OECD*publishing*

Please cite this paper as:

Deetjen, U., E. T. Meyer and R. Schroeder (2015), “Big Data for Advancing Dementia Research: An Evaluation of Data Sharing Practices in Research on Age-related Neurodegenerative Diseases”, *OECD Digital Economy Papers*, No. 246, OECD Publishing.
http://dx.doi.org/10.1787/5js4sbddf7jk-en

OECD Digital Economy Papers No. 246

Big Data for Advancing Dementia Research

AN EVALUATION OF DATA SHARING PRACTICES IN RESEARCH ON AGE-RELATED NEURODEGENERATIVE DISEASES

Ulrike Deetjen, Eric T. Meyer,
Ralph Schroeder

Big Data for Advancing Dementia Research

An evaluation of data sharing practices in research on age-related neurodegenerative diseases

oiioiioii
oiioiioii
oiioiioii

UNIVERSITY OF OXFORD

OECD
BETTER POLICIES FOR BETTER LIVES

Ulrike Deetjen | Eric T. Meyer | Ralph Schroeder

March 2015

Big Data for Advancing Dementia Research

Ulrike Deetjen | Eric T. Meyer | Ralph Schroeder

Oxford Internet Institute, University of Oxford

March 2015

OECD: Elettra Ronchi, Christian Reimsbach-Kounatze

OECD International Advisory Group: Robin Buckle (Chair), Philippe Amouyel, Neil Buckholtz, Giovanni Frisoni, Yves Joanette, Richard Johnson, Miia Kivipelto, Martin Rossor, Donald Stuss, Yoshiaki Tojo (see Appendix for affiliations)

All interviewees represent their own personal views and not necessarily the official standpoint of any organisation. This report should not be reported as representing the official views of the OECD or of its member countries.

Please cite this report as: Deetjen, U., E. T. Meyer and R. Schroeder (2015), “Big Data for Advancing Dementia Research”, Paris, France: OECD Publishing.

Abstract

Dementia currently affects more than 44 million people worldwide and is a major economic burden projected to increase over the next decades. To date, there is no cure for dementia, or treatment to prevent disease progression. Making better use of big data and sharing it among the research community may accelerate dementia research, as it offers the promise of larger and wider datasets that enable new insights from both well-established and novel sources and types of data.

This report addresses the question of how big data can be used and shared more efficiently for dementia research. This primarily concerns large datasets of genetic, imaging, clinical or proteomic data obtained in medical studies, but also linkage to the vast amounts of routinely collected data within and outside of the health system.

For this analysis, we examine four data sharing initiatives: ADNI, AddNeuroMed, UK Biobank and the Swedish Brain Power studies. For each of these case studies, we interviewed leading researchers from academia as well as general experts from government, funding bodies and industry.

We found a range of data governance models in terms of design, access policies, interoperability, quality assurance and ownership. We also identified more deep-seated structural challenges to using and sharing data: These go beyond technical and consent-related challenges, and include the need for a favourable ecosystem in terms of the legal basis, collaboration across all stakeholders and sustainable funding, as well as the people challenge of appropriate skills, aligned incentives and mindsets.

We use these findings as a launching point for outlining next steps, with a specific focus on how public policy may contribute to unlocking the power of big data to advance dementia research in the coming years.

Executive Summary

Dementia affects about **44 million individuals today**, a number that is expected to nearly double by 2030 and triple by 2050. With an estimated **cost of 604 billion USD annually**, dementia represents a major growing economic burden for both industrial and developing countries, in addition to the significant physical and emotional burden it places on individuals, family members and caregivers.

Currently, **neither a cure for dementia nor a reliable way to slow down its progress exists**. Improving prevention and diagnosis, and discovering potential ways of treatment and cure requires better understanding of the mechanisms underlying neurodegeneration. However, these mechanisms are complex, and influenced by a range of genetic and environmental influences that may have no immediately apparent connection to brain health.

In December 2013, the **G8 Global Dementia Summit** in London identified the better use of available data, resource sharing and researcher collaboration as key priorities. With the ambition to find a cure or disease-modifying therapy by 2025, the G8 health ministers mandated the **OECD** to report on **how big data can be used and shared more efficiently for dementia research**.

This report is one part of that **OECD** work, and is aimed at a wide audience of policymakers, funders, the private sector and researchers. The results will be reported to the **World Dementia Council** and presented to the G7 health ministers at the First WHO Ministerial Conference on Global Action Against Dementia in Geneva in March 2015.

The promises of big data for dementia research

At the core of dementia research are data obtained purposively in medical settings, such as images; clinical, genetic, proteomic and biological data; and cognitive tests or surveys. **Big data approaches to research add new types of data and ways of analysing them** to this repertoire, including data from electronic medical records, registries and other routine health data, from online patient platforms, but also from retailers and mobile phone providers for social and lifestyle insights.

Both **broad data** (relating to the number of individuals represented in a dataset) and **deep data** (an indication of the number of measures and granularity of those measures related to each individual) are part of the big picture for dementia research. Both of these types of data represent challenges for researchers in terms of generating, linking, using and sharing data in a way that respects individual rights to privacy without unnecessarily constraining dementia research.

To address these challenges, we need to **better understand current and emerging practices in data sharing and governance**. To this end, the authors of this report interviewed 37 leading experts from academia and beyond, with a particular focus on four case studies of data sharing initiatives: ADNI, AddNeuroMed, UK Biobank and the Swedish Brain Power studies. These cases represent a number of important examples of efforts to create, federate, catalogue, and share data, and cover the spectrum of ongoing data sharing activities in terms of geographic coverage, general or dementia-specific focus, size, maturity, openness by design and linkage to routine data.

ADNI (Alzheimer's Disease Neuroimaging Initiative) was established in 2004 to better predict and monitor Alzheimer's Disease (AD) onset and progression, establish global standards to measure cognitive changes, and share data across the international research community. North American ADNI includes imaging, genetic, and clinical data from nearly 2,500 research participants, with various other countries having established spin-off initiatives.

AddNeuroMed was a European initiative started in 2005 that followed a relatively traditional scientific research model, initially planning to use the dataset only within the collaboration, but sharing it more broadly later in response to specific requests. Deep data on 700 patients were collected, mostly used by European researchers.

UK Biobank is a massive initiative to collect longitudinal data on 500,000 UK volunteers that combines research data with routine data from electronic medical records. After the initial data collection between 2006–2010, the data were recently made available to researchers. UK Biobank not only relates to dementia, but also to a variety of conditions as participants develop diseases over the course of the 25+ year data collection.

The **Swedish Brain Power** network ties together a variety of national longitudinal population-based health studies that take advantage of Sweden's remarkably complete and traceable health information, which is connected using a unique and widely used identification number for all residents.

Current data governance of data sharing initiatives

Today, **availability** of data for dementia research is largely in place, with our case studies jointly covering a wide spectrum of observational studies and population-based studies with both broad and deep data. Various other sources relevant for dementia research exist both within and outside of the medical realm. **Interoperability** of these datasets remains a challenge due to a wide variety of data collection methods across studies, as well as legal and consent-related protections designed to prevent linkage. While routine data from within the health system is incorporated in UK Biobank and the Swedish Brain Power data, for example, linking to big data from outside the medical realm is still in its infancy.

Accessibility to the data of our four case studies follows a variety of models, from fast online access with low-effort applications as for ADNI, to staged application procedures, to models relying on direct collaboration. Not all of these models scale well, and hybrid models may need to be put in place incorporating aspects of both lending and reading library models. **Ownership** rights often prevent users from redistributing these datasets, which makes further data aggregation and linkage difficult but allows for some control over the data and the uses to which it is being put to be retained.

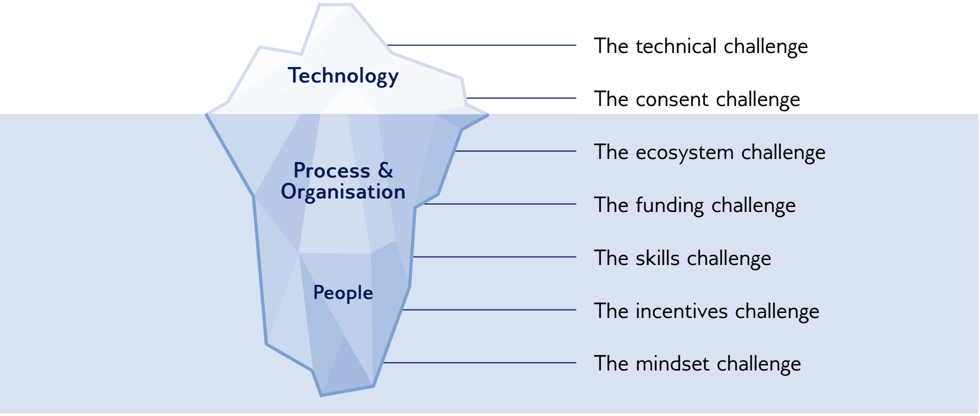
Specifically with purposively collected data where **quality** can be influenced, (meta) data management needs to be a priority from the early design stages. Governance models for user-provided documentation systems and user-enhanced data being integrated back to the resource can reduce duplication of effort. For routine data, new analytical approaches may help to pull signal out of the noise.

Traceability is no major issue for connecting research participants across the research process or across multiple data sources, especially if it is designed for from the beginning. For big data in particular, traceability needs to be balanced with **privacy/security**, which has been taken seriously by all data sharing initiatives. While consent for each of our case studies covers traditional scientific research, innovations in open sharing, integrating big data from outside the medical realm and enabling new forms of crowdsourcing still present significant challenges.

We conclude there is no single way to build and share data, as each of our case studies has made valuable contributions to the field of dementia research. However, "openness by design" helps – both in terms of setting up the structures accordingly, obtaining consent for linking traditional and potentially novel types of routine data, informing participants adequately about risks and consequences associated with their participation, establishing governance structures that allow for responsible use and sharing of data in a changing context and environment, and in terms of aligning motivations of researchers to contribute.

Challenges to using and sharing big data in dementia research

Beyond data governance, our interviewees revealed a range of more deep-seated challenges to harnessing big data for dementia research as illustrated by the iceberg analogy. Hurdles in relation to technology are evident, but can largely be overcome. Below the surface are issues around how data collection, analysis and sharing are managed, as well as underlying people-related challenges.



Above the surface is the **technical** challenge in relation to mechanisms for sharing data securely and the need for common standards to pool data more easily. Establishing robust yet flexible core data standards can make data more sharable by design and save researchers time and effort. **Consent** needs to be set up in a way that it is understood by individuals and protects them against data misuses with effective enforcement mechanisms, but without unduly hampering the potential of research and routine data from a variety of sources and the ability for scientists to collaborate beyond borders and across time.

Below the surface are challenges of process and organisation, including the need to ensure a favourable **ecosystem** for research with stable and beneficial legal frameworks, and links to pharmaceutical companies and other private organisations for exchange of data and expertise, and connecting research findings to future prevention strategies and treatments. Similarly important is sustainable **funding** for data infrastructures, while funders at the same time can also have considerable influence on how research data, in particular, are made available.

The most fundamental level relates to the people involved in dementia research: the scientists, but also the policymakers, regulators, private partners, patients and research participants. We need more people with appropriate **skills** to manage big data, apply their imagination, and employ novel analytical approaches. These people must be trained, be connected across disciplines, and be given **incentives**: Currently, there are relatively few incentives or career rewards that accrue to data creators and curators. Rather than a one-size-fits-all model of research and publication, additional ways to recognise the value of shared data must be built into the system. Finally, everyone involved must shift their thinking to adopt a **mindset** towards responsible data sharing, collaborative effort, and long-term commitment to building the two-way connections between basic science, clinical care and the increasingly fluid boundaries of healthcare in everyday life.

Advancing dementia research **requires addressing "all of the iceberg"**, as only by tackling the challenges at all levels jointly, will change happen – without being held back by the weakest link in the chain. Some competing interests may need to be balanced with each other: While privacy concerns about digital, highly sensitive data are important and should not be deemphasised as a subordinate goal to advancing dementia research, they can be balanced with the openness required through releasing data in a protected environment, allowing people to voluntarily "donate data" about themselves more easily, and establishing governance mechanisms that safeguard appropriate data use for a wide range of purposes, especially in instances in which the significance of data changes with its context of use.

Next steps for unlocking the value of big data for dementia research

There is **no lack of data for dementia research, but we need to exploit it more effectively**: Sharing data globally across research teams and tapping into new data sources is a prerequisite for resources to be exploited more fully. At the same time, collaboration between dementia researchers and those disciplines researching factors “below the neck” (such as cardiovascular or metabolic diseases) is important – while collaboration with engineers, physicists or innovative private sector organisations may prove fruitful for tapping into new sources and skill sets.

It is worth highlighting that **no one nation has it all, but complementarities exist**. Global data sharing and collaboration can help to exploit data more fully and leverage more researchers to spend time on analysing, rather than collecting new data. Dementia is a disease that concerns all nations in the developed and developing world. To enable global collaboration, it is crucial that just as diseases do not respect national boundaries, neither should research into dementia and funding of data infrastructures be seen as purely a national or regional priority.

Dealing with big data is not entirely new, but requires **some new operational procedures with larger and more complex types of data**. As data are combined from different research teams, institutions and nations, funded by a variety of organisations, and even combined with data from outside the medical realm, new access models will have to be developed that make data widely available while protecting privacy as well as the personal, professional, and business interests of the data originator.

To fully capture its potential, big data requires **thinking outside of the box**. It requires imagination to consider what data sources to use, and how to link in big data being generated routinely across all facets of everyday life, ranging from mobile phone data, to customer data, to tracking data, to government data. All of these have potential for understanding the behaviour and environment of dementia patients not only after diagnosis, but for prevention, early identification and diagnosis, or even to retrospectively analyse the years leading up to diagnosis. For this, we need to develop a culture that promotes trust between people who form part of the data, and those capturing and using the data.

At the same time, big data also offers **new forms of potential involvement for individuals**. Actively involving people in contributing to research by donating their data, participating in consumer-led research, and engaging as citizen scientists can capture valuable user-generated data and yield unexpected benefits. People have a strongly vested interest in their health and the health of their loved ones, and empowering them to be active contributors to science is a way to alleviate the helplessness that many may feel, while also improving the future for themselves, their families, and others who will be touched by dementia.

Finally, we need an **ongoing dialogue about new ethical questions that big data raises**. We will need to discuss the direct and indirect benefits to participants engaged in research, when it is appropriate for data collected for one purpose to be put to novel uses, and to what extent individuals can make decisions especially on genetic data, which may have more far-reaching consequences. The scientific need to use longitudinal data to understand diseases may also need to be balanced with the fundamental right to privacy and the “right to be forgotten”.

Recommendations: How public policy can help

Policymakers and the international community have an integral **leadership role to play in informing and driving the public debate on responsibly using and sharing data** alongside with researchers, funders and other stakeholders. More directly, public policy can help to stimulate more innovative uses and sharing of data through a variety of supporting initiatives detailed as follows.

First of all, **funding needs to support dementia and research infrastructures for using and sharing data**. If data sharing and more widespread use of routine data is to become the norm, we must fund data sharing and infrastructures for doing so, and recognise this as one of the essential costs of good science.

Policy should stimulate **collaboration between public and private actors**. Public-private partnerships, in-kind donations of data and expertise, government tax incentives for contributions to science, and other innovative mechanisms can help make data for dementia research available – both in relation to pharmaceutical companies, but also supermarket chains, mobile phone companies or start-ups.

There needs to be **investment in future health-/bioinformatics talent** and increased collaboration with data experts outside dementia research. Universities will need to offer opportunities and funding for education in data science and related areas, create multi-disciplinary centres of excellence, and focus on interdisciplinary, multi-institution and multi-country research.

On an international level, **guidelines for consent and Institutional Review Boards (IRB) or Ethics Review Committees (ERCs)** need to be agreed on. Reducing uncertainty about whether consents obtained for medical research allow data to be shared beyond an institution, collaboration, or nation, can lower the barriers to sharing while still protecting research participants. Going forward, we need to obtain routinely and purposively collected data in a future-proof way, with governance mechanisms to give confidence that uses of the data remain consistent with an ethical framework that reflects the spirit in which an individual agreed to his or her data being used.

Also beyond national boundaries, a **stable and beneficial legal framework** must be ensured. We need policies that protect citizens against any undue exploitation of their data that they would not want, but must balance data protection and privacy rights with making medical advances in the interest of everyone. Legislation also needs to account for the growing global research communities in terms of funding and making best use of human and data resources.

It is worth emphasising that all of these recommendations do not just apply to dementia research, but are especially pronounced in this case due to the potential of data from outside the medical realm for advancing dementia research, the current state of research in relation to the aim of having a disease-modifying therapy by 2025, and the high personal, societal and economic importance to improve prevention, diagnosis, treatment and cure across the globe.

Contents

List of Figures

List of Tables

List of Abbreviations

AD	Alzheimer's Disease
ADNI	Alzheimer's Disease Neuroimaging Initiative
AIBL	Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing
BBMRI	Biobanking and Biomolecular Resources Research Infrastructure
CAMD	Coalition Against Major Diseases
CDISC	Clinical Data Interchange Standards Consortium
CEO	Chief Executive Officer
CLSA	Canadian Longitudinal Study on Aging
CoEN	Centres of Excellence in Neurodegeneration
CPI	Critical Path Institute
CSF	Cerebrospinal fluid
EADC	European Alzheimer's Disease Consortium
EDPI	European Dementia Prevention Initiative
EMA	European Medicines Agency
EMIF	European Medical Information Framework
EMR	Electronic Medical Record
ERC	Ethics Review Committee
EU	European Union
EUR	Euro (currency)
FDA	Food and Drug Administration (US)

GAAIN	Global Alzheimer's Association Interactive Network
GAIN	Genetic Association Information Network
GBP	British Pound (currency)
GP	General Practitioner
GPRD	General Practice Research Database
ICHOM	International Consortium for Health Outcomes Measurement
IDAD	International Database on Aging and Dementia
IMI	Innovative Medicines Initiative
IRB	Institutional Review Board
JPND	Joint Programme – Neurodegenerative Disease Research
LONI	Laboratory of Neuro Imaging
MCI	Mild Cognitive Impairment
MOU	Memorandum of Understanding
MRC	Medical Research Council (UK)
MRI	Magnetic Resonance Imaging
NACC	National Alzheimer's Coordinating Center
NHS	National Health Service (UK)
NIAGADS	National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site
OECD	Organisation for Economic Co-operation & Development
PET	Positron Emission Tomography
PI	Principal Investigator
SBP	Swedish Brain Power
SEK	Swedish Krona (currency)
SNAC-K	The Swedish National study on Aging and Care in Kungsholmen
UK	United Kingdom
UK BiLEVE	UK Biobank Lung Exome Variant Evaluation (project)
US	United States
WHO	World Health Organization

Note: Throughout the report, interviewees are referred to by name only. For a list of their affiliations, positions and relationships to each of the four case studies, please refer to the Appendix (p. 98).

1

Setting the stage: How can big data help dementia research?

Currently, about 44 million individuals worldwide suffer from dementia, making it one of the most prevalent age-related neurodegenerative diseases. This number is expected to nearly double by 2030, and triple by 2050. On average, one new case of dementia is diagnosed every four seconds (Alzheimer's Disease International, 2014).

While dementia – with Alzheimer's Disease (AD) as the most common form – is a devastating diagnosis for individuals and caregivers within their families, it also represents a major growing economic burden. The costs of dementia are estimated at 604 billion USD annually, a number that is expected to rise drastically over the next decades. Only a small proportion of these costs stem from direct medical expenditure (16%), largely coming instead from informal and formal care that people with dementia receive (Alzheimer's Disease International, 2010).

To tackle the challenge of successfully fighting dementia, the G8 nations came together at the Global Dementia Summit in London in December 2013. At this meeting, UK Prime Minister David Cameron called for action with respect to the various challenges in dementia research: from the "market failure undermining dementia research and drug development" to the need for more investment (UK Department of Health, 2014). Specifically, he highlighted the issue of data sharing and collaboration:

> *[There is] a lack of collaboration and openness with different scientists all over the world using different data and trying different approaches but frankly not really working together enough.*
>
> – **David Cameron**
> Global Dementia Summit, 2013

Purpose of this report

This report responds to several actions that the G8 health ministers committed to in the G8 Dementia Summit Declaration in December 2013 in order to accelerate international dementia research. Specifically, we aim to inform the aims to "identify strategic priority areas, including sharing initiatives for big data, for collaboration and cooperation", "encourage open access, where possible to all publicly funded dementia research" and to "take stock of our current national incentive structure for research", all of which may contribute to "the ambition to identify a cure or a disease-modifying therapy for dementia by 2025" (G8 UK: Global Action Against Dementia, 2013).

In previous research, the Organisation for Economic Co-operation and Development (OECD) has defined a range of areas for action to unleash the power of big data for dementia research and more efficient data sharing. These include the need for compatible data governance frameworks, financing for innovation and sustainability in the long term, means for sharing and linking data, managing patient consent, timely dissemination of findings and data, as well as open data strategies to accelerate innovation (OECD, 2014).

Building upon this groundwork, this report specifically focuses on using and sharing different kinds of data for dementia research. For this purpose, we have examined four case studies of data sharing initiatives – ADNI, AddNeuroMed, UK Biobank and the Swedish Brain Power studies – to examine how data are used and shared across research groups, universities and countries. This report is intended to give an overview of existing efforts and practices. At the same time, we use the findings from these case studies and insights from other experts as a basis for developing recommendations to advance dementia research going forward.

The scope of this report is the area of biomedical research, with links to patient care being made where these intersect, without however covering the care aspect more extensively. While the focus lies specifically on dementia, we also intend to inform research practices for age-related neurodegenerative diseases in general, and suggest ideas for using and sharing big data across medical domains.

1.1 Dementia and the current state of research

Dementia is a syndrome of chronic or progressive nature in which cognitive functioning deteriorates beyond what might be expected from normal ageing. It affects different aspects of everyday life, and comes in a variety of forms, often as a result of other diseases or injuries that directly or indirectly affect the brain (World Health Organization and Alzheimer's Disease International, 2012). The most common types are Alzheimer's Disease, vascular dementia, dementia with Lewy bodies, and frontotemporal dementia.

Currently, neither a cure for dementia nor a reliable way to slow down its progress exists. Researchers have a range of hypotheses about the processes causing dementia and a good understanding of the contributing risk factors, but more research is needed to develop mechanisms to stop or decelerate its onset:

> *The key aim of dementia research is to identify the mechanisms underlying neurodegeneration because identifying those will provide drug targets, and we need drug targets that will engage very early in the dementia process. In one sense neurodegeneration is with us to stay but we want to slow it down so much that nobody dies from it. I think that is the target. And I think it's completely do-able.*
>
> **– John Gallacher**

The multi-faceted nature of dementia makes understanding this and other age-related neurodegenerative disorders a complex undertaking. For example, dementia is believed to be influenced by a multitude of factors "below the neck", but also genetic and environmental influences:

> *We have insights from cardiovascular diseases, and use that for the dementia research, and that has been helpful. We need a multifactorial approach, one that respects life course and incorporates many new modifiable risk factors. It is good that we are collaborating not only with dementia researchers, but also researchers from other fields.*
>
> **– Miia Kivipelto**

Researchers have looked into a variety of data sources to find biomarkers: measurable indicators of disease status and therapeutic effect. Their aim is to detect early biological changes even before the clinical symptoms start, with the aim of better understanding the mechanisms, and ultimately being able to intervene and influence the condition's onset.

For now, it is clear that discovering biomarkers is a complex task. Further data analysis is required, while personalised medicine as a data-enabled way of tailoring treatments to individuals may also offer benefits for advancing dementia research:

> *It's important that we stop looking at trying to find one cure for one disease and trying to find one biomarker. [...] It's so heterogeneous and complex that we probably need to look for a combination of biomarkers. And maybe we can use different therapies for different types of groups. I think the reason why so many trials have failed is maybe the medicine works on some subjects but not on others, because you can have a similar clinical manifestation, but the causes are slightly different. [...] We need to think that maybe it's more complex than we think.*
>
> **– Eric Westman**

1.2 The promises of big data

With advances in data management (falling costs of obtaining and storing data, the routine collection of data becoming ubiquitous, and research collaboration in networked environments being possible) and technological progress in the medical sciences, new opportunities for dementia research have arisen. With the question of how to define big data being debated widely across contexts (Schroeder, 2014), we adopt a broad definition of big data, including new forms and sources of data, but also novel analytical approaches as outlined below.

What is big data, and why is it relevant for dementia?

At the core of dementia research are **datasets obtained purposively in medical settings**, such as images from Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET), clinical, genetic, proteomic and biological data from blood or cerebrospinal fluid (CSF), and cognitive tests or surveys. These datasets, which may not necessarily be "big data" in the sense of data that accumulates at high velocity and occurs in a high variety of forms and structures, but rather only "big" in terms of volume, also represent the focus of this report.

By big data, we also understand the **inclusion of routine data** linked in from electronic medical records (EMR), registries or other data within the health system. These data have already been, and will continue to be, collected for a variety of non-research purposes, and offer advantages in terms of longitudinal availability:

> *The economics of research is not great, and the beauty of routine data is that data can be deployed for a secondary use: it has already been paid for. If you really want to help the world of dementia, you need big longitudinal data for historic research, and to find the root cause of what happened to people ten years ago who have been diagnosed with dementia today.*
>
> – **Nicolaus Henke**

Moreover, big data also goes **beyond measures traditionally obtained** in medical settings due to a changing context of what is medically relevant:

> *Dementia is an outcome rather than a disease. Onset is influenced by biological factors, education, exposures and lifestyle. Outcomes often are determined as much by social context as they are by any treatments delivered within the confines of the medical care establishment. The benefits of big data lie in the potential for learning from our experience in this complex ecosystem – creating a "learning medicine". Because of the broad range of interdependencies, what is medically relevant data now has quite fluid boundaries, and researchers' perceptions of what is important also will be changing over time.*
>
> – **Paul Matthews**

On a few occasions, we look at big data in the sense of **un-/structured data obtained on-/offline in a routine manner**: for example, mobile phone data, loyalty card data, or banking data. While our case studies (and dementia research more widely) have not yet made use of these data sources, they may give insights into lifestyle habits, which are increasingly being recognised as factors influencing neurodegeneration. Moreover, this data may help to identify individuals earlier:

> *I understand loyalty card and consumer data very well, so the thing that really rang home for me was that before you're diagnosed and perhaps have early signs, you are starting to develop coping strategies. One of the coping strategies in the early stages is to start showing more regular or habitual behaviour: So people start doing the same thing over and over again, as a way of having a routine. And it struck me that that's where loyalty card data could be hugely valuable.*
>
> – Clive Humby

Especially with the **Web 2.0 and greater participation** from individuals in content production and sharing, those affected by early stages of dementia or mild cognitive impairment (MCI) and their caregivers may become data providers, and feed back data into research to strengthen the links between cure and care:

> *There is certainly the case for a big social engagement agenda, a model of open data for patients with dementia, online communities sharing information about the impacts of therapies on their conditions and day-to-day lives, enabling real-time research.*
>
> – Nick Seddon

Finally, beyond a specific kind of data we also include **novel analytical approaches**, such as data mining in which finding correlations – which may not necessarily imply causality – is the overarching paradigm (Mayer-Schönberger and Cukier, 2013). While these techniques have to be followed up by rigorous medical research and clinical trials, they may serve as a point of departure for further investigations:

> *When you go into Google, they actually look at everything you're doing. You already have a profile which means you get an advertisement which is for you. Also for disease processes this kind of analysis could be used to predict [...] which people are going to get dementia and maybe why. The common disorders probably have a lot of different factors, and it might be that if we use these approaches then we might be able to look at the combination of factors which lead to the disorder in different individuals. [...] It's not enough, you always need to confirm data or go in maybe more detailed old-fashioned analysis, but this is one new method which might be important. I generally don't think it's one thing which would solve the problem. We need to have combinations of different ways of analysing data and doing studies.*
>
> – Ingmar Skoog

What are different forms of big data in dementia research?

