enhancing their information infrastructure to improve health data. This process includes:

1. Assessment of the health information system and its governance;
2. Integrating health data infrastructure across care settings and social services;
3. Incorporation of new and emerging technologies and modes of analysis; and
4. Building the workforce needed to operate a 21st century health infrastructure system.

Council Recommendations: The OECD Recommendation on Health Data Governance outlines principles for harmonised health data governance across countries, promoting national frameworks and trans-border cooperation. Notably, the published report “Health Data Governance for the Digital Age: Implementing the OECD Recommendation on Health Data” reviews the implementation from 2016-2021, noting progress but also ongoing efforts in data sharing, accessibility, quality, interoperability, and privacy protections. Efforts will continue through 2022-27, focusing on cybersecurity, harmonising health data governance for multi-country projects, and improving global health data interoperability.

[33] https://www.oecd.org/health/health-data-infrastructure.htm

The OECD Declaration on Government Access to Personal Data Held by Private Sector Entities, adopted in 2022, aims to improve trust in cross-border data flows by clarifying how national security and law enforcement agencies can access personal data under existing legal frameworks. It represents a major commitment by the 38 OECD countries and the European Union, and it is open for adherence by other countries.

Partnerships and Collaborations

The OECD regularly holds minister-level meetings and forums as platforms for a discussion on a broad spectrum of thematic issues relevant to the OECD charter, member countries, and non-member countries. The target audience for the organisation includes policymakers, healthcare administrators, medical researchers, data governance professionals, and international health organisations.

2.4.2. Health RI

Introduction

Health-RI[34] is a non-profit foundation based in the Netherlands, dedicated to creating an integrated data infrastructure for health information. Health-RI's mission is to improve the health of citizens and patients by enabling data reuse through a comprehensive data infrastructure that supports research, policy development, and innovation. By providing infrastructure for secure data exchange, harmonisation, and analysis, Health-RI ensures that researchers have access to diverse data sets. Health-RI aims to improve healthcare, advance medical research, and develop personalised medicine.

Key Programs and Initiatives

Health-RI's strategy follows three lines of action:

1. Collective Action: Optimising the conditions for building a national health data infrastructure by advocating for standardisation and policy alignment across institutions.
2. Building a National Health Data Infrastructure: Facilitating initiatives and collaborations to develop a cohesive, nationwide health data network.
3. Providing Services and Tools: Supporting researchers and data managers by ensuring that infrastructure services, tools and data are easy to locate and use.

Output and analytics

Health-RI provides a suite of advanced services to support integrated health data management and research. These include a high-throughput data pipeline service for designing and managing complex data workflows, a medical imaging data platform for storing and analysing imaging data, and a genomic data portal for accessing and visualising genetic information. Additionally, the infrastructure includes an interoperable case report form generator for standardised data collection and a support service desk for technical assistance and coordination.

[34] https://www.health-ri.nl/

Partnerships and Collaborations

Health-RI collaborates with various public and private initiatives, both nationally and internationally, that are involved in the reuse of research and healthcare data. Notably, Health-RI collaborates closely with European research infrastructure organisations that are part of the European Strategic Forum for Research Infrastructures (ESFRI).

2.4.3. National Human Genome Research Institute-funded Electronic Medical Records and Genomics (eMERGE)

Introduction

The Electronic Medical Records and Genomics (eMERGE)[35] Network is a consortium of U.S. medical research institutions, funded by the National Institutes of Health (NIH). It unites experts in genomics, statistics, ethics, informatics, and clinical medicine to conduct genomics research using electronic medical records, focusing on discovery and clinical implementation. The Network promotes the broader adoption of genomic medicine by encouraging affiliate memberships and openly sharing its resources.

Key Programs and Initiatives

The eMERGE Network integrates genomic data with electronic medical records (EMRs) to advance personalised medicine, improve patient outcomes, and foster international collaboration in genomic research[36]. Since its inception, it has progressed through several phases: Phase I (2007-2011) demonstrated EMR feasibility for genomic research, Phase II (2011-2015) expanded into pharmacogenomics and returned genetic results to clinicians, and Phase III (2016-2020) sequenced 109 genes in 25,000 patients, integrating these results into EMRs. The current phase (2020-2025) focuses on genomic risk assessment and management across ten clinical sites and a coordinating centre.

eMERGE operates through specialised workgroups targeting critical aspects of genomic research and implementation, including comprehensive risk assessment, EHR workflow and infrastructure, phenotyping, provider uptake and outcomes, PRS validation and evaluation, and addressing recruitment, retention, and ethical issues.

Output and analytics

The eMERGE Network has developed 68 electronic phenotypes and published 770 papers. In terms of advanced analytics tools, the network offers a resource library featuring tools for clinical decision support, phenotype analysis, and educational content on genetics. These tools and resources are publicly available to facilitate research planning, feasibility assessment, and the integration of genomic data into clinical practice.

[35] https://emerge-network.org/
[36] https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE

Partnerships and Collaborations

The eMERGE Network collaborates with numerous clinical sites and network affiliates to advance genomic research using electronic medical records. eMERGE encourages further collaborations and welcomes new affiliate members to join its efforts in genomic medicine.

2.4.4. Genome Institute of Singapore (GIS)

Introduction

The Genome Institute of Singapore (GIS) is a premier research institute established in 2001 under the Agency for Science, Technology, and Research (A*STAR). Located in Singapore, GIS is dedicated to advancing genomic science and its applications in healthcare, biotechnology, and the life sciences. The institute aims to translate research findings into practical benefits for human health and economic growth.

Output and analytics

The Genome Research Informatics & Data Science (GRIDS) Platform at GIS provides essential computing support to GIS research groups, with a dedicated team of 15 staff for the platform. The platform develops and maintains computationally agnostic data analysis pipelines and integrated data management systems. To address the challenges of transitioning to cloud-based solutions, GRIDS deploys cloud-enabler ecosystems like RONIN[37] and promotes infrastructure-agnostic workflows such as nf-core[38].

Partnerships and Collaborations

GIS engages in extensive collaborations with international and local research institutions, universities, hospitals, and industry partners. In terms of commercialisation, GIS is developing a cutting-edge suite of software[39], pipelines, and databases for bioinformatics, genomics, and clinical data analysis. These tools are available for commercial licensing, and GIS is open to joint software/database development, modification and maintenance of projects with industry partners.

2.5. Summary

The environmental scan highlights the critical components and emerging tools and technologies shaping the National Framework for Advanced Health Analytics Infrastructure. Key insights reveal that scalable infrastructure is essential for supporting AI and analytics in health, facilitating efficient data processing, secure storage, and real-time analysis. National initiatives in Australia, such as NCI, Pawsey and ARDC, exemplify the integration of high-performance and cloud computing resources to drive innovation across

[37] https://ronin.cloud
[38] https://nf-co.re
[39] https://www.a-star.edu.sg/gis/our-science/gis-software-for-commercial-licensing

various sectors, including health and medical. There is an increasing popularity of TPUs and GPUs due to the large computational resources required for advanced AI and analytics.

In the health domain, digital technologies and AI advancements have transformed patient experiences through integrated and personalised treatments. Investments in research infrastructure to support advanced analytics are being made by several countries, demonstrating a strong commitment to scientific research and technological innovation. The recently launched National Artificial Intelligence Research Resource (NAIRR) pilot aims to democratise access to advanced computing resources for researchers. Balancing disclosure risks with data usefulness, enhancing privacy-protecting technologies, and improving informed consent processes are key considerations. Synthetic data generation and federated learning are emerging technologies that can address data privacy and accessibility challenges, promoting collaborative research without compromising sensitive information. The environmental scan also indicates that effective governance frameworks are needed to support data sharing, linkage, and analysis while balancing public benefit with privacy and security requirements. International cooperation, transparency, accountability, and adherence to FAIR principles should be considered in the development of data governance frameworks and fostering a cultural shift toward reproducible research practices.

3. Survey

An open online survey was created in order to get a broad understanding of the current state of use of health data and/or advanced analytics within the Australian scientific community. The survey was promoted to ADSN, the Statistical Society of Australia (SSA), Machine Learning Community of Practice for Australia (ML4AU CoP), Australian Bioinformatics And Computational Biology Society (ABACBS) and Data Management Association Australia (DAMA) members as well as through institutional mailing lists, the ARDC social media channels and personal networks. In addition to publicly distributing the survey and inviting responses, targeted stakeholder groups were approached through direct mails with an anonymous link to the survey. This included circulating in the relevant government and health networks.

The survey responses are not probability samples from defined sample frames; it was not possible with the resources and time available to conduct a survey of this nature. In several cases, the target populations are relatively small, and an attempt was made to reach all members of these populations. To this end, the survey was sent around through newsletters and social media channels, and recipients were encouraged to forward the survey to others. However, without sample frames response rates cannot be determined, which should be considered when the quantitative results are examined.

The survey questions are listed in Appendix A.3. They were grouped into five main themes/areas of interest: data privacy and ethics; version control; data integrity; compute resources and advanced analytics. The results of the survey are detailed in Appendix A.4. The main findings, under each theme, are discussed in the following sections.

3.1. Results

3.1.1. Overview of Respondents

The survey had a total of 156 valid responses where a valid response is considered a response that progressed through at least 50% of the survey and took more than one minute to answer the survey. There were 110 invalid responses that were not considered further. Note that a response was not required for any of the questions, hence the total number of responders for each individual question varies.

The majority of the responders (n=110, 71%) have a PhD or MD. The field that the responders obtained their highest qualification tended to be the health or biomedical area (n=74, 47%) rather than the fundamental areas like Mathematics, Statistics, Computing or IT (n=46, 29%). There was a high representation of the university sector with 77% (n=120) of responders being affiliated with a university.

Additionally, respondents were asked a series of questions about their skills in programming and analytics. While many rated their basic maths/statistics skills as at least good, more advanced analytics and computing skills were less represented (see Table 3.1 below). Note that in this context, we use AI/ML to represent the larger field of advanced health analytics which also includes advanced mathematical and statistical methods.

Table 3.1 Skills in programming and analytics.

Skill	None-Some	Good-Excellent
Applied Statistics/Mathematics	42 (27%)	114 (73%)
AI/ML	98 (63%)	58 (37%)
R/Python programming	78 (50%)	78 (50%)
Cloud/HPC computing	88 (56%)	68 (44%)
Data Engineering/Databases	93 (60%)	63 (40%)

At present, data used within their organisation is internal only for just 11 respondents, with 45 using both internal and external data sources and 87 respondents stating that only external data sources are used. Similarly, when asked about the location of data and the necessary compute resources, out of the 152 respondents who answered this section, 59% stated the data was co-located and a similar number stated that the data needed to be moved in order to be analysed.

3.1.2. Privacy and Ethics

All respondents agreed that data privacy and ethics is a concern with 91% (n=142) rating it as very or extremely important. Likewise, the majority also agreed that the main issues in this area are obtaining the authority to access relevant data (72%, n=113); understanding their legal rights and responsibilities (58%, n=91) and secure data storage (67%, n=105) and transfer (65%, n=101). Additional concerns included re-identification risks, communication/explanation of privacy and ethics to non-researchers and community members (participant trust), legislation around data privacy and ensuring data privacy once it has been shared.

Current approaches being used by their organisations included sharing files via the cloud (n=93, 60%) and using secure centrally maintained data repositories with read-only access (n=83, 53%). Use of encryption when transferring data was available for 37% (n=58) of respondents, while a similar number reported the availability of a research data champion or other central contact for help and advice. Only 28 respondents (18%) indicated synthetic data was in use, while others mentioned the use of secure analytic environments (n=9, 6%).

Respondents were also asked what they thought is important with respect to data privacy and ethics in an ideal world. Responses are summarised by theme in Table 3.2 below.

Table 3.2 Response summary of data privacy and ethics in an ideal world grouped by theme.

Theme	Recommendations
Access	• Significant reforms are necessary to mitigate risk-averse policies that hinder timely access to valuable research data. • Implement systems like the UK model where synthetic data is used for developing syntax, which is then applied to actual data by a trusted third party. • Credentialed and approved researchers should have the infrastructure to access and analyse sensitive medical data, to the required security level without interfering with daily operations. • Focus on de-identification and reducing the storage of personal data by entities that don't need it, protecting against misuse. • Ensure data is freely available to researchers and linked to identifiers like the Individual Healthcare Identifier (IHI) to enhance AI-enabled prediction and clinical trial matching. • Enable sharing of identified data across state boundaries under appropriate ethical and governance approvals.
Consent	• Educate consumers on consent for data sharing and clearly differentiate data use in healthcare, advertising, etc.
Community of Practice	• Facilitate in establishing a consensus on acceptable risks, restrictions, and access safeguards. • Facilitate in community adoption of standardised practices with high-quality metadata including licensing details.
Custodian	• Individuals have access to their own data and can authorise, prevent and revoke access by others (informed consent and transparency as to who has access to their data and for what purposes). This information should be clear and understandable. • Healthcare providers have mutual access to coordinate care and researchers have access to anonymised/synthetic data.
De-identified	• De-identify all health data and publish only summary data/models. • Address challenges in de-identifying data, as studies show 48% of US census data can be re-identified.
Environments	• Ensure availability of Secure/Trusted Research Environments and support federated activities nationally and internationally. • Consolidate datasets within the same analysis environment to minimise data transfer.
Ethics	• Researchers must understand ethical and privacy requirements for sensitive health data, supported by well-resourced ethics committees for outreach,

	audit, and enforcement. Ensure ethics committee members are skilled in recognising levels of risk. • Implement a centralised ethics approval framework with opt-in/out options for data sharing. • Address challenges associated with perception of ethics approvals as a barrier, by highlighting good governance. Education on dangers of privacy breaches and potential harms.
Procedures	• Implement clear policies, procedures, and workflows for privacy provisions. • Provide accessible, authoritative advice and education for compliance with data privacy principles and legislation. • Establish a common framework for health data governance across state jurisdictions, distinguishing research from business operations in hospitals. • Implement centralised approval for collecting de-identified patient data, simplifying researcher access without multiple ethics approvals. • Facilitate transparency in processes related to collection, usage, sharing and storing of data.
Regulation	• Implement oversight with mandatory breach reporting, severe consequences, and ethics committees empowered to block studies. • Establish standard guidelines, improved accountability, and regular audits to ensure compliance with privacy laws. • Focus policing on commercial misuse of data, ensuring personal identity protection from overreach by banks and tech companies through stringent regulations. • Harmonise legislation and regulations across jurisdictions/organisations and ensure compliance with 21CFRpart11, GDPR, NHMRC, etc.
Risk Management	• Adopt risk-based, proportionate data management approaches (balance data privacy with accessibility). • Address challenges with overly strict privacy interpretations that impede research. • Implement trust measures for people/institutions with a proven track record to allow greater use of more timely data.
Security	• Ensure secure, de-identified data storage with robust backup and access controls. • Enable easy, secure sharing among collaborators with additional authentication. • Grant access to de-identified data to approved researchers for public good, with systems to protect privacy. • Improve education on data security and confidentiality. • Enhance communication between Trusted Research Environments (TREs). • Implement centralised, secure storage for research groups, with strong encryption and continuous security monitoring. • Use secure cloud locations for project access across organisations, avoiding insecure methods like email.

	• Adopt data minimisation, purpose limitation, and strong security measures. • Facilitate timely access to secure data environments.
Standardised	• Implement a coordinated national approach with standardised systems (data storage, platform for consistent formatting, function, and access) and clear, recognised standards. • Adopt a common standard like OMOP, enabling federated analyses while data remains local. • Foster international cooperation to address data privacy challenges by harmonising standards, facilitating cross-border data transfers, and assisting in enforcement actions.
Training	• Improve training in data management and storage. • Facilitate greater understanding on where responsibility and accountability sits, and how these translate internationally. • Support for researchers to de-identify data and results using state-of-the-art methods and technology.

3.1.3. Version Control

All respondents considered version control as important with 72% (n=112) rating it as very or extremely important. The most common approach currently in use was file naming conventions (n=97, 62%). Nine respondents stated they were unsure what methods were in use, or that none were in their organisation. Important tools for researchers were linked to reproducibility, with minimum standards for metadata (n=90, 58%); compatible protocols (n=83, 53%) and uniform use of program(s) (n=74, 47%) being highly regarded.

3.1.4. Data Integrity

Data integrity and quality were viewed as being at least very important to almost all respondents (n=146, 94%). In contrast, around half the respondents (n =77)had only, at best, a moderate amount of confidence in the data they typically access. About 29% of those respondents felt that the documentation available about the data contained enough detail (overall half of the respondents agreed with this statement). Of those that thought the documentation was lacking, 80% (n=61) thought the variable definitions were lacking and similarly 79% (n=60) thought information about missing data was lacking.

In terms of access to benchmark or synthetic datasets, only half the respondents thought this would be useful for their work (n=81, 52%). Most do not currently access data that is behind a paywall (n=128, 82%). However, if the ideal[40] dataset required payment to access, most would consider using an alternative, free but less ideal dataset (n=139, 89%). This is mainly due to potential issues with paying for access, and again most (n=130, 83%) would consider abandoning the project due to lack of funds.

[40] where ideal means high quality data, with detailed documentation and excellent integrity.

3.1.5. Compute Resources/Infrastructure

In terms of the compute resources currently available in the health data analytics space, the majority of respondents are provided with access by their workplace (n=144, 94%). Of those, 58 respondents thought that the resources available were not sufficient for their needs (40%). Of the 96 respondents who indicated which computing facilities they (or their colleagues) access, 76 (80%) use at least one commercial provider often as well as an NCRIS facility (e.g. NCI). Funds to pay for this access generally come from research grant funds (45%, n=69) or institutional agreements (35%, n=54).

3.1.6. Advanced Analytics

At present, most respondents use simple queries, or basic analysis on health data. Many have never used AI/ML and almost half have never used Deep Learning or Large Language Models (LLMs). A summary of techniques used on health data is in Table 3.3 below:

Table 3.3 Techniques used on health data.

Technique used	Never	Most of the time/Always
Database queries/data extraction	14 (10%)	69 (47%)
Exploratory data analysis	7 (5%)	86 (59%)
Standard analyses (e.g. regression)	6 (4%)	83 (56%)
AI/ML	51 (35%)	30 (21%)
Deep Learning/LLMs	71 (49%)	21 (14%)

In terms of access to advanced analytics capabilities and training, 65 (43%) respondents were either unsure or did not think their organisation has sufficient capabilities in-house, even though the majority of respondents are affiliated with the university sector. Areas of need in advanced analytics include AI/ML expertise; distributed machine learning; cyber security and using cloud or HPC computing environments.

71 respondents (47%) stated their organisation does not provide access to formal training in advanced analytics. While, of those who do have access, 30 (38%) said the training was not sufficient for their needs in this space. The areas in which respondents would like to see more training are summarised in Table 3.4 below.

Table 3.4 Favoured training areas.

Training Area	Number (# out of 100 respondents)

AI/ML methods	71
Advanced (bio)statistics methods	62
Federated learning	52
Using version control software or GitHub	50
Cloud computing	44
Specialised mathematical models	41
Data encryption	32
Accessing NCI/HPC resources	31
Ethics	30

Roughly half the respondents either have access to or utilise informal training opportunities such as meetups and hacky hours. The preferred methods for engaging with others about advanced analytics are either online (n=104, 70%) or at conferences or workshops (n=98, 66%). Many respondents (n=84, 56%) are not members of any community of practice (CoP) for either advanced analytics in general or health in particular. Biocommons, VicBioStat and the Statistical Society of Australia events and workshops as well as internal institutional events/series were common responses, with 90 (60%) rating the need for such CoPs as at least very important.

Finally, respondents were asked to provide any additional comments which are summarised by theme in Table 3.5 below.

Table 3.5 Additional comments by theme.

Theme	Recommendations
Access	• Support Australians to take part in building "vertical" open-access international databases. • Affordable research platform for data sharing across organisations, access to high-performance compute at low or no cost. • Address challenges to streamline approval processes, currently this lengthy process prevents students from partaking in research projects.
Collaboration	• Collaborate with technical experts rather than developing internal expertise. • Promoting trusted data sharing for regional and global collaboration. • Address challenges in finding common needs with silos e.g. HPC, linked admin data, and clinical health data researchers

Community of Practice (CoP)	• Biostatisticians have consensus on robust study methods (TRIPOD, TRIPOD+AI, PROBAST/CHARMS), but data science practices often conflict, creating challenges in health research. • ARDC could lead by highlighting reporting standards for Australian researchers.
Data	• Make datasets available in Observational Medical Outcomes Partnership (OMOP) for easier collaboration to save data wrangling time. • Create complex synthetic datasets for teaching • Establish national benchmarks/templates for data storage and metadata quality. • Localise models developed on North American clinical data, with ARDC leading efforts to unlock the potential of existing data in Australia. • Develop nationally representative data assets, including detailed clinical data from hospitals, primary care, aged care, and prisons. • Enhance cross-jurisdictional and national data linkage and integration with clear national leadership.
Exemplar	• The ABS DataLab infrastructure business model is exemplary. Noted were BioCommons services (Galaxy,NextFlow, etc) and linked data services such as PHRN.
Governance	• Hold organisations receiving government funds accountable for client satisfaction through third-party assessments, addressing the unmet needs in the health research sector. • Establish national data governance frameworks and standards for interoperability, consistency, and quality. • Ensure researchers accessing data have clear, relevant research questions and proficiency in data analysis.
Guidelines	• Provide senior leadership with guidelines on data privacy, coding scripts, and excluded information. • Offer machine learning guidelines and healthcare use cases with aggregated, unidentifiable data. • Educate researchers on available resources and access methods, using approachable terms for those unfamiliar with Python, Jupyter, Linux, and HPC. • Ensure researchers accessing data have clear, relevant research questions and proficiency in data analysis.
Models	• Emphasise understanding and mastery of basic methods in ML before advancing to complex models. • Look deeper into the needs for foundational models, especially sovereign AI, synthetic data, federated learning.
Support	• Data science expert to support data repository aggregating from several family service organisations to optimise benchmarking and practice.

