Neuronet

IMI2 821513 – NEURONET

Efficiently Networking European Neurodegeneration Research

WP3 – Tools & Services

D3.7 Final version on guidance tools on data/sample sharing and use

Lead contributor	Lennert Steukers	Janssen Pharmaceutica
Other contributors	Andrew Owens	King's College London
	Andrew Peter McCarthy	Eli Lilly
	Angela Bradshaw	Alzheimer Europe
	Anthony Brookes	University of Leicester
	Carlos Díaz	Synapse
	Chris Bintener	Alzheimer Europe
	Cindy Birck	Alzheimer Europe
	Emma Dodd	Roche
	Francoise Le Vacon	Biofortis Mérieux NurtiSciences
	Jean Georges	Alzheimer Europe
	Judi Syson	University of Edinburgh
	Kirsten Emmert	University of Kiel
	Manuela Rinaldi	Janssen Pharmaceutica
	Martin Hofmann-Apitius	Fraunhofer
	Niamh Connolly	Royal College of Surgeons in Ireland
	Nigel Hughes	Janssen Pharmaceutica
	Nikolay Manyakov	Janssen Pharmaceutica
	Pieter Jelle Visser	Vumc & Maastricht University
	Rodrigo Barnes	Aridhia
	Serge Van der Geyten	Janssen Pharmaceutica
	Walter Maetzler	University of Kiel

Document History

Version	Date	Description
V0.1	28/04/2021	First draft
V0.2	05/05/2021	Second draft
V0.3	16/06/2021	Third draft
V1.0	21/01/2022	Final version

Contents

Definitions and abbreviations

Partners of the NEURONET Consortium are referred to herein according to the following codes:

1. **SYNAPSE**: Synapse Research Management Partners SL
2. **NICE**: National Institute for Health and Care Excellence
3. **AE**: Alzheimer Europe
4. **JANSSEN**: Janssen Pharmaceutica NV
5. **LILLY**: Eli Lilly and Company Limited
6. **ROCHE**: F. Hoffman – La Roche AG
7. **TAKEDA**: Takeda Development Centre Europe LTD (*terminated partner*)
8. **SARD**: Sanofi-Aventis Recherche & Développement
9. **PUK**: Parkinson's Disease Society of the United Kingdom LBG
10. **TAKEDA AG**: Takeda Pharmaceuticals International AG

AD: Alzheimer Disease

ADDI: Alzheimer Disease Data Initiative

ADNI: Alzheimer's Disease Neurodegenerative Initiative

ANM: AddNeuromed

BD4BO: Big Data for better outcomes

CA: Consortium Agreement

CDA: Confidential Disclosure Agreement

CDM: common data model

CESR: Center for Effectiveness and Safety Research

CSA: Coordination and Support Action

Consortium: The NEURONET Consortium, comprising the above-mentioned legal entities.

Consortium Agreement: Agreement concluded amongst NEURONET participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

CRO: contract research organisation

DMP: Data management plan

EEA: European Economic Area

EHDEN: European Health Data & Evidence Network

EMIF-AD: European Medical Informatics Framework- Alzheimer's Disease

EPAD: European Prevention of Alzheimer's Dementia

ETL: extract, transform and load

EU: European Union

FAIR: findable, accessible, interoperable and re-usable

GA: Grant Agreement

GAAIN: Global Alzheimer's Association Interactive Network

GDPR: General Data Protection Regulation

Grant Agreement: The agreement signed between the beneficiaries and the IMI JU for the undertaking of the NEURONET project.

IMI: Innovative Medicines Initiative

ICF: Informed Consent Form

IP: Intellectual property

ML: Machine learning

ND: Neurodegenerative Disorders

OHDSI: Observational Health Data Sciences & Informatics

OMOP: Observational Medical Outcomes Partnership

ORD: Open Research Data

PCORI: Patient Centered Outcomes Research Institute

PI: Principle Investigators

Project: The sum of all activities carried out in the framework of the Grant Agreement.

RADAR-AD: Remote Assessment of Disease and Relapse – Alzheimer's Disease

RWD: real world data

SCB: Scientific Coordination Board

SHDN: FDA's Sentinel within a shared health data network

SME: Small to medium enterprise

WG: Working group

Work plan: Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out, as specified in Annex I to the Grant Agreement.

WP: Work Package

Glossary of terms

Data custodian	A person that manages the actual data
Data sharing	**Data sharing** is the practice of making research data available to other investigators.
Data standardisation	**Data standardisation** is the critical process of bringing data into a common format that allows for collaborative research, large-scale analytics, and sharing of sophisticated tools and methodologies.
Data harmonisation	**Data harmonisation** involves transferring data from a source system, often a proprietary one, to a common data representation, such as OHDSI's OMOP CDM. This process can vary in complexity depending on how the source data is structured, how the information is coded (*or not coded*), language, volume of data, and other factors.
Data steward	A person within an organisation who is responsible for the quality of the data and the correct usage of data.
ETL	**ETL** is short for **e***xtract,* **t***ransform,* **l***oad*, three database functions that are combined into one tool to pull data out of one database and place it into another database. • **Extract** is the process of *reading data* from a database. In this stage, the data is collected, often from multiple and different types of sources. • **Transform** is the process of *converting the extracted data* from its previous form into the form it needs to be in so that it can be placed into another database. Transformation occurs by using rules or lookup tables or by combining the data with other data. • **Load** is the process of *writing the data* into the target database. The ETL process is often used in data warehousing.
FAIR principle	The **FAIR** Data Principles are a set of guiding principles in order to make data "findable", "accessible", "interoperable" and "reusable".
GDPR	**The General Data Protection Regulation (EU) 2016/679** is a regulation in EU law on data protection and privacy in the European Union (EU) and the European Economic Area (EEA). It also addresses the transfer of personal data outside the EU and EEA areas. The GDPR aims primarily to give control to individuals over their personal data and to simplify the regulatory environment for international business by unifying the regulation within

	the EU. Superseding the Data Protection Directive 95/46/EC, the regulation contains provisions and requirements related to the processing of personal data of individuals (formally called data subjects in the GDPR) who reside in the EEA, and applies to any enterprise—regardless of its location and the data subjects' citizenship or residence—that is processing the personal information of data subjects inside the EEA.
OHDSI	**Observational Health Data Sciences & Informatics** is a multi-stakeholder, interdisciplinary collaboration to bring out the value of health data through large-scale analytics. All their solutions are open-source.
OMOP CDM	**Observational Medical Outcomes Partnership.** The OMOP Common Data Model (CDM) allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format.

Abstract

There is an **urgent need** to maximise the utility of data in the neurodegeneration field. Data sharing could help to increase the understanding of the causes, treatment, prevention and care of neurodegenerative diseases.

Neuronet is a Coordination and Support Action (CSA) aiming to support and better integrate projects in the Innovative Medicines Initiative (IMI) Neurodegenerative Disorders (ND) portfolio. A Neuronet Working group (WG) 'Data sharing and reuse' was established, which consists of subject matter experts in data sharing and re-use, participating in IMI ND projects and/or Neuronet members (*Figure 1*).

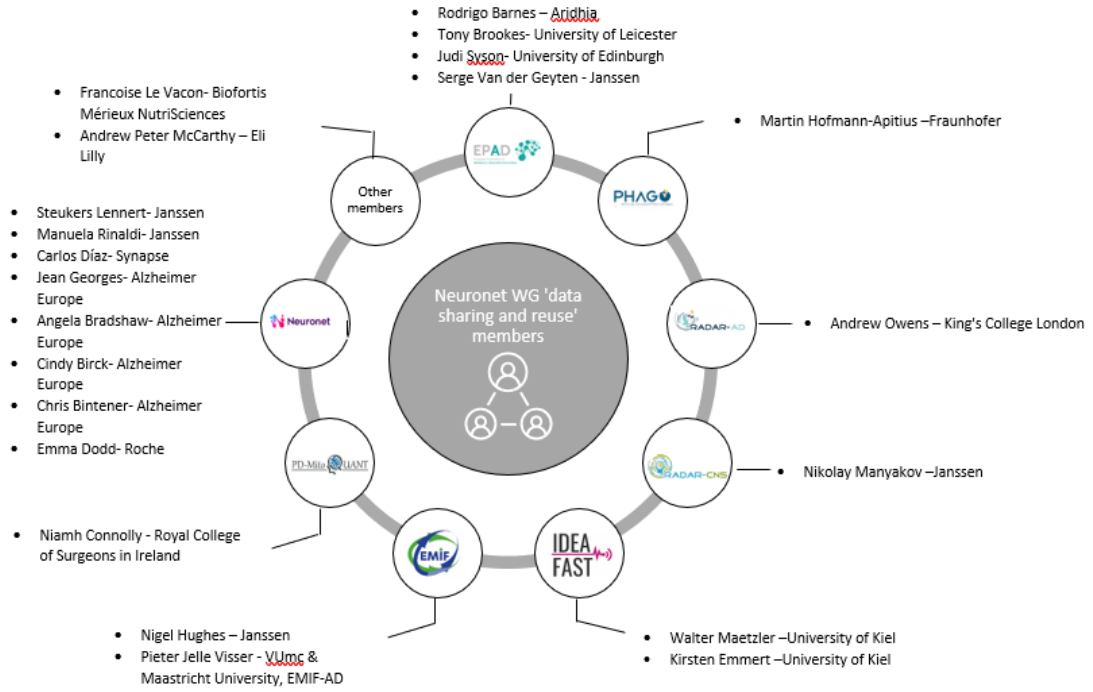

Figure 1. Neuronet WG 'Data sharing and reuse'-members

With this deliverable, Neuronet provides some learnings from the Neuronet WG 'Data sharing and reuse'. Neuronet aims to facilitate the sharing of data **amongst IMI projects**, and with **other interested research programs at European and global level.**

In the current deliverable, practical challenges of efficient data sharing are discussed. The **obtained learnings** from the various IMI projects in the ND portfolio when it comes to data sharing are being integrated into the discussion. The aim is to provide the research community with best practices in order to enable efficient sharing and access to data, while taking into consideration all relevant data sharing barriers (e.g. General Data Protection Regulation (GDPR), legal, intellectual properties (IP), ethical, societal, technical).

