IMI2 101034344 – EPND

European Platform for Neurodegenerative Diseases

WP2 – Legal & Ethical Aspects

D2.1 – Data Protection Impact Assessment

Lead contributor (partner organisation)	6 – UNILU		
Other contributors	1 – UM	2 – ULEIC	3 – CHUV
	13 - UMCG	16 – Aridhia	23 – SARD
	25 – UCB	26 – GV	

WP2	
Due date	31/08/2022
Delivery date	August 2022
Delivery type	Report
Dissemination level	Public

Description of Action	
V1.0	15/10/2021

Document History

Version	Date	Status
V1.0	25/05/2022	Draft
V2.0	07/04/2022	WP review
V3.0	23/08/2022	PMT review
V4.0	06/09/2022	Final

Summary

A DPIA is a useful tool for identifying and mitigating privacy risks, demonstrating compliance, and fostering trust in health research and data sharing. We have conducted a community DPIA looking at how data protection principles map to dementia sample and data sharing networks employing the ADWB. This DPIA has the following aims: to develop a shared understanding of the flows of personal data in the EPND context; to provide analysis to support Cohorts to conduct their own DPIAs where needed; and to inform the design and operation of EPND, such as whether it will offer central, federated, or hybrid IT platforms. This exercise aligns closely with other mapping exercises, including the WP1 Framework and WP8 Data Management Plan. Key elements of this exercise are to identify GDPR roles (e.g., controllers), privacy risks, privacy safeguards, and privacy-by-design opportunities for Cohorts and platforms. Partners who implement an instance of the ADWB platform in a particular organisational and technical environment will be expected to complement this DPIA by conducting local security risk assessments (e.g., Aridhia, University of Luxembourg). A future need was identified to develop a general EPND data security framework, as well as more specific guidelines on anonymisation and pseudonymisation. This DPIA is a first version and should be refined over time as the aims and operations of EPND are defined in more detail. This effort should be seen as part of a broader conversation with Cohorts, partners and other stakeholders to determine the aims, structure, processes, and requirements of the EPND.

I. Introduction

The EPND project T2.2 is described as follows: "WP2 will carry out a Data Protection Impact Assessment of the EPND Platform in the first year, to identify possibilities for improvement, which we will implement thereafter (D2.1 – M10 - UNILU (lead), UCB, BBMRI-ERIC). A timely data protection impact assessment (DPIA) will be essential to inform the Phase 1 review, and to demonstrate to cohorts and other stakeholders that the platform respects the fundamental rights of data subjects. The DPIA of the ADWB platform will assess the potential impacts on the fundamental rights of data subjects. It will focus on the platform's novel metadata and record-level search functionalities. The DPIA will consider levels of identifiability, appropriate technical and organisational safeguards, and processes to ensure respect for the rights of data subjects. The DPIA will be designed as a tool to provide confidence to cohorts, as well as a reference template for their local risk assessments. As the exact design, functionality, and governance of the platform will still be under development, this DPIA will need to be a living document that will be refined over the course of the project."

A. What is a DPIA?

The main purpose of a DPIA is to identify, assess, and help address risks to the rights and freedoms of data subjects arising from data processing. As such, DPIA is a risk-focused assessment that complements the controller's overall risk management framework. A DPIA is legally required by the GDPR for ***processing of personal data that is likely to pose a high risk to the fundamental rights of natural persons***. It may also be prudent to conduct a DPIA even where processing does not pose a high risk. In general, the greater the risk to the rights and freedoms of data subjects arising from the intended data processing, the greater the need for carrying out a DPIA. A DPIA can also be seen as a tool for building and demonstrating regulatory compliance.[1] Having a well-articulated DPIA shows the data controller's efforts towards ensuring compliance and may help limit liability in certain cases. Additionally, the data controller may use a DPIA to help foster trust among external stakeholders by, for example, actively engaging relevant groups (especially data subjects) in the process of preparing the DPIA, and/or making parts of the DPIA publicly available.

B. When is a DPIA legally required for a data processing activity?

The GDPR includes some criteria indicating that processing is likely to result in a high risk to the rights and freedoms of natural persons, which would mean a DPIA is required. These criteria are elaborated in the national data protection legislation of Member States, by guidance of the European Data Protection Board (EDPB)[2], as well as in criteria lists developed by national data

[1] *Article 29 Data Protection Working Party "Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679" (2017)*

[2] *EDPB Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research (2.02.2021)*

protection authorities (DPAs). It is the responsibility of the data controller to decide whether a DPIA should be carried out (unless the processing is already covered by an existing DPIA).

The GDPR Article 35(3) provides a non-exhaustive list of criteria where a processing operation is likely to result in high risk. The relevant criteria with the most interpretive uncertainty for EPND is the **processing on a large scale of special categories of data**.

EDPB: The EDPB has endorsed a list of nine criteria for assessing whether processing operations are likely to result in high risks. In general, a DPIA is likely to be required when two or more of these criteria are met. Additionally, the EDPB has endorsed a (non-exhaustive) list of data processing activities not requiring a DPIA. The criteria with the most potential relevance for EPND include:

- Processing of special categories of personal data (e.g., health and genetic data) [generally applicable to EPND]
- Data processed on a large scale [depends]
- Matching or combining datasets from different contexts [depends]
- Data concerning vulnerable data subjects, e.g., children or patients [depends]
- Innovative use or applying new technological or organisational solutions [depends]

(Additional) country-specific guidance: In accordance with the article 35(4) of the GDPR, national supervisors/DPAs establish and make public lists of the (kinds of) processing operations for which a DPIA is required. Optionally, DPAs may also specify the kinds of processing operations that are exempt from a DPIA, in accordance with the article 35(5) of the GDPR. While the lists published by national authorities are similar to, and largely reiterate the criteria prescribed by the EDPB, they may include additional criteria or processing operations not explicitly addressed.