A common distinction in research data is between **broad data** and **deep data**. While broad data relates mainly to the number of individuals contained in a dataset, deep data is about how many and what measures and are included per individual. Technological advances have created a favourable ground for working with both types over the last decades: The cost (and also speed) of genotyping has fallen by 60% annually between 2001 and 2011 (OECD, 2014), new developments in MRI/PET imaging have been achieved, and storing large-scale data has become cheaper. At the same time, researchers have developed an appreciation for having both broader and deeper data, which are useful especially when used jointly:

> *You need both size and detail. You need to have large enough numbers so that your random errors are small and enough information about the exposures. It's the combination, the genotyping information, the environmental information, the clinical measures – and then enhancing that with imaging – that is valuable.*
>
> **– Sir Rory Collins**

Especially with routine data, data may come in different forms and a variety of structures. For this data, novel analytical approaches are especially fruitful, but are also sometimes seen as more demanding due to the high number of stakeholders involved, complexity and data quality:

> *There is a lot of noise in medical data. [...] Clinical observation of the human body is often subjective or not standardised and all patients can be argued to be different. The documentation is also often unstructured even in EMRs, [where] the data complexity is high and signal to noise often low. That's why the big data approaches must be different and are more demanding in medicine than they are in many other facets of life.*
>
> **– Stefan Larsson**

Larger datasets, independently of whether the size results from broader or deeper data, create new challenges in relation to how the data are collected and managed, which we will revisit in our case studies. However, in line with Borgman (2012), while big data offers new promises and size enables statistical confidence, "small data" also still often plays a key role depending on research question:

> *What's important though is not to get seduced by scale. [...] Big data is a very contemporary buzz word at the moment, and can actually lead to rather big errors if we don't think carefully about how we handle it. [...] So it's about how we share and make use of that data, and big data is an example of knowledge or information complexity, [while] we should not ignore smaller, targeted datasets in this conversation.*
>
> **– John Williams**

What do we mean by sharing big data?

With the recent Open Data movement, more datasets have been put into the public domain, particularly with regard to data aimed at increasing government transparency and participation (Huijboom and Van den Broek, 2011), while protecting information flows has equally been a topic of many debates (Gutwirth et al., 2013). Yet, openness is not a simple binary condition, and hence **data sharing means different things to different people**.

An **organisational** question is, for example, whether there needs to be a collaboration between those using the data and those who collected it. This has been the classic model for centuries and continues to be so, but here we focus on sharing that can occur without necessarily expecting active collaboration as a prerequisite.

From a **technical** perspective, sharing means "having access to the data". This may be through physical exchange of the dataset via storage mediums or networks or getting remote access to a server with the dataset and analytical tools.

Finally, from an **access policy** point of view, our definition of sharing includes open access datasets at one end of the spectrum, as well as those only being shared following an application and review process of varying degrees of detail.

We will take a closer look at the advantages and disadvantages of technical and organisational models of sharing throughout the analysis of our case studies.

Why is sharing big data important for dementia research?

Sharing big data for dementia research promises to unlock value through two levers: It has the potential to create deeper, broader, more complete and simply better data, and allows more researchers to use these data in innovative ways.

Better data through sharing partially creates value through pooling data and thereby obtaining larger datasets, which are necessary to detect biomarkers with statistical significance. Similarly, sharing data also creates the possibility to replicate findings across multiple cohorts, which is often necessary to validate findings:

> *The value is in the statistical power. If you want to get to really big numbers, you need to combine data across projects. It also allows you some variation in the methodology, so that the replication of a finding becomes more robust.*
>
> – **Arthur Toga**

Value may also come from linking to other (routine) data, such as diagnoses from other medical care settings or information about social circumstances from other sources. With some serendipity in the research process, even seemingly irrelevant data may lead to new lines of thought to be followed up with other research:

> *[Combining data] gives new classification ideas. It gives new diagnostic ideas. It gives new ideas about perhaps new biomarkers for some specific dementia*

> *diagnosis. You could see that frontotemporal lobe dementia has quite another pattern biochemically than Alzheimer's. I think by having large databases and being able to combine things [...] we could perhaps even see that some existing drugs are slowing down the progression of the disease. I mean, it comes to a lot of surprises.*
>
> – **Bengt Winblad**

More researchers using the shared data allows for more thorough exploitation and examination of the data by those who otherwise would not have access, including researchers from other disciplines or data scientists bringing their analytical knowledge into the medical world. This enables getting better insights from the data, while also using the funding obtained to collect the data more efficiently:

> *So the biggest advantage [of sharing data] is clearly exposing data to a wider community who can bring analytical expertise, ideas and resource time, and get better use – or at least more use – of the data.*
>
> – **Simon Lovestone**

> *Turning depletable resources or samples or images into accessible information has a number of advantages, but one major one is that it democratises the resource. [...] What we want are people with the imagination to use the resource. That's the whole point of UK Biobank: it increases the range of imaginations that can be applied to it.*
>
> – **Sir Rory Collins**

Even if sharing medical data is not risk-free, it enables the resources to be used more frequently, something that manifests itself in more publications from the shared data, all of which may contribute in one way or the other to tackle dementia:

> *How do you weigh the benefits of greater sharing against the disadvantages of greater sharing? [For ADNI] we have now eight years of data in which we can see, at least in terms of the large number of publications, some concrete metrics on what you might consider to be the advantages of sharing. While there is always the possibility that sharing could lead to a breach of privacy, that has not happened thus far.*
>
> – **Robert Green**

1.3 Lessons from other disciplines and areas

Of course, using and sharing (big) data is not an issue specific to the domain of dementia, but one that has long been debated in other contexts. Complex and large datasets have long been used in the world of the natural sciences, for example, from which a few initial lessons can be learnt.

Data sharing across the disciplines

Recent years have seen an array of initiatives to promote data sharing across disciplines. These range from creating international alliances to build social and technical bridges to enable open data sharing (Research Data Alliance, 2014) to policies around open access publication with datasets, to funding bodies providing incentives for or mandating data sharing (National Institute of Health, 2003).

Research looking at barriers to sharing research data found that the reasons included insufficient time, lack of funding, a lack of rights to make the data public, no place to put the data, a lack of standards, missing requirements from the sponsor and that there was "no need" for the data to be shared more widely (Tenopir et al., 2011). More fundamentally, scientists rightly want credit for their discoveries, and there are persistent concerns that releasing data can result in one being scooped by others, even if this is a rare occurrence in practice.

A related challenge is that incentives for sharing are misaligned, as most academic credit still accrues mainly to people who write highly cited academic papers or file lucrative patents. There are few mechanisms to document or reward the creation of valuable infrastructure such as databases and the data that fill them.

Finally, there are practical hurdles: Sharing data is difficult, since shared data must be more fully and clearly documented than internal data often are (with many studies relying on institutional memory for aspects of how to use their data), and creating metadata and exhaustive documentation are expensive tasks that have rarely been budgeted for at the outset of a project. These activities also generally occur at the very end of the project, when resources have been largely spent.

In addition, health research comes with its own complications, both for research data and other routinely collected data. Most saliently, human data needs to be protected to ensure patient confidentiality and privacy, and to comply with the consent provided. This has to be balanced with making the data widely available, so as not to hamper progress in medical research.

Over the last decade, there have been successes in sharing medical research. One relatively early example in the US was the Genetic Association Information Network (GAIN) project, started in 2006 to share phenotypic and genotyping data on a variety of medical conditions. One innovative feature of this effort was that all researchers (whether or not they were part of the project) were provided with the data simultaneously. Unlike the typical pattern of the originating researchers having exclusive access to results for a period of time, everyone (including pharmaceutical companies) had the same access, and contributing researchers had just a six month window of exclusive right to publish (Meyer and Schroeder, 2015).

A more recent example in the UK is the Farr Institute of Health Informatics Research, pushing forward the use of EMR data, research data and various sources of routinely collected data, and enabling links between the National Health Service (NHS), universities and industry for data sharing. The centre also aims to innovate methodology, build capacities and promote a new culture of engagement between the public, patients, clinicians and researchers (Farr Institute, 2015), which is crucial especially when using big data from a variety of sources in the future.

Big data across the disciplines

Beyond sharing, other areas and disciplines provide insights into how big data is used. Earlier projections estimated that even with rapid growth in recent years, big data is still in the process of reaching its full potential (Groves et al., 2013).

Across academic disciplines, there has been variable uptake of big data approaches. Fields like high-energy physics have been dealing for years with massive amounts of data generated by technologies such as the Large Hadron Collider, but big data approaches are not limited to these large projects. An example from astrophysics arose from the difficulty of dealing with increasing quantities of image data generated by the Sloan Digital Sky Survey, for example. While purely computational approaches work well with many kinds of data, with photographs of galaxies the human eye and brain are better at the pattern recognition required to categorise the galaxies. The Galaxy Zoo project (Lintott et al., 2008) was established to solve this problem by enlisting the aid of citizen scientists, who participate in science by classifying galaxies online. Since this early very successful project, the creators have expanded the Zooniverse to include projects on climate, nature, history, and other areas that can benefit from contributions by citizen scientists.

The social sciences have also been using big data approaches, ranging from forecasting human behaviour using global news media (Leetaru, 2011), to predicting movie ticket sales from Wikipedia activity (Mestyán et al., 2013), to using Google search data to predict everything from flu (Lazer et al., 2014) to house prices (Wu and Brynjolfsson, 2013). Likewise, the humanities have been using big data for analysing millions of books digitised as part of the Google Books scanning project (Michel et al., 2011), or enlisting hundreds of contributors to make an annotated Wiki edition of Thomas Pynchon's novels (Schroeder and Den Besten, 2008).

Specifically for research and development in the health area, big data may be a crucial enabler for personalised medicine or new forms of analysing trial data (Manyika et al., 2011). However, integration of different data sources in healthcare is difficult, both as data capture may be distributed across stakeholders in the health system, and due to the (desired) separation of patient data from data outside the medical realm, pointing to issues around privacy and security. Big data for biomedical research raises a whole set of new ethical issues, covered extensively in the Nuffield Council on Bioethics (2015) report.

Before we dive into the details of how using and sharing big data can contribute to advancing dementia research, let us first explore what resources for dementia research exist, and introduce our four case studies for this report.

2

The landscape of available resources today

Data relevant to dementia research have been collected over the last decades, and shared and aggregated in various ways. While a complete account of efforts goes beyond the scope of this report, this section provides an overview of some of the major resources available today, and introduces a selection of the most salient data sharing initiatives as summarised in Figure 1.

2.1 The evolution of data sharing initiatives

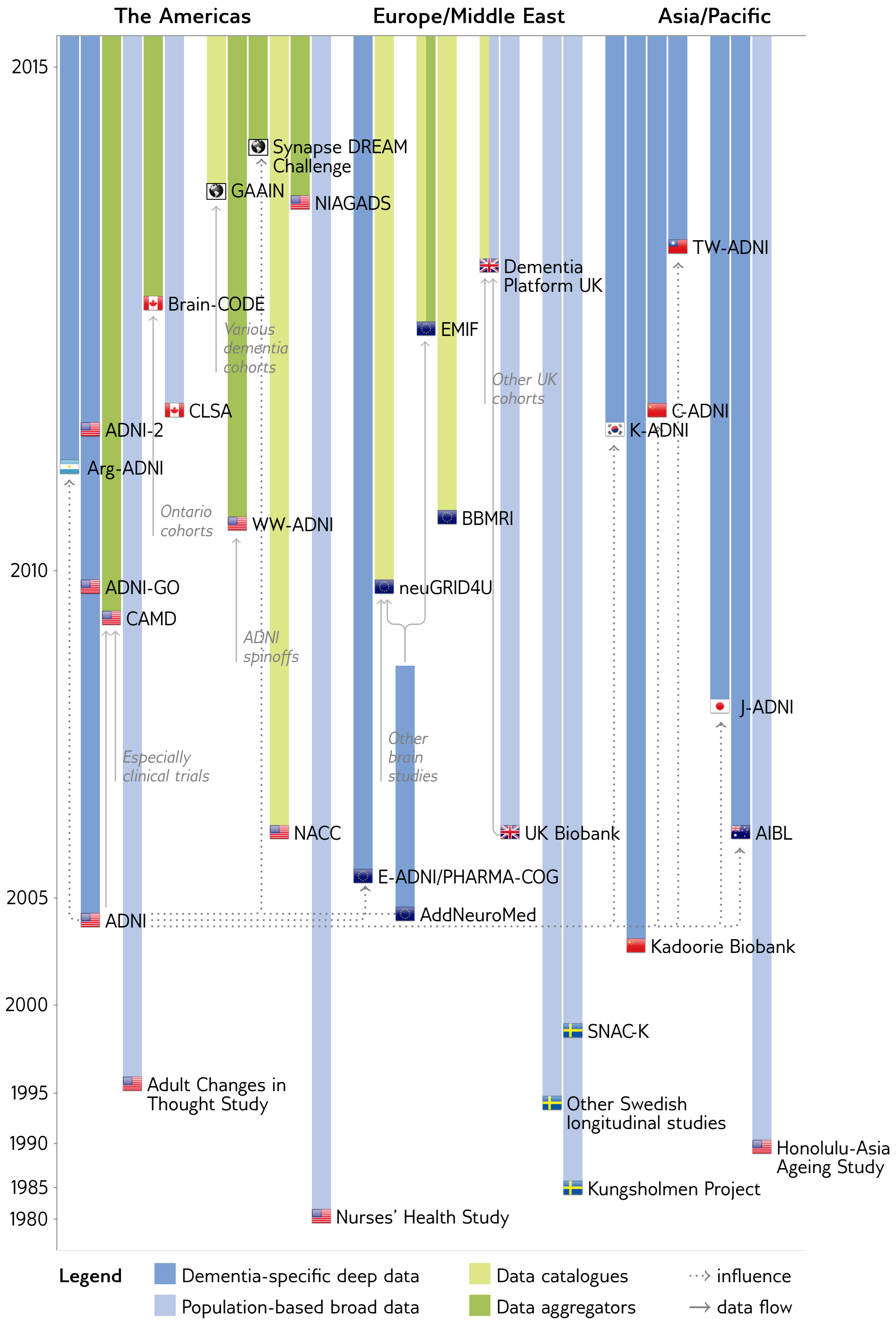
Historic roots

Historic accounts of dementia can be traced back to ancient times, but medical analyses into dementia started in the last century, with Alois Alzheimer's discovery based on one of his patients in 1906 among the most notable events. In the latter half of the 20th century, scientists developed the now prevailing understanding that dementia is commonly found in older people, but is not a normal part of ageing (Katzman, 1976; Ritchie and Kildea, 1995; World Health Organization and Alzheimer's Disease International, 2012).

While the early analyses were based on specific cases only, the first large-scale data collection efforts relevant to dementia research started in the middle of the last century. While these studies may not include the richness of information that **deep data** features today, countries like Sweden count among the richest resources of longitudinal **broad data** in terms of population-based studies, longitudinal cohort studies and administrative data that can be linked.

In the 1970s and 1980s, a number of large epidemiological studies were started in different parts of the world. For example, the Honolulu-Asia Aging Study, the Nurses' Health Study, the Adult Changes in Thought Study and the Kungsholmen Project all gave new insights into ageing, and are the foundation for a lot of the current evidence on dementia (Alzheimer's Association, 2014).

Figure 1: Evolution of selected major data sharing initiatives



Timeline approximates beginning of data collection/foundation (without preparatory phases)

The new millennium

The beginning of the new millenium marked several new developments relevant for data sharing and dementia research.

First of all, initiatives to collect observational, dementia-specific **deep data** started both in the US, with the Alzheimer's Disease Neuroimaging Initiative (ADNI), and in Europe, with the AddNeuroMed project for the discovery of novel biomarkers for Alzheimer's Disease. ADNI soon began inspiring spinoffs in other regions, beginning in the EU with E-ADNI and the Pharma-COG studies for predicting the cognitive properties of new drug candidates for neurodegenerative diseases in early clinical development, and AddNeuroMed adopting the imaging protocols that ADNI used. In 2006, ADNI protocols were also used in the Australian Imaging, Biomarker and Lifestyle Flagship Study of Ageing (AIBL), with many more countries (Japan, Argentina, Korea, China, Taiwan) to follow.

With the appreciation that many diseases originate from more than a single defective gene, and partly as a result of decreasing costs for obtaining and storing large amounts of data, a new era of biobanks started at the same time (Swede et al., 2007). While biobanking is not a new phenomenon, the novel aspects are the digitalisation of data instead of storing tissues only, and the magnitude of population-based **broad data**. Large initiatives such as the UK Biobank or the Chinese Kadoori Biobank collect data on upwards of 500,000 individuals, with the aim of creating large prospective cohorts that gain value in the future.

Beyond data collection and sharing initiatives, the new millennium also featured more collaboration between public and private actors. ADNI and AddNeuroMed piloted public-private partnerships, among others, with the pharmaceutical industry, which at that time faced the costly challenge of numerous clinical trials for dementia drugs failing in the late stages. In 2005, the Critical Path Institute (CPI) was founded to unite scientists from the Food and Drug Administration (FDA), industry and academia to improve drug development and the regulatory process for medical products, and initiated the Coalition Against Major Diseases (CAMD). The Innovative Medicines Initiative (IMI) created similar bridges in Europe.

Moreover, a growing number of consortia and networks of collaboration started to coordinate research and funding. Country-level initiatives such as the Swedish Brain Power network or groups based on the French Sarkozy plan appeared. Other networks that came into being include the European Alzheimer's Disease Consortium (EADC), the Centres of Excellence in Neurodegeneration Research (CoEN), or lately also European initiatives with a global focus, such as the EU Joint Programme on Neurodegenerative Disease Research (JPND).

Recent developments

In recent years, dementia research has seen two major developments aimed at making data more available and bringing together **broad and deep data**.

Data catalogues to enhance transparency and make data with specific characteristics findable were introduced. For example, neuGRID4U is a web portal to help find datasets on neurodegenerative diseases based on specific characteristics of the participants, aiming to become the "Google for brain imaging" (NeuGRID4U, 2014). Similarly, the Global Alzheimer's Association Interactive Network (GAAIN) brings together clinical, genetic, imaging and proteomic data from various sources.

Moreover, **data aggregators** – often with further data enhancement – were established in various parts of the world. For example, the UK dementia platform integrates 22 different cohorts, one of them the UK Biobank cohort with 500,000 individuals, and will add further data over the coming years. With a similar aim, the Canadian Longitudinal Study on Ageing (CLSA) has created a new prospective cohort including broad and deep data for researching age-related diseases.

Of course, some recent initiatives such as the European Medical Information Framework (EMIF) also combine elements of data catalogues and aggregators. The spirit of aggregation is also evident in the biobanking area, with the European Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) aiming at aggregating data from smaller, traditional biobanks. On a local level, BRAIN-Code at the Ontario Brain Institute makes data from all over Ontario accessible.

With crowdsourcing being closely related to approaches in big data analytics, datasets have also been aggregated to bring in researchers from beyond the realm of dementia. For example, the Synapse DREAM Challenge has offered a prize for ordinary users with data analytics skills to build and analyse models of dementia based on the North American ADNI data and test them on AddNeuroMed data.

More generally, funders and charities have increasingly recognised the importance of using and sharing data for dementia research, for instance with the Alzheimer's Association funding GAAIN. Several business leaders have initiated movements that have contributed to raising awareness and delivering ideas about how data may be combined and shared more efficiently. Most notably, this includes the Global CEO Initiative on Alzheimer's Disease led by George Vradenburg in the US, and the UK-based Evington Initiative around Marc Bolland, Richard Cousins and Terry Leahy.

2.2 Selected case studies

From among these existing data sharing initiatives for dementia research, we selected ADNI, AddNeuroMed, UK Biobank and the Swedish Brain Power studies (as an umbrella term for the various longitudinal datasets in Sweden) as case studies. The six key considerations in selecting the cases included:

- **Geographic coverage:** With a balance between single country and international initiatives
- **Focus:** With representatives of dementia-specific, brain-focused or cross-conditional data repositories (also in relation to broad versus deep data)

- **Size:** With a variety of sizes as measured by number of participants, data providers, public/private partners and the amount of funding to date
- **Maturity:** Considering the start of data collection in relation to the lifespan of the initiative, and how long data has already been shared
- **Openness by design:** Degree to which the initiative had the intention to share data from the outset
- **Linkage to routine data:** Whether the case includes an established connection to other routine data in the health system, such as EMRs, registries, or big data from outside the health system

In selecting the four case studies, we aimed at including cases that would represent the spectrum of ongoing activities, but that would also differ on dimensions that influence how data are collected and shared. Table 1 summarises the main characteristics for all case studies, each of which will be introduced subsequently.

Table 1: Overview of case studies

	ADNI	AddNeuroMed	UK Biobank	SBP studies
Geographic coverage	🇺🇸 National (+ spinoffs)	🇪🇺 International (6 EU countries)	🇬🇧 National	🇸🇪 National
Focus	Dementia	Dementia	Universal	Brain-focused
• Broad data			✓	✓
• Deep data	✓	✓	(✓)	(✓)
Size				
• Participants	835[1]	700	500,000	up to 50,000[2]
• Data providers	58 (US)[1]	6 (Europe)	22 (UK)	n/a
• Partners	28[1]	18	4	6
• Approx. USD funding	150 million[1]	22 million *18 million EUR*	131 million *84 million GBP*	26 million[3] *200 million SEK*
Maturity	established	late stages	early stages	established
• Data collection	2004–	2005–2008	2006–	1960s–[2]
• Data access	since 2005	since 2009	since 2013	ongoing[2]
Openness by design	✓		✓	(✓) *within Sweden*
Linkage to routine data			(✓) *EMRs (hospital)*	(✓) *EMRs, registries*

[1] Data refers to US ADNI only. WW-ADNI is a more recent effort to combine country initiatives.

[2] Depends on specific study within the Swedish Brain Power network.

[3] Relates to network funding (without studies) only, hence comparability to other cases is limited.

ADNI (Alzheimer's Disease Neuroimaging Initiative)

Our first case study, ADNI, which also served as a pilot for this project, is an initiative founded by Michael Weiner in the US in 2004. Its aims are to help predict and monitor the onset and progression of Alzheimer's Disease, to establish global standards to identify and document cognitive changes, and to share data across the international research community. The North American ADNI has received more than 150 million USD of funding from 28 members of the public-private partnership, and includes imaging, genetic and clinical data from medical examinations and cognitive tests.

For the North American ADNI, data are obtained at 58 sites across the US, with 2,469 participants examined since its first iteration, and 835 participants enrolled in 2014 – both with Alzheimer's Disease, MCI and healthy controls. It has produced country spinoffs across the world, which the initiative WW-ADNI aims at uniting. Currently, the North American ADNI and the Australian ADNI (AIBL) are accessible via the Laboratory of Neuro Imaging (LONI) at the University of Southern California. Japan is supposed to follow soon. In 2014, ADNI counted around 5,400 single users, of which about 80% came from academia, with biotech, government, pharma and scanner manufacturers making up the other 20%.

ADNI has made a contribution to at least 1,971 publications worldwide, making it the most highly referenced data source of the four cases we examined. Using Scopus data gathered by searching for references to ADNI, there has been a consistent upward trend in publications and citations. These publications include authors from the US (n=1,138), Europe (n=982), Canada (n=133), China (n=120), Australia (n=84), and a range of countries with fewer publications. These papers have in turn been cited over 15,000 times, with over 5,000 citations in 2013 alone, indicating that ADNI will have continuing influence in the coming years both in terms of direct contributions to publications and the impact those publications have.

About 70% of publications do not have any author who is an ADNI contributor, showing the success of widespread open data sharing. ADNI is also the most globally distributed resource in terms of author affiliation of publications, partly because of the ADNI country initiatives. There are relatively few authors located in non-ADNI countries, which suggests the potential for engaging new researchers if national barriers to sharing data can be more readily overcome, both in terms of regulation but also in terms of mindset, which often results in researchers looking primarily within their own national context for resources and partners.

AddNeuroMed

At around the time ADNI was created, European researchers led by Simon Lovestone as the principal investigator (PI) started the AddNeuroMed initiative, which ran as a public-private partnership. While it contained both pre-clinical and clinical elements for dementia research, about 700 patients from 6 centres across

Europe were enrolled in the clinical part between 2005 and 2008. While ADNI and AddNeuroMed are similar in terms of the number of individuals enrolled, and how broad the data are, they differ in depth of the data. This also relates to differences in funding, with AddNeuroMed's funding of around 18 million EUR (around 22 million USD) being a magnitude smaller than ADNI. Based on ADNI's initiative to establish common global standards around imaging, AddNeuroMed adopted the imaging guidelines from ADNI.

In contrast with ADNI, AddNeuroMed adopted the traditional model of doing research and therefore was not intended to be open from the outset. Currently, the AddNeuroMed dataset can either be obtained by emailing the PI, or is accessible via research platforms including EMIF and neuGRID4U.

AddNeuroMed publication metrics are a bit less clear-cut than the other cases. Searching for publications making reference to AddNeuroMed on Scopus only yields 31 results. However, using Google Scholar (which is less precise but more inclusive both in terms of the parts of the document it searches, and the types of documents indexed) shows 469 publications of some type referencing the AddNeuroMed data. The authors of the Scopus papers are heavily weighted toward AddNeuroMed project participants, although there have also been authors outside the consortium publishing on the data. The authors are largely located in Europe (n=192), with a much smaller number in North America (n=13). Citations to these publications (n=704) have remained relatively steady from 2012-2014.