Executive summary

This deliverable provides an introduction of the European Union (EU) landscape regarding data sharing and its relevance and importance for EU-funded projects.

Within the Neuronet WG 'Data sharing and reuse', various organisational, legal, data protection, socio-psychological and technical challenges were identified that hamper efficient data sharing. Based on the obtained learnings from this WG, the following insights/recommendations are provided to address these challenges:

Organisational challenges

- When considering data sharing, there is a hierarchy of relationships starting from the direct relationship between a clinician or clinical researcher and a patient or study participant.
    - To overcome any organisational hurdles, it should **be clear what role each party has and that each party has the organisational** basis to commit to that role.

Legal challenges

- To overcome legal hurdles when sharing data between two IMI projects or (third) parties and sectors, the IMI2 project FAIRplus is developing **legal templates** for cross-consortia agreements to maximise the value/impact of data generated by IMI projects.

Data protection challenges

- Another challenge is the uncertainty around data protection rules for researchers. There is no **clear lawful basis for secondary data processing (i.e data reuse)** in GDPR and there is also no clear definition of **"scientific research**".
    - To overcome these challenges**, clear guidance** is needed based on sound ethical principles, to support researchers.

Psychological/social challenges

- Another common challenge is **social/motivational barriers** to data sharing.
    - To overcome social/motivational barriers, trust, trustworthiness and reliability are of paramount importance to facilitate data sharing. More specifically, researchers are only willing to share the data based on the profile and the reputation that the counterpart institution has built.
    - To overcome **motivational barriers for academic researchers**, methods to better incentivise data sharing can be a solution.

Technical challenges

- Several technical challenges also hamper data sharing. The fragmentation of the data landscape is an important problem that hampers interoperability and incentivises new research projects to come up with yet more de novo developments.
    - To increase findability and accessibility of data, several initiatives from funders and EU Commission have been initiated to make data '**Findable, Accessible, Interoperable and Reuseable'** (**FAIR).** For example, the IMI project **FAIRplus** has been launched in 2019 to tackle this problem.
    - Another challenge is the **lack of medatata,** meaning the lack of awareness of which data is available in the research field. To overcome this challenge, **several cataloguing initiatives** have been developed within IMI projects (e.g. EMIF

Catalogue, ROADMAP Data Cube, ELIXIR-LU/eTRIKS Data Catalogue). However, it is important that research data sets should not be analyzed without taking into account some companion data to avoid misinterpretation of what they mean – leading to potential errors in derived results.

- There is also is a need for data producers to annotate, document and provide meta-data that are as informative, efficient and actionable as possible. To overcome this problem, **human input** into such meta-data provision remains key.
- Another issue is that the legal norms (GDPR) for pseudonymisation, anynonymisation and consent when sharing data, remain open and offer limited practical guidance to researchers. To overcome this problem, **ethical guidance is needed for researchers.**
- **Another challenge is related to data harmonization.** Up to 70-80% of data management efforts are spent curating (real world) data prior to conducting any analysis. As such, data harmonization is about creating a single source of truth, ensuring complementarity of diverse data, removing errors and inconsistencies, and aligning on assumptions, syntactic and semantic interoperability. To overcome these challenges, some recommendations are provided by the WG regarding the harmonisation of real-world data (RWD) from diverse sources (e.g. registries and cohorts).

In addition to the identified data sharing challenges and insights/recommendations for these challenges, this document also outlines some practical examples from different IMI projects (EMIF/EHDEN and RADAR-AD).

1 Introduction

There is a wealth of scientific data buried in the archives of hospitals, academic institutions, the pharmaceutical industry, and others that has not yet been leveraged. The sharing of data useful for research and clinical practice is increasingly viewed as a moral duty, especially in the neurodegeneration field where major breakthroughs and interventions being brought to market are still pending[1].

There is broad agreement in the research community on the value of data sharing[2-5] and on basic models for data sharing infrastructure (centralised, federated, hybrid models, etc).

- One advantage of sharing data is that it keeps researchers from having to **"reinvent the wheel"**—and to repeat the work that previous investigators have already done. Data sharing could really increase the speed of scientific discovery. It is important to highlight that there is a need to have access **to validated** data sets, which are of high importance in the Artificial Intelligence (AI) field but also in the research field. A paper by Birkenbihl et al.[6], showed that thorough investigation of real patient-level data is imperative to assess a data landscape.
- Data sharing can also be crucial in carrying out systematic reviews in a particular field. Often, it is necessary in a systematic review to reanalyse at least some of the data in the studies being reviewed, and that is only possible if the original researchers make their data available to other scientists.
- Another benefit of sharing data broadly is that you can bring more intellectual power and people from different disciplines and different perspectives into analysing data.
- Editors of international medical journals have labelled data sharing a highly efficient way to advance scientific knowledge. The combination of even larger datasets into so-called "Big Data" is considered to offer even greater benefits to science, medicine, and society. Several international consortia have now promised to build grand-scale, "Big Data"-driven translational research platforms to generate better scientific evidence regarding disease etiology, diagnosis, treatment, and prognosis across various disease areas[3].

However, despite the willingness and general ethos of data sharing to advance the field, in practice, it still proves to be quite challenging to provide an adequate framework for doing so that deals with the various organisational, legal, ethical, socio-psychological and technical challenges that hamper efficient data sharing.

Over the last years, numerous initiatives have been launched, that could lead to potential improvements in data sharing. Data sharing is increasingly required **by funders and publishers** to increase the reuse and reproducibility of research, and return on investment. To support this, publisher actions have included the publication of data availability statements with research articles. PLOS, Springer Nature and many other publishers now have journal **data policies** that require or recommend data availability statements and data sharing[3].

1.1 European framework for data sharing

1.1.1 Open Access to Scientific publications

The EU recognises that policy actions and governance frameworks can play a major role in encouraging data sharing. Since 2002, the **European Commission** has also proposed several initiatives to enhance the sharing of data that is generated in EU-funded research projects[7]. Under Horizon 2020, each Beneficiary must ensure open access to all peer-reviewed publications including the right to download(ing) and print(ing). A machine-readable electronic copy of the published version must be stored in a repository for **scientific publications** together with bibliographic metadata providing the name of the Action, project acronym and grant number (*Article 29.2 of the Model Grant Agreement*)[8].

A similar provision was included in the Grant agreement for the IMI2-Joint Undertaking, which was launched at the same time as Horizon 2020. Article 29.2 '**Open access to scientific publications'** of the IMI2 JU Grant Agreement details the obligations related to the provision of **open access** to peer-reviewed publications[9].

To further support Open Access, the European Commission has recently launched **Open Research Europe**, the open access publishing platform for scientific articles that present the results of research funded by Horizon 2020, and soon Horizon Europe[10]. Open Research Europe champions open science principles by immediately publishing articles, followed by transparent, invited and open peer review with the inclusion of all supporting data and materials. Ultimately, Open **Research Europe** will give everyone, researchers and citizens alike, free-of-charge access to the latest scientific discoveries.

1.1.2 Open Research data and FAIR principles of data sharing

To complement its Open Access policy, Horizon 2020 included provisions aimed at facilitating data sharing (*Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable and wherever possible, interoperable to specific quality standards*). In 2014, a set of high-level principles using the acronym FAIR '**Findable, Accessible, Interoperable and Reusable'** were developed and these data principles have been widely endorsed by European research funders[11].

In line with the FAIR principles, a restricted Open Research Data (ORD) Pilot was launched **as part of the 2014 Horizon 2020 Work Programme**. The ORD pilot required the underpinning data for scientific publications to be deposited in a research data repository, together with the information necessary to analyse and interpret the data.

Optionally, Horizon 2020-funded beneficiaries could provide further raw or curated data, such as unprocessed image files or databases. All funded projects were required to provide **a Data Management Plan (DMP)**, an essential step towards embedding data sharing principles in Projects at an operational level. To help projects to manage data in a FAIR way, guidelines were also developed.

