



ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



Universitat Autònoma
de Barcelona

DataSHIELD advances

Transformative Extensions for Privacy-Preserving Big Data
Analysis in Health and Biosciences

Thesis submitted by:

Xavier Escribà Montagut

for the degree of:

Doctor in Bioinformatics

Doctorate Program in Bioinformatics

Thesis developed at the Barcelona Institute for Global Health (ISGlobal)

Supervisor:

Dr. Juan Ramón González Ruiz

Universitat Autònoma de Barcelona

Barcelona, 2023

Director:

Juan Ramón González Ruiz^{1,2}

1. Barcelona Institute for Global Health, ISGlobal, Barcelona, Spain.
2. Department of Mathematics, Universitat Autònoma de Barcelona, Cerdanyola, Spain.

*Al pare i a la mare,
per estar sempre radere.*

*Al meu germà,
per marcar el camí.*

Agraïments

Va ser ben bé al mig d'una època ben confusa que va començar aquest període de la meua vida, al mig d'una pandèmia mundial. Després d'estar tancats i no saber ben bé cap on anava res, va ser el que ha estat el tutor d'aquesta tesis, el Juan Ramon, JR pels amics, qui va confiar en mi i em va brindar l'oportunitat de fer un doctorat al seu costat. Tot i que no tenia gens clar si aquest era el camí que volia seguir, em va ficar al davant un projecte que només podia anar bé, i no sols això sinó que em va oferir les millors condicions possibles perquè pugues emprendre aquesta època amb totes les comoditats possibles, tenint en compte la situació en què estàvem.

Al mateix moment que tot això començava, va ser també un moment de grans canvis i millores en l'àmbit personal. Després de vint anys d'amistat i massa històries per recordar, el meu millor amic, el Torres, va tornar d'un Erasmus a Milà una mica accidentat per la pandèmia, i va tornar per instal·lar-se al meu pis de Barcelona, pis en el qual ell va començar la seva carrera professional set dies després que jo comences el doctorat. No va ser una època molt llarga que vam estar vivint junts, però ens ho vam passar tan bé com vam poder, inclús vam començar a anar al gimnàs i mirar pel·lícules... algunes millors que d'altres. Tota aquesta època també la vam compartir amb l'Àlex, amb qui ja feia un any que vivia; quan el vaig conèixer era el seu primer cop vivint fora de casa, massa jove per tindre el cap enlloc, ara ja parlo d'ell com el meu amic que és actor professional, i quina sort la meua de poder-ho dir, només fa falta que ens trobem un any a parroquies altes per celebrar-ho.

Passada aquesta època, el Torres i jo vam decidir tornar a Lleida, doncs amb el teletreball preferíem estar més a prop de la nostra família i amics. Jo vaig tornar a casa dels meus pares, ell es va llogar un apartament al mateix barri, ens podem veure de balcó a balcó, no fos cas que ens trobéssim a faltar. Després de quasi set anys vivint fora per estudiar, vaig tornar a casa, on els meus pares em van acollir amb les mans obertes, no recordava el bé que es viu (i es menja) a casa dels pares, on es donen fenòmens paranormals com un calaix on deixes la roba bruta i apareix neta, hi ha coses que no es poden pagar amb diners, aquestes coses només passen a casa dels pares. Aquest període junts ens hem fet companyia els uns als altres, hem fet mil coses junts i ho trobaré a faltar quan s'acabi.

Aquesta època per Lleida també ha estat un gran retrobament d'amistats. Després d'estar estudiant a fora, ens vam retrobar els de sempre als llocs de sempre. També ens vam retrobar els d'avegades per passar a ser els de sempre. Hem fet esmorzars d'aquells que són dinars, dinars que són sopars i sopars que acaben de dia. Ens hem reunit per fer calçotades que haurien d'estar penades per llei, inclús vam intentar fer un club de cine que va fracassar. Algun viatge en cotxe massa apretats, cuinar un xai sencer sense saber en absolut el que estàvem fent. Una passejada per la montanya, unes surtides en bici i escalar algun dia; que no tot en aquesta vida és menjar, també s'ha de suar una mica. Tot el que hem viscut junts són petits records que formen una gran història que no puc esperar a veure com segueix.

Es ist auch wichtig, meinen Freund Paul zu erwähnen, seit wir uns in Warschau kennengelernt haben, haben wir den Kontakt durch das aufrechterhalten, was uns ursprünglich verbunden hat, Live-Musik. Wie glücklich, weiterhin dies mit dir genießen zu können und das nicht zu verlieren, was uns verbindet.

La història de la gent que m'ha acompanyat al llarg d'aquesta etapa s'acaba en aquest últim any, on hem compartit amistat i lloc de feina amb el David. Amb el David feia anys que ens coneixíem, i tot va ser un dia fent una cervesa que li vaig comentar que hi havia una nova plaça de doctorat al grup de recerca, i doncs a partir d'aquí tot va anar ben de pressa. Aquest últim any ens hem fet un bon fart d'anar junts al cotxe a l'oficina de Barcelona, quina sort haver pogut estar tan ben acompanyat aquest últim any, perquè tot ha estat molt més fàcil amb ell.

Aquesta ha estat la meua història personal al llarg del doctorat, on tothom que hi ha participat ha estat una peça clau i ha estat directament relacionat en l'èxit d'aquest. Sense tot aquest suport al darrere, no hauria estat possible.

“The only source of knowledge is experience.”
—Albert Einstein

Abstract

In an era where data privacy is crucial, conducting reproducible and secure data analysis in a collaborative context among multiple research centers is a challenging task. With the growing relevance of various types of highly sensitive data such as clinical, epidemiological or omics (genomics, transcriptomics, exposomics, ...), the need for research through federated analysis has become a critical necessity, especially due to the sensitive nature of the data involved, which raises significant privacy and ethical concerns. To address this problem, this PhD thesis aims to extend the capabilities of DataSHIELD, a federated analysis platform. This PhD thesis provides advanced methods and tools like ShinyDataSHIELD, *