UK Biobank

UK Biobank is an initiative led by Sir Rory Collins that has collected broad data on 500,000 individuals in the UK with deep data about a range of conditions to facilitate data use for a multitude of different purposes in the future. Additionally, UK Biobank data are being enhanced in various ways, for example with the addition of imaging data at an unprecedented scale (100,000 individuals, repeat imaging for 10,000 individuals). A special feature of the UK Biobank study is that individuals are not only followed up over a long period of time (at least 25 years), but also that most of this follow-up is done through linkage with other forms of routine health data from EMRs in secondary and – in the near future – primary care.

As of December 2014, about 1,000 users had registered with UK Biobank, with about 200 research applications having been made, out of which about 50 have been provided with the data so far. UK Biobank attempts to grant access to all bona fide researchers, with rejections being made when depletable resources are at risk of being used up (or not used for convincing reasons). If individuals have to be reapproached for further data collection, application decisions may be postponed in order to bundle efforts.

As UK Biobank data has only been made accessible recently, and prospective cohorts gain value over time, relatively few publications are currently based on UK Biobank data. According to Scopus, 109 publications make reference to "UK

Biobank" in the title, abstract, authors, or affiliations of the publication. The UK Biobank website lists 27 publications using the Biobank resource, 8 publications by the UK Biobank staff, and 9 about the UK Biobank as a resource. The majority of the publications are from 2014, so little can be concluded yet in terms of publication impact, although there are a total of 996 citations to the publications in this set. The top 10 author affiliations are all located in the UK (total n=79), although there are a small number of publications with authors in the US (n=15), Canada (n=7) and other parts of Europe (total n=26). For a resource such as this, with the UK focus embedded in the name, it will be interesting to see in the longer term if the data are used to contribute to studies with more international, non-UK involvement.

Swedish Brain Power Studies

The Swedish Brain Power Network was established in 2005 as a research network to facilitate cooperation within Sweden. We use it as an umbrella term for the various longitudinal population-based studies in Sweden, such as the Kungsholmen study (continued as the Swedish National study on Aging and Care, SNAC-K), the Gothenburg MCI and H70 study, the OCTO-Twin Study, the Betula study on aging, memory and dementia, as well as broader prospective cohort studies like LifeGene.

What is special about the Swedish longitudinal data – similar to UK Biobank – is the ability to link data collected in longitudinal studies to other forms of routine data, such as EMRs and disease-specific registries to which health organisations within the country must submit data. Sweden also has a system whereby the whole of the population uses a personal identification number, to which a number of records in addition to medical ones – such as tax records – can be linked (see Axelsson and Schroeder (2009) for more details).

The publication data for the Swedish Brain Power Studies requires searching for a variety of projects that fall within the broader umbrella of the study (see above). Searching for publications making reference to these studies on Scopus yields 224 results. The authors are heavily weighted toward Swedish authors (65%, or 209 of the 321 authors), although there have also been authors from the United States (n=39, 12%), Italy (n=18, 6%) and a range of other mainly European and (to a lesser extent) North American authors. The number of publications increased by approximately 50% after the formation of the Brain Power network in 2005, with 84 publications from 1995-2004 and 132 publications from 2005-2014. Citations to these publications (n=3,694) increased slightly after 2005 and have remained relatively steady since, at about 250-350 citations per year.

All of the introduced initiatives have created valuable assets for current and future dementia research. The next chapter reviews current data governance, focusing on the selected case studies, with the aim of crystallising their key differences.

3

Current data governance

In the first part of our analysis, we examined "what makes good, usable data" for dementia research. The intention of this chapter is to juxtapose different data governance models adopted across our case studies, with the aim of highlighting a spectrum of possibilities and their benefits and drawbacks.

Data governance comprises technical and organisational aspects of how data are collected, stored, made accessible and tracked. We synthesise our findings for each of the four case studies along seven dimensions, as summarised in Table 2.

Table 2: Data governance dimensions (OECD, 2014)

Dimension	Description
Availability	Data needed for analysis must be available
Accessibility	Data should be accessible to those who need it
Interoperability	Data must be semantically/syntactically interoperable
Quality	Data must be accurate and complete
Traceability	Data must have a trail of data from its source
Privacy/security	Data must be kept secure/non-identifiable to others
Ownership	The rights to data use must be agreed

Of course, these governance dimensions are closely interrelated and in some cases must be balanced against each other. Also, for each of the dimensions, it is not a dichotomous question of being "fulfilled" or not, but rather a question of the chosen model, as will be presented throughout this chapter.

3.1 Availability

Finding: Large amounts of data relevant to dementia research are available, both within and beyond our case studies. Going forward it will be important to combine a longitudinal perspective before and after diagnosis on a large number of individuals, with a wide range of measures beyond the realm of dementia research to be included.

All four data sharing initiatives contain a variety of measures relevant for dementia research, and each has made valuable contributions to a range of research questions, as evidenced by their publications. Key differences in relation to data availability are the overall number of individuals, longitudinal coverage in terms of long follow-up periods and pre-diagnosis data, and the inclusion of sufficient metrics to make the study "future-proof" and be able to treat dementia as a whole body issue.

Number of individuals

As outlined earlier, medical research increasingly needs large numbers to find small effect sizes with a sufficient degree of certainty, also because few of the diseases with unknown onset mechanisms currently being researched are likely to be based on a single gene variant or a simple set of environmental risk factors. While this can be achieved through pooling, large studies such as UK Biobank with 500,000 individuals (and imaging for 100,000) already come with much broader data than smaller scale projects such as AddNeuroMed or ADNI, which (excluding data from the worldwide spinoffs) only include a few thousand individuals:

> *The imaging on 100,000 people will be unprecedented. When we first contacted the imaging community and we said we wanted to image 100,000 people, many responded saying "Very interesting but your email has got an error, you meant 10,000?" No, no. It took 5 years really for the imaging community to understand why it needed to be 100,000. Only a small proportion of the imaged individuals would develop any particular disease of interest, and therefore to have enough cases you have to have at least 100,000 imaged. [...] There is nothing else like that planned so in terms of competitive advantage it's scale and depth, particularly with the addition of the imaging.*
>
> – **Sir Rory Collins**

Of course, as implied in this statement, numbers may be smaller for dementia-specific research, where all participants by definition have the condition of interest. However, as the case of Swedish Brain Power studies shows, registry data – as a form of routinely collected data – is also available for large numbers of individuals, enabling cost-effective research with large numbers:

> *In our svedem [dementia] register we have now 50,000 patients followed. We have affiliated all memory clinics and I would say around 60 percent of the general practitioners who do assessments. If the clinic does not fill in this data then it gets punished and has to pay a sum of money to the county council. [...] So if you want to use large databases, I'd say you can do that without much money, because every patient is researched and we have the advantage to combine these registers instantly over the country.*
>
> – **Bengt Winblad**

At the same time, the discussion of dementia-specific versus the general population points to the important issue of how and when individuals were selected as research participants.

Longitudinal coverage and selection of individuals

ADNI and AddNeuroMed specifically focus on dementia research and provide valuable deep data including imaging, genetic, clinical and proteomic data. While both studies include individuals with Alzheimer's Disease, MCI and healthy controls, they cannot follow up individuals over a longer period as people die from the condition. Related is the caveat that individuals in advanced stages of the disease are already a selected subgroup of the population. Due to the lack of availability of longitudinal data, earlier events in their lives cannot be traced back.

In contrast, UK Biobank and the Swedish Brain Power studies with their population-based approach offer the advantage that researchers can look longitudinally at early biomarkers of those who later develop dementia:

> *UK Biobank, for example, is great because you've got a huge number of patients with a lot of standardised data, and they're followed up over a long period of time. So we're getting to early disease markers rather than looking at people who are in various stages of advanced disease.*
>
> – **Richard Dobson**

It is worth highlighting that this longitudinal data, as in the case of UK Biobank, does not necessarily have to come from repeated examination of the research participant. It may also be obtained based on routine data collected in primary or secondary care, or other sources of metrics of relevance for dementia research.

Inclusion of metrics

One of the challenges is the inclusion of the right measures that are useful beyond specific research ideas, which was challenging for AddNeuroMed given the initiative was not intended to be made open widely from the beginning:

> *If you were asking what would I have done if we'd have gone back at the time, yes, I would have added loads more. We would have put in PET imaging. We tried to do CSF but [...] it was much harder to do than we thought at the time. I'd have added longer follow-ups. There was all sorts of things which we wanted to have but we didn't have the funds to do, or we didn't know we should have wanted to have them but now we do.*
>
> **– Simon Lovestone**

A related challenge inherent to longitudinal studies is that, as medicine and medical technology progress, it is difficult to ensure that datasets are "future proof":

> *The challenge with most longitudinal studies is that they get outdated before they even begin. [...] So we said what are the most important emerging research questions in the area of ageing. [...] We generated over 200 questions which was impossible for us to accomplish, so we prioritised which are the most important questions that we would like to address in the next 10 years.*
>
> **– Parminder Raina**

Similarly, while the Swedish Brain Power studies comprise datasets with follow-up periods of several decades, some possibilities in imaging have only developed over the last years. Follow-ups via EMRs ensure that further measures are added over time, though routine data do not entail a great depth, of course, and hence may be combined with deep data obtained within research projects.

To achieve the necessary depth, UK Biobank is adding further data on their cohort, for example through activity tracking and whole body imaging. Additionally, avoiding a focus on dementia as a single condition may also help to broaden data collection beyond what is deemed relevant for dementia research:

> *It's not just brain imaging. Dementia is a whole body issue. The simplistic view of dementia would be "let's image the brain". I would suggest that that's actually wrong. [...] There may be things in the images of the rest of the body that are relevant to dementia.*
>
> **– Sir Rory Collins**

Untapped big data sources

While all studies have collected useful and important data that have already advanced dementia research, it is worth looking at other sources of longitudinal data with a wide range of measures beyond the specific realm of dementia research.

One of these avenues may come from other routinely collected big data, that has already been collected in the past and that thereby also offers a longitudinal perspective. Potential sources may be loyalty card data or wearable technologies such as activity trackers, giving insights into individual lifestyle habits, but also helping to pick up signals for detecting dementia:

> *You could combine banking records, mobile phone records, and then potentially some measure of activity through things like FitBit. And if you had peoples' shopping patterns, you'd have some idea of their diet as well. So you could potentially harvest a massive amount of data and then start to look for patterns, and potentially pick up early signs of dementia.*
>
> – **John Drew**

Making data from outside the medical realm accessible may at least in principle enable new methods of data collection, for example about diet. While UK Biobank has no plans to do so, it is worth highlighting the benefits:

> *For most chronic diseases, and dementia is a good example, researchers now are increasingly understanding that what people do, as compared to what doctors do or don't do, is probably more important. [It would be] valuable to find out more about what people are doing directly, whether that would be directly from participants, or linking into systems that sort of capture data in novel ways [...] Loyalty cards are a fantastic example. We know diet plays a huge part in long-term health and the development of chronic conditions. We know as epidemiologist that the methods we have, the questionnaire-based methods, are problematic as they are not very reliable, so actually looking at the question in a different way may be just the way we need to do that.*
>
> – **Tim Sprosen**

Big data from outside the medical realm also has its limitations, of course. What data are available and the potential they have may have to be carefully evaluated, as some data may also be overestimated:

> *Another area was the idea of whether you could look at constituents of products that people buy, but I think it's too hard to do. The data on products is much more hit and miss and the idea that a product has got a certain protein in it or similar looks very difficult to discover. I don't think this data exists at the moment. Technically it looks like it's known, but it isn't known. It's not necessary to run the business, so you don't collect it – but that may also be an idea.*
>
> – **Clive Humby**

To conclude, new forms of big data offer a value proposition due to their routine character in the sense of having already been collected over some period of time; methodologically, in terms of their collection as a byproduct without specific actions from participants; and due to extending beyond the medical sphere. However, in current dementia research they remain largely unused.

The value of this data may be especially harnessed in combination with existing medical data – while recognising that their combination, especially when the data are being used across different contexts, also creates new questions in terms of privacy and data protection, which we will cover later in this chapter.

3.2 Accessibility

Finding: With our case studies using a range of access procedures, key questions about data access are scalability, the levels of detail required for applications, traceability of data access, funding models, and how to balance fast access with delays due to data cleaning and potential embargo times. Two technical/organisational approaches, the reading and the lending library model, provide different benefits and drawbacks.

Access to data resources for dementia research varies according to what the procedures are; in terms of effort and detail required for application; and time between application and actual data access needed. Our case studies present a spectrum from fast and easy access online (ADNI) to collaboration with physical presence (Swedish Brain Power studies).

Access procedures

In terms of ease and speed of access to the data, ADNI has redefined the gold standard, with an application process to verify the bona fide status of the researcher requiring only a short description of the intended research. More than anything, this acts as a barrier to impede undue access and acts as a spam filter. Scrutiny is paid if researchers do not apply from institutional email accounts, or if doubts exist about the dementia-related intention of the data use. However, with this low bar, ADNI providers and users agree that data access is easy and quick:

> *There are no barriers at all, any qualified scientist in the world has total access to all ADNI data, without embargo. [...] Everyone has the same access as I do.*
>
> – **Michael Weiner**

> *Getting access to the ADNI data is fairly simple. You must agree to terms in the ADNI data use agreement stating that you will not try to identify participants, not redistribute data, agree to comply with local (institution) IRB rules in use of data and cite ADNI in any presentations and include ADNI among authorship. In addition you must provide an experimental plan and list of collaborators. In my experience the turnaround is quite quick, making it pretty simple process.*
>
> – **Mette Peters**

Access to AddNeuroMed is similarly easy, though with less sophisticated processes: Access is granted through email contact with the PI Simon Lovestone. As mentioned earlier, the dataset is also available through EMIF and the neuGRID4U platform.

Access to UK Biobank data already requires a somewhat more complicated four-step process, which, given that releasing data has only recent started, may accelerate as the access processes become more established:

> *The process has been quite time-consuming: filling in the form, waiting for approval for the initial application, then writing a fuller application and then learning how to use the website to click on the variables and select the variables and then paying for the data, and then finally having it approved, and then having the material transfer agreement sent out and found by different people in different institutions, pulling all that together. But I'm not really moaning about that because it's such an excellent resource [...] It could obviously be a bit more efficient but it's not really surprising if this is such a massive project and this is just the first year or two of people requesting data, then it's inevitable that there are going to be some teething problems, and I can see that things are being streamlined. So I'm very happy with it really.*
>
> – **Daniel Smith**

At the other end of the spectrum of the access procedures model are the Swedish Brain Power studies. The network fosters close collaboration within Sweden, with projects only being funded by the Swedish Brain Power network if it at least involves two partners, for instance.

Also outside of the country, Swedish brain researchers collaborate broadly with others all over the world and also with partners from the pharmaceutical industry. In terms of getting access, this focus on collaboration means that researchers are usually asked to come to Sweden, look jointly at how to best use the intended dataset, and then take the dataset with them to their institutions afterwards:

> *We don't put our database on the web, as does ADNI, for example. We consider this unethical because we have clinical data with our diagnosis. So we cannot put this on the web. But we made it possible to do collaboration. [...] Usually we have always a researcher coming here for a period just to get things going. And then the study is working, it's provided for a period. So the person can continue to work in his or her own home. But if [...] you need really a lot of data, then I would really like to have the analysis here.*
>
> – **Laura Fratiglioni**

While we do not seek to identify the best model, some characteristics of getting access can be highlighted here.

The first is the question of scalability. Emailing the PI to get access to the dataset works very well for a single project, it can hardly be extended to a larger scale. Even more so, while the Swedish model of accessing the data in person ensures proper understanding of the data and close collaboration with researchers who know the resource (which in itself is very desirable), Swedish data may not have its full potential exploited as a consequence:

> *The trouble is that we have too much data. And we can't even make the best use of it now because we need more people, need more funding to support more people to work with all this data we have been collecting.*
>
> – **Linda Hassing**

A second question is whether access procedures need to be complicated and detailed based on the wide consent that especially ADNI and UK Biobank have been given in order to enable "health-related research in the public interest", or whether the ADNI model of low bar access is sufficient:

> *In UK Biobank the area of concern is around [...] how tightly to define what research people do in a sense that they get approval to do a specific piece of research. [...] You don't want to constrain people unreasonably from applying their imagination to the data. On the other hand you want to ensure that the resource is not used for purposes for which consent wasn't given. We have very wide consent [...] so one can argue that does allow a lot of freedom in terms of what people can do. We are now going through a review process as to whether there are some applications that are so clearly low risk that we could have a much more streamlined approval process.*
>
> – **Sir Rory Collins**

Then there is a question of traceability of what was done with the data. While ADNI and UK Biobank keep a precise register of who applied and who published based on the datasets, with more informal procedures as in AddNeuroMed, little is known about its applications:

> *[The emails] dribble through, and every time we say yes. We don't manage it very well, we've never had any funds to do it so it's really expensive to give data out. You know, we do it gratis, as it were. What we don't do is follow up and find out what people have done with the data, you know, we haven't got the resource for that and there's a material cost.*
>
> – **Simon Lovestone**

This point also highlights the importance of funding not only the study itself, but the continued financial means to manage and monitor data access. Financing may either come through original funding (as in ADNI), specific funding through the establishment of a network for data sharing activities (as in the Swedish Brain Power Network), or from partial charging for data access (as in UK Biobank) – with potential combinations of any of these.

Access models

Beyond the specific access procedures, it is useful for our purposes to discuss different access models. Beyond collaboration where physical presence is required (as in the case of the Swedish Brain Power studies), we can identify two main models: The lending library model where datasets are exchanged (as has been adopted by ADNI, AddNeuroMed and UK Biobank) and the reading library model, where researchers get access to a remote machine with the data (as envisioned by the Dementia Platform UK, incorporating UK Biobank data, for example):

> *A lending library sends you the datasets, you do the analyses, if you are doing any original research with samples you have to provide back to them the data which it adds to its library. [...] In a reading library you have to go there either in person or via a portal to actually look at any of the datasets and that can cause problems especially if you are also going to be submitting data as well.*
>
> **– Ian Hall**

This highlights the key advantage of the lending library model: it gives more flexibility in terms of what can be done with the data, for example in terms of combining data with other sources. On the other hand, the reading library model may allow for more control over the data, and offer benefits in terms of the size of datasets and the traceability of analyses performed on them:

> *At the moment data is transferred to the researcher. [But] we're getting into volumes of data where that kind of physical transfer is just not going to be possible. UK Biobank are just now doing genetic analysis and using an accelerometer to measure physical activity, and all of those [...] require a different solution than basically sending people data.*
>
> **– Tim Sprosen**

> *[Another aspect] is accountability because by having all the analyses done centrally we can monitor, if you like, audit all the analyses. It's not that you want to routinely do that but the point is you can if you need to. There are two good reasons for this. One is scientific; we would like to be able to reproduce exactly someone's analyses if they are questioned. And, obviously, there's a social responsibility, that we know exactly what use is being made of the data.*
>
> **– John Gallacher**

Both access procedures and technical access models of distributing the data may have an influence on data sharing. The key here will be to strike a balance between acceptable hurdles that prevent misuse of the data as far as possible and enable trust for those depositing the data, and hurdles that still create appropriate efforts and low constraints as to how research is being done. For big data in particular, this also means finding suitable mechanisms for accessing datasets of growing size.

Timing access

A final aspect is timing access to data. In ADNI and UK Biobank, data are made available to other researchers at more or less the same time as those who are part of the data sharing initiative.

The difficulty here is striking a balance between avoiding keeping the data closed for so long it reduces the value it brings to dementia research, and the time that is needed for reasonably cleaning the data:

> *For me the limit of six months or one year, is too short. If I had to put a limit in that sense, a sort of embargo for my group, it should ideally be ten years because we have so much data to understand. So what I want to be sure of is that when I provide the data, the data needs to be cleaned. [...] And we are the only ones who can clean it in a nice way.*
>
> – **Laura Fratiglioni**

At the same time, timing is also very dependent on the question of whether the reseachers involved in the collection should have preferential access to the data for use in publications as compensation for the effort and resources that they put into collecting the data. AddNeuroMed, for example, was not designed to be open from the beginning, and has only opened up the data since the original collaborators have published from it:

> *During the active phase of the study we had access so that the consortium was using first the data but then there have been a lot of agreements and collaborations with groups that have not been involved basically in collecting the data. [...] That of course was to ensure that when we had the results, the people who have been working for that get the credit [by being] involved in the publications.*
>
> – **Hilkka Soininen**

As shown throughout this section, the way data are accessed – in terms of procedures, models and timing – has implications for a range of other questions as regards getting credit for data collection, ownership of the data, redistribution of datasets and quality bars for releasing data, all of which we will revisit throughout the remainder of this chapter. The next section will focus on how data can be combined with data from other sources.

3.3 Interoperability

Finding: Interoperability in the sense of common standards is recognised as important, but still presents challenges in actual practice. Linkage to routinely collected data in the health system is done for several population-based studies, but we did not find examples of non-medical big data being integrated.

Agreeing on standards and protocols means that data can be pooled and compared more easily, and consistent information gathering enables effective linkage and secondary data analysis (OECD, 2014). We consider two aspects of interoperability: Data gathering according to common standards, and the ability to link data to other datasets in the health system.

Standards for data gathering

ADNI has the ambition to develop a standard for imaging that would be followed internationally, and has been recognised by many interviewees as having attained this goal, which is an achievement for the scientific research community.

For example, AddNeuroMed has adopted imaging standards for their MRI data, so that data from ADNI and AddNeuroMed can be combined. This has been evidenced in several publications and in the usage of ADNI as the main dataset for the recent Synapse DREAM challenge, in which AddNeuroMed was used as the test dataset.

Within the different international spinoffs of ADNI, pooling nevertheless has its challenges, and is one of the main aims of the worldwide project WW-ADNI. While sharing data may create larger datasets that enable greater statistical power and allow for researching differences between countries, thus demonstrating the value in international standardisation, this has not necessarily always been achieved:

> *To start the ADNI-like initiative in my country, we put some knowledge that we had looking at the US-ADNI website, so we downloaded all the protocols for the scanners and we implemented them locally [...]. One of the technical points to further discuss and improve to make giant steps forward for dementia research is the possibility to make cross-references about the variables of all the different datasets such as US-ADNI, I-ADNI, EU-ADNI, Argentina ADNI, J-ADNI and AIBL. It's quite difficult to harmonise all these, but at the mid of 2018, it should be possible to do these kind of aggregated queries and a sort of data federation as well as data sharing. That is something that is just started thanks to the Alzheimer's Association.*
>
> – **ADNI data manager from country initiative**

The project WW-ADNI now looks at harmonising these structures, so that the datasets can potentially be pooled, exploiting the fact that they were collected with similar protocols in different parts of the world. For some cognitive tests where cultural differences play a role, however, this may be harder to do:

> *Cognitive research is difficult with cultural and linguistic differences. [...] Neuroimaging is easier in a way to be in a common protocol, image is image and we have similar physiology. But when you get into the cognitive tests, that is more subjective and there are nuances. How do you normalise the general population differences in areas like calculation or memory tests?*
>
> – **Yoshiaki Tojo**

It is worth highlighting that AddNeuroMed, which collected data from six different European countries, faced similar problems for parts of the data:

> *For AddNeuroMed, we had some problems in comparing some neuropsychological tests between clinically defined groups since different tests were used in patients and controls. For the future I hope a more standardised approach will be used in different studies and countries to facilitate the comparison.*
>
> – **Patrizia Mecocci**

Beyond internal standardisation within studies, some standards may also help to improve comparability of data and pool data. In general, the lack of standardisation may be overcome, and has been seen as a surmountable hurdle by most of our interviewees – however, it would be desirable to have at least some level of harmonisation beyond the imaging realm:

> *I think it's important that we try to harmonise the data collection in a good way all over so that it's possible to combine all the population-based cores with the research cores in a good way. And I don't only mean the harmonisation of MRI, which of course is very important, but also the clinical evaluation and so on is very important [...] So there are many things to work on standardisation; not only MRI, but everything.*
>
> – **Eric Westman**

In many cases backward standardisation is achievable, but is also very complex, requires lots of resources and makes studies prone to errors. The Dementia Platform UK which uses 22 different cohorts, the UK Biobank cohort the largest among them, has combined data even though considerable effort was needed to pool the cohorts:

> *Yes, [matching the definitions was] extraordinarily difficult, and it's a work in progress which is another reason why it's challenging to get the quick wins. [...] It is often very difficult to harmonise what is apparently the same*

> *variable measurement between studies and we just have to put time and money into it.*
>
> – **John Gallacher**

Linkage to other datasets

UK Biobank is a good example of linkage to other routine datasets, as individuals have agreed to data from their EMR being used for follow-ups, as also have other population-based cohort studies, such as the Canadian Longitudinal Study on Ageing (CLSA). This allows for tracking of conditions developed by individuals at later stages in their lives without the need for active follow up:

> *Our participants consented to us re-contacting them to ask them new questions or invite them to take part in new sub studies, but also consented to us following their health through linkages to their health records. A lot of the follow-up that we do is from the participant's point of view passive and because of the comprehensive coverage of national health records [...] we can link readily to their health records both in primary and secondary care, which provides coded diagnostic information about them.*
>
> – **Cathie Sudlow**

Of course, one has to keep in mind that linking data may create further process hurdles, as other routinely collected data may be held by other players in the health system, but none of our interviewees had already gone through the full process of doing this linkage. Beyond this linkage actively being provided by UK Biobank, further linkage by the researcher is actively discouraged by UK Biobank, with each data user receiving a dataset with different unique identifiers to prevent further linkage – an aspect that relates to privacy and security as an important aspect of data governance.

An extraordinary example of linkage within the health system is Sweden, not only due to its many longitudinal studies dating back to the middle of the last century, but also due to the possibility to link in condition-specific registries and routinely collected data from Swedish primary and secondary care settings:

> *In the Nordic countries, we have the unique personal number and everything is based on that, like all the hospital registers, drug registers, medical records, everything. [...] So it has been possible to use it and link various sources of data. In that way we have been able to create very big and comprehensive longitudinal databases. There has also been the tradition to share the data.*
>
> – **Miia Kivipelto**

In contrast, ADNI and AddNeuroMed data are not linked to other forms of routinely collected information in the health system. This linkage was not intended from the outset, and hence has neither been considered in the consent, nor is it easily possible, as the individual has already been deidentified at the data collection site.

None of our case studies, and to our knowledge no other large initiative already incorporates other forms of big data in the sense of routinely collected data outside of the health system, as introduced in section 3.1. While technically the UK Biobank and the Swedish longitudinal studies would be able to do so, there are no concrete plans for this. We will return to this point in our analysis of the deep-seated dynamics of sharing data and exploring new sources in the next chapter.

3.4 Quality

Finding: Quality of data and approaches to data quality differ. While big data analytic approaches accept data quality as a challenge, at least when collecting data purposively, quality needs to be a priority from the early design stages. Documentation and metadata are at acceptable levels, but have room for improvement to ensure wider use of the data. Governance models for adding user-enhanced data back to the main dataset reduce duplication of effort.