To mitigate concerns around intellectual property loss, data privacy and national security issues, opt-outs were allowed from the ORD Pilot – but only if a reasonable explanation was provided

by the projects. In addition, projects were allowed to apply different degrees of data sharing, from fully Open Access data, to restricted/controlled access, or fully Closed data.

Although there is a dissemination element to sustainable data sharing, sustainability does not mean 'publishing everything' nor does it involve making all data available to everyone. The chosen Consortium approach needs to be well considered e.g. 'open access to data' may either positively or negatively impact sustainability and as a result does not have to be in full.

From 2017, **the ORD Pilot was** made the **default option** for all Horizon 2020-funded projects, paving the way for widespread sharing of data.

In 2018, a **cost-benefit analysis** of FAIR research data was published by the European Commission. In this report, the Commission cited a figure of EUR10.2 billion as the annual cost of **NOT** having FAIR research data[12].

2 Defining how 'data sharing' challenges inhibit science

Neuronet has established a Neuronet WG 'Data sharing and reuse', which consists of subject matter experts in data sharing and re-use participating in IMI Neurodegeneration projects (EPAD, PHAGO, RADAR-AD, RADAR-CNS, IDEA-FAST, EMIF, PD-Mitoquant) and/or NEURONET members (*Figure 2*).

Neuronet has also collaborated with other IMI projects outside the ND field such as BigData@Heart, EHDEN and FAIRplus, to obtain their learnings.

The Neuronet WG focussed on discussing several data sharing challenges in the ND field at large. These challenges have been captured in Figure 2 and will be explained in detail in the next paragraphs.

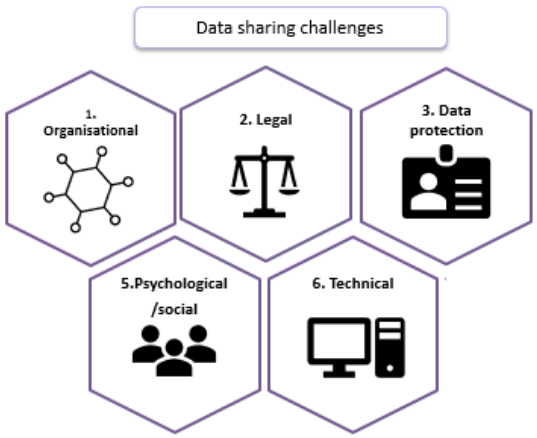

Figure 2. Data Sharing Challenges in the ND field

2.1 Organisational challenges

Clinical studies usually involve a complex hierarchy of relationships starting from the direct relationship between a clinician or clinical researcher and a patient or study participant, linked to research institutions or healthcare organisations. In turn, these institutions or organisations may participate in regional consortia, provide data to a repository, or be involved in data sharing networks. Within these frameworks, data sharing agreements become multi-layered documents that build on the initial agreement between patient/study participant and the clinician/clinical researcher.

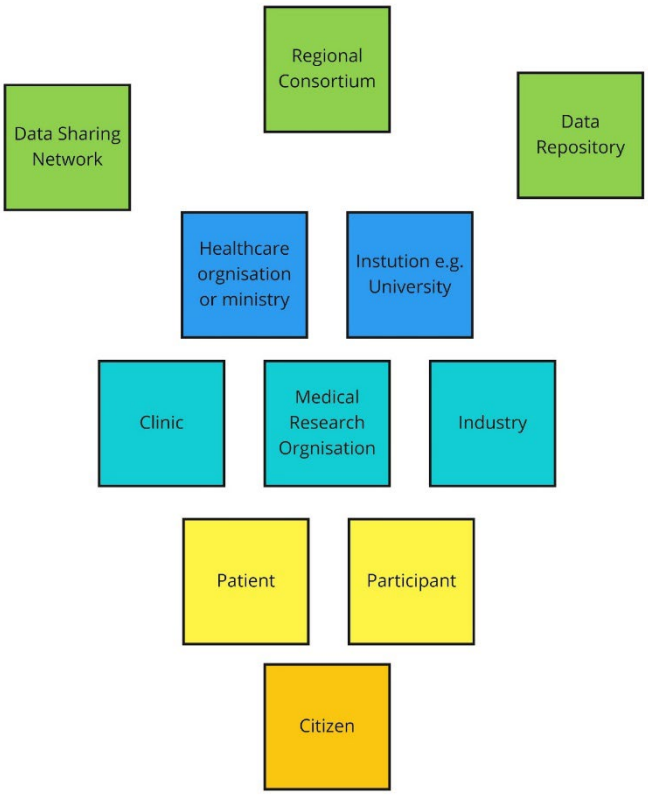

Figure 3. Organisational hierarchy

Interactions between stakeholders at different levels of this hierarchy can impact data sharing in different ways. For example, the citizen (as both patient and/or study participant), will have certain perceptions and needs in relation to data sharing and consent. The research organization or clinic attended by the patient or research participant may employ diverse legal and organizational models. Indirect organizational ownership by larger legal entities (a clinic may be part of a private healthcare system, for example) may introduce unexpected obstacles to concluding a data sharing agreement. Finally, data sharing may be influenced by initiatives such as data repositories or data sharing networks, which may have their own rules, guidelines or organizational frameworks (*Figure 3*).

2.1.1 Identified challenges

As a result, efforts to scale out data sharing to drive biomedical research face organizational challenges at numerous levels. There may be challenges linked to the ownership of data as 1) citizens have rights as data subjects in the study and 2) citizens have relationships with organisations to whom they are donating their data (= mutual dependence on data or reciprocity). There may also be challenges linked to individual studies, and how they are internally structured and governed; or there may be challenges between institutions participating in a consortium project, particularly when consortia involve organisations in different sectors. Regional or global initiatives such as DPUK or ADDI may also face specific organizational challenges.

2.1.2 Insights from the Neuronet WG regarding the organisational challenges

The following insights can be provided by the Neuronet WG on how to overcome organisational challenges:

- To understand the organisational challenges for data sharing, it is important to consider how individual actors are placed within the organisational model, what laws they may be subject to, and what aspects of data sharing they can control (*Table 1*).

Table 1. Different actors in the organisational model

	Legal basis	**Data sharing degrees of freedom**
Citizen	National and international law	Can give consent to data sharing models, case by case. Can control downstream use of data (under GDPR)
Clinical Researcher	Staff contract, professional qualification	Constrained by regulation if clinical, also by parent organisation
Medical Research Organisation	Legal entity, subject to regulation in legal territory, e.g. as a charity or registered as a data controller	High degree of freedom Acts as data controller on receipt or creation of data Can share data with researchers or subcontractors Can take custody of 3rd party data on behalf of researchers Can initiate and collaborate on projects with data sharing
Pharmaceutical company	Legal entity, subject to regulation in legal territory including company law	High degree of freedom Acts as data controller on receipt or creation of data Can share data with internal researchers or subcontractors Can initiate and collaborate on projects with data sharing

Consortium	Partnership agreement	Partnership agreement establishes a clear and usually constrained framework for data sharing inside and outside the protocol of a study
Data sharing network	May be a legal entity (often not)	If legal entity, can contract data processors and facilitate and host data sharing agreements

- The clearest source of organisational hurdle is the ability of individual parties to act in the roles required by legal frameworks. Under the GDPR, for any data sharing transaction the roles of individual parties should be clear. Participants or patients in the clinical study are the data subjects, while the data controllers or processors are usually the project Project Investigator (PIs), managers or a clinical research sponsor at a research institution. Legal entities providing technical services to process data can also act as processors. Other defined roles in data sharing agreements include data custodians (the person who manages the actual data); data stewards (the person within the organisation who is responsible for the quality and correct usage of the data) and data recipients (the entity or individual data is disclosed to).
- Organisational challenges may arise when the roles of different parties to a data sharing agreement are less defined or unclear. For any given data sharing discussion or relationship arrangement, it should **be clear which role each party belongs to and that each party has the organisational** basis to commit to that role. In addition, organisations and individuals may not invest time in suitable training to operate effectively within the legal data sharing framework.
- A capability maturity model is an approach from organisational improvement in IT. A similar approach could be taken to data sharing. Less mature or less experienced organisations may need to invest in process and training efforts to be more effective at data sharing.
- Finally, at each organisational level, different considerations apply – from privacy concerns (for participants), competitive concerns (for organisations) and intellectual property concerns. Moreover, each organisational unit or individual will have their own set of priorities, which are not always obvious or straightforward to negotiate. Interactions and discussions to clarify needs and constraints are necessary to overcome these challenges.

2.2 Legal challenges

2.2.1 IMI Legal framework allowing data sharing

There are essentially two main agreements structuring the IMI grant and collaboration between consortium partners. The Grant Agreement is the main legal document underpinning the project's execution – effectively, a contract between the participants and the IMI JU. The Grant Agreement mainly provides information on the grant (parties, duration, start date, budget, etc.)

and obligations of the Participants towards the IMI JU (such as reporting requirements), as well as the intellectual property framework and other legal conditions.