	• Address critical need for sustaining advanced domain-specific research software development, possibly through co-funded positions to retain essential expertise. • Need for available expertise, as opposed to training (e.g. Kubernetes support).
Tools	• A robust and scalable hardware infrastructure is essential for advanced analytics workloads. This includes high-performance computing resources, specialised hardware accelerators (such as GPUs and TPUs), and cloud-based platforms. • Creating a virtual space where a group of people can collaborate to do research using modern techniques determined solely by them.
Training	• Establish high-quality panel for population health and health marketing research. • Investing in education programs that help build a skilled workforce for advanced analytics techniques e.g. data science, machine learning, artificial intelligence, and related fields. • Provide concise information for healthcare clinicians on analytics capabilities and practical considerations.

4. Sectoral Dialogues through Workshops/Interviews

Workshops and interviews were conducted to inform an understanding of the use of health data and advanced analytics within the Australian scientific community.

4.1. Design

To enable a deeper understanding of the analytical needs of academics, industry personnel, government officials and clinicians, a mixed methods approach was utilised to integrate information and findings from focus groups and supplementary interviews.

A total of four focus groups/workshops and three interview sessions were held between Dec 2023-July 2024. The first workshop was conducted at the ADSN conference in Dec 2023. Subsequent workshops were conducted online. The open workshop invitations were posted on the Eventbrite platform and disseminated through ADSN and ARDC social media channels. Selected targeted invites were also sent to government agencies and other related associations and organizations.

The participants were introduced to the ARDC framework goals and mission and presented with the initial Advanced Analytics survey results that identify the needs and gaps in data analytics and infrastructure in Australia. Mentimeter was used during the workshops to capture a) challenges faced during the analytical life cycle, b) ranking of infrastructure based on personal importance and finally c) justification of ranking. This resulted in an array of information and ranked infrastructure categories which were used to build smaller focus-groups during the workshop. Each focus group was moderated by 2-3 ADSN and ARDC leads and explored specific infrastructure categories in detail, with notes taken on the findings. Additional interview sessions were facilitated after the workshops to capture ideas and challenges from participants who were unable to attend the workshops. The interviews were open in-depth discussions on the different topics explored in the workshops. The data and findings from these workshops and interviews were synthesised to develop the framework.

4.2. Results

The findings from the workshops and interviews are summarised below, detailing the challenges and recommendations identified by researchers and stakeholders in the advanced health analytics infrastructure landscape.

4.2.1. Challenges

The following Table 4.1 expands on the challenges identified during the workshops and interviews. These challenges include difficulties related to data access, management, governance, and the integration of advanced analytics tools and methodologies.

Table 4.1 Challenges in the advanced analytics infrastructure landscape.

Theme	Challenges
Underpinning Infrastructure	• Researchers do not want to pay for cloud services. • Infrastructure disappears after the duration of study • Strong need for affordable, high-performance computing environments to handle sensitive healthcare data. • Healthcare and Personally Identifiable Information (PII) have additional jurisdictional and organizational requirements
Access	• Barriers in accessing data:bureaucratic hurdles, regulatory restrictions, and inconsistent data sharing policies across institutions. There is a critical need for timely access to high-quality data. Delays in obtaining data can hinder research progress and the ability to respond to urgent health issues. • Lack of access to suitable data assets as well as skills for processing data are highlighted. Accessing, processing and managing sufficient, relevant and quality data to apply AI to clinical problems are persistent issues. • Reference data assets have improved greatly across the nation, but still face major barriers (e.g. state-held vs Commonwealth held datasets). • Barriers for use of LLM for health and medical data, where opportunity is to get structured data from unstructured. The barriers can be largely grouped around sensitivity of data and the reliability of outputs. • Researchers are increasingly relying on opt-in online panels to understand prevalence of disease, prevention (e.g. screening), health risk behaviours and changes in trends; these have been shown to be unreliable and misrepresentative. • Constraints in sharing de-identified data with overseas collaborators (e.g. unit level mortality outcome data can't be shared with overseas collaborators due to AIHW specification) • Researchers use data from other countries with easier access at national level. However, such data may not be relevant for the Australian population.
Community of Practice (CoP)	• Academics are currently incentivized to publish, as opposed to actually improving health outcomes. Communities of Practice could change this through a culture to some extent, mobilizing academics to solve problems instead. • There is limited space for imparting practical wisdom, where researchers can collaborate, share insights, and provide feedback. There is a need for

	socio-technical infrastructure, including Communities of Practice (CoP) for Health Analytics areas using existing or new CoPs.
Data management	• No national standards in curation, governance and release for data. • Data collected in systems are for operational purposes, not research. Quality enhancement is needed for insights to be drawn. • Data security and release is not standardised increasing the risk of data breach. • No commercially available code repository suitable for sensitive data, e.g. GitHub doesn't meet legal privacy requirements. • Gaining representative clinical data for uncommon health conditions from primary care is challenging. • Need to reapply to ethics committees in different jurisdictions using different systems. • Generating synthetic datasets that accurately reflect the diverse and complex nature of Australian health data is difficult.
Governance	• Limited ability to choose which secure environment to use - often mandated by data custodians. • National Federated Learning Infrastructure barriers include hospital policies, data custodians' governance requirements (e.g., 5Safe framework), requiring extensive national coordination and training. • Engaging data custodians poses a challenge, as does governance across institutional and state boundaries, particularly for data acquired within hospital services. • Diverse perspectives among university ethics committees on data sharing benefits versus risks highlight the need for a national approach to ethics in data governance. • The rapid growth of the AI market outpaces the workforce's ability to interpret models and integrate them into decision-making processes. • A privacy breach would be disastrous, highlighting the need to recognise risks and ensure a secure environment.
Linked data assets	• Address data custodian restraints for cohort study data linked with national datasets e.g. NDI/ACD. • Before standardisation, it's crucial to clarify current data collection and storage practices. • Data silos within and between health services prevent data integration for hospitals, clinicians, and governments.

	• Inconsistent radiology data across sites may be problematic for future use in AI. • Biospecimens and data from biobanks, cohorts, trials, and researchers are siloed and often inaccessible to the research community.
Models	• National Federated Learning Infrastructure challenge lies in the overhead for collaboration among diverse groups and organisations, requiring leadership from an entity like ARDC. This effort is too extensive for a single research grant but would benefit many. • Addressing the black box problem: Many AI techniques, particularly deep learning algorithms, are often considered black boxes due to their lack of interpretability
Tools	• Clinicians and researchers face difficulties with command-line and scripting tools, yet early stage infrastructure is hard to make user-friendly. • Health and medical datasets are initially identifiable, but vary in the implementation and interpretation of de-identification. • Amazon Web Services (AWS) tools are primarily business-oriented and often do not align with the specific needs and objectives of researchers and organisations. • Lack of access to tools and platforms for ease of learning and use. • Foundational Models/ GenAI was mentioned in the context of processing Electronic Health Records (EHR) data processing and co-pilots
Training	• The survey indicates a strong interest but weak technical base, highlighting the need for national investment in training on topics like Federated Learning.
Lifecycle Specific Issues - Problem Formulation	• It is recognized as the most challenging step in the life cycle. Researchers often struggle with clearly articulating the problem they intend to solve, which can lead to inefficiencies and misdirected efforts in the analytics process. Problem formulation stage also encompasses a range of issues including understanding limitations of the techniques and tools, resource planning, ethics and governance. Addressing risks associated with data privacy, security, ethical use is a key challenge. High setup costs for each new research project is a new barrier to health research and clinical practice improvement.
Lifecycle Issues - Data Acquisition, Processing and Management	• There is interest in synthetic data. Differing views about synthetic data, which is deemed applicable to some use cases. Both techniques, tools, and guidelines for handling synthetic data responsibly are needed to address key questions and ensure these data products are suitable for specific research purposes. • Shortfalls in skills and training as well as infrastructure to manage and process sensitive data are important. Sensitive data is locked up in siloes and often can not be sent across organizational and jurisdictional boundaries

	• Ensuring data assets are FAIR and compliant
Lifecycle Issues - Model Development and Validation	• Shortfalls in skills and resources including data, tools and other infrastructure for model development and validation were reported. • The level of validation required of the models in real-world settings to ensure • their accuracy and reliability has been hard to achieve in the research settings. • Lack of supporting infrastructure for tool validation, despite available resources for development; effective translation and implementation are crucial for real value. • Need for infrastructure for managing large volume of sensitive data (Pipelines, Lifecycle, Data Management) • National Federated Learning Infrastructure challenge lies in the overhead for collaboration among diverse groups and organisations, requiring leadership from an entity like ARDC. This effort is too extensive for a single research grant but would benefit many.
Lifecycle Issues - Deployment and Translation	• Gap between technique and practice; only a small proportion of models developed are deployed. • Concern around inappropriate usage of advanced analytics (mismatch between models and problem, aiming for technically valid solution to a real clinical problem). • Risk of writing proposals which lead to inappropriate methods. Pilot projects are ineffective as they often have similar requirements to full-scale projects.

The participants also rated the urgency and importance of key infrastructures on a notional scale, and the average normalized scores are shown in the chart below in Figure 4.1. Note that the bubbles are arbitrarily sized for visual appeal and hence go over the range of scale [0,1].

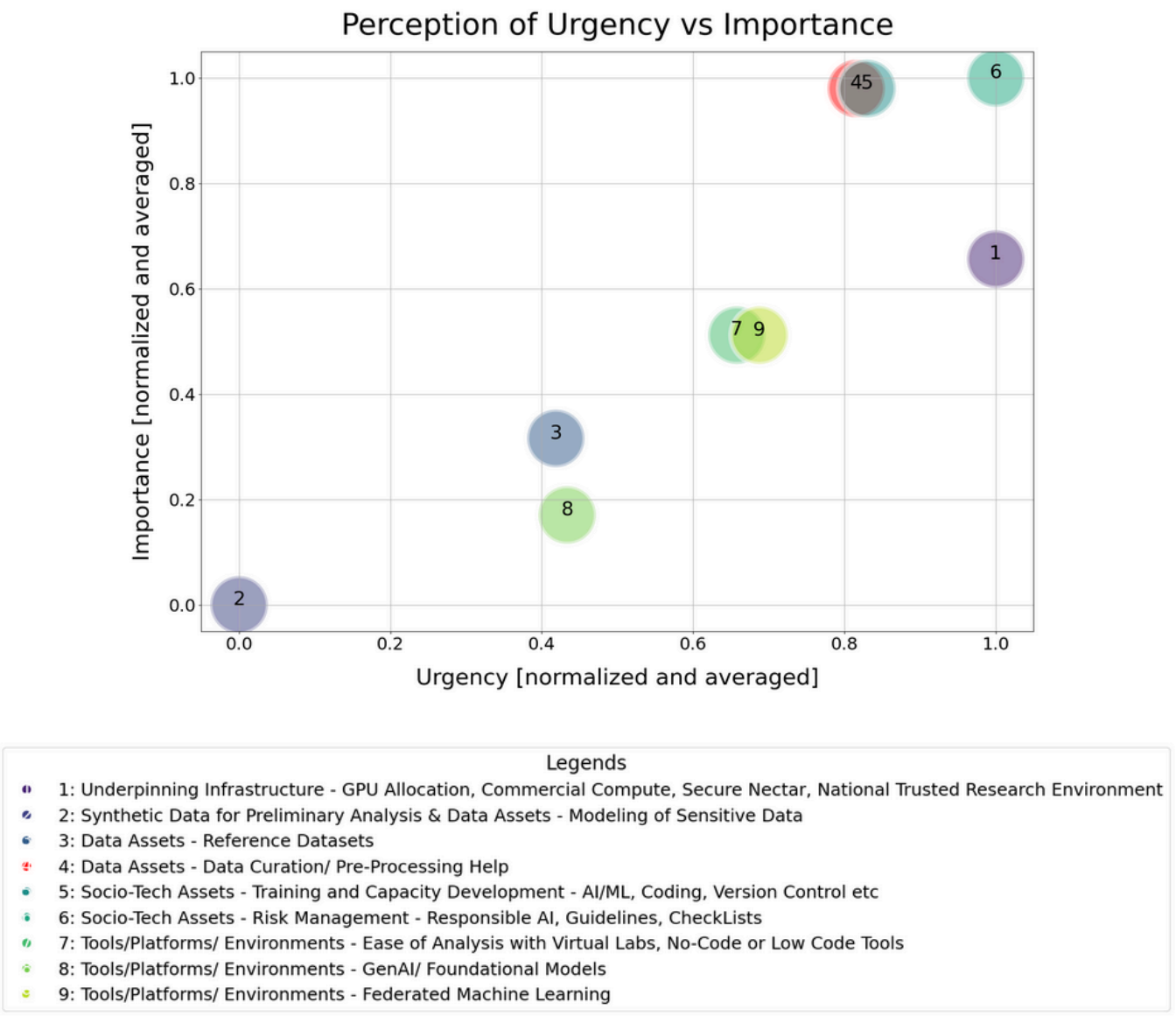


Figure 4.1 Average scores of urgency and importance of the infrastructure opportunities

4.2.2. Recommendations

The following Table 4.2 summarises the recommendations identified during the workshops and interviews, outlining key strategies to enhance the infrastructure and capabilities.

Table 4.2 Recommendations for the advanced analytics framework.

Theme	Recommendations
Access	• Access compute in a form that also allows for identified data to be used (health services data). • National implementation of AI tools or systems suitable for sensitive data, but accessible to everyone.
Community of Practice (CoP)	• Harness existing communities of practice (or establish new ones as needed) that encourages researchers to collaborate, share insights, and receive feedback on problem formulation, best practices, and lifecycle practice efforts, focusing on practical issues rather than just publication outcomes. • Undertake investigation as to current use of advanced analytics tools across disciplines (e.g. Biostatisticians have a good community of practice).
Data Acquisition Processing and Management incl. Data Access	• The development of national reference data assets, including curated datasets and synthetic data, is essential to provide researchers with reliable and standardized data sources. • National online panel to access accurate and representative data quickly, with ability to link health records securely. • National approach to develop consensus and frameworks for consistent EHR data curation, governance, and release across health services. • Commercially available solution for code repository for highly protected data, with CI/CD, automation, etc. • Searchable portal of health datasets, with guidelines on metrics and curation, and document of limitations and metadata. This makes datasets more findable, promotes reuse/reproducibility and helps to identify gaps in the datasets. • Prioritise inclusion of state and territory hospital, aged care and census data in the ABS Person Level Integrated Data Asset (PLIDA) data asset. Valuable for exploring issues of access and equity. • National ethics application (incorporate data access and sharing, and AI in research). • Develop a series of synthetic datasets as benchmarks to accelerate tool development and testing without needing ethics approval.
Underpinning Infrastructure Environment	• Strong need for secure, trusted and high-performing computing infrastructure. The implementation may take a combination of high-performance computing (HPC), Trusted Research Environments (TREs) as well as for parts of Nectar as well as commercial infrastructure if needed. It would also require parts of Nectar to handle sensitive healthcare data as a priority. Security measures should be in place to protect data privacy and ensure compliance with legal and

	ethical standards. Continuous monitoring and strong encryption are essential. As secure underpinning infrastructure is essential for enabling advanced analytics, work towards a security accreditation (e.g., ISO 27001) to provide a secure infrastructure for handling sensitive healthcare data. Key items are: • Security accreditation of the infrastructure, e.g. ISO 27001 • A unified virtual research environment leveraging cloud or national HPC to provide specialised tools to all researchers. • Cutting-edge methods require innovative infrastructure; TREs must adapt to keep pace. • TREs with features such as on-demand GPUs (only when required), LLMs, R-Studio, VS Code, and user-friendly tools like Tableau.
Socio-Tech Assets and Tools/ Models Exemplars	• Guidelines TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) and TRIPOD-AI are standards for reporting the development and validation of prediction models. • HeSANDA (Health Studies Australian National Data Asset) program/PeopleRDC is aimed at creating a searchable repository of health/medical research datasets. • UK's Clinical Practice Research Datalink (CPRD), a comprehensive database of electronic medical records (EMR) from general practitioners (GPs) that has supported numerous impactful research projects • OHDSI (Observational Health Data Sciences and Informatics) is a global community that develops tools and best practices for large-scale analysis of health data. • OMOP (Observational Medical Outcomes Partnership) is a widely adopted common data model used to standardise healthcare data from multiple sources. • Neurodesk is an accessible, flexible and portable data analysis environment for reproducible neuroimaging (reported as useful but complex). • The interim CDC (Centers for Disease Control) may provide insights into establishing an effective data-sharing culture and practices. • Cogniti.ai for creating teaching and learning tools.
Governance	• Establishing consistent data governance frameworks can facilitate easier access to data by standardizing the processes and requirements across different institutions and jurisdictions. Establish a National Data Ethics Committee to standardise ethical guidelines and promote consistent understanding and acceptance of AI technologies across research institutions. • Balancing the application of advanced analytics methods with training the workforce to understand their functionality. • Fail-safe mechanisms for privacy breaches, similar to Good Samaritan laws for medical professionals helping under non-insured context.
Linked data assets	• Harmonised health records and national data repositories.

	• Integrated Electronic Health Records (EHR) Systems (e.g., more accessible records from a range of stakeholders). • Include broad data types (health, infrastructure, land use, satellite data) to track health outcomes based on environmental factors. • Standardise the way that data is captured, collated and modelled. • Chronic disease registry. • Consistent and secure method of transferring large data from hospital to researchers (e.g. X-rays, CT scans). • Centralised storage for large, high-frequency physiological datasets. • Establish a national biospecimen unique identifier system and governance framework to facilitate seamless sharing and utilisation of biospecimens and associated data.
Models	• Establish a National Federated Learning Infrastructure to learn from otherwise fragmented sensitive data, use case EHR. This data network should streamline the setup of new projects by pre-establishing ethics, governance, and IT protocols, with the flexibility to link local detailed datasets as needed for individual projects. • OpenSAFELY model where algorithms are shipped and data stays local; provide framework for data owners to implement the compute capability necessary. Ensures data never leaves the secure environment, mitigating risks of data breaches during transfer. • Establish sovereign foundational model for Australia (or fine-tunings of global LLM to leverage resources). • AI/ML models, including foundational models, could also be employed to support data management and training
Tools	• User-friendly workflows (no-code), and interface (e.g. to use LLMs productively). • National De-Identification & Participant Privacy Toolset(s): toolset and operational guidelines would help to address issues of de-identification and transferring custodianship. Automatic implementation in a TRE would enable greater scalability. • Coding and training platform, similar to Google Colab, that integrates with secure data storage and computation systems, is accessible to all Australian researchers, and linked to a code repository for sharing and collaboration. • Leverage national infrastructure to handle large datasets, such as imaging, by implementing a secure API for on-demand access to advanced computational capabilities ("bursts" of high-end infrastructure). • Private mini-Data Commons to assist researchers in making their project data FAIR during their work. • MRFF (Medical Research Future Fund) and philanthropic projects are reinventing the wheel in efforts to make data FAIR. Such projects are often trans-institutional. A consistent, reusable framework is needed.

	• Develop and implement responsible AI guidelines and tools for model development and validation. • Microscopy/multi-omics unified research infrastructure.
Training	• Introductory ML courses for clinical researchers. • Information on the essential components (e.g., outcome column in data) for launching initiatives in local health contexts.
Lifecycle Stage Practice Support - Problem Formulation to Translation	• There is a need for training and resources to help researchers improve their skills in all stages of the lifecycle including problem formulation, data pre-processing and management, model development and validation and deployment. • Ensure that problem formulation includes responsible AI (e.g. ethical) considerations and is sensitive to the context of the research. • Encourage researchers to engage with stakeholders during the problem formulation phase to ensure that the defined problems are relevant and meaningful. • Consider providing resources including guidelines and expertise to support collaboration. • Establish standard protocols for model validation, including cross-validation, external validation, and continuous performance monitoring. Conduct pilot projects and real-world testing to assess practical utility and robustness. • Ensure responsible AI and good technical practices are encouraged throughout the modelling lifecycle process. Develop tools, processes and platforms that assist researchers. This includes interactive guidelines, automated checks, and integration with co-pilots and examples of well-defined cases from problem to deployment. • Centralised Methodological Hub with resources specific to local context, to support the transition from research to clinical care and reuse prior work effectively. • Infrastructure for bridging the gap between data and problem (e.g. tools and supportive technologies developed for studying dementia could apply to prenatal, but information about key clinical questions does not spread between specialties).

5. Synthesis and Recommendations

5.1. Synthesis

The comprehensive analysis of the environmental scan, survey results, and workshops/interviews reveals a multifaceted landscape for advanced health analytics infrastructure in Australia. The key findings highlight the importance of robust infrastructure, collaborative tools, secure data management, effective governance frameworks and training.

