IMI2 821513 – NEURONET

Efficiently Networking European Neurodegeneration Research

WP3 – Tools and Services

D3.2 - First version on guidance tools on data/sample sharing and use

Lead contributor	Lennert Steukers (4 – Janssen)
Lead contributor email	lsteuker@its.jnj.com
Other contributors	Carlos Díaz (1- SYNAPSE) Manuela Rinaldi (4 – Janssen) Angela Bradshaw (3- Alzheimer Europe) Nigel Hughes (4 – Janssen) Pradier Laurent (8- SARD) Rodrigo Barnes (Aridhia) Lavergne Sidonie (Biofortis Mérieux NutriSciences) Niamh Connolly (Royal College of Surgeons in Ireland) Martin Hofmann-Apitius (Fraunhofer)

Due date	30/12/2019
Delivery date	02/06/2020
Deliverable type	R
Dissemination level	PU
DoA version	V2
Date	29/05/2020

Contents

Document history

Version	Date	Description
V0.1	23/03/2020	First draft
V0.2	04/05/2020	Second draft
V0.3	11/05/2020	Third draft
V0.4	19/05/2020	Fourth draft
V0.5	28/05/2020	Fifth draft
V1.0	02/06/2020	Final version

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
8. **SARD**: Sanofi-Aventis Recherche & Développement
9. **PUK**: Parkinson's Disease Society of the United Kingdom LBG

AD: Alzheimer Disease

BD4BO: Big Data for better outcomes

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

ETL: extract, transform and load

EU: European Union

FAIR: findable, accessible, interoperable and re-usable

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

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

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

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 sharing	**Data sharing** is the practice of making research data available to other investigators.
Data standardization	**Data standardization** 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.[1]
Data harmonization	**Data harmonization** 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.[2]
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.[3]

OHDSI	**Observational Health Data Sciences & Informatics** is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All their solutions are open-source.[1]
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.[1]

Publishable summary

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. WP3 “Tools and Services” aims to develop tools and services to support the IMI ND projects in areas where unmet needs have been identified or to address cross-project challenges identified within the past and present portfolio. One of the issues, in fact not limited to the IMI ND portfolio, are best practices around data sharing and re-use of valuable data. Neuronet intends to compile, share learnings, and provide recommendations on data standards, harmonization, and sharing of data to ensure best practice, reduce duplication of effort, and create resources that will be of value to existing and future IMI ND projects. Some learning points will also be useful for IMI projects unrelated to neurodegenerative diseases, both based on past data sets or producing new ones.

This deliverable focuses on developing specific guidance to aid projects on data sharing policies and tools, incentives, value propositions, infrastructural solutions, etc. With the support of the “Data sharing and re-use” WG, Task 3.2 will develop guidelines aimed at facilitating the sharing of and access to data, biological tools, and other materials amongst IMI projects, as well as with other interested researchers at a European and global level.

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 to its maximum. 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. Editors of international medical journals have labeled 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. [4]

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 technical, ethical, legal, financial, cultural and even psychological issues that typically hamper data sharing. Another challenge is the variability in how data are being collected (lack of standardization) and the format of these datasets (lack of harmonization). In some cases, these data exist in paper or PDF format only and, consequently, are difficult to mine, search, interpret, and analyse.

Also within and between various IMI projects (beyond neurodegeneration) similar challenges as listed above have been identified and hamper project progress. Various IMI consortia have been created to address parts of the global challenge of data identification, standardization, harmonization and eventually large-scale community use (e.g. EMIF, EHDEN, FAIRplus…). The aim of the NEURONET WG is to consolidate these learnings and develop guidelines aimed at facilitating the sharing of and access to data, biological tools and other materials amongst IMI projects, and with other interested research programs at a European and global level. We aim to collaborate with similar initiatives, such as the BigData@Heart consortium, to synergize as much as possible. The current deliverable is 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] expose where additional efforts are needed for the development of a governance framework for international data sharing in health research.

2 Methods

2.1 Working groups within NEURONET

As part of its activities, NEURONET has established several thematic WGs to act as cross-project spaces for experts to discuss their experiences, share lessons learnt, align on common issues, debate about 'hot topics' in the field and identify priorities and opportunities for synergy and collaboration across projects, providing NEURONET with expert advice on key topics and areas.