There are some legal hurdles to overcome when sharing data between Beneficiaries, between IMI consortia, or with Third Parties.

Considerations in terms of an appropriate framework are, without limitation:

- ownership of the data,
- access rights to the data + conditions (Royalty-Free, Fair & Reasonable) + use limitations (for the agreed Purpose only),
- time limits for requesting and exercising access rights,
- right to sublicense access rights (to Affiliates, subcontractors, Third Parties),
- the parties involved, their role (Data Controller or Processor) and location (in/outside EEA),
- privacy restrictions,
- ownership of / access rights to results generated with the data.

Some of these agreements are devised as multi-party agreements between all partners in both consortia. While in some cases this may be mandatory due to the respective Consortium Agreements, in some cases the process could be streamlined by focussing on which institutions actually own (or control) the data, and which will use the data on the receiving end. Similarly, these agreements could be limited to a specific purpose, and not be wide-ranging, to simplify and accelerate the process.

A certain notion of a 'quid pro quo' can also be useful to incentivise both ends. An honest appraisal of whether any additional work will be needed to enable data sharing, and compensations for such work, can help.

The Beneficiaries of two IMI Consortia can enter into a Collaboration Agreement (Cross-Consortia Agreement) in order to share some of their datasets for a specific purpose (for performing their project or for making the project results sustainable). However, since all beneficiaries of both projects frequently need to approve and sign a dedicated collaboration agreement, this often leads to a very time-consuming process causing major delays and sometimes completely undermines timely collaboration. Data agreements can also be executed with other Beneficiaries (and their Affiliates), Associated Partners (e.g., data contributing parties), Linked Third Parties, Third Parties and other stakeholders (e.g. Data Sharing Agreements, Data Processing Agreements).

2.2.2 Challenges identified through the Neuronet survey

Neuronet conducted a survey about previous cross-project collaboration attempts between IMI ND projects. Projects were asked to provide information on: 1) the topic of the collaboration; 2) whether the results of the collaboration were satisfactory or not; 3) whether legal support was required to materialise the collaboration, and 4) whether there were any specific obstacles hindering the collaboration[13].

Table 2. Collaborations between IMI projects

Project 1	Project 2
AMYPAD	EMIF-AD
AMYPAD	EPAD
PHAGO	ADAPTED
PHAGO	EBISC
PHAGO	EPAD
PHAGO	IMPRIND
EMIF-AD	EPAD
EMIF-AD	AETIONOMY
AETIONOMY	EPAD

The following results were obtained through the Neuronet survey:

- Out of the 9 collaboration attempts (*Table 2*), 6 were materialised (totally or partially) and 3 were unsuccessful.
- The main obstacle for collaboration reported by the projects were **the long delays due to the nature of preparing a collaboration agreement and collection of signatures**. As a consequence of such delays, the data was available only at the end or, in some cases, even after the end of the project[13].

2.2.3 Insights from Neuronet WG on how to overcome legal challenges

The following insights can be provided by the Neuronet WG on how to overcome legal challenges:

- Obtain internal approval from business, IP and regulatory groups for sharing specific (preclinical/clinical) data sets with external parties (e.g., check if the data are proprietary or in-licensed, check use restrictions in the applicable ICF);
- Discuss/assess the requirements for data privacy and data transparency (e.g., what data are required to achieve the Project Objectives (data minimisation principle), what is the appropriate data format : pseudonymized, anonymized or synthetic – this is a balanced decision between data protection and scientific value);
- Involve legal teams to prepare a (Material &) Data Transfer Agreement (side agreement to the CA / GA).
- If standards for data agreements (data licenses) could be made available and IMI projects only have to sign up to a standard “data” agreement, this would prevent preventing time wasting on legal discussions.
- As legal discussions in IMI projects can take several months/years, adequate resources should be included in IMI projects (for legal staff) to solve this issue.

The following Insights were provided from the FAIRplus project:

- To overcome the legal hurdles when sharing data between two IMI projects, the IMI2 project FAIRplus is developing **legal templates (CDA, collaboration agreement)** for cross-consortia agreements to maximise the value/impact of data generated by IMI projects. The legal templates are in the final stages and will be accessible on the FAIRplus website[14].

2.3 Data protection challenges

2.3.1 Sharing of personal data

Another important consideration is that legal norms specified for the sharing **of personal data** for health research have been developed in the EU, most notably those set out in the GDPR (EU 2016/679). Under this new legislation, individuals will receive more information about how and why their personal data are being collected, used, disclosed, transferred and retained. They also have the right to obtain a copy of their personal data, to have the data transferred in a portable format to another entity of their choice, or to request that their personal data can be erased under specific circumstances. Where consent is necessary, requests for consent must be presented in a more easily understandable and accessible form, and it must be easy to withdraw consent. However, these rules and regulations remain **open to interpretation** and offer limited practical guidance to researchers[1,7].

2.3.2 Identified data protection challenges

Researchers are critical of the fact that there is no **clear lawful basis for secondary data processing (i.e data reuse) in the GDPR** – instead, the choice of legal basis is left to individual researchers or institutions to determine. There is also no clear definition **of "scientific research";** Recital 159 lists a series of examples, but whether the rights of data subjects are likely to "*render impossible or seriously impair the achievement of these specific [research] purposes*" is left open to interpretation[1,7].

Similar learnings can be observed from Bigdata@Heart, an IMI2 project that aims to develop a Big Data-driven translational research platform of unparalleled scale and phenotypic resolution in order to deliver clinically relevant disease phenotypes, scalable insights from real-world evidence and leads for drug development and personalised medicine. To accomplish this, BigData@Heart will combine data from a large variety of already existing databases to perform advanced analytics. The aim of Work Package 7 is to deal with all the relevant ethical and legal issues in order to establish a **sustainable governance** for the data infrastructure during the project and beyond[15,16].

2.3.3 Insights/learnings regarding data privacy challenges

The following **insights/learnings were provided from Big data@Heart[15,16]:**

- BigData@Heart contacted Principal investigators of 48 participating databases via e-mail with the request to send any kind of documentation that possibly specified the conditions for data sharing. Documents were qualitatively reviewed for conditions pertaining to data sharing and data access. The following learnings were provided:
  - A governance system cannot fall back on prespecified local policies.
  - The role and value of **local data protection officers** is currently underestimated, and their involvement is key for data sharing.
  - There remains a lot of administrative work at the local centres, which is still a hurdle. There is a preference **for 'networks' (federated approach)** instead of a centralized governance structure, to decrease the administrative burden.

2.3.4 How to move forward with data sharing for secondary research?

A main barrier to data sharing among researchers is the lack of clarity around legal and regulatory policies and practices.

Before human data collected in a primary study can be shared with other researchers **for secondary research**, it should be evaluated whether the consent forms under which the data were collected permit such sharing for secondary research.

As an example, ADDI has **made a decision tree** to determine if consent forms permit sharing data with third parties for secondary research on Alzheimer's Disease[2]. This decision tree will help researchers to analyse the consent forms to determine whether consent forms permit the desired sharing. If this decision tree indicates that your desired sharing or uses of the data are precluded by the consent form, [legal/administrative] colleagues should be contacted to explore possible alternatives[17].

2.4 Psychological/ Social challenges

2.4.1 Identified psychological/social challenges

When discussing the ownership of data, both the psychological/social and motivational aspects of data sharing play a role. A survey was conducted to understand how important it is for researchers that their data are discoverable. On a scale of 1-10, the average rating was 7.3 indicating that they find it important that their data are discoverable[3].

However, researchers also report **several obstacles** to data sharing in practice. Requiring investigators to make their data—and often their software—available to other researchers puts a tremendous burden on the investigators. A recurring theme in **surveys are social or motivational barriers** to data sharing. Notably, the well-established scientific system of individual reputation and rewards, and the notion of data as the new "gold" can generate an exaggerated sentiment of ownership and competitive 'loss' associated with sharing. This can create barriers, sometimes implemented as over-complicated access processes[7].

The social dynamics of sharing come into play at each organisation level or individual, which will have their own set of priorities. At each level, different considerations come into play - privacy concerns, competitive concerns, and intellectual property concerns.

Besides the motivational barriers, researchers – and in particular, researchers working on clinical studies – also indicate the **financial and time cost of data sharing** as a key challenge to overcome. To share data well, it **takes time, effor**t **and money**.[7]

Trust, trustworthiness and credibility are of paramount importance to facilitate sharing in IMI projects – these are crucial elements in the case of consortia, where by definition a degree of sharing and collaboration is implicit in the work plan.