5.1.1. Underpinning Hardware Infrastructure

A robust and scalable hardware infrastructure is essential for advanced analytics workloads. This includes high-performance computing resources, specialised hardware accelerators (such as GPUs and TPUs), and cloud-based platforms for flexible and scalable computing. In the interviews and sectoral dialogues, participants highlighted the challenges in securing sufficient computational resources and the necessity for innovative funding strategies. This includes the ability to dynamically allocate resources such as GPUs and TPUs based on demand. The national initiatives, such as NCI, Pawsey, and ARDC, provide a solid foundation (Section 2.1.5), but there is a clear demand for expanded capacity and accessibility. The survey results show that while many respondents have access to some form of compute resources through their workplace or national initiatives, a significant portion indicated that these resources are not sufficient for their needs (Section 3.1.5). As secure compute infrastructure is essential for enabling advanced analytics, it is important to progress towards a security accreditation (e.g., ISO 27001) of parts of Nectar to provide a secure infrastructure for handling sensitive healthcare data.

5.1.2. AI-integrated Infrastructure

Advancements in AI are transforming health analytics, especially in medical imaging, drug discovery, predictive analytics, and personalised medicine (Section 2.1). There is a need for developing a cohesive ecosystem for utilizing advanced analytics tools and platforms to facilitate analytics. This includes AI and machine learning tools tailored for healthcare applications. The environmental scan revealed existing platforms like BioData Catalyst and Alteryx that support complex data analyses through machine learning and deep learning tools. The compliance with privacy and regulatory requirements varies by platform. The interviews and sectoral dialogues reinforce the importance of advanced AI techniques for handling sensitive data securely and leveraging varied datasets in AI model training (Section 4.1). Also, survey respondents highlighted the need for training and infrastructure to support techniques such as federated learning and other AI methods (Section 3.1.6). Federated learning enables the integration of diverse datasets, improving the robustness and generalisability of AI models (Section 2.1.3).

Notably, Generative AI has the potential to transform health data analytics by enhancing data acquisition, which improves accessibility, and by automating complex analytical processes. Furthermore, it can facilitate improved inference and communication by synthesising diverse data types into clear insights and reports (Section 2.1.3). There is a need to harness the transformative potential of Generative AI.

5.1.3. Data Access and Management

The survey and interviews indicated that developing platforms and tools to facilitate data access and sharing can lower barriers to entry for researchers and organisations seeking to leverage advanced analytics. This may involve data repositories, data sharing agreements, and secure data exchange platforms. Confidence in data quality was moderate, with a need for better documentation and benchmark datasets (Section 3.1.4). Improving data management practices, including the standardisation of data curation, metadata, and access protocols, can significantly enhance the usability and reliability of health data for research purposes.

5.1.4. Trusted Research Environments (TRE)

The analysis presented the need for tools that facilitate modeling, simulation, and data analysis, such as virtual labs and codeless environments, which are essential for enabling researchers to engage in advanced analytics without requiring extensive coding skills. Platforms like BioData Catalyst, Datapine, and Alteryx highlight the potential enhancement to research capabilities (Section 2.2). The development and support of TREs can provide secure, controlled environments where sensitive health data can be analysed while maintaining strict privacy and security standards. Challenges identified include the need for security accreditation (e.g., ISO 27001), high setup costs, and a lack of user-friendly interfaces and on-demand resources like GPUs (Section 4.2.1 and 4.2.2). Survey responses also indicated the need for infrastructure that supports secure data access and sharing among collaborators (Section 3.1.2).

5.1.5. Governance and Standards

Fragmented data governance frameworks and inconsistent policies across jurisdictions and institutions were identified as significant barriers to effective data management and sharing. Harmonising data privacy regulations and developing national standards for data curation, governance, and release are necessary steps to address these challenges (Section 2.3.4). Establishing national-level data governance frameworks and standards can help ensure interoperability, consistency, and quality of data used in advanced analytics. This includes metadata standards, data classification schemes, and data sharing protocols. The OECD's current focus on harmonising health data governance for multi-country projects, as well as cybersecurity and improving global health data interoperability, further emphasises the importance of cohesive and standardised governance frameworks (Section 2.4.1).

5.1.6. National Reference Data Assets

The analysis showed that effective data curation, the development of vocabularies, and the creation of analytic reference datasets, including synthetic data, are foundational elements that support research activities (Section 3.1.6). The creation and maintenance of national reference data assets were highlighted as important for advancing health analytics. This involves not only the development of datasets but also ensuring their quality, accessibility, and interoperability (Section 2.1.2 and 2.1.3). Ensuring that all national reference data assets adhere to the FAIR principles includes developing metadata standards, creating searchable catalogs, and providing tools for easy data discovery and access. The interviews indicated a strong interest in synthetic data, highlighting the need for a clear methodology to address key questions and ensure these data products are suitable for specific research purposes (Section 4.2).

5.1.7. Privacy and Ethics

Data privacy and ethics emerged as major concerns. Researchers expressed concerns about obtaining the necessary authority to access data, understanding their legal rights and responsibilities, and ensuring secure data storage and transfer (Section 3.1.2). The need for a centralised ethics approval framework was highlighted, along with the importance of developing risk-based data management approaches that balance privacy with accessibility. The establishment of a national data ethics committee could play a pivotal role in promoting consistent understanding and acceptance of AI technologies across research institutions (Section 4.2.2).

5.1.8. Training and Guidelines

Investing in education, training, and capacity building programs can help build a skilled workforce capable of utilising advanced analytics techniques effectively. This includes training in data science, machine learning, artificial intelligence, and related fields (Section 3.1.6). Additionally, establishing support systems to help researchers de-identify data and navigate complex data privacy issues was identified as a need.

There was also an emphasis on comprehensive guidelines to support the ethical and effective implementation of advanced analytics. As indicated by the literature, these guidelines could address best practices, data privacy, informed consent, and the responsible use of AI in clinical settings (Section 2.3.5). The development of these guidelines should involve input from a wide range of stakeholders, including researchers, clinicians, ethicists, and policymakers, ensuring they are comprehensive and applicable across various healthcare contexts. Moreover, ongoing education and training programs are necessary to keep healthcare professionals updated on the latest guidelines and best practices (Sections 3.1.2 and 3.1.6). The environmental scan also pointed to the importance of standardised reporting and documentation practices to ensure reproducibility and transparency in research (Section 2.3.2). The need for technical support, as opposed to capacity development, was highlighted (Section 3.1.6). The

environmental scan revealed collaboration in developing infrastructure through joint partnerships, such as those exemplified by GIS, which provide necessary expertise and resources (Section 2.4.4).

5.1.9. Translation and Community of Practice

The analysis highlighted the need for infrastructure that bridges the gap between research methodology and clinical practice, ensuring that research findings are effectively translated into practical healthcare applications. Concerns were raised about the inappropriate usage of advanced analytics, where there can be a mismatch between models and real clinical problems (Section 4.2.2). The interviews highlighted the need for a centralised methodological hub with resources specific to the local context, to support the transition from research to clinical care, ensuring that information about key clinical questions is effectively disseminated (Section 4.2.2).

The survey and interviews highlighted the need for a national community of practice (CoP) to support the implementation of advanced analytics solutions in healthcare. Despite the current lack of widespread participation, the majority rated the need for such CoPs as at least very important (Section 3.1.6). A national CoP could be successful by incentivising researchers based on implementation success rather than just publication outcomes (Section 4.2.2). This shift in focus can drive the practical application of research findings, ensuring that advanced analytics are effectively translated into healthcare improvements.

5.2. Recommendations

The synthesis of the findings from the environmental scan, survey results, and sectoral dialogues has led to the formulation of several key recommendations. These recommendations aim to address the identified gaps and needs in the advanced health analytics infrastructure in Australia:

1. Scalable Hardware/Cloud Resources: Increase investment in scalable, secure (e.g., ISO 27001) and high-performance computing resources, including GPUs and TPUs, to support advanced AI and analytics workloads. Ensure that these resources are accessible on-demand to meet the dynamic needs of researchers.
2. Federated Learning Platforms: Establish national federated learning infrastructure to facilitate the secure analysis of decentralised datasets while maintaining data privacy. This includes developing interoperable systems and governance frameworks to support collaborative AI research.
3. Synthetic Data Generation: Promote the use of synthetic data generation in appropriate contexts to address data privacy concerns and enhance data accessibility. Develop clear methodologies to ensure that synthetic data accurately reflects real-world complexities and is suitable for various research purposes.
4. Foundational models: To harness the potential of Generative AI models like GPT-4 and BERT in health research, robust infrastructure and integration into existing systems are essential.

Specialised training and ethical guidelines are also needed to ensure researchers can use these technologies effectively and responsibly.

5. Standardised Data Curation: Implement national standards for data curation, metadata, and access protocols to improve data quality and usability. Develop searchable portals for health datasets with comprehensive documentation and metadata.
6. Secure Data Environments: Develop and support Trusted Research Environments (TREs) that provide secure, controlled access to sensitive health data. Ensure these environments are user-friendly and equipped with necessary computational resources.
7. National Data Governance Framework: Establish a national data governance framework that harmonises data privacy regulations and standards across jurisdictions. This framework should facilitate data sharing, linkage, and analysis while ensuring compliance with privacy and security requirements.
8. Centralised Ethics Approval: Develop a centralised ethics approval framework to streamline the process of obtaining approvals for research involving sensitive health data. This framework should include provisions for data access, sharing, and AI in research.
9. Advanced Analytics Training Programs: Invest in education and training programs to build a skilled workforce capable of utilising advanced analytics techniques. This includes training in data science, machine learning, artificial intelligence, and related fields.
10. Support Systems for Researchers: Establish support systems to assist researchers in using infrastructure technology and data privacy. Provide resources and guidelines to ensure compliance with data privacy principles and legislation.
11. National Community of Practice: Establish a national community of practice (CoP) to support the implementation of advanced analytics solutions in healthcare. This CoP should incentivise researchers based on implementation success and practical application of research findings.
12. Collaborative Research Initiatives: Foster collaborative research initiatives that bring together academia, industry, and government to leverage shared resources and expertise. Promote international collaborations to enhance the global impact of Australian health research.
13. Risk-based Data Management: Adopt risk-based data management approaches that balance privacy concerns with the need for data accessibility. Develop clear guidelines for informed consent and data sharing to build public trust.
14. National Data Ethics Committee: Establish a national data ethics committee to promote consistent understanding and acceptance of AI technologies across research institutions. This committee should provide oversight and guidance on ethical issues related to data privacy and AI.

15. Methodological Hub: Create a centralised methodological hub with resources specific to the local context to support the translation of research findings into clinical practice. This hub should provide tools and frameworks to bridge the gap between research and healthcare application.
16. Implementation Strategies: Develop frameworks to support the translation of advanced analytics methods into real-world healthcare solutions.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 1-3.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R. (2023) 'Gpt-4 technical report'. arXiv preprint arXiv:2303.08774.

Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020) 'Principled Artificial Intelligence: Mapping Consensus In Ethical and Rights-based Approaches To Principles For AI', SSRN Journal. Available at: https://doi.org/10.2139/ssrn.3518482.

Agustono, D. (2023) 'Artificial Intelligence in Human Resource Management Practices', KSS. Available at: https://doi.org/10.18502/kss.v8i9.13409.

Alagumariappan, P., Dewan, N., Muthukrishnan, G., Raju, B., Bilal, R. and Vijayalakshmi, S. (2020) 'Intelligent Plant Disease Identification System Using Machine Learning'. Available at: https://doi.org/10.3390/ecsa-7-08160.

Arora, A., Belenzon, S., Cioaca, L.C., Sheer, L. and Zhang, H. (2023) 'The Effect of Public Science on Corporate R&D' (No. w31899). National Bureau of Economic Research.

Aung, Y., Wong, D. and Ting, D. (2021) 'The Promise Of Artificial Intelligence: A Review Of The Opportunities And Challenges Of Artificial Intelligence In Healthcare', British Medical Bulletin, 1(139), pp. 4-15. Available at: https://doi.org/10.1093/bmb/ldab016.

Australian Government, Department of Education. (2021) 2021 'National Research Infrastructure Roadmap'. Available at: https://www.education.gov.au/national-research-infrastructure/2021-national-research-infrastructure-roadmap.

Baird, A. and Schuller, B. (2020) 'Considerations for a More Ethical Approach to Data in AI: On Data Representation and Infrastructure', Front. Big Data, 3. Available at: https://doi.org/10.3389/fdata.2020.00025.

Bak, M., Madai, V. I., Celi, L. A., Kaissis, G. A., Cornet, R., Maris, M., ... & McLennan, S. (2024) 'Federated learning is not a cure-all for data ethics', Nature Machine Intelligence, pp. 1-3.

Baker, M. (2017) 'Scientific computing: Code alert', Nature, 541(7638), pp. 563-565.

Barone, L., Williams, J. and Micklos, D. (2017) 'Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators', PLoS Computational Biology, 13(10), p. e1005755.

Berente, N., Howison, J., King, J. L., Ahalt, S. and Winter, S. (2018) 'Organizing and the cyberinfrastructure workforce'. Available at: https://ssrn.com/abstract=3260715).

Briganti, G. (2023). 'A doctor's guide to foundation models'. Available at: https://doi.org/10.31219/osf.io/5zg3q

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D. and Nath, V., 2022. 'Monai: An open-source framework for deep learning in healthcare'. arXiv preprint arXiv:2211.02701.

Chaudhry, S., Pazouki, A., Schmitz, P., Hillery, E., & Kee, K. (2022). 'Understanding Factors that Influence Research Computing and Data Careers'. Practice and Experience in Advanced Research Computing. https://doi.org/10.1145/3491418.3530292

Crowson, M. G., Moukheiber, D., Arévalo, A. R., Lam, B. D., Mantena, S., Rana, A., ... & Celi, L. A. (2022) 'A systematic review of federated learning applications for biomedical data', PLOS Digital Health, 1(5), p. e0000033.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2018) 'Bert: Pre-training of deep bidirectional transformers for language understanding'. arXiv preprint arXiv:1810.04805.

Eppler, M. (2023) 'The Benefits and Dangers of Artificial Intelligence in Healthcare Research Writing', UTJ, 1(7), pp. 01-02. Available at: https://doi.org/10.31491/utj.2023.03.006.

Gabriel A. Brat, Joshua C. Mandel, & Matthew B.A. McDermott (2024). 'Do We Need Data Standards in the Era of Large Language Models?'. NEJM AI, 1(8), AIe2400548.

Gonzalez, M., Capman, J., Oswald, F., Theys, E. and Tomczak, D. (2019) '"Where's the I-o?" Artificial Intelligence And Machine Learning In Talent Management Systems', PAD, 3(5). Available at: https://doi.org/10.25035/pad.2019.03.005.

González-García, J., Estupiñán-Romero, F., Tellería-Orriols, C., González-Galindo, J., Palmieri, L., Fagaralli, A., Pristās, I., Vuković, J., Misinš, J., Zile, I. and Bernal-Delgado, E. (2021) 'Coping with interoperability in the development of a federated research infrastructure: achievements, challenges and recommendations from the JA-InfAct', Archives of Public Health, 79, pp. 1-18.

Guan, J. (2019) 'Artificial Intelligence In Healthcare and Medicine: Promises, Ethical Challenges, And Governance', Chinese Medical Sciences Journal, 0(0), p. 99. Available at: https://doi.org/10.24920/003611.

Gunasekeran, D., Tseng, R., Tham, Y., & Wong, T. (2021). 'Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies'. NPJ Digital Medicine, 4. https://doi.org/10.1038/s41746-021-00412-9

Hampton, S.E., Jones, M.B., Wasser, L.A., Schildhauer, M.P., Supp, S.R., Brun, J., Hernandez, R.R., Boettiger, C., Collins, S.L., Gross, L.J. and Fernández, D.S. (2017) 'Skills and knowledge for data-intensive environmental research'. BioScience, 67(6), pp.546-557. Available at: https://doi.org/10.1093/biosci/bix025

Ijiga, A. (2024). 'Ethical considerations in implementing generative ai for healthcare supply chain optimization: A cross-country analysis across India, the United Kingdom, and the United States of America'. International Journal of Biological and Pharmaceutical Sciences Archive, 7(1), 048-063. Available at: https://doi.org/10.53771/ijbpsa.2024.7.1.0015

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A. and Bridgland, A., 2021. 'Highly accurate protein structure prediction with AlphaFold'. nature, 596(7873), pp.583-589.

Knawy, B., McKillop, M., Abduljawad, J., Tarkoma, S., Adil, M., Schaper, L., Chee, A., Bates, D., Klag, M., Lee, U., Kozlakidis, Z., Crooks, G., & Rhee, K. (2022). 'Successfully Implementing Digital Health to Ensure Future Global Health Security During Pandemics: A Consensus Statement'. JAMA network open, 5 2, e220214 . https://doi.org/10.1001/jamanetworkopen.2022.0214

Kourou, K., Exarchos, K., Papaloukas, C., Sakaloglou, P., Exarchos, T. and Fotiadis, D. (2021) 'Applied Machine Learning in Cancer Research: a Systematic Review for Patient Diagnosis, Classification and Prognosis', Computational and Structural Biotechnology Journal, 19, pp. 5546-5555. Available at: https://doi.org/10.1016/j.csbj.2021.10.006.

Kuleto, V., Ilić, M., Dumangiu, M., Ranković, M., Martins, O., Păun, D. and Mihoreanu, L. (2021) 'Exploring Opportunities and Challenges of Artificial Intelligence and Machine Learning in Higher Education Institutions', Sustainability, 18(13), p. 10424. Available at: https://doi.org/10.3390/su131810424.

Leung, R. (2023) 'Using Ai–ml To Augment the Capabilities of Social Media For Telehealth And Remote Patient Monitoring', Healthcare, 12(11), p. 1704. Available at: https://doi.org/10.3390/healthcare11121704.

Levine, A., et al. (2020). 'Synthesis of diagnostic quality cancer pathology images by generative adversarial networks'. The Journal of pathology, 252(2), 178-188. https://doi.org/10.1002/path.5509

Lu, Y., Wang, H., & Wei, W. (2023). 'Machine Learning for Synthetic Data Generation: a Review'. ArXiv, abs/2302.04062. https://doi.org/10.48550/arXiv.2302.04062

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H. and Liu, T.Y., 2022. 'BioGPT: generative pre-trained transformer for biomedical text generation and mining'. Briefings in bioinformatics, 23(6)

Malsia, E. (2024). 'Generative artificial intelligence in health system management: transformative insights'. Journal of Service Science and Management, 17(02), 107-117. Available at: https://doi.org/10.4236/jssm.2024.172005

Mckay, F., Williams, B., Prestwich, G., Bansal, D., Hallowell, N. and Treanor, D. (2022) 'The Ethical Challenges Of Artificial Intelligence-driven Digital Pathology', The Journal of Pathology CR, 3(8), pp. 209-216. Available at: https://doi.org/10.1002/cjp2.263.

Modiba, M., Ngulube, P. and Marutha, N. (2023) 'Infrastructure For the Implementation Of Artificial Intelligence To Support Records Management At The Council For Scientific And Industrial Research In South Africa', Esarbica J, 41, pp. 159-171. Available at: https://doi.org/10.4314/esarjo.v41i.11.

Muhamedyev, R., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A. and Yelis, M. (2022) 'Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges', Mathematics, 15(10), p. 2552. Available at: https://doi.org/10.3390/math10152552.

National Academies of Sciences, Engineering, and Medicine (2024a) 'Toward a 21st Century National Data Infrastructure: Managing Privacy and Confidentiality Risks with Blended Data'. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/27335.

National Academies of Sciences, Engineering, and Medicine (2024b) 'Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources'. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/26804.

Neretin, O. and Kharchenko, V . (2022) 'Ensurance of Artificial Intelligence Systems Cyber Security: Analysis of Vulnerabilities, Attacks and Countermeasures', Vìsnik Nacìonal'nogo unìversitetu "L'vìvs' ka polìtehnìka". Serìâ Ìnformacìjnì sistemi ta merežì, 12, pp. 7-22. Available at: https://doi.org/10.23939/sisn2022.12.007.

Netherlands eScience Center (2024) 'Netherlands eScience Center Technology Forecast 2024', Zenodo. doi: 10.5281/zenodo.10635719.

Novak, L. (2023) 'Clinical Use Of Artificial Intelligence Requires Ai-capable Organizations', Jamia Open, 2(6). Available at: https://doi.org/10.1093/jamiaopen/ooad028.

OECD (2019) Enhancing access to and sharing of data: reconciling risks and benefits for data re-use across societies. OECD Publishing.

Olczak, J., Pavlopoulos, J., Prijs, J., IJpma, F., Doornberg, J., Lundström, C. and Gordon, M. (2021) 'Presenting Artificial Intelligence, Deep Learning, and Machine Learning Studies to Clinicians and Healthcare Stakeholders: An Introductory Reference with a Guideline and a Clinical AI Research (CAIR) Checklist Proposal', Acta Orthopaedica, 5(92), pp. 513-525. Available at: https://doi.org/10.1080/17453674.2021.1918389.

Pethani, F. (2021) 'Promises and Perils of Artificial Intelligence in Dentistry', Aust Dent J, 2(66), pp. 124-135. Available at: https://doi.org/10.1111/adj.12812.

Precedence Research (2022) 'North America Digital Health Market Size to Surpass US$ 151.88 Bn by 2027', GlobeNewswire, 9 February. Available at: https://www.globenewswire.com/en/news-release/2022/02/09/2382062/0/en/North-America-Digital-Health-Market-Size-to-Surpass-US-151-88-Bn-by-2027.html.