According to Article 35(3), a DPIA is more likely to be required when special categories of data are processed on a large scale. Although the GDPR does not define what constitutes large-scale data processing, WP29 in its guidelines on DPIAs lists the following factors to be taken into account in assessing the scale of processing:

- the number of data subjects concerned, either as a specific number or as a proportion of the relevant population;
- the volume of data and/or the range of different data items being processed;
- the duration, or permanence, of the data processing activity; and
- the geographical extent of the processing activity.

However, as per EDPB guidance,[3] it must be emphasised that the scale of the processing is but one criterion in ascertaining the need for a DPIA. For example, a high-risk processing activity would likely require a DPIA regardless of the scale of processing. Therefore, in practice, the question of scale of the processing can be viewed as secondary, relative to the other questions listed here.

[3] EDPB Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research (2.02.2021).

Where, following a careful consideration of the questions above, it cannot be clearly established whether a DPIA is required, data protection authorities recommend that data controllers adopt a proactive approach and carry out a DPIA nonetheless.

C. DPIA Method

A DPIA is a process encompassing the following components (GDPR Art 35(7)):

1. Systematically describe the envisaged processing operations and their purposes, considering the nature, scope, context and purposes of processing.
2. For the personal data being processed, determine if legal criteria for high-risk processing are met.
3. Assess the necessity and proportionality of processing in relation to the purposes.
4. Assess the risks to the rights and freedoms of the data subjects.
5. Describe the measures envisaged to address the risks.
6. Decide on whether or not to proceed with processing (or if additional consultation is needed), considering the residual risk of processing after implementation of the envisaged mitigation measures.
7. Document all information relevant to the DPIA.

Under certain circumstances, data controllers are obliged to consult representatives of data subjects (where appropriate). This aligns with best practices in biomedical research. The EPND also plans to engage various stakeholders, such as representatives of persons with neurodegenerative disease. These engagement activities will explicitly address data protection issues and the results can be incorporated into future versions of the DPIA. Furthermore, the EPND could explore recommendations about when and how controllers (e.g., Cohorts) may need to be engaged.

Data controllers may also be obliged to consult the relevant supervisory authority in relation to the intended processing. This is the case where after carrying out a DPIA, the controller has identified a significant residual risk associated with the intended processing and the controller is unable to mitigate the risk.[4] This DPIA will reach some general conclusions, and can make recommendations to controllers (e.g., Cohorts) to determine when and how supervisory authorities should be engaged.

D. Challenges for DPIAs of Data Sharing Networks and Platforms

The EPND aims to deploy the Alzheimer's Disease Workbench (ADWB), an IT platform that enables the discovery and sharing of samples and data across an ecosystem of neurodegeneration Cohorts and researchers (and their research organisations). This platform will support a complex processing ecosystem in connection with Cohorts and public/private research organisations. This ecosystem may involve a range of organisations, including biobanks, data repositories, data hubs (nodes of EPND), cloud service providers, sample processing laboratories, bioinformatics tool

[4] GDPR Article 36(1).

developers and standards bodies. This raises three fundamental challenges for ensuring data protection risks are effectively identified and mitigated:

Challenge 1: Who is the Controller (Legally Responsible for DPIAs)?

One threshold issue in the case of scientific research platforms and data sharing networks, is that it may not be clear at the outset ***who is the controller for what aspects of processing***, and therefore responsible for carrying out the DPIA. A traditional DPIA maps processing from the perspective of a single controller. For a network, a DPIA must map processing involving multiple organisations, and must ask a series of "meta" questions at the outset: What organisations or types of organisations are involved in the processing life cycle? What are their GDPR roles (controller/joint controller/processor/third party)? Which actors are legally required to conduct a DPIA (or to assess if they need to conduct one), for what parts of the processing chain? It is not easy to answer even these threshold questions. First, it may be necessary to divide the data lifecycle into separate phases, each with a different purpose and controller (or joint controllers). This is required in complex processing chains, but the divisions between stages and purposes are not always obvious in scientific resource-sharing contexts. Second, there is legal uncertainty over what level of involvement of the ADWB in data governance may trigger controllership.

Challenge 2: Describing the Envisaged Processing Operations and their Purposes.

A second threshold issue is defining what is the set of processing operations that are conducted on the EPND platform. Presumably, the EPND will provide a general-purpose (possibly federated) IT platform, which will enable different forms of collaboration, each with its own purposes and requiring different sets of services and tools. This means that the processing cannot be fully defined at the platform design stage.

Challenge 3: Federated and Hybrid Approaches

A third emerging issue is the increasing organisational and technical complexity of federated networks and IT platforms. At the early stage of designing networks and platforms in EPND, many design choices concerning the setup of the infrastructure are not yet fully clarified. Instances of the IT platform may be hosted in different environments, e.g., ADDI/Aridhia, at the University of Luxembourg, and/or on-premise in the Cohort's local IT environment. A "hybrid" or "flexible" approach is also proposed offering different options to Cohorts. The infrastructure may also be separated into stages across the data lifecycle. For example, a Cohort may establish a local repository for data harmonisation, storage and access, but then will allow users to export the data to a central secure processing environment for their analysis. Or, certain simple statistical analyses will be conducted through a federated analysis with the data staying on-premise at each Cohort, while for other more sophisticated analyses data may be temporarily pooled in a central

environment. Different platform designs may (or may not) affect GDPR role assignments, exacerbating the challenge above.

The aim is not to resolve these uncertainties in this document, but rather to work with them, keeping in mind that as implementations become more concrete, so can future versions of the DPIA.