The expected **WG results** are, among others:

- more consistent and informed decision-making,
- improved re-use of results,
- enhanced networking across projects and more exposure of expert knowledge,
- awareness and homogeneous application of standards.

Dynamics within WGs prime free and non-judgemental discussions with the intention of leveraging all the knowledge that is presently scattered in different projects. Conclusions from WG meetings will be elevated to the NEURONET Scientific Coordination Board (SCB), formed by ND project leaders, who in turn may recommend specific actions to be funded through new topics in IMI or via other mechanisms.

2.2 "Data sharing and re-use" NEURONET working group

This WG focuses on developing specific guidance to aid projects on data sharing policies and tools, incentives, value propositions, infrastructural solutions, etc. With the support of the "Data sharing and re-use" WG, Task 3.2 will develop some guidelines aimed at facilitating the sharing of and access to data, biological tools, other materials amongst IMI projects, and other pertinent research programs at a European and global level (*D3.7. Final version on guidance tools on data/sample sharing and use).*

2.2.1 Members of the "Data sharing and re-use" NEURONET working group

The "Data sharing and re-use" WG consists of subject matter experts in data sharing and re-use participating in IMI Neurodegeneration projects (*Table 1*), and NEURONET members (*Table 2*).

Table 1: Subject matter experts in data sharing and re-use

Members	**Academia/EFPIA/small company**	**IMI project**
Rodrigo Barnes	Aridhia	EPAD/AMYPAD
Niamh Connolly	Royal College of Surgeons in Ireland	PD-MITOQUANT
Martin Hofmann-Apitius	Fraunhofer	PHAGO & AETIONOMY
Nigel Hughes	Janssen	EMIF-AD/EHDEN
Sidonie Lavergne	Biofortis Mérieux NutriSciences	SGG
Nikolay Manyakov	Janssen	RADAR-CNS
Andrew Owens	King's College London	RADAR-AD
Andrew Peter McCarthy	Eli Lilly	RADAR-AD
Agustín Ruiz	Fundació ACE	ADAPTED
Pieter Jelle Visser	VUmc & Maastricht University	EMIF-AD
Serge Van der Geyten	Janssen	EPAD
Judi Syson	University of Edinburgh	EPAD

Table 2: NEURONET members

NEURONET Members	**Beneficiary**
Angela Bradshaw	Alzheimer Europe
Carlos Díaz	SYNAPSE
Emma Dodd	Roche
Jean Georges	Alzheimer Europe
Manuela Rinaldi	Janssen
Lennert Steukers	Janssen

2.2.2 Frequency of meetings

Regular quarterly meetings are held for this working group, preferably via teleconference.

- The first TC for the WG was held on the 29th of November 2019.
- A face-to-face WG meeting was organized on the 26th of February 2020 at the J&J offices in Diegem, Belgium.

3 Scope setting of the working group

Based on the first TC (held on the 29th of November 2019) and the first F2F meeting (held on the 26th of February 2020), the scope of this working group was defined and driven by the WG participants. The topics that were identified for this working group are explained below in more detail.

3.1 Data sharing topics to be addressed

Over the years, significant investments by both funders and pharmaceutical companies have created significant amounts of data that could be used to significantly accelerate research, e.g. biomarkers including clinical outcome measures. However, these valuable data resources remain in silos, hard to be searched and accessed by the research community. What the field could significantly benefit from is a set of agreed principles to enable sharing and access to data, taking into consideration all relevant barriers (e.g. General Data Protection Regulation (GDPR), legal, intellectual properties (IP), ethical, regulatory, financial, societal).

3.1.1 Legal (e.g. collaboration agreement)

There are some **legal hurdles** to overcome when sharing data between two IMI projects. Two IMI projects can enter into a **collaboration agreement (legal document)** for the purpose of sharing data between both consortia. As it can take several months to finalize a collaboration agreement, this WG will capture the lessons learnt from these collaboration agreements. Importantly, 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 the data, and which are going to be using 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 those, can help.

Within this WG, NEURONET will also provide templates derived from existing collaboration agreements. Appendix 1 includes the publishable summary of the report on the AMYPAD governance and relationship with EPAD, which is part of Deliverable 1.1 of the IMI AMYPAD project.[5] Appendix 2 includes the publishable summary of the report on the set up of the EPAD/AMYPAD collaboration framework, which is part of Deliverable 4.1 of the IMI AMYPAD project.[6]

3.1.2 Organizational (e.g. "honest broker" model)

The financial aspect of data maintenance is another big problem in IMI projects. Within such projects, it should clearly be defined which assets should be made sustainable and who will be responsible for sustaining them.