Ramim, M. and Hueca, A. (2021) 'Cybersecurity Capacity Building Of Human Capital: Nations Supporting Nations', OJAKM, 2(9), pp. 65-85. Available at: https://doi.org/10.36965/ojakm.2021.9(2)65-85.

Salehjahromi, M. et al. (2024) 'Synthetic PET from CT improves diagnosis and prognosis for lung cancer: Proof of concept', Cell Reports Medicine, 5(3), p. 101463. Available at: https://doi.org/10.1016/j.xcrm.2024.101463.

Schmitz, P. (2021) 'Advancing the workforce that supports computationally and data intensive research', Computing in Science and Engineering. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9492830.

Simhadri, N. (2023) 'Awareness Among Teaching on AI and ML Applications Based on Fuzzy in Education Sector at USA', Soft Computing. Available at: https://doi.org/10.1007/s00500-023-08329-z.

Sivathanu, B. and Pillai, R. (2018) 'Smart HR 4.0 – How Industry 4.0 Is Disrupting HR', HRMID, 4(26), pp. 7-11. Available at: https://doi.org/10.1108/hrmid-04-2018-0059.

Srivastava, P. and Grosel, C. (2024) 'Four keys to successful digital transformation in healthcare', Innosight, February. Available at: https://www.innosight.com/insight/digital-transformation-in-healthcare.

The Economist (2024) 'Universities are failing to boost economic growth', The Economist, 5 February. Available at: https://www.economist.com/finance-and-economics/2024/02/05/universities-are-failing-to-boost-economic-growth.

Velev, D. (2023) 'Challenges of Artificial Intelligence Application for Disaster Risk Management', Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-M-1-2023, pp. 387-394. Available at: https://doi.org/10.5194/isprs-archives-xlviii-m-1-2023-387-2023.

Wilkinson et al. (2016). 'The FAIR Guiding Principles for scientific data management and stewardship'. Scientific Data. 3 (1) doi:10.1038/SDATA.2016.18

Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., … & Shah, N. (2023). 'The shaky foundations of large language models and foundation models for electronic health records'. NPJ Digital Medicine, 6(1). Available at: https://doi.org/10.1038/s41746-023-00879-8

Yoon, J., Drumright, L., & Schaar, M. (2020). 'Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN)'. IEEE Journal of Biomedical and Health Informatics, 24, 2378-2388. https://doi.org/10.1109/JBHI.2020.2980262

Zhang, Z., Li, G., Xu, Y. and Tang, X. (2021) 'Application of Artificial Intelligence in the MRI Classification Task of Human Brain Neurological and Psychiatric Diseases: A Scoping Review', Diagnostics, 8(11), p. 1402. Available at: https://doi.org/10.3390/diagnostics11081402.

Appendix

Appendix 1. People RDC National Priority Areas and Health Research Funding Priorities

National priority areas

The national priorities related to health research and supporting infrastructure are addressed in the following national strategies:

- 2021 National Research Infrastructure Roadmap Exposure[41]
- National Medical Research & Innovation Strategy[42]
- National Science & Research Priorities[43]
- National Climate Resilience & Adaptation Strategy[44]
- Blueprint for Critical Technologies[45] and The Action Plan for Critical Technologies[46]

Looking across these high-level strategic documents, a number of areas of focus can be identified. These are summarised below:

- Development of biologics and medical devices
- Pre-clinical and clinical research
- Diverse national health datasets and integrated digital platforms
- Enabling research and research translation to support better models of health care and services for all sections of the community
- Cutting-edge treatments with genomics and genetic engineering

Health research funding priorities

National health research is largely funded through NHMRC and MRFF.

The NHMRC research priorities[47] are dementia, mental health and health impacts of environmental change.

[41] https://www.education.gov.au/national-research-infrastructure/2021-national-research-infrastructure-roadmap
[42] https://www.health.gov.au/sites/default/files/documents/2021/11/australian-medical-research-and-innovation-strategy-2021-2026.pdf
[43] https://www.industry.gov.au/sites/default/files/2018-10/science_and_research_priorities_2015.pdf?acsf_files_redirect
[44] https://www.awe.gov.au/sites/default/files/documents/national-climate-resilience-and-adaptation-strategy.pdf
[45] https://www.pmc.gov.au/sites/default/files/publications/ctpco-blueprint-critical-technology.pdf
[46] https://www.pmc.gov.au/sites/default/files/publications/ctpco-action-plan-for-critical-technology-amalgamated.pdf
[47] https://www.nhmrc.gov.au/research-policy/research-priorities

MRFF research Priorities outlined in the Australian Medical Research and Innovation Priorities 2021-26[48] highlight the importance of:

- Research and research translation that is multidisciplinary, cross-sector and cross-jurisdiction on a national scale
- Enabling research and research translation in primary care settings and clinicians in the health care system; and indigenous researchers
- Bringing together diverse datasets from priority populations - through data linkage along with data storage and analytics
- Data platforms, Applied AI, novel decision tools and end-user digital utility

[48] https://www.health.gov.au/sites/default/files/documents/2021/11/draft-australian-medical-research-and-innovation-priorities.pdf

Appendix 2. Detailed Analysis of Selected Platforms

A.2.1. NHLBI BioData Catalyst® (BDC)

Introduction

BioData Catalyst (BDC)[49] is a cloud-based ecosystem funded by the National Heart, Lung, and Blood Institute (NHLBI) in the U.S. that aims to accelerate scientific discoveries by providing researchers access to data, advanced analytical tools, and computational resources. BDC offers data, analytic tools, and workflows in a secure environment for researchers to access, share, and analyse scientific data.

As one of NHLBI's data repositories, this platform allows researchers to share data from NHLBI-funded research, enabling reproducible research and reuse of data for further scientific advancement. BDC focuses on data related to heart, lung, blood, and sleep conditions and supports scientific studies aimed at understanding, preventing, diagnosing, and treating diseases in these areas.

Features and Capabilities

Data Integration and Discovery: Gen3 is a data platform for building data commons that enables approved researchers and partner organisations to search for harmonised genomic and phenotypic datasets, and export selected data to analytical workspaces. It includes tools that allow researchers to standardise and combine data from different sources, ensuring consistency and compatibility. The Patient Information Commons Standard Unification of Research Elements (PIC-SURE) user interface gives investigators the ability to search available data and conduct feasibility queries, allowing for data cohorts to be built and visualised in real-time. BDC also includes a metadata catalogue that helps users find and understand the data available within the ecosystem.

Model Building and AI Capabilities: BDC provides a suite of analytical tools that support complex data analyses, including statistical modelling, machine learning, and AI. Researchers can build, train, and deploy machine learning models using pre-configured workflows and customisable pipelines. The platform leverages cloud computing to provide scalable computational resources, allowing users to perform intensive data processing and model training without local infrastructure limitations. Furthermore, collaborative workspaces powered by Seven Bridges and Terra enable large-scale genomic data analysis. Researchers can utilise pre-built and pre-loaded workflows or import their own data and models, using tools like RStudio and Jupyter Notebooks for interactive analysis.

Access and Cost

Access to hosted controlled data on the BDC ecosystem is managed through NIH Database of Genotypes and Phenotypes (dbGaP) approvals. Users must have dbGaP approval to access specific studies and log in to BDC platforms using eRA Commons credentials, authenticated by iTrust. BDC adheres to internationally recognised data access and release policies, ensuring broad access while restricting

[49] https://biodatacatalyst.nhlbi.nih.gov/

controlled data to authorised users. Users are responsible for complying with Data Use Agreements, Institutional Review Board policies, and other Data Use Limitations when uploading or downloading data on the BDC ecosystem.

BDC hosts a number of datasets available for analysis to users with appropriate data access approvals. Users are not charged for the storage of these hosted datasets; however, if hosted data is used in analyses, users incur costs for computation and storage of derived results.

Security and Compliance

All data within the BDC is de-identified, with additional access restrictions to ensure that only authorised users can reach sensitive information. Data authorisations are overseen by Data Access Committees. Data transfers are secured using TLS encryption, and users must log in with their eRA accounts to access or download data. NHLBI BioData Catalyst® (BDC) adheres to standards and laws including HIPAA (Health Insurance Portability and Accountability Act), FISMA (Federal Information Security Management Act), and NIST (National Institute of Standards and Technology), ensuring compliance within their respective jurisdictions.

User Support and Community

The BDC is a continually evolving ecosystem that invites community contributions through various methods such as submitting resources, providing feedback via the contact form, and responding to Requests for Comment (RFC). BDC offers extensive support resources, including documentation, tutorials, and user guides to help researchers navigate and utilise the platform effectively.

The target audience for the BDC includes a broad range of researchers and data scientists who are focused on biomedical research, particularly in the areas of heart, lung, blood, and sleep disorders. The platform is specifically designed for those who need access to large datasets and computational tools to conduct their research. The BDC Fellows Program offers the opportunity for early-career researchers to deepen their engagement with cutting-edge data science and biomedical research. These fellowships support research projects that leverage the BDC data ecosystem and computational resources.

A.2.2. Datapine

Introduction

Datapine[50] is a private company that offers business intelligence (BI) software for data analytics and visualisation. Datapine provides a platform that focuses on business intelligence by enabling the collection and analysis of large-scale data across various aspects of healthcare such as costs, pharmaceuticals, clinical data, and patient behaviour. The aim is to drive improvements in medicine, patient care, research, and education within the healthcare industry.

[50] https://www.Datapine.com/healthcare-analytics

Features and Capabilities

The main features of Datapine are the modern healthcare dashboard and healthcare analytics. The scope of healthcare analysis includes financial planning, evaluating hospital performance, enhancing patient satisfaction, improving communication through real-time access to patient data and medical history, and forecasting for effective healthcare facility management.

Datapine integrates data from various sources, such as Electronic Health Records, patient management systems, financial databases, and more, using a variety of custom data connectors. This configuration centralises information in a single access point, allowing for streamlined exploration through an intuitive drag-and-drop interface. Datapine includes tools for data cleaning, transformation and preparation to ensure that data is accurate and ready for analysis. Users can create customisable, interactive dashboards to visualise complex healthcare data intuitively. The platform also allows for the automation of reporting processes, ensuring that stakeholders receive regular, up-to-date reports without manual intervention. See Figure A.2.1 below for data visualisations with the dashboard.

Figure A.2.1 Datapine dashboard.

Access and Cost

The service operates on a flexible pricing model, allowing customers to select from a range of subscription plans. Access is managed within the interface by manually granting users permission to view data, dashboards, and reports.

Security and Compliance

As a German engineering company, Datapine is subject to stringent global data security and privacy laws. It ensures customers retain sole ownership of their data, limiting access strictly to authenticated account users. The company's data centers feature robust physical security measures including electronic access controls and video surveillance. Data transmission is secured via SSL and SSH protocols, with added protections against SQL injections. Datapine hosts its services on dedicated servers with a German provider, avoiding cloud solutions, and only stores customer data with explicit permission. Security practices include comprehensive testing and monitoring for vulnerabilities, conducting threat assessments, penetration testing, and regular security reviews to ensure robust protection against potential threats. Considerations regarding legal implications arise when considering the use of Datapine (hosted in Germany) for handling personal data in Australia, highlighting potential issues and solutions in cross-border data management. Datapine does not explicitly advertise or state compliance with HIPAA, NIST or FISMA on its website or in its official documentation.

User Support and Community

Datapine provides comprehensive documentation online detailing its features and step-by-step tutorials on creating interactive dashboards and automated reports. The company also offers video tutorials covering how to connect, explore, analyse, present, and share data effectively. Additional resources include articles tailored for business owners, departmental managers, and analysts. Personal consulting and coaching services are also available to support users from the implementation phase through to the practical use of the software. The target audience for Datapine includes business owners, departmental managers, and analysts across a wide range of industries.

A.2.3. Alteryx

Introduction

Alteryx[51] provides a comprehensive data analytics platform tailored to the needs of healthcare organisations. It is designed to enhance decision-making and streamline data integration, preparation and analytics in the healthcare sector. The platform enables users to derive actionable insights that can improve patient outcomes, operational efficiency, and compliance. Alteryx is designed to make advanced analytics accessible to healthcare professionals without the need for coding or extensive technical expertise.

[51] https://www.alteryx.com/solutions/industry/healthcare

Features and Capabilities

The Alteryx platform provides a suite of enterprise analytics tools, supporting processes for data integration and preparation, alongside advanced capabilities for predictive and prescriptive analytics, geospatial analysis, and AutoML[52]. Alteryx provides various solutions tailored to different needs and environments. The Designer Cloud enables data transformation and analytics directly in the cloud, while the on-premises Designer offers drag-and-drop analytics solutions. The Intelligence Suite and Machine Learning tools offer predictive modelling and text mining capabilities. Additionally, the platform supports scaling and automation of reporting, with collaborative tools for data governance and workflow management. Specialised functionalities include the Intelligence Suite for insights from unstructured data and FIPS-compliant (Federal Information Processing Standards) versions for secure environments.

Case study on healthcare pricing and revenue analysis: The Healthcare Corporation of America (HCPA) used the Alteryx Designer to clean data from diverse sources, adjust for pricing changes over a four-year period, and evaluate impacts to revenue such as growth, seasonality and COVID. Key to their success was the Designer's fuzzy matching capabilities, which effectively resolved mis-mapped revenue data across the health system. The integration of Alteryx with Snowflake and Tableau enabled real-time insights, enhancing the client's ability to monitor revenue and make timely adjustments.

Case study on public health data management: The Washington State Department of Health faced increased demands to accelerate its analytics capabilities to the cloud during the COVID-19 pandemic. The scale of incoming data from outdated systems overwhelmed their existing processes. Alteryx Designer Cloud, implemented within their CEDAR (Cloud Environment for Data Analytics and Reporting) platform on Microsoft Azure, allowed data scientists to rapidly prepare and standardise data for analysis, significantly shortening the time to insight and reducing duplicative work.

Access and Cost

Alteryx is a paid service with access managed through its interface, featuring a granular permissions model for security. The Administration Portal enables monitoring of usage reports, auditing of data updates, tracking of workflow runs, user activity, and data lineage.

Security and Compliance

Alteryx ensures data security with encryption both at rest and in transit. It adheres to industry-standard frameworks such as the US National Institute of Standards and Technology Cybersecurity Framework (NIST CF) and a Security Incident Response Team (SIRT), aimed at protecting information assets against unauthorised access and incidents. Alteryx's global privacy program ensures compliance with data protection laws like the EU General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), addressing the privacy of user and customer data through rigorous practices and controls. Additionally, Alteryx's analytics and server environments are compatible with Federal Information Processing Standards (FIPS), meeting the security requirements as outlined by the National

[52] https://www.alteryx.com/products/alteryx-machine-learning

Institute of Standards and Technology (NIST) and in accordance with the Federal Information Security Management Act (FISMA). In the official documentation it states that customers must not submit any patient, medical, or other protected health information regulated by HIPAA or similar laws. Alteryx holds no liability for such health information, despite any other provisions in the agreement.

User Support and Community

Alteryx provides a comprehensive community hub that includes a variety of resources to enhance learning and engagement. The platform offers a range of micro-certifications and full certifications for products such as Alteryx Designer Desktop, Alteryx Cloud, and Alteryx Server. Users can participate in digital learning opportunities that feature weekly challenges and guided curriculum learning paths, join online groups and attend events like summits. Additionally, the hub offers extensive resources like videos, interactive lessons on data science from beginner to advanced levels, and e-books on topics such as modern analytics, generative AI, and data quality. Alteryx provides access to open-source libraries hosted on GitHub, including Woodwork, Compose, Featuretools, and EvalML. Alteryx also offers branded merchandise for its community members.

The target audience for Alteryx includes a wide range of professionals, offering tools for both technical users engaged in advanced analytics and non-technical users who require easy access to data analysis without programming skills.

A.2.4. UK biobank

Introduction

The UK Biobank[53] is a comprehensive biomedical database and research resource containing de-identified genetic, lifestyle, health information, and biological samples from half a million UK participants. It is the most widely used dataset of its kind and is globally accessible to approved researchers from academic, commercial, government, or charitable settings who are conducting health-related research in the public interest. UK Biobank is advancing modern medicine by enhancing the understanding of the prevention, diagnosis, and treatment of various serious and life-threatening illnesses, including cancer, heart disease, and stroke.

Features and Capabilities

The Research Analysis Platform (RAP), enabled by DNAnexus technology and powered by Amazon Web Services (AWS), is designed to accommodate the vast and growing scale of the UK Biobank resource. This platform allows researchers to access the dataset in the cloud at no cost, although there are associated costs for compute, data storage for analyses, and data egress of permitted data. RAP includes a variety of preconfigured tools for data processing, statistical analysis, and machine learning, such as JupyterLab for Python, RStudio Workbench, and more. Researchers can create and customise their own workflows using a range of programming languages and software tools available on the platform. It also provides

[53] https://www.ukbiobank.ac.uk/

tools for creating interactive dashboards and data visualisations to help interpret and present research findings effectively.

The UK Biobank collects extensive data on 500,000 participants, including environmental, lifestyle, health and genetic information. This data encompasses brain, heart, and full-body imaging; whole genome and exome sequencing; linkage to electronic health records; biomarkers; physical activity and accelerometer data; and online questionnaires covering diet, work history, cognitive function, and mental health. Blood, urine, and saliva samples are also collected, enabling comprehensive health and well-being research.

Access and Cost

To access the UK Biobank database, registered researchers must apply via the Access Management System (AMS), providing a summary of their research, the required data-fields, and a description of any new data or variables generated. The application process includes steps for adding collaborators, obtaining necessary approvals, and paying access fees. The fee structure is designed to cover only the incremental costs of servicing access applications. UK Biobank offers a reduced access fee for students and researchers from low and middle-income countries.

Security and Compliance

Trained personnel conduct background checks on researchers applying for access, ensuring they have a professional history of high-quality health-related research and work for a legitimate organisation. Researchers are also screened against international sanctions lists. In-house scientists evaluate whether the research proposal benefits public health without causing harm. If there are concerns, the proposal is reviewed by UK Biobank's expert Access Committee, which may seek ethical advice.

Data security measures comply with the UK General Data Protection Regulation (GDPR), requiring researchers to store, process, and use data securely, with restricted access. The cloud-based Research Analysis Platform (RAP) allows secure access and analysis without the need to download data, with specific large datasets restricted to the RAP to enhance security.

User Support and Community

An online community forum supports researchers using the platform by providing a space to ask questions, share experiences, and receive support from the UK Biobank and DNAnexus teams. The forum aims to facilitate broad and diverse health research by encouraging knowledge sharing and collaboration among researchers. Extensive documentation and user guides are available to help researchers understand the dataset and navigate the resources.

The target audience for UK Biobank includes researchers, scientists, and healthcare professionals from academia, industry, and charitable organisations involved in health-related research and data analysis, aiming to advance scientific discoveries, improve public health, and enhance medical research.

A.2.5. Database of Genotypes and Phenotypes (dbGaP)

Introduction

The Database of Genotypes and Phenotypes (dbGaP)[54] is a repository sponsored and managed by the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH) (US). The database was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

Features and Capabilities

Information in dbGaP is organised hierarchically, including accessioned objects, phenotypes, molecular assay data, analyses, and documents. Public metadata, summary data, and study-related documents are freely accessible on the dbGaP website, while individual-level data requires a Controlled Access application. Scientists worldwide can access public data and apply for controlled access by demonstrating research objectives and data protection capabilities.

The FTP site includes a directory for each study, with subdirectories for every version and analyses. Each study version contains directories for documents, phenotype variable summaries, manifests, and release notes. Manifests describe available files by consent category, while release notes detail file histories and changes. Variable summaries and data dictionaries are provided as XML files with an accompanying XSL file for browser viewing. The documents directory contains .zip files with XML files, images, and PDFs, which may be separated into multiple .zip files if numerous.

Access and Cost

Publicly accessible metadata, summary level data, and documents related to studies can be accessed via the dbGaP website. Individual-level data are accessible to Senior/NIH Investigators via a Controlled Access application. Staff and trainees such as graduate students and postdoctoral fellows are not permitted to submit for access requests. Funding for the open access charge was provided by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Security and Compliance

Data distribution by dbGaP follows NIH policies for managing Genome-Wide Association Studies (GWAS) data. Individual-level data is available only to authorised users via the dbGaP Authorised Access System (dbGaP-AA), which handles request submissions, reviews, approvals, and secure data downloads. Data is grouped by consent groups, each with distinct Data Use Limitations (DULs), and requests must match these DULs. A research use statement explaining how the research objectives conform to the DULs is required for access and is publicly posted for transparency. Each data file has an embargo release date, and results cannot be published before this date. Non-NIH users need an NIH eRA Commons account with a Principal Investigator role to access the system. The dbGaP is designed to comply with relevant data protection and security standards (HIPAA, FISMA, NIST).

[54] http://www.ncbi.nlm.nih.gov/gap

User Support and Community

Limited resources are available online, including guides on how to submit and Frequently Asked Questions. The target audience includes researchers, scientists, and data analysts who are seeking to submit or access health and genomics data.

Appendix 3. Survey Questions

The following questions are about you, your education and current employment.

