Reliable AI and Data Optimization

D4.1 – Report on the legislations for privacy, IPR, real or synthetic data/metadata

Submission date:

30/06/2025

Funded by the European Union

Project Number:	101135800		
Project Acronym:	RAIDO		
Project Title:	Reliable AI and Data Optimization		
Start date:	January 2024	**Duration:**	36 months
Deliverable: D4.1 – Report on the legislations for privacy, IPR, real or synthetic data/metadata			
Work Package:	WP4		
Lead partner:	VITO		
Author(s):	Burakcan Izmirli (VITO), Elfi Goesaert (VITO), Erik Bollen (VITO) Zisis Batzos (SID) Sofiane Lagraa (FTS), Geoffroy Robin (FTS) Stylianos Klados (ADR) Cédric Clévy (UMLP) Antonis Porichis (THL) Nikolaos Ntampakis (MINDS) Nikolaos Nikoloudakis (PPC)		
Reviewers:	Elfi Goesaert, Raffaella Santucci, Serena Rufino		
Due date:	30/06/2025		
Deliverable Type:	Report (R)	**Dissemination Level:**	Public (PU)
Version number:	v2.0		

Document History

Version	Date	Author	Description
0.1	01/03/2025	Burakcan Izmirli (VITO), Gokhan Ertaylan (VITO)	First release of the Table of Contents.
0.2	01/04/2025	Burakcan Izmirli (VITO)	Finalization of the Table of Contents.
1.0	01/05/2025	Burakcan Izmirli (VITO), Elfi Goesaert (VITO), Erik Bollen (VITO), Zisis Batzos (SID), Sofiane Lagraa (FTS), Geoffroy Robin (FTS), Stylianos Klados (ADR) Cédric Clévy (UMLP), Antonis Porichis (THL), Nikolaos Ntampakis (MINDS), Nikolaos Nikoloudakis (PPC)	Preparation of the first full draft.
1.1	01/06/2025	Elfi Goesaert (VITO), Raffaella Santucci (LOGOS), Serena Rufino (LOGOS), Burakcan Izmirli (VITO)	Completion of internal review.
1.2	07/06/2025	Burakcan Izmirli (VITO), Elfi Goesaert (VITO), Erik Bollen (VITO), Zisis Batzos (SID), Sofiane Lagraa (FTS), Geoffroy Robin (FTS), Stylianos Klados (ADR) Cédric Clévy (UMLP), Antonis Porichis (THL), Nikolaos Ntampakis (MINDS), Nikolaos Nikoloudakis (PPC)	Content revisions following internal review.
1.3	15/06/2025	Burakcan Izmirli (VITO), Elfi Goesaert (VITO), Erik Bollen (VITO), Zisis Batzos (SID), Sofiane Lagraa (FTS), Geoffroy Robin (FTS), Stylianos Klados (ADR) Cédric Clévy (UMLP), Antonis Porichis (THL), Nikolaos Ntampakis (MINDS), Nikolaos Nikoloudakis (PPC)	Finalization of document content.
2.0	25/06/2025	Burakcan Izmirli (VITO)	Final checks and submission preparation.

Table of Contents

Abbreviations

Abbreviation	Full Term
AEPD	Agencia Española de Protección de Datos (Spanish Data Protection Agency)
AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
CJEU	Court of Justice of the European Union
CNIL	Commission Nationale de l'Informatique et des Libertés (French Data Protection Authority)
DA	Data Act
DER	Distributed Energy Resource
DGA	Data Governance Act
DP	Differential Privacy
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EHDS	European Health Data Space
EHR	Electronic Health Record
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reuse
FLOPs	Floating Point Operations
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HDS	Hébergeur de Données de Santé (Health Data Host)
HDPA	Hellenic Data Protection Authority
IP	Intellectual Property
IPR	Intellectual Property Rights
JRC	Joint Research Centre (European Commission)
LLM	Large Language Model
LOPDGDD	Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (Spanish Organic Law on Data Protection)
MR	Méthodologies de Référence (Reference Methodologies)
OT	Operational Technology
PET	Privacy Enhancing Technology

PGx	Pharmacogenomics
PII	Personally Identifiable Information
PU	Public (Dissemination Level)
R	Report (Deliverable Type)
RAG	Retrieval-Augmented Generation
RNN	Recurrent Neural Network
SBPM	Similarity-Based Privacy Metric
SNDS	Système National des Données de Santé (French National Health Data System)
SPE	Secure Processing Environment
TTP	Trusted Third Party
VAE	Variational Autoencoder
WP	Work Package

Disclaimer

Executive Summary

This deliverable provides a comprehensive legal and regulatory assessment for the RAIDO project, establishing a compliance framework for the reuse of data in the development of reliable and regulation-aware artificial intelligence. The report analyses key European legal instruments, including the General Data Protection Regulation (GDPR), the AI Act, the Data Act (DA), and the European Health Data Space (EHDS). It offers a detailed comparative analysis of the national data protection laws and practices in Belgium, France, Spain, and Greece to guide the consortium through the complex landscape of cross-border data processing.

A central finding of this report is the significant legal fragmentation across Member States regarding the use of personal data for scientific research. While the GDPR provides a common foundation, national implementations of its research derogations vary substantially. The analysis reveals key differences in the requirements for technical and organizational safeguards, the role and necessity of regulatory or ethical oversight bodies, and the specific conditions under which data subject rights may be limited. This jurisdictional divergence presents a primary challenge for multi-partner projects like RAIDO, demanding a carefully harmonized approach to data governance.

The report also evaluates the role of Privacy Enhancing Technologies (PET), particularly synthetic data and digital twins, to mitigate privacy risks. It concludes that while these technologies are powerful tools for enabling data-driven innovation, they are not a simple legal solution. Their application requires a careful, risk-based assessment to manage the trade-off between data utility and privacy, and to determine whether the resulting data can be legally considered anonymous or remains personal data under the GDPR. Furthermore, the analysis highlights the forward-looking impact of the new AI Act, which will impose significant obligations on the project's use cases, especially those potentially classified as high-risk.

To navigate these challenges, this deliverable puts forward a set of practical guidelines and a clear compliance strategy for the RAIDO consortium. It recommends a proactive "privacy-by-design" methodology, which includes mapping data flows, clarifying roles, and adopting the "highest common standard" of the relevant national laws. By using established compliance checklists and maintaining robust documentation, these guidelines provide a clear path for the consortium to mitigate legal risks and ensure the project's development process is both ethical and compliant.

1. Introduction

In the evolving landscape of data-driven research and artificial intelligence, the ability to reuse data in a legally compliant, ethically sound, and practically feasible manner has become a critical challenge. The RAIDO project aims to address this challenge by building reliable AI systems that are transparent, privacy-preserving, and regulation-aware across different national contexts. This deliverable, D4.1, focuses on the legal foundations necessary for achieving these goals. It presents a detailed analysis of the European and national legal frameworks that govern the reuse of real, personal, and synthetic data in scientific research settings.

This report begins by mapping the core regulatory instruments that RAIDO must engage with, including the General Data Protection Regulation (GDPR) [1], the Artificial Intelligence Act (AI Act) [2], the Data Act (DA) [3], and the European Health Data Space (EHDS) [4] initiative. These instruments provide the legal scaffolding for how data can be collected, processed, reused, and shared, particularly in cross-border research collaborations. Since each Member State interprets and implements these frameworks differently, the deliverable offers a comparative assessment of national approaches in Belgium, France, Spain, and Greece. This comparative lens allows RAIDO partners to better understand jurisdiction-specific constraints, enabling them to design data governance strategies that are both legally compliant and operationally coherent.

The deliverable also explores the use of synthetic data and digital twins as alternatives to real data, discussing their legal status, privacy benefits, and limitations. By addressing the intellectual property challenges surrounding synthetic data, and by evaluating best practices for privacy-preserving data generation, the report contributes to a deeper understanding of how innovation and compliance can coexist.

Overall, D4.1 provides the RAIDO consortium with the legal clarity and practical recommendations needed to operationalise trustworthy AI across sectors and jurisdictions. It serves as a foundational document that informs not only RAIDO's internal development processes but also the project's contributions to shaping a more harmonized European data space.

2. Legal and Regulatory Assessment of Data Reuse for Scientific Research

Scientific research increasingly depends on the availability and reuse of diverse datasets, many of which include personal or sensitive information. Within RAIDO, ensuring that such reuse is lawful, ethical, and aligned with European and national regulations is essential for enabling trustworthy AI development.

This section provides a comprehensive assessment of the legal frameworks that shape data reuse in scientific contexts. It outlines both the overarching European regulations and also presenting the tools and strategies developed to navigate this complex landscape effectively.

2.1. Regulatory Landscape and Methodology

This subsection outlines the regulatory context and methodological approach adopted to assess legal compliance challenges related to data reuse in RAIDO. It begins by identifying the core European regulatory frameworks that govern the processing of personal, sensitive, and synthetic data for scientific purposes. This legal mapping is essential to ensure that the project's AI systems are developed in accordance with applicable data protection and AI-specific regulations. In addition, the section presents the rationale for developing LexAid-EU, a multilingual, AI-powered legal assistant designed to help interpret and operationalise these frameworks in practice across different national jurisdictions.

2.1.1. Regulatory Frameworks and Rationale

To support its vision of trustworthy and green AI, RAIDO engages with a set of European regulatory frameworks that define how data and artificial intelligence can be lawfully and ethically used. These frameworks provide the foundation for RAIDO's platform design, ensuring that all components, from data curation to model deployment, comply with current legal standards and respect fundamental rights. The selection of frameworks is driven by the project's need to enable cross-border data reuse, safeguard personal data, and develop reliable AI systems that operate within clearly defined legal boundaries.

General Data Protection Regulation (GDPR)

The GDPR is the core legal framework for personal data protection in the European Union. It sets out conditions for the collection, processing, and reuse of personal and sensitive data. For RAIDO, the GDPR is particularly relevant in the context of scientific research. Article 89 [1] introduces the possibility for Member States to apply derogations for research purposes, provided that specific safeguards are implemented. Since national interpretations of these derogations vary, RAIDO partners must navigate different legal landscapes. This deliverable includes a comparative analysis of these national approaches in Belgium, France, Spain, and Greece to support consistent and lawful data reuse.

Artificial Intelligence Act (AI Act)

The AI Act introduces a legal framework for the development and use of artificial intelligence systems based on their level of risk. High-risk AI systems are subject to detailed requirements related to transparency, robustness, data governance, and human oversight. Several RAIDO demonstrators, particularly those in healthcare,

smart systems, and robotics, may fall within the high-risk category. The AI Act informs how RAIDO designs its models and risk mitigation strategies to comply with future regulatory obligations.

Data Act (DA)

The Data Act aims to establish clear rules for accessing and sharing both personal and non-personal data across sectors. It supports fair use and reuse of data generated by devices, platforms, and services. In the context of RAIDO, this regulation helps define the legal basis for sharing and reusing data from decentralized environments, while ensuring that ownership and access rights are respected.

European Health Data Space (EHDS)

The EHDS is a legislative proposal that seeks to create a secure and standardized framework for accessing and reusing electronic health data. One of its focuses is on enhancing secondary use for research, innovation, and public policy. EHDS is particularly relevant to RAIDO's health demonstrator and future-oriented data sharing infrastructure. It emphasizes strong governance, transparency, and patient empowerment in line with RAIDO's approach to responsible AI.

Intellectual Property Rights (IPR) Frameworks

RAIDO also considers the implications of intellectual property rights (IPR) on data access, reuse, and synthetic data generation. This includes database rights, copyright, and data ownership issues [5]. Understanding these legal constraints is essential for managing permissions, licensing, and reuse policies across the project's real and synthetic datasets or results.

2.1.2. LexAid-EU: AI-Powered Framework for Legal and Regulatory Compliance

RAIDO's vision is to enable trustworthy and green AI by providing a comprehensive, integrated platform that addresses all aspects of data and model development from automated data curation and efficient model design to explainability, compliance, and energy-aware deployment. A fundamental requirement for this vision is the ability to ensure that AI systems are legally compliant by design, particularly when operating across jurisdictions with different interpretations of complex regulations such as the GDPR, the AI Act, and the Data Act.

To support this need, LexAid-EU was developed as a multilingual, AI-powered legal assistant that enables consistent, explainable, and verifiable interpretation of European and national legal texts. LexAid-EU was purposefully designed to address the regulatory fragmentation that RAIDO seeks to overcome, particularly in the context of building reliable and regulation-aware AI systems. LexAid-EU is built on a retrieval-augmented generation (RAG) [6] architecture that combines large language models (LLMs) with a legal knowledge base composed of authoritative sources, including EU-

level regulations, national laws, and rulings from the Court of Justice of the EU (CJEU) [7]. This enables the system to answer legal questions with jurisdiction-specific precision, ensuring that all responses are traceable to legal documents rather than relying on unsupported general-purpose AI outputs.

The tool supports RAIDO's objectives by strengthening the platform's compliance-by-design capabilities. It enables partners to examine the legal conditions under which sensitive data may be reused for research, determine whether obligations such as consent or transparency apply in each jurisdiction, and navigate the regulatory landscape related to synthetic data and digital twins. These are all key components of RAIDO's data governance strategy. LexAid-EU currently supports legal corpora in English, Dutch, Greek, French, and Spanish, in line with the jurisdictions and languages represented in the RAIDO consortium. Its modular architecture allows for the continuous integration of newly adopted legislation and evolving legal interpretations. The system is evaluated through a dedicated framework that tracks hallucination rates, contextual accuracy, and multilingual performance to ensure reliable and auditable outputs in legally sensitive contexts.

Ultimately, LexAid-EU reinforces RAIDO's mission to deliver trustworthy AI by providing an explainable and legally grounded reasoning layer within the platform. It enables both technical and non-legal stakeholders to navigate complex regulatory environments confidently and supports the project's broader goal of developing reliable, transparent, and regulation-compliant AI systems with real-world societal value.

2.2. Analysis and Legal Considerations for Data Reuse in Scientific Research

Having outlined the general regulatory landscape, this section now provides a detailed analysis of the specific legal considerations for data reuse in a scientific research context. It examines the conditions for reusing personal data under GDPR, offers a comparative analysis of national laws, explores the legal status of technological alternatives like synthetic data, and concludes with a practical compliance framework for the RAIDO project.

2.2.1. Reusability of Personal Data for Scientific Research Under GDPR Derogations

The foundation for reusing personal data for scientific research within the EU is the General Data Protection Regulation (GDPR). While the GDPR provides a specific pathway for such reuse through a system of derogations, the conditions are complex. This subsection breaks down the key legal requirements for using personal data under these derogations, the cross-border challenges that arise, and the different ways that Member States interpret these rules.

2.2.1.1. Legal Requirements for Reusing Personal Data in Scientific Research

The reuse of personal data for scientific research within RAIDO must follow several regulatory requirements, primarily those set out in the GDPR, but also obligations introduced by other legal instruments such as the AI Act, the Data Act, EHDS, and applicable intellectual property laws. Together, these frameworks define the conditions under which personal data can be reused in a legally and ethically sound way, especially in projects involving cross-border data sharing and artificial intelligence.

Scientific research is explicitly recognized by the GDPR as a valid ground for data processing, provided that certain safeguards are applied. However, when data is reused in the context of AI model development or in sensitive domains like health, additional rules may apply.

The key requirements for lawful reuse of personal data in RAIDO include the following:

Legal Basis for Processing

A valid legal basis under Article 6 of the GDPR must be identified. While consent is possible, it is not always practical in research settings. In many cases, processing based on public interest or legitimate interest is more suitable. When dealing with sensitive data, such as health information, an additional legal condition under Article 9 [1] is needed. Most often, Article 9(2)(j) is used, allowing processing for scientific research provided that safeguards are in place.

Compatible Purpose

Data originally collected for another purpose can be reused for research if the new purpose is compatible. According to Article 5 and Recital 50 of the GDPR, scientific research is usually considered a compatible use, as long as appropriate safeguards under Article 89 [1] are implemented.

In cases where the data are used to train or evaluate AI systems, compatibility must also be considered in light of the AI Act, especially for high-risk systems. The intended use of the data must remain within acceptable legal and ethical boundaries.

Data Minimization

Only the data necessary for the research purpose should be reused. This means selecting only relevant variables, and where possible, applying techniques such as pseudonymization. The AI Act reinforces this by requiring that training data used in high-risk AI systems be representative and free from bias [3].

Safeguards

Article 89(1) of the GDPR requires safeguards to protect data subjects [1]. These may include:

- Pseudonymization or anonymization
- Encryption and access controls

- Data protection impact assessments (DPIAs)
- Role-based access management
- Separation of identifiers from research variables

Such measures are also supported under the AI Act and the EHDS, particularly when sensitive data, such as electronic health records, are reused.

Transparency and Accountability

Even when direct contact with data subjects is not possible, transparency must be considered. Articles 13 and 14 of the GDPR apply, but may be adapted if informing individuals requires disproportionate effort [1]. The Data Act also stresses the importance of transparency in data access, especially where datasets are generated from platforms, devices or third-party sources. All decisions regarding reuse, legal basis, and safeguards should be well documented.