Three different patterns should be considered for the information governance models:

- Enterprise: single lead organisation, mostly insiders (staff, affiliates), one set of data sharing rules;
- Consortium: multiple organisations, opt-in, agreement rules for data sharing (e.g. pre-competitive consortium);
- Ecosystem: multiple, overlapping data sharing arrangements (which might be competitive and federated) with a common environment or platform that supports honest brokers' access. An "honest broker" is an entity that keeps multiple sets of private information, but distributes parts of those sets to other entities who should not have access to the entire set, according to permissions specified by the data donors.

We will evaluate in different IMI projects which Information governance models were chosen by the data owners and provide possible solutions for future set-ups, which may serve as time-saving inspiration for new projects.

3.1.3 Technical (e.g. databases, infrastructure)

The WG will also develop some recommendations/decision trees regarding the type of infrastructures that can be used for data sharing and those that can be of use for upcoming IMI projects. Broadly, lessons from projects indicate two basic models: centralised databasing (e.g. Transmart installations in the EMIF-AD project, or the HARMONY project), and federated data access (e.g. PREPAD system in the EPAD project, or the EHDEN project). Both have distinct pros and cons that projects need to be aware of. These considerations are key for further data sharing with other projects or the broader community, and indeed for the scalability of efforts, dependencies, interoperability, maintenance overhead, sustainability, etc.
Actually, one of the key aspects that will have to be evaluated in this WG is whether projects have a sustainability plan in place for the maintenance of the databases, and how infrastructural choices have affected such plans.

3.1.4 Political (e.g. "FAIR" principle)

This WG will evaluate whether data in several IMI projects are "Findable", "Accessible", "Interoperable" and "Reusable", according to the "FAIR" principle – and which are core components of the Data management plan (DMP).

Based on this principle, IMI projects are encouraged to submit a DMP that includes information on[7]:

- the handling of research data during and after the end of the project,
- what data will be collected, processed, and/or generated,
- which methodology and standards will be applied,
- whether data will be shared/made open-access, and
- how data will be curated and preserved (including after the end of the project).

A DMP is required for all projects participating in the extended Open Research Data (ORD) pilot, unless they opt out of it. However, projects that opt out are still encouraged to submit a DMP on a voluntary basis.

Key to make data sets 'findable' is the notion of meta data (see section 3.1.8 below), 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 an important concept – in that sense, transparency in procedures to request and grant data access are paramount, and examples can be found for example in the EPAD project and its Data Research Access Committee.

3.1.5 Data protection (GDPR)

Another important consideration is that legal norms specified for the sharing of personal data for health research have been developed in the European Union (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. Striking in this regard is that the GDPR itself stresses the importance of adherence to ethical standards, when broad consent is put forward as a legal basis for the processing of personal data. For example, Recital 33 of the GDPR states that data subjects should be allowed to give "consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research". In fact, the GDPR actually encourages data controllers to establish self-regulating mechanisms, such as a code of conduct. To foster responsible and sustainable data sharing in translational research platforms, ethical guidance and governance is therefore necessary.[3] The learnings of the data protection issues & GDPR, will be discussed into more detail in the "Patient Privacy and Ethics" NEURONET WG.

3.1.6 Psychological/ Social (e.g. Trust)

When discussing the ownership of data in the WG, both the psychological/social and motivational aspects of data sharing will be addressed. Initial discussions highlight the importance of implementing a 'culture' of sharing (beyond an 'obligation' to share) to maximise effectiveness. 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. Trust is of paramount importance to facilitate sharing – this is a crucial element in the case of consortia, where by definition a degree of sharing and collaboration is implicit in the work plan. In those situations, vicious and virtuous circles can be generated easily, and a "snowball" effect is sometimes observed. The more data are shared, the better the predisposition of others in the group to share.

IMI EHDEN has recently worked out a research quid pro quo for sustainability framework based on trust and relevance (see

Figure 1). Openess, transparency and collaboration will be 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.