1. What is your highest qualification? (Bachelor degree - Postgraduate diploma/certificate - Masters - PhD/MD - Other)

2. In which field is your highest qualification? (Mathematics/Statistics - Computing/IT - Health/Biomedical - Other)

3. What is your current primary position? (Manager/Director/Executive/Other leadership role - Analyst/Professional/Technical officer - Academic Researcher - Student (Masters/PhD) - Other)

4. In which industry do you currently work? (select all that apply; University - Medical research - Government department/agency - Biomed/Biotech industry - Other)

5. In your organisation/work: (select all that apply; All data used is internal - Data from other sources used - Computing resources are co-located with the data - Data needs to be moved to the compute location - None of the above)

6. Rate your analytics and/or programming skills below: (None - Some - Good - Very good - Excellent)

- 6.1. Applied statistics/mathematics
- 6.2. Artificial Intelligence (AI) and Machine Learning (ML)
- 6.3. R/Python programming
- 6.4. Cloud Computing/HPC environments
- 6.5. Data Engineering/Databases

The following questions focus on the issues surrounding ethics (often required to access health data) and data privacy (i.e. how to keep individual level data safe and secure when accessing and utilising them).

7. How highly do you rate data privacy/ethics as a concern? (Not at all important - Slightly important - Moderately important - Very important - Extremely important)

8. What do you see as the main issue(s) around privacy/ethics? (select all that apply; Obtaining authority to access relevant data - Understanding legal rights and responsibilities - Secure data storage - Transferring data (secure/reliable/available) - Other)

9. What are the current approach(es) in your organisation? (select all that apply; Secure centrally maintained data repository (read-only access) - Use of synthetic data - Data encryption (SFTP/MFT) - Cloud file sharing - Research Data Champions/central contact point for help - Other)

10. In an ideal world, what do you think needs to happen with respect to data privacy? (open text field)

The next few questions are concerned with version control, i.e. different versions/updates of software, hardware and data, and the potential impact on analyses.

11. How highly do you rate version control as a concern? (Not at all important - Slightly important - Moderately important - Very important - Extremely important)

12. What are the current approaches used by you/your organisation? (select all that apply; GitHub/GitLab/BitBucket/etc - AWS CodeCommit/AzureDevOps/etc - Data Version Control software - Metadata - Filenaming conventions - Other)

13. What tools or resources would help in this area? (select all that apply; Minimum standards for metadata (e.g. MIAME/MIASE) - Compatible protocols - Uniform use of program(s) - Other)

The following questions focus on the quality and integrity of health data that you use or would like to use in your analytics projects, as well as the availability of such data.

14. How highly do you rate data integrity and quality as a concern? (Not at all important - Slightly important - Moderately important - Very important - Extremely important)

15. How confident are you in the quality and integrity of the data you typically use/access? (None at all - A little - A moderate amount - A lot - A great deal)

16. In your experience, do you think the documentation available is detailed enough for you to feel confident about using a dataset? (Yes - No)

16.1. Since you selected "no" to the previous question, what information do you find is often missing? (select all that apply; Variable definitions - Population/sample details - Information about missing data/observations - Metadata - Other)

17. Would the availability of benchmark and/or synthetic datasets be of use to you/your organisation? (Yes. Please describe the type of dataset desired - No)

18. Do you currently access data behind a paywall? (Yes - No)

19. If you needed to pay to access your ideal dataset, would you: (Yes - Maybe - No)

19.1. Search for a different, freely available (but less ideal) dataset instead

19.2. Have no problem with paying for access

19.3. Pay for the access but funding would be tight

19.4. Need to abandon the project due to lack of funds

The following questions are concerned with the computing resources that you access in order to carry out your analyses. These compute resources may include in-house or external computing clusters, cloud computing or access to supercomputers through HPC providers (commercial and public).

20. Does your workplace provide access to the compute resources required for your work? (Yes and it is sufficient - Yes but not sufficient - No)

21. Do you access or use any of the following computing facilities? (select all that apply; NCI Gadi - Pawsey Setonix - Pawsey Nimbus cloud - ARDC NeCTAR cloud - MLeRP (GPU based cloud) - DUG - Amazon Web Services (AWS) - Google - Azure - Other - None of the above)

22. How do you pay for access to commercial providers? (select all that apply; Industry funds - Research grant funds (ARC/NHMRC/MRFF etc) - Institutional agreements - Other - I don't pay or I don't know)

The final section of questions relates to skills and training in advanced analytics, including machine learning, artificial intelligence and other complex statistical/mathematical models, as well as other specialist technical expertise.

23. How often do you use the following techniques on health data? (Never - Sometimes - About half the time - Most of the time - Always)

- 23.1. Database queries/extracting data
- 23.2. Exploratory data analysis
- 23.3. Standard analysis (e.g. linear regression)
- 23.4. AI/ML
- 23.5. Deep learning/LLMs

24. Does your organisation have sufficient advanced analytics capabilities in-house? (Definitely not - Probably not - Might or might not - Probably yes - Definitely yes)

24.1. If no, what area(s) of advanced analytics are needed? (Cyber security and encryption - Distributed machine learning (e.g. federated learning) - AI/ML expertise - Secure Research Environments (SRE) - Trusted Research Environments (TREs) - Other)

25. Does your organisation provide access to formal training in advanced analytics? (Yes - No)

25.1. If yes, is this training sufficient for your needs? (Yes - No)

26. In which areas would you like training to be available? (select all appropriate; Data encryption - Artificial Intelligence/Machine Learning methods - Advanced (Bio)Statistics methods - Specialized mathematical models (e.g. agent-based modelling, systems dynamics, optimization etc) - Federated learning (i.e distributed machine learning) - Cloud computing - Accessing NCI/HPC resources - Ethics - Using version control software or GitHub - Other)

27. Do you have access to and/or utilise other informal training opportunities such as a slack channel, meetup or hacky hour? (Yes - No)

28. How do you prefer to engage with others about advanced analytics? (select all that apply; Conference/Workshop - Online - Informal meet-ups - Other)

29. Do you participate in any community of practice for learning and development for advanced analytics in general and in health in particular? (Yes - No)

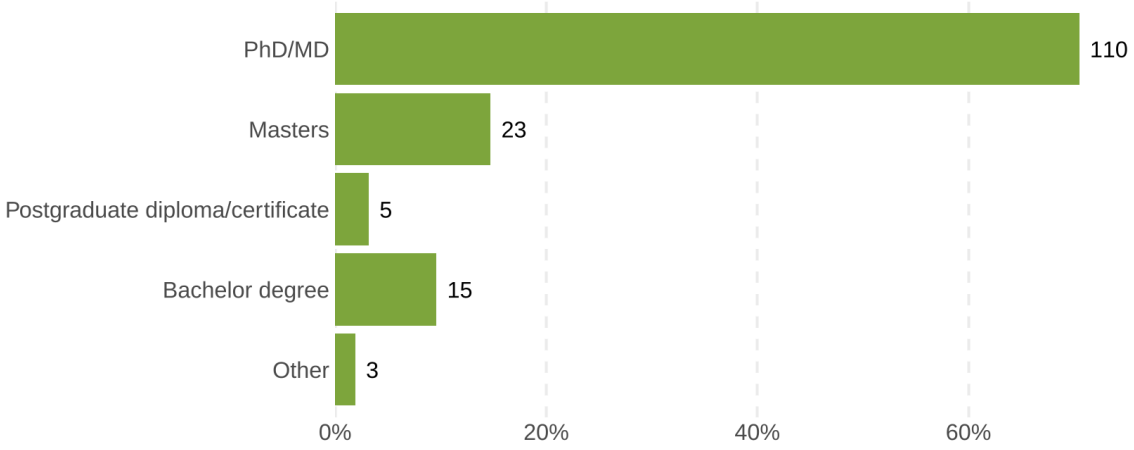
30. How do you rate the need for such a community of practice? (Not at all important - Slightly important - Moderately important - Very important - Extremely important)

31. Are there any other comments you would like to make about what you think the ARDC Advanced Analytics Framework should support or provide? For example, underpinning hardware infrastructure, national reference data assets, tools & environment reference programs or national-level cultural and coordination assets. (open text field)

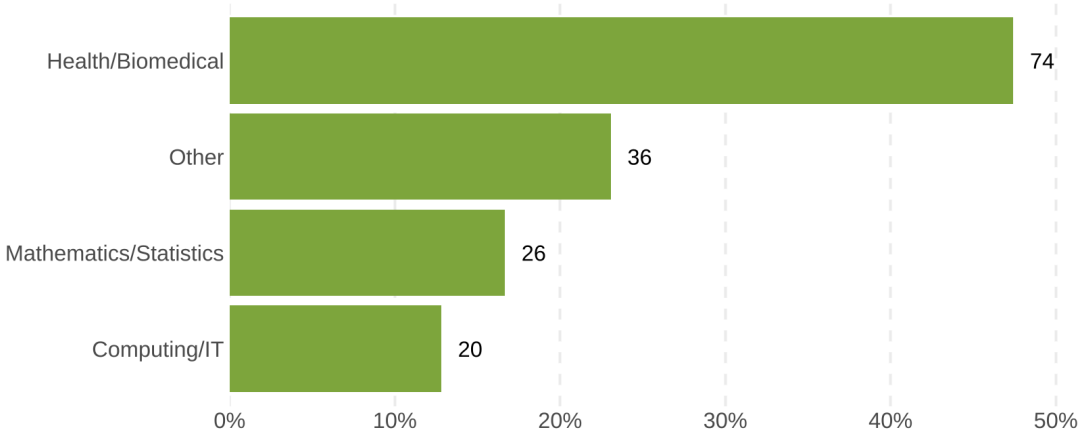
32. Are you willing to be contacted further about the design of the advanced analytics framework for the ARDC people research data commons? (Yes - No)

Appendix 4. Survey Responses

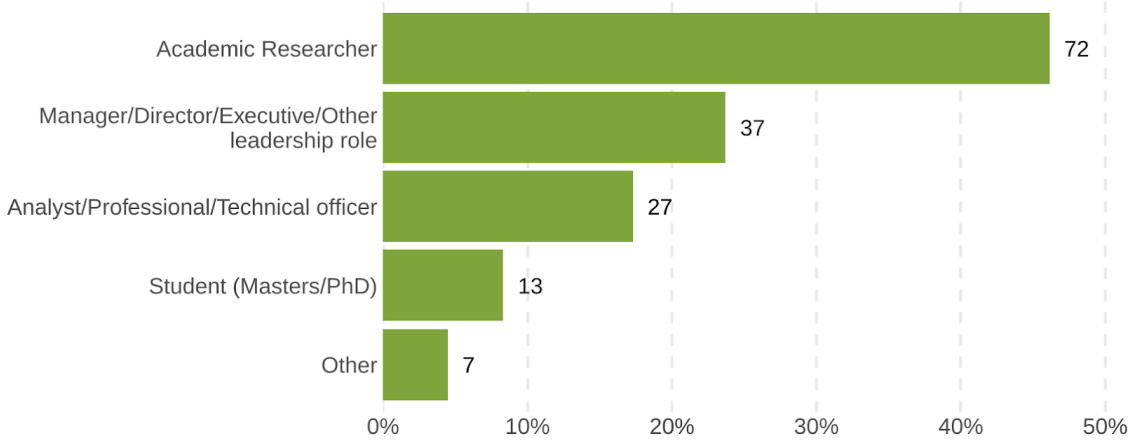
1. What is your highest qualification?



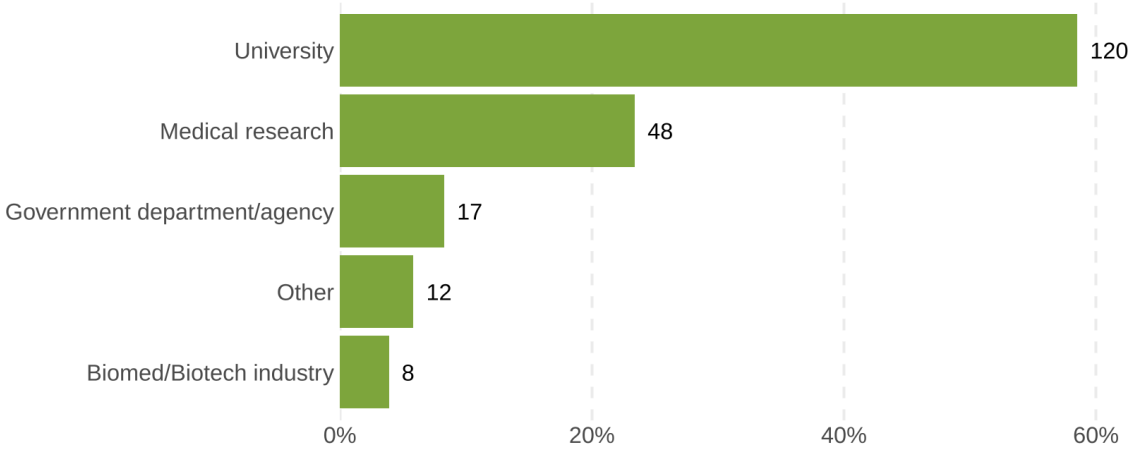
2. In which field is your highest qualification?



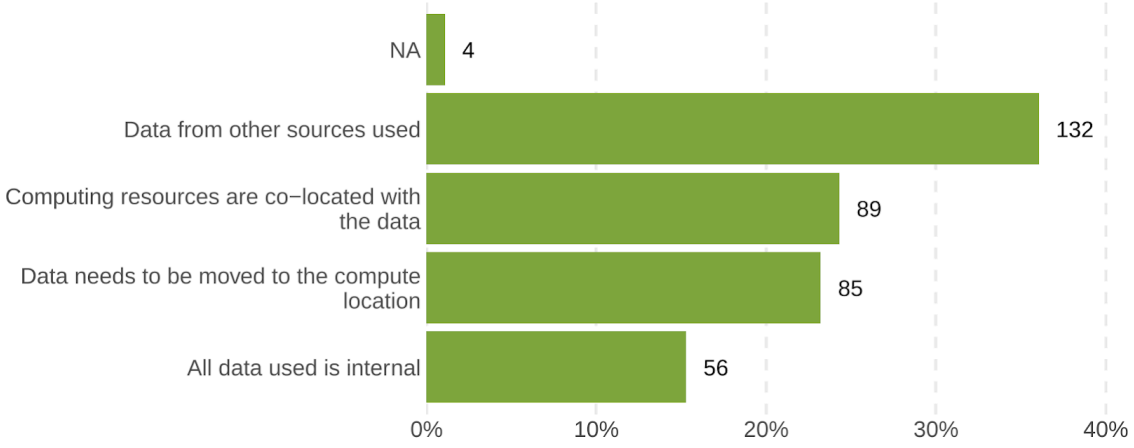
3. What is your current primary position?



4. In which industry do you currently work?

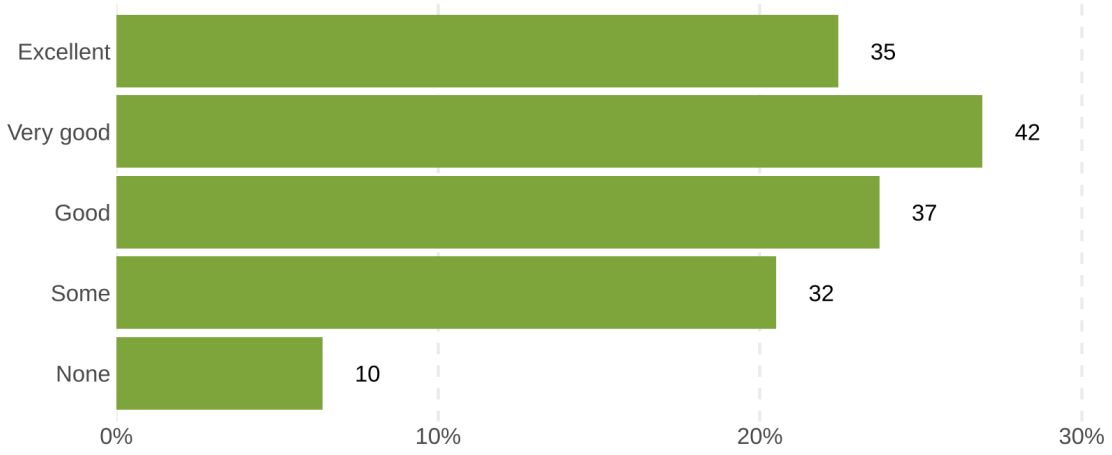


5. In your organisation/work:

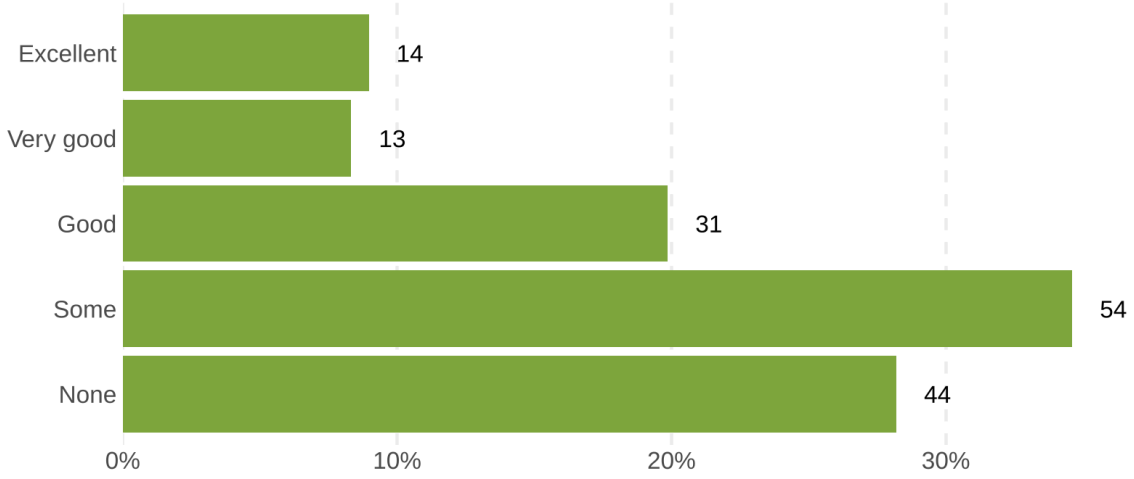


6. Rate your analytics and/or programming skills below:

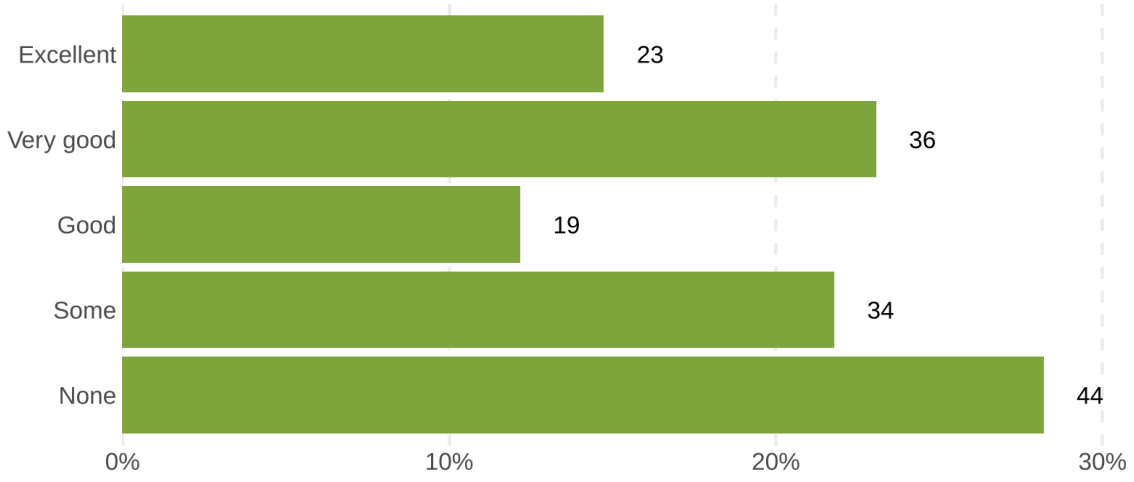
6.1 Applied statistics/mathematics



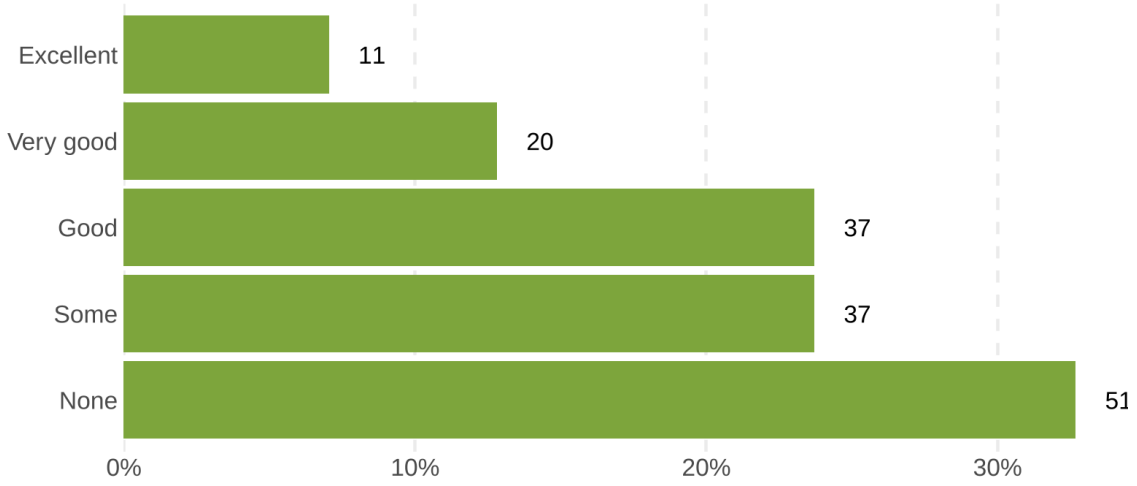
6.2 Artificial Intelligence (AI) and Machine Learning (ML)