Limitation of Data Subject Rights

The GDPR allows certain rights - such as access, rectification, and objection - to be limited in the context of scientific research, but only where applying those rights would make the research impracticable. These limitations must always be justified and proportionate. National legislation may differ on how these rights are restricted, which is why RAIDO includes a comparative analysis of Member State approaches.

Intellectual Property and Reuse Restrictions

Finally, even when reuse is lawful under data protection law, it must still respect intellectual property rights. This includes database protection, copyright, and licensing terms. The Data Act also defines the rights and obligations of data holders and users, which may affect how RAIDO partners access and combine datasets. These considerations are especially important when using external or synthetic datasets.

Taken together, these requirements ensure that personal data can be reused in a way that supports scientific progress while remaining compliant with legal standards. In RAIDO, they provide the foundation for designing trustworthy AI systems that respect both individual rights and the broader public interest.

2.2.1.2. Challenges in Data Reuse Across RAIDO Jurisdictions

Building on the general principles outlined at 2.2.1.1, GDPR establishes a framework for the protection of personal data within the European Union while recognizing the importance of scientific research through specific derogations. Article 89 [1] of the GDPR provides safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes. However, the implementation of these derogations varies significantly across Member States, creating a complex regulatory landscape for multi-jurisdictional research initiatives such as RAIDO [8].

In practice, researchers engaging in cross-border projects such as RAIDO must interpret and implement these derogations not only in accordance with the GDPR but also within the specific procedural and cultural contexts of each national legal system. This can create uncertainty, especially when data must be transferred or reused across institutional or jurisdictional boundaries.

Belgium

The Act of 30 July 2018 [9] implements GDPR derogations and allows for the processing of special categories of personal data for scientific research when specific conditions are met. These include the implementation of a data management plan, pseudonymization where feasible, and a prohibition on publishing identifiable information without explicit consent. Belgian law is notable for requiring a "necessity test" when derogating from data subject rights, and researchers must document why certain rights (e.g., access or rectification) would hinder the scientific objective.

In collaborative projects, Belgian institutions often expect formal documentation of compliance measures. Ethics committees or data protection officers may request evidence of pseudonymization protocols or ask that a trusted third party handle the re-identification keys, reflecting a strong commitment to both accountability and transparency.

France

The French implementation, through the amended Act No. 78-17 on Information Technology, Data Files and Civil Liberties, [10] establishes a regime requiring prior authorization from the Commission Nationale de l'Informatique et des Libertés (CNIL) [11] for health research that does not meet the criteria for the Méthodologies de Référence (MR) [12]. The French framework imposes additional requirements for data protection impact assessments and emphasizes the role of the designated data protection officer in overseeing research data processing activities.

Spain

Spain's approach, codified in Organic Law 3/2018 on Protection of Personal Data and Guarantee of Digital Rights [13], takes a middle ground, allowing for broader exceptions for scientific research while requiring additional safeguards, including pseudonymization when possible. The Spanish framework places particular emphasis on the principle of data minimization and requires explicit documentation of necessity for any identified data.

Greece

Finally, the Greek implementation through Law 4624/2019 [14] adopts a relatively permissive approach toward scientific research, allowing broader processing of special categories of data for research purposes. This Law permits such processing when adequate safeguards are implemented, particularly emphasizing anonymization and pseudonymization techniques.

These jurisdictional variations present concrete challenges for the RAIDO project. Data reuse protocols must be designed with sufficient flexibility to accommodate differing national interpretations of key provisions. While some partners may require formal involvement of data protection officers or prior approval from ethics committees before accessing datasets, others may permit reuse under more permissive or implicit regimes. Differences also arise in how data subject rights such as access, rectification, or objection can be limited in the context of scientific research [8]. These discrepancies affect not only legal certainty but also practical aspects like project timelines and resource planning. Furthermore, aligning technical safeguards such as pseudonymization standards, data minimization practices, and access controls across multiple jurisdictions adds further complexity.

2.2.1.3. National Interpretations of GDPR Derogations in Scientific Research

The legal basis for national variations is Article 89 of the GDPR. Article 89(1) mandates the use of appropriate safeguards, such as pseudonymisation and data minimisation, when processing personal data for research purposes. Recital 156 clarifies that Member States are permitted to establish additional rules. Building on this, Article 89(2) allows for derogations from specific data subject rights, namely those in Articles 15 (access), 16 (rectification), 18 (restriction), and 21 (objection), but only "insofar as such rights are likely to render impossible or seriously impair" the achievement of the research objectives [1].

The European Data Protection Board (EDPB) [8] and European Data Protection Supervisor (EDPS) [15] have identified several measures that qualify as "appropriate safeguards." These include prior ethical review, technical separation and pseudonymisation, Data Protection Impact Assessments (DPIAs), transparency registers, contractual controls for data transfers, and oversight by an independent supervisory authority.

Belgium

Belgium implements the research exemption regime in Title 4 (Articles 186-208) of its Data Protection Act of 30 July 2018 [9]. The law sets out several substantive conditions. It establishes a three-level data rule, obligating controllers to default to anonymous data where possible, followed by pseudonymised data, and only using identifiable data as a last resort (Art. 197–199). When data is pseudonymised, the re-identification key must be held by an independent trusted third party (TTP) bound by professional secrecy (Art. 198–203). Furthermore, any justification for derogating from data subject rights must be documented in the Article 30 record of processing activities before data collection begins (Art. 191). Derogations from the rights of access, rectification, erasure, and objection are permitted only when exercising these rights would seriously hinder the study. Notably, Belgium does not derogate from the transparency obligations under Articles 13 and 14 of the GDPR, meaning subjects must still receive initial notice that their rights may be limited (Art. 193).

France

In France, the legislative approach delegates significant operational detail to the national data protection authority, the CNIL [11]. The amended Data Protection Act (*Loi Informatique and Libertés*) empowers the CNIL to publish "méthodologies de référence" (MRs) [12], which are reference methodologies that standardise safeguards for recurring research scenarios. For health research, for example, MR-001 (for research with consent) and MR-003 (for research without explicit consent) embed the Article 89 derogations; once a controller signs a conformity declaration, the CNIL treats the research as authorised. For health research specifically, a 2024 practice note clarifies that researchers using the French national health data system (SNDS) [16] must encrypt data extracts and maintain access logs for five years. The rights to access, rectification, restriction, and objection may be limited if the research serves the public interest and the technical and organisational controls specified in the MR are followed, such as pseudonymisation, limited data retention, and DPO oversight. The CNIL retains the power to conduct ex-post audits and can lift the derogation if it finds that safeguards have lapsed

Spain

Spain's Organic Law 3/2018 (LOPDGDD) [13] deems that further processing of data for scientific purposes is compatible with the initial purpose of collection and creates a specific research-only regime for health data that reflects Article 89. Operational details are provided in subsequent decrees, such as Royal Decree 957/2020 [17] for observational drug studies, which makes an ethics committee opinion and a DPIA compulsory. The Spanish framework mandates key safeguards and defines the conditions for limiting rights. Data must be pseudonymised with functional separation, meaning the team holding the re-identification key cannot be the same as the research team (Art. 631-633). Furthermore, a favourable review from an ethics committee is a pre-condition for invoking any derogation from data subject rights (Art. 640(g)). Spain also expressly permits “broad consent” for biomedical research, provided the reuse of data aligns with the initial project’s objectives or is approved by the same ethics committee, pursuant to Article 20 of Ley 14/2007 [18, p. 14].

Greece

Greece transposed the Article 89 derogations through Articles 29–31 of Law 4624/2019 [14]. Article 30 of the law permits restrictions on GDPR Articles 15, 16, 18, 20, and 21 when exercising these rights would obstruct the research and suitable safeguards are in place. The Greek law profile mandates pseudonymisation or statistical aggregation "wherever the purpose can be fulfilled thereby" (Art. 30 §4) and requires controllers to maintain detailed access logs that are available to the Hellenic DPA upon request (Art. 30 §5). A notable national specificity is that children aged 15 and over may provide consent themselves for the processing of their data in a research context. The framework for limiting rights is based on a necessity and proportionality

test; for instance, a derogation is not justified if a data subject's right can be respected by providing only aggregated results. The law also instructs controllers to inform data subjects about the processing "to the extent that this does not frustrate the research" (Art. 30 §3).

Table 1: Summary of National GDPR Derogation Frameworks

Country	Rights Most Often Derogated	Key National Trigger	Mandatory Technical Control
Belgium	15, 16, 18, 21	The controller must justify in the record that the right would "make the research impossible or seriously impair it." Art. 191(2).	Pseudonymisation by a trusted third party; researchers may never hold the key. Art. 198–203.
France	15, 16, 18, 21 (health), plus consent waiver	CNIL "méthodologies de référence" (MR-001 to MR-006) act as pre-approved templates; conformity to an MR replaces need for case-by-case authorisation.	Data-management-plan annex + encryption of the SNDS extract; CNIL can still audit at any time.
Spain	15, 16, 18, 21 (health)	LOPDGDD Art. 19/20 & 631–640: derogation only if data are pseudonymised and an ethics committee has issued a favourable report.	Functional separation between the team that pseudonymises and the team that analyses; mandatory DPIA for re-identification risk.
Greece	15, 16, 18, 20, 21	Law 4624/2019 Art. 30: rights may be limited where exercise would "render impossible or seriously impair" the research and safeguards are "proportionate."	Obligation to minimise data at source and to keep a log of every access (auditable by the Hellenic DPA).

2.2.2. Comparative Analysis of GDPR Derogations Across RAIDO Member States

This comparative analysis aims to provide clarity on how the GDPR's research derogations are implemented across RAIDO Member States. By systematically contrasting the legal and procedural frameworks in Belgium, France, Greece, and Spain, this section highlights the practical implications for researchers who process personal data under divergent national regimes. The comparison also identifies shared patterns and significant divergences in safeguards, oversight, and limitations of data subject rights.

2.2.2.1. Framework for Comparing GDPR Derogations in Member States

This section proposes a structured method to compare how different EU Member States implement the research derogations under Article 89 GDPR. Article 89(2) GDPR allows national laws to limit certain data subject rights (e.g. access, rectification, restriction, objection) for scientific or historical research or statistical purposes, provided appropriate safeguards are in place [19]. The framework below defines key

comparison dimensions, each reflecting a legal or practical aspect of these derogations—and explains why each dimension is critical for analysis. This repeatable framework will enable consistent comparison across countries and help identify where national approaches converge or diverge in balancing data protection with research needs.

To examine the Member State derogations related to research, the following criteria are proposed:

- **Scope of Research Purposes & Public Interest**: This dimension clarifies the scope of "research" (e.g., whether the derogation applies equally to private and public-interest research) to ensure that comparable scenarios are being analysed. Identify how broadly “research” is defined and whether the derogation is limited to research in the public interest. For example, some laws interpret scientific research “in a broad manner” (including privately funded research) [19] while others may require a public interest justification.

- **Data Subject Rights Limitations**: This criterion is proposed to extend and compare the limitations of the subject research rights. This process can showcase how each jurisdiction balances individual rights with research objectives. Determine which data subject rights can be derogated and under what conditions. Article 89(2) explicitly permits derogation from rights of access, rectification, restriction, and objection under strict necessity. Some countries may also limit erasure or portability rights in research contexts (e.g. via related provisions) [19].

- **Required Safeguards**: Safeguards are the cornerstone for allowing reduced rights—comparing requirements (pseudonymization, encryption, access controls, etc.) highlights differences in safety nets for data subjects across jurisdictions. Document the protective measures mandated to justify derogations. GDPR emphasizes techniques like anonymization or pseudonymization as safeguards. National laws often require researchers to use anonymized data if possible, or pseudonymized data otherwise [9] . Some introduce concepts like a “trusted third party” to separate identities from research data.

- **Ethical and Regulatory Oversight**: Review data research approval or oversight mechanisms. Some countries mandate ethical approvals for certain research. For example, the French National Commission on Informatics and Liberty (CNIL) requires that health data research must be pre-approved or conform to CNIL standards [20] Another example is that Spain’s law involves ethics committees for biomedical research [21].

- **Legal Basis and National Conditions**: Understanding the legal frameworks (consent, public interest, legitimate interest, etc.) and the research requirements in each country can identify the exceptions and prerequisites for

their implementation. Additionally, identify any specific legal bases or national laws that facilitate research processing. Certain laws explicitly authorize the processing of sensitive data for research purposes without consent, provided that strict criteria are satisfied (e.g., Greece permits this when the researcher's interest exceeds the data subject's interest, accompanied by safeguards for the subject's rights). Others may necessitate a specific legal statute or public interest directive.

- **Data Minimization and Retention Rules**: Identify legal regulations on data storage and anonymization. Many regimes oblige controllers to anonymize data as soon as the research purpose allows and to keep personal identifiers separate during research [22]. National laws may permit longer retention of research data (for reproducibility) despite the GDPR's storage limitation, but with safeguards. This criterion is suggested based on the premise that research validity must be balanced with privacy on certain types of data.
- **Transparency to Data Subjects**: Check requirements to inform data subjects about research uses and rights. Some laws require notifying individuals if their data will be used for research (e.g. Belgium requires that data subjects are informed upfront if data will be anonymized and which rights might be limited).
- **Publication and Further Use of Research Data**: This criterion is related to the use of data after the research has ended. Comparing legislation across different countries shows how they prevent published findings from unjustifiably exposing personal data. Thus, rules for disseminating research results containing personal data should be investigated.

Those are the proposed criteria for comparing derogation in GDPR between member states, in RAIDO's case: Belgium, France, Greece, and Spain. In the next section, those criteria will be used to examine the differences between those four member states, to reveal the GDPR derogations.

2.2.2.2. Legal Differences in GDPR Derogations Across Belgium, France, Spain, and Greece

In this section, the above framework is applied to compare how Belgium, France, Spain, and Greece have implemented the GDPR's research derogations. Each country's approach is outlined – highlighting national laws (and relevant articles), any sector-specific provisions (especially for health data research), and noteworthy Data Protection Authority (DPA) practices or guidance. The key similarities and differences are then synthesized, supported by a comparative summary table. This analysis provides insight into the fragmentation (or harmony) of research-related privacy rules within the EU and offers practical awareness for the RAIDO project partners conducting cross-border research.

Belgium

Belgium's implementation of Article 89 is encapsulated in its Act of 30 July 2018 [9]. The law establishes a detailed framework for research and statistical processing with strong safeguards:

- **Scope & Basis**: The Belgian law affirms that further processing for scientific or historical research or statistical purposes is deemed compatible with original purposes (per GDPR) if safeguards are applied. Controllers may rely on this compatibility to use data for research without needing a new legal basis, though if special categories are involved, Article 9(2)(j) GDPR (research in public interest with safeguards) is typically invoked.
- **Data Subject Rights Limitations**: Belgium permits derogation from data subject rights in research contexts, but conditions apply. If exercising a right (e.g. access, rectification) would "render impossible or seriously hinder" the research purpose, the controller can restrict that right. Uniquely, Belgian law requires the controller to document the reasons for any such limitation in the record of processing (justifying why allowing the right would impair research). This establishes accountability: a DPA or auditor can subsequently examine the rationale behind the restriction of rights. Belgium's approach does not automatically suspend rights; instead, it implements a necessity test on a case-by-case basis, documented for oversight.
- **Safeguards & Pseudonymization**: Belgian law is strict in requiring data minimization for research. Article 197 of the Act mandates that controllers "shall use anonymous data" for research or statistical purposes whenever possible [9]. Only if the research cannot be achieved with anonymized data may pseudonymised data be used; and only if pseudonymisation is impossible (or would defeat the research) may identifiable data be used. In practice, this means researchers must plan to strip identifiers at the earliest stage. The law also defines a "trusted third party" concept, an independent entity or unit to perform pseudonymisation so that researchers do not see identifying information. The original data controller or a trusted third party must separate identity keys from the dataset, and researchers are prohibited from re-identifying individuals (de-pseudonymising) except if necessary for the research and with Data Protection Officer (DPO) consultation. These measures exceed the GDPR's baseline, effectively operationalizing the "pseudonymise where possible" mandate of Article 89(1).
- **Oversight & Ethical Review**: Belgian law does not require prior authorization from the DPA for research uses of data. Instead, oversight is built into the process via mandatory DPO involvement for high-risk research. If research is likely high risk (e.g. large-scale sensitive data), a DPO must be appointed and a Data Protection Impact Assessment (DPIA) would be required under GDPR. Additionally, sectoral laws (outside the Data Protection Act) may impose ethics

committee approval for certain types of research (especially biomedical research on humans), but those are separate from data protection law. The Belgian DPA has issued guidance emphasizing adherence to the legal framework rather than creating a licensing regime. For instance, in a 2020 case, the Belgian DPA fined an NGO for repurposing public Twitter data for research without proper safeguards; the DPA held that the controller "could not invoke the "scientific research" exception because they failed to pseudonymise the data or implement Article 89 safeguards" [23]. This enforcement stance effectively backs the law's requirements by showing non-compliance has consequences.

- **Publication Rules**: Belgium's law tightly controls dissemination of personal data from research. Controllers may not disseminate non-pseudonymised data from research unless one of a few exceptions applies (e.g. the data subject has consented, or the data was made public by the data subject themselves, or a specific legal provision allows it). Even disseminating pseudonymised data must comply with any stricter conditions in other laws (such as confidentiality of certain archives) [9]. These rules ensure that research findings do not become a backdoor to expose personal information. Typically, research publications in Belgium can include aggregate or anonymized results freely, but publishing any identifiable data (even indirectly identifiable) requires either consent or a strong public-interest justification under law.