Beyond data being available, accessible and interoperable, its quality is a very important aspect. While big data analytics are also being designed to mine sources of varying quality, good documentation, metadata and user-enhancement of the data help to develop the resource and ensure its wide use.

Quality of the data itself

Quality assurance may be difficult especially when data are collected in different places, as in the case of ADNI for example. While quality controls are in place and circumstances (equipment, skills) within countries are often somewhat homogeneous, ADNI's country spinoffs may have different levels of quality controls:

> *In general, the quality in US-ADNI is generally very good because every image has been previously checked by the ADNI quality control staff. It's more difficult for the other ADNI-like local initiatives because there is no specific quality assurance and maybe also the quality control procedures are not well standardised.*
>
> – **ADNI data manager from country initiative**

Going back to the question of broad and deep data, for some analyses quality of the data may be as important as its size, and while big data offers exciting prospects in terms of size and breadth, there is also a right to exist for small data. This also depends on the specific question:

> *I would like to highlight again that it's not always a question of how big is the database, it's an issue, but it's not always that. Quality is very important and sometimes it may be better that you have a small amount of data, if it's harmonised and done in the same way. It depends on the question.*
>
> – **Miia Kivipelto**

As mentioned earlier in relation to access timing, there is a trade-off between assuring quality and releasing the data. For example, AddNeuroMed only released part of the dataset for the DREAM challenge, whereas it was kept private before as there was no need to release it – and because it had not been cleaned. And, of course, quality is never going to be perfect:

> *You would probably have to put in a lot of effort to make it completely perfect for anyone to just go in and analyse it. It's also so that you always need to do the data collating, checking. Even if we try to be as careful as possible, every time really you analyse you find something which is not completely perfect, and I think that's the same for most studies.*
>
> – **Ingmar Skoog**

To allow users to assess quality, it is also important to document known data quality challenges, and include measures that actually allow for this kind of documentation, such as source data from the medical devices with which the measures were taken, something that is provided in ADNI and UK Biobank, for example.

Metadata

Keeping metadata, i.e. data about the data, up-to-date is important in order to make data findable and ensure that units of analysis or scales used in the measurements are made explicit. This may be an enabler for pooling datasets, as it may help to indicate how a certain measure in a dataset was taken:

> *Sometimes you think you're getting the same thing but you're actually measuring completely different data, which is the risk if you just go into different datasets. So then you really have to decide how [you] can compare let's say variable A in one study with variable A in another study – because they could look differently even if they tried to measure the same thing.*
>
> – **Ingmar Skoog**

At the same time, metadata of course contributes to findability. Data within catalogues such as GAAIN, neuGRID4U or EMIF, where one may look for datasets with specific features of individuals, or specific measures taken, can be found more easily based on good and accurate metadata:

> *[In EMIF] we've established a fingerprint browser that has 26 meta variables and nested within that are a total of 800 variables which we collect on cohorts. So this might be age. "Do you collect the age of your subjects?" So everybody does. It might be MRI. "Do you have MRI?" If you do is it serial MRI, have you passed it for hippocampal volume?" and so on and so forth. So we don't ask for the actual data, we just ask for the metadata, the headline variables. It's a catalogue of what's in Europe.*
>
> – **Simon Lovestone**

Documentation

In addition to metadata, good documentation is important to actually allow for data use, and is often challenging due to the additional efforts beyond the data collection that do not directly add value to the research outcome. We found that for all our case studies, the documentation was generally acceptable, though with room for enhancement especially in terms of how data fields could be interpreted.

In the case of the Swedish Brain Power studies, documentation exists for most studies. For the Kungsholmen/SNAC-K study, documentation has been translated into English, and its codebook and the questionnaires are available online. For others, however, coming to Sweden to collaborate with the researchers may be the only way to use the data in a knowledgable way. This indicates how a lack of documentation may contribute to scalability issues and the data not being used more widely. For some projects, knowledge may even be in the hands of single individuals, without funds available for improving the documentation. This not only creates hurdles for sharing, but also for using the data at all:

> *You need a lot of documentation behind every variable to be able to use it in the correct way. [...] One of the practical problems is that you have people collecting data and then [...] you have to make sure that the database is in good shape so other people can use it. Some of our databases started maybe thirty years ago and people are starting to leave [...] with a lot of knowledge and we are not funded really to have people working just with supporting the database and keeping it up to date.*
>
> – **Linda Hassing**

Especially in relation to sharing, it is also important to have the documentation accessible to those considering using the data. This will not just help to identify the suitability of the resource, but also allow users to be more precise about which part of the dataset they need and how to use it properly – which was mentioned as an advantage of UK Biobank both by users and providers of the resource:

> *For a resource of the scale of UK Biobank to be useful to researchers they need to be able to see in some way what data was collected on all half a*

> *million people, what data did we collect on some participants, what data have we collected subsequently. So we devoted a lot of the [UK Biobank] website for researchers to be able to answer those questions. [...] It's not helpful if we just have very superficial documentation on the website which is what a lot of older cohorts have.*
>
> – **Tim Sprosen**

One particular challenge relates to making documentation also more accessible for people outside of the field of dementia research, which created hurdles in the Synapse DREAM Challenge, for example:

> *Everything is reasonably well annotated, if you know what you're looking for. But for someone that is outside of the field of Alzheimer's research, which we believe many of the challenge participants are, then it becomes much, much more difficult to find what you need. If you want to encourage unique innovative approaches from people that aren't necessarily in the field, there needs to be an easier way of utilising this data.*
>
> – **Mette Peters**

The originators of this crowdsourced open challenge ended up creating additional documentation in order for their participants to be able to use the data.

Enhancements to the dataset

A final question is how user-generated enhancements to the dataset should be managed in terms of quality. For example, while ADNI does not allow users to upload their own data and reintegrate them into the main database, UK Biobank actively encourages its users to do so – even if the processes may not have been fully developed from the outset:

> *If you work with UK Biobank samples you have to provide the data back to them. Now that's absolutely fine and we don't have any difficulty with that at all. [...] The issue is that they need to be able to receive these data and store it in a manner so it can then be distributed to other investigators. When we started doing the UK BiLEVE project I felt the logistics of data submission had not been fully worked through at UK Biobank. Obviously you need physical infrastructure to store all the data – because we're talking a lot of data, terabytes, so there has to be some sort of central facility to store all the data, and this requires additional resource. [...] I think it's fair to say that arrangements for the process of data submission has moved more slowly than we expected despite the good will of all concerned. I think this is mainly a resourcing issue.*
>
> – **Ian Hall**

These enhancements are managed in addition to the main dataset, and can be searched and requested separately. At the same time, users are actively encouraged to rebuild datasets if there is a suspicion that a data addition has not been done properly:

> *You can't police the way in which the data is used. [...] Trying to do a scientific review of each proposal would, first of all, be an extraordinarily big task to do but secondly you would probably stop people from doing imaginative things that might turn out to be worthwhile, you would tend to be conservative. Probably some people will do things which are not scientifically appropriate and that's again where, in Biobank, what we said is the data go back into the resource so that if people think that someone has done something wrong they can reanalyse it and demonstrate that. This is part of the approach that we have adopted; but we recognise it is an experiment in access.*
>
> – **Sir Rory Collins**

The discussion of enhancements again shows the tension between opening up progress – at the potential expense of some quality – and trying to drive quality to perfection. It is worth highlighting the approach that UK Biobank has taken: ensuring quality in the core resource, while leaving extensions to a certain extent up to self-governing networks in which users would correct other users' errors, which shows some parallels to the way Wikipedia works, for example.

3.5 Traceability

Finding: Traceability of research participants does not represent major problems, and neither does tracing participants to other forms of routinely collected medical data for those studies designed for it from the outset. Linking routine data from outside the medical realm remains a challenge. Finally, ideally, it should be possible to trace what data were collected and how data were analysed.

Traceability may relate both to questions as to the origins of the data, but also the extent to which decisions about the data and definitions are made transparent, and how analyses can be tracked.

Tracing participants across stages

As opposed to big data which may be collected in unstructured ways from a variety of sources, traceability is not seen as a major hurdle for any of the four case studies, in all of which individuals have a joint identifier. While ADNI does not use a single identifier across country initiatives, the chances of one patient having participated in several hospitals may not present a major scientific issue.

Of course, what has been used as unique identifiers across case studies differs, with those who are linking to other data in the health system using national identifiers at their core (even though these are not openly shared with researchers – see the next section).

To link other forms of big data such as loyalty card data, having a single identifier may be much more challenging. However, even for aggregate analysis of this data without going down to the individual level, this is far less a technical issue than a consent issue, with legal safeguards prohibiting linkage:

> *With data tagging you can create data lakes. This data can be analysed through machine learning algorithms. Our ability to link even unstructured datasets has increased a lot, and is therefore relatively cheap. But it does not sit well with current legislation in Europe. So either you go and get consent, which is very hard – the whole point of using routine data for aggregate analyses is not to ask everyone individually – or you would need to find jurisdictions which take a different approach and learn from them.*
>
> **– Nicolaus Henke**

As mentioned earlier, UK Biobank, which uses the individual's NHS number to link data within the health system only provides pseudonymised data with a unique key for a specific dataset, hence creating high bars for tracing individuals elsewhere or in other researchers' versions of UK Biobank data.

Transparency of decisions and changes

Less than the data itself, it may be more challenging to trace the way the data were collected and decisions about how the measures were set up. This may involve changes made to data over the years, or to how elements were constructed:

> *What I was lacking is maybe the rationale for some of the tests, so some of the cognitive tests were so novel I think some of them, so it wasn't clear why certain things were chosen but I guess that's probably recorded somewhere else and I know these things are often decided by committees over long periods of time.*
>
> **– Daniel Smith**

A similar aspect is version control, which may be especially necessary when trying to make data accessible in a timely manner, and potentially correcting errors later:

> *Yes, unfortunately the idea that as soon as the data are collected they are pushed to the top means that they are accessible for the world. Sometimes if I [...] do a request to get all the ADNI account today, and somebody else is doing the same request one month later, it's not similar to that.*
>
> **– Jean-François Mangin**

> *Information is periodically updated and some of the data is preliminary. You are made aware of that through the data use agreement, but we were concerned that if we simply pointed challenge participants to LONI people would end up with different data versions. The data we needed was also spread across multiple files creating additional concerns about data differences. [...] A better versioning system and a simpler way of querying the LONI database would have made that process much simpler.*
>
> – **Mette Peters**

Finally, ensuring replicability is at the core of statistical analyses. While it is to a large extent the researcher's responsibility to keep and document an audit trail, the reading library model, as envisioned by the Dementia platform UK for example, may also allow for tracing the analyses performed on the data.

At the same time, increased traceability and the ability to link data may also pose threats to privacy and security, as will be explored in the following section.

3.6 Privacy/security

Finding: Each of the case studies has taken privacy and security very seriously, thereby minimising any misuse. By and large, there have not been major limitations based on how they were consented for our case studies, though a wide range of open questions exists especially for legacy data. Combining data with big data from outside the medical realm and new forms of crowdsourcing presents major challenges.

Privacy and security lie at the heart of medical research, and should not be neglected as a subordinate goal to advancing dementia research, especially as this may open up a slippery slope of using the data for other purposes which are at variance with the individuals' interest. At the same time, privacy and security considerations may also hamper dementia research if considered to be paramount.

All data sharing initiatives that serve as case studies in this report have taken privacy and security seriously, approaching it with a combination of obtaining consent, de-identification and/or secure environments, and restrictions in the data use agreements for researchers. Except for the Japanese case outlined below, there have not been any major problems in this regard with any of the four cases we examined.

In 2014, the Japanese ADNI faced some challenges in terms of alleged data mishandling, both in relation to inaccurate managing of consent forms by one of the 38 centres in which data were collected, and due to potentially endangering the scientific validity of the results. A third-party review made recommendations to annotate some of the variables, in combination with filling the gaps in the consent obtained, but cleared the data for use in research:

> *For the alleged data mishandling, it is clear by the review of the commission that they found some human errors, but that it is not jeopardising the value of the data to be disseminated. So the data should be available for the international research community in a couple of months.*
>
> – **Yoshiaki Tojo**

The case of J-ADNI shows that proper data management of research data is crucial both for maintaining the trust of research participants, but also for ensuring that the data can be regarded as scientifically valid.

Beyond data management and proper management of consent procedures, two other important issues are detailed below: where data are de-identified, and the models that have been adopted to ensure that consent does not unnecessarily constrain research.

De-identification approaches

Datasets may never be fully de-identified, as through combination of attributes, a chance of identifying an individual always persists. However, all of our case studies have taken a careful approach to ensuring research participant protection as far as possible. For example, ADNI and AddNeuroMed have decided to de-identify patients at the point of data collection:

> *We don't have the key to patient or subject identity and all data are anonymised, so one of the ways we protect sensitive data is that the database itself does not have in it the key to identify anyone, that's never uploaded into the end database used for data distribution. That key is kept at the acquisition site, so we couldn't unlock it if we wanted to. We do not have patient identifying information.*
>
> – **Arthur Toga**

In addition to technical measures, a governance aspect is that scientists agree not to try to identify participants, or contact them in case an identity was revealed:

> *With this robust model of data sharing, we have certainly been concerned about privacy and about potential misuse of the data, but thus far, we are not aware of any such examples. Part of the data use agreement is that data users state that they will not seek to identify subjects. In the early days of ADNI, we were concerned that people could reproduce cases from the imaging or DNA data, but to my knowledge we have had no abuses of this.*
>
> – **Robert Green**

In contrast, the identity of UK Biobank participants are centrally known, which, as outlined earlier, is also one of the prerequisites in order to be able to link data to other routine data in the health system, but will not be shared with data users.

Consent models

All case studies obtained consent from participants for several aspects of collecting, linking and using data. All of our case studies reported that there were generally no major hurdles in what could be done with the data in terms of data analysis by researchers at universities based on the consent obtained.

Constraints due to the way consent was obtained may be a particular issue with condition-specific data that was not designed to be open from the outset. However, for AddNeuroMed, the only case study fulfilling this condition, the consent was interpreted to be broad enough to allow its distribution to other researchers.

In the case of Sweden, consent may be less clear in some cases, so that the way datasets are shared through collaborations is also partially motivated by being careful to comply with the consent that was obtained:

> *Then of course it comes to question, what has the patient consented to. [...] And what we have done now is more to let people come here and work on it, because then like we have at least a bit better coverage and informed consent. But I mean, ten years ago they didn't think about the enormous possibilities that having CSF from 10,000 people – what that means.*
>
> – **Bengt Winblad**

The Swedish case exemplifies well the problem of historic consent. Consent may be interpreted within the context of when it was written, but may also take into account progress and a changing social and technological context. One area of conflict relates to specificity: For example, if individuals in the past consented to their blood being shared, would this allow sharing of genetic data derived from blood samples? Especially due to the aforementioned rapid decline in sequencing speed and cost, could the individual who consented to their blood being used have understood or predicted the consequences related to sharing genomic data?

In addition, different national legal traditions complicate the interpretation of consent in a globalised research sphere, with differences in whether something must either be specifically allowed or just not specifically forbidden to be accepted. For example, as regards consent that includes "sharing data with investigators from this institution" – does the fact that there is no additional "but not with researchers from elsewhere" prohibit sharing? Especially for consent obtained before the widespread availability of technology, there may a wide range of answers to this question. In Sweden, researchers are therefore made adjunct to the group:

> *So far what we have done is to have very loose collaborations. [...] The people who work with us could be said to be adjunct to the original research group, which means that we do have consent, but we could say that we have no consent that people [from] outside come in and just use the data.*
>
> – **Ingmar Skoog**

Both ADNI, AddNeuroMed and UK Biobank have obtained relatively broad consent, which is important as beyond the highly problematic nature of reconsenting data, when it comes to cohorts in advanced stages of dementia, individuals may either lack the mental capacity to reconsent for themselves, or may have died.

Consent can be revoked by individuals, though these cases are rare in practice. For example, for UK Biobank, individual data may be removed for all analyses from that point onwards, but individuals often simply want to reduce the burden of further data collection rather than withdraw their consent completely:

> *Withdrawing from the studies completely, so participants contacting us and saying they no longer want to be part of the study [...] those [cases] are extremely rare and well under 1%. A large proportion of participants who do decide to withdraw actually only withdraw from being re-contacted but they are quite happy for us to continue to follow their health through their records, as long as it doesn't involve us contacting them at all.*
>
> – **Cathie Sudlow**

UK Biobank is a very good example of how having individuals consent to follow-ups via routine data reduces drop-out and participant burden. However, with broad consent given, it is important to ensure that data are used for the public good and that consent is properly interpreted based on an evolving social context. This requires other governance structures to be built up in order to balance the interests of all involved stakeholders in the light of changing circumstances:

> *To address these challenges, the consent given by participants for use of their data in UK Biobank is broadly given. Because of that, UK Biobank has instituted an unusually rigorous kind of process for an ongoing ethical balancing of benefits and risks of data use. This involves not only an internal governance committee, but also a standing external ethics and governance panel, which regularly reviews the approach to consent in the context of the type and use of data and the evolving social context. This is a kind of structure that, in my experience, doesn't exist in other studies.*
>
> – **Paul Matthews**

When it comes to big data that extends beyond the medical realm and that is stored for longer periods of time, this ethical oversight is important for participants and researchers. Just two examples are the blurring boundaries of what it means to provide consent for data to be used for "any health-related research", and to what extent an individual can in fact give consent about genetic data:

> *How far can you take the concept of health-related data? What kind of data would this in the end cover? [For some data] it would be quite an intelligent interpretation of health-related data. Another take on this [...] is if Biobank is proposing to do something which it thinks is within the spirit*

> *of the governance framework or the spirit of the consent [...] then arguably Biobank might be justified in proceeding with this, communicating this to its participants [who] would then have the opportunity to opt out.*
>
> – **Roger Brownsword**

> *To what extent does prior agreement and the consent form hold after death, and to what extent does it not? A very salient question in the era of genomics is what if subjects die, and genomics was never mentioned in the earlier consent form, and these genomic data have implications for other family members.*
>
> – **Robert Green**

Completely new challenges have also arisen as new ways of distributing and sharing data, through data catalogues and data aggregators, for example, have appeared in the past decade. For some of the cohorts included in the Dementia Platform UK, consent was inconclusive with respect to whether they could be included in the platform, as data aggregators were neither covered nor excluded by the consent obtained decades ago. Surveys among participants of other genetic data sharing initiatives showed that participants valued the opportunity to decide whether their data should be included in aggregated databases (Ludman et al., 2010). A related issue is the re-contacting/re-consenting of individuals for trials, which however extends beyond the scope of this report.

Similarly, new barriers exist for new ways of extending the realm of researchers using the data beyond established organisations. For example, when the Synapse DREAM Challenge was set up in order to crowdsource biomarker identification, distributing the ADNI data to the challenge participants presented unexpected difficulties, to a large extent based on how the data were consented:

> *Due to different data use agreements based on differences in data consent, we had to go through a lengthy process of MOUs to get approval for how we wanted to distribute the data to the challenge participants. We did not have approval to share the ADNI data through Synapse due to their re-distribution clause based on how the data was consented. [...] Challenge participants ended up having to access two of the datasets through Synapse where each had a different type of data use agreement and individually had to go through the process of getting approval by ADNI in order to get access to the challenge data in LONI.*
>
> – **Mette Peters**

Similarly, consent may not allow linking in other forms of routinely collected information. While one may argue that this is the very purpose of consent forms – to prevent, for example, linking of credit card data to medical data – this at the same time also reduces the possibilities that might be gained from big data:

> *UK Biobank has asked participants about employment and will capture data from hospital and GP visits, but does not intend [...] to link in any way to financial records. That would be something out of the scope of the data collection. Any intent to re-identify people for information about personal financial or personal circumstances also would be fundamentally in violation of the volunteer consent and the investigator agreements for Biobank.*
>
> – **Paul Matthews**

In summary, while consent has not limited the main activities done with the data in our four case studies in any major way, there are a number of questions about how to balance consent between patients wanting their data to be used for medical research and unwanted restrictions to research that may be created by consent, but also ensuring appropriate protection for patients data. Additionally, new developments in global sharing do create new questions in how previous consent should be interpreted, and how consent should be obtained to minimise similar barriers in the future.

3.7 Ownership

Finding: Datasets created by data sharing initiatives may often not be shared further, a condition that makes data aggregation difficult, but that retains a certain level of control over who has access. It is good practice to have the data sharing initiative acknowledged on any publications (not as co-author), without the initiatives being involved in quality control of the paper.

With data being easily distributable and copyable as opposed to physical samples, new questions of ownership arise, affecting the questions to what extent data can be redistributed, and how the "data owners" should be properly acknowledged.

Redistribution of datasets

In general, ownership of data has many facets. Ownership in the context of data created for dementia research is generally seen to be within the initiative that created it, with those who helped build it being the owners of that particular dataset – which may or may not come with preferential access.

As outlined in the previous section in relation to the Synapse DREAM Challenge, redistribution of datasets may be difficult based on consent. However, beyond that, it may also not necessarily be in the interest of the data sharing initiative to have data redistributed, as for example for ADNI:

> *Anybody can have the data, but we prohibit re-distribution and require acknowledgement about the data source in any publication. The reason we do that is to protect the integrity of the data, track data use and provide data*

> *value metrics to the funders. Data integrity can be lost if the data is taken from ADNI and modified by someone else and then given to a third party. There have been a few occasions where people have not read the data use agreement or disagreed with it, and violated those rules. We try to monitor ADNI data use and when such a violation is discovered we call them on it and say you signed this agreement, those are the rules, please, please stop. And so far they always have.*
>
> – **Arthur Toga**

Similar approaches are taken by the other initiatives, though for the specific case of the Synapse DREAM Challenge, these were relaxed:

> *No, it's not an issue for us. [AddNeuroMed] gave out the data fully anonymised to be used for purpose. So once we're confident [...] that what Synapse is doing will meet those requirements, [there are] no restrictions on it other than a publishing restriction, which is that we just want proper acknowledgement for the funders principally.*
>
> – **Simon Lovestone**

In addition, UK Biobank uses a specific identifier for the released datasets which – beyond making linkage to other datasets impossible – also give the released dataset a unique fingerprint. Other technical approaches again relate to the access model, with the reading library creating an additional barrier to redistribution.

Acknowledgement of creators/owners

Proper acknowledgement is another issue pertaining to ownership. ADNI, AddNeuroMed and UK Biobank all have established rules with which the initiatives are acknowledged in papers, mainly in the acknowledgement section and keywords.

However, the issue of giving personal credit remains (Gardner et al., 2003). For sharing, there would need to be some credit assigned to individuals and their institutions:

> *The world is going more and more to this open direction and possibly we need to keep in mind that collecting data in these kind of databases costs a lot of money and trouble. Of course, the researchers also want to have some credit for that, so that we are not doing this kind of thing for fun.*
>
> – **Hilkka Soininen**

While in the case of the Swedish Brain Power studies data owners become co-authors, as they actually contributed to the publication in collaborations, there was widespread agreement among the interviewees that co-authorship for "just" providing the data would not be appropriate:

> *I have always felt, personally, that you can give data to people and let them do with it what they want, or you can share data with them and collaborate with them and be part of the paper. You do one or the other, rather than just give the data and have your name appended. Many people feel uncomfortable about that. They want your name on it, they actually get quite negative if you say, look I don't want to be listed on it because I am not actually an author, I am not involved in it, I haven't read the paper, [but] I am happy to give you the data.*
>
> – **Sir Rory Collins**

Another aspect in this regard is that quality control of resulting articles may almost be impossible, or at least hardly scalable. Those initiatives which do not work based on the collaboration model (ADNI, AddNeuroMed and UK Biobank) have taken the conscious decision to leave quality control to established peer-review mechanisms and just formally ensure that the initiative is properly acknowledged:

> *We don't check the publications in great detail because we haven't conducted the science so it is not really our place to do that; that is more something for the peer review and specialist scientific community to make a judgment on. We make sure they have acknowledged the use of UK Biobank and quite often researchers include UK Biobank in the title or it tends to be in the abstract. We ask them also to include that acknowledgement in the body of their abstract as well to make it easier for us to search for the publications that are coming out for using the resource.*
>
> – **Cathie Sudlow**

The distinction between providing the data and doing the research is also important in case conflicting research is published based on the same data, without being obvious what is right and wrong:

> *To what extent are you going to allow more than one investigator, or more than one team, to ask the same question? And, by extension, do you care if data from your study are used to support conflicting or even contradictory messages in the marketplace of ideas. I have talked to leaders of other studies who have said, "I would never do this because we don't want to be in a position that one person publishes a paper that says A, and another person publishes a paper that says A is wrong." In ADNI, we have quite consciously chosen to let scientific ideas, even conflicting scientific ideas, emerge and compete.*
>
> – **Robert Green**

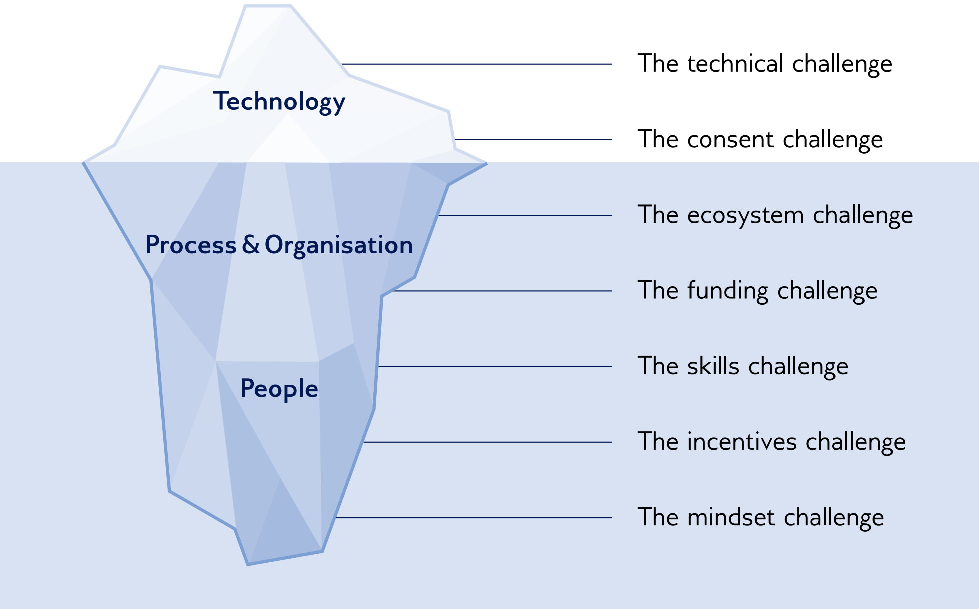
Again, this highlights an important aspect of scalability: Checking the quality of each publication coming from the data would require significant resources. In contrast, leaving quality control to established mechanisms enables more publications to be published without bottlenecks created by the data provider.