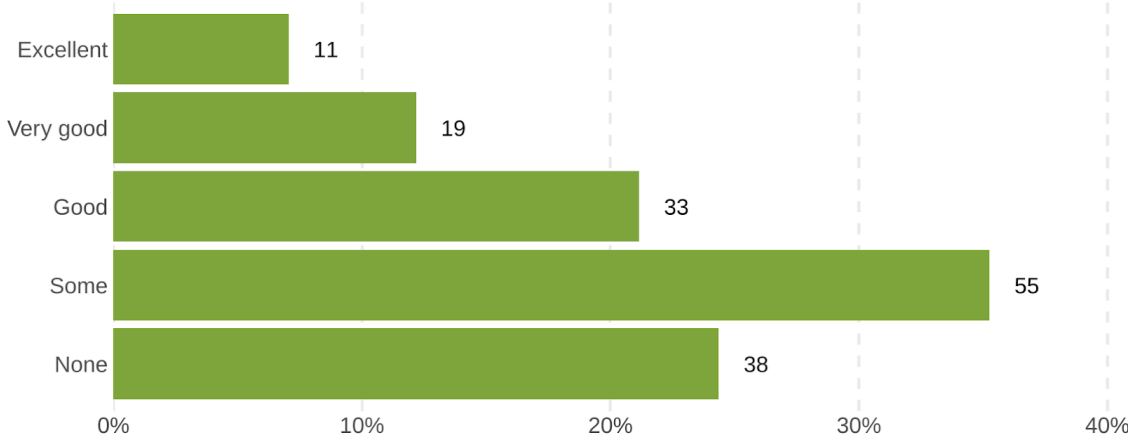
6.3 R/Python programming



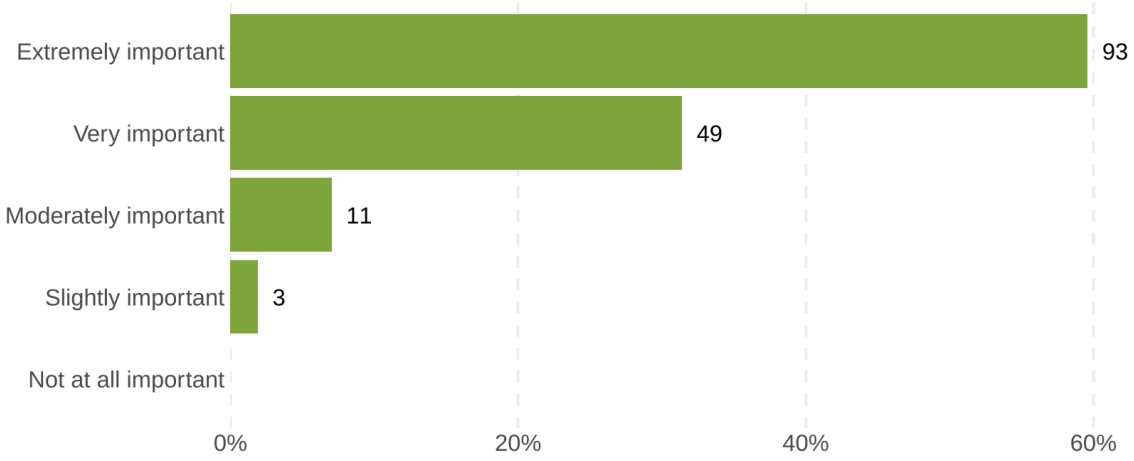
6.4 Cloud Computing/HPC environments



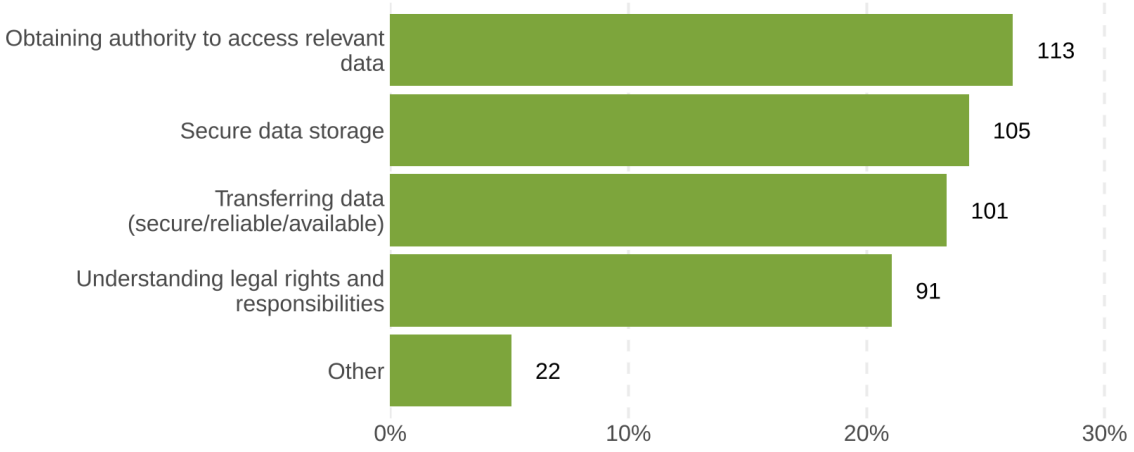
6.5 Data Engineering/Databases



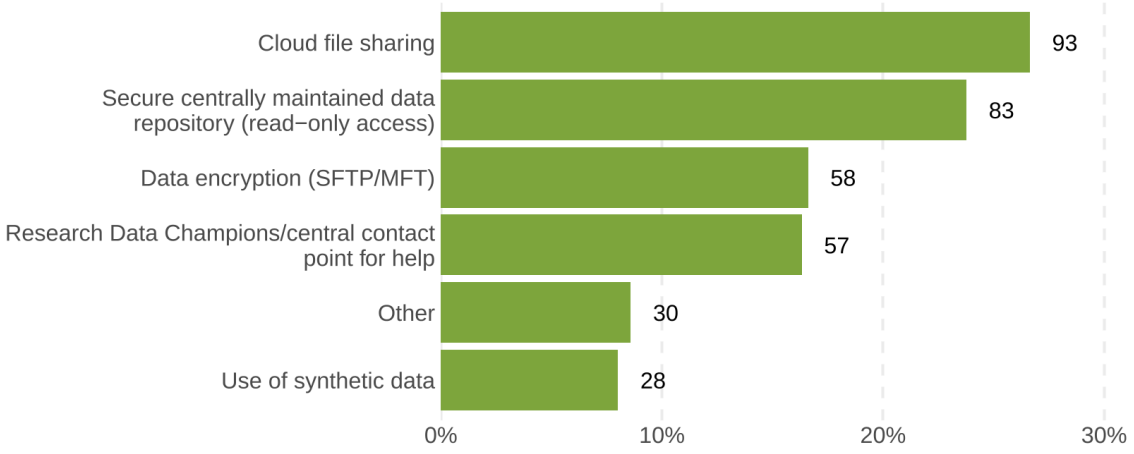
7. How highly do you rate data privacy/ethics as a concern?



8. What do you see as the main issue(s) around privacy/ethics?



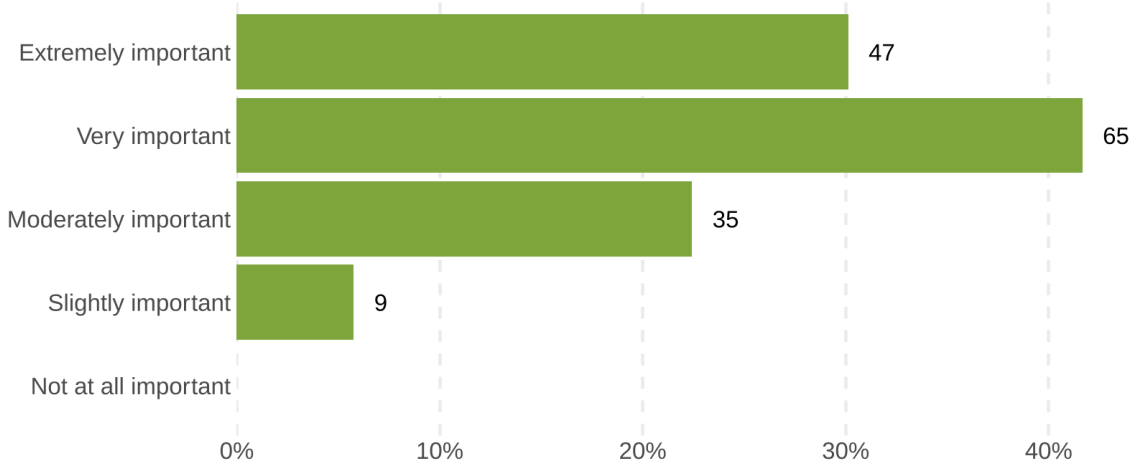
9. What are the current approach(es) in your organisation?



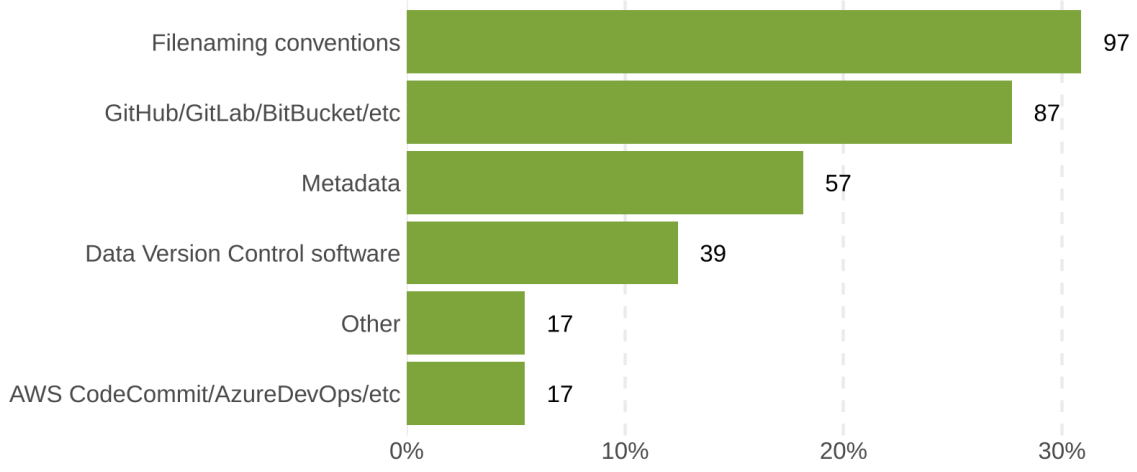
10. In an ideal world, what do you think needs to happen with respect to data privacy?

50% of the respondents (n=78) provided a text response. De-identification and security were commonly mentioned as important. Responders frequently characterised an ideal world for data privacy will have a harmonised regulation and standardised practice across jurisdictions and sectors to facilitate data sharing while ensuring compliance with privacy laws.

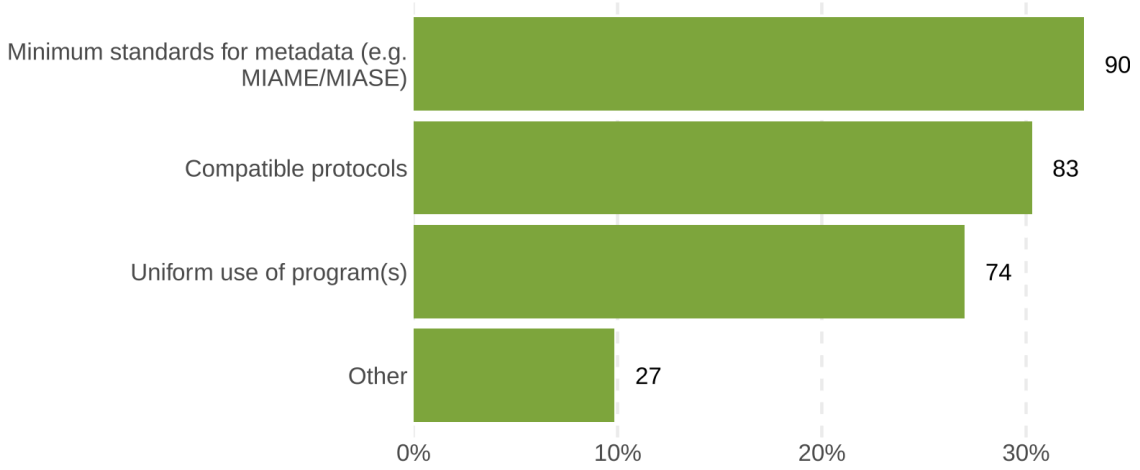
11. How highly do you rate version control as a concern?



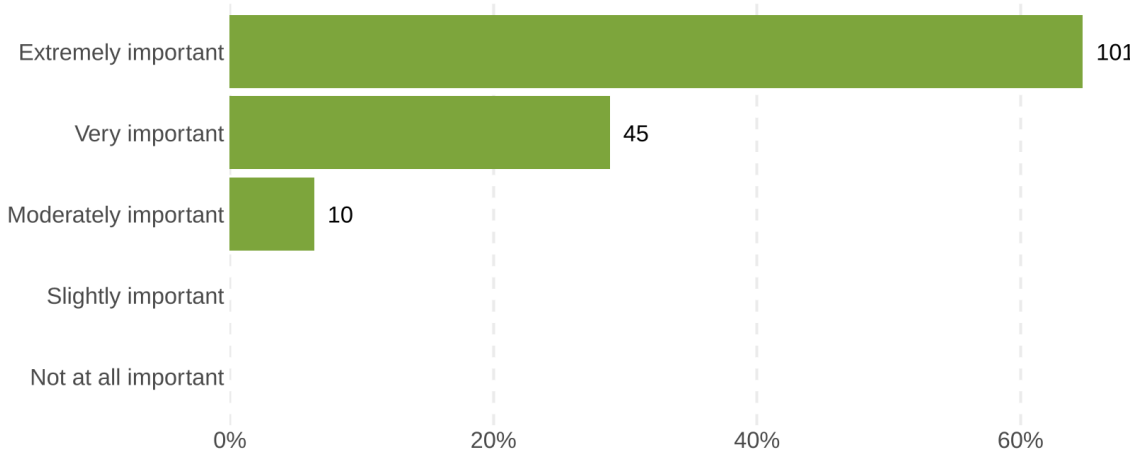
12. What are the current approach(es) in your organisation?



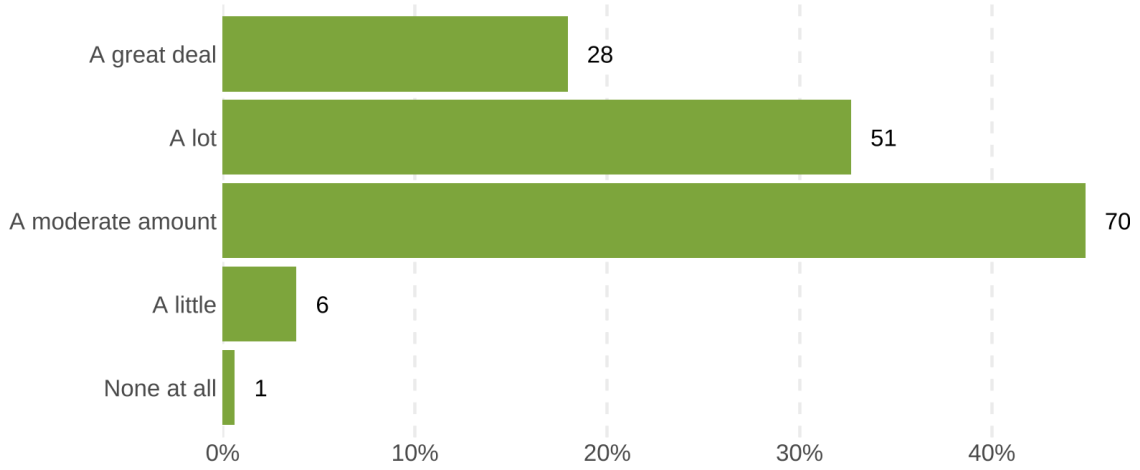
13. What tools or resources would help in this area?



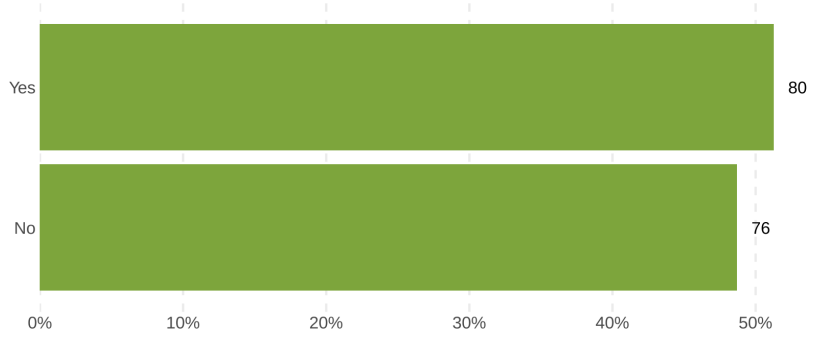
14. How highly do you rate data integrity and quality as a concern?



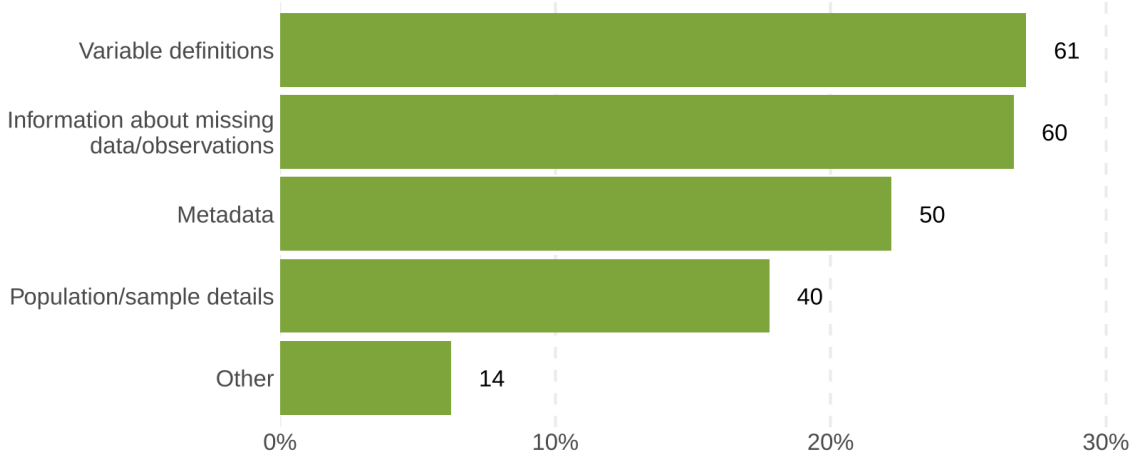
15. How confident are you in the quality and integrity of the data you typically use/access?



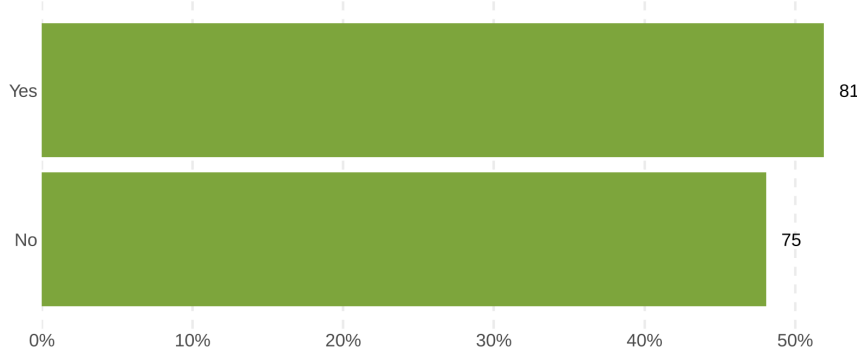
16. In your experience, do you think the documentation available is detailed enough for you to feel confident about using a dataset?



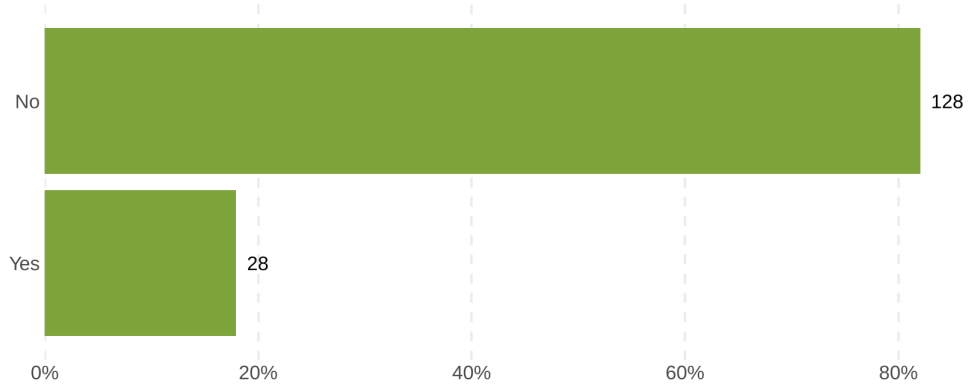
16.1. Since you selected “no” to the previous question, what information do you find is often missing?