Belgium exemplifies a high-safeguard model for research derogations. It leverages detailed statutory obligations (anonymize/pseudonymize by default, trusted third parties, documentation of necessity) rather than case-by-case prior approvals. Data subject rights can be restricted, but only when strictly necessary and with justification on record. The Belgian DPA's practice underscores that controllers must earn the research exemption by demonstrably applying all required safeguards. This rigorous approach aims to enable research (including further use of existing data) while maintaining public trust through strong privacy protections.

France

Formalized control and a separation between general research and the highly regulated health/medical research sector define France's approach for GDPR derogations in research. Along with guidelines and standards published by the French DPA (CNIL), the main legal instruments are the French Data Protection Act (Loi n° 78-17, as amended) and implementing decrees [11].

- **Scope & Sectors**: French law embraces the GDPR's broad definition of scientific research (covering public or private research). However, France created sector-specific regimes: notably, processing of health data for research is subject to additional laws and CNIL regulations. For non-health scientific research, the general GDPR-based rules apply (with possible derogations per Article 89), whereas for health/medical research, special authorization

procedures exist. This split is why CNIL publishes separate guidance for research outside health versus health research [24].

- **Data Subject Rights & Derogations**: Under the French Data Protection Act, data subject rights can be restricted for research purposes in line with GDPR Article 89(2). For example, the law permits that the right to information or access may be deferred or limited if providing it immediately would require disproportionate effort or introduce bias into research, provided that appropriate safeguards are in place. In practice, CNIL has indicated that researchers can rely on these derogations only if they have implemented measures like pseudonymization and if informing data subjects directly would be impracticable or impair the research goals. Nonetheless, France has a strong culture of transparency – even when a formal right is derogated, CNIL often expects some public communication about the research or consultation with ethical committees [24]. Notably, consent is not always required for research use of personal data in France (GDPR allows other bases), but if researchers choose to rely on consent, all usual rights apply. Thus, many research projects in France invoke either "task in public interest" as a legal basis (if backed by public research mandate) or "legitimate interests", coupled with Article 89 safeguards to limit certain rights.

- **CNIL Oversight Mechanisms**: A hallmark of the French system is CNIL's prior involvement in sensitive research. For most medical research using personal data (especially involving health or genetic data), French law requires either: (a) specific authorization from CNIL, or (b) compliance with a CNIL-approved research methodology. CNIL has issued standardized methodologies (referred to as Méthodologies de Référence (MR) that cover common research scenarios. For instance, MR-001 outlines conditions for certain interventional biomedical studies; if a controller pledges to follow MR-001 to the letter, they need not obtain individual CNIL authorization. If no MR covers the study, then the researcher must file an application and CNIL may grant authorization.

- **Safeguards in Practice**: French law and CNIL guidelines stress pseudonymization, encryption, and data minimization as prerequisites for derogations. Controllers should use coded or pseudonymous data whenever possible in research, and separate identifying info. CNIL has underscored the importance of this by correcting researchers who mislabeled data as "anonymized" when it was merely coded, reminding that pseudonymized data is still personal data and requires full GDPR protection. For example, in 2023 CNIL publicly reprimanded two research institutes for failing to conduct DPIAs and for informing patients that their data was anonymized when it was not (it was only pseudonymised). This shows CNIL's expectation: even in research, accuracy in terminology and rigorous safeguards are required. Moreover, a Data Protection Impact Assessment (DPIA) is obligatory for many research projects in France (especially those involving health data) as they likely meet

criteria of large-scale sensitive data processing. CNIL's standard methodologies explicitly require a DPIA as part of compliance. Security measures (access controls, encryption in transit/storage, etc.) are also mandated by sectoral rules (e.g. health data hosts must be approved HDS (Hébergeur de Données de Santé) hosts).

- **Data Subject Engagement**: While rights may be limited, French practice often involves data subjects or their representatives in oversight. For medical studies, ethics committees ensure that participants (or the public, for secondary use of data) are not unfairly kept in the dark. Typically, informed consent is still obtained from participants in interventional studies as per ethical requirements (this is separate from GDPR consent; it's a broader research ethics consent). If consent is permitted for some retrospective studies using existing data, CNIL's authorization will require that patients were informed through some public notice or that their data was used under an approved framework. Additionally, the French Public Health Code may require that data subjects have not objected to the use of their data (an opt-out mechanism) before researchers can rely on the derogations.

- **Retention and Publication**: France allows longer retention of personal data for research than ordinary processing, acknowledging the need for reproducibility of scientific results. The law (via Article 36 of Loi 78-17 and Decree 2019-536) provides that personal data can be kept for as long as necessary for the research, possibly beyond the initial purpose, as an exception to the storage limitation principle [11]. However, CNIL expects justification of duration and periodic re-evaluation of the need to keep identifiable data. Regarding publication, French rules align with ethical norms: published research should not identify individuals unless specifically authorized. Aggregate results, anonymized statistics, or case studies with pseudonyms are the norm. If a researcher ever needed to publish identifiable information (rare in scientific publications), they would likely need the person's consent.

France's approach uses regulatory safeguards and pre-approvals. By requiring DPIAs and CNIL authorization/standards (especially for health data research), France ensures that data subject rights and interests are considered at the design stage of research. The derogations for data subject rights exist (e.g. a researcher might not have to grant an access request immediately if it compromises study integrity), but such situations are typically covered by the protocols approved by CNIL. In essence, France strikes a balance by outsourcing the judgment to CNIL and ethics bodies: if they are satisfied that a project meets legal and ethical criteria (with robust pseudonymization, security, proportionality, etc.), then the project may proceed with certain relaxed obligations.

The cost is more upfront bureaucracy for researchers, but the benefit is a high assurance to data subjects and regulators that derogations will not be abused. This contrasts with Belgium's approach of detailed rules in law; France instead puts weight on ex-ante oversight and documented methodologies to enforce Article 89 conditions.

Spain

Spain implemented GDPR's research provisions in its Organic Law 3/2018 (LOPDGDD) [13] with a focus on facilitating health and scientific research through pseudonymisation and ethical oversight. A key element is the law's Additional Provision 17, which specifically addresses the processing of health data for scientific research.

- **Lawful Basis & Consent**: Under Spanish law, processing personal data for scientific research (especially health/biomedical research) is recognized as an important public interest. The LOPDGDD clarifies that such processing can be done without consent under GDPR Article 9(2)(j), as long as appropriate safeguards (per Article 89(1)) are applied, and the research is duly authorized or reviewed as required by other laws. Spain's approach leans on the idea that pseudonymised data use is lawful for research: The law explicitly states that the use of pseudonymised personal data for public health and scientific research "shall be considered lawful", provided the pseudonymisation is carried out by someone other than the research team [13].
- **Pseudonymisation via Third Party**: The mention of pseudonymisation by a third-party other than the research team, mirrors the Belgian concept of a trusted third party. In practice, for multi-centre studies or use of hospital records for research, Spanish procedures involve an intermediary (e.g., the health authority or a data custodian) who pseudonymises data before researchers access it. The researchers receive data that cannot directly identify individuals, and they commit not to attempt re-identification.

  By structuring data flows this way, Spain ensures the researchers operate on data that is significantly de-identified, which is a cornerstone safeguard in Article 89(1).
- **Ethical Oversight and DPO Involvement**: Spain's sectoral regulations (outside the LOPDGDD but referenced by it) require robust ethical oversight for research on personal data. The Biomedical Research Law 14/2007 and subsequent norms mandate that any research on health data (especially without consent) must be approved by a Research Ethics Committee. LOPDGDD complements this by requiring that such ethics committees include a Data Protection Officer (DPO) when assessing proposals. This ensures data protection considerations (data minimization, security, proportionality) are evaluated alongside the ethical issues. The committee can impose conditions on the research to enhance privacy (e.g., require extra anonymization steps or

follow specific security standards). Additionally, the Spanish Data Protection Agency (AEPD) has endorsed a Code of Conduct for scientific research in recent years [25], which provides practical rules and best practices for researchers handling personal data. Adherence to this code (developed with input from scientific bodies) can demonstrate compliance with both ethical and legal expectations.

- **Data Subject Rights**: In Spain, data subject rights may be limited in research contexts similarly to GDPR's allowances, but the LOPDGDD puts conditions. For instance, if research is conducted with pseudonymised data and re-identification is not possible by the researcher, the rights of access, rectification, etc., might be exercised through the intermediary holding the identifying data. The law appears to allow that data subjects' rights can be satisfied in an indirect manner in such cases, and if the effort to fully honor a right (like telling someone if they are in a dataset) is disproportionate, the right can be restricted provided that the research has public interest and approvals. Importantly, the right to withdraw consent doesn't apply if consent was not the basis in the first place (many Spanish research projects rely on legal permission rather than consent). However, ethically, individuals often are given an opt-out option for use of their medical records in research, and if someone opted out, researchers must respect that.

- **Transparency**: Spanish law requires clear public transparency about research uses of personal data. The LOPDGDD and the science ethics frameworks encourage researchers to publish a notice or registry of research projects (for example, a public registry of clinical studies or data use) so that data subjects or the public are informed that their data might be used. While not every individual is directly contacted (especially if millions of medical records are involved), the research must be documented such that oversight authorities and the public can know it is happening. Additionally, if researchers need to contact individuals (say for additional data or follow-up), they then have to comply with GDPR rights fully.

- **Security and Other Safeguards**: Controllers must implement technical and organizational measures akin to other countries: secure storage, role-based access (only researchers who need to see certain data can see it), and agreements binding researchers to use data only for the specified study. The LOPDGDD's Additional Provision 17 likely sets out criteria for security and pseudonymisation that mirror the Spanish National Security Scheme for health data. The AEPD's guidance also emphasizes that any results of research should preferably be anonymized or aggregated. If individual-level data are published or shared, they should not be directly identifiable.

- **Publication and Further**: Spain allows results to be published as needed for scientific validity, but identifiable personal data should not be published unless consented to by the individual or absolutely necessary (which is rare and would require additional justification). Typically, Spanish researchers publish only anonymized results. If a researcher wanted to publish a participant's quote or case detail, they would either anonymize it or obtain that participant's consent. Moreover, if researchers intend to reuse data from one study for another, they are generally required to seek either a new ethics approval or an amendment to the original approval, ensuring that the new use is compatible with the original consent or authorization. This ties into GDPR's purpose limitation and compatible use test, which Spanish law explicitly addresses: it states that further processing of personal data for scientific research purposes shall not be considered incompatible with the initial purposes, in accordance with Article 5(1)(b) GDPR. This provision in the 2018 law gives legal assurance that data collected for one reason (e.g. clinical care) can be further processed for research, as long as Article 89 safeguards are observed.

Spain's approach is somewhat a middle ground: it does not impose a case-by-case regulatory approval for each study (as France does), but it relies on a combination of legal provisions and ethical oversight to manage derogations. The emphasis on third-party pseudonymisation and involvement of DPOs in research ethics committees shows Spain's focus on practical safeguards. By legally endorsing pseudonymised-data research, Spain facilitates large-scale studies (like in public health, epidemiology) while attempting to shield personally identifying details from researchers. Data subject rights are curtailed only to the extent that the research setup itself prevents identifying individuals easily; if a person does exercise rights (e.g., finds out they were in a study via public info and asks for erasure), the request may be refused during the study if it would ruin the research integrity, but usually the data would be anonymized after the research anyway, addressing such concerns.

Spain's AEPD encourages a proactive stance (codes of conduct, consultation with the DPA for novel situations) to ensure researchers interpret these derogations correctly. In essence, Spain leverages pseudonymisation and ethical governance as the ticket to use the GDPR's research exceptions, thereby safeguarding privacy while enabling valuable research, especially in healthcare.

Greece

Greece implemented GDPR's research derogations through Law 4624/2019 [14], which supplements the GDPR. The Greek law takes a clear stance on when research processing is allowed and how data subject rights are curtailed, essentially embedding a balancing test and strict subsequent anonymization requirement.

- **Permission for Research Processing**: Greek law explicitly allows the processing of special categories of personal data (sensitive data) for scientific or historical research or statistical purposes without consent, under a condition:

the processing must be necessary for the research purpose, and the controller's legitimate interest in performing the research overrides the data subject's interest in privacy. This is effectively a statutory balancing test aligned with GDPR Article 9(2)(j) and Recital 159. It means that if, for example, important research cannot be accomplished using anonymous data, and it addresses a significant public interest, it can proceed even on sensitive data (like health records) provided the individuals' privacy interests are not unduly harmed (and appropriate safeguards are used). Controllers must also implement suitable and specific measures to protect the data subject's legitimate interests [5] in such cases, these measures echo Article 89(1) safeguards.

- **Data Subject Rights Restrictions**: Pursuant to Law 4624/2019, when personal data are processed for research or statistical purposes, data subject rights can be restricted if fulfilling the request would seriously impede the research and if the restriction is necessary to achieve the research purpose. Specifically, the law confirms derogations from the rights of access, rectification, restriction, and objection (Articles 15, 16, 18, 21 GDPR) in line with GDPR, and even notes that the right to data portability (Art. 20) does not apply in archiving/public interest contexts. Notably, Greece adds detail that if providing access "would entail a disproportionate effort" for research, the right of access need not be honored, which is very much in the spirit of GDPR Recital 63 and Article 15 limitations for disproportionate effort. However, these derogations are not absolute: Greek law requires that such limitations only occur "where such rights are likely to render impossible or seriously impair" the research and are necessary for that reason. This is essentially the GDPR test restated.

- **Anonymization Requirement**: A standout provision in the Greek law is the requirement that personal data used for research should be anonymized as soon as the research purposes allow. The law says that unless it conflicts with the data subject's legitimate interests, identifiers should be removed when they are no longer needed for the ongoing research. In the interim, identifying information should be kept separately from the research data (classic pseudonymization) and can only be combined with the dataset if required for research purposes. This means Greek law sets a trajectory: use personal data for research temporarily and move toward anonymization as the project progresses or at least once it concludes. It puts an obligation on controllers to not indefinitely keep data identifiable if it is not needed. This goes beyond GDPR (which implies it but does not explicitly mandate eventual anonymization), highlighting Greece's cautious approach to long-term privacy.

- **Publishing Research Results**: Greece explicitly regulates the publication of research results containing personal data. According to Article 30(4) of Law 4624, a controller may publish personal data from research only if either (a) the data subjects have given their explicit consent, or (b) publication is necessary

for presenting the research results. Even in the latter case, where it is deemed necessary (e.g., quoting a person's statement as part of qualitative research findings), the law requires that the results be pseudonymized before publication. In practice, this ensures that research publications do not include identifiable data unless individuals agree. For instance, publishing a photograph or real name from a study would need consent; otherwise, perhaps a pseudonym or blurred image must be used. This Greek rule underscores respect for individuals' privacy even at the final dissemination stage.

- **Other Safeguards and Practices**: Security measures are mandated by general Greek and EU law (e.g., controllers must implement appropriate security per Article 32 GDPR). In addition, Greek law requires that if a Data Protection Officer (DPO) is designated (for high-risk processing), the DPO should advise on the pseudonymization/anonymization methods used in research. The law entrusts the DPO with giving opinions on the effectiveness of safeguards when research involves high risks or novel techniques. Furthermore, Greece does not have a requirement to seek DPA approval for research projects, but the Hellenic DPA has the power to issue guidance. For example, the HDPA has emphasized that controllers remain accountable for demonstrating how they meet Article 89(1) safeguards. If audited, a research institution should be able to show documentation of why identifying data was needed, how and when it will anonymize them, and how it ensured individuals' rights and freedoms were considered (through a DPIA, for instance, if required by GDPR criteria).

- **Example and Enforcement**: There have not been high-profile fines in Greece solely on research derogations so far, but consider a scenario: if a Greek hospital used patient data for research without consent, Greek law allows it only if the project has public interest merit and anonymization will follow. If the hospital failed to anonymize afterward or used the data for an incompatible purpose, that would violate Law 4624 and GDPR, and the HDPA could sanction it. This ex-post enforcement threat incentivizes compliance even without a prior approval step.

Greece's framework is characterized by clear legal conditions for research use and a forward-looking anonymization mandate. It essentially says: you may use personal data for valuable research and even override some individual rights, but you must protect those individuals' interests through measures and you should erase or anonymize data as soon as you can. The balancing test in the law means researchers and authorities must consider the societal benefit of the research versus privacy intrusion; it is not a free pass.

Compared to other countries, Greece's approach is a bit more explicit in law about the rules (similar to Belgium's detail), and it shares elements with them: like Belgium/Spain, it stresses pseudonymization and eventual anonymization; like France, it demands necessity and proportionality. One can say Greece tries to codify the spirit of GDPR's Recital 156-159 in its national law to avoid ambiguity.

Comparative Analysis and Key Differences

All four countries provide for derogations under Article 89 GDPR, but their implementations vary in strictness, procedure, and emphasis. Below the key similarities and differences across Belgium, France, Spain, and Greece are presented along the dimensions of the framework.