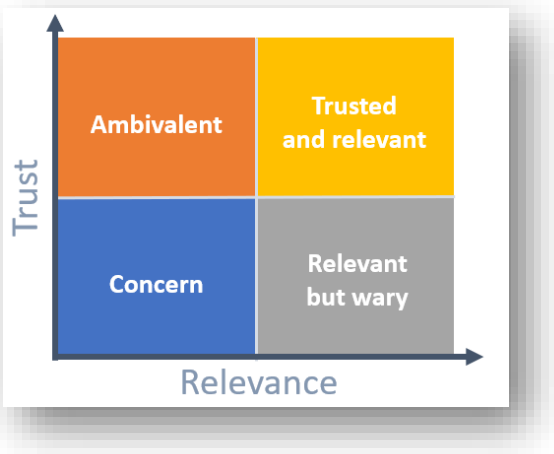

Figure 1: IMI EHDEN has worked out a research quid pro quo for sustainability framework based on trust and relevance.

3.1.7 Ethical

The "Data sharing and re-use " WG will further align this topic with the NEURONET "Patient privacy and Ethics" WG.

3.1.8 Lack of meta data

It is important to identify all existing data that may have resulted and are available from IMI projects, to share high-level information about such datasets and also supporting a metadata-driven catalogue for FAIR data. Several cataloguing initiatives have been developed within IMI neurodegeneration projects (e.g. EMIF Catalogue, ROADMAP Data Cube). These allow to detect 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. Problems associated with these initiatives relate obviously with long-term maintenance, precision of the information contained, and lack of representation power, so it is important that users understand the limitations that may apply. Beyond representing characteristics of the data or the data source, meta data are also important in terms of annotating data with important information that provides context, limitations, etc. Indeed, certain data sets should not be analyzed without taking into account some companion data to avoid misinterpretation of what they mean – leading to potential erros in derived results (e.g. microbiota taxonomic profiles & stool consistency & transit). Shared semantics (a common basis for uniform meta data used across several IMI projects) do exist in neurodegeneration research, but adherence to usage of these metadata templates (ontologies, terminologies, thesauri, data models) is usually poor. On the longer run, the situation might improve if annotations with meta data are done by machines.

3.1.9 Subjective data

Particular attention should be given to sets of subjective data, such as pain levels or certain dietary parameters. The area of patient-reported outcomes is increasingly important and attributed to pervasive use of smart devices and increasing responsibility about one own's health. These data sets present distinct challenges that should be studied in this WG. Lessons learnt by projects like RADAR-AD and RADAR-CNS are bound to be essential to develop any WG recommendations on these particular data.

3.2 Different tools for data sharing based on several identified challenges

Different data sharing tools were discussed in this WG based on the challenges that were identified. An initial framing of data sharing tools was presented as a set of challenges faced by researchers providing or requesting data (

Table 3). It would be useful to the community to understand and incorporate common strategies to address those challenges. The pros and cons of each tool could be evaluated by learning what works and improving or avoiding what doesn't.

Table 3: Different tools for data sharing based on several challenges

Challenge	Context	Tool	Benefit	Drawback
Privacy	Transfer between parties	*Pseudonymisation*	Differential privacy	Risk of re-identification
Participant retention	Where is my data used?	*Consent management, Audit trail*	Link data donation to research and clinical outcomes	Disparate system integration Standardising project metadata
Participant rights	Right to deletion	*Data management platforms*	Respecting the participant	Impact on study Potential conflict due to misunderstood consents
Data sovereignty	Legal territory for data custody	*Federated data sharing*	Record level data stays in situ	Monolithic approaches
Lack of data	Machine learning	*Synthetic data*	Can do 90%? of work before real data use	Tuning fidelity of synthetic data
Lack of metadata	Data reuse	*FAIR data services*	Automated discovery	Manual effort to define

The feedback from the WG was to develop this approach into a knowledge base of tools and strategies that could be documented and shared with the Neuronet community. In order to be useful, this knowledge should expand on the initial framing, so that researchers could find some solutions from different angles. Each solution could be documented as a high level pattern with additional features:

- requirement to be met (challenge, context),

- tool instances/products (specific solutions avaiable in the market or open source),
- relevant user personas (researcher, participant, data controller, service provider),
- sources of funding.