2.4.2 Insights from the Neuronet WG on how to overcome these challenges

The following insights were provided by the Neuronet WG:

- To overcome **motivational barriers of researchers**, methods to better incentivise data sharing can be a solution. Some examples on how to improve this are listed here:
    - IMI projects are not funded to share data. Systems should be in place to assure data sharing capabilities survive the originator project. Also, projects could for instance be funded to generate reusable data. A possible solution can be to encourage funders to provide additional budget specifically for data transfer and harmonisation efforts of researchers. Therefore, researchers could act as data stewards and can be fully credited when their data are re-used[6].
    - There should be some career prospects for researchers when sharing data. For instance, a metric could be introduced that counts the amount of data shared under FAIR principles for researchers.
    - Publicly celebrating role models could be considered for researchers when sharing data (e.g. award for clinical data sharing).
- To facilitate data sharing between researchers in IMI projects, trust, trustworthiness and credibility are of paramount importance.
    - More specifically, researchers are only willing to share the data based on the profile and **the reputation that you/your institution has built.**
- When considering the patient perspective, **trust and trustworthiness** also play an important role. Trusting requires a leap of faith. It requires accepting the uncertainty and the risk that the clinicians will act to the best of their ability and most definitely, in good faith. It is also important for clinician researchers to be trustworthy, and not merely reliable. For most patients, their data is shared to and by institutions and organisations they may not even be aware of, so there has to be not only trust from a data subject, but also proven trustworthiness of an interested research entity, and reliance on the working of processes, policies, procedures and technologies.
- Although it has taken tremendous time to obtain data in the ND field, things are slowly changing. Due to COVID, the importance of sharing data between research projects was reinforced and **there is nowadays a much higher preparedness and willingness to share data with researchers and policymakers to advance the science.**

2.5 Technical challenges (e.g. databases, infrastructure)

Technologies to capture, manage, discover, standardise, visualise, analyse and generally exploit data in multiple ways are continually being improved globally, both in terms of functionality and computing power. As a consequence, very rarely can one think of technology being itself the limiting factor when it comes to data sharing. However, several technical issues have to be considered and these are listed in the next subsections.

2.5.1 Lack of FAIR data as a technical challenge

Fragmentation of the data landscape in projects is also a significant issue, hampering interoperability and incentivising new research projects to come up with yet more *de novo* developments. This results in a high number of solutions that are not maintained or further developed, which affects the associated datasets. The need to 'reinvent the wheel' every time also leads to a sizeable number of rudimentary solutions, as every project tries to fulfil its particular needs under constraints of budget and time.

To gain most benefit from research data in IMI projects, data should be available to researchers. Key to making data sets **'findable'** is the notion of meta data, which in turn is key (but typically underused) to provide understanding to future users about the context in which data were collected, limitations to their applicability, and interpretation notes, all of which can hugely affect re-usability**. Accessibility** is also an important concept – in that sense, transparency in procedures to request and grant data access are paramount.

2.5.1.1 Insights from FAIRplus project

To address the aforementioned issues, the IMI2 project FAIRplus was launched in 2019. The aim of FAIRplus is to develop guidelines and tools to make data **FAIR**. FAIRplus aims to increase the discovery, accessibility, and reusability of data from selected projects, as well as internal data from Pharmaceutical industry partners[14].

The following **insights/learnings were provided from FAIRplus (during a Neuronet WG TC):**

- Within FAIRplus, two main tools (FAIR CMMI and FAIR cookbook) are in development that **will enable researchers to assess the FAIR level of datasets**, to understand the benefits of achieving a higher level of FAIR, and to follow a process and guidelines on how to actually make data sets more FAIR.
- **The FAIR cookbook** aims to collate protocols for making data FAIR. To make data accessible in the long run, FAIRplus[14] is applying their efforts to the Elixir IMI data catalogue at the University of Luxembourg[18], which will be a searchable **metadata repository** for IMI data.

2.5.2 Lack of metadata as a technical challenge

It is important to identify **all existing data** that may have resulted and are available from IMI projects, and to share high-level information about such datasets to support a metadata-driven catalogue for FAIR data. Several cataloguing initiatives have been developed within IMI neurodegeneration projects (e.g. EMIF Catalogue, ROADMAP Data Cube). More broadly than the ND field, the ELIXIR-LU/eTRIKS Data Catalogue is a data catalogue that is being developed for large research initiatives such as IMI and H2020 that centralises metadata of ongoing and past projects[19].

All these catalogues enable the detection of **the existence of data, without accessing the data themselves** – thus providing very useful ways to facilitate requests to whoever holds pertinent data for access to the data sets of interest. Providing online access to a database or enabling visualisation of data can also be understood as meaningful sharing (and, indeed, there are technologies allowing data to be captured in those situations). Access to metadata or data discovery approaches (i.e. revealing the existence of data, but not the data themselves) can also inadvertently become forms of sharing unless carefully designed and implemented. Problems associated with these initiatives relate to long-term maintenance, precision of the contained information, and lack of representation power, so it is important that users understand the limitations that may apply.

The use of cloud technologies to cope with ever-increasing amounts of data presents additional challenges in terms of physical hosting that needs to be aligned with local requirements (e.g. Europe) and in terms of security, that can affect perception on actual control of the data, which in turn can affect the psychological aspects related to enthusiasm for data sharing.

2.5.2.1 Insights from the Neuronet WG

Despite a big drive in recent years towards 'open' solutions and 'open' data, and the production of many (sometimes overlapping) online repositories and catalogues, adoption and re-use of tools and data relies heavily on adequate provenance, context and application domain.

- Support systems for data sharing will not be effective unless they also elaborate on the nature of the data, where they come from, the purpose for which they were collected, etc., all factors that can decisively affect further analysis and interpretation. Generally, this implies a need for data producers to annotate, document and provide meta-data that are as informative, efficient and actionable as possible[20]. Despite developments in terms of semantic web technologies, etc., human input into such meta-data provision remains key in many areas, involving huge efforts that are often underestimated. In the long run, the situation might improve if annotations with **meta data are done by machines.**
- As data volumes exponentially grow, so does the need to be able to meaningfully visualise and understand such data. Innovation in terms of ways of interacting with data in some cases lag behind the pure generation of massive amounts of data.

2.5.3 Pseudonymisation, anonymisation and consent as a technical challenge

One of the standard ways to share sensitive data in which the privacy of the subjects must be respected is to "deidentify" the data before they are released. There is not a precise definition of deidentified data, and much debate and uncertainty about what constitutes deidentified data remain. There are also some ethical and data protection requirements that add a further level of complexity: for individual participant data from clinical studies. Data access committees need to manage requests and ensure that participants are not re-identified from their data**.**

The GDPR rules on pseudonymisation, anonymisation and consent are currently open for interpretation. Without a consistent framework to manage pseudonymisation, anonymisation and consent, many research institutions and data protection officers are understandably hesitant to share data for secondary research[1,7].

2.5.3.1 Insights from Big data@Heart

The following **insights/learnings were provided from Big data@Heart[17,18]:**

- To ensure responsible use of data in BigData@Heart as well as similar research projects, good governance of data sharing and data access is critical. So far, no blueprint of a broadly accepted governance framework exists. Within BigData@Heart, a governance is needed to manage privacy and confidentiality issues, to ensure valid informed consent for data sharing, to determine who will decide about the data access and to promote social justice and public trust.
- Since the legal norms (GDPR) remain open and offer limited practical guidance to researchers, **ethical guidance is needed to create, reinforce and reproduce social norms and institutions.** The limitations to preserve anonymity and confidentiality of shared data are recognised. In order to truly safeguard the rights and interests of participants, future work should concentrate on the development of measures to establish public trust in data sharing activities, at all levels of (de-)identification.

2.5.4 Data standardisation/harmonisation as a technical challenge

Considerable time, perhaps 70-80%, is spent curating (real world) data prior to conducting analysis, and this is further complicated when working with data from multiple sites, in multiple platforms, across multiple languages (human and machine).

- There are fundamental tensions between quantity and quality, derived from the need to curate data, an effort that frequently can be more intensive than data production itself, especially in environments prone to noise. Since the definition of 'quality' is often a moving target that pertains to the specific research question, data processing becomes a complex process.
- Data harmonisation is about creating a single source of truth, ensuring complementarity of diverse data, removing error and inconsistencies and aligning on assumptions, syntactic and semantic interoperability. A number of approaches can be used (with varying pros and cons) to harmonise data, usually with three operations, extract, transform, and load (ETL).

Depending on the source data being transformed, this can be resource intensive, and some argue that the act of harmonisation can impact on the subsequent analysis due to the imposition of a specific structure. Fidelity of the harmonisation, i.e. if there has been any appreciable loss from source to harmonised data, needs to be evaluated to substantiate the veracity of performed analysis. A relatively straightforward data warehouse, a repository for the ETL output is a common approach, and increasingly a data lake or cloud, where the ETL can become ELT, so transformation can occur prior to analysis from the diverse loaded data in the lake. Intrinsic to the ETL process is audit and data hygiene, with collaborative evaluation of a dataset with those who have domain expertise, and those who can perform the ETL (can be one and the same, but also often not), providing revealing insights into data characterisation (i.e. completeness, consistency and coverage), as well as the assumptions underpinning the source data.