II. Scope and Aims of the EPND DPIA Framework

Our proposal is to develop a **DPIA framework for the EPND sample/data sharing network** as a whole, rather than focusing narrowly on the activities of a particular support platform. We seek to take advantage of an opportunity to coordinate compliance processes across organisations to sustainably ensure high levels of data protection. This opportunity is provided for by the GDPR, which allows a DPIA to be carried out for a specific data processing activity or multiple related data processing activities. Repeating similar compliance processes across multiple organisations in a consortium or network may lead to duplication of effort without necessarily increasing protection of data subjects. Such a DPIA framework can also capture risks particular to complex processing environments in which multiple organisations are involved in processing personal data, in multiple different roles, such as accountability gaps between organisations.

The scope of the DPIA is focused primarily on the protection of personal data from data subjects (namely patients and research participants of the EPND Cohorts). It also touches on the processing of personal data of platform users (e.g., Cohort team members and researchers) as part of business or operational processes. General data security risks that affect for example scientific integrity or commercially confidential information will also need to be addressed by the EPND, but are outside the scope of this document.

The aims of this DPIA are the following:

1) To assist EPND partners – including Cohorts, platform providers, and research institutions – to understand their personal data flows, associated risks, and appropriate safeguards.
2) To serve as an external communication tool with publics, patient groups, and regulators, providing assurance that data protection is an EPND priority.
3) To serve as a resource to support EPND Cohorts to conduct DPIAs of their sample/data sharing activities and demonstrate their own compliance.
4) To inform the design of the EPND platform, by identifying any key data protection risk implications of different alternatives (e.g., central, federated, hybrid). Where EPND offers different onboarding options to Cohorts, this DPIA can similarly inform Cohorts' choice.

The DPIA focuses on key "stages" of the sample/data sharing lifecycle that are likely to be involved in most EPND use cases: publishing Cohort descriptions, data transfer (to an on-premise

or external IT platform), data transformation/harmonisation, storage, sample/data discovery, access, and analysis. The exact location of data processing or operations involved for each of these stages, as well as their order, may differ depending on the use case and ultimate platform design.

A DPIA in these contexts should be seen as preliminary, and may require several future iterations. It should serve as part of broader information gathering activities about partners, their intended roles and resources, as well as part of a broader conversation about the aims, means, risks, and controls that will comprise the EPND.

We conclude by identifying opportunities to revise this framework over time, as well as to encourage EPND partner organisations to incorporate it into their local risk assessments while elaborating on more concrete details. For example, Cohorts may be able to provide more detail about the specific populations they are working with and data types they plan to provide, based on their particular Member State regulatory framework. Organisations responsible for operating data platforms may be able to perform more detailed technical security risk assessments, providing additional assurance to end users. The EPND may also develop more detailed guidelines and frameworks for data security and pseudonymisation/anonymisation.

III. Description of the Processing

A. Description of Data Types

Cohort-level Information

The EPND will collect and publish information about the participating Cohorts. A majority of this information will be basic administrative information, perhaps describing different entities, such as the hosting institution, the Cohort name, the PI and main contact, the scope of samples, demographic data and clinical data that could be provided, data dictionaries (describing fields and variables), dataset provenance, and access and use conditions. There may also be some aggregate-level statistics provided, describing the data distributions and completeness.

Individual-level and Sample-level data

We make some generic assumptions about the types of biosamples and data that may be shared through 1+MG based on existing landscape surveys.[5] The exact data types and flow may differ depending on the use case. More concrete information about what sample and data are available will be developed in iterative versions of the EPND *Data Management Plan* and the work of WP3

[5] See, for example, Birkenbihl, 2020; https://adata.scai.fraunhofer.de/, https://datacube.roadmap-alzheimer.org/data.php

on minimum data standards. Ideally, information will be obtained directly from the Cohorts about the kinds of samples and data they are planning to make available, which can make this description more concrete.

- Population types: Healthy, Mild Cognitive Impairment, Alzheimer's Disease (and diagnostic criteria), Parkinson's Disease, Dementia Lewy Bodies
- Longitudinal data collection (interval, retention rates)
- Demographic data (sex, age, education, ethnicity)
- Lifestyle
- Data concerning health, including diagnosis, medication use, comorbidities, family history
- Cognition, neuropsychological test results
- Standardised clinical and motor performance rating scales
- Biomarkers (APOE, MMSE, CDR, CDR0SB, Hippocampus, A-beta, t-Tau, p-Tau)
- Genetic data of varying quantity (e.g., from genotype - single genetic biomarkers such as APOE or GBA variants to Whole exome/whole genome data)
- Other omic data (metagenomic/microbiome, proteomic, transcriptomic)
- Sample related information (e.g., storage temperature, quality and quantity of the biological material)
- Imaging (MRI, PET)

The example below provides an overview of the data modalities available at Cohorts participating in various neurodegenerative disease platforms