4

Structural challenges and recommendations

Based on our analysis of the four case studies' data governance, we have identified more deep-seated root issues that hold back using and sharing data for dementia research. The technology-related challenges are often talked about and can be overcome. Below the surface are issues around how data collection, analysis and sharing are managed, as well as underlying people-related challenges. None of these are easy to solve, and none alone will be sufficient to advance dementia research, but all of them can be tackled in a manner that acknowledges their systematic nature.

Figure 2: Structural challenges to data sharing



4.1 The technical challenge

Recommendation: To address the technical challenge, one may set up flexible standards for a minimum dataset used in dementia research, combined with employing better analytical approaches to help find specific datasets without the need for rigid ontologies. Documentation in wikis may help to share the burden of documentation and enhance it. A combination of the reading and lending library access models may encourage more data sharing going forward.

While there was widespread agreement among the interviewees that the technical challenges may be overcome, it is worth spelling some of these out and delineating potential ways of approaching them. Addressing these issues may not only save time and effort of researchers, but also enable further data sharing.

Create flexible standards for measures

Creating standards is always a trade-off: While innovation should not be stifled and it may neither be easy nor desirable to just establish "one way of doing things", some structure may be helpful to foster data comparability and enable pooling:

> *My preference is to have a number of standard procedures made available that individual investigators can pick and choose. So if they are going to measure blood pressure, these are the sorts of guidelines that we would suggest. If they are going to measure anxiety these are the different sorts of questionnaires that we would suggest. And it is a matter of suggestion because who knows where science takes you? We really want the clever sort of left-field ideas and the last thing you want to do is constrain those. [...] A library of standard procedures which people can dip into is helpful.*
>
> – **John Gallacher**

Creating these flexible standards helps to give different ways of measuring items a proper name, making differences evident and allowing conversion of one to another if required. Of course, it would be a mammoth task to synchronise every variable of potential importance to dementia – especially with its increasingly fluid boundaries – and also it does not reflect the big data mindset, where data are increasingly analysed without being able to determine all of its features from the outset. Yet determining a set of common options could be done at least for the most important dementia-related variables and general demographic and lifestyle-related variables, in order to establish a minimum set of common, global standards:

> *There are ways of doing it, but maybe at least to have some minimal sets of variables which could be combined would be good to have. There are rules for it, but [...] there are several guidelines, probably different in each country and different depending on what you do.*
>
> – **Eric Westman**

Ideally, these standards should go beyond the borders of dementia research due to the complex "whole body" nature of dementia as an outcome of a variety of factors. This shows the importance of existing institutions like the Clinical Data Interchange Standards Consortium (CDISC), an organisation that is tackling the issue of a lack of standardisation in medical research, or more specific ongoing ititiatives such as the International Database on Aging and Dementia (IDAD) and the European Dementia Prevention Initiative (EDPI), which are useful as an enabler of more international collaboration in dementia research (Solomon et al., 2014).

With this, we do not propose the creation of rigid ontologies that stop individuals from advancing measures in medical research, but rather making an effort towards standardisation from the outset where possible. Where available, existing standards may be used, such as the standards ADNI created in the neuroimaging area. Large studies using certain ways of measuring items are often creating an implicit pull for others to use the same measures:

> *You don't want researchers to adhere to a standard for its own sake. There must be some sort of incentive or reason for adhering to them. For example, if you have major studies adhering to guidelines and you want your data to be comparable then you will adhere to them also.*
>
> – **John Gallacher**

For the space of clinical data and other routinely collected data in the health system, some standardisation may also be helpful, especially in cross-country research. For example, with many current and new initiatives being funded at an EU level, it is important to make sure that the data are also truly "European", instead of researchers sticking to data within their countries. Clinical data tends to have grown historically within countries, with little value in making the data more compatible between them:

> *At the moment it is not easy to use data from most of the European community funded studies. I work mostly on data that are produced in Italian networks. It is much easier to work with Italian colleagues and to share data, because when you put together the clinical data from several datasets, they are usually similar. [...] Also when you share your database with other databases, you find that not all the clinical evaluations are comparable, and you waste a lot of time in trying to figure out how to perform comparisons between one dataset and another.*
>
> – **Patrizia Mecocci**

However, with some data structures undergoing changes via other initiatives such as the Cross-border Care Initiative, a multitude of motivations for increasing comparability of clinical data is coming into existence – and is partially already being put into practice, e.g. with consortia like the International Consortium for Health

Outcomes Measurement (ICHOM) developing global ways of measuring outcomes for different diseases. As with data collected in research studies, a minimum dataset may be very beneficial:

> *We should put in a minimum dataset in all European countries. And then, similar to what we do within Sweden, I'm sure in one year without any money we would have much more data than we would have ever. But it will be an easy way, I think, to come up with a stronger – not only in Sweden but a European collaboration.*
>
> – **Bengt Winblad**

> *The ICHOM working groups typically define no more than ten outcome metrics by disease to establish a minimal sufficient set of metrics as a global standard. In addition the teams identify a set of risk adjustment factors to enable correct comparison of quality of care. A given institution will often for different reasons want to add their own specific metrics, but the ICHOM set would be the base and common language enabling benchmarking. Only a year after the first four sets were launched we're seeing a very large interest in using them. The ICHOM team is working closely with some of the best centres in the world across all continents. We already see that these prominent centers are followed by others, as they want to compare themselves to the best.*
>
> – **Stefan Larsson**

It should be highlighted that this is something that may be especially important for dementia research in the future. At the same time, where possible, trying to match standards for existing data is useful to make sure that these data are not lost for research, with the caveat that this may not be possible for every single existing dataset due to the effort involved in doing so.

Document data and make it findable

Common standards may also help to make data more findable, as for catalogues like GAAIN, neuGRID4U or EMIF, which work based on the metadata in terms of overall characteristics of the study and in relation to what measures were used. Additionally, ways of accessing the data may also be linked more closely to previous journal publications (Howe et al., 2008), thereby not only increasing transparency of the research process but also simplifying finding the data for further analysis.

However, especially when talking about big data as a research paradigm, developing predefined schemes for finding data is not currently the state-of-the-art (Bollier and Firestone, 2010). Another approach may be to deploy machine-learning algorithms to index data and make it findable, similar to approaches used to query large amounts of unstructured material in search engine technology.

In a similar manner, it may be useful to consider new approaches to documentation as well, as documentation is a key to making the data usable once other researchers from within and outside the field have discovered it. As mentioned earlier, as part of the Synapse DREAM Challenge, an additional set of ADNI documentation was created which may also be useful for other researchers using the data. While documentation is often an activity that only indirectly adds value to the research process, data sharing initiatives might consider rather using wiki-style documentation, which has proven to produce high-quality work through continuous contributions from individuals (Su et al., 2013).

> *There's no e-learning section where a user can grasp and download useful instructions to correctly use the platform. I think there should be the need to provide additional and well structured materials in order to teach the new users how to use these advanced platforms collecting millions of data. Moreover - it should be thought to add ad-hoc video tutorials, Wiki pages, and Specific Support Centres. I know that on the US-ADNI website there is a frequently asked questions section (FAQ) but I think some additional improvements should be done to really empower the future research on Alzheimer's and Dementia.*
>
> – **ADNI data manager from country initiative**

With this, documentations that have been created by users – as for example as part of the Synapse DREAM Challenge – may also be fed back and thereby made available to other users.

Wiki structures may also help to document ongoing directions of work with the dataset, as has been done in other genomic projects (Howe et al., 2008). While this is partly already being done, with for example ADNI or UK Biobank documenting the published work on their website, it may help to bring together individuals working on similar issues – or even researchers from different fields of study working with data – and at the same time avoid duplication of efforts:

> *Every month there are more and more projects being listed and approved and some of them are quite similar. [...] I know Biobank are doing this now, they are putting people in touch with each other who seem to be requesting to do similar projects, but I think across the depression dementia field there's potential probably for more interaction and collaboration.*
>
> – **Daniel Smith**

> *It could be nice to have a website, where all researchers who are working on this big data can put which kind of data are available, which kind of studies are going to be performed, and what is already published. Because sometimes you are going to apply for a study, and after a while you discover that these data were already used for a similar study. And it is really disappoint-*

> *ing for a researcher, who already worked on the dataset for a specific paper, to discover too late that you have wasted time for processing data uselessly.*
>
> – **Patrizia Mecocci**

Find new models of sharing

Another question is to what extent new forms of data sharing enable more sharing and use. As the Nuffield Council on Bioethics (2015, p. 60) report states, "as neither anonymisation nor compliance with consent offer sufficient privacy protections in data initiatives, additional controls on the use of data – on who is permitted to access them, for what purposes, and how they must conduct themselves – are therefore required." Beyond organisational aspects covered later in this report, this also relates to finding new models of accessing and sharing data.

Earlier we described the models of the lending and reading libraries, which relate to whether data are physically transferred or accessed via remote access. The reading library model may be less flexible in terms of what can be done with the data by the researcher, most notably for linking the data to other sources. However, a secure environment for sharing data may also help to spur further data exchange, for example for making pharmaceutical companies release data:

> *Rather than expecting pharmaceutical companies to put it entirely in the public domain, you put it into a trusted third party and that trusted third party would provide a safe place for the data to be analysed. [...] You get rid of a lot of the risk by making sure that the data was shared within a trusted third party environment where the people who then were looking at it would have constraints on the way they behaved consistent with the regulatory environment of drug development.*
>
> – **Derek Hill**

Likewise, openness may also affect patients' willingness to participate in research. More than the exact details of technical sharing it may be most important that there is some protection, and data are not just floating freely online. The participation effect should also be considered:

> *We have extremely confidential data, we ask people if they had attempted suicide, we ask people about their sex life, and if we would put the data open that would be a danger. They also come into hospital and get examined, which makes people [think] they have the same protection you have in a hospital. So [with] completely open datasets where people just go into the Internet, I'm a little bit afraid that the response rate would be very low.*
>
> – **Ingmar Skoog**

Additionally, these two models may not necessarily only work in isolation. For example, while ADNI generally has the policy to release data in the lending library model, the Synapse DREAM Challenge had major problems in redistributing the data to their participants. An additional reading library model could consist of ADNI data being allowed to be redistributed in a secure environment while still keeping up the current model of data distribution for all other requests to ADNI:

> *Providing access to data and tools through central cloud computing resources [...] would make data access much simpler and may in fact provide better data protection if people do not have to download data independently to their own system.*
>
> – **Mette Peters**

It may be that in the long run there need to be hybrid models, in which anonymisable data can be distributed and shared with trusted researchers using relatively quick and easy mechanisms (along the lines of ADNI), but other more sensitive data that can be linked to individuals is only available via secure environments.

The real challenge for hybrid models of data access is that the two do not stand in isolation to each other, but allow researchers to move back and forth between the two as their research needs change and evolve. For instance, a researcher who finds patterns in anonymised data that merit further examination but require access to more sensitive data, would be able to apply for access that would allow an extension of the original study, rather than starting from scratch. Likewise, a researcher working in a secure environment may be able to run analyses on granular data that can be transformed based on these analyses into aggregated data that they can take away with them for further analysis.

The general idea of thinking of new ways to access and distribute data goes hand in hand with the OECD Privacy Guidelines (2013) and are at the core of the OECD draft Recommendation on Digital Security Risk Management for Economic and Social Prosperity (“Security Risk Recommendation”), to be finalised in 2015.

4.2 The consent challenge

Recommendation: Standardising consent is a crucial enabler for global collaboration. The consent challenge may further be alleviated by obtaining open consent combined with oversight, ideally also for linkage to routine data to reduce the follow-up burden on participants. One key for this successful model is making research value and risks understandable to participants, managing public trust and allowing individuals to be involved and contribute.

The key issue for consent is that research on dementia should not be unnecessarily impeded now or in the future, while at the same time respecting and protecting individual rights to privacy.

Standardise consent and IRB/ERC approvals

Not having standardised consent forms may create restrictions when data are pooled, with different parts of the data being consented differently. Standard forms of consent, ideally with standard interpretations of what is meant by definitions to date, would streamline approvals without less protection for individuals. At the same time, standardised consent may further contribute to ensuring that all important aspects are covered in a consistent and understandable way for individuals. Standardised consent also concerns related areas with respect to what the procedures are in case of breaches, what the penalities are, and who enforces consent if required.

The example of the North American ADNI, which obtained approvals for all 58 data collection sites, shows the importance for reducing local variation if there is one dataset to be shared. The example also illustrates the need for clearer language:

> *The language that is considered appropriate for consent is constantly changing, and in the US, it is frequently based on local IRB judgment, which is clearly variable. An example might be if one consent process says, "I consent to allow all of my data to be shared with investigators," is that sufficient? Or should the term "all of my data" be broken down into "my imaging data", "my cognitive data", "my genetic data"? If you do break it down into genetic data, is that sufficient? Or should you say "my genotype markers", "my sequencing data"? And positions on this could be different in different IRBs, and they have evolved through the years, generally evolving in the direction of greater specificity. [For ADNI,] we tried very hard to work with the local IRBs and to maintain a standard language about the sharing across the sites.*
>
> – **Robert Green**

For data sharing on a global scale, this also involves international standardisation of consent. This may be an area where academic research may learn from the private sector, with these challenges being commonplace in pharmaceutical trials:

> *A very good model for sharing data from academic-led studies across national borders is the way that global clinical trials are done, whereby all the data can be transferred and used for the statistical analysis plan. [...] It has to be in the context of an international agreement and I don't know how far off we are doing it. I suspect that it might be as much as anything about sharing the practice used in industry sponsored trials with academic trials and that might help. [...] There's often less awareness of the challenges of sharing data internationally within the academic community than there is in the pharma community.*
>
> – **Derek Hill**

However, model consent forms are part of the solution only – IRB/ERC processes will also have to be streamlined. While ethical oversight of research is an important responsibility, finding ways of increasing the efficiency of IRB approvals may be valuable, for example by reducing the need for approvals by every IRB/ERC involved:

> *The issue is that historically if one centre will be providing 50 samples of this and another centre will be providing 100 samples with some other condition then the mechanism to do that would previously have been that each of those sites would have to obtain separate local ethical and R& D approvals [...] this is potentially hugely time consuming. So I think the idea that it can be managed at the national level is quite encouraging.*
>
> – **Ian Hall**

> *And I can also see the differences when we initiated, for example, clinical trials in different countries, how different the ethical committees are working, and I think it's very counterproductive. [...] I think you have to start on national levels. What we have done now in Swedish Brain Power is that if we do a project involving three or four universities, then it's enough if one ethical committee in one university takes the decision and informs the other. And we do the same in clinical trials.*
>
> – **Bengt Winblad**

With respect to some of the difficulties involved in large-scale research involving data sharing among different actors, IRBs/ERCs may also have to be educated to ensure that they act in concert. While there are controversial topics in research ethics for which answers have to be found at a more overarching, international level, there is little reason for divergence of local IRBs/ERCs:

> *Concerning ethics boards, it seems to be a bit who is sitting there, what is the level of knowledge that they have, or what they're understanding for that type of research. There are also many new laws and regulations in the medical field. It's surely important that all the people on the ethics board are informed about the complexity of the research that is out there so that there are no unnecessary delays or problems.*
>
> – **Miia Kivipelto**

For doing research with the data, it is yet to be discussed how specific IRB/ERC approvals have to be. On the one side, it is important to ensure that data are not being put to improper uses justified by the beneficial purpose of dementia research. On the other side, rigidity in terms of specifying precisely which analyses are going to be done may restrict scientific discovery and add unnecessary burden to the researcher community:

> *We got ethical report maybe ten years ago to do a very open approval. The ethical approvals are changing now. They are going more into more detail:*

> *they want to know exact analysis, which questions are you going to relate to which questions, and I think there is something happening now that is not good for science. You never know which type of analysis you are going to do in detail. You can't know that before.*
>
> – **Linda Hassing**

It is a balancing act to streamline the IRB/ERC process without taking away competencies and ensuring that IRBs/ERCs still fulfil the important purpose for which they were created. At the same time, the reduced overhead work may further contribute to better use of data, and also free up further time that researchers can spend on doing research.

Choose open models of consent with oversight

The broad consent approach that our case studies have taken – coupled with oversight about how the data are being used taking into account the evolving social and technical context – may be a good way to strike the balance between enabling research and individual protection.

As outlined earlier, the combination of the two is important as the context changes over time, so that what consent meant when it was obtained and what is reasonable to infer from that today may not be clear. Consequently, "where a person providing data about themselves cannot foresee or comprehend the possible consequences when data are to be available for linkage or re-use, consent at the time of data collection cannot, on its own, be relied upon to protect their interests" (Nuffield Council on Bioethics, 2015, p. 75). This is also the case due to new forms of data (which may just come from data that could not previously be extracted from biological or clinical data), new models of sharing data, or additional routine data collection. To improve accountability, keep interested individuals informed and improve the dialogue across data sharing initiatives, decisions taken by ethics governance may be published on the initiative's website, for example.

With this in place, the main issue for consent may not be whether it is broad or specific, but rather whether individuals understand what they consent to and whether they are reasonably protected from misuse of the data:

> *Participants have consented to the use of the data and samples for use by bona fide researchers for health-related research. [...] That authorisation is broad. [...] My view is that broad authorisations are okay, provided the participants understand the breadth of the authorisation that they are giving. A consent to a broad range of purposes is no better or worse than a consent to a narrow range of purposes. It's whether you've really given a free and informed consent to whatever it is you're saying yes to.*
>
> – **Roger Brownsword**

Biobanks may pose challenges and may never fully be able to protect anonymity, particularly if they include genetic components (Greely, 2007). The key here is that individuals understand the benefits and potential risks, and are able to make their own decisions whether the former outweigh the latter:

> *The challenge for us as a research community is to develop more effective ways of communicating how data will be used, along with the benefits of use and the protections against misuse, to ensure that volunteers to this or similar data initiatives understand what they are consenting to. We also need to develop better ways of identifying any misuse of data and letting people potentially affected be aware of it.*
>
> – **Paul Matthews**

Another element of this is that participants may also give partial consent (for example, to participate in the research, but not be followed up via their EMR, or not to be recontacted again), and also have a choice about certain data not to be included in the research if they do not feel comfortable doing so.

In the end, this model may present a compromise to ensure that consent does not obstruct what both researchers and participants want – to make the best use of the data without undue constraints, but maintaining an appropriate level of individual protection:

> *Of course, consent is there to protect the privacy of the person the data is coming from, that is completely reasonable. But it can be difficult to apply modern, innovative ways of looking at data with overly restrictive data use conditions. I think we need to pause and ask if these data use conditions are really what the study participant wants. I believe patients and patient advocates often want a more open consent than the ways that these repositories have been built. We are good at explaining to study participants how we will implement security protocols to protect data, but not how excess data security may impede science and discoveries that may lead to faster cures.*
>
> – **Mette Peters**

It is worth highlighting here that alternative approaches such as dynamic consent have been proposed. These rest upon the idea that patients are given more control about how exactly their data is used. However, recent research comparing the use of broad consent versus dynamic consent in biobanking concluded that broad consent with ethical oversight may be the preferred model, while patient communication should also be improved (Steinsbekk et al., 2013). Also from the clinician's perspective, patients often expect their data to be put to a good use:

> *I think patients would expect you to do that, even if they haven't directly consented to it. I can understand the concerns about data privacy, but I think we have got the balance completely wrong. [...] My experience of*

> *patients is they simply expect you to make the best use of the data that you have got, and if the best use involves sharing it with other people, that is what they would expect – and frankly, I think that is our ethical obligation.*
>
> **– Rupert McShane**

While we do not argue that researchers should decide what is best for a patient by going over their head, the point here is that consent should not get in the way where the interests of individuals and researchers are congruent, but where the way consent was obtained acts as an unnecessary constraint.

Finally, for existing studies, it may in some cases be appropriate and worthwhile to re-consent to avoid rich longitudinal data collected over the past decades being discarded. At least going forward, however, consent should be obtained in a future-proof way that finds a compromise between enablement and individual protection.

Engage the public and manage trust

Giving broad and open consent also means that researchers must actively work on keeping individuals in the loop, partially as an obligation of researchers, but also to encourage individuals to "donate their data" for research purposes.

Research relies a lot on individuals' willingness to do something good for society, which may only indirectly benefit themselves. Countries like Sweden in particular have always had a climate of trust in research and people wanting to contribute:

> *I think the key thing is that actually the main reason why people take part in our research is not to benefit themselves, it's basically the same reason why they give money to charity, donate blood or do voluntary work – it's more to do with altruism and humanity than it is to do with personal benefit.*
>
> **– Tim Sprosen**

> *I think elderly people are very motivated to take part in research and express that, "I want to contribute to future generation's health. I realise perhaps it's not for me but if I do something now people will get it better in the future." So I think they really are positive taking part, we could say.*
>
> **– Bengt Winblad**

At the same time, many value being involved in ongoing research, also keeping in mind that some of the participants or their caregivers may have a high personal interest in being informed:

> *We do not have the problem. We take a sample and they agree. They come. So we do not really have to go through the voluntary stage because of the very big participation in Sweden. We try, of course, to encourage them. For example, every two years we do a day with all our participants and report*

> *to them what we have done with their data. [...] And usually there is a participation of around 1,000 persons in the whole day. And it's a day that we like when we are researchers too because it's a good feeling. It's a feeling that we are doing something together with the participants.*
>
> – **Laura Fratiglioni**

Especially when accessing routine data on an ongoing basis, thereby reducing costs and burden on participants, communication needs to be trust-enhancing in order to allow these data to be used in beneficial ways with individual consent. Incidences like the care.data episode in the UK or other cases in which data security could not fully be protected may potentially put patient confidence at stake, but also just the general notion of big data being exploited for commercial reasons:

> *In France we are always suspicious in the idea of finding some part of your clinical records to become a large scale research, maybe frightening these days with all these stories of Google and stuff like that. [...] Europe is a mess on it. You know, from France, my platform is in several different projects, it's already difficult to try to push forward the idea that this should lead to a national [platform].*
>
> – **Jean Francois Mangin**

However, often it may also be a question of how honest researchers are with the public about their intentions, and how something is communicated to the public in a way that makes clear that the data are being put to an appropriate use:

> *Using routine data, or getting people to wear devices or capture information – you would always find people who see this as an invasion in their privacy, but I think if you're very very clear with people and with the public about the question that you're trying to address – and dementia is a great one because it very much captures the public's imagination – and actually say, if you want to help us with this, these are the things that we want to do, and these are the risks and these are the benefits [they will participate]. I chair a research ethics committee and we should not see our role to protect people from researchers, taking a paternalistic perspective, but rather ensure the risks and benefits of the research are presented clearly and fairly so that people can decide for themselves whether or not they want to take part.*
>
> – **Tim Sprosen**

> *I think the way you frame the question already determines the answer you get. So if you say we want to do things with your medical records people get quite nervous. Whereas if we say we're trying to crack the problem of dementia and cognitive health, will you help? Then I think many people would most likely say, "Yes, I will. Tell me how."*
>
> – **John Drew**

The importance of communication should not be confused with deceiving participants, of course. On the contrary, participants need to understand the purposes and associated risks of their data being used, but also the potential benefits to making an informed decision. Concerning data included from outside the medical realm – which may be more sensitive and may not be covered by consent at the moment – individuals might also be willing to "donate their data". As a corollary, this might help to further increase individual participation and strengthen the dialogue with the research community:

> *So what would it take for this to be much more consumer-led? At the moment it's producer-led. It's about principal investigators and scientists who lead major research programmes, building them up and competing with other centres and that's the way the thing advances or doesn't. At the moment it feels like the message from behind a locked door is "Please wait another twenty years and we may or may not have a breakthrough". In the meantime people think "is that it?" So I can imagine that a big data initiative will at least signal the idea that you can do something useful to contribute. It's this idea of turning the tide. If everyone opts in to something and we get enough people then it starts to get interesting.*
>
> – **John Drew**

With data collection efforts such as UK Biobank receiving widespread interest, a subgroup of these individuals may even be prepared (and consent to) having their data linked by UK Biobank to enable further insights to be derived:

> *Another way might be to start with UK Biobank, you could go back to that group of people and say "Would you be prepared to allow us to match your consumption data with your health data". [...] The thing that struck me is that there are so many people [...] that somehow have been affected by it, I think you'd find a lot of people in their middle age who would happily give their consent.*
>
> – **Clive Humby**

Finally, involving individuals may also extend beyond the research space into the area of care, where those affected and their caregivers can contribute their data back into the research cycle. This may take the form of social networks or other online resources in which patients contribute data with the intention of advancing research and managing their own or their family member's care.

To summarise, engaging the public may help to gain the trust for open consent, but also provides access to new forms of accessing data, and feeds back results to those who are affected. At the same time, making the area more consumer-led may also improve the dialogue between researchers and research participants.

4.3 The ecosystem challenge

Recommendation: Using and sharing big data relies on an appropriate environment for doing so. Legislation must not impede medical progress, but provide a framework for data to be used beneficially. Equally important is linking to regulators and the private sector for exchange of data and expertise, and for streamlining the link between academic findings and future treatments.

To make sure that data can be used and shared with the aim of advancing dementia research, there should be a supportive ecosystem of legal safeguards, regulatory bodies and the private sector.