- The use of a common data model (CDM) to support harmonisation and interoperability, for instance within a standardised, modular and extensible collection of data schemas, has gained considerable ground in recent times. Harmonisation of vocabularies is integral to this process, especially within CDMs such as OMOP (Observational Medical Outcomes Partnership). The FDA's Sentinel within a shared health data network (SHDN), the OMOP CDM within a federated or distributed network, the Kaiser Permanente CESR (Center for Effectiveness and Safety Research) virtual data warehouse, or the PCORI (Patient Centered Outcomes Research Institute) CDM, are examples of such approaches, facilitating collaboration and harmonisation of diverse data for analytics, in particular and for example, via a standardised analytics stack from OHDSI (Observational Health Data Sciences & Informatics) initiative, utilising the OMOP CDM.
- Data harmonisation is a necessity, and furthermore, in the context of collaborative neurological projects, moving to a FAIR construct for their data, agreement on the harmonisation approach is critical in the longer term for success with regards to ensuring a common purpose (i.e. analytical outputs), efficiencies of scale, longevity and sustainability, and return of investment. In the short term it is a socio-technical construct with regards to the need to collaborate, investment of both human (e.g. domain and infrastructure expertise on a given dataset) and machine resources to achieve a state of interoperability. Unless specifically resourced the ETL and harmonisation of neurological data, diversely collected, stored and analysed, will be difficult, and requires utilisation of specific expertise, knowledge, and skills. Within the IMI2 Big Data for Better Outcomes (BD4BO) initiative, individual projects, such as HARMONY in haematological cancers, are mapping to the OMOP CDM, in this case via a pooled (centralised) SHDN, with PIONEER in prostate cancer working on mapping to the OMOP CDM via elements of a pooled SHDN and a federated SHDN, a hybrid model, or in the case of EHDEN (European Health Data & Evidence Network) a federated or distributed SHDN. The EHDEN project is unique in utilising certified small to medium enterprises (SMEs) to undertake the ETL with Data Partners, whilst working symbiotically with OHDSI on methodological, tools and use case development.
- Within neurodegeneration and real world data (ND-RWD) use specific examples exist in neurology and IMI, such as the EMIF-AD (European Medical Informatics Framework – Alzheimer's Disease) experience, where AD registries were harmonised via a variant of the OMOP CDM, utilising a specific variable set, can provide direction as to a future path more widely. Initially a number of AD registries were involved in the ETL work to assist with the project's research aims, initially using the tranSMART data warehouse, and then the OMOP CDM variant, but this work unfortunately stopped at the end of the IMI project (May 2018). Interoperability with external projects, such as the Global Alzheimer's Association

Interactive Network (GAAIN), which utilises a CDISC intermediary for ETL/harmonisation, was also envisaged by EMIF-AD, and would have potentially led to international interoperability for AD data. As many of the AD registries were not dynamic, the historical data within projects, such as EMIF-AD, could still be valuable, especially as harmonised datasets.

- More recent developments in neurodegeneration, such as the Multiple Sclerosis Data Alliance (MSDA), are pointing to wider initiatives to create federated data networks to support critical research from both patient and clinician perspectives. This was further illustrated by a rapid harmonisation and analysis project by MSDA and collaborators to research risk of MS therapy with regards to COVID-19, spread over weeks, not months. Using a standardised ETL process enabled acceleration of research under pandemic conditions, but also reinforced what is achievable in general utilising a common data model (in this case OMOP).
- Other challenges, in particular for semi and unstructured data, which require additional work, such as natural language processing, also need to be addressed to release even more potential data for study, which will add to structured data for harmonisation and standardised analytics. Contemporary developments in methods, tools, and resources for working with all such data will only increase the resolution of RWD for evidence and insights into neurological and all other diseases. Supporting such use cases as machine learning, are wholly dependent on training and validation sets, which can be challenging for certain diseases and populations in terms of availability and in particular representativeness. As such, harmonisation and interoperability of diverse datasets will become an even more pressing need.

Two basic models for data sharing infrastructure have been tried in the past:

- Centralised. This approach is based on the existence of a central location that gathers data from a number of data generators and provides access mechanisms to a number of data users. The scheme can be reproduced on a number of levels, so that e.g. data generators can be gathering data from several sites; similarly, central nodes can become part upstream of bigger meta-platforms. This model has in general advantages in terms of clarity of who is responsible to custode and organise data, following in some cases an “honest broker” paradigm where trust and clear terms and conditions become key underpinning factors. It also has disadvantages in terms of implying transfer of data to another location, which can be affected by problems of legal, ethical, governance and psychological nature and therefore requires an appropriate governance model.

- Federated. This approach relies on data being kept at source with data source as final arbiter on its use, and devising instead mechanisms to process and analyse such data in a distributed manner – generally by allowing data custodians to run specific software on site, then share the results of such processing only for centralised analyses. This generally has advantages in terms of compliance with local legal and ethical rules and regulations, as the data don’t have to be transferred anywhere (reinforces the need for metadata & FAIR principles overall!), it can also help engagement through an enhanced value proposition where a multiplicity of data generators of diverse origins are needed to achieve critical mass. It can also have disadvantages in terms of diluted responsibility, reliability and persistence of data, audit trail and also regarding the establishment and operation of unified access mechanisms for potential data users.

Hybrids models are possible, e.g. a federated system for discovery or high-level interrogation of data coupled with a centralised system for selective centralised data sharing, or a generally federated system coupled with periodic transfers of subsets of data to a central database, etc.

Example: ADDI is adopting a Common API based on open standards for federated networks of data sharing. This emphasises the importance of metadata to support data discovery and FAIR principles but also codifies what federated queries and federated computation mean.

2.5.5 Insights from the WG regarding data harmonisation

Incorporating higher dimensional data, in particular genomics, provides additional challenges, in part due to the intrinsic application of that data, i.e. understanding and ensuring clinical meaning, as well as understanding of the linkage between genomic and phenotypic data. Making informed and aligned assumptions on harmonising this data is critical for its usefulness in subsequent analysis and interpretation. For harmonising ND-RWD from diverse sources, e.g. registries and cohorts, the following recommendations ought to be considered. Ultimately, harmonising data is an inevitable requirement of working with ND-RWD, and short term pain is worthwhile for more efficient and reliable longer term gain in addressing research and scientific need:

- There needs to be a common understanding of the focus and standardised querying required for the common research proposed in a collaboration – what are the questions?
- Harmonising data can facilitate, via a CDM, standardised analytics to support higher reproducibility, transparency, rigour, and confidence in research outputs, so it is a means to an end, and not an end itself
- Utilising the ETL process to generate deeper insight into individual datasets while harmonising is an excellent opportunity to have a feedback loop to the source for verification and improvements
- During an ETL process, e.g. to the OMOP CDM, there should be a clear process for working between those knowledgeable of the source data and those responsible for the ETL, and clear verification and evaluation steps. Semi or fully automated steps and tools, with output reports during sequential steps and at the end of the ETL phase are important
- With ND-RWD it is likely there will be a subset of variables harmonised, perhaps for specific queries, or for an ongoing programme of research. Aligning on what will be harmonised is of paramount importance, and what is realistic (e.g. long tail or frequency table-led) akin to understanding an exam question
- Verification and evaluation of the fidelity between source data and harmonised data is good practice, in part with appropriate tools (integral to the OMOP CDM ETL process), but also in conducting validation studies, for instance by re-running protocols previously run in source data in the harmonised data
- Utilising standardised analytical tools assists with the preceding recommendation, and also assists with error detection with regards to whether an issue is with the source/harmonised data or the analysis, in particular with e.g. higher dimensional data

- Sharing the harmonisation/ETL process, scripts, tools, and methods across the collaboration is helpful in ensuring complementarity of approach, even with a centralised ETL, while also educating relevant parties in the inherent steps and outputs
- Harmonising may be a one off process, for instance with historical or static datasets, quite often with ND-RWD. With more dynamic datasets, the frequency of updates will need to be agreed, depending on the scope and scale of those datasets, and the ETL approach (e.g. to a CDM) could be semi or fully automated.
- In any case, harmonisation should not be a hard barrier or prerequisite for collaboration or federation. Useful insight on data divergence or degree of harmonisation can be obtained in a federated set up. This accelerates efforts to harmonise data or map onto common data elements.

2.6 Practical examples for specific datasets

2.6.1 Sociotechnical Construct for working with real world data in Neurodegenerative disorders – IMI EMIF/EHDEN

Utilisation of real world data (RWD), captured in the main for clinical primary use in a normative, observational setting outside of a clinical trial, for insight and evidence generation is not new, but technological advancement and expectations have seen a remarkable expansion[21].

The use of RWD is disease agnostic, i.e., generic capture of clinical and allied data from diverse sources generated about a patient, whether phenotypic, genotypic or both, is enabling the realisation of new insights into our own biology, right through to real world outcomes of therapeutic interventions on disease progression. With reference to ND, due to their intractability of treatment, the lack of cure, genetic or familial associations, and the need to both understand their pathogenesis to a molecular level, as well as identify potential therapeutic targets over longitudinal timelines of years to decades, the need for large-scale patient data is paramount.