17. Would the availability of benchmark and/or synthetic datasets be of use to you/your organisation?

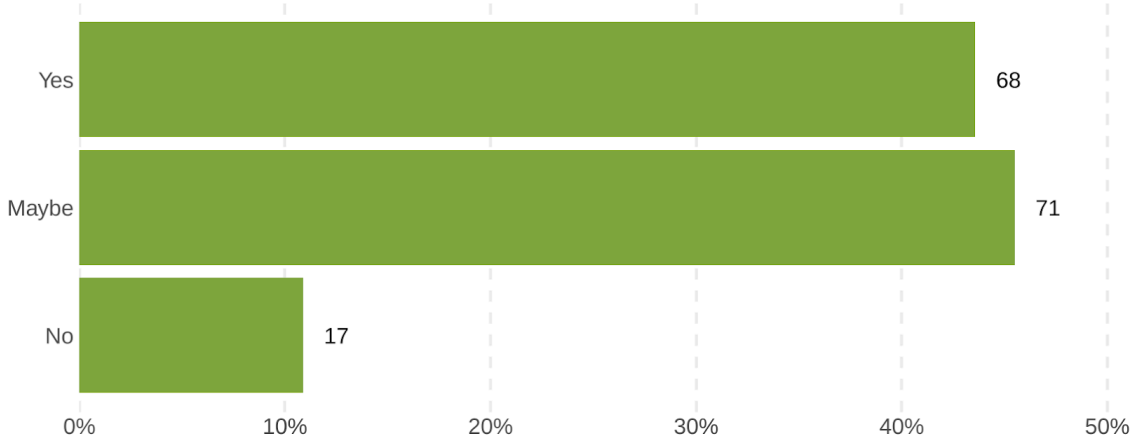


18. Do you currently access data behind a paywall?

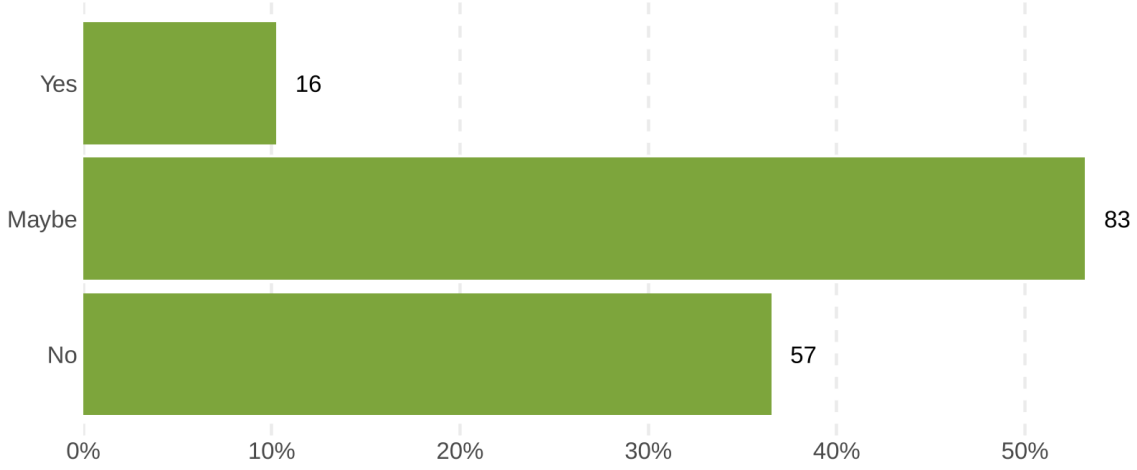


19. If you needed to pay to access your ideal dataset, would you:

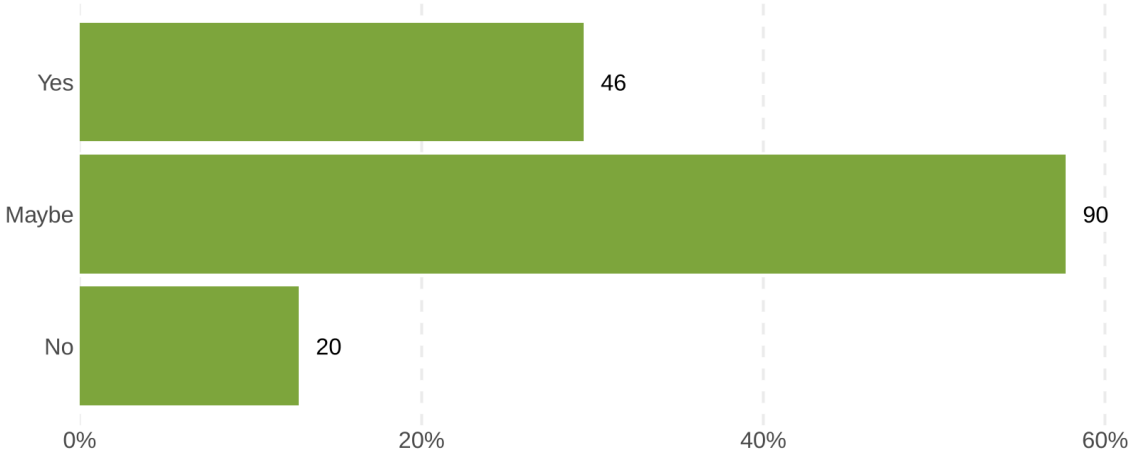
19.1. Search for a different, freely available (but less ideal) dataset instead



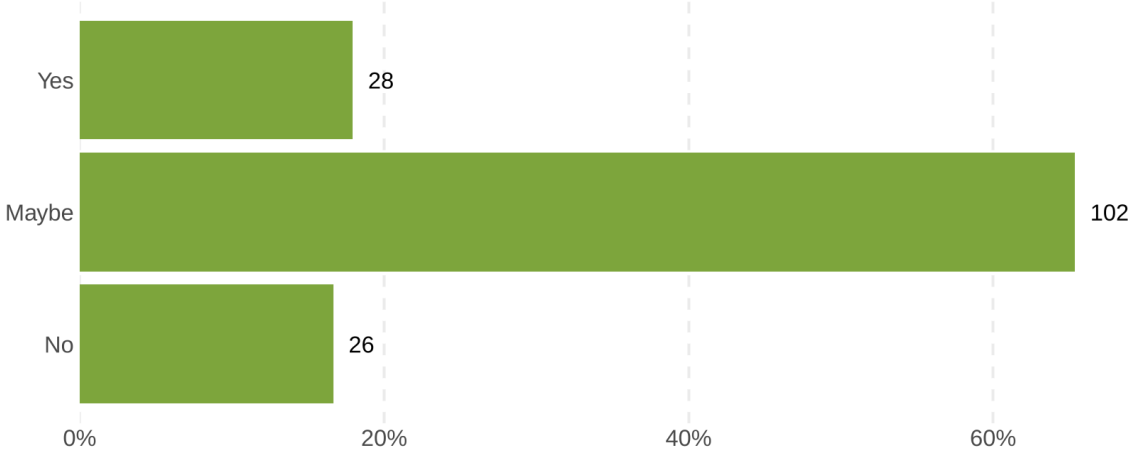
19.2. Have no problem with paying for access



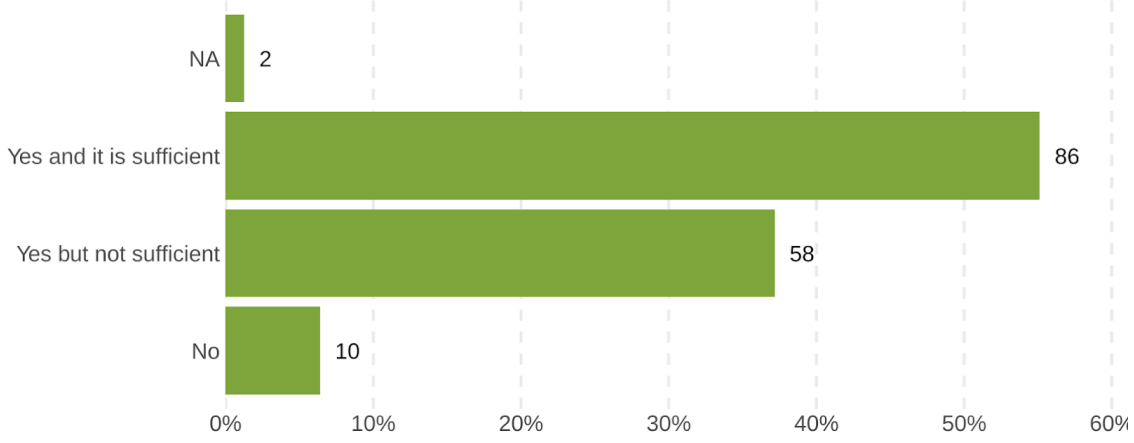
19.3. Pay for the access but funding would be tight



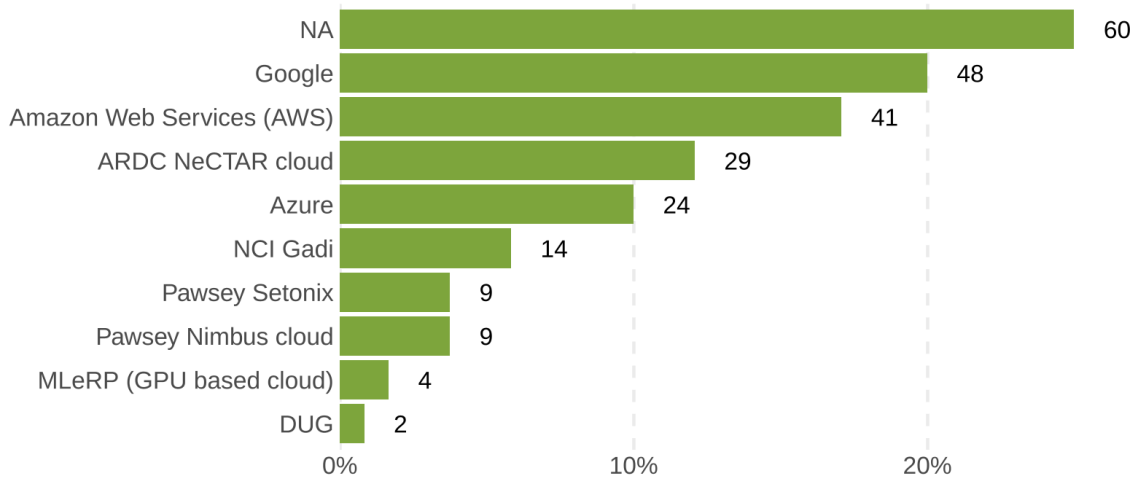
19.4. Need to abandon the project due to lack of funds



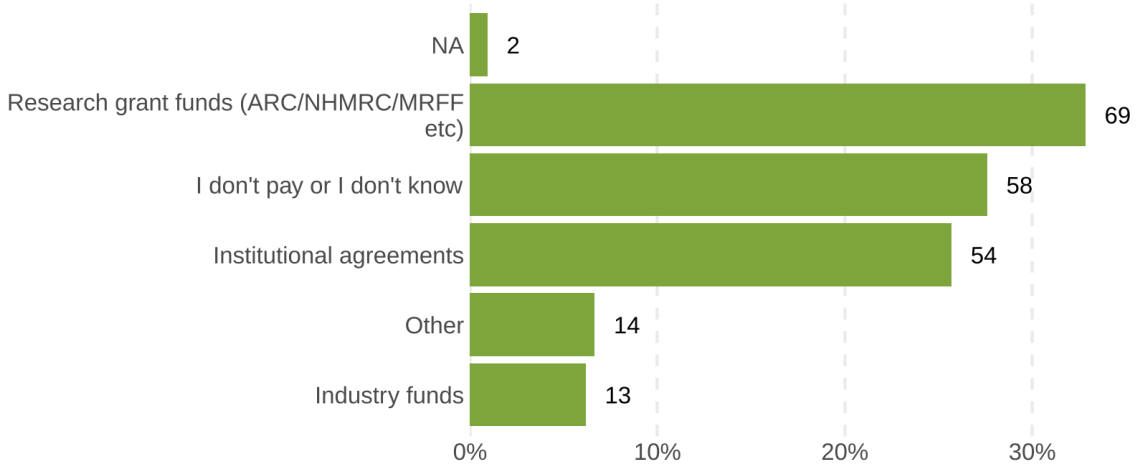
20. Does your workplace provide access to the compute resources required for your work?



21. Do you access or use any of the following computing facilities?

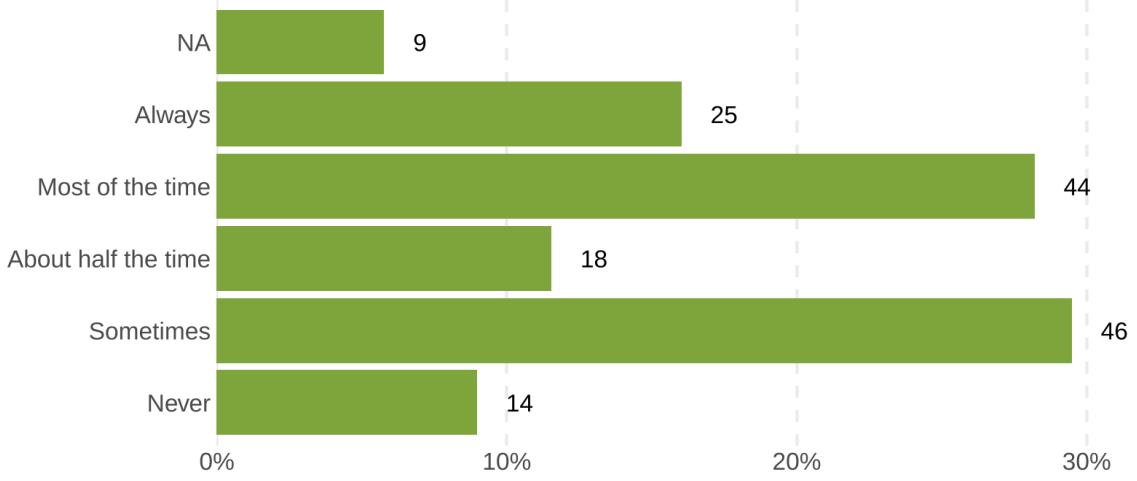


22. How do you pay for access to commercial providers?

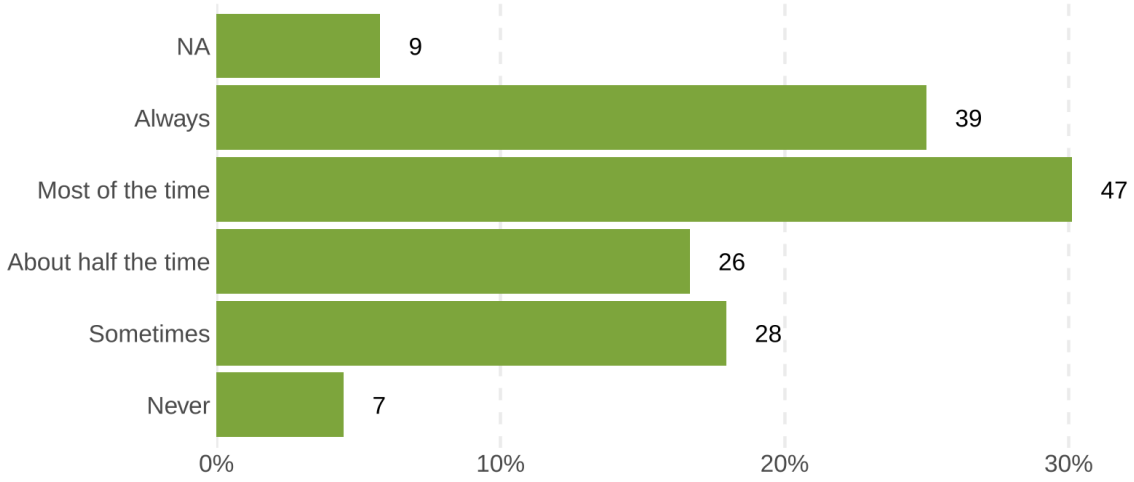


23. How often do you use the following techniques on health data?

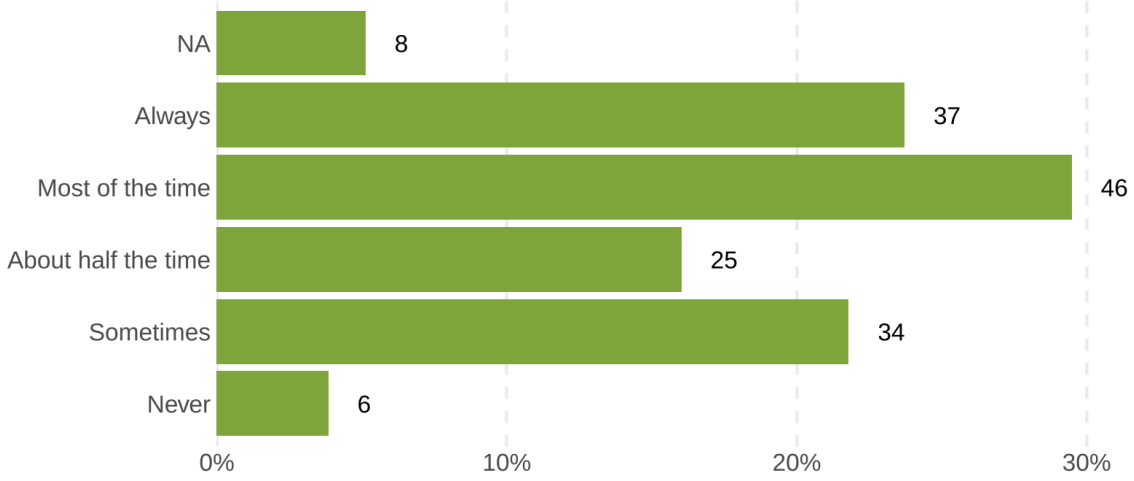
23.1. Database queries/extracting data



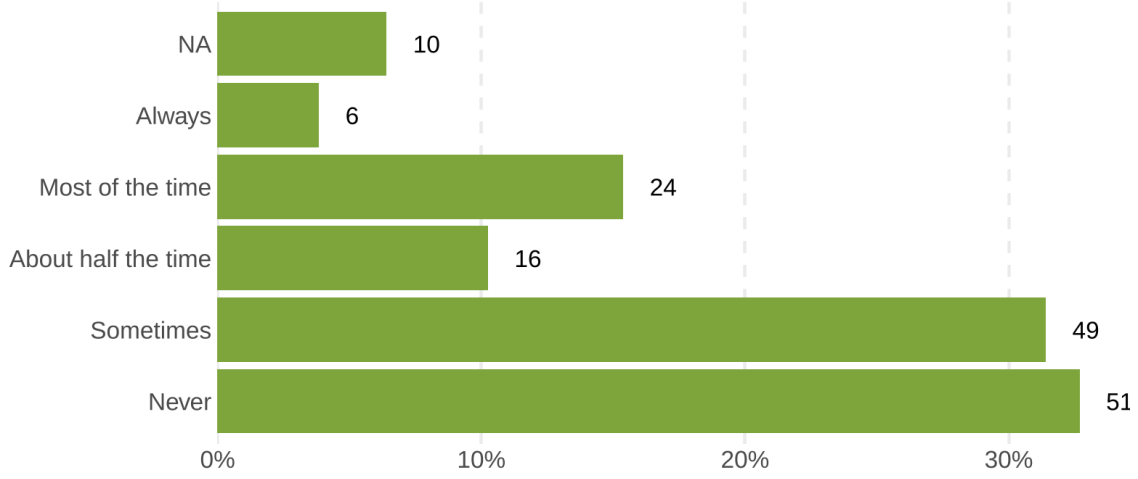
23.2. Exploratory data analysis



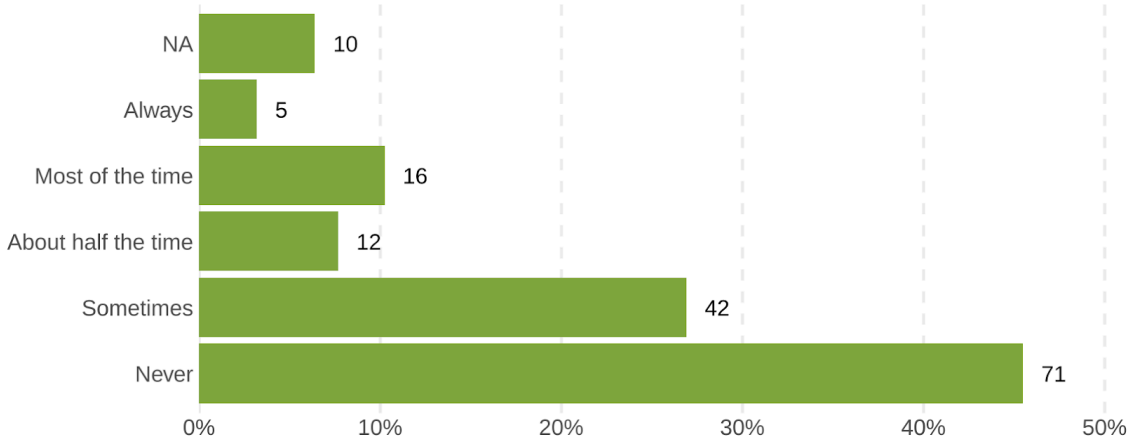
23.3. Standard analysis (e.g. linear regression)