Secure a favourable legal environment

Building on the question of consent, there is the need for a favourable legal environment that stimulates research and reflects the balance between openness required for research and individual rights to privacy. This applies for example to the EU, where rules for medical research are relatively strict based on the tradition of data protection:

> *Research has to be stimulated and has to be done without too much bureaucracy. But now there are coming a lot of rules, that you may not take materials from the brain for post mortem studies, you may not do this and that. It's of course a lot of new legal aspects that have been brought in after we joined the European Union.*
>
> – **Bengt Winblad**

Treating pseudonymised data as equivalent to identifiable data in terms of consent would create major barriers to data flows within and across nations:

> *One potential risk is that if the EU Data Protection Directive requires that for every specific additional research project you have to go out and reconsent people this could prevent a lot of research from taking place. I personally think as long as the initial consent process is robust it is unnecessary to reconsent but there is a risk this could happen. If we have additional inappropriate regulatory structures imposed upon us with which we have to comply they may limit the usefulness of these resources.*
>
> – **Ian Hall**

> *That's fantastically sensitive in Europe right now and there is legislation going through the process I'm deeply worried about [...]. If that commentary was accepted in its current form obviously the top level of EMIF would have to end immediately. But actually so would epidemiology, I mean, it is extraordinary*

> *what that would do to research, it would be the single biggest stepback for public health that there has ever been – it would be disastrous.*
>
> – **Simon Lovestone**

While the legislation is crucial in order to provide a legal framework, it is especially important to make sure that there are no unintended consequences as a side effect:

> *I'm quite an admirer of the attempt that was made in Europe to articulate a set of data protection principles, basically principles about transparency and proportionality and that sort of thing, but I think that the Data Protection Directive got off on completely the wrong foot by anchoring this to the idea of privacy. [...] I mean really, it would be the final irony if data protection were the problem for big biobanks. Wasn't the whole raison d'être of data protection originally to allow information to flow across borders so that markets could operate in Europe as though they were a single market?*
>
> – **Roger Brownsword**

The approach to routine data may also have to be rethought to enable research to be conducted with all available data in a country:

> *In Germany, for example, they destroy hospital data for security reasons after three years. As a result, doctors do not see longitudinal patient histories of people with chronic disease (which accounts for more than half of disease burden). In countries like Estonia, big data has been successfully used for a while now. They have set out an innovative vision, and enacted it through policy. NHS England is setting up a data centre, adopting a similar approach to its very successful Hospital Episode Statistics and its long standing GP data research programme, GPRD, in which almost 10% of GPs participate. By linking this with health outcomes, climate, crime, deprivation, traffic data and other risk factors, we can identify new relationships which will drive insights and improvements. These examples need to be promoted to show other countries what is possible, and also to demonstrate the benefit of big data.*
>
> – **Nicolaus Henke**

It is worth highlighting that the law may not be sufficient in its own right, as compliance with the law does not mean that something is morally right and should be done, also because the law lags behind more recent shifts of what is deemed appropriate in a changing societal context (Nuffield Council on Bioethics, 2015). However, in its main spirit, the legal environment needs to follow the same logic as informed consent: Patients should be protected, and it is important to do so. At the same time, research needs a stable legal environment in which it can operate, without hurdles that needlessly constrain medical research.

Link in pharmaceutical companies and regulators

Another aspect of the ecosystem challenge is further collaboration between private and public actors in the space of drug development. Initiatives like ADNI and AddNeuroMed have successfully piloted public-private partnerships and shown their feasibility in terms of funding, which we will revisit in the next section.

Beyond funding, data collected in the pharmaceutical industry may also contain relevant insights, and making these data available may be a way of supporting further research in the academic space:

> *[We should] continue the ongoing and constructive dialogue with pharma companies about the release of their data. I think there are understandable reasons why they may be reluctant, not least of which being the data aren't necessarily curated in a form suitable for release and they perhaps are reluctant to invest in this level of curation. But nevertheless the idea of control data from pharma companies being available alongside information on failed trials i.e. trials in which the target has not been effective, would be highly valuable. It just saves everybody reinventing wheels. I think that is something that we should encourage.*
>
> – **John Gallacher**

> *An increasing problem is that we are losing the research possibility [in these well] investigated cases because the companies believe everything belongs to them and we believe everything belongs to us. [...] And companies, you know, have their protections. They say, "We have to run this study, which takes four years, then you can have access to this material." Of course, sometimes we could agree on something else, but it's very variable how these discussions go.*
>
> – **Bengt Winblad**

At the same time, the collaboration between the two has advantages for improving the translation of research findings into prevention strategies and actual treatments:

> *This is really about developing new treatments, and academics are unlikely to do that. After they've published their Nature paper they move on to something else. And so what you need is a part of the platform which is really committed to developing treatments: so okay, we have a signal. We're now going to take it through to see whether we have a treatment. And that will benefit the public far more than the next Nature paper.*
>
> – **John Gallacher**

However, the prospect for pharmaceutical companies to transform research findings into a treatment are fundamentally at variance due to the negative research economics in the field of dementia research, also referred to as the market failure problem. The risk/reward balance cannot be relied on to drive drug development, after several trials failed in the late stages, resulting in losses of millions for the pharmaceutical companies.

While the market failure problem in its entirety goes beyond the scope of this report on using and sharing data for dementia research, it should be highlighted here that some collaboration and data sharing across all actors may contribute to a more favourable ecosystem for research.

On the one hand, this is through making data from the pharmaceutical space also available for researchers. As private pharmaceutical companies also have an obligation towards their shareholders, tax incentives may help as a way of encouraging sharing without making it necessarily mandatory:

> *To encourage data sharing, governments could say [to pharma companies] "you only get this research and development tax incentive if you are going to share the data". So although technically governments won't be paying for data sharing, in practice it will provide a financial incentive. So that's the sort of incentive I have in mind. Pharmaceutical companies could choose not to share data but actually you'd make it worth their while, and [if they] don't want to share for whatever reason, they haven't broken the law, they just have a financial penalty in terms of the way their taxes are treated.*
>
> – **Derek Hill**

Furthermore, shared data has the potential to further contribute to advancing the field of dementia research. For example, through CAMD, ADNI data has already contributed to the process for ensuring regulatory buy-in from the FDA in the US and the European Medicines Agency (EMA). This is crucial as their approval of tools or biomarkers for clinical outcome assessments helps industry to ensure that they can apply these with confidence, and that the drug review process will not be impeded by reviewing the science behind the biomarker:

> *In 2011, CAMD was successful in achieving a positive regulatory qualification opinion for the use of imaging biomarkers for Alzheimer's trials. The publication on this success with the regulators as authors was published in 2014. Importantly, that regulatory decision was only possible because the ADNI data on imaging was pivotal in order to enable the regulatory decision. Without ADNI data we would not have been successful.*
>
> – **Diane Stephenson**

At the same time CAMD experience also highlights that there is not just one way to collect data, as ADNI may not be a reflection of the actual patient population. This shows the importance of making more data accessible, but also that there are several ways of building resources:

> *One issue that [the FDA] is concerned about is that the patients that enroll in ADNI may or may not reflect the true patient population in a clinical trial. So the FDA has asked us to go out and get data and provide them individual patient level data from different observational studies throughout the world, but more importantly clinical trials in which the biomarkers have been used according to the application intended. And so that's in fact where we're stuck. To date, we have not been successful at acquiring biomarker data from relevant clinical trials; some of the challenges are due to the fact that we are targeting an early AD population in which many of the trials are ongoing. The success of CAMD's AD clinical trial simulation tool, also catalysed by ADNI data, was enabled due to industry contribution of patient-level placebo data from clinical trials of mild/moderate stages of AD.*
>
> – **Diane Stephenson**

The CPI has also received patient-level control-arm data from 24 remapped studies with individuals in several stages of neurodegeneration from nine CAMD member companies and organisations. The standardisation and pooling of clinical trial data facilitates the analysis of data across multiple studies as another source of insights for dementia research. The CAMD database is an example of what can be achieved by standardisation and integration of clinical trial database from industry-sponsored AD trials.

Get other private companies on board

Beyond the pharmaceutical industry, linking in the private sector has a crucial role to play in tackling dementia in many other ways.

First of all, a lot of data from outside the medical realm are in the hands of private companies, who may be interested in making the data available or helping to find a group of people who would agree to their data being used:

> *Provided there is a consent mechanism, I think you would easily find that the big organisations would participate, because they want to be seen to be socially responsible. I don't think we would have any problem at all with the Sainsbury's and the Tesco's and the banks to start doing this sort of thing if there was something to do and there was a public opinion that this was a good thing.*
>
> – **Clive Humby**

> *Amazon or Google – if there is a substantial storyline that big data are quite useful, or sheds some new light on dementia research, they would probably be willing to do that because it's not their domain, it's not jeopardising their business, rather it increases their brand and the socio-economic value.*
>
> – **Yoshiaki Tojo**

On the one hand, the data analytics skills found in data-rich companies with data mining expertise will be valuable, as we will discuss in more detail as part of the skills challenge in section 4.5. Beyond skills, also a certain "private sector drive" may help to focus efforts and find an answer:

> *I see their input as critical and highly valued. It's good to have the energy of business around the table. The academics are very happy [...] to discuss the nuances of experimental medicine – we can have a whole afternoon session on this; we love it. Our industry partners bring us back and say: Okay, but what's the answer? And so it's really helpful to have them, if you like, encouraging us to draw a practical conclusion.*
>
> – **John Gallacher**

The difficulty in this regard may again be to make it worthwhile for companies to be involved. While large corporations may be enticed by improving their social responsibility vis-à-vis the public, smaller start ups might also be motivated to contribute through open challenges with prizes. This might also be a way in which consumer-led initiatives could make an important contribution:

> *If you took a different view the Evington Initiative could say we're going to have a £100,000 prize for the best app for cognitive health and research. We know we've got lots of twenty year research programmes and that's good but we want something else to balance it. We're going to create an incentive for people to come up with innovative stuff.*
>
> – **John Drew**

For example, a team at University College London has developed an app, the Big Brain Game, now with more than 60,000 users who are playing games, and thereby producing data for scientific research. While this highlights an interesting way to collect data from individuals, a collaboration between academic actors and private start-ups may again help to build up participant trust, as there may be lower trust when the suspicion of commercialisation appears:

> *The question that [participants] tend to ask most frequently is that they have concerns about commercial research, so if you say you have an important research question, you will get a different response if it was done by a university than if it was done by a pharmaceutical company, for example.*
>
> – **Tim Sprosen**

Of course, private partners may also be important for funding data sharing initiatives, particularly with respect to infrastructure, which may be compromised due to other priorities in the cure and the care area:

> *There should be a more flexible way to be in line with private companies and obtain funding. Otherwise a shrinking budget of the public health of some*

> *countries, money on the care side may take away money from the long-term based research, especially from funding infrastructure. So here we need some clever ways how these infrastructures will be valued, so value-for-money vis-à-vis money for the current patient.*
>
> – **Yoshiaki Tojo**

This point also extends into the wider question of how data sharing can be funded sustainably, and how funding in itself may be a lever to increase data sharing and use, as discussed in more detail below.

4.4 The funding challenge

Recommendation: The funding challenge consists of ensuring that data sharing is sustainable, and that more investment helps to advance dementia both directly and indirectly. Funding and proper data management may ideally be built in from the beginning in order to open up data later, while funders also have a critical role to play in mandating and incentivising how data are used and shared.

Funding of research in general has received widespread attention, with several charities operating in the field of dementia research, and specificially Alzheimer's Disease. Nevertheless, dementia is not where it could be, with other diseases being funded particularly through charitable donations to a much larger extent (Luengo-Fernandez et al., 2010). However, with the current structures and the market failure challenge, dementia research in itself may not be very "investable":

> *I think the lack of money in the space is probably a symptom of the problem, although it becomes a problem in itself as well.*
>
> – **Nick Seddon**

> *[In the Evington Initiative] initially the business leaders felt that they wanted to do something and that more money was the thing that would help solve it. And very quickly they found that it was not that straightforward and that, actually, the research is quite stuck so giving more money may not help. That led to asking how do you make this sector investable? Because people that are very clever, wealthy, used to big investment would look at this and say it's like a jungle, why would I ever go in there?*
>
> – **John Drew**

While funding as such extends beyond the scope of data, by the funding challenge we specifically refer to the fact that beyond research, sharing data may also need more funding to enable resources to be put to a larger variety of uses, and creating sustainability in an environment of scarce resources. The challenge here is of course having limited resources:

> *Resources are limited – and any funds spent to support data sharing take away from other scientific projects.*
>
> – **Michael Weiner**

Of course, looking at the bigger picture and a long-term perspective, investing in infrastructure offers the promise of reducing costs due to being able to exploit data more thoroughly and avoiding duplication of effort.

Fund data sharing in a sustainable way

First of all, for smaller studies that have not been designed with the aim of building a resource it is important to make sure that data management is planned for opening up the data from the very beginning, and to make sure that data sharing is not neglected at the end of the study due to a lack of funding:

> *[Funders should] force, or encourage, to adhere to standards and to have good data management practice in place right from the start. I'm afraid many cohorts are recruited and a lot of money spent on recruiting these cohorts and data being generated and the data is stored on Excel sheets on the post doc's computer under their desk – and you don't need to mess up clinical data much before it gets completely befalsed or the study is useless.*
>
> – **Richard Dobson**

Getting data management right from the beginning is important, as low quality may be one of the reasons why data are being held back, with transparency increasing the risk that faults in publications may be discovered:

> *I suspect subconsciously [researchers] might be worried about the fidelity of the data, how good the data is. Because there's nothing like having other people rummaging around in your data to discover maybe it's not quite as good as you thought it was.*
>
> – **Simon Lovestone**

After opening up data, it is also important to highlight that there may be some cost for continued data sharing, which often may get forgotten. It is important to actually budget for continued resourcing from the outset, as often there is no money left in the end to open the data to the wider public, and the resource cannot be used as well as it could be:

> *The key issue is going to be ensuring that the platforms are resourced in a way such that all of the data can be accessed and analysed by the research community in a rapid and efficient manner. [...] I've seen it in the past with these sorts of projects, where the data are all there and the expectation is that they can be accessed, but the approval processes take six months or*

> *more to work through. Even then, once a proposal is approved there may be no resource available to actually deliver the data in the format that is required to the individual who wants it.*
>
> – **Ian Hall**

> *The problem is sustainability [...] as the data cleaning and data dissemination and keeping this data available for the broader community costs money to do, and it's not necessarily rewarded scientifically or academically.*
>
> – **Yoshiaki Tojo**

One way to do that is by charging a small fee for data access, as UK Biobank does for example. The other way may be to recognise the importance of infrastructure as an enabler for research, and specifically fund it. For example, the Alzheimer's Association has funded GAAIN on the basis that finding data for dementia research is important for progress in this area. Data sharing as an enabler makes it very attractive for funders:

> *The funders can see that money they have put in is potentially going to be an investment in a huge amount of research productivity that follows [...] It has been very successful in terms of a funding model so not only is it highly cost effective to study half a million people at the same time, as long as you have scalable methods, the per participant costs are very much lower [...] if you can justify the sort of level of investment that is required by ensuring that the data are made widely available.*
>
> – **Cathie Sudlow**

At the same time, getting prospective cohort studies to be funded may be difficult, as the low-hanging fruit or quick wins may be hard to realise, given the resource only gains value over time.

> *So we have to communicate with funders that these models take time to deliver because much of the scientific community is just not geared up for operating at this level. [...] And if you've got researchers who need to deliver grant income by a certain date, and funders who need to see that the platform is successful by a certain date, there can be a conflict of time pressures which is a challenge to developing collaborative infrastructure.*
>
> – **John Gallacher**

The reference to the way the scientific community works also highlights its relationship to the underlying incentive structure in the academic system, which we will revisit later.

Encourage sharing through tying research funding to it

Beyond data sharing initiatives needing continued funding, funders are in a position to mandate sharing from the outset, thereby leading to clear agreements from the beginning, which may be changing the system in which researchers operate today:

> *It's not the fault of the individual researcher who doesn't want to share. I think that's how the structure is built up at the moment. If you have the funding source require them to share from the beginning, then I think it's much easier. [...] I think that it's the whole general system of funding. You need to get funding, but you also need to publish to get more funding.*
>
> **– Eric Westman**

Funders like the Wellcome Trust are already mandating data to be shared, thereby also setting standards for others to open up data:

> *We're very clear that [...] all our researchers have to fill in a data sharing section on the application form where they are mandated to explain how they will make their data available to the wider community. And I think it's setting standards which many, many funders and increasingly publishers now have. That is gradually pushing communities, certainly using human data, to address these issues and ensure that that knowledge can be shared widely and for the common good.*
>
> **– John Williams**

Also for government funding, the prevailing view is that data should be opened up, as for example in the case of Sweden:

> *All of the Swedish research councils have a similar policy. They want really that your database is open for also other researchers. And it's open because the money comes from all the people, because this comes from taxation – so these data are not my data when I collect it. Of course, it's my project, it's my baby. But I am very, very aware of the fact of that I'm using money from all people [...] in Sweden.*
>
> **– Laura Fratiglioni**

This comment about using money from national taxpayers to share data globally, and thus potentially benefit the whole world, highlights the fact that most scientific funding is nationally or regionally allocated and the default mechanisms for sharing are frequently through national organisations following rules written within the legal framework of that country. On the one hand, taxpayers expect the research they fund to help their country first and foremost, but on the other, the potential fears of free-riders scooping up data that will not benefit those who funded it seems less acute than the real need, identified throughout this report, to scale up research to a global scale if the mechanisms behind dementia are to be more fully understood and with a benefit to the whole world.

Finally, funders may not just help to force sharing, but can also help to shape a broader dialogue with the involved parties, which beyond funding also contributes to building an ecosystem in which data can be used and shared efficiently:

> *Yes, we can fund research, but a funder like Wellcome – because we fund into culture and society, we hold exhibitions, we promote public dialogue – we are in a position to help shape a wider conversation about the disorder, about how we should study it, about the societal implications of wanting to trawl through and explore complex datasets. We can bring people together as a convener to talk about these issues and potentially think about framing and innovative applications. So it's that value added piece that a funder can contribute, and then not just an individual funder, but funders acting in partnership together to inform the debate.*
>
> – **John Williams**

4.5 The skills challenge

Recommendation: To address the skills challenge, further capacity building in bioinformatics and interdisciplinary exchange between medical experts and data scientists are required. Crowdsourcing insights through open challenges may be a way to expose the data to a wider community beyond the medical realm for some datasets.

Using big data for advancing dementia research relies upon the right skills to do so. While interdisciplinary research due to the cross-specialty character of dementia is important, large datasets also require novel skills:

> *It's when you flip it the other way round and say, here's a huge morass of information, what do we want to do with it? Will we find spurious correlations? How do we even begin to approach that? [...] And it's going to need some smart data scientists to engage with it and think about how we ask the questions and how we then interrogate the data.*
>
> – **John Williams**

Find or train more health-/bioinformaticians

A common theme across interviewees was the lack of well-skilled individuals who can analyse the data, but also understand the medical side behind it:

> *The funding for biomedical research in the United States has been going down. This is reducing the calibre and number of trainees, because they realise that if there is no funding there will be no jobs. Ultimately we will*

> *pay the price for this at some point. Other countries are making significant, healthy and appropriate investments in this kind of science, and they should be applauded.*
>
> – **Arthur Toga**

While data sharing can level balances between those who have data and those who do not, educating more bioinformaticians, health informations and related interdisciplinary educational paths and tapping into global talent pools will be important to harness big data:

> *We understand that there's a problem, but I don't think we realise how big a problem it is. […] Right the way through from really young kids in infant school, the teaching of maths is shocking, we don't have enough young people coming through our universities, and a big problem for us is immigration. […] If we could tap into the informaticians coming out of Bangalore and Shanghai, better, it would be very good for Europe.*
>
> – **Simon Lovestone**

There has been a growing recognition by universities around the world that the broad area of data science is an area of acute need and likely job growth. However, data science can be applied to many domains, including business, finance, and other highly lucrative options, so talented young people with interests in gaining data skills must be encouraged to apply them to medical research by offering them incentives which go beyond the strictly financial, such as the excitement and satisfaction possible in careers at the cutting edge of medical research.

Bring together talent from different disciplines

As important as the need to train more health-/bioinformaticians, may also be the need to increase collaboration across disciplines. Of course, this first of all includes medical researchers from across disciplines, for example, within imaging, or within different areas related to dementia, with the intention of integrating risk factors "below the neck" more closely. However, for handling and gaining value from big data, it is especially important to bring in those with a data-driven background:

> *My experience in this kind of topic is that it works best when the people with the different skills are actually working really closely together. […] Epidemiologists can create fantastic databases but they may not be the best people to pull out the signals particularly when the data become incredibly complicated. We are struggling with how we turn data into information.*
>
> – **Sir Rory Collins**

> *The problem may be more on the researchers' side, who do not have enough literacy or at least network for the big data science. So what big data is available, how big data should be handled and how big data should be combined*

> *with the deep data the researchers usually handle. So there would need to be more cross-disciplinary research with computers scientists or social network scientists, those kinds of groups that are handling those types of data on a daily basis. Their insights to the deep scientific research in dementia research would be helpful.*
>
> – **Yoshiaki Tojo**

Bringing people together from different backgrounds may not just be helpful with exploiting existing big data in the medical system, but also with integrating new sources. Currently, one of the reasons for big data from outside the medical realm not being used more widely may also be the lack of interdisciplinary collaboration:

> *The simple answer [why data from outside the medical system is not used more] is probably that most of the discussions in terms of epidemiological studies are generally people in the medical profession talking to other people in the medical profession. They understand each other very easily because they speak the same language, but we are speaking to the wrong audience and what we need to do is to have a meaningful discussion directly with the public about the risks, safeguards and benefits to all us of research being an integral part of health. There are positive developments in this direction – for example, the updated NHS Constitution refers to research being part of the central role of the health service.*
>
> – **Tim Sprosen**

This collaboration may happen with single research groups in universities. However, establishing multidisciplinary centres of excellence for research on neurogeneration may also contribute to achieving better results from big data:

> *It needs collaboration in the way that [we have] centres where basic researchers work in a close environment with the clinical trial units. We need to have multidisciplinary work [and] centres of excellence that really cover everything and would very quickly come to results but also have a back translation from knowledge of the patients back to the basic science.*
>
> – **Bengt Winblad**

Crowdsource other people's skills and creativity

A third way to effectively use skills may be to open up data to those beyond the medical sciences. Past open challenges and data hackathons showed that collaboration between those with medical knowledge and those with deep data mining expertise is very fruitful in knowledge discovery (Celi et al., 2014). For dementia, this has been done through the Synapse DREAM Challenge. Of course, the results from open challenges will have to be followed up with proper medical knowledge and by no means replace rigorous controlled experiments:

> *I don't believe so much in the value of abstract big data analytics for advancing the science of dementia treatment. I feel that you are likely as anything else to get invalid conclusions out of that sort of activity. [...] The findings from such big data approaches would probably not convince either regulators or the executives in pharma companies making decisions about investing in clinical trials.*
>
> – **Derek Hill**

Also in relation to the previous discussion on access models in chapter 3, some data may not be suitable for crowdsourcing. This may not only refer to medical data, but also big data from outside the medical realm which may be queried for unintended secondary uses:

> *[Loyalty card data] could be crowdsourced, but I think that's when the organisations would get more nervous. [...] What is to stop someone from using the data for a completely different purpose? [...] So I think the problem with crowdsourcing is that the data is too sensitive. You've got to do this in a closed environment, in some form of a data-safe environment, because the data itself is commercially sensitive.*
>
> – **Clive Humby**

To summarise, crowdsourcing may present an opportunity for some data (or parts of data) to be exposed to a wider community – as just one of the many ways in which the pieces for advancing dementia research are coming together.

4.6 The incentive challenge

Recommendation: The incentives challenge needs to be addressed to make building and sharing resources worthwhile. Acknowledgement may be given, for example, through providing a citation for the dataset that counts for academic metrics, or recognising contributions to building resources in hiring and promotion. Moreover, a model of adding data and feeding it back to the main resource may help to establish a win-win situation for data collectors and users.

As shown throughout this report, the question of reward is fundamental to the question of sharing and opening up data, going beyond the technical and organisational aspects of data sharing:

> *In the end, if a principal investigator wants to share, and can find a website to support their data, they can share. But there is no reward for data sharing. It's more work. Most investigators just want to get their papers out.*
>
> – **Michael Weiner**

Make building and sharing attractive

Currently, there are few incentives to share data the way academia is set up, with recognition being mostly attributed to publishing papers. Incentives are a crucial underlying mechanism both for building resources of future value and for sharing data more generally:

> *So we are talking about very expensive studies. And when you get funding for a big study like a study of dementia, longitudinal with a huge amount of people, people are very concerned to publish as much as possible but you are not really willing to give away a dataset to make others do what they want to. [...] If you are the principal investigator you want to have the control at least for several years to do what you have planned to do.*
>
> – **Linda Hassing**

As discussed earlier, funders may play an important role in making sure that data are made available. In general, creating very clear arrangements upfront that data have to be shared may solve the problem:

> *I think the funders can have a really important role to play by putting pressure on researchers that they give money to share their data, and I don't mean a clause 15.8.3 that says "You will share data", I mean a discussion upfront in person, "You understand you only get this funding if you make your data open". That – and funding to enable it to happen, as sharing data properly is not cost free – would help. I don't see any reason why that shouldn't happen.*
>
> – **Simon Lovestone**

Potentially funders could even consider creating financial incentives to sharing:

> *You need to have a real lever to encourage academic data sharing. So, for example, some of the funding bodies at the moment withhold some (e. g. 10%) of the grants until you send in your final report, and you could use the same approach to encourage timely data sharing: academic researchers wouldn't get a chunk of the cash from an academic grant unless they put the data into a trusted third party. I think it has to be that brutal.*
>
> – **Derek Hill**

However, one has to keep in mind that often funds are insufficient for cleaning the data properly before being released – which may be exacerbated by holding back part of the grant. Similarly, structures to create pressure alone will not help. At the core, researcher acknowledgement will need to be addressed:

> *Some of the pressure [of funders] making data from resources freely available is constructive; however sometimes it does not necessarily really reflect the*

> *huge investment by the researchers who built the resource. Funders and various other people talk about how this work should be acknowledged but they don't actually say how it will be acknowledged. [...] So I think that unless that is addressed you will get coercion but you won't get collaboration generally in the concept of wide access.*
>
> – **Sir Rory Collins**

One further way of making sharing attractive may be to create a mutual value-add, a win-win situation, through enhancements that the data user makes to the overall resource. This has been the practice in several informal collaborations:

> *[We] struck up a relationship with the PI by introducing ourselves and saying, "We'd love to generate this data on your cohort and pay for that," in order to buy our way into a collaboration. [...] So we offered to generate protein data on a subset [...] to increase their coverage in terms of the data that they have, and in return they allowed us to perform an analysis to get a first authorship paper and then to get leading or highly ranked authorships on follow up.*
>
> – **Richard Dobson**

As outlined earlier, UK Biobank is in the process of institutionalising this form of data enhancement in order to keep developing the resource with external funding, and then make enhanced data available to others. These enhancements may be managed as a separate addition to the main dataset, with other researchers having the opportunity to recreate it if they suspect that there may be errors in it.