Regarding data subject rights, all four countries allow the restriction of data subject rights (access, rectification, etc.) during research, when necessary, Belgium, Spain, and Greece embed this in law with similar wording to GDPR. France also allows it but manages it through CNIL-approved protocols rather than explicit detailed law provisions. A common thread is that the right to access can be deferred if it imposes disproportionate burdens or invalidate research. Greece and Belgium explicitly mention this proportionality consideration in their statutes. None of the countries absolutely eliminate data subject rights, the derogations are conditional everywhere, preserving the essence of GDPR's protections.

For pseudonymization and anonymization, all four countries mandate safeguards, but Belgium and Greece legislate very granular requirements (e.g. Belgium's stepwise mandate: anonymize or pseudonymize data immediately upon collection, while Greece mandates to anonymize as soon as possible). Spain's law explicitly requires pseudonymization by a party separate from researchers, which is conceptually similar to Belgium's trusted third-party model. France also strongly encourages pseudonymization, and in practice CNIL will not authorize a study that doesn't adequately pseudonymize data. In terms of differences: Belgium and Spain codify an approach requiring "pseudonymisation by a third party"; Greece requires eventual anonymization in law; France uses CNIL guidance to ensure at least pseudonymization, but not via a specific third-party mechanism (institutions themselves implement it, often following CNIL's methodologies).

France is distinct in requiring regulatory or ethical approval (ex-ante) for many research activities (especially health data). Spain and Greece rely on ethics committees by law (especially for health research), but not DPA pre-approval. Belgium does not require prior DPA approval, instead leaning on internal accountability (DPO, documentation) and potential ex-post enforcement.

Regarding national specificities: Belgium introduced the concept of a "trusted third party" to handle pseudonymization keys. France uses Reference Methodologies (MR) to streamline recurring types of research, essentially templates of compliance that, if followed, grant a green light. Spain integrated data protection into its Biomedical Research Law, including requiring a DPO in research ethics committees, showing a

fusion of data protection and bioethics oversight. Greece explicitly ties the derogation to a requirement to override data subject's interest and insists on subsequent anonymization, highlighting a philosophy of temporary use of personal data for research.

In all four, research on health (sensitive) data is permitted but treated with extra caution France and Spain have the most formal requirements for health data (CNIL authorization in FR, ethical approval in ES). Belgium and Greece allow it under Article 9(2)(j) GDPR with safeguards, without a special permit, but Greece explicitly says no consent is needed if the research necessity test is met. Belgium's law similarly doesn't require consent for research if done under these safeguards, though other Belgian sectoral laws (like regarding human subject research) might require informing patients or obtaining ethical consent. All require pseudonymization for health data, while Spain and Belgium even require a separate entity to pseudonymize if possible.

Finally, transparency should also be considered. Belgium requires that when data is collected from data subjects for research, they must be told upfront if the data will be anonymized and that their rights might be limited. France's approach is to ensure data subjects (especially patients) are informed about research uses either through consent processes or public notices. Spain similarly emphasizes transparency via research registers and the information given to patients in healthcare about potential uses of their data. Greece likely follows GDPR's general rules (e.g. if data are not collected directly, the requirement to inform data subjects can be waived under certain conditions, which would align with Article 14(5)(b) if providing the info is impossible or involves disproportionate effort for research). So, while using derogations, controllers still often publish privacy notices or research summaries for accountability.

To sum up, Belgium and Greece have detailed legal frameworks that controllers must follow, whereas France and Spain rely more on administrative and ethical governance to oversee research uses. Nevertheless, the end goals are aligned – ensuring that research can progress for societal benefit while upholding fundamental data protection principles. A notable similarity is the emphasis on pseudonymization in all four countries as a key safeguard, reflecting consensus that pseudonymizing data to the extent possible. These variations underscore the importance for RAIDO's partners to consult local counsel and DPA guidance when conducting cross-border research, what is automatically allowed in one country might require formal approval in another.

2.2.2.3. Summary of Commonalities, Differences, and Legal Challenges

The comparative analysis of GDPR derogations across Belgium, France, Spain and Greece highlights a shared legal commitment to facilitating scientific research through the structured reuse of personal data. All four countries acknowledge that research, particularly when serving the public interest, merits tailored data protection measures that balance individual rights with collective benefit.

However, the practical application of Article 89 GDPR varies significantly across jurisdictions. These differences affect not only legal certainty but also the design and implementation of cross-border projects such as RAIDO.

Across all four Member States, certain foundational principles are consistent. Each legal system permits the limitation of specific data subject rights when necessary for scientific research. The rights to access, rectification, objection and restriction may be curtailed, provided that appropriate safeguards are in place and the limitation is proportionate to the research objective. These safeguards often include pseudonymization, access control, purpose limitation, data minimization, and in some cases anonymization at a later stage. Such protective mechanisms are considered prerequisites for applying derogations and are directly linked to the broader accountability obligations under the GDPR.

Despite these shared principles, the regulatory pathways differ. Belgium relies heavily on statutory precision and internal documentation. Its law requires that pseudonymization be implemented whenever anonymization is not feasible and places strong emphasis on using trusted third parties to separate identifiers. France, on the other hand, builds its research governance on formalized oversight through the CNIL. This includes either case-by-case authorizations or strict adherence to predefined research methodologies. Spain integrates data protection directly into biomedical ethics oversight, requiring third-party pseudonymization and ensuring that data protection officers are involved in ethical review processes. Greece adopts a relatively permissive legal stance by allowing sensitive data to be processed without consent when a legitimate interest in research outweighs individual privacy interests. However, it simultaneously imposes a forward-looking obligation to anonymize personal data as soon as research purposes allow.

These jurisdictional differences produce concrete challenges for RAIDO. For example, the same dataset that may be reused under Belgium's internal safeguards might require prior authorization in France or additional ethical clearance in Spain. Furthermore, the standard of what constitutes sufficient pseudonymization or necessity for derogating from data subject rights is interpreted differently, potentially leading to inconsistent expectations across project partners. Transparency obligations also vary. While Belgium requires that individuals be informed at the outset when data are subject to anonymization and limitations on rights, other countries rely more heavily on public notices or research registries to satisfy information requirements, especially in cases of secondary data use. From an operational standpoint, the divergence in national rules affects timelines, institutional roles, and technical measures. Some partners may need to consult their data protection officers before gaining access to datasets, while others may operate under broader institutional approvals. Aligning these procedures within a unified research framework requires proactive coordination.

To manage these complexities, RAIDO must adopt a harmonized and comparative legal approach. This involves early-stage mapping of national derogation regimes, the use of shared compliance templates, and clear protocols for applying safeguards. It is also essential to ensure that the most stringent national requirement among the participating jurisdictions is respected when processing is intended to occur across borders. Such a “highest common standard” approach offers both legal robustness and operational clarity.

Ultimately, the goal is to foster legal interoperability without compromising the integrity of research or the rights of individuals. The GDPR provides the overarching framework, but its practical implementation is shaped by national law and institutional practice. For cross-border initiatives like RAIDO, legal alignment is not only a matter of compliance but a prerequisite for ethical and sustainable research collaboration.

2.2.3. Synthetic Data and Digital Twins as GDPR-Compliant Alternatives

This section examines the emerging role of synthetic data and digital twins as innovative mechanisms for achieving GDPR compliance while supporting data-driven innovation. Synthetic data offers a method to reproduce statistical patterns of real datasets without directly exposing personal information, while digital twins provide dynamic representations of real-world systems with predictive capabilities. Both technologies promise enhanced privacy protection and operational efficiency but also introduce new challenges related to privacy risk assessment, legal compliance, and data governance. The following subsections explore generation techniques, privacy evaluation frameworks, and legal considerations associated with their deployment in practice.

2.2.3.1. Definition and Role of Synthetic Data and Digital Twins in Privacy Protection

Synthetic data offers a promising path towards reconciling the need for data-driven innovation with the growing demands for privacy. By algorithmically generating artificial datasets that mimic the statistical properties of real data, synthetic data aims to unlock valuable insights without exposing sensitive information [26]. However, achieving this promise requires careful consideration of various challenges, including balancing utility and privacy, navigating complex data structures, and addressing legal and ethical considerations [26]. Rigorous evaluation and robust privacy-preserving techniques are essential to ensure that synthetic data effectively safeguards privacy while enabling meaningful analysis. This section highlights key research addressing the challenges and applications of synthetic data for privacy protection. It explores key aspects of synthetic data for privacy preservation: generation methods, associated risks, evaluation of privacy guarantees, regulatory compliance (e.g. GDPR), and potential applications.

Synthetic Data Generation Approaches

The authors in [27] explore the effectiveness of synthetic data for privacy-preserving clinical risk prediction, focusing on its potential to replace real data in various aspects of clinical model development. The study uses state-of-the-art privacy-preserving generative models to create a synthetic version of ever-smokers from the UK Biobank, and then develops prognostic models for lung cancer using both real and synthetic datasets. The use case of lung cancer risk in ever-smokers from the UK Biobank is leveraged to demonstrate the effectiveness of synthetic data throughout the medical prognostic modeling pipeline, even without eventual access to the real data, and to evaluate the consequences of different data release approaches within the healthcare system.

The authors in [28] propose a Privacy-preserving Generative Adversarial Network (PPGAN) model to address the privacy leakage issue when GANs are applied to private or sensitive data, such as patient medical records. The PPGAN model achieves differential privacy in GANs by adding well-designed noise to the gradient during the model learning procedure and introduces the Moments Accountant strategy to improve the model's stability and compatibility by controlling privacy loss. The use case is generating high-quality synthetic data that retains the required data accessibility. The authors in [29] propose CTAB-GAN+, a GAN designed to generate synthetic tabular data that addresses privacy concerns and regulatory constraints, particularly the GDPR. The work focuses on improving upon state-of-the-art tabular data synthesizers by enhancing data utility in both classification and regression domains, improving training convergence, targeting mixed continuous-categorical variables, and imposing strict privacy guarantees through DP stochastic gradient descent. The use case is generating synthetic tabular data that closely mirrors real data for knowledge development while ensuring data privacy and complying with regulations, particularly in industries such as banking, insurance, and manufacturing.

For example, digital twins have been increasingly used in hospital operations to simulate patient flows and optimize resource allocation, raising important questions about how patient privacy is maintained throughout such simulations.

Privacy Risk Assessment and Quantification Framework for Synthetic Data

The authors in [30] introduce Anonymeter, a statistical framework designed to quantify privacy risks in synthetic tabular datasets, addressing the challenge of sharing sensitive information while preserving privacy. The framework focuses on evaluating the singling out, linkability, and inference risks, which are critical indicators of factual anonymization according to data protection regulations such as the GDPR. The paper demonstrates the effectiveness of Anonymeter through experiments that measure privacy risks in data containing intentional privacy leakages, as well as in synthetic data generated with and without differential privacy, using datasets such as Adults, US Census, and Texas Hospital Discharge Data. On the other hand, the authors in [31] explore whether training with synthetic data truly protects privacy, a claim often

made by methods without formal differential privacy guarantees. The work investigates four training paradigms—coreset selection, dataset distillation, data-free knowledge distillation, and synthetic data generated from diffusion models—and uses membership inference attacks as a privacy auditing method. The use case is to train machine learning models, particularly based on the well-known CIFAR-10 dataset, using synthetic data as a privacy-preserving alternative to training directly on sensitive data, while rigorously evaluating the privacy leakage of these models. The findings caution against relying on empirical approaches without careful and rigorous evaluation, as they can provide a false sense of privacy. The authors in [32] investigate the reliability of similarity-based privacy metrics (SBPMs) used in real-world synthetic data deployments, which aim to guarantee privacy by testing the statistical similarity between synthetic and real data.

The authors demonstrate the inadequacy of SBPMs by providing counter-examples where severe privacy violations occur despite passing the privacy tests. The use case involves generating synthetic tabular data, with the goal of releasing it in a privacy-friendly manner, while the attacks aim to reconstruct sensitive information from the synthetic data, highlighting the vulnerabilities of SBPMs in protecting privacy.

Legal Challenges of Synthetic Data Protection under GDPR

The authors in [33] examine the legal challenges surrounding synthetic data protection, particularly focusing on the GDPR. The paper discusses the potential of synthetic data, generated through machine learning algorithms from original real-world data, to provide privacy-preserving alternatives to traditional data sources. The use case involves leveraging synthetic data to maintain the statistical characteristics of real-world data while ensuring confidentiality and privacy, addressing the limitations of the personal/non-personal dualist approach under the GDPR and advocating for clear guidelines that prioritize transparency, accountability, and fairness in the generation and processing of synthetic data. Finally, the existing research works explore the use of synthetic data as a privacy-preserving alternative to real data across various applications, including clinical risk prediction, tabular data sharing, medical record synthesis, and AI model training, while emphasizing the critical need for rigorous privacy evaluations. They highlight the limitations of similarity-based metrics in guaranteeing privacy and advocate for the adoption of differential privacy and other robust techniques to mitigate the risk of sensitive information leakage. Important challenges include optimizing the trade-off between data utility and privacy, addressing the complexity of real-world data, navigating legal and ethical considerations, and managing computational costs associated with synthetic data generation and evaluation.

The findings caution against relying on empirical methods without careful scrutiny. They also emphasize the importance of transparency, accountability, and fairness in the development and deployment of synthetic data solutions.

Digital Twins in Privacy Protections

In parallel, digital twins are virtual replicas of physical entities, such as machines, systems, or even human beings, that provide real-time data insights and predictive capabilities. The use of digital twins involves significant privacy protection challenges, as they often process vast amounts of personal and sensitive data. Ensuring data privacy in digital twins requires compliance with regulations such as the GDPR, which mandates principles such as data minimization, purpose limitation, and security measures [34] [35]. Organizations must implement robust security protocols, including encryption and multi-factor authentication, to safeguard the integrity, confidentiality, and availability of data within digital twins. In fact, the authors in [35] present a systematic literature review of privacy and security challenges in digital twin implementations, employing the PRISMA methodology to analyse 47 publications. The study identifies six primary groups of challenges: data privacy, data security, data management, data infrastructure and standardization, ethical and moral issues, and legal and social issues. These ethical and moral issues include dilemmas such as balancing transparency with security, ensuring algorithmic fairness, and addressing concerns about autonomy when decisions are increasingly delegated to AI-driven digital twin systems. The use case is to provide diverse insights from research and industry about major digital twin challenges, with emphasis on privacy and security. It aims to offer a better understanding of these concerns to researchers and practitioners. This enables them to proactively address these issues in the development and deployment of digital twin technologies across various sectors, such as emerging technologies, construction, healthcare, and manufacturing. A survey was proposed in [36] to analyse the cybersecurity threats to digital twins in advanced manufacturing, focusing on data collection, data sharing, machine learning/deep learning, and system-level security and privacy. It explores vulnerabilities arising from the integration of technologies like cyber-physical systems and the Industrial Internet of Things.

The use case is to provide solutions to these threats, aiming to establish greater trust in digital twins by addressing security and privacy concerns across various stages of their lifecycle, including model updates, decision-making processes, and uncertainty quantification to enable reliable and secure decision-making in advanced manufacturing and resilient supply chain operations.

While digital twins offer unmatched accuracy for simulating and enhancing physical systems, the paper in [37] acknowledges the serious security issues arising from the combination of these technologies. The use case is to investigate the security and privacy challenges associated with digital twins technology, emphasizing the need for authenticity, privacy, and accessibility to protect virtual representations and ensure the secure and dependable functioning of digital twins systems, ultimately promoting their wider adoption. In [38] Privacy Enhancing Technologies (PETs) are used in the design of Digital Twins for smart cities to address the challenging privacy issues introduced by data-rich models. The use case is to ensure essential data protection requirements

are met by converging privacy preservation mechanisms with digital twins as part of the initial system design, mitigating privacy-based challenges while maintaining the value and utility of digital twins in optimizing urban ecosystems, improving healthcare, and enhancing decision-making.

2.2.3.2. Compliance and Legal Considerations for Synthetic Data

This section presents the status of synthetic data under the GDPR and provides guidance on compliance when using or generating synthetic data. Synthetic data refers to data artificially generated, often using models or algorithms trained on real data that reproduce the statistical characteristics or structure of real-world data without directly copying it [39]. RAIDO, with its focus on data-driven research, may leverage synthetic data to augment real datasets or protect individual privacy.

Here, whether and when synthetic data is considered “personal data” under GDPR is analysed, relevant positions from regulators and scholars are presented, and best practices to ensure legal compliance and effective data protection are recommended. The goal is to understand topics like anonymization thresholds, with practical guidance for RAIDO’s technical partners who might develop or use synthetic datasets.

GDPR and Synthetic Data

A central question is whether synthetic data is subject to GDPR at all. GDPR only applies to personal data, defined as any information relating to an identified or identifiable person. Properly generated synthetic data may be so far removed from the original individuals that it is no longer related to an identifiable person, thus making it anonymous data, which falls outside GDPR’s scope. However, not all synthetic data achieves this level of anonymity.

If synthetic data is generated in such a way that individuals from the original dataset can no longer be identified, even indirectly, in the synthetic data, then the synthetic data is effectively anonymized. Data are only anonymous if there is an insignificant risk of re-identification [40]. In the context of AI models and synthetic data, the European Data Protection Board (EDPB) recently reiterated that for outputs to be considered anonymous, one must ensure there is no reasonable means to link them back to any real person. For example, if RAIDO generates a synthetic dataset of smart meter readings, and no one can link those synthetic readings back to a specific household or individual with any plausible effort, then the synthetic dataset is not personal data. In such a case, GDPR would not apply to the synthetic data itself, though it would apply to the generation process, as discussed below.