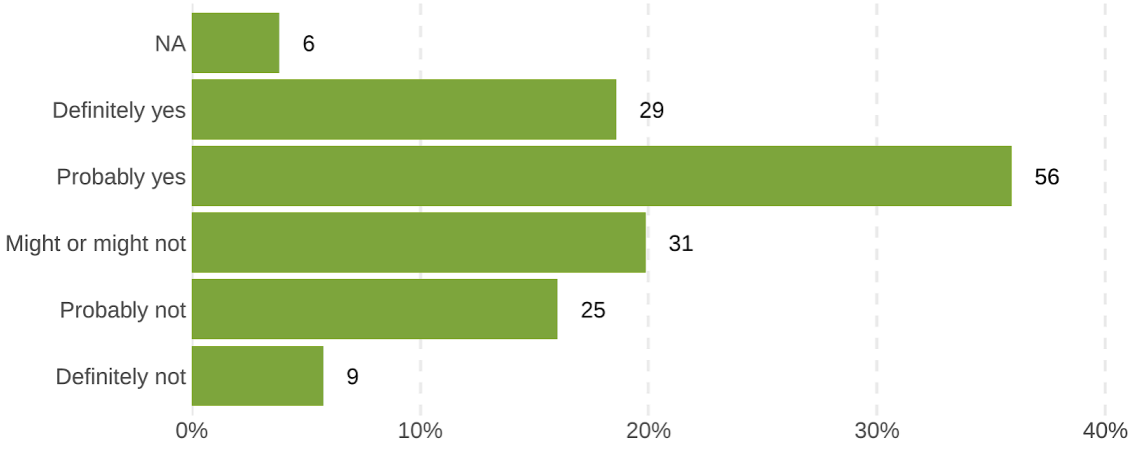
23.4. AI/ML



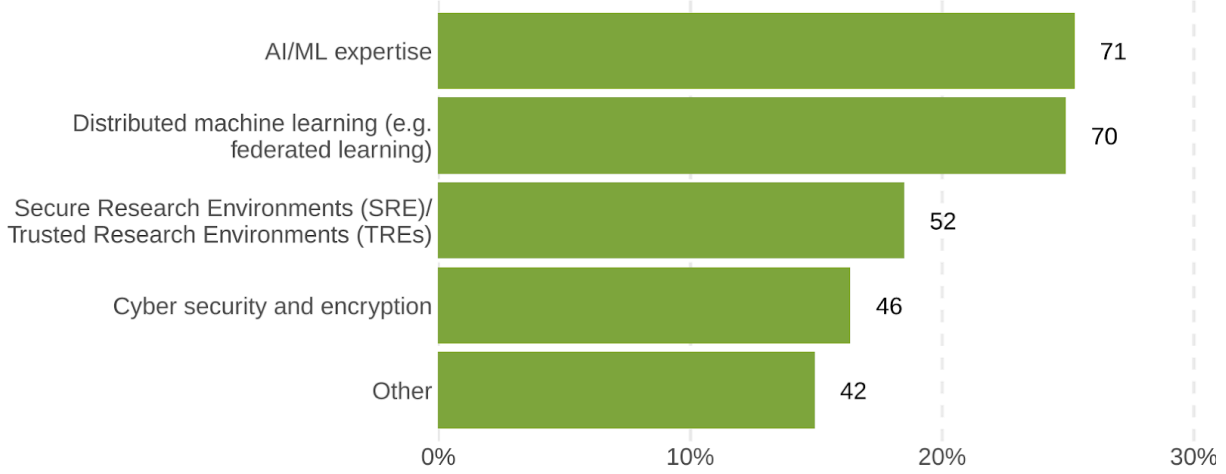
23.5. Deep learning/LLMs



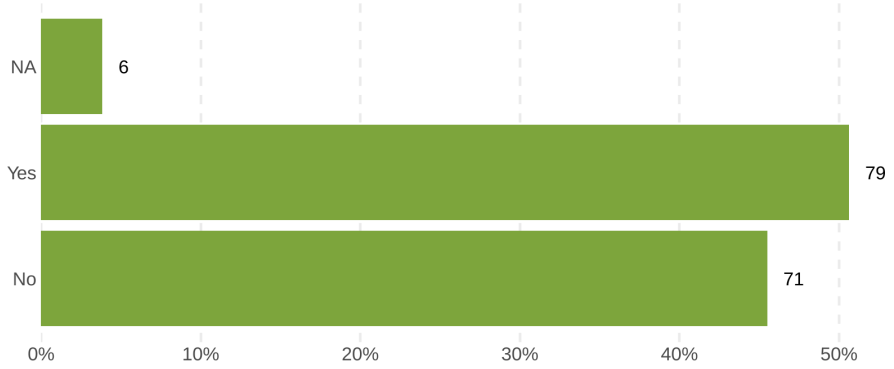
24. Does your organisation have sufficient advanced analytics capabilities in-house?



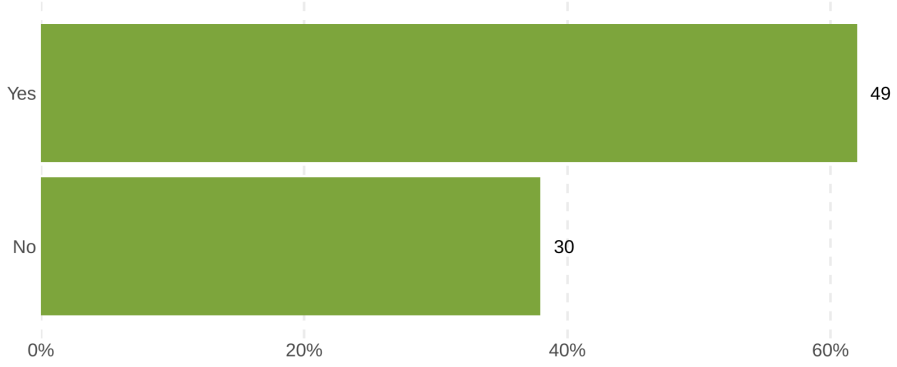
24.1 If no, what area(s) of advanced analytics are needed?



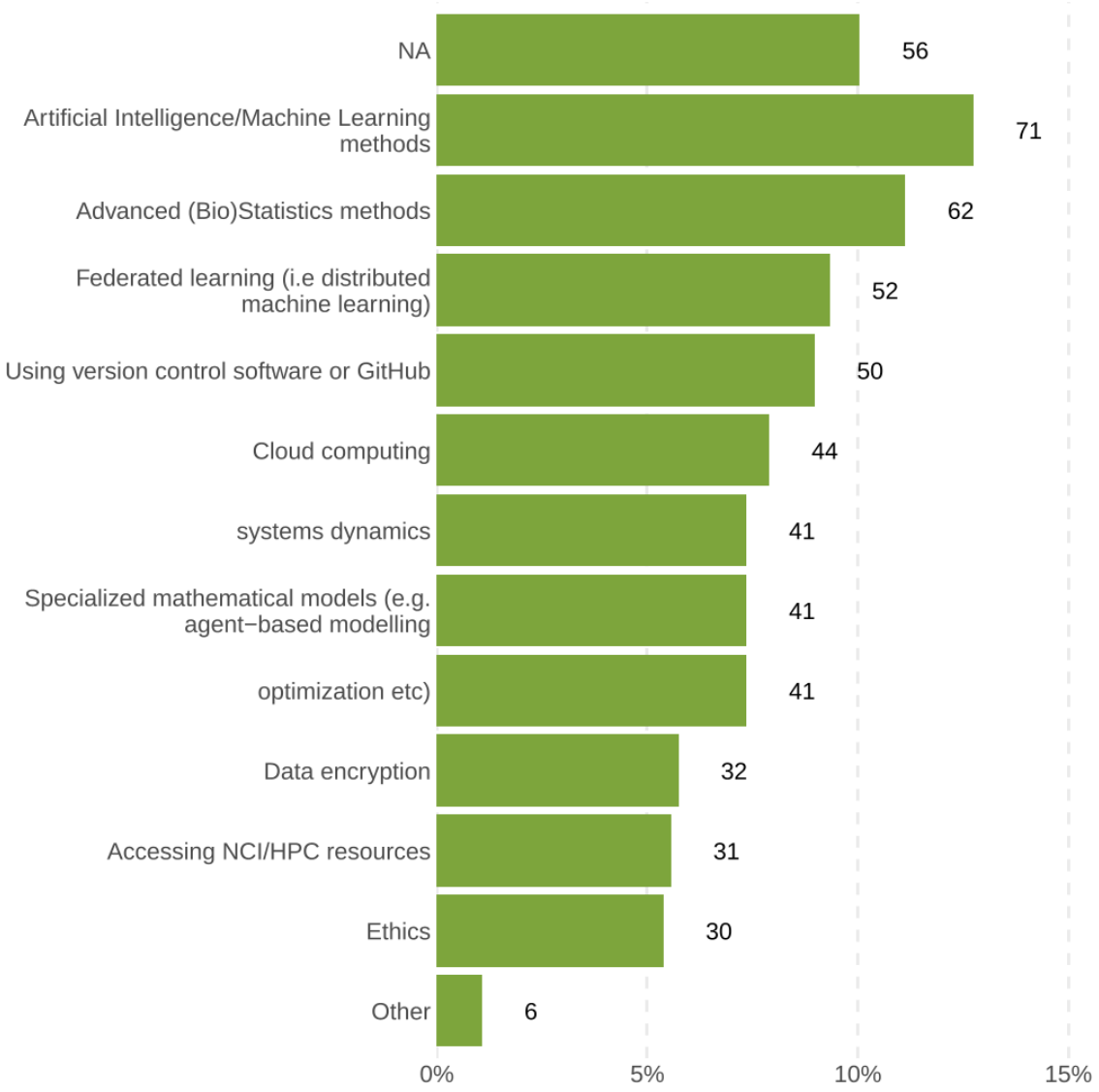
25. Does your organisation provide access to formal training in advanced analytics?



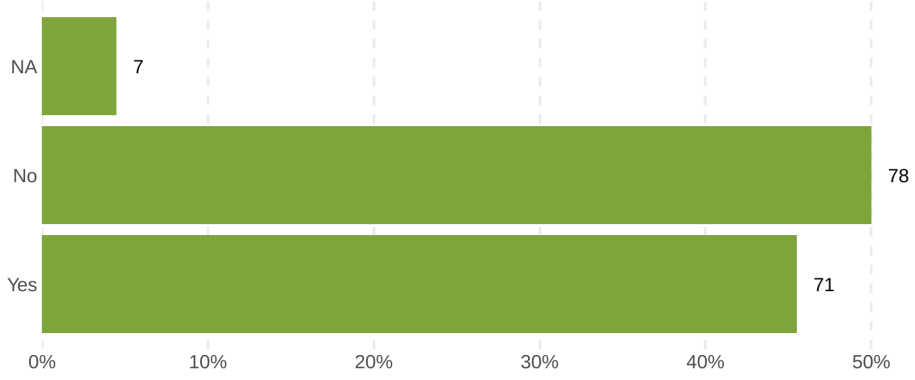
25.1 If yes, is this training sufficient for your needs?



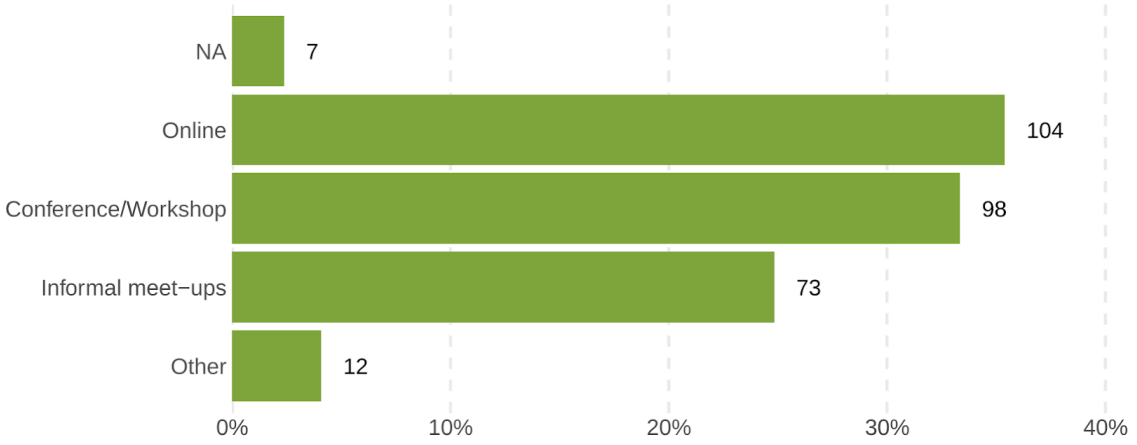
26. In which areas would you like training to be available?



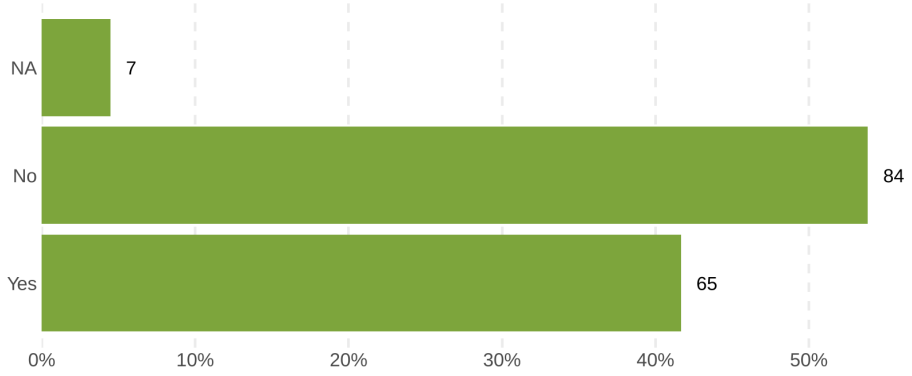
27. Do you have access to and/or utilise other informal training opportunities such as a slack channel, meetup or hacky hour?



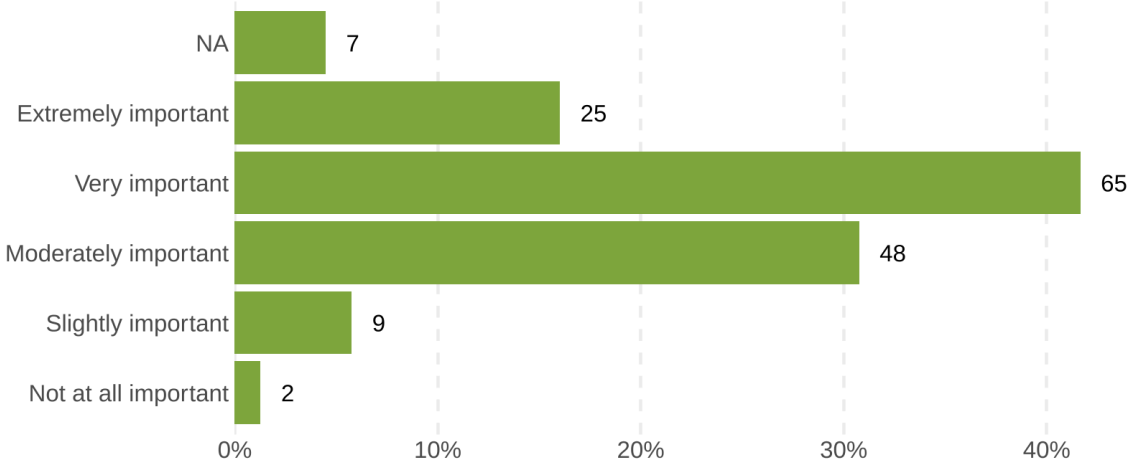
28.How do you prefer to engage with others about advanced analytics?



29. Do you participate in any community of practice for learning and development for advanced analytics in general and in health in particular?



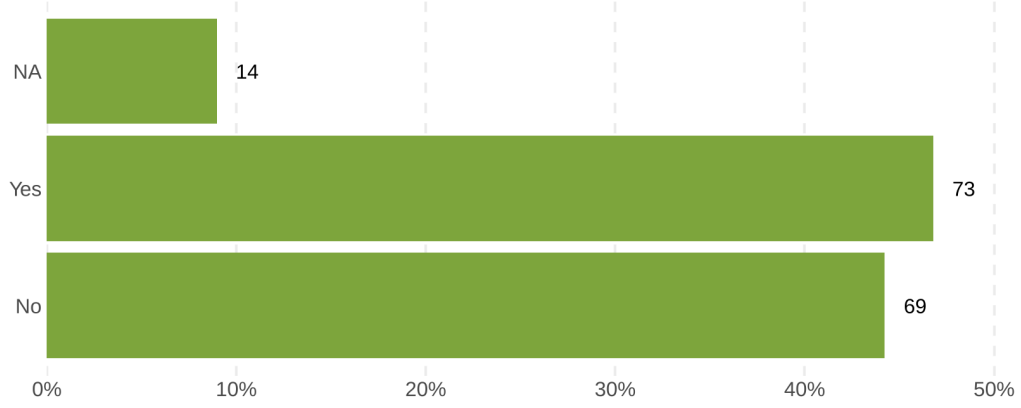
30. How do you rate the need for such a community of practice?

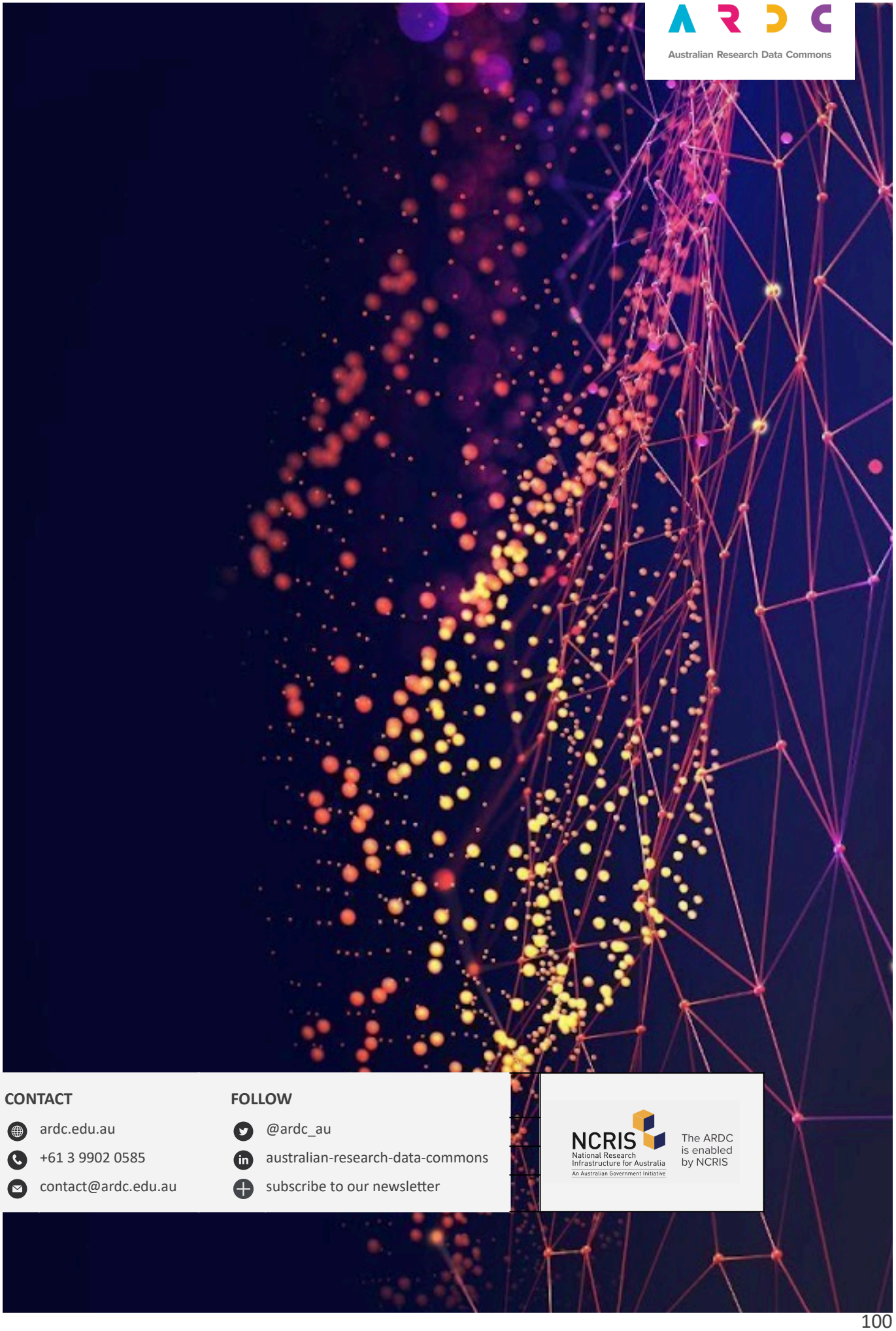


31. Are there any other comments you would like to make about what you think the ARDC Advanced Analytics Framework should support or provide? For example, underpinning hardware infrastructure, national reference data assets, tools & environment reference programs or national-level cultural and coordination assets.

There were 50 respondents (32%) that provided a text response of which 44 (28%) were meaningful (i.e. removing responses like “No”, “N/A”, and so on). The responders commonly raised the need for robust national infrastructure to support data analytics, including high-performance computing (HPC), standardized health data assets, and centralized health data repositories. Another point that was commonly emphasised was on training and education for researchers and healthcare professionals, particularly in foundational skills and understanding advanced analytics method.

32. Are you willing to be contacted further about the design of the advanced analytics framework for the ARDC people research data commons?





CONTACT

ardc.edu.au

+61 3 9902 0585

contact@ardc.edu.au

FOLLOW

@ardc_au

australian-research-data-commons

subscribe to our newsletter

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

The ARDC is enabled by NCRIS