Manage acknowledgement

Much of the underlying issues in relation to data sharing are about incentives. In academia particularly, collected data are one of the key assets for a researcher:

> *People's careers are tied up in this, I absolutely understand that, and I don't think we as a community have paid enough attention to how to deal with that. We're very good about saying, you know, "It's not your data, give us your data, I want access to your data", but actually how are we going to respect the time that these people have spent on acquiring this data?*
>
> – **Simon Lovestone**

The challenge here is that while the academic reward system works based on publications, many interviewees expressed it would neither be fair that those creating resources would suddenly have "300 or 400 publications overnight" due to being co-author on publications which they only provided the data for, while authorship would also create a sense of necessary quality checking of the paper, which again may not be scalable:

> *For authorship, you can't just be giving data, you also really should give some sort of an input, and you should also be helping with the writing up the data and things like that. So we generally don't think it's enough just to collect the data, but you should always be asked if you're in, if you are willing to analyse and give comments and suggestions for the paper. And also that you agreed on the final version.*
>
> – **Ingmar Skoog**

Our Swedish case studies also showed that one of the reasons for relying on the collaborative approach of releasing data may be a lack of researcher acknowledgement in other modes of releasing the data:

> *PIs are putting much effort in establishing and maintaining high-quality cohorts. So how to acknowledge these efforts when data is shared is an important question. And I think that's why so far we have had the tradition that it's a collaborative approach if someone wants to use the data and there has been not so much open access data.*
>
> – **Miia Kivipelto**

Of course, acknowledgement is not only needed for sharing datasets that have already been collected, but especially for large cohort studies which require substantial investments by people involved in them without immediate rewards. However, indirect benefits to creating resources such as knowing the resource very well and shaping it exist, but work in a much more implicit long-term way:

> *I have had to think quite carefully how I ensure that [junior researchers] get personal recognition and career progression through doing this work. Part of that comes from being involved with a big resource like UK Biobank and being able to say that they have worked in the UK Biobank team, because that has a certain cachet. Part of it comes through ensuring that not only are they involved in resource building but that some of their time can be used for conducting research. And part of it comes because if you are involved in building a resource you tend to have a better idea of what it is valuable for, in terms of research.*
>
> – **Cathie Sudlow**

Beyond what is already being done – mentioning the data sharing initiative in the acknowledgement section, abstract or keywords – other ways may help to ensure better individual acknowledgement. For example, Rohlfing and Poline (2012) suggest publishing the dataset as papers are published, which may then get cited in the papers using the data and count towards academic performance metrics, which often include a mixture of publications and citations (such as the h-index).

In addition, other research has recommended to find distinct career paths for helping with curating data and making them available to a wider community (Howe et al., 2008), or treating data as equivalent to publications (Gardner et al., 2003), as well as creating additional incentives by recognising data sharing efforts in academic hiring and promotion decisions beyond publications, journal impact factors and citations (Piwowar et al., 2008).

These measures may not transform the way academia works in the short term, but may gradually shift the balance for the academic community to acknowledge the effort that individuals have put into building large-scale resources.

4.7 The mindset challenge

Recommendation: The mindset challenge occurs at the most basic level and may be addressed through sharing success stories and initiatives transforming how science is done. Moreover, it is important to create a "big data mindset" to relevant data outside the medical system and novel analytical approaches.

Create a sharing culture

The mindset challenge occurs at the most fundamental level of how we will approach data analysis and data sharing in the future:

> *Widespread data sharing is not currently part of the culture. [There is] not sufficient enthusiasm for data sharing by the scientific community.*
>
> – **Michael Weiner**

At the same time, the initiatives introduced here present models that have the potential to redefine thinking in this space by sharing success stories, forcing other initiatives to open up and creating a mindset towards responsible data sharing:

> *Imaging data has a special impact on people because pictures communicate so well. UK Biobank is developing an imaging component of unprecedented scale. The most well-known – although much smaller – example is ADNI. [But] I think that the importance of ADNI actually doesn't at all lie in the dataset that they have developed, although this is very useful. What they did is find incentives for researchers to work together, address ethical issues involved in consent for open sharing of imaging data and demonstrate the value that can follow. It brought the best research groups across the US together. People who had hitherto only competed with each other became part of a team. This has set a higher standard for us all.*
>
> – **Paul Matthews**

> *I think UK Biobank is redefining the way in which we develop resources and share data at least to some extent; because it is so big some smaller resources are at risk of being squeezed out of the picture unless they start to make their data available in a similar way.*
>
> – **Cathie Sudlow**

Similarly, these initiatives may be pushing researchers to reconsider their principles. Several of our interviewees highlighted the importance of "the deal being clear upfront" to make sure everybody is on the same page, especially if there would not be any preferential access to the dataset by those who created it:

> *There were people in our team initially who were really not willing to give up that intellectual property and we made a decision that they are not part of the team at the outside. And [...] the same people who went away, [...] now they are coming back. So I think it was just a bit of a culture shift that we had to do, and it is still an experiment; we hope people use it because we have made it open.*
>
> – **Parminder Raina**

Research on data sharing in the area of cancer research found that sharing detailed data increased the citation rate of the original papers, thereby benefitting the researcher who has shared the data (Piwowar et al., 2007). In the end, it may be about getting the message out that data sharing is positive, which may in turn slowly turn the tide. Similarly, success stories about data sharing initiatives should be spread more widely to push others to rethink the way science is done:

> *I think what we are not doing a good job of is sharing success stories. [...] When I go out and talk about Alzheimer's Disease the first thing I do is talk about ADNI because it showed that in order to enable success you need to do two things: You need to agree on consensus data standards and open sharing of all the individual patient level data to anybody that can request access. [...] And those two things, if that happened globally in clinical trials and academic centres throughout the world we would be in a totally different space. New initiatives demonstrate there is positive movement in this area.*
>
> – **Diane Stephenson**

Manage the fear of being scooped

Closely related to this is managing the fear of being scooped: that somebody might download all the data and exploit them before those who collected them can do. This may be the major issue why data has not been shared to a greater extent:

> *The biggest challenge: Fear! Some data owners may have an initial reluctance to share, because they feel that they've put their heart and soul and time and effort and money into collecting these data – and why should they share*

> *it with anybody? We are experiencing a transitional period in big science, where the data sharing motivation is slowly being realised by people, but it's incomplete yet. So people need time to get used to the notion, they need to understand and see how they can benefit professionally and how it accelerates the pace of discovery. I believe it is best to use a positive motivator rather than the threat of a negative one. But it's an educational process.*
>
> – **Arthur Toga**

There may the worry of "someone downloading all the data and publishing everything", which may apply to individual researchers as much as to whole countries:

> *If UK has made a significant investment [...] and then if you open that platform to anybody who wants to apply, then the possible risk is that a big group in the US or elsewhere ask for all the data, get given approval, download everything and then scoop the UK scientific community in terms of all the interesting findings. You'd actually argue from a scientific point of view that that isn't a problem because actually what really matters is that we maximise the use of the resource. But if the UK has made a major investment in these platforms then you could argue the UK science base should have the opportunity to maximise its use of them first.*
>
> – **Ian Hall**

As a compromise, there might be a very short embargo period during which researchers have exclusive access to the resources they created. This period should strike a balance between giving the researchers involved in the collection a head-start, but without keeping the resource closed for an extended period of time:

> *If you are three months ahead of everybody else and they know that you are doing this, it is unlikely they are going to move really fast to try and catch up with you, as by the time they've accessed the datasets and undertaken analyses it is likely that you will have published the paper anyhow. It varies a bit from one dataset to another but my gut feeling is three to six months is a reasonable period of time without it inhibiting the rest of the scientific community. But then the model has to be in principle that the dataset is open to the whole scientific community such that everybody gets the maximum benefit from them.*
>
> – **Ian Hall**

Of course, this period is of no use for larger population-based studies, where cohorts only gain value over time. Again, with proper acknowledgement mechanisms, those who built the resource may in fact benefit from sharing – with chances of having a large population-based study perfectly analysed within a short amount of time being very small anyway. Sometimes the indirect benefits of sharing may be less evident to researchers:

> *It is going to be important [...] to convince clinicians and clinical scientists that, if they include their data in your jointly larger dataset the findings may be more significant, so the papers they will be on will be more important than if they'd published only data on their own smaller dataset. Therefore I guess that's one of the things that will be important: just to convince people of the value of cooperating.*
>
> **– Stefan Larsson**

Create a big data mindset

Finally, one of the biggest challenges in terms of using and sharing big data is a shift in the mindsets and a conversion to a new paradigm in research.

> *There is much data out there and now when we are using so much technical equipment, I believe it may open new opportunities in the future. Maybe we have been thinking in a too narrow way what data we can access and some of it goes back to the question about the informatics and how we can capture this data. But I believe we should try to use that kind of data in much broader terms.*
>
> **– Miia Kivipelto**

As an analogy from the past, the value of non-medical data has often not been fully recognised until it was in fact analysed:

> *One of the studies that has been really important for the development of our knowledge about dementia was [...] when we showed that a rich social network could be a protective factor for dementia. When I started in Kungsholmen, I was completely not interested in all these social factors. I'm an MD, so it was far away from me. And I remember, "Oh my God, how many questions we have about social networks!" I was annoyed about this because it costs – until I started to be more involved. And then it was only for a series of events that I came up with some idea that maybe the social networks could be relevant.*
>
> **– Laura Fratiglioni**

Similarly, big data analytics, in which a certain degree of lower quality is more accepted based on the way the data were obtained, may clash with the quality demands that academics from the medical research tradition may expect:

> *Some of the longitudinal data has not been shared up until the DREAM Challenge. The Synapse team take a different approach, and I'm paraphrasing their view but it's broadly "Don't worry too much about the quality control, give us the data in all of its mess and let the community deal with it. So long as everybody is aware of where the problems lie let's see if we can crowdsource a solution," as it were. And there was a degree of a culture*

> *clash there which we had to manage. We managed it but it was interesting – it's a different approach to data.*
>
> **– Simon Lovestone**

Moreover, for other data that would traditionally not be considered medical – and even more so for routinely collected data obtained outside of the medical realm – there is a high degree of scepticism among the research community:

> *[Using consumer data from outside the medical system] kind of remained an idea. I think the scientific community was much more focused on established things like Biobank and the samples. But I can imagine it would be powerful to do both [...] in parallel, and it depends what kind of Biobank samples that you have, but to potentially have a massive longitudinal cohort that includes tracking consumer data as well as scientific data.*
>
> **– John Drew**

Although academics are incentivised to apply innovative thinking, sometimes the lack of a past history may prevent new forms of research being done, and using big data from outside the medical realm may not be of interest to researchers:

> *I think [the problem is] finding someone who wants to get their hands dirty and start analysing the data. [...] Whilst [in discussions with academics] this was seen as "Yes, that might be helpful", it didn't feel like it was on anyone's agenda. I guess you need an academic who has an interest in the topic. I was talking to a couple of academics, and I think because of the research course on previous research, if there's never been any previous research, then it's much harder to start something - that's the impression I got.*
>
> **– Clive Humby**

However, with new opportunities continuously coming up, approaches to science also change over time, with excitement for new opportunities gradually taking over scepticism, thereby broadening the realm of finding new ways of advancing dementia research:

> *The big new challenge which is really going to be transformative [...] will be streaming data from pervasive computing opportunities and then we really scale, and that's really exciting. A degree of scepticism, but nonetheless I think the excitement outweighs the scepticism in the field at the moment.*
>
> **– Simon Lovestone**

While a precise assessment of all levers that may come from new sources, new approaches and other technological innovations, this should be looked at more closely in future research, as the value expected from these sources will be a crucial driver for developing the mindset for these sources to be unlocked.

The way forward

Computers are incredibly fast, accurate and stupid. Humans are incredibly slow, inaccurate and brilliant. Together they are powerful beyond imagination.

– **attributed to Albert Einstein**

Using big data efficiently with novel fast and accurate analytical approaches, while allowing more humans with brilliant ideas across disciplines to work on it, is crucial if we are to push the boundaries of imagination in pursuit of the ambitious goal of finding a cure for dementia by 2025 as set out by the G8 nations.

Dementia research has advanced over the last years and decades, and with the resources and ways of finding and sharing data built over recent years, important foundations have been laid. As we have seen, there have been successful efforts from many dedicated scientists and their teams to advance the science (prevention, diagnosis and treatment) of dementia by establishing shared data and robust research infrastructures.

However, our research with leaders in the field has generated some overarching themes and observations relevant for using big data for advancing dementia research going forward. In this final chapter, we will synthesise the major recurrent themes that have occurred throughout this research, and highlight the insights relevant for suggesting a way forward, both in terms of medical research practice and in terms of the role public policy can play for advancing dementia research.

5.1 Steps on the path ahead

The way to finding a cure for dementia may still be a long way ahead. However, by enabling global collaboration and exchange, addressing systemic questions and transitioning towards a new research paradigm, the process of scientific discovery can be facilitated to support the science of dementia research going forward.

Collaborating globally

There is no lack of data for dementia research, but we need to exploit it more effectively. As discussed throughout this report, (big) data of relevance for dementia research comes in various formats: Large population-based studies, smaller studies that can be pooled, routine data in the health system, and other routine data from various areas of everyday life that have not yet been harnessed.

> *I think it's time to understand what all these things are and how they link together in the progression of the disease [...] People are building up more and more, but [...] I'm not sure we need so much more data, to be honest.*
>
> – **Bengt Winblad**

The abundance of data highlights the importance of sharing and combining them. Sharing across research teams is a prerequisite for resources to be exploited more fully, as opposed to keeping the data behind closed doors while collecting more elsewhere. This also includes combining new forms of routinely collected data, which promises untapped potential especially in combination with research data.

At the same time, collaboration with other disciplines will be important due to the multifaceted nature of dementia. This includes collaboration with other brain disorder researchers, but especially with those researching risk factors "below the neck". Finally, collaboration with engineers, physicists or innovative private sector organisations may prove fruitful for including new sources and applying new ways of analysing the data.

No one nation has it all, but complementarities exist. As already set out in the G8 Dementia Summit Declaration, dementia is a disease that concerns all nations in the developed and developing world, so that the economic challenge intensifies as life expectancy increases across the globe. At the same time, global collaboration is required to achieve progress on the dementia challenge.

In order to enable global collaboration, it is crucial that just as diseases do not respect national boundaries, neither should research into dementia be seen as purely a national or regional problem or priority. It is inevitable that many funding bodies will continue to operate mainly within national political boundaries, but nationally funded research can be made more valuable and have a bigger impact for everyone if the data can be combined with research or routine data originating in other countries. To enable this will require additional incentives to share and legal frameworks that protect both researchers and research participants.

At the same time, global collaboration will help to compensate the imbalance of resources: Despite some general scarcity of highly-skilled health informaticians and bioinformaticians, some countries have a larger pool of researchers, but they do not necessarily have the deep nor broad data available, and need to spend time and effort on data collection. On the other hand, countries such as Sweden have huge amounts of data but lack the human resources to analyse them fully:

> *I think we are sitting on goldmines, but we don't have any people who really are taking out the data. [...] Perhaps in collaboration with other countries, that they can bring in the researchers.*
>
> – **Bengt Winblad**

There is no single way to build and share data – but "openness by design" helps. In this report, we have introduced key differences in terms of access procedures, access models and kinds of data included. We have shown that different models come with different benefits and drawbacks, mostly around scalability and data made available. Nonetheless, there is not one single model, rather a multitude of approaches should be enabled in the marketplace of ideas:

> *I think that there is not only one model how it should be done, but I really see different countries and studies for different purposes, and what can be done with the data that is there. If it is not possible to share data, we can share the results and experiences from different studies and groups to get a comprehensive picture of the research questions.*
>
> – **Miia Kivipelto**

ADNI for instance is unprecedented in terms of the ease of getting access and providing high quality imaging data that set standards. Then again, it may neither offer the broad data to make claims with the same confidence enabled by big numbers, nor may it allow us to look at factors outside of dementia research or longitudinal data. Then again, resources such as UK Biobank or the Swedish Brain Power studies may offer precisely these benefits, but are not as easy to access, as they contain larger amounts of data, may need more protection due to the nature of their participants, and in the case of Sweden, due to the obtained consent.

While it may be tempting to consider combining the best elements of each model, it is worth highlighting that some of these characteristics are interrelated and conflicting in various ways: Making data openly available online may prevent research participants from participating. Also, while UK Biobank will eventually contain unprecedented longitudinal data, initiatives such as ADNI and AddNeuroMed led to research findings within a much shorter time frame.

Just one feature, at the risk of stating the obvious, will help promote sharing and using big data to a greater extent in the future: Openness should be designed for from the beginning – in terms of how the data are collected, how data are managed, how links to other (routine) datasets are planned for, how participants are adequately informed about risks and consequences associated with their participation, and how governance structures for responsible use and sharing of data in a changing context and environment are set up. We do not advocate for all questions to be solved from the beginning – in a changing environment, new questions will come up and will have to be dealt with. However, it is important that funding for sharing exists beyond the actual project, and that there is a common understanding among researchers and participants about how data will be made available to others from the outset.

Reducing barriers and aligning incentives

Advancing dementia research requires addressing "all of the iceberg". It will be crucial to address the more obvious technical questions, manage consent so that it protects individuals without hampering research, create an ecosystem in which data can be shared legally and all involved actors collaborate, get funding to build resources and enable sharing, build up the skills needed to share and exploit data, create incentives for people to do so, and align the mindsets towards sharing and incorporating other datasets.

Only by addressing all of these challenges jointly will it be possible to continue the positive trajectory created by the data sharing initiatives during the last years. Failing to address part of the problem may result in a lack of action – the weakest link in the chain principle.

Recognition and reward for data creation, curation, and sharing efforts are the foundation for making change happen. People-related challenges are usually the hardest to tackle. Spurring collective action for fighting dementia will only be possible if individual incentives exist for people to build resources and to share data for others to use. This may include funding and financial incentives, but at the most fundamental level it starts with providing career incentives and giving recognition for activities that benefit the field as a whole. In the current world of research, there are relatively few incentives or rewards that accrue to data creators and curators. Most academic recognition goes to authors of publications reporting key findings or to patent holders.

Rather than a one-size-fits-all model of research and publication, additional ways to recognise the value of shared data must be built in to the system. As outlined earlier, treating datasets like publications and thereby incorporating them in existing incentive mechanisms or acknowledging data sharing efforts for hiring and promoting decisions may contribute to make sharing more worthwhile for everyone. Inspiration for these new models may come from other domains: For instance, it is worth considering whether the model film makers use to give credit to a wide range of contributions to a film could be adapted (Meyer et al., 2015).

Some challenges will have to be addressed from multiple directions. For example, metadata may be improved to improve findability, but more innovative approaches of finding data that do not only rely on manually created metadata are available. Similarly, while data quality and medical knowledge is crucial in dementia research, additional insights might be created from data outside of the medical realm – analysed by those who specialise in data analytics of any kind, and followed up by those with rich knowledge in dementia research.

Similarly, by engaging the public and involving them to a greater extent, data from novel sources may be made available for dementia research. This engagement may happen through mechanisms such as letting individuals "donate their data" from existing sources (for example, loyalty card data), but also to engage them via apps or tracking devices – potentially being promoted by medical staff. Engaging the public in this way may not only help to raise awareness, but also encourage further participation in research projects, increase charitable spending, and integrate care by including those with dementia and their caregivers. At the same time, offering the opportunity to be involved may help to increase transparency about the uses to which the data are being put.

In some cases, competing interests may need to be balanced with each other. For instance, while privacy is and will remain an important consideration for health information, there are ways to balance privacy with the openness required to combine data across sources and across time. These range from having data held securely by trusted organisations and made available via secure mechanisms on the researchers' side, to allowing people to voluntarily contribute data and information about themselves more easily via simplified consent processes and innovative platforms such as PatientsLikeMe, for example, which allows individuals to share conditions, symptoms and treatments for a variety of medical conditions such as dementia.

Likewise, there may be tradeoffs between increasing speed of data access and discovery against the time required for thorough data cleaning. In some cases, it may be that faster (and often cheaper) yet somewhat messier data can be used to find new patterns, which can then later be tested more thoroughly using carefully cleaned (and thus expensive) data. In this respect, we again highlight the importance of building in quality from the beginning, and creating a basic documentation that can be enhanced by those using and extending the data.

There are some no-regret-decisions for enabling data sharing. For instance, establishing core data standards with robustness yet flexibility allow researchers not only to make their data more sharable from the start, but also make designing their databases faster and easier than if they are regularly re-inventing data structures from scratch. These standards, as well as contributed data, can be documented more clearly in low-cost, low-barrier resources such as wikis, taking effort that would otherwise be internal to a project and sharing it so that others can not only understand their methods, but also build upon them for future work.

Transitioning to a new research paradigm

Dealing with big data is not entirely new, but requires some new operational procedures with larger and more complex data. It is important not to understate the many years of work that have already gone into assembling high-quality and important datasets that support dementia research, some of which we have discussed in this report. Combining these data with new data that are currently being collected, and especially with *new types* of data, will require research teams to budget for additional personnel who have expertise in working with big data in terms of its size, but also its structure and novel analytical approaches.

As data are combined from different research teams, from different institutions and nations, funded by a variety of organisations, and even combined with commercial data or other forms of data exhaust which were not originally intended for medical research, new access models will have to be developed that make data widely available while protecting privacy as well as the personal, professional, and business interests of the data originator. Organisations such as UK Biobank are already making great strides in this area, but scaling similar efforts up beyond the national level will require further effort. Establishing flexible standards for data and associated metadata will be important, as will ways of not only ensuring quality but also documenting the quality assurance procedures. Trust both in the data and in the process are essential if scientists are going to adopt big data approaches more widely.

Research teams will increasingly need to clarify data standards and practices at the beginning of projects, with an eye toward allowing new uses and combinations to emerge as the science develops. This can involve setting up databases designed to be sharable from the start, adhering to standards that are agreed to be examples of best practice, and avoiding over-reliance on institutional memory (frequently contained in the heads of a few individuals) instead of accessible documentation.

Big data requires thinking outside of the box. It requires imagination to consider what other data sources to use, for example linking in routine data that has low cost and high longitudinal availability, and understanding how to better mine data with “big data analytics”. There are innumerable sources of data being generated in the world, ranging from mobile phone data, to customer data, to tracking data, to government data, and at least some of these have potential for understanding the behaviour and environment of dementia patients not only after diagnosis but potentially even retrospectively in the years leading up to diagnosis.

These data are being exploited for many purposes, but rarely for their potential contributions to medical research. In many cases, the value of the data to the original owner would not be diminished, and could potentially even be enhanced, by using it as part of our understanding of specific medical conditions and health more generally. Making this happen requires bringing together the right stakeholders and helping them communicate their domains to each other. People with bridging skills that enable them to speak the language of different stakeholders will be particularly valuable in such an ecosystem.

Using these new data sources requires effective communication with the individuals involved, with a combination of better communication of the risks involved even with anonymised data, but also communicating the purpose of exploiting new data sources and enabling individuals to “donate their data”. Currently it may in fact be difficult for individuals to make their data available to the wider research community even if they wanted to (Friend and Norman, 2013).

Big data offers new forms of potential involvement for individuals. Researchers have largely moved on from treating research participants as passive "subjects". Going forward, we can take this even further by actively involving people in contributing to research such as with ideas like "donate your data", as sources of ideas in consumer-led research, and as citizen scientists. People have a strongly vested interest in their health and the health of their loved ones, and empowering them to be active contributors to science is a way to alleviate the helplessness that many may feel, while also improving the future for themselves, their families, and others who will be touched by dementia.

This involvement is not limited to the confines of the healthcare setting, but transcends boundaries into everyday life and between cure and care. This relates to using data more effectively from "bench to bedside" and the other way around in a feedback loop that can establish a virtuous cycle: care data more directly being used as part of the toolkit for basic science, and the findings of scientific research more rapidly being translated into prevention, detection, treatment, while also linking back to the care realm.

Big data creates new ethical questions that will have to addressed. To what extent can or should research participants expect to "get" something from their participation in scientific research (whether in the form of compensation, or free health checks, or other direct benefits as opposed to the more nebulous and long-term benefit of having contributed to science)? What are the expectations when the contribution is not direct, such as donating blood samples, but indirect, such as agreeing to having one's mobile phone activity available to scientists? What are the ethical consequences of using data collected for one purpose but later put to entirely unanticipated uses? How does the scientific need to use longitudinal data to understand diseases such as dementia that develop over decades square with the calls for the "right to be forgotten" or conversely contribute to a social and political climate where nothing is ever forgotten, for good or for ill?

These are among the many questions about big data and health that philosophers, ethicists, and legal experts will be grappling with for many years to come. It is important to have this dialogue and to keep it going in a changing environment – both between the experts and with the wider public, as society as a whole is a key factor in enabling new research and participating in it:

> *In going forward with big data in healthcare research, we need to keep everyone involved who is affected. A parallel discussion needs to begin for better public understanding of how best healthcare is delivered – whether it is something that is done just in hospitals and clinics or something that also is part of the way we structure our society.*
>
> – **Paul Matthews**

5.2 Policy levers to support action

While researchers working on advancing dementia research are at the core, public policy has a crucial role to ensure that framework conditions to promote data sharing and collaboration are in place, and take on **leadership in informing and driving the public debate on responsibly using and sharing data**. Defining what is acceptable use of biomedical data goes beyond issues of consent, the law and technical questions, but is fundamentally driven by what is seen as acceptable and appropriate (Nuffield Council on Bioethics, 2015).

In addition, governments can promote action for using and sharing data directly and indirectly. The specific areas cut across all of the seven challenges we identified to a certain extent, and are outlined below.

Funding needs to support dementia and infrastructures for sharing data. While it may seem patently obvious that all scientific research requires funding, it is still worth reiterating the need to fund *data sharing and infrastructures for sharing data* as part of successful research proposals. Data sharing must not be left as an unfunded afterthought to research, but should be an integral part of research practice, and funders must ensure that researchers budget for this and must be willing to pay for this as part of the essential costs of good science.

Government can lead this effort and play a role in this both by funding research and infrastructures directly, but also by setting aspirations and stimulating interest from other players (see next point) despite an unfavourable risk/reward balance and the current market failure situation:

> *I think there are a number of ways in which government can galvanise interest and then let other agencies push – whether those are state funding bodies, private companies or charities.*
>
> – **Nick Seddon**

Policy should stimulate collaboration between public and private actors. Further collaborations between these stakeholders can be in the form of public-private partnerships, through in-kind donations of data and expertise, through government tax incentives for contributions to science, and any number of other innovative mechanisms that help to make data for dementia research available. Potential for collaboration will come from existing players in the medical research realm (such as pharmaceutical companies), but also from new players (such as supermarket chains or mobile phone companies) or entirely new organisations such as start-ups which bring a can-do attitude and new skills to the field.

Funders can encourage these forms of collaboration by highlighting opportunities for new kinds of partners to become involved in science, and enabling these partnerships to develop. With this, funders are also in a position to support findings being translated throughout the research cycle into prevention strategies, diagnosis, treatment and care:

> *[It is mostly] about how we share the knowledge that emerges from the data. Because at the end of the day, what you're looking for is meaning. And it's that meaning that's going to give us the insight that's going to lead to the next well-funded experiment that's going to, hopefully, lead to the next well-funded drug target and ultimately to the therapy.*
>
> – **John Williams**

It is worth highlighting again that through funding data sharing, institutions like the Critical Path Institute with the CAMD initiative have already pushed forward regulatory approvals in the field of dementia research. Data sharing does not only benefit the research as such, but also benefits the whole field by supporting those activities further downstream on the way to treatment and prevention strategies.