Working with RWD is a sociotechnical construct, in as much as there is a technical requirement to identify, curate, analyse fit for purpose data, but within a sociological framework of governance, ethics, policy and law to ensure citizens/patients are protected sufficiently, data is closed enough, but also open enough to allow research purposes. Within Europe specifically, the introduction of the GDPR in 2018, and the Data Governance Act, 2020, reflect the importance to Member States and European citizenry in assuring their rights in a digital society, whilst protecting them and their families and carers, whilst facilitating research for bona fide intended purposes[22].

Within IMI, there have been a considerable number of ND-related projects, such as for Alzheimer's disease, like EMIF-AD, RADAR-AD, AMYPAD or EPAD, as well as other ND diseases such as in RADAR-CNS, PD-MIND and PRISM. There has been a cumulative experience of working with ND-RWD that inform the future approach to governance and protection within the sociotechnical construct for ND research.

From this operational and real world experience to date a number of lessons learned will inform both broad and granular recommendations for ethical practice in working with ND-RWD. What needs to be considered for stakeholders working with ND-RWD:

- There needs to be a **clear and transparent ethical overview**, both in terms of legal aspects, but also in terms of guidance as to ensuring ethical tenets, such as autonomy, nonmalifence, beneficence, and justice in preserving the rights of the individual in conducting ND-RWD related research. A particular ethical consideration that complicates ND-RWD research and data sharing is capacity and consent; ideally, consent should be broad enough to enable data re-use and data sharing at a point in the future where the subject may no longer have legal capacity, but stringent enough to ensure the rights of the data subject are always respected.
- An understanding of the **motivational aspects for patients/citizens to participate in research is important**, or at least the motivation to support or consent to the use of their data within large-scale, non-identifiable research is important. This is dependent on transparency of communications and purpose, and whether this is wholly altruistic, or linked to incentives, whilst understanding the balance of benefits versus risk[23].
- Beyond the data subjects, there are also **motivational considerations that affect researchers**, and can preclude the effective sharing of health data. Qualitative studies[24-25] indicate that these motivational considerations are linked to systemic disincentives to openly share research data, and an absence of standard processes to credit data generators. There are also legitimate sociotechnical concerns that can make researchers less motivated to share health data; who should absorb the cost of data sharing, and how can we ensure that data collected for a specific purpose is used in a way that respects the original study aims?
- Citizen/patient engagement, or indeed public involvement overall is growing in maturity of purpose and methods, partly due to digital tools, but assurance of being able to assert views, opinions and concerns needs to be central to the overall debate on how best to utilise RWD. Within ND this also needs to include consideration of the role of advocates or representatives that can opine on behalf of those with ND who are unable to do so themselves, whether a relative, carer or legal guardian, especially related to consent.
- There is an existing data protection framework in most EU Member States to support research, with variable adherence to more recent legislation, such as GDPR, in part due to differing maturity of systems, but also due to derogation of interpretation by Member States. Essentially, at least 27 interpretations, with considerably more access points, dialogue points and compliance requirements impinge on the speed and reproducibility of being able to conduct research with ND-RWD (and RWD in general)
- ND research can be and is complex, and there can be overlap between basic, translational research, clinical studies and RWD research. Consequently, there can be a need for instance to re-contact individuals, with appropriate governance, to augment data with e.g., new samples or new information. This can pose technical challenges in protecting individual identity or at least reducing recognisability whilst serving appropriate research requirements. The ethical balance of such need versus individual rights, whilst adhering to local and wider law or policy is an essential prerequisite evaluation.
- Risk management with relevant mitigation strategies needs to be a recommended requirement with regards to security, protection and failure management, especially with regards to protecting the sanctity of individual data and identity for instance in the occurrence of a data breach or incursion
- The need for trust by a research participant, or their representatives, as opposed to trustworthiness of the research interest entity is a bi-directional and active process by all actors in this regard. All projects have experienced the need to support transparency

in this whole process of trust/trustworthiness, and recent work within IMI2 project the European Health Data & Evidence Network (EHDEN) on a 'concentric circles' framework may also be informative for working with ND-RWD. IMI EHDEN has recently worked out a research quid pro quo for sustainability framework based on trust and relevance. Openess, transparency and collaboration are the key drivers to gain the necessary trust. Relevance needs to be based on relative usefulness of data, research outputs and accessible technology. Ultimately, trust and relevance are based on successful outputs.

- Other than for rarer diseases, individual data is of little benefit, but aggregation of individual data (and samples) in addressing ND research is the norm rather than the exception. Historical, contemporary and prospective data capture systems, whether for translational insights through to optimising clinical studies, need to incorporate appropriate protection, both technically and via policy adherence.
- Enhanced access and sharing typically requires opening information systems in order that data can be accessed and shared. This may expose parts of an organisation to digital security threats which can lead to incidents that disrupt the availability, integrity or confidentiality of data and information systems.

Based on these principles derived from experience and practice in e.g., IMI-related ND-RWD projects, a number of **recommendations** can be suggested:

- Anyone working with ND-RWD needs to obtain ethical advice, or indeed the use of an ethics advisory board to support appropriate and adherent research in the context of societal norms. A balance is required between risk and benefit for the individual, a cohort and society at large regarding any research utilising ND-RWD.
- Legal guidance needs to be sought to ensure alignment with e.g., GDPR, Data Governance Act, and derogated Member State interpretations and laws, as well as local institutional requirements.
- Intended use and purpose of any research needs to be transparent to all involved, complying with local and regional consent requirements, as well as governance needs, inclusive and up to publication of findings, positive or negative
- Depending on the nature of research, the opportunity to include meaningful patient and public involvement to provide guidance and direction on the use of ND-RWD within the bounds of bona fide research should be explored. This should be inclusive of representing sociocultural norms and diversity
- It would be sensible to develop overarching code(s) of conduct to ensure consistent application of approaches that meet ethical and data protection requirements across projects that use ND-RWD rather than multiple and individual approaches. Numerous guidance exists within Europe to support use of RWD *per se*, and can be incorporated into research practice with regards to nuances of working specifically ND-RWD.

To avoid digital security threats, IT systems should be in place that will allow that data can be accessed and shared.

Working with ND-RWD is a complex challenge, and though many of these challenges are germane to RWD use, there are nuances reflective of this therapeutic domain. Any initiative

needs to reflect this in its research practice, but also incorporating appropriate frameworks for transparent, ethical and sensitive research which include patient/citizen views, balancing benefit, motivations, and risk is of paramount importance.

2.6.2 Datasets and use of remote measurement technologies: the RADAR- AD experience

The area of patient-reported outcomes is increasingly important and attributed to pervasive use of smart devices and increasing responsibility about one's own health.

Smart devices track a wealth of activity on the people who choose to wear them, such as daily activity patterns and levels, calories burned, sleep patterns, and weight. While wearing smart devices are personal choices, the world is moving toward gathering data about people to conduct research. Realistic discussions about privacy and confidentiality will have to take into account the coming changes in the ways in which data are collected, the types of data that are collected, and the attitudes that people have about their data being collected[4].

Remote Assessment of Disease and Relapse – Alzheimer's Disease (RADAR-AD) is investigating how mobile technologies can improve understanding of functional decline in Alzheimer's Disease (AD). Due to the nature and complexity of data collection and the variety of data types in RADAR-AD, the RADAR-AD Consortium established a DMP prior to the start of participant enrolment. The DMP includes policies, actions and ethical scrutiny regarding the governance of any data being produced during the RADAR-AD research project. The DMP aims to manage the data produced by the patients and, caregivers and the aggregated data produced for analysis and maintenance purposes. This document provides the actions, policies and principles as data is created, updated, maintained and searched. The policy in the DMP follows the FAIR principles of data management, providing information on;

i. How study data is handled during the project lifetime
ii. What data types will be collected or computed
iii. Standards and ethical policies for study data
iv. The storage and retention of data during and after the project.

In addition to prospective data collection, one of RADAR-AD's tasks is to select and use relevant longitudinal dementia datasets for statistical modelling. We initially had access Alzheimer's Disease Neurodegenerative Initiative (ADNI) and AddNeuromed (ANM) data. The next step was to apply for access to other cohorts, contacting the data providers of 12 datasets. This action was very time-consuming **due to legal, ethical and privacy concerns** related to data sharing.

An appropriate data sharing plan, including establishment and maintenance of data access committee and data access and sharing policy, and the selection of a sustainable data repository implementing the FAIR principles has been developed, agreed upon, implemented and communicated by RADAR-AD. **Data sharing and interoperability** is paramount to the success of the RADAR programme. The framework supporting this data sharing (i.e., the type of data to be shared and access governing data sharing) had been established in line with IMI2 Intellectual Property (IP) policy and considering the overall approach agreed upon in the other RADAR projects. EFPIA members and consortia partners are committed to sharing all data (clinical, biosensor etc.) available to, or generated by the RADAR program amongst all members of a RADAR topic, and across topics, as required. In addition to data, RADAR constituents also share domain practices and expertise developed with respect to data management procedures, usability, regulatory and policy pathways etc. across the RADAR program and externally as

required by IMI policy and procedures. It is expected that any system built within the RADAR programme adheres to well-accepted data standards, where applicable, to ensure compatibility and interoperability with other systems both within the RADAR programme and more widely. The developed solutions, irrespective of whether leveraging the foreseen facilitating common platform infrastructure or built independently from it, should, in any case, allow for cross-analysis, data stream sharing and aggregated visualisation across all RADAR-AD solutions, as well as in combination with pre-existing solutions such as those being elaborated under RADAR-CNS. It is paramount to the value of the project deliverables that they do not result in vertical, ad-hoc solutions.

Learnings from RADAR-AD:

- Acquiring access to the data sets is time consuming due to legal and ethical issues underlying each of the data sets. It was indicated that improving this process will facilitate better research by promoting collaboration and multifaceted working.
- The developed solutions, irrespective of whether leveraging the foreseen facilitating common platform infrastructure or built independently from it, should, in any case, allow for cross-analysis, data stream sharing and aggregated visualisation across all RADAR-AD solutions, as well as in combination with pre-existing solutions such as those being elaborated under RADAR-CNS.

3 Discussion & Conclusion

The purpose of this deliverable was to lay the foundation of our plan to generate insights into pre-specified conditions for data sharing that [1] help to respect original agreements between data subjects and researchers, [2] uncover site-specific legal, social, financial and ethical conditions for data sharing and [3] identify where additional efforts are needed for the development of a governance framework for international data sharing in health research.

Although there is broad agreement in the research community on the value of data sharing, there are still some challenges associated with data sharing. This deliverable provides a high-level landscape of the common challenges and dimensions in the realm of data harmonization, sharing and efficient use thereof.

Although it has taken tremendous time to obtain data in the ND field, it seems that things are slowly changing. Due to COVID, the importance of sharing data between research projects was reinforced and **there is nowadays a much higher preparedness to share data, with researchers and policymakers to advance the science.**

In a recent publication by Birkenbihl et al.[18], **the current AD landscape** was assessed through investigation and curation of accessible cohort data sets on the data level (rather than relying on metadata and/or literature). Nine of the major clinical cohort study data sets available in the AD field were traced down, accessed, investigated, and compared. This paper comprehensively describes the acquired data and shows which data modalities were found in the data sets as well as their overlaps with other studies. Also, the longitudinal follow-up on the biomarker level was assessed and demonstrated to what extent current AD data are covering the progression of the disease. The content of these data sets was compared with the reported findings of metadata-based approaches. All their results have been made available through ADataViewer (https://adata.scai.fraunhofer.de)[26], an interactive web-portal that allows researchers to explore the AD data landscape generated based on the investigated data sets[6].

It is important to highlight the following obtained learnings from the Neuronet WG:

- When sharing data, it should be clear what role each party has and that each party has the organisational basis to commit to that role.
- To overcome legal hurdles when sharing data between two IMI projects or (third) parties and sectors, legal templates for cross-consortia agreements (developed by FAIRplus) could be used to maximise the value/impact of data generated by IMI projects.
- The social dynamics of sharing come into play at each organisation level or individual, which will have their own set of priorities. At each level, different considerations come into play - privacy concerns, competitive concerns, and intellectual property concerns.
- Besides the motivational barriers, researchers – and in particular, researchers working on clinical studies – also indicate the financial and time cost of data sharing as a key challenge to overcome. In general, clinical researchers do intend to support at least their own ongoing research through those core agreements but the legal and ethical framework in which they operate faces continuous pressures which create pressure and raise hurdles to change. To support researchers, a clear guidance is needed based on

sound ethical principles and there should be a lawful basis for secondary data processing (i.e; data reuse) in GDPR.

- Also time is a factor, that should be considered when sharing data. One example of change over time is the move from snapshot datasets to real time data flows from digital devices. Data sharing and consent agreements suitable for a trial with discrete time points, clear curation and publication steps may not stretch to stream datasets. With data collected from mobile apps over effectively continuous periods, for example, - there will be pressure to use the data outside the original protocol.
- Trust, trustworthiness and credibility are of paramount importance to facilitate sharing in IMI projects – these are crucial elements in the case of consortia, where by definition a degree of sharing and collaboration is implicit in the work plan.
- Up to 70-80% of data management efforts are spent curating (real world) data prior to conducting any analysis. To overcome these challenges, some recommendations were provided by the WG regarding the harmonisation of real-world data (RWD) from diverse sources (e.g. registries and cohorts).
- Enhanced access and sharing typically requires opening information systems in order that data can be accessed and shared. To avoid digital security threats, IT systems should be in place that will allow that data can be accessed and shared.
- Finally, in order to advance much more efficiently, a mindset change in the research community is very much needed – data 'collaboration' will be crucial for future success. Hence, there the term "data collaboration" should be considered instead of "data sharing".

Besides the identified challenges and the obtained learnings from the Neuronet WG, also a clear outline is needed within IMI projects on how to maintain data resources and platforms (assets of their project). Funders and research insitutions should provide support **to sustain data** resources and platforms when their research funding period ends, to ensure that these valuable resources and tools can continue to be used and shared. Therefore, systems should be in place to assure data sharing capabilities survive the originator project.

An example of how to sustain data in an IMI project can be provided for the European Prevention of Alzheimer's Dementia (EPAD) project. To collect a wide range of cognitive, clinical, neuroimaging and biomarker data to help further our understanding of the early stages of Alzheimer's disease, EPAD began a longitudinal cohort study (LCS) that screened more than 2,000 participants. After the end of EPAD project (October 2020), all data from the EPAD LCS study have now been incorporated on the ADDI platform (AD workbench), which will provide even greater value to the global neuroscience research community[27].

Another example is the IMI project eTox, that has accomplished an effective synergic sharing of historical toxicological data within the pharmaceutical industry. It created a series of models to support toxicity prediction. Both data and models are integrated in the platform developed in the project, the **eTOXsys**, which is a powerful system to access the **eTOX** data and the predictive models. The eTOX IMI grant finished on December 2016 and the project entered into its sustainability phase with SME partners leading the commercial exploitation of eTOXsys. A user board with representatives of the different partners oversees the maintenance and exploitation processes[28].

4 References

1. Kalkman et al. (2019) Responsible data sharing in a big datadriven translational research platform: lessons learned. BMC Medical Informatics and Decision Making; 283 (19)
2. Principles and Obstacles for Sharing Data from Environmental Health Research. Workshop Summary (2016)
3. White paper Practical challenge for researchers in data sharing (https://www.springernature.com/gp/open-research/open-data/practical-challenges-white-paper)
4. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229003
5. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101
6. Birkenbihl et al. Evaluating the Alzheimer's disease data landscape. Alzheimers Dement. 2020 Dec 16;6(1):e12102.
7. Data sharing in dementia research - the EU landscape (from Alzheimer Europe)
8. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf
9. https://ec.europa.eu/research/participants/data/ref/h2020/other/mga/jtis/h2020-mga-imi_en.pdf
10. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/other/comm/open-research-europe_official-launch_en.pdf
11. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016).
12. Cost-benefit analysis for FAIR research data : policy recommendations. http://op.europa.eu/en/ publication-detail/-/publication/d3766478-1a09-11e9-8d04-01aa75ed71a1/language-en (2019).
13. https://www.imi-neuronet.org/wp-content/uploads/2020/04/NEURONET_D1.2_Final-3.pdf
14. https://fairplus-project.eu/
15. S. Kalkman. Responsible data sharing in a big data-driven translational research platform: lessons learned. BMC Medical Informatics and Decision Making volume 19, Article number: 283 (2019)
16. https://www.bigdata-heart.eu/
17. https://www.alzheimersdata.org/-/media/files/addi/addi_data_permission_decision_tree.pdf
18. https://datacatalog.elixir-luxembourg.org
19. https://elixir-luxembourg.org/sustainability-data/
20. Badawy, R. et al. Metadata Concepts for Advancing the Use of Digital Health Technologies in Clinical Research. Digit Biomark 2019 Oct 7;3(3):116-132
21. Resnic F. Medical Devices in the Real World. N Engl. J Med 2018 Feb15; 378(7): 595-597
22. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A233%3AFIN
23. Manta C. Digital Measures That Matter to Patients: A Framework to Guide the Selection and Development of Digital Measures of Health. Digit Biomark 2020 Sep 15;4(3):69-77
24. Fecher B. A reputation economy: how individual reward considerations trump systemic arguments for open access to data. Palgrave Communications volume 3, Article number: 17051 (2017) (https://www.nature.com/articles/palcomms201751)
25. Rosenbaum L. Bridging the Data-Sharing Divide — Seeing the Devil in the Details, Not the Other Camp. N Engl J Med 2017; 376:2201-2203 (https://www.nejm.org/doi/full/10.1056/NEJMp1704482?query=featured_data-sharing)

26. https://adata.scai.fraunhofer.de
27. https://www.alzheimersdata.org/
28. https://www.etoxsys.com/