If synthetic data contains patterns or remnants that allow linkage to individuals, then it remains personal data. A common pitfall is overly accurate synthetic data that unintentionally embeds unique combinations from real records, hence making it possible to trace that record back to the real person. For instance, if a synthetic health record happens to match a real patient’s rare disease and birthdate, someone with background knowledge might guess the identity. In such cases, the synthetic data is

essentially pseudonymized data, the direct identifiers are fake, but an individual is still "identifiable" through correlation with outside information. GDPR treats pseudonymized data as still personal data. Consequently, if synthetic data is not properly anonymized, it must be assumed it is within GDPR's scope and handled accordingly.

Sometimes synthetic data is generated from a model trained on personal data, and the model might inadvertently memorize certain original data points, a process named model memorization [41]. If those data points resurface in synthetic output, that output is personal data. Even absent memorization, one must consider attribute inference. Regulators have noted that even derived or aggregated data can be personal if it still allows singling out individuals. Thus, RAIDO should treat synthetic data with caution unless a rigorous evaluation confirms anonymization.

In summary, synthetic data can fall outside GDPR if it is truly anonymized. However, given the high standard for irreversibility, many synthetic datasets will still be treated as personal data. RAIDO should assume GDPR applies unless and until proven otherwise through robust anonymization evidence. Next, we consider the generation process and use cases, which also bring legal considerations.

European Data Protection Board and Academic Perspectives on Synthetic Data

Over the past few years, data protection authorities and scholars have increasingly discussed synthetic data as a tool for privacy-preserving data use.

The GDPR's Recital 26 encourages anonymization as a way to use data without privacy risk [1]. The Article 29 Working Party (predecessor to European Data Protection Board EDPB) in its landmark 2014 Opinion on Anonymisation Techniques considered data generation methods, like random data generation, as one way to anonymize. More recently, the EDPS on synthetic data highlights that synthetic data generation can be a form of data protection by design, allowing organizations to share or analyse data without using real personal data. This supportive view sees synthetic data as a promising privacy-enhancing technology when properly applied.

The EDPB in 2024 (Opinion 28/2024 on anonymization in the AI context) stress that controllers often overestimate the anonymity of synthetic data. Moreover, the same reference states that high-utility synthetic data, which preserves granular correlations, may carry a higher re-identification risk. In an opposite way, synthetic data that is highly randomized protects privacy better but may be less useful for analysis. In addition, the EDPB considered scenarios of training AI on personal data and later using the model outputs. The EDPB clarified that if personal data is effectively anonymized during model training (e.g., the model no longer contains personal data and cannot output personal data), the GDPR would not apply to the model's outputs. Nonetheless, if any personal data is retained within the model (even implicitly), the outputs may still be personal. In the context of synthetic data, this implies that if RAIDO trains an AI, such as a Generative Adversarial Network (GAN), on personal data to generate a synthetic dataset, one must evaluate whether the model has neglected the

personal data or is instead retaining and reproducing it. The EDPB establishes scenario-based evaluations, as referenced in the opinion, to direct this assessment.

Compliance Best Practices and Legal Recommendations

This section presents best practices identified from the state of the art regarding the data protection principles and to maximize privacy. Moreover, legal recommendations are also presented. The practices and recommendations are derived from research of the state of the art regarding data anonymization and synthetic data and will be presented in broader categories.

- **Consider DPIAs for Synthetic Data Generation**

Whenever RAIDO generates synthetic data from personal data, they should be treated as a high-risk processing that likely warrants a DPIA. A DPIA will help systematically evaluate risks, such as the re-identification of synthetic data, and the misuse of them. Many DPAs (like CNIL) consider innovative data uses as likely needing DPIAs. Performing a DPIA demonstrates accountability and helps design controls (e.g., decide to withhold certain rare attributes from the synthetic data because the DPIA flagged re-identification risk).

- **Robust Privacy Safeguards in Generation**

Source data should be ensured that it is as anonymized as possible when training generative models. Direct identifiers should be removed and possibly even certain indirect ones not needed for the model. This way, even if the model memorizes something, it is less likely to be a direct identifier. Integrating differential privacy should also be considered into synthetic data generation. Differential privacy techniques add controlled noise to outputs, giving mathematical guarantees that any single individual's data has limited influence on the synthetic result. This significantly lowers re-identification odds, at a slight cost to precision. After the generations of synthetic data, re-identification tests should be performed (sometimes called attacker simulations). For instance, if synthetic records match real ones or if an algorithm can distinguish synthetic vs real in a way that links to identities. Also, outliers should be checked, as EDPS noted, synthetic data might miss or undercount outliers. but if it does not, and an outlier from real data appears in synthetic data, that could be a red flag for privacy. A synthetic dataset should be used if it passes these tests. These procedures are recommended to be documented as part of demonstrating compliance.

- **GDPR Consideration at the Early Anonymizing Stages**

It is recommended for synthetic data to be considered as personal. This conservative approach ensures that accidental compliance gaps are avoided. This is recommended because even if the synthetic data itself becomes anonymous, the process of creating it involves personal data. As a result, the process is still fully subject to GDPR. For example, if RAIDO partners share real data with a team that will synthesize it, that sharing is a GDPR processing operation that needs a legal basis and safeguards.

- **Continuous Documentation Process**

For compliance, it is recommended to maintain documentation on how synthetic data was created and validated. This includes the DPIA, the algorithms used, the privacy tests conducted, and why it is concluded that the data is or is not personal data. If an EU data protection authority ever inquires (for instance, if a data subject complains or an audit occurs), this documentation will be vital to demonstrate that RAIDO took a diligent, state-of-the-art approach consistent with EDPB guidance and GDPR's accountability principle. It will also help address any concerns from ethics boards or the EC reviewers about the rigor of RAIDO's data protection measures.

Based on the above, it is recommended to RAIDO consortium to treat synthetic data generation as a privacy-by-design measure that still requires legal application. The identification of a GDPR legal basis for using the original data to create synthetic data is also recommended. It is also recommended to consider any cross-border processing, if RAIDO partners in different countries exchange data to build models, complies with Chapter V of GDPR. This implies the usage of synthetic data to minimize transfers of raw personal data, and the utilization of EU-standard contractual clauses or intra-group agreements as needed for any transfers of the original data.

Finally, the RAIDO Project should stay updated with any new guidelines that may be published by official vendors. By rigorously ensuring that synthetic data is either truly anonymized or treated with the same care as real personal data, RAIDO can harness its benefits (like improving AI models and enabling collaboration) while upholding the legal and ethical standards expected by the GDPR, the project's legal experts, and the European Commission reviewers.

2.2.3.3. Intellectual Property Considerations for Synthetic Data and Digital Twins

The integration of synthetic data and digital twins within the RAIDO framework raises important considerations concerning intellectual property (IP) rights. These technologies offer novel ways to support privacy-preserving data processing, yet their legal status in terms of ownership, protection, and reusability remains partially unsettled under European law. As the consortium generates and applies these tools in research and innovation activities, it is essential to clarify the intellectual property implications at both dataset and model levels.

Synthetic data is typically generated using algorithms trained on real datasets, often containing personal or proprietary information. Although the resulting data may not contain any actual records from the original dataset, it may still reflect the structure, relationships, or statistical distributions of the source. This raises the question of whether the generated data qualifies for protection under the EU Database Directive (Directive 96/9/EC) [42, p. 96] or copyright law. If the generation process involves a substantial investment in verifying, presenting, or structuring data, the resulting collection of synthetic dataset may be protected as a sui generis database. However, this protection applies only to the collection and not to the data itself, and only if the

collection is original and non-trivial in its arrangement or if the selection of features itself demonstrates sufficient intellectual input.

Similarly, the models used to generate synthetic data, such as generative adversarial networks (GANs) or diffusion models, may themselves be subject to copyright or licensing conditions, particularly when built upon open-source or third-party frameworks. According to the European Commission's IP Code and the guidance from the European Union Intellectual Property Office (EUIPO) [43], ownership of algorithmic outputs depends not only on the training data but also on the architecture of the model and the licensing regime under which it was deployed. If the model was developed entirely within the RAIDO consortium, IP rights will typically follow consortium agreements. However, if models incorporate pre-existing tools or datasets, appropriate license compatibility checks and attribution are necessary.

In the context of digital twins, the situation becomes more complex due to the layered nature of the technology. Digital twins typically involve real-time data collection, dynamic simulation, predictive analytics, and user interfaces. Each of these components may originate from different partners or vendors, creating a shared ownership structure. In such cases, it is critical to determine who holds the rights over the model, the simulation logic, and the output generated during its use. Under current EU IP frameworks, particularly the Copyright Directive (Directive 2001/29/EC) [44], the rights of co-creators must be clearly outlined, and agreements must specify access, usage, and potential commercial exploitation of digital twin systems.

To address these concerns and ensure legal clarity within the RAIDO project, several best practices are recommended:

- Define the holders, the access rights, or ownership if possible, and licensing status of both the synthetic datasets and the digital twin components in the project's data management and consortium agreements.
- Conduct an early IP assessment for each dataset or model, including a review of the provenance of training data, licensing terms of underlying tools, and the originality of generated outputs.
- Ensure that any third-party software or datasets used are compliant with relevant licenses, such as the GNU General Public License (GPL), MIT, Apache, or Creative Commons, and verify whether these allow for redistribution, modification, or commercial reuse.
- Maintain metadata and documentation detailing data lineage, model configurations, and applicable rights to enable transparency and reproducibility.
- Align data and model sharing practices with the European Open Science Cloud (EOSC) FAIR [45] principles, while also respecting any commercial sensitivities or legal restrictions.

In addition, the European Data Strategy and initiatives such as the Data Governance Act (Regulation (EU) 2022/868) [46] increasingly emphasize the need to balance data innovation with the protection of rights holders, particularly in cross-border research collaborations. The RAIDO consortium is advised to stay informed of evolving legal interpretations and case law concerning synthetic datasets and AI-generated outputs, including those emerging from the Court of Justice of the European Union (CJEU) on database protection and authorship in AI contexts.

In conclusion, while synthetic data and digital twins provide valuable tools for privacy-conscious data science, their deployment must be accompanied by rigorous intellectual property management. By proactively addressing legal and licensing dimensions, the RAIDO project can support open collaboration and innovation while respecting the rights and contributions of all stakeholders.

2.2.3.4. Privacy Benefits and Limitations of Synthetic Data for Research

The utilization of synthetic data and digital twins represents a promising approach to mitigate privacy risks associated with the processing of personal data while maintaining research utility. Synthetic data, defined as artificially generated information that statistically resembles but does not contain actual personal data, offers significant potential for GDPR compliance in scientific research contexts [47]. Digital twins, as computational models that simulate real-world systems or processes, can similarly enable research without direct processing of personal data [48]. Within the RAIDO framework, these technologies warrant thorough examination as privacy-enhancing methodologies.

Synthetic data provides meaningful privacy protection by structurally breaking the connection to directly or indirectly identifiable individuals. It can retain key statistical relationships from the original dataset while substantially lowering the risk of re-identification. The RAIDO platform's data enrichment and generation capabilities, particularly through diffusion architectures and digital twins, present opportunities to generate high-fidelity synthetic datasets that maintain analytical value while enhancing privacy protection. Beyond privacy protection, synthetic data also plays a critical role in supporting the development, validation, and benchmarking of AI models. This is particularly relevant in healthcare and social sciences, where data access is often restricted due to legal and ethical constraints, synthetic datasets provide a legally compliant alternative to test model robustness and fairness. For instance, the European Commission’s Joint Research Centre has emphasized synthetic data as an enabler for creating controlled testing environments that minimize the use of real personal data (JRC Technical Report, 2021) [49].

However, applying synthetic data methods still poses limitations. A key challenge is the trade-off between privacy and utility. As synthetic data becomes more representative of the original, the risk of inadvertently exposing personal information increases. Recent research demonstrates that model inversion attacks and membership inference techniques can potentially extract sensitive information from

generative models used to create synthetic data, particularly when trained on small or distinctive datasets [50]. While differential privacy techniques mitigate memorization risks, recent work by the European Data Protection Supervisor warns against overreliance on algorithmic guarantees alone. Studies have shown that even without direct identifiers, synthetic datasets generated by GANs or diffusion models can encode patterns that indirectly allow reconstruction of sensitive features, particularly in imbalanced datasets. This reinforces the necessity for comprehensive adversarial testing. Techniques such as membership inference, linkage analysis, and k-anonymity scoring should be included in the RAIDO platform's validation process.

The integration of GDPR principles into synthetic data generation processes represent a critical consideration for RAIDO implementation. It is important to assess the project's digital twins to ensure that rare or outlier cases do not result in synthetic outputs that could be linked to identifiable individuals. Similarly, the generative diffusion architectures proposed for controllable image generation require privacy evaluation before resulting datasets can be confidently deemed outside GDPR scope. In addition to synthetic data, digital twins can facilitate privacy by enabling federated research environments. Instead of centralizing sensitive datasets, digital twins can simulate research scenarios based on distributed inputs, aligning with GDPR principles such as data minimization and purpose limitation. This approach is gaining traction in biomedical and energy domains, where real-time simulations can replace raw data. In conclusion, synthetic data and digital twins offer promising GDPR-compliant alternatives within the RAIDO framework, potentially enabling more extensive research data sharing. However, their implementation requires careful technical design, legal analysis, and governance frameworks to ensure genuine privacy protection rather than merely creating a perception of GDPR compliance. Privacy-preserving technologies like synthetic data and digital twins must also be situated within broader ethical frameworks. The High-Level Expert Group on AI [51] stresses that fairness, explainability, and contestability are core requirements for trustworthy AI. Synthetic data may inadvertently reflect historical biases or introduce skewed distributions, which could compromise equity in research. Accordingly, RAIDO should establish accountability mechanisms, such as regular audits and stakeholder oversight, to ensure the ethical use of these technologies.

2.2.3.5. Feasibility of Publicly Available Synthetic Data for RAIDO Use Cases

The integration of artificial intelligence and Machine Learning into energy systems has significantly enhanced capabilities in grid management, particularly in forecasting electricity generation and consumption.

However, these advancements are often constrained by the limited availability of high-quality, real-world data. Privacy concerns, regulatory barriers, and data sparsity—especially in under-monitored or emerging regions—impede the effective training and evaluation of forecasting models.

Synthetic data has emerged as a viable solution to these challenges by replicating the statistical and temporal characteristics of real-world energy data, thereby enabling the development and validation of forecasting algorithms without compromising data privacy or accessibility.

Pilot 1 – Energy Sector

The energy sector, which includes power generation and transmission, and distribution and consumption requires careful management of sensitive information because it forms a vital infrastructure system. The sector operates with operational data, which comprises precise power flow readings, grid conditions, asset wellness metrics, and customer utilization records. The vital nature of this data for maintaining grid stability and operational efficiency creates significant national security threats together with economic disruptions when unauthorized parties access or disclose it. The sector faces an urgent requirement for high-quality datasets to drive innovation through data-driven technologies such as predictive maintenance and demand response and grid optimization systems while protecting security and privacy. The application of synthetic data has become a suitable solution for this situation.

Synthetic data represents artificially created data that duplicates statistical patterns and operational behaviours of real-world datasets without exposing any actual confidential information. Through this method energy operators together with researchers and technology developers can both share information and perform analysis and innovation tasks without facing the dangers of exposing real data. Various modern programs and undertakings demonstrate that synthetic data presents both viable and beneficial solutions for this specific field.

The energy sector can use synthetic data because of recent improvements in privacy-preserving machine learning and differential privacy techniques which provide mathematical data anonymity guarantees during synthetic data generation. Synthetic data protection systems reduce the chance of attackers identifying sensitive information even when they obtain access to the data because the risk of re-identification remains very small.

Synthetic data provides the energy sector with a workable solution to combine data-driven innovation requirements with rigorous security and privacy protocols. Synthetic data serves as a fundamental element to support safe collaborative research and development in the sector while enhancing grid resilience and speeding up the adoption of smart sustainable energy systems.

Synthetic data refers to artificially created data that retains the statistical integrity and structural patterns of real datasets. Several methodologies exist for generating synthetic data:

- Rule-based simulations: domain-specific simulators generate data based on physical laws or predefined behaviours.

- Statistical models: approaches such as autoregressive models or copulas simulate temporal dependencies and correlations.
- Generative Machine Learning models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) or Diffusion Models.

Recent advancements in synthetic data generation for energy forecasting include:

- TimeGAN: Combines RNNs and GANs to generate realistic multivariate time series, widely applied in load and solar forecasting [52].
- CTGAN: Tailored for structured tabular data, useful for replicating smart meter datasets [53].
- Physics-Informed Generative Models: Integrate domain-specific constraints (e.g., power flow equations) to ensure realistic synthetic outputs.
- Transfer Learning and Domain Adaptation: Pretraining models on synthetic data and fine-tuning on limited real-world datasets.