There needs to be investment in future bioinformatics talent and in increased collaboration with data experts outside dementia research. Some of this will be in the form of encouraging universities to offer opportunities and funding for training and education in data science and related areas, but it can also involve creating multi-disciplinary centres of excellence, and a focus on funding interdisciplinary, multi-institution and multi-country research. In general, it is important to develop the potential of the individuals and the *networks that they are part of* in ways that put data sharing at the centre of the scientific culture.

On an international level, there is the need for guidelines on consent and IRB/ERC approval processes. There is currently too much uncertainty about whether consent for medical research allows data to be shared beyond institutions, collaborations, or nations, which may unnecessarily hold back research and extracting value from data on a larger scale.

Standardised consent forms, developed by researchers who work in the field, as well as clear, coherent and consistent principles for IRBs/ERCs will help scientists design research protocols that protect research participants and institutions from harm and loss while enabling data to be shared more widely. One source of inspiration might be the Creative Commons community, which has successfully simplified the requirements for sharing copyrightable information in a way that is also very clear, understandable, and legally robust.

For these IRB/ERC guidelines, it is also important that IRBs/ERCs jump on the data sharing bandwagon that funders have already initiated by mandating data sharing or providing incentives for it. There should be no hurdles to responsible data sharing except where individual research participants or patient rights are in fact violated, so that incongruences between funders and IRBs/ERCs cease to exist:

> *For ethical applications, we need to say we are protective for our data. But to get grants we need to say that the data are very open which makes it very confusing for researchers today.*
>
> – **Ingmar Skoog**

On an international level, a stable and beneficial legal framework must be ensured. We need policies that protect citizens against any *undue* exploitation of their data that they would not want, but must balance this right to privacy with the right of millions of people with dementia and their caregivers – and millions more to come – to find a cure.

With respect to research data, legislation also needs to account for the growing global research communities. As outlined earlier, diseases do not know borders, countries need to collaborate in terms of funding and making best use of human and data resources, and legislation must support this changing research landscape. This is especially true for Europe, with integration being one of the key aims of policymakers, but where data protection policies rather lead to new hurdles:

> *In Sweden, politicians want it to be easier for researchers [...], but at the European level there are movements in a different direction to make it more difficult to get other datasets [...] because of the confidentiality problem.*
>
> – **Ingmar Skoog**

In addition, the legal framework needs to be stable: If beyond the situation of a market failure there is a legal risk for research data not being useful, funding and investment in the sector will be an even more arduous task than it already is. Not to mention as well that research participants and the public more generally rely on some stability when making choices about contributing to science, and might no longer do so if they sense this is not the case.

All of these policy levers to support using and sharing big data are not specific to dementia only. However, dementia is an area of research in which the benefits of using big data and sharing it more widely are especially apparent. This is due to the highly complex nature of influence factors from across medical research areas, the fact that different kinds of data from within and outside the medical realm may be relevant, and the challenge of engaging all stakeholders in the health system despite the market failure situation.

Not least of these is the potential with respect to personalised medicine, as there may not be just one answer to the question of what causes dementia and how we can stop it. Not only is dementia itself an umbrella term for different diseases, but many treatments also work differently with different people. Using a multitude of data-driven approaches may help to shed more light on partial answers being combined into a larger picture.

All of this is coupled with the high importance of dementia due to its growing economic burden in an ageing society, and the devastating personal burden it creates for affected individuals and their families. By no means will better data sharing be a sufficient guarantee for having a cure – or a way to slow the disease's onset – by 2025. However, we need to exploit all potential avenues for making progress, and support research with policy initiatives that create the best environment for making discoveries. Most importantly, there is an urgent need for action:

> *If we don't do a few things right, recognise that we need to do personalised medicine, achieve consensus on data standards which still isn't happening, and get agreement to share the data for trials [...] we're going to be in 5 years from now looking back and saying the same things we're saying now.*
>
> – **Diane Stephenson**

Big data – and sharing it efficiently among the different actors – has a critical role to play in advancing dementia research. Used efficiently in an environment that is beneficial to research, it is one of the main bearers of hope to advance the current understanding of dementia, and to develop strategies for healthy ageing and fighting neurodegenerative diseases in the future.

Appendices

List of interviewees

All interviewees represent their own personal views and not necessarily the official standpoint of any organisation.

ADNI (Alzheimer's Disease Neuroimaging Initiative)

Prof. Michael W. Weiner, MD	ADNI Principal Investigator Professor of Radiology and Biomedical Engineering, Medicine, Psychiatry, and Neurology University of California, San Francisco, CA, United States
Robert C. Green, MD, MPH	Chair, ADNI Data and Publications Committee Associate Professor of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States
Prof. Arthur W. Toga, PhD	ADNI Informatics Core Leader Professor of Neurology, Laboratory of Neuro Imaging, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States
Jean-François Mangin, PhD	Director of CATI platform Service Hospitalier Frédéric Joliot, Orsay; IN-SERM ERM, Orsay; Institut Fédératif de Recherche, Paris, France
Diane Stephenson, PhD	User of ADNI data Executive Director for Coalition Against Major Diseases (CAMD), Critical Path Institute, Tucson, AZ, United States
Klaus Romero, MD, MS, FCP	User of ADNI data Director of Clinical Pharmacology, Critical Path Institute, Tucson, AZ, United States
Mette Peters, PhD	User of ADNI data Principal Scientist, Sage Bionetworks, Seattle, WA, United States
Hiroshi Matsuda, PhD	J-ADNI data manager Director General, Integrative Brain Imaging Center (IBIC), National Center of Neurology and Psychiatry (NCNP), Tokyo, Japan
Anonymous	ADNI data manager One of the ADNI country initiatives

AddNeuroMed

Prof. Simon Lovestone, PhD, MRCPsych	AddNeuroMed Principal Investigator Professor of Translational Neuroscience, Department of Psychiatry, University of Oxford, United Kingdom
Prof. Patrizia Mecocci, MD, PhD	AddNeuroMed site leader Professor of Gerontology and Geriatrics & Director of the Center of Brain Aging and Dementias in the Elderly, University of Perugia, Italy
Richard Dobson, PhD	AddNeuroMed data manager & user of ADNI data Senior Lecturer, King's College London & Head of Bioinformatics at the NIHR Biomedical Research Centre for Mental Health, London, United Kingdom
Prof. Hilkka Soininen, MD, PhD	Data provider to AddNeuroMed Professor in Neurology, Institute of Clinical Medicine & Neurology, University of Eastern Finland, Kuopio, Finland
Eric Westman, PhD	AddNeuroMed & Swedish Brain Power data user Associate Professor, Department of Neurobiology, Care Sciences & Society, Karolinska Institute, Stockholm, Sweden

UK Biobank

Prof. Sir Rory Collins, FMedSci FRCP	UK Biobank Principal Investigator Professor of Medicine and Epidemiology, Nuffield Department of Population Health, University of Oxford, United Kingdom
Prof. Cathie Sudlow, DPhil, FCRP	UK Biobank Chief Scientist & Dementias Platform UK Steering Committee member Professor of Neurology and Clinical Epidemiology, University of Edinburgh, United Kingdom
Prof. Roger Brownsword	Chairman, UK Biobank Ethics & Governance Council Professor of Law, King's College London and Bournemouth University, United Kingdom
Prof. Paul Matthews, OBE, MD, DPhil, FRCP, FMedSc	UK Biobank Steering Committee & Dementias Platform UK Executive Group Edmond and Lily Safra Chair and Head of Division of Brain Sciences, Department of Medicine, Imperial College, London, United Kingdom

John Gallacher, PhD

Director of Dementia Platform UK
Lecturer, Cardiff University School of Medicine, University of Cardiff, United Kingdom

Tim Sprosen, DPhil

UK Biobank Steering Committee member & former Chief Scientist
Associate Professor & Head of Innovation, Regulation & Knowledge Transfer, Nuffield Department of Population Health, University of Oxford, United Kingdom

Prof. Daniel Smith, MD

UK Biobank data user
Professor of Psychiatry, Institute of Health and Wellbeing, University of Glasgow, United Kingdom

Prof. Ian Hall, DM FRCP

UK Biobank data user
Executive Dean & Professor of Molecular Medicine, Faculty of Medicine and Health Sciences, University of Nottingham, United Kingdom

Swedish Brain Power (SBP) studies

Prof. Bengt Winblad, MD, PhD

Head of Swedish Brain Power network
Professor of Geriatrics, Department of Neurobiology, Care Sciences & Society, Karolinska Institute, Stockholm, Sweden

Prof. Laura Fratiglioni, MD, PhD

Principal Investigator of SNAC-K/Kungsholmen
Professor of Medical Epidemiology, Department of Neurobiology, Care Sciences & Society, Karolinska Institute, Stockholm, Sweden

Prof. Ingmar Skoog, MD, PhD

Principal Investigator of SBP studies in Gothenburg
Professor in Psychiatry, Department of Psychiatry and Neurochemistry & Director of the Centre for Health and Ageing (AGECAP), University of Gothenburg, Sweden

Prof. Miia Kivipelto, MD, PhD

Principal Investigator of several SBP studies
Professor in Clinical Geriatric Epidemiology & Director for Research, Development and Education, Aging Research Center and Alzheimer's Disease Research Center, Karolinska Institute; Department of Geriatrics, Karolinska University Hospital, Stockholm, Sweden

Prof. Linda Hassing, PhD

Swedish Brain Power data user
Professor in Psychology, Department of Psychology, University of Gothenburg, Sweden

Other experts from government and industry

Nick Seddon	Health & Social Care Advisor to PM David Cameron, The Prime Minister's Office, 10 Downing Street, London, United Kingdom
John Williams, PhD	Head of Science Strategy, Performance and Impact, Wellcome Trust, London, United Kingdom
Yoshiaki Tojo	Director, Commerce & Information Policy Bureau, METI Program Advisor for IT Integration & Innovation Platform, NEDO, Japan & Chief Executive Adviser, JETRO, San Francisco, CA, United States
Nicolaus Henke, MPA	Director and Global Head of Healthcare Systems & Services, McKinsey & Company, London, United Kingdom
John Drew	Principal and Head of McKinsey Hospital Institute, McKinsey & Company, London, United Kingdom
Stefan Larsson, MD, PhD	Senior Partner and Managing Director, Global Sector Leader for Health Care Payers and Providers, The Boston Consulting Group, Stockholm, Sweden
Clive Humby	Founder of DunnHumby (creator of Tesco Clubcard loyalty programme) and Chief Data Scientist of Starcount, London, United Kingdom

Other academic experts

Prof. Parminder Reina, PhD	Principal Investigator, Canadian Longitudinal Study on Aging (CLSA), Professor at Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, ON, Canada
Prof. Derek Hill, PhD	CEO of IXICO plc & Professor of Medical Imaging Science, University College London (UCL), London, United Kingdom
Rupert McShane, MD FRCPsych	Dementia Clinical Network Lead, Department of Psychiatry, University of Oxford, United Kingdom

Methodology

Case study selection

The case studies for this report have been selected purposively, with the aim of achieving variety in geographic coverage, a dementia-specific or cross-conditional focus, size, maturity, openness by design and linkage to other routine data. More details on case study selection approach are available in chapter 2.

Interviewee selection

We conducted 37 interviews with scientific and technical experts involved with each of the four case studies, as well as general experts from academia, industry and government. The interviews were conducted face-to-face or over the telephone/Skype between July 2014 and January 2015.

Interviewees were selected using purposive sampling methods, since it was most important in this study to reach decision-makers and key implementers and users rather than a random cross-section of stakeholders.

To achieve a diversity of views, we included both individuals internal and external to the data sharing initiative. On the internal side, this includes individuals from all levels: From the Principal Investigator or leaders of specific working groups to data managers working with data on a day to day basis. On the external side, we spoke both to users of the data – where available, from different backgrounds including academia and industry – and data providers from clinical settings.

This study received ethics approval from Oxford University Central University's Research Ethics Council (Reference: OII C1A 14–025). Informed consent was obtained from all interviewees orally or in writing. Interviewees were given the option of choosing between four levels of confidentiality:

- **No restrictions:** The individual and the organisation are identified and quoted in reports about this research
- **Minimum:** The individual and the organisation are identified, but some parts of the interview were off the record at request
- **Medium:** No identification of the individual, but of the organisation
- **Maximum:** Complete anonymity for the individual and the organisation

Interview guide

The specific questions and topics addressed during the interviews differed in order to accommodate differences between those internal and external to an organisation, different levels of medical or technical expertise, and between interviewees for specific case studies and general experts. However, all interviews followed the following common framework:

Stage	Topics covered / example questions
Beginning	Introduction, informed consent and questions Interviewee involvement *“Can you please give me an overview of how you are involved with [initiative/dementia research]?”*
Specific initiative	Current benefits of data sharing and specific initiative *“From your point of view, what is the biggest current benefit of [initiative] to the research community?”* Data governance dimensions (availability, accessibility, interoperability, traceability, quality, privacy/security, ownership), e.g. *“What are the current barriers are to a researcher or research group using data from [initiative]?” – “How well can data of [initiative] be linked to other routine data collected in the health system?” – “How would you describe the approach to individual privacy and security of [initiative]? – “How must data use be attributed in potential publications? Are data providers co-authors?”* Future benefits of data sharing and specific initiative *“What do you think the biggest potential benefit of [initiative] is, and what would have to happen for that benefit to be realised?”*
Dementia research in general	Wider context in dementia research *“In an ideal world, what will be most important going forward to advance dementia research?”* Hurdles to data sharing *“In terms of data sharing, what do you see as the best way forward to accelerate dementia research?”* Prioritisation of aspects *“Out of the different aspects we discussed earlier, which is most problematic when it comes to data sharing for dementia research?”* New forms of big data in dementia research *“What is your view on using other routinely collected big data for dementia research?”*
Ending	Other important aspects *“Is there anything else that you think may be relevant that I have not asked about?”* Next steps and follow-up questions

Analytical strategy

All interviews were recorded, transcribed and analysed using standard methods in qualitative research. While the data governance framework (OECD, 2014) in chapter 3 was used in the interview guide and served as an analytical framework, the structural challenges introduced in chapter 4 emerged from the interviews.

Additional information was obtained from the initiatives’ websites and previous research, as well as discussions at OECD workshops in Toronto (September 2014) and Paris (December 2014), where interim findings of this research were presented.

About the authors

Ulrike Deetjen is a doctoral researcher at the Oxford Internet Institute, University of Oxford. Her thesis supervised by Prof. Rebecca Eynon and Prof. John Powell looks at Internet use and health outcomes, using an innovative approach of combining secondary medical datasets with survey and census data. She spent several years working as a business technology consultant at McKinsey & Company, advising healthcare organisations and pharmaceutical companies around the globe primarily on questions of business intelligence and data management. She can be reached at ulrike.deetjen@oii.ox.ac.uk (ulrike.rauer@oii.ox.ac.uk until June 2015)

Prof. Eric T. Meyer, PhD is Senior Research Fellow and Associate Professor as well as Director of Graduate Studies at the Oxford Internet Institute, University of Oxford, where he has been on the faculty since 2007. He has research interests in collaboration using digital tools in the sciences and arts, social informatics, information science, and social aspects of science and technology. He has published over 30 journal articles and over a hundred other publications (including reports, book chapters, and conference papers) as well as the book Knowledge Machines: Digital Transformations of the Sciences and Humanities (with Ralph Schroeder, MIT Press 2015). Prior to coming to Oxford, Eric T. Meyer spent ten years as the national research coordinator for a collaboration of 13 US universities on the genetics of bipolar disorder. He can be reached at eric.meyer@oii.ox.ac.uk

Prof. Ralph Schroeder, PhD is currently Professor and MSc Programme Director at the Oxford Internet Institute, University of Oxford, where he has been on the faculty since 2004. He was formerly Professor in the School of Technology Management and Economics at Chalmers University in Gothenburg (Sweden). His publications include Rethinking Science, Technology and Social Change (Stanford University Press, 2007) and Being There Together: Social Interaction in Virtual Environments (Oxford University Press, 2010), as well as over 100 articles. Ralph Schroeder is also the author of An Age of Limits: Social Theory for the 21st Century (Palgrave Macmillan 2013) and, with Eric T. Meyer, of Knowledge Machines: Digital Transformations of the Sciences and Humanities (MIT Press 2015). He can be reached at ralph.schroeder@oii.ox.ac.uk

---

The Oxford Internet Institute (OII) is an interdisciplinary department of the University of Oxford and a world-renowned academic centre for the study of the societal implications of the Internet founded in 2001. The OII has undertaken extensive qualitative and quantitative empirical research on how new technologies and platforms are adopted and shaped, conducted comparative analyses of innovation systems, policies and strategies across countries and organisations, and provided strategic advice to policymakers to help with planning and forecasting.

List of OECD & International Advisory Group

OECD

Elettra Ronchi, PhD	Senior Policy Analyst, OECD Digital Economy and Policy Division, Paris, France
Christian Reimsbach-Kounatze	Economist, OECD Digital Economy & Policy Division, Paris, France

OECD International Advisory Group

Rob Buckle, PhD (Chair)	Director of Science Programmes & Director, UK Regenerative Medicine Platform, MRC, United Kingdom
Prof. Yves Joanette, PhD	Director, CIHR Institute of Aging & CIHR International Collaborative Research Strategy for Alzheimer's Disease, University of Montreal, Canada
Prof. Philippe Amouyel, MD, PhD	Director General Fondation Nationale de coopération scientifique Maladie d'Alzheimer, France
Prof. Donald Stuss, PhD, CPsych, ABPP-CN, FRSC, FCAHS	President and Scientific Director, Ontario Brain Institute & Professor in Psychology, Neurology and Rehabilitation Science, University of Toronto, Canada
Giovanni Frisoni, MD	Deputy Scientific Director, IRCCS Fatebenefratelli, Brescia, Italy
Yoshiaki Tojo	Director, Commerce & Information Policy Bureau, METI Program Advisor for IT Integration & Innovation Platform, NEDO, Japan & Chief Executive Adviser, JETRO, San Francisco, CA, United States
Prof. Martin Rossor, MD FRCP	Professor of Clinical Neurology, Institute of Neurology, University College London, United Kingdom
Neil Buckholtz, PhD	Director, Division of Neuroscience, National Institute of Ageing, United States
Prof. Miia Kivipelto, MD, PhD	Professor in Clinical Geriatric Epidemiology & Director for Research, Development and Education, Aging Research Center and Alzheimer's Disease Research Center, Karolinska Institute; Department of Geriatrics, Karolinska University Hospital, Stockholm, Sweden
Richard Johnson	Chairman, Business Industry Advisory Committee, Science & Technology Committee, OECD

Bibliography

Alzheimer's Association (2014). Prevention and Risk of Alzheimer's and Dementia. Available online at http://www.alz.org/research/science/alzheimers_prevention_and_risk.asp [retrieved on 29/11/2014].

Alzheimer's Disease International (2010). The Global Economic Impact of Dementia. World Alzheimer Report 2010, London, United Kingdom.

Alzheimer's Disease International (2014). Dementia and Risk Reduction: An Analysis of Protective and Modifiable Factors. World Alzheimer Report 2014, London, United Kingdom.

Axelsson, A.-S. and R. Schroeder (2009). Making it Open and Keeping it Safe: e-Enabled Data-Sharing in Sweden. *Acta Sociologica 52*(3), 213–226.

Bollier, D. and C. M. Firestone (2010). *The Promise and Peril of Big Data*. Washington, DC, USA: Aspen Institute, Communications and Society Program.

Borgman, C. L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology 63*(6), 1059–1078.

Celi, L. A., A. Ippolito, R. A. Montgomery, C. Moses, and D. J. Stone (2014). Crowdsourcing Knowledge Discovery and Innovations in Medicine. *Journal of Medical Internet Research 16*(9), e216.

Farr Institute (2015). About the Farr Institute of Health Informatics Research. Available online at http://www.farrinstitute.org/centre/London/6_About.html [retrieved on 24/01/2015].

Friend, S. H. and T. C. Norman (2013). Metcalfe's law and the biology information commons. *Nature Biotechnology 31*(4), 297–303.

G8 UK: Global Action Against Dementia (2013). Dementia Summit Declaration. Available online at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/265869/2901668_G8_DementiaSummitDeclaration_acc.pdf [retrieved on 24/01/2015].

Gardner, D., A. W. Toga, G. A. Ascoli, J. T. Beatty, J. F. Brinkley, A. M. Dale, P. T. Fox, E. P. Gardner, J. S. George, N. Goddard, et al. (2003). Towards Effective and Rewarding Data Sharing. *Neuroinformatics 1*(3), 289–295.

Greely, H. T. (2007). The Uneasy Ethical and Legal Underpinnings of Large-scale Genomic Biobanks. *Annual Review of Genomics and Human Genetics 8*, 343–364.

Groves, P., B. Kayyali, D. Knott, and S. Van Kuiken (2013). The "big data" revolution in healthcare. *McKinsey Quarterly*.

Gutwirth, S., R. Leenes, P. De Hert, and Y. Poullet (2013). *European Data Protection: Coming of Age*. Springer.

Howe, D., M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, et al. (2008). Big Data: The Future of Biocuration. *Nature 455*(7209), 47–50.

Huijboom, N. and T. Van den Broek (2011). Open Data: an International Comparison of Strategies. *European Journal of ePractice 12*(1), 1–13.

Katzman, R. (1976). The Prevalence and Malignancy of Alzheimer Disease: a Major Killer. *Archives of Neurology 33*(4), 217–218.

Lazer, D. M., R. Kennedy, G. King, and A. Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis.

Leetaru, K. (2011). Culturomics 2.0: Forecasting Large-scale Human Behavior using Global News Media Tone in Time and Space. *First Monday 16*(9).

Lintott, C. J., K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, et al. (2008). Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society 389*(3), 1179–1189.

Ludman, E. J., S. M. Fullerton, L. Spangler, S. B. Trinidad, M. M. Fujii, G. P. Jarvik, E. B. Larson, and W. Burke (2010). Glad you Asked: Participants' Opinions of Re-consent for dbGap Data Submission. *Journal of Empirical Research on Human Research Ethics 5*(3), 9.

Luengo-Fernandez, R., J. Leal, and A. Gray (2010). The Economic Burden of Dementia and Associated Research Funding in the United Kingdom. Report produced by the Health Economics Research Centre, University of Oxford for the Alzheimer's Research Trust.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers (2011). Big data: The next frontier for innovation, competition, and productivity.

Mayer-Schönberger, V. and K. Cukier (2013). *Big data: A Revolution that will Transform how we Live, Work, and Think*. New York, NY: Houghton Mifflin Harcourt.

Mestyán, M., T. Yasseri, and J. Kertész (2013). Early Prediction of Movie Box Office Success based on Wikipedia Activity Big Data. *PloS one 8*(8), e71226.

Meyer, E. T., J. Cowls, and R. Schroeder (2015). Roll the credits: Valuing differential contributions to knowledge in the big data era. Paper presented at FORCE2015 conference, Oxford, UK.

Meyer, E. T. and R. Schroeder (2015). *Knowledge machines: digital transformations of the sciences and humanities*. Cambridge, MA: MIT Press.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. (2011). Quantitative Analysis of Culture using Millions of Digitized Books. *science 331*(6014), 176–182.

National Institute of Health (2003). Final NIH Statement on Sharing Research Data. Available online at `http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html` [retrieved on 17/01/2015].

NeuGRID4U (2014). What is NeuGRID? Available online at `https://neugrid4you.eu/background` [retrieved on 30/11/2014].

Nuffield Council on Bioethics (2015). The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues. Report released in February 2015.

Organisation for Economic Co-operation and Development, OECD (2014). Unleashing the Power of Big Data for Alzheimer's Disease and Dementia Research: Main Points from the OECD Expert Consultation on Unlocking Global Collaboration to Accelerate Innovation for Alzheimer's Disease and Dementia. OECD Digital Economy Papers, No. 233, OECD Publishing, 20–21 June 2013, Harris Manchester College, Oxford, United Kingdom.

Piwowar, H. A., M. J. Becich, H. Bilofsky, R. S. Crowley, et al. (2008). Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. *PLoS medicine 5*(9), e183.

Piwowar, H. A., R. S. Day, and D. B. Fridsma (2007). Sharing Detailed Research Data is Associated with Increased Citation Rate. *PloS one 2*(3), e308.

Research Data Alliance (2014). About the Research Data Alliance. Available online at `https://rd-alliance.org/about.html` [retrieved on 05/12/2014].

Ritchie, K. and D. Kildea (1995). Is Senile Dementia "Age-related" or "Ageing-related"? Evidence from Meta-analysis of Dementia Prevalence in the Oldest Old. *The Lancet 346*(8980), 931–934.

Rohlfing, T. and J.-B. Poline (2012). Why Shared Data should not be Acknowledged on the Author Byline. *Neuroimage 59*(4), 4189–4195.

Schroeder, R. (2014). Big Data and the Brave New World of Social Media Research. *Big Data & Society 1*(2), 2053951714563194.

Schroeder, R. and M. Den Besten (2008). Literary Sleuths Online: e-Research Collaboration on the Pynchon Wiki. *Information, Community and Society 11*(2), 167–187.

Solomon, A., F. Mangialasche, E. Richard, S. Andrieu, D. Bennett, M. Breteler, L. Fratiglioni, B. Hooshmand, A. Khachaturian, L. Schneider, et al. (2014). Advances in the Prevention of Alzheimer's Disease and Dementia. *Journal of Internal Medicine 275*(3), 229–250.

Steinsbekk, K. S., B. K. Myskja, and B. Solberg (2013). Broad Consent versus Dynamic Consent in Biobank Research: Is Passive Participation an Ethical Problem? *European Journal of Human Genetics 21*(9), 897–902.

Su, A. I., B. M. Good, and A. J. van Wijnen (2013). Gene Wiki Reviews: Marrying Crowdsourcing with Traditional Peer Review. *Gene 531*(2), 125.

Swede, H., C. L. Stone, and A. R. Norwood (2007). National Population-based Biobanks for Genetic Research. *Genetics in Medicine 9*(3), 141–149.

Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame (2011). Data Sharing by Scientists: Practices and Perceptions. *PloS one 6*(6), e21101.

UK Department of Health (2014). G8 Global Dementia Summit: Global Action Against Dementia, 11 December 2013. Available online at `https://www.gov.uk/government/publications/g8-dementia-summit-global-action-against-dementia/g8-dementia-summit-global-action-against-dementia-11-december-2013` [retrieved on 29/11/2014].

World Health Organization and Alzheimer's Disease International (2012). Dementia: a Public Health Priority. Available online at `http://apps.who.int/iris/bitstream/10665/75263/1/9789241564458_eng.pdf` [retrieved on 29/11/2014].

Wu, L. and E. Brynjolfsson (2013). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In *Economics of Digitization*. University of Chicago Press.

Big Data for Advancing Dementia Research

Oxford Internet Institute, University of Oxford
bigdatadementia.oii.ox.ac.uk
March 2015