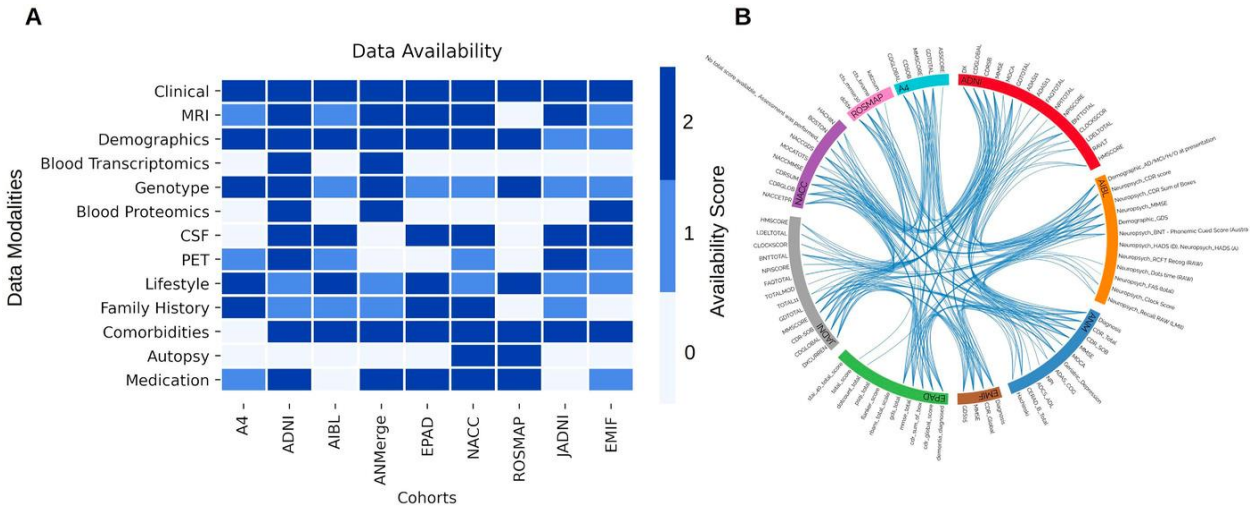

Figure 1. Birkenbihl et al. Evaluating the Alzheimer's disease data landscape. Alzheimers Dement. 2020 Dec 16;6(1):e12102.

It is assumed that within the EPND, additional biomolecular data may be generated over time from biospecimens (plasma, cerebrospinal fluid, and stool) through projects and will constitute part of the Cohort data available.

B. Description of EPND Organisational Framework

The following description is provided in D1.1 – Platform Framework and Roadmap.

"The EPND system will be the technology layer of the European Platform for Neurodegenerative Disease. It is designed to maximise the responsible discovery, access and use of data and samples across the public and private European neurodegenerative research community. The proposed design builds on existing infrastructure such as institutional data banks and biorepositories alongside the AD Workbench (ADWB) operated by Alzheimer's Disease Data Initiative (ADDI). It comprises multiple component services from consortium partners in a federated architecture, with a central entry point for users. A set of interoperability standards allow the system to evolve as new technical components become available."

"[The EPND will] Establish a network that provides central and federated access to high quality samples and data from over 60 cohorts by combining existing data discovery and sharing initiatives made interoperable as part of the AD Workbench. In addition to enabling the sharing and access of data and samples, the platform will provide central and federated capabilities to ensure unified interpretation of biomarker data."

Organisational Roles*	Individual Roles
Host institutions of Cohorts ("Data Provider")	Cohort PI Team members
Host institutions of Researchers ("Research Institution")	Research
Platform hosts (either on-premise with Cohort and/or EPND central (Aridhia/UNILU)).	Data Steward Data Engineer EPND Operator

* Assignment of GDPR roles (controller/processor) to these organisation types is addressed below.

IV. Mapping Data Flows to GDPR Roles/Principles

For illustration purposes, a use case is provided to inform mapping of GDPR roles to a representative data flow.

A. Use Case

- ADWB is hosted in a "trusted cloud" located in the EU/EEA. That is, no international data transfers take place under the GDPR (assume this could be either ADDI/Aridhia/Microsoft Azure or University of Luxembourg).
- Two dementia cohorts deposit pseudonymised data in the ADWB platform for secure hosting (including transformation/standardisation (pre-processing), publication of metadata in data catalogue, storage, remote access and remote analysis).
- A researcher from the consortium identifies data and sample collections of interest in the data catalogue needed to conduct a specific research project. This researcher determines the question to be asked to the data, i.e., the purpose of data processing.
- Each Cohort retains full control over access to its samples and data (decisions made by Cohort's local Data and Sample Access Committee (DSAC)). The researcher sends an access request to ADWB, which is forwarded to the two DSACs. The DSACs both grant access to the relevant samples and data.
- The researcher is granted access to the collected and generated data remotely (direct access) in the ADWB secure research environment which provides tools and security.
- The researcher downloads the results of the analysis.

B. Mapping of GDPR Roles / Accountability for GDPR Principles

	Cohorts (Host Organisation)	**ADWB Platform**	**Researcher (Recipient Organisation)**
Main source/reference of obligations	GDPR, national data protection laws/tissue laws	GDPR, Processor Agreement	GDPR, Data Use Agreement
GDPR role *FOR COLLECTION, SHARING, PUBLICATION OF PERSONAL DATA (e.g., Cohort PIs, research teams) AS PART OF EPND OPERATIONS, FOR THE PURPOSE OF FACILITATING SCIENTIFIC COLLABORATION*	Short term: EPND Consortium (or a representative Partner) Long term: EPND legal entity		
GDPR role *FOR TRANSFORMATION, METADATA PUBLICATION, AND STORAGE FOR THE PURPOSE OF FAIRIFICATION*	X (Each Cohort is sole controller)	Processor on behalf of each Cohort (even if it offers a "take-it-or-leave-it" service)	n/a
GDPR role *FOR PROVIDING ACCESS* (for the purpose of disclosing data for research)	X (Each Cohort makes access decision and is sole controller for the disclosure)	Processor on behalf of each Cohort	