The National Grid ESO in the United Kingdom has analysed synthetic data application for power system research by creating simulated grid state data which accurately mimics real measurement temporal and spatial relationships. The research utilizes GANs along with machine learning techniques to develop datasets which preserve vital power system modelling metrics including power flow distributions and voltage levels and frequency variations. The synthetic data enables predictive algorithm testing for fault detection and demand forecasting and grid optimization purposes while protecting operational insights that adversaries could misuse [54].

The InterConnect Project funded by the European Union uses synthetic data simulation to accelerate energy sector digitalization through secure data sharing. The project generates synthetic data which reproduces actual household power consumption behaviours through realistic patterns of appliance usage. The synthetic data approach enables GDPR compliance while enabling the development and testing of smart grid technologies including automated demand response and distributed energy resource (DER) management systems. The project uses synthetic data that resembles actual customer information to minimize privacy concerns and boost development of energy management systems [55]. Researchers at Lawrence Livermore National Laboratory (LLNL) established synthetic data systems which support power grid cyber-physical security investigations. The datasets present realistic interactions between IT and operational technology (OT) networks to enable cybersecurity research while protecting actual operational vulnerabilities from exposure. The growing cyber threats directed at critical infrastructure make this capability essential [56].

Pilot 2 – Smart Farming

In the field of smart farming, the integration of artificial intelligence (AI) and synthetic data has gained significant traction due to its cost-effectiveness and scalability. Traditional data collection methods in agriculture are often prohibitively expensive and

time-consuming, making synthetic data an essential component for advancing AI-driven solutions.

Generative adversarial networks (GANs) have been used to produce synthetic temperature data for greenhouses in Murcia, Spain [57]. This approach was necessitated by the limited quantity and quality of real-world temperature datasets, which were insufficient to meet the rigorous standards required for AI models designed for commercial applications. Synthetic data generation mitigates the high costs associated with extensive real-world data collection while ensuring models are adequately trained for deployment.

The use of synthetic data for detecting diseases in tomato plants (*Solanum lycopersicum)* has been explored in recent research [58]. By leveraging synthetic datasets, they developed a low-cost methodology for identifying plant diseases. This innovation provides an accessible solution for farmers, particularly in resource-limited settings, to monitor and manage crop health. Bayer employs synthetic data to simulate field conditions for predicting crop yields and disease outbreaks [59]. These virtual simulations facilitate advanced R&D efforts and improve decision-making processes, highlighting the transformative potential of synthetic data in precision agriculture. Beyond the aforementioned examples, synthetic data is being adopted by other industry leaders, including Ceres Imaging, Bosch Deepfield Robotics, and Taranis [60]. These companies utilize synthetic datasets for applications ranging from water stress detection to smart harvesting and pest monitoring. As the demand for sustainable and efficient farming practices grows, synthetic data remains a pivotal tool in bridging the gap between technological innovation and practical agricultural solutions.

Pilot 3 – Healthcare

In the healthcare domain, RAIDO's Pilot 3 explores the feasibility of using synthetic data in personalised pharmacogenomics (PGx), particularly where access to real-world patient data is limited due to ethical, legal, or operational constraints. Given the sensitivity of genomic and clinical records, the reuse and sharing of such data across research teams and institutions remain limited. As a response, synthetic datasets generated using diffusion models and digital twins provide a viable alternative, enabling AI training and testing without compromising data privacy.

In particular, publicly available health-related synthetic datasets such as Synthea™ [61] (an open-source synthetic patient generator developed by MITRE), UK Biobank simulation models [62], or MIMIC-Syn [63] (a synthetic counterpart of the real-world MIMIC-III database) may serve as proxies for evaluating PGx-relevant applications. These datasets simulate realistic electronic health records, disease trajectories, medication histories, and laboratory values, which can support AI-based summarisation, explainability testing, and adverse event prediction without exposing identifiable information. For instance, Synthea includes demographic and clinical data

generated using publicly available statistics and could be extended with PGx-like properties through domain-specific digital twins.

In RAIDO Pilot 3, synthetic data is leveraged in several stages of the AI development pipeline. It is used to pretrain models for bias detection, summarisation of PGx reports, and edge deployment of pharmacogenomic recommendations. The use of synthetic data in early phases reduces dependence on real patient data and allows rapid prototyping, robustness checks, and bias audits before clinical fine-tuning. Publicly available synthetic datasets and specialised simulators also facilitate scenario testing in underrepresented groups, such as individuals with rare genetic variants or complex co-morbidities. This staged approach reduces ethical risks during the early development phases and supports iterative validation of fairness and interpretability metrics before clinical deployment.

The feasibility of this approach is further enhanced by recent EU-funded initiatives and guidance. The European Commission's Joint Research Centre (JRC) [49] has promoted synthetic data as a privacy-enhancing enabler for AI development in regulated domains. Similarly, the EDPB Guidelines 07/2020 [8] recognize the role of anonymized synthetic data in privacy-preserving data sharing. From a legal and operational perspective, using publicly available synthetic datasets in RAIDO aligns with data minimization and privacy-by-design principles under the GDPR, especially when the synthetic generation pipeline is well-documented and evaluated for re-identification risks.

Pilot 4 – Robotics and Industry 5.0

Robotics and Industry 5.0, which is explored within RAIDO through Pilot 4**,** can derive great benefits from using synthetic data for AI model training. The collection of real data in the physical world can be a very complex and expensive process that often gives rise to safety issues**.** In the case of plant fibre characterisation considered within the project, real data collection requires the use of sophisticated micro-mechatronic setups that are very hard to operate. Thus, each sample gathered can take a long time, placing inherent limitations on the amount of real data that can be practically collected.

The use of synthetic data allows for numerous advantages, such as the construction of a multitude of scenarios corresponding to different situations that may occur in the data collection process, whether planned, possible or even sometimes potentially unexpected [64]. Most importantly it allows for virtually unlimited amounts of data that can complement the real datasets which are expensive to collect.

For RAIDO Pilot 4, synthetic data are being generated based on Finite Element Models that are currently used to model plant fibre mechanical properties. These data are then used in combination with real measurements of plant fibres obtained through tests carried out in micro-mechatronic setups.

Plant fibre characterisation using imaging and micro-mechatronic setups is a highly niche application that is only now surfacing as a viable method for improving bio-based composite manufacturing. As such, to the best of our knowledge, there are no public datasets available for this task. Public datasets investigating the mechanical properties of plant fibre reinforced composites, such as [65], [66] are scarce and only consider the finished composite structure not the plant fibres themselves as RAIDO does.

2.2.3.6. Integrating Privacy Principles into the Data Generation Process

Integrating privacy principles into the data generation process is crucial for building trust, complying with GDPR, and minimizing the risk of data breaches and privacy violations. The key principles for implementing and integrating privacy principles into the data generation process them are:

- **Data Minimization**, a fundamental aspect of GDPR and broader privacy principles, dictates that only data essential for a specific, legitimate purpose should be collected and processed. This principle aims to mitigate the risk of privacy breaches and misuse of personal information by limiting the amount of data retained. Implementing data minimization involves thoroughly assessing data requirements before collection, ensuring that only necessary attributes are gathered.It is crucial to resist the urge to collect data "just in case" and to regularly review existing data holdings to eliminate unnecessary information. By adopting data minimization, organizations show a commitment to responsible data handling and reduce potential privacy risks [1].

- **Purpose Limitation**, a core principle of data protection regulations like GDPR, mandates that personal data be collected only for specified, explicit, and legitimate purposes. It prohibits further processing in ways that are incompatible with those initial purposes. This principle ensures individuals maintain control over their data and prevents organizations from using the data for unforeseen or unrelated activities without explicit consent or a legal basis. To adhere to purpose limitation, organizations must clearly define the intended use of data at the point of collection, document this purpose, and implement technical and organizational measures to restrict data processing to these specified activities. Any new use of the data requires a reassessment of its compatibility with the original purpose and, if necessary, obtaining fresh consent or identifying a new legal basis for processing.

- **Transparency**, a fundamental aspect of ethical data handling and legal compliance, requires organizations to be open and honest with individuals about how their personal data is collected, used, and protected. This principle involves providing clear, concise, and easily accessible information about data processing activities, including the types of data collected, the purposes for which it is used, the recipients of the data, and the rights individuals have regarding their data. Achieving transparency involves making privacy notices

readily available, implementing user-friendly consent mechanisms, and proactively communicating about data practices. By embracing transparency, organizations build trust with individuals, empowering them to make informed decisions about their personal data and exercise their rights effectively [1].

- **Data Security**, a critical concern in the digital age, involves implementing technical and organizational measures to protect personal data from unauthorized access, use, disclosure, alteration, or destruction. This principle requires a multi-layered approach, including robust access controls, encryption of data both at rest and in transit, regular security assessments and penetration testing, as well as the development of incident response plans to address potential data breaches. Moreover, data security demands ongoing vigilance, with continuous system monitoring and adaptation to evolving threats. By prioritizing data security, organizations protect the privacy and confidentiality of individuals' information, maintain trust, and comply with legal and regulatory obligations.
- **Data Accuracy**, a vital component of responsible data management, requires organizations to ensure that the personal data they hold is precise, complete, and current. This principle aims to prevent errors and inaccuracies that could result in unfair or discriminatory outcomes for individuals. Maintaining data accuracy involves implementing validation checks at the point of collection, regularly cleansing and correcting data, and providing individuals with the opportunity to review and rectify their information. Additionally, organizations should establish procedures to promptly address and resolve data inaccuracies. By prioritizing data accuracy, organizations not only comply with legal requirements but also enhance the reliability and trustworthiness of their data-driven decisions.
- **Storage Limitation**, a fundamental principle in data protection, mandates that personal data be retained only for as long as necessary to fulfil the specific purpose for which it was collected. This principle aims to reduce the risk of data breaches and misuse by limiting the time data is kept. Implementing storage limitation involves defining clear data retention policies that specify retention periods for different types of data based on legal, regulatory, and business requirements. Organizations should also employ automated mechanisms to delete or anonymize data when it is no longer needed. Regularly reviewing and updating data retention policies is essential to ensure ongoing compliance and responsible data management.
- **Accountability**, a key aspect of data protection, requires organizations to demonstrate compliance with data protection principles and regulations. This involves implementing technical and organizational measures, conducting privacy impact assessments, appointing a Data Protection Officer (DPO) where necessary, maintaining records of processing activities, and training employees

on data protection principles. By fostering a culture of privacy and responsible data handling, organizations build trust with individuals and regulators.

While implementing and integrating privacy principles into the global data generation process, specific techniques for data generation, especially when dealing with sensitive information or adhering to privacy regulations such as GDPR, are crucial for creating useful datasets while minimizing privacy risks. The key techniques are described as follows:

- **Synthetic Data Generation**. Synthetic data generation creates artificial data that mimics the statistical properties and patterns of real data without containing any personally identifiable information (PII). This technique is often used for training machine learning models, testing software, and sharing data with external partners without compromising privacy. It involves analyzing real data to understand its distributions, correlations, and relationships between variables, and then using these statistical models to generate new, artificial data points. While synthetic data eliminates the risk of exposing real PII and can address data scarcity and bias issues, its quality depends on the accuracy of the statistical models used [67].

- **Data Masking/Pseudonymization.** Data masking or pseudonymization replaces sensitive data with realistic but non-identifiable substitutes to protect individuals' identities while preserving the data's utility for analysis and testing. Techniques include substitution, shuffling, number variance, encryption, and tokenization. This approach reduces the risk of identifying individuals [68].

- **Differential Privacy**. Differential privacy adds carefully calibrated noise to data to protect individual privacy while still allowing for useful statistical analysis. It ensures that the presence or absence of any single individual's data does not significantly affect the analysis results. This technique involves adding random noise to the data and tracking the privacy budget to balance privacy protection and data utility. While differential privacy provides a strong privacy guarantee, it can be complex to implement and may reduce the accuracy of the analysis [69].

- **Generalization**. Generalization replaces specific values with more general categories to reduce the granularity of the data, making it more difficult to identify individuals. Examples include replacing specific ages with age ranges or specific locations with broader geographic regions.

  This technique is relatively simple to implement and reduces the risk of identifying individuals, but it also reduces the level of detail in the data, which may affect the accuracy of analysis.

- **Suppression**. Suppression involves removing or redacting sensitive data fields, which is the most basic form of data protection. Examples include removing social security numbers or redacting names and addresses. While

suppression eliminates the risk of exposing the suppressed data and is simple to implement, it can significantly reduce the utility of the data and may not be sufficient to protect privacy if other data fields can be used to identify individuals.

- **Choosing the Right Technique**. The best data generation technique depends on factors such as the sensitivity of the data, the intended use of the data, legal and regulatory requirements, and available resources. Establishing clear data governance policies, conducting thorough risk assessments, being transparent with individuals, and regularly reviewing and updating data generation practices are key considerations for minimizing privacy risks and complying with legal requirements. Often, a combination of techniques is the best approach to balance privacy protection and data utility [67].

2.2.4. Defining a Regulatory Framework for Data Use in RAIDO Use Cases

There are several regulatory aspects that affect the RAIDO use cases. This ranges from European legislation and property rights to national obligations imposed by the countries for each use case. In this section, we are going to focus on the European legislation which has a major impact and is the basis for most applicable national legislation.

2.2.4.1. Key Legislative Areas Affecting Data Use in RAIDO

All pilots work with data, which puts the Data Act in the centre of attention. In some cases, the Data Governance Act might be equally applicable if the institution is a public or governmental institution, or if their data is used. This is interesting to consider because it might not only impose requirements on the pilot but also provide access to additional resources that were not previously accessible.

As soon as personal data or data categorised as “special” is involved, the General Data Protection Regulation and its national implementations need to be considered. When dealing with electronic health data, the newly published EHDS is applicable as well. The use cases also focus on AI applications, which means that the recently adopted AI Act is applicable. Although it is not yet fully in force, it will be in force within one year, and the use case pilots will have to pay attention to the consequences it entails.

Next to the data and AI-related regulations, long-standing intellectual property rights regulation will play a role in the use cases as most software and some data fall within its scope. For this reason, it must be taken into account to ensure compliance when using third-party contributions and, on the other hand, it might provide opportunities to exploit the products, systems, and results produced by the pilots.

2.2.4.2. Regulatory Requirements Under GDPR, EHDS, AI Act, Data Governance Act and Data Act

In this section, the requirements and opportunities of each legislative area are summarised with respect to the pilots. First, the Data Act and Data Governance Act will be considered, after which GDPR is discussed, followed by the EHDS, and finally the AI Act.

Data Act

The Data Act lays down rules to create a data ecosystem where data of connected products or related services that is in the possession of one party is made accessible to the data subjects in question. This means that this legislation has two ways of impacting the pilots. (1) The data produced by systems developed and made available by the pilots, when they are considered connected products, need to be made available to the users. (2) Pilots that want to use connected products are given the freedom to also use the data the connected product produces.

In both cases, a connected product must be in play, and looking at the definition [3] of it in the Data Act, it is described as an item that 'generates, obtains or collects data about its use or environment, is able to communicate the data and whose primary purpose is not storing, processing or transmitting the data for a party other than the user itself'. This means there should be an item involved that would generate or collect interesting information, but its original intent is not related to the data. Take for example smart watches that can collect information interesting for health studies or smart energy meters that collect data about the energy production and consumption. While interpretation of the exact definition is under discussion, so far it seems that 'item' should be a physical device, so it can be assumed that, for example, apps will not count.

It is assumed that point one is only in a limited way applicable to the pilots since the pilots do not intend to build and produce devices. In case they would, however, or in case they are using devices or tools, the data these devices generate need to be made available to the final end-user. The Energy pilot might be impacted by this, depending on the data sources and processing purposes. In exceptional cases and where it is related to the safety or interest of the public, this generated data must also be made available to the public authorities and government. However, this is an exceptional situation, and it is assessed to be out of scope for this text.
The regulatory requirements imposed by the Data Act seem very limited thus far, but it is easy to see that the pilots that are working with connected devices can take advantage of this regulation to access data that might be useful to them.

Data Governance Act

The Data Act and the Data Governance Act are two legislative frameworks that support the construction of a European data economy and the sharing of data through common data spaces, and to establish trust in the data sharing process. Where the

Data Act focuses on unlocking data from connected devices/products, the Data Governance Act looks at unlocking public sector data, setting up trusted data intermediaries, and promoting the sharing of data for altruistic purposes [46]. Public sector data has been the focus of the Open Data Directive in the past, but not all types of data were available for re-use under the Directive due to protective measures. That is why the DGA lays down the rules for sharing all public sector data and ensures the necessary safeguards for sharing. Next, the DGA defines data intermediaries; these are trusted, neutral organisations that negotiate between data holders or data subjects (these can be individuals or organisations) and data users on the terms of use of the data, the necessary safeguards, the financial agreements, and so on. With regards to data subjects, data intermediaries can help them negotiate access to their personal data by third parties or help them manage their GDPR rights (e.g., data cooperatives). Considering the impact of the DGA on RAIDO, there will be limited to no impact on the RAIDO platform itself, since it is no public sector body that holds data and does not take up a data intermediary role. However, considering the whole data flow, there might be an impact on some of the use cases if public sector data is involved.