3.3 Data standardization/harmonization

Considerable time, perhaps up to 70-80% of data management efforts, is spent curating (real world) data prior to conducting any analysis. This is further complicated when working with data from multiple sites, in multiple platforms, across multiple languages (both human and machine). As such, data harmonization is about creating a single source of truth, ensuring complementarity of diverse data, removing errors, inconsistencies and aligning on assumptions, syntactic and semantic interoperability. A number of approaches can be used (with varying pros an cons) to harmonize 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 harmonization can impact the subsequent analysis due to the imposition of a specific structure that is not completely adapted to the question researchers want to answer. Conversely, a harmonization effort that is completely developed in a bespoke way around a specific research question may render the effort unusable when further, different research questions want to be posed to the same question. Fidelity of the harmonization, i.e. if there has been any appreciable loss from source to harmonized data, needs to be evaluated to substantiate the veracity of any 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 (extract, transform, and load) can become ELT (extract, load and transform), so transformation can occur prior to analysis from the diverse loaded data in the lake. Intrinsic to the ETL process is one of audit and data hygiene, with collaborative evaluation of a dataset by 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 harmonization and interoperability, for instance within a standardized, modular and extensible collection of data schemas, has gained considerable ground in recent times. The FDA's Sentinel within a shared health data network (SHDN), or the OMOP (Observational Medical Outcomes Partnership) 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 harmonization of diverse data for analytics, in particular and for example, via a standardized analytics stack from OHDSI (Observational Health Data Sciences & Informatics) initiative, utilising the OMOP CDM.

Data harmonization is a necessity in a world where people, systems and structures are increasingly interconnected and interdependent. The recent unfortunate example of the COVID-19 pandemic has shown very clearly how important it is to be able to access sizeable data sets in a reliable, fast way for real-time analysis. Furthermore, in the context of neurological collaborative projects, moving towards a FAIR construct for their data, an agreement on the

harmonization approach is on the critical path 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 on investment. In the short term, it is a socio-technical construct with regards to the need to collaborate, and 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 harmonization 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 are mapping to the OMOP CDM. HARMONY in haematological cancers is mapping data via a pooled (centralised) SHDN; PIONEER in prostate cancer is working on mapping via elements of a pooled SHDN and a federated SHDN, a hybrid model; and EHDEN (European Health Data & Evidence Network) is utilizing a federated or distributed SHDN. The EHDEN project is unique in utilising certified small to medium enterprises (SMEs) to undertake the ETL with so-called Data Partners (institutions holding relevant health data), whilst working symbiotically with OHDSI on methodology, tools and use case development, to create a real-world evidence ecosystem that thrives on its own and is not dependent on the project in the medium to long term.

Specific examples exist in neurology and IMI, such as the EMIF-AD (European Medical Informatics Framework – Alzheimer's Disease) experience, where AD registries were harmonized via a variant of the OMOP CDM, utilising a specific variable set. These experiences 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/harmonization, was also envisaged by EMIF-AD that 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 be still valuable, especially as harmonized datasets.

Other challenges, in particular for semi- and unstructured data that require additional work such as natural language processing or dietary data, also need to be addressed to release even more potential data for study, which will further complexify structured data for harmonization and standardized analytics. Contemporary developments in methods, tools, and resources in working with such data will only increase the resolution of real world data for evidence and insights within neurological disorders and any other disease. Supporting use cases such as machine learning, which are wholly dependent on training and validation sets, can be challenging for certain diseases and populations in terms of availability and in particular representativeness. As such, harmonization and interoperability of diverse datasets will become an even more pressing need.

3.4 Discuss specific usage scenarios

Within this WG, several specific usage scenarios were proposed and will be further discussed in detail during the course of the WG.

3.4.1 Use case #1: Using independent patient-level test sets for model validation.

- Machine learning models are being trained and tested on essentially the same data set (e.g. in Alzheimer's: ADNI).
- It is important to share record level patient data in order to validate (systematically) any "prediction" made on one of the "over-analysed" data sets, such as ADNI. Data sharing is required for **independent validation of trained** models.

3.4.2 Use case #2: Estimating the generalizability of models

- Models / classifiers / signals & patterns identified in a data set can be generalized only if the model / pattern / classifier performs similarly well on a wide spectrum of relevant data.
- The use case #2 is highly related to use case #1, but specifically highlights the need for data sharing if we want to apply the new knowledge / predictions / machine learning (ML) models in the real world.