GDPR Role *FOR RESEARCH PROJECT*	Generally = n/a Collaborator = joint control	Processor on behalf of Researcher	Generally = sole controller (unless Cohort(s) are research collaborators, hence acting as joint controller)
Lawfulness (Consent) OR	Arts. 6(1)(a) + 9(2)(a) + rec 33[biobank consent] GDPR	n/a	Arts. 6(1)(a) + 9(2)(a) + rec 33 [project covered by broad consent] GDPR OR Arts. 6(1)(f) + 9(2)(j) GDPR
Lawfulness (Research Exemption)	Arts. 6(1)(e) + 9(2)(j) GDPR (for repository or initial research project) OR Art. 6(4) GDPR compatible further processing (research purpose)* *if legal basis extends	n/a	Arts. 6(1)(f)+9(2)(j) GDPR
Research Ethics Approval	Original cohort approval covers sharing, OR Cohort obtains a general approval for repository/biobank, OR, (in rare cases) REC must approve each access request	n/a	For the project, if required by applicable local norms For the project, if required by a cohort's DSAC as a prerequisite for granting access (In rare cases) Required to seek approval from one or both Cohorts' REC
Purpose Limitation	Access policy Data processing agreement DSAC Review Data use agreement	No use for own purpose Access only to Cohort authorised users	Use for approved purposes only
Data minimisation	Only Safe Data submitted to platform (de-identified, coded) DSAC Review - Only relevant data (#, fields) provided May review Safe outputs	May assist with the following (TBD): structuring datasets in accordance with disease-specific data standards; robust de-identification as a service; double coding service; safe metadata release; Assists with reviewing and assessing Safe outputs	Assists by only requesting relevant samples/data Selects data outputs for extraction/publication
Data retention	Establishes global retention policy (storage in ADWB) Establishes project-specific retention policy	Assists Assists/Enforces	n/a (removal can be technically enforced)
Security	Secure transfer to ADWB	Secure storage/access Authentication and authorisation	Secure connection to platform
Breach Reporting	Obligation, where high-risk, to report to DPA/data subject	Obligation to report to Cohort	Obligation to report to Platform/Cohort

Transparency obligations (vis-a-vis data subjects)	Responsible unless the Art. 14(5)(b) GDPR exemption applies. Must mention: - Dementia research purposes - Types of Data - Recipients: Academic/ Commercial researchers -Recipients: Platform - If legal basis is 6(1)(f), the legitimate interest must be specified	n/a	n/a Exception applies for the research project 14(5)(b)
Accountability - DPIA	Likely. Responsible for DPIA covering provision of access if high-risk processing	May assist w/ template or description of safeguards	Unlikely. Responsible for DPIA of the project if high-risk processing
Accountability - Record keeping	Responsible	May assist with register of access/use Must keep a detailed ROPA (Art 32(2))	Responsible
Data Subject Rights (Point of Contact)	X (Responsible for acting as a direct point of contact)	Direct to Cohort	Direct to Platform/Cohort
Data Subject Rights (Access to Information on Processing)	X (Member State law research exceptions may apply)	Maintains register of access/use	n/a (pseudonymised data only; possibly also research exceptions apply)
Data Subject Rights (Access to Copy of Data)	X (unless Member State law research exception applies)	Assists	May be responsible if retaining newly generated personal data concerning the data subject
Data Subject Rights (Withdrawal/Objection/Erasure)	Responsible (unless research exception applies)	Assists	n/a (withdrawal can be technically enforced)
GDPR International Transfers	Responsible for ensuring transfers take place in a GDPR-compliant manner.	Assists; Must seek permission from cohort to use third country subprocessor under a valid Standard Contractual Clauses agreement.	n/a

A. **Potential GDPR Compliance Services Provided by the EPND**

Purpose Limitation	Documenting consents/lawful basis/REC approvals of Cohorts. Tracking, publication of metadata on access/use conditions. Central Access Review Service: - Check requests fall within authorised purposes.
Data Minimisation	General De-identification service General Double Coding service General De-identification service (metadata publication) Safe Outputs service (record-level search): - For a single Cohort. - Aggregating across multiple cohorts.

	Central Access Review Service: - Determine relevant data for a single cohort. - Determine relevant data for multiple cohorts. - Monitoring for duplication of research projects and encouraging collaboration. Safe Outputs service (analysis results). - For a single Cohort - Aggregating across multiple cohorts.
Retention Period	Retention-period tracking.
Transparency/ Accountability	DPIA Template Security. Central Access Review Service: - Record-keeping (access requests, approved projects)

B. Potential Variations

Here we identify certain foreseeable variations on the use case that might affect the assignment of GDPR roles. The analysis here is not definitive. Some of these variations will be further explored in the WP2 White Paper.

Implications of a central Data and Sample Access Committee (DSAC): All or some Cohorts delegate access review to a central DSAC. The organisation hosting central DSAC may be implicated as a (joint) controller for the disclosure of data to Data Users.

Secure Processing Environments: Where users are required to analyse data in an EPND-certified secure processing environment, the Data User would remain the sole controller, with the platform provider as a processor. This requirement simplifies due diligence and contracting regarding the user's security framework.

Implications of Limiting Access to Query/Algorithm: Users query or send algorithms to the Cohort. Cohorts run on hidden data. This has unclear consequences on controllership.

Federated Approaches: Cohorts require "on-premise" secure processing environments. Data never leaves each Cohort, only aggregate results (that can be deemed anonymous) are exported, so there is no external data processor. Cohort must manage data and keep data secure across the lifecycle. A coordinating centre is needed to receive analysis queries, distribute them to the sites, and to aggregate the results. This has unclear consequences on controllership.

Sample Sharing: Traditionally, samples are sent directly to the user or its designated laboratory. The laboratory is a processor for the data user. Generated data are then sent directly to the data user. Requestors may also request samples from Cohort biobanks. In theory, an analogy to the trusted research environment for biobanking would be a trusted laboratory. The samples are sent to a designated laboratory for processing on behalf of an authorised user. The resulting molecular data is uploaded directly to the ADWB platform by the laboratory.

International Transfers: How will the EPND support collaboration between EU researchers and those in the broader EEA (Norway, potentially UK) as well as third countries (including the USA)? What kinds of sample/data flows will be necessary for these types of collaborations? Does data and platform localisation on EU servers, or in EU organisations mean no GDPR international transfers occur? What GDPR transfer mechanisms are feasible where international transfers are necessary (e.g., Standard Contractual Clauses)? What kind of information needs to be provided to data subjects about such transfers for different tiers of countries, e.g., EU/EEA where GDPR applies, adequate third countries, and other third countries (with a mention of the risks as well as safeguards in place)?

V. Necessity and Data Minimisation

All processing through the EPND will be pursued for the purpose of scientific research to better understand neurodegenerative diseases, and to develop means to better prevent, predict, diagnose, prognose, and treat them.

Data minimisation is a key data protection principle that reduces the personal data processed to what is strictly necessary to achieve these purposes. As a general rule, only anonymised or pseudonymised samples and data shall be processed through EPND. Cohorts are responsible to appropriately pseudonymise samples and data at the earliest possible stage of the lifecycle (e.g., before transfer to a central ADWB platform). Cohorts must also ensure that cohort-level information published through the EPND data catalogue for the purposes of discovery must contain only aggregate, anonymous data (vis-a-vis the participants).

The Cohorts will generally need to maintain a link back to the individual participant's identity. This will generally be necessary for the Cohort's local needs as well as for the proper functioning of the EPND (e.g., to enable re-contact, longitudinal data collection or linkage, data enrichment).

Individual records are generally presumed to be unique, and thus potentially linkable. Discovery of individual-level / sample-level data shall ensure data minimisation of both the "discovery dataset" exposed through search interfaces (e.g., a separate, limited dataset anonymous-in-context), as well as the "export rules" of the API (e.g., only aggregated results).

The Data and Sample Access Committee (DSAC) - whether a central DSAC established by the EPND, or a distributed DSAC hosted by local Cohorts - shall review consistency of project purposes and the scientific relevance of the resources selected as part of the access request.

Access by the user will generally be limited to a secure processing environment, hosted either by a central ADWB platform or on-premise with a local Cohort, limiting the creation of dataset copies.

EPND functionality / stage of processing	Data Minimisation
Publication of cohort-level information (metadata)	- Cohort must ensure cohort-level information - particularly any descriptive statistics - is aggregated and does not allow re-identification of individual data subjects. - EPND should consider identifiability when designing the cohort-level information model. - To this end, generalisation/obfuscation techniques can be applied to the data. - Updates to Cohort statistics should be done at regular intervals to avoid allowing individual data to be inferred from the change between the new and old statistics.

Creation of individual-level/ sample-level data set / transfer to EPND platform (where applicable)	- In principle only pseudonymised samples and data shall be processed through EPND.
	- Cohorts are responsible for ensuring samples and data are appropriately pseudonymised, and for the security of the re-identification mechanism.
	  - Cohorts depositing data in the ADWB platform shall pseudonymise data before the transfer.
	  - Cohorts managing data on-premise shall work with a locally pseudonymised research data set.
	- EPND should provide guidance to Cohorts on appropriate de-identification and pseudonymisation techniques.
	- The AWDB platform should provide assistance with de-identification and (double) pseudonymisation processes.

Data standardisation/ storage	- EPND will develop a Data Model with a common set of data fields/variables describing individuals and samples according to scientific relevance.
	  - Individuals (*demographic, clinical phenotype, interventions*)
	  - Samples (*visit number, timing, method used, sample type, volume/size*)
	- The number of the data fields/variables will likely exceed 100
	- Where possible, the EPND Data Model should establish a controlled vocabulary (to limit processing of unexpected data elements).

Discovery (record-level)	- The EPND will select a subset of fields from the Data Model (approx. 5-15 elements) most relevant for discovery to create a Discovery Dataset Model. - Discovery datasets will be pseudonymised, minimal representations of individuals and samples. - (Safe inputs) Cohorts, with the assistance of the ADWB platform (where applicable), shall create separate discovery datasets that are anonymous-in-context (considering the safeguards in place and who practically has access). Only these discovery datasets will be exposed through API for discovery purposes. - Fields (e.g., age) that are potentially indirectly identifying will be obfuscated (e.g., generalised). - (Safe outputs) Export parameters shall generally only allow outputs in aggregate statistical form, with minimum aggregation thresholds (e.g, rule based, such as n>5 minimum threshold, or through a case-by-case assessment).

	- The EPND shall continuously assess the identifiability of the discovery dataset model, as the model expands over time to enable richer queries responding to community needs.
Data access (including access control by Cohort DSAC or central DSAC)	- The EPND shall provide an end-to-end information management system allowing granular sample/data discovery, selection, request, and drafting of the access agreement (enabling data minimisation). - Cohorts and the EPND shall generally aim to store data in a manner that enables granular access to subsets of data. - Where metadata are unclear, selection could be further refined through negotiation with the Cohort (e.g., by email or by a negotiator platform). - Requestors shall be required to scientifically justify in the project plans their choice of data subjects, data types, and samples. - The DSAC (local or central) shall review project aims, descriptions and selected resources for scientific relevance and feasibility. - The DSAC shall ensure the resources selected, alongside their use conditions and restrictions, are accurately described in the data access agreement.
Data analysis (in on-premise or ADWB secure research environment)	- Platform shall ensure that only the resources described in the access agreement are made available to the Data User. - User access should generally be limited to a secure processing environment, limiting the dissemination of copies of data. - The responsible DSAC or Cohort shall ensure only anonymous data (results) are exported from a secure processing environment (process may be dependent on the type of remote access granted - direct v.s. indirect/algorithmic).
Business processing of personal data by EPND	- Any collection or publication of personal data from Cohort PI/team members or from Requestors/Users must be done based on consent or syndicated from publicly available websites/data catalogues.

VI. Identifying and Assessing Risks to the Fundamental Rights and Freedoms of Data Subjects

Having established appropriate controls for data minimisation and purpose limitation above, the remaining focus is on ensuring the security of the data processed in the context of EPND. Recall that the scope of this DPIA focuses on the risks to the rights and freedoms of data subjects, namely the patients and/or research participants comprising the EPND Cohorts. There are a broader set of risks to data security confronting EPND that are beyond the scope of this exercise. In particular, general cybersecurity considerations are beyond the scope of this DPIA, but are clearly crucial for both the protection of the personal data of data subjects, and the integrity of broader data types (and other resources) as well as the rights and interests of a broader set of stakeholders. Future work in EPND will be done to elaborate such a cybersecurity framework, premised on shared responsibility across Cohorts, platforms, and users. Vis-a-vis data subjects, the primary risk category is loss of confidentiality, with secondary risks relating to data integrity and availability.

Loss of Confidentiality	Any form of unauthorised access to or divulgence of data subjects' personal data. Loss of confidentiality may arise from various sources, including: human error (e.g., an accidental or intentional disclosure of personal data to an unauthorised third party); malicious intent (e.g., cyberattacks, unauthorised reversal of pseudonymisation, re-identification of data subjects by establishing linkages across databases); or data breaches caused by non-human factors (e.g., failure of vulnerable systems and applications). The **likelihood** of multiple events may need to be considered before a harm could be realised: e.g., unauthorised access, individual re-identification (breach of identifiers, linkage with other datasets, attribute inference attacks), and finally a misuse of the compromised personal data. The nature of the event leading to the loss of confidentiality, alongside the amount and sensitivity of the personal data compromised, will determine the **severity** of the event on data subjects. In certain cases, particularly where large amounts of sensitive health and genetic data are accessed without authorisation, re-identified, and misused, individual data subjects may suffer serious consequences, including psychosocial harm, discrimination, and financial repercussions in the areas of insurance and employment. Loss of confidentiality can also lead to potential breaches of data protection principles (e.g., purpose limitation - misuse), or an inability of data subjects to exercise their data subject rights where applicable (e.g., withdrawal, objection, erasure). There is an interplay between data minimisation and the risk of loss of confidentiality. Where data are robustly pseudonymised and/or anonymised (as described in the section above), the likelihood of unauthorised re-identification will be low, even if there is, for example, an unauthorised disclosure or access to the data.

Loss of Data Integrity	Any event that compromises the quality, accuracy or comprehensiveness of personal data. Such events include unwanted data modification, data damage, data deletion, or loss of access to the data. Similar to loss of confidentiality, these events can be caused by different factors such as human error, malicious intent, and system vulnerabilities. Loss of data integrity is a concern for researchers using the data, as it will affect the validity of their findings. Clearly this is an important risk for EPND to consider. However, in this DPIA exercise we focus on the risk of a loss of data integrity *to the data subjects*. In principle, with controls in place for purpose limitation (see above), a loss of data integrity is unlikely to have serious consequences for data subjects, because the data is not generally being used to make decisions that directly affect their rights and interests. An exception is that a loss of integrity of certain data types (pseudonyms) in the EPND may lead to an inability for individuals to exercise their data subject rights where applicable (e.g., withdrawal, objection, erasure). A downstream harm resulting from loss of data integrity in EPND may arise where data processed in the EPND are later accessed by individuals from the Cohort and used for their own (health-related) purposes. Another downstream harm is where data of potential clinical relevance are fed back through EPND to individuals (assuming such a return policy applies), leading to risks that individuals inappropriately rely on inaccurate information. The likelihood of these harms will depend in the first instance on the EPND and Cohort policies regarding individual data access and return of individual findings of clinical relevance (yet to be determined). Where access/return is foreseen, risks of harm may be mitigated by adopting appropriate disclaimers and confirmatory testing.
Loss of Data Availability	Loss of data availability is a general cyber-security concern for the EPND platform, but does not generally present risks to the rights and interests of data subjects in this context. The only exception is that a loss of data availability may lead to an inability for individuals to exercise their data subject rights under the GDPR, where applicable (e.g., withdrawal, objection, erasure).

In the following table, we map these risks to the various processing stages across the sample/data sharing lifecycle, as well as provide baseline risk assessments, proposed controls, and residual risk assessments (once controls are applied). The controls listed will need to be further assessed to ensure they are proportionate and feasible, which may in turn depend on the ultimate design of the EPND network.

EPND functionality / stage of processing	Risk	Source(s) of risk / Controls Applied	Likelihood Before Controls **(After Controls)**	Severity Before Controls **(After Controls)**
Publication of cohort-level information by Cohort/EPND	Loss of Conf	Accidental public release of personal data if cohort-level information (aggregate statistics) is inappropriately aggregated. Updates between cohort level data render newly added data subjects identifiable (e.g., from n=100 to n=101). **Controls: browse-wrap terms of use on EPND data catalogue prohibiting misuse of published business data; add noise; only update catalogue in larger intervals.**	Low **(Low)**	Low **(Low)**
Transfer of data from Cohort to EPND platform (where applicable)	Loss of Conf	Accidental or intentional data breach in transit. **Controls: encryption in transit; data minimisation (see above); API authentication; point-to-point security controls between known parties; default deny rule on firewalls.**	Mod **(Low)**	High **(High)**
Data standardisation and storage (on-premise or ADWB platform)	Loss of Conf	Breach of platform security. Breach of authentication/authorisation controls. Breach of Cohort's access credentials (e.g., password). **Controls: Data processing agreement (Cohort-Platform); encryption at rest; restrict access to authorised individuals in segregated workspaces (Cohort team members or EPND authorised individuals assisting Cohort); access logging and monitoring; Cohort data steward security training; penetration testing; data minimisation (see above).**	Mod **(Low)**	High **(High)**
	Loss of Int	Unauthorised/unintended modification of data. Virus/malware being released through the (local) EPND application (e.g. virtual assistance, data FAIRifier). **Controls: QC/QA processes; limits on data subject access and quality disclaimer; signing/certification of EPND applications.**	Mod **(Low)**	Low **(Low)**
Sample/data discovery (record-level)	Loss of Conf	Inappropriate installation of API. Misconfiguration of the underlying infrastructure (backend components). Breach of authentication/authorisation controls applied to API (where applicable). Multiple queries leading to individual data extraction. **Controls (*Where deemed proportionate based on risk assessment of discovery inputs and outputs)*: API testing; API monitoring; data minimisation of input data; browse-wrap terms of use prohibiting re-identification of individuals; user authentication (w/**	Mod **(Low)**	Low **(Low)**

		public ID) + click-wrap agreement or consortia member White Lists (with activity governed by consortia agreements); logging and monitoring of search queries for suspicious behaviour; multiple query auditing/tracking/denial; DSAC approval of user / project / specific search query; Adding incremental noise to the results.		
Sample/data access (Cohort or central DSAC)	Loss of Conf	Wrong datasets or resource types approved. Fraudulent identity / misleading justification of access provided to the DSAC. **Controls: QC/QA of access/use conditions descriptions; DSAC SOPs; auditing of decisions; access management incident response plan.**	Mod **(Low)**	Low **(Low)**
Data analysis (in Cohort on-premise or central ADWB secure research environment)	Loss of Conf	Breach of platform security. Breach of authentication/authorisation controls. Breach of data user's remote access credentials (e.g., password). Unauthorised data export by user, application, or unauthorised party. **Controls: Data Access/Use Agreement to ensure user respects data protection principles, as well as security and ethical principles (e.g., no re-identification, no export); user access should generally be limited to a secure ADWB processing environment; access logging, auditing, and monitoring; restrict access to authorised individuals (user); user security training; penetration testing (analysis environment); user breach reporting obligations; data minimisation - summary outputs (see above).**	Mod **(Low)**	High **(High)**
	Loss of Int	Unauthorised/unintended modification of data. Analysis application has ransomware, viruses etc. **Controls: Provide users read-only access; prohibit reporting of individual findings of health relevance to data subjects (without confirmatory testing/disclaimers); provide EPND analysis apps (providing standard analyses in which only the query/configuration can be changed); EPND review of external code; certificates for dockers/applications from authorised registries; sandboxing of containers (no network access/limited file access); antivirus software applied to container.**	Mod **(Low)**	Low **(Low)**
Business Processing of Personal Data by EPND	Loss of Conf	Browse-wrap terms of use on EPND data catalogue prohibiting misuse of published business data.		

Conf= Confidentiality; Int = Integrity; DSAC = Data and Sample Access Committee; ADWB = Alzheimer's Disease Work Bench.

Conclusion/Recommendations

The EPND only intends to process data to advance research to better understand, prevent and treat neurodegenerative diseases. The intended processing within the EPND therefore presents limited risks to the rights and interests of data subjects (participants and patients). The EPND must establish safeguards (e.g., data sharing agreements) to demonstrate that processing will be limited to these intended purposes.

The EPND can generally achieve its intended goals by processing only pseudonymised or anonymised data. Cohorts should pseudonymise data before including them in the EPND (or at the earliest opportunity thereafter). Only Cohorts should retain the link back to the individual patient or research participant. The EPND should provide additional guidance and tools to Cohorts on appropriate data pseudonymisation/anonymisation techniques.

Where the EPND processes pseudonymised and thus personal data, it should take advantage of additional data minimisation opportunities, such as access review processes that ensure user access is limited to data fields necessary to answer their research questions.

Users will by default be provided with access to data in a trusted research environment - either hosted by a central platform (e.g., ADDI/Aridhia or University of Luxembourg) or hosted by the individual Cohort (on-premise). Cohorts should only seek to host data on-premise where they have the sophistication and resources to implement the federated appliance while maintaining a robust level of data security.

In the EPND White Paper, WP2 should further consider the GDPR implications of the following EPND design decisions: establishing a central Data and Sample Access Committee (as opposed to leaving decisions to local Cohorts); on-premise hosting of data by Cohorts (as opposed to central hosting); and the possibility of GDPR transfers to third countries (e.g., when using platforms hosted by US cloud providers) and the conditions under which these are appropriate.

Any central platform (such as the ADWB) will generally act as a GDPR Processor, as the platform will be processing data on behalf of Cohorts as controllers for the stages of data transformation and storage. Where a data user pursues a project in the trusted research environment (possibly in collaboration with a Cohort), the data user will generally be a Controller for the analysis and the platform will process data on behalf of the data user as a processor. In the White paper, WP2 should further consider the GDPR roles of Cohorts, the ADWB platform, and users across the sample/data lifecycle, in particular the implications where the EPND establishes a central Data and Sample Access Committee.

The EPND should develop a general data security framework that identifies risks beyond data protection, such as risks to scientifically/commercially confidential information, or to scientific integrity, and establishes controls to protect against those risks. The EPND general data security framework should also establish minimum security requirements selected from the list provided above that are proportionate and feasible considering the costs of implementation and the impact on useability.

Authors: Adrian Thorogood and Davit Chokoshvili (UNILU, LU)

Contributors: Andre Dekker (UM, NL), Anne Bahr (Sanofi, FR), Anthony Brookes (ULEIC, UK), Birgit Schaffhauser (CHUV, CH), Charles Campbell (Aridhia, UK), Hanna Cwiek-Kupczynska (UNILU, LU), Irina-Afrodita Balaur (UNILU, LU), Leslie Weston (Gates Ventures, DE), Mark Jones (UCB, UK), Matt Clement (Gates Ventures, US), Morris Swertz (UMCG, NL), Niranjan Bose (Gates Ventures, US), Pieter Jelle Visser (UM, NL), Philippe Ryvlin (CHUV, CH), Phil Scordis (UCB, UK), Rodrigo Barnes (Aridhia, UK), Stephanie Vos (UM, NL), Venkata Pardhasaradhi Satagopam (UNILU, LU), Vijay Sureshkumar (Gates Ventures, US)