General Data Protection Regulation

The GDPR has been tackled in great detail higher up in this document, since it is the legislation with the biggest impact when dealing with any personal data. GDPR is not in effect when not dealing with natural persons or with fully anonymised data. However, with regards to anonymised data, with current technological evolutions like the big advances in AI, the threshold for proving that data are truly anonymous has become quite high. This was also highlighted in the section on synthetic data and the GDPR. In addition, personal data might still remain absorbed in AI models, e.g., due to the involvement of personal data in the training set. Likewise, while the GDPR does not deal with non-personal data, any involvement of or link to personal data involves the GDPR again, and given the rising connectedness of databases, the number of mixed datasets (personal and non-personal) is increasing as well [1].

Within the RAIDO pilots, not many of them involve personal data, the pharmacogenomics case being the main exception. In the energy case, there are no personal data involved at this point, but data from simulated households is involved. If the pilot evolves into a full real-world case, personal data might therefore be involved at a later stage. For example, the electricity usage data of a household and the EAN identifiers are also considered personal data.

Any future cases making use of the RAIDO platform might very well involve personal data in some way, so it is important to set up the right processes and checks to make sure the basic principles and the data subject rights are being dealt with. In addition, the full data flow needs to be mapped to make sure the data coming into the RAIDO platform are treated correctly, and to distinguish between all the different roles and responsibilities in the data flow. Key is to know who the data controller is at each stage and which party acts as a processor, if applicable. Note that data storage is also a

form of data processing, and given the assumptions of considering both synthetic data and AI models with personal data as input as potentially containing personal data, this will impact the RAIDO platform as well as the other parties involved in the data processing.

A final important point on the GDPR is that while a large section above is dedicated to data in a research context, and RAIDO is dealing with a research context at this point, the pilots involve data that at some point will be part of a real-world business case. This means it is important to ensure the necessary safeguards are implemented for both research and non-research contexts.

European Health Data Space

The EHDS builds upon the general data space frameworks DA and DGA, and the safeguards under GDPR, to include specific requirements for health data. It distinguishes between two main types of data usage: primary use, which has a focus on health care and direct improvement of a patient's health, and secondary use, which involves the processing of electronic health data for research, innovation and policy purposes.When dealing with primary use, the main focus of the EHDS is getting all member states on a similar maturity level with regards to electronic patient data and their exchange through electronic health records (EHRs), and of medical devices and wellness apps that connect to these EHRs. On secondary use, EHDS defines data holders and categories of data that are required to be shared, and the supporting trustworthy infrastructure that makes this data sharing possible. Health Data Access Bodies will be set up to deal with data findability, access requests and access grants, and data processing will occur in Secure Processing Environments (SPE). There are requirements, both for primary and secondary use, on interoperability and data quality [4]. In the current RAIDO project, only one pilot falls under the EHDS requirements. As a scientific study in this phase, it falls mostly under the requirements for secondary use (Chapter IV) [4], but as the goal of the study is to develop tooling to help physicians make better decisions in which medicine best suits the patients, and for patients to be better informed, at some point it will need to fulfil the requirements of Chapter II as well**.**

Artificial Intelligence Act

Compared to the first three legislative documents, which might have little or no impact at all for most of the current pilots, the AI Act will have far-reaching implications for all pilots, the RAIDO platform itself, and future cases using RAIDO. The core of each pilot includes the development of AI systems and therefore the AI Act is the number one regulation to take into account within the European Union. Even if the development of the system is considered outside of Europe, as soon as the system is made available within the Union it must comply with the requirements laid down in it.

First, there are practices concerning AI systems that are prohibited. This means the entity responsible for the system must ensure they are not violating one of these prohibitions. This is not the case for the pilots discussed within the project or the

RAIDO platform itself, but it is good to track the intention of the AI systems being developed. For reference, prohibitions concern AI systems that manipulate human decisions, behaviour or freedoms; that exploit any vulnerabilities of natural persons; that evaluate or classify people over periods of time leading to discriminatory treatment; that infer emotions of a person in the area of work or education; that categorise people on biometric data to deduce any special kind of personal data (see also GDPR definition of personal data); and that use real-time biometric identification in public spaces. For a full understanding, Article 5 of the AI Act should be consulted [2]. A second important category of AI systems is the high-risk AI system group. An AI system is considered to be high-risk if it is a safety component of a product covered by a European Harmonisation law such as for vehicles, aviation and medical devices. A full list of products is given in Annex I of the AI Act. In addition, systems are considered high-risk if they fall within the context of biometric identification or categorisation, critical infrastructure, education, employment or access to self-employment, access to private or public benefits, law enforcement, migration or asylum and the justice or democratic processes. The complete details can be found in Annex III of the AI Act [2].

For several pilots, the AI system might fall within this high-risk category, e.g., as they are related to the energy grid which is considered critical infrastructure (energy pilot) and the health care of people if it would impact the access to treatment or healthcare (pharmacogenomics pilot). For these systems, the use and performance must be assessed together with the risks it poses to the users. Even the misuse of the AI system needs to be considered, to evaluate whether it is reasonably possible that it might impact society or a natural person negatively in their rights, safety or freedoms.

In cases where the AI system is considered to fall within the high-risk category, the provider (i.e., the developer of the AI system) has to ensure compliance of their AI systems by conducting risk assessments; mitigating known and reasonably foreseeable negative impacts; undergoing a third-party conformity assessment; applying thorough data governance and data quality validation; providing proper technical documentation; providing information to the deployer and ensuring transparency; ensuring automatic record-keeping; foreseeing human oversight and intervention possibilities; and achieving an appropriate level of accuracy, robustness and cybersecurity.

Thirdly, the AI systems might fall within the scope of general-purpose AI models or limited-risk AI systems. The first group are general AI models that are made available that might have high-impact capabilities and therefore have systemic risks, which amongst others means it is trained with more than 10^25 floating point operations (FLOPs) [70]. For comparison, the generative text models such as GPT-3 fall below this threshold, GPT-4 is around the threshold, and Gemini Ultra is considered to be above the threshold. Developers of such models must mainly provide clear documentation of the intent and usage of the model. In addition, they need to consider possible risks and the impact they have, which means there are mainly transparency

obligations. There are fewer transparency obligations for models that are made available open-source and for free. The second group, AI systems with limited risks, are AI systems that have mainly a generative character and are able to produce content such as humans do. Providers and deployers of these systems must comply with transparency requirements to ensure it is clear for people that they interact with an AI system or that the content is created by an AI system. For example, the use of chatbots needs to be made explicit to the user of the chat such that the user is aware of the AI system interaction.

In this last paragraph, the exceptions to the requirements given before are discussed. The biggest exception is given to research of high-risk AI systems, for which almost all high-risk requirements are not applicable anymore. However, the chances of the pilots being able to call in this exception is small since commercial partners are involved and the pilots have a clear market and economical application. To deal with this, regulatory sandboxes provide a space for development of AI systems impacted by the AI Act to support innovation and development by reducing initial compliance. The time-scoped initiative comes with other reporting requirements and at the time of writing, no such sandbox has been set up, to the best of our knowledge. However, it is assessed that the first regulatory sandboxes will likely be established soon for certain domains. Initial information might become available by September of 2025 on a European level, although some countries such as Spain might be ahead in the process. It is therefore recommended to postpone a decision regarding this until the end of this year if any high-risk systems are part of the pilots. If that is not possible, it is recommended to work within the confines of Article 60 of the AI Act [2] that provides means to test high-risk systems in real-world applications.

2.2.4.3. Intellectual Property Rights in Data Use and Compliance

For participants to assess their compliance with current Intellectual Property rights regarding the usage of data, it first needs to be clarified whether or to what extent any IP rights exist on the data. This question can be split for the data that is gathered for the development of the AI system and the data produced by the AI system when it is put in place.

Since the interest in data has risen sharply in the last decade, the discussion about data ownership and IP on data has been ongoing. Recent discussions between domain experts show that this is still a grey area, and the interpretation needs to be clarified in future court decisions [5]. However, certain trends are visible and can be built on in this work. Many experts seem to agree that data itself is not easily protected by copyright law, if possible, at all, making it hard to have IP rights on data itself. However, a collection of data, where it can be considered to be a database, undoubtedly has IP rights associated with it via the Database Directive [42] and the sui generis [71] right. This right exists for the collection, the structure and arrangement of the collection of data, and not on the data itself.

Related to the IP rights on data is the question of who the owner of the data is. In the exploration document of Hans Graux, titled “What is data ownership, and does it still matter under EU data law?”, this exact question is tackled in the current legislative landscape [71]. One of the interesting conclusions it presents is that the concept of data ownership is hard to define and might not be as relevant anymore. Instead, it is more useful to see who the holder is and who has data access rights. Not coincidentally, these terms are central to how the Data Act is set up. For example, if an organisation would like to use health data produced by a smartwatch, who is the data owner of the data generated by the watch? The user of the watch? The company which developed the watch and is providing the service that collects the data? In the Data Act, this is solved by defining the user to be the data subject and the company gathering the data the data holder. The Data Act clearly stipulates that the user, as a data subject, has data access rights to the data of which they are the subject. In addition, they could give an organisation the right to access the same data. Even though the company physically holds the data, it is not the only one with rights on the data.

When looking at the broader context of rights on data, there are possibilities to protect data via trademark protection, trade secret protection, and patent protection. However, even in these cases, additional requirements must be met to be able to use these protection mechanisms. For example, protection by patent right results not in the protection of the data itself but of how it can be used. For data to be protected by the trade secret framework, there must be confidentiality in the first place, and claims can only be made against breaches.

All together, it is very hard to provide a one-size-fits-all answer to all the data access questions for the RAIDO pilots and the RAIDO platform itself, especially if it stores and transforms data on the platform. For each process, it must be assessed which data is involved, who holds the data at that point in time, and who can provide data access rights to it. In addition, the Database Directive, sui generis right, or a patent or trade secret might mean that to obtain these access rights, fees, obligations, or restrictions might apply. There are trade secrets involved in at least one of the pilots, so apart from the data, the general presence of trade secrets and its impact on data processing needs to be taken into account. A second question is which rights apply to the data that is generated by the AI system. A similar conclusion can be made based on the observations in the previous section. However, a ruling of the EU Court in 2004 [72] implied a restrictive interpretation of the sui generis right: that it applies only to the investment made in obtaining data and not to resources used for the creation of new data, which could be an AI system in one of the use cases. In other words, it is not possible to claim the sui generis right on data produced by the AI system. The question of whether copyrighted data might be used and what the impact is on the output of an AI system if that data is used as training data has been tackled recently. First, the Digital Single Market Directive clearly states that data mining, training AI systems with data, is considered lawful when: (1) it is done by a research organization for the purposes of scientific research, [73] or (2) it is done as long as the data which was

“mined” was accessed lawfully and the copyright owner has not expressly prohibited this use. It has also been clarified that the output of generative AI can be copyright protected insofar as it is generated as "AI-assisted content." This means purely AI-generated output is not protectable, and AI-assisted content is only protectable provided it meets the required level of creative human input.

Given the great uncertainty about how IP rights can be applied to data and how this impacts the usage of AI applications, pilots must treat this subject with special care. Each pilot must examine the rights applicable to the data they want to use and ensure they comply with the applicable requirements, possibly being IP rights, patent rights, trademark rights, or trade secrets. Next, there needs to be an evaluation on which data is going to the RAIDO platform and how they are processed there. These requirements, the format of the output, and the set-up of the AI system itself will then in turn determine what rights can be applied to the output of the AI system. While most discussions on IP and AI look at data input and output, the algorithms used to develop a model might also be subject to IP. In general, mapping all components, roles, and responsibilities in the process and seeing which claims or legislations apply is the best approach. Considering the roles of data holder, data user, and data subject and their access rights is easier to apply than the concept of data ownership. In addition, this is more in line with the GDPR, which might also impose additional requirements in this context.

2.2.4.4. Legislative Gaps and Recommendations for RAIDO Compliance

In the previous sections, the different legislative areas were discussed. For each of these, the gaps that still exist in light of RAIDO will be summarised. In addition, specific recommendations are listed that can be followed to increase the chance of a successful compliance evaluation. The biggest questions arise concerning data usage in AI systems development. Before the implementation of the AI Act, the legislative landscape seems not fully harmonised when looking at which rules apply to the data intended for AI systems development. Secondly, the rules applicable to the data generated by the AI system depend heavily on the legal situation of the input data (personal or non-personal, public sector or private, from connected devices or not). However, current EU legislative publications and court rulings are starting to provide case-by-case clarity. The recommendations for RAIDO pilots are to carefully examine the data flow and each role in the flow: who is the data holder, are there personal data involved, who are the data subjects, etc.

It is also important to note which types of AI tools are developed and by whom, what input data are involved, and where and how the RAIDO platform is involved. For each step, the applicable legislation can then be assigned to better assess which rules apply to the data output. The AI Act, in contrast, leaves almost no gaps, and although not all implementation details are clear yet, it might require the partners to fulfil a long list of requirements. The chances of the AI practice in the current pilots being prohibited are slim, but it is recommended to verify the goal of the AI with the list of prohibited

practices, especially if these slightly change throughout the lifetime of the pilot. Several pilots might be likely to fall within the category of an AI practice that entails a high risk and must therefore meet the requirements laid down in the AI Act. Checking whether the AI system is considered high-risk should be one of the first steps. The requirements have a far-reaching impact which needs to be considered from the start. For other pilots, a good view on the AI system's capabilities and limitations is sufficient to comply with the requirements laid down in the AI Act. However, it is recommended to, even then, follow the good data-quality and management techniques required for high-risk AI systems. They will improve the system's quality, reliability, and trustworthiness, which are all key elements in the pilots presented here. In that respect, it will be beneficial if the checklist of trustworthy AI systems is followed. These will have the same effect and provide additional robustness to the pilots. In the next section, a more detailed description will be given.

Lastly, exceptions within the AI Act can be applicable, for example in the case of purely research-based AI model development, and the regulatory sandboxes provide an interesting tool to deal with testing high-risk systems. However, due to the recent adoption of the AI Act, there is still a big gap between the description and implementation of the regulatory sandboxes [70]. Therefore, it is not possible to rely on them yet.

The GDPR has a major impact on personal data and is mainly focused on data protection. It speaks only in limited terms on AI, mainly with regards to automated decision making and profiling, but its requirements on data protection affect the entire data flow when dealing with health data, unless the data is fully anonymised—a requirement that is increasingly difficult to meet. Therefore, a very thorough examination of the data flow and the roles and responsibilities of each party involved is recommended [74]. The parties must define which actor is the data controller (which can differ throughout the flow and stage of the project) and who might be the data processor. Synthetic data and AI models should also be considered as potentially containing personal data.

Concerning the Data Act and Data Governance Act, their impact on the pilots is likely limited. The recommendation is to be aware of their existence and the possibility of them giving the partners additional opportunities for access to more data. If a pilot falls under the definition of a connected product, it needs to be able to answer data subject requests on access to the data, and if public sector data is involved, it needs to answer to the requirements of the Data Governance Act. In future cases, these Acts might be more applicable when dealing with connected or public sector data. In those cases, again the impact of the legislation in the data flow needs to be considered. Since these acts are recent, there might be a need to contractually determine actions partners must take to fulfil future requirements, because different aims of partners might result in conflicting applications of the acts.

2.2.4.5. Practical Compliance Guidelines for RAIDO Partners

Practical legal compliance must be determined on a case-by-case basis since the details of the use case or pilot are very important. In addition, each pilot falls within the jurisdiction of a different country, with each having specific implementation differences. Although the information and guidelines given before are a good starting point, they will not guarantee total compliance.

Within the goals of this project, it is, however, advisable to consider the checklist of Trustworthy AI [75]. The trustworthiness of the AI systems aligns with the goals set out in RAIDO. To obtain Trustworthy AI, the realisation of trustworthy processes must be built on the foundations of trustworthy AI and assessed on a regular basis. If these elements are applied conscientiously, Trustworthy AI can be obtained. The foundations of Trustworthy AI are “Respect for human autonomy,” “Prevention of Harm,” “Fairness,” and “Explicability.” A development will follow these four foundations if it tries to find the balance between the following realisation principles: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity and non-discrimination, Societal and environmental wellbeing, and Accountability.

To find this balance between the principles, as they might be contradicting in some cases, the Assessment List for Trustworthy AI (ALTAI) provides questions that can be posed for each principle. Answering these will help to establish more trustworthy AI, assess legal compliance more easily and, in addition, reach the project goals. Therefore, a specific practical guideline is to assess the development of the pilots with the provided questions and assess the RAIDO platform implementations with the checklist as well. Answering it will give a good idea of the state of the AI systems. Be aware that answering the questions will not entail legal compliance, nor does legal compliance entail trustworthiness. Both need to be actively sought by assessing the state and adapting the system and development accordingly.

Specifically, it is advised to address the following steps:

- Map the data flow from start to finish for each pilot and project.
- Identify all parties and the roles they assume in the data flow.
- Look at the GDPR as soon as any personal data is involved (this includes synthetic data and AI models potentially trained on personal data).
- Define the controller(s) and processor(s) in each stage of the flow.
- Look at the EHDS as soon as health data is involved.
- Look at the Data Act as soon as the pilot interacts with, or is, a connected product.
- Define data holders and their responsibilities within the flow.
- Look at the Data Governance Act when the public sector is involved.

- Assess whether the work is purely scientific or if it has any commercial aspects. If purely scientific research, the exceptions for the AI Act and GDPR might apply.
- Use the assessment list of Trustworthy AI and define the following roles: Who is the developer of the model? Who is deploying the model? Who is using the model? Which data is involved at each processing stage, and is data altered in some way?
- Assess the risks and see how the AI Act applies.
- Look at potentially IP-protected material or trade secrets. Is IP-protected material used and is the flow compliant? Is IP protection possible on elements in the flow, and do you need to protect them from the start?

The steps outlined above provide a practical roadmap for the RAIDO consortium to navigate the complex web of legal and ethical obligations inherent in the project. This list should not be viewed as a one-time compliance check, but rather as a framework for continuous due diligence and risk assessment throughout the entire data and model lifecycle. By systematically addressing these points, partners can ensure that the principles of data protection, transparency, and fairness are embedded into their workstreams, directly contributing to the project's core mission of developing trustworthy and robust AI systems. Ultimately, this proactive and structured approach to compliance is fundamental to mitigating legal risks and ensuring the successful and sustainable impact of the RAIDO project.

Conclusion

This deliverable, D4.1, has provided a comprehensive legal and regulatory assessment for the reuse of data within the RAIDO project, extending beyond analysis to provide the consortium with tangible tools and actionable strategies. The report confirms that while the GDPR establishes a harmonized foundation for data protection, significant legal fragmentation exists in how RAIDO Member States apply research derogations, creating a complex compliance landscape for cross-border collaborations. To address this challenge, this deliverable has contributed a systematic comparative framework for evaluating these national differences, which was applied to carefully detail the legal regimes in Belgium, France, Spain, and Greece.

As a key innovation to navigate this complexity, LexAid-EU was introduced; an AI-powered legal assistant developed to support the project's compliance-by-design methodology. By providing verifiable, jurisdiction-specific interpretations of complex regulations, it empowers both technical and non-legal stakeholders to confidently address compliance challenges.

Furthermore, the report offers an in-depth evaluation of advanced technologies like synthetic data and digital twins, concluding that while they are powerful privacy-enhancing tools, their use demands careful evaluation of re-identification risks and

diligent intellectual property management. To translate these complex requirements into practice, this deliverable puts forward a set of practical compliance guidelines for RAIDO partners, including a clear roadmap for mapping data flows, defining roles, and utilizing the ALTAI checklist to ensure trustworthiness.

By delivering not only a comprehensive legal analysis but also a reusable assessment framework, practical operational guidelines, and the innovative solutions presented herein, D4.1 equips the RAIDO consortium with the essential resources to mitigate legal risks and embed ethical considerations into its core processes. These contributions enable the consortium to achieve its central mission: to develop innovative, reliable, and trustworthy AI systems that are not only technologically advanced but are built upon a robust and verifiable foundation of compliance with European values and law.

References

[1] European Union, 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation)'. in L119. Official Journal of the European Union, May 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[2] European Parliament and Council, 'Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)', no. L, 2024/1689. Official Journal of the European Union, 2024. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[3] European Parliament and Council, 'Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)', no. L, 2023/2854. Official Journal of the European Union, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32023R2854

[4] European Commission, 'Proposal for a Regulation on the European Health Data Space'. European Commission, Apr. 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0197

[5] WIPO Secretary, 'WIPO CONVERSATION ON INTELLECTUAL PROPERTY (IP) AND FRONTIER TECHNOLOGIES: Summary of Fourth Session', World Intellectual Property Organization, Geneva, Switzerland, Sep. 2021. [Online]. Available: https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=558874

[6] P. Lewis *et al.*, 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks', *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[7] Court of Justice of the European Union, 'CURIA - Case-law of the Court of Justice'. 2025. [Online]. Available: https://curia.europa.eu/jcms/jcms/j_6/en/

[8] 'Guidelines 05/2020 on Consent under Regulation 2016/679'. 2020. [Online]. Available: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en

[9] Autorité de protection des données, 'Act of 30 July 2018 on the protection of natural persons with regard to the processing of personal data'. Belgian Official Journal, Sep. 2018. [Online]. Available: https://www.dataprotectionauthority.be/publications/act-of-30-july-2018.pdf

[10] GDPRhub, 'Data Protection in France'. Mar. 2025. [Online]. Available: https://gdprhub.eu/Data_Protection_in_France

[11] CNIL, 'Commission Nationale de l'Informatique et des Libertés (CNIL)'. 2025. [Online]. Available: https://www.cnil.fr/en

[12] CNIL, 'Les méthodologies de référence pour la recherche dans le domaine de la santé (Reference methodologies for research in the health field)'. 2024. [Online]. Available: https://www.cnil.fr/fr/les-methodologies-de-reference-pour-la-recherche-dans-le-domaine-de-la-sante

[13] Gobierno de España, ‘Ley Orgánica 3/2018 de Protección de Datos Personales y garantía de los derechos digitales’. Dec. 2018. [Online]. Available: https://www.uspceu.com/Portals/0/docs/transparencia/normativa/legislacion-general
[14] Greek Parliament, ‘Law 4624/2019 on the Protection of Personal Data’. Government Gazette A/137, Aug. 2019. [Online]. Available: https://kglawfirm.gr/wp-content/uploads/2019/09/Newletter-GDPR_4624_2019_clean.pdf
[15] ‘European Data Protection Supervisor - Official Website’. 2020. [Online]. Available: https://edps.europa.eu
[16] CNIL, ‘Practical note on data security requirements for research using the French national health data system (SNDS)’. [Online]. Available: https://www.cnil.fr
[17] Spain, ‘Real Decreto 957/2020, de 3 de noviembre, por el que se regulan los requisitos de realización de estudios observacionales con medicamentos de uso humano’, Boletín Oficial del Estado. [Online]. Available: https://www.boe.es/buscar/doc.php?id=BOE-A-2020-13823
[18] Spain, ‘Ley 14/2007, de 3 de julio, de Investigación biomédica’, Boletín Oficial del Estado. [Online]. Available: https://www.boe.es/buscar/doc.php?id=BOE-A-2007-12945
[19] Bird & Bird LLP, ‘Guide to the GDPR – Derogations and Special Conditions’. [Online]. Available: https://www.twobirds.com/-/media/pdfs/gdpr-pdfs/81--guide-to-the-gdpr--derogations-and-special-conditions.pdf
[20] C. & B. LLP, ‘The French CNIL Reminds Two Medical Research Organizations of Their Data Protection Obligations’. [Online]. Available: https://www.insideprivacy.com/data-privacy/the-french-cnil-reminds-two-medical-research-organizations-of-their-data-protection-obligations/
[21] C. M. Romeo-Casabona, ‘The New European Legal Framework on Personal Data Protection and the Legal Status of Biological Samples and Biobanks for Biomedical Research Purposes in Spanish Law’, in *GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation across Europe*, S. Slokenberga, O. Tzortzatou, and J. Reichel, Eds., Cham: Springer International Publishing, 2021, pp. 363–378. doi: 10.1007/978-3-030-49388-2_20.
[22] D. L. A. Piper, ‘Data Protection Laws of the World - Greece’. [Online]. Available: https://www.dlapiperdataprotection.com/index.html?t=law&c=GR
[23] A. Legal, ‘Belgian DPA fines NGO Desinfolab for GDPR breach’. [Online]. Available: https://www.activemind.legal/guides/fine-desinfolab/
[24] CNIL, ‘La recherche scientifique hors du domaine de la santé’. [Online]. Available: https://www.cnil.fr/fr/recherche-scientifique-hors-sante
[25] A. E. de P. de Datos, ‘Farmaindustria Code of Conduct Regulating the Processing of Personal Data in Clinical Research’. [Online]. Available: https://www.aepd.es/documento/farmaindustria-code-conduct-regulating-processing-personal-clinical-en.pdf
[26] C. A. Trappe, C. M. Sivalli Campos, C. de Freitas Oliveira, J. G. S. Tavares da Silva, L. Y. T. Uchimura, and M. F. Figueiró, ‘Scoping review of evidence synthesis: Concepts, types and methods’, *PLoS One*, vol. 20, no. 5, p. e0323555, 2025, doi: 10.1371/journal.pone.0323555.
[27] Z. Qian, T. Callender, B. Cebere, and et al, ‘Synthetic data for privacy-preserving clinical risk prediction’, *Scientific Reports*, vol. 14, p. 25676, 2024, doi: 10.1038/s41598-024-72894-y.
[28] Y. Liu, J. Peng, J. J. Q. Yu, and Y. Wu, ‘PPGAN: Privacy-preserving Generative Adversarial Network’, *arXiv preprint*, 2019.

[29] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, 'CTAB-GAN+: Enhancing Tabular Data Synthesis', *arXiv preprint*, 2022.
[30] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, 'A Unified Framework for Quantifying Privacy Risk in Synthetic Data', *arXiv preprint*, 2022.
[31] Y. Zhao and J. Zhang, 'Does Training with Synthetic Data Truly Protect Privacy?', *arXiv preprint*, 2025.
[32] G. Ganev and E. De Cristofaro, 'The Inadequacy of Similarity-based Privacy Metrics: Privacy Attacks against "Truly Anonymous" Synthetic Datasets', *arXiv preprint*, 2024.
[33] A. Beduschi, 'Synthetic Data Protection: Towards a Paradigm Change in Data Regulation?', *Big Data & Society*, vol. 11, no. 1, pp. 1–5, 2024, doi: 10.1177/20539517241231277.
[34] 'GDPR and Digital Twins: Managing Data Privacy in Virtual Replicas'. [Online]. Available: https://www.gdpr-advisor.com/gdpr-and-digital-twins-managing-data-privacy-in-virtual-replicas/
[35] M. Kustelegea, R. Mekovec, and A. Shareef, 'Privacy and security challenges of the digital twin: systematic literature review', *Journal of Universal Computer Science*, vol. 30, no. 13, pp. 1782–1806, 2024, doi: 10.3897/jucs.114607.
[36] A. Zemtsov *et al.*, 'Security and Privacy of Digital Twins for Advanced Manufacturing: A Survey', *arXiv preprint*, 2024.
[37] D. M. Botin-Sanabria, A.-S. Mihaita, R. Peimbert-García, M. Ramírez-Moreno, R. G. Ramírez-Mendoza, and J. H. Lozoya-Santos, 'Investigation of Digital Twin Technology for Secure and Privacy Preserving Networking', in *Procedia Computer Science*, 2023, pp. 398–406.
[38] M. Ahmed-Assalemi, A. Al-Khateeb, and A. Agounon, 'Privacy-enhancing technologies in the design of digital twins for smart cities', *Network Security*, vol. 2022, no. No. 7, pp. 7–17, 2022, doi: 10.12968/1533-4858(22)70046-3.
[39] European Data Protection Supervisor, 'TechSonar: Synthetic Data'. 2022. [Online]. Available: https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data
[40] Wilson Sonsini Goodrich & Rosati, 'EU Privacy Regulators Confirm That Legitimate Interest is a Valid Legal Basis for AI Model Training and Deployment'. 2024. [Online]. Available: https://www.wsgr.com/en/insights/eu-privacy-regulators-confirm-that-legitimate-interest-is-a-valid-legal-basis-for-ai-model-training-and-deployment.html
[41] Tonic AI, 'Understanding Model Memorization in Machine Learning'. [Online]. Available: https://www.tonic.ai/guides/understanding-model-memorization-in-machine-learning
[42] 'Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases'. 1996. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A31996L0009
[43] 'European Union Intellectual Property Office - Official Website'. 2024. [Online]. Available: https://euipo.europa.eu
[44] 'Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society'. 2001. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32001L0029
[45] M. D. Wilkinson and others, 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data*, vol. 3, p. 160018, 2016, doi: 10.1038/sdata.2016.18.
[46] European Union, 'Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 (Data Governance Act)'. in L152. Official Journal of the European

Union, Jun. 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R0868
[47] I. S. Association, 'Synthetic Data'. 2021. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/governance/iccom/IC21-013_Synthetic_Data.pdf
[48] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, 'Digital Twin in Industry: State-of-the-Art', *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2019, doi: 10.1109/TII.2018.2873186.
[49] E. C. J. R. Centre, 'Artificial Intelligence and Privacy: Artificial Data Generation and Privacy Enhancing Technologies'. 2021. [Online]. Available: https://publications.jrc.ec.europa.eu/repository/handle/JRC125952
[50] M. Slokom, P.-P. de Wolf, and M. Larson, 'Exploring Privacy-Preserving Techniques on Synthetic Data as a Defense Against Model Inversion Attacks', in *Proceedings of the 26th International Conference on Information Security (ISC 2023)*, 2023. [Online]. Available: https://ir.cwi.nl/pub/33674
[51] H.-L. E. G. on A. Intelligence, 'Ethics Guidelines for Trustworthy AI'. 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[52] J. Yoon, D. Jarrett, and M. van der Schaar, 'Time-series Generative Adversarial Networks', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019, pp. 5508–5518. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c9efe5f23c2b9a5b38e0e0f1b62d94f0-Abstract.html
[53] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, 'Modeling Tabular Data using Conditional GAN', in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019, pp. 7333–7343. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b8a3a7978450c7c985d1d-Abstract.html
[54] National Grid ESO, 'ESO and Alan Turing Institute use machine learning to help balance GB electricity grid'. [Online]. Available: https://www.neso.energy/news/eso-and-alan-turing-institute-use-machine-learning-help-balance-gb-electricity-grid
[55] InterConnect Project Consortium, 'About the InterConnect Project'. [Online]. Available: https://interconnectproject.eu/about/
[56] Lawrence Livermore National Laboratory, 'Defending Critical Infrastructure'. [Online]. Available: https://gs.llnl.gov/energy-homeland-security/cyber-and-infrastructure-resilience/defending-critical-infrastructure
[57] E. Morales-Garcia and others, 'Synthetic Temperature Data Generation for Greenhouse AI Models', *Journal of Agricultural Data Science*, vol. 5, no. 2, pp. 45–58, 2023.
[58] R. Klein and others, 'Low-Cost Disease Detection in Tomato Plants Using Synthetic Data', *Computers and Electronics in Agriculture*, vol. 200, pp. 105–120, 2024.
[59] Bayer Crop Science, 'Simulating Field Conditions with Synthetic Data'. 2023. [Online]. Available: https://www.cropscience.bayer.com
[60] Ceres Imaging, Bosch Deepfield Robotics, and Taranis, 'Industrial Applications of Synthetic Data in Smart Farming'. 2023.
[61] M. Corporation, 'Synthea - Synthetic Patient Population Simulator'. 2023. [Online]. Available: https://syntheticahealth.github.io/synthea/
[62] U. K. Biobank, 'UK Biobank Data Simulation Resources'. 2023. [Online]. Available: https://www.ukbiobank.ac.uk/
[63] M. I. T. Lab, 'MIMIC-Syn: Synthetic Critical Care Data'. 2021. [Online]. Available: https://physionet.org/content/mimic-syn/

[64] H. Deng, 'Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems: A Primer', United Nations Institute for Disarmament Research (UNIDIR), 2023. [Online]. Available: https://unidir.org/publication/exploring-synthetic-data-artificial-intelligence-and-autonomous-systems-primer
[65] S. S. Kumar, 'Dataset on mechanical properties of natural fiber reinforced polyester composites for engineering applications', *Data in Brief*, vol. 28, p. 105054, 2020, doi: 10.1016/j.dib.2019.105054.
[66] M. A. H. Alharbi, S. Hirai, H. A. Tuan, S. Akioka, and W. Shoji, 'Dataset on mechanical, thermal and structural characterization of plant fiber-based biopolymers prepared by hot-pressing raw coconut coir, and milled powders of cotton, waste bagasse, wood, and bamboo', *Data in Brief*, vol. 30, p. 105510, 2020, doi: 10.1016/j.dib.2020.105510.
[67] M. Goyal and Q. H. Mahmoud, 'A Systematic Review of Synthetic Data Generation Techniques Using Generative AI', *Electronics*, vol. 13, no. 17, p. 3509, 2024, doi: 10.3390/electronics13173509.
[68] F. Kohlmayer, R. Lautenschläger, and F. Prasser, 'Pseudonymization for research data collection: is the juice worth the squeeze?', *BMC Medical Informatics and Decision Making*, vol. 19, p. 178, 2019, doi: 10.1186/s12911-019-0905-x.
[69] R. Cummings *et al.*, 'Advancing Differential Privacy: Where We Are Now and Future Directions for Real-World Deployment', *Harvard Data Science Review*, vol. 6, no. 1, 2024, doi: 10.1162/99608f92.d3197524.
[70] E. Commission, 'Questions and Answers on the Artificial Intelligence Act'. 2024. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/qanda_24_3032
[71] H. Graux, 'What is Data Ownership, and Does it Still Matter under EU Data Law?' 2020. [Online]. Available: https://data.europa.eu/sites/default/files/report/What%20is%20data%20ownership%2C%20and%20does%20it%20still%20matter%20under%20EU%20data%20law.pdf
[72] Court of Justice of the European Union, 'Case C-444/02, Fixtures Marketing Ltd v Organismos Prognostikon Agonon Podosfairou (OPAP)'. 2004. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62002CJ0444
[73] European Parliament and Council, 'Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market'. 2019. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790
[74] European Data Protection Board, 'Guidelines 05/2020 on profiling and automated decision-making under Regulation 2016/679'. 2020. [Online]. Available: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-profiling-and-automated-decision-making_en
[75] European Commission, 'Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment'. 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101135800