3.4.3 Use case #3: Establishing global models

- Very often, clinical studies are highly biased (with respect to inclusion and exclusion criteria, variables, and measured assays). Usually, there is not a single study that captures all relevant information (e.g. on co-morbidities, on all concurrent medications and supplements, on all dietary habits) and as a consequence, important information is lost on relevant co-variables (confounders).
- Global models combine and unify variables and measurements across a variety (ideally: all) of relevant clinical studies. Global models allow for complementation of controlled study data by observational real-world-data.
- Establishing "global models" requires **sharing of a wide spectrum of relevant data** that all lack bits and pieces here and there, but bear the potential to establish a "bigger picture" across a huge number of observations of shared data unified in a global model.

3.4.4 Use case #4: Establishing meta-models

- Very often, clinical studies are under-powered. Thousands of potentially “interesting” observations are lost every year because the number of observations are simply too small for a real clinical study.
- **Smaller cohorts** can contribute to the large “**meta-cohort”** and may influence the distribution between parameters in the global cohort. Data sharing of many such smaller cohorts will fit the distributions between variables in the global cohort towards the real (ground truth) distribution. Data sharing in this sense enables sampling at global scale with small sample collections.

3.5 Sample sharing

It will need to be discussed whether sample sharing will be included in this WG or whether a separate WG will be initiated that will specifically focus on this issue.

4 Conclusion

The purpose of this deliverable was to lay-out 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] expose where additional efforts are needed for the development of a governance framework for international data sharing in health research.

This document provides a high-level landscape of the common challenges and dimensions in the realm of data harmonization, sharing and efficient use thereof. In the next few months, the WG will develop, as much as possible, guidelines and recommendations tailored to specific ND issues such as validation of disease models due to lack of appropriate datasets or variety of cognition test batteries and different thresholds of cognition scores used for diagnostic labelling of states leading to study data interoperability issues. As such, our results will not only be of high value to the IMI Neurodegeneration portfolio, but also to any other initiatives that have the ambition of, or rely on establishing Big Data-driven research platforms.

5 Annexes

5.1 Annex I: Publishable summary of D1.1 of the AMYPAD project

The Report on the AMYPAD governance and relationship with EPAD is part of Deliverable 1.1 of the AMYPAD project.[5] The Collaboration agreement in respect of the EPAD and AMYPAD projects is included as an annex of D1.1.

This deliverable aimed to describe the governance structures common to the AMYPAD and EPAD projects, as well as the processes needed for a smooth and dynamic collaboration among partners of both projects. This alignment was especially important considering the multiple connections with EPAD that are embedded in the AMYPAD project.

This was achieved via:

- Joint governance bodies for strategic management.
- Ensuring data flows across projects according to the respective needs, via the Data Oversight Committees of both projects and other appropriate governance bodies.
- Synergistic activities at the executive and management levels in order to share and maximally align procedures, workflows, timelines and resources.
- Joint analysis of dependencies between both projects – particularly regarding the protocol of the EPAD Longitudinal Cohort Study (EPAD WP4), derived logistics (AMYPAD WP2), data collection and analysis (AMYPAD & EPAD WP4 and WP5), ethics and dissemination strategy (AMYPAD & EPAD WP6).
- Setting up of task forces and specific teams across related Work Packages.

These tasks were facilitated by the numerous partners in both Consortia. The specific terms are solidified in a Collaboration Agreement between both projects. This document will complement the Grant Agreement and the Consortium Agreement, providing a legal framework to all the actions, procedures and joint governance bodies described in this deliverable.

For more information: info@amypad.org.

5.2 Annex 2: Publishable summary of D4.1 of the AMYPAD project

The Set-up of the EPAD/AMYPAD Collaboration Framework is part of Deliverable 4.1 of the AMYPAD project.[6]

The aim of this deliverable was to define the EPAD/AMYPAD Collaboration Framework including legal, regulatory, data management, research governance, ethics and funding provisions, thereby ensuring seamless integration of both projects without compromise to the science, timelines and the successful overall delivery of the two projects.

For more information: info@amypad.org.

6 References

(1) https://www.ohdsi.org/data-standardization/

(2) https://www.ehden.eu/

(3) https://eur-lex.europa.eu/eli/reg/2016/679/oj

(4) Kalkman et al. (2019) Responsible data sharing in a big datadriven translational research platform: lessons learned. BMC Medical Informatics and Decision Making; 283 (19)

(5) https://amypad.eu/wp-content/uploads/2017/01/D1.1-AMYPAD.pdf

(6) https://amypad.eu/wp-content/uploads/2017/10/D4.1-AMYPAD.pdf

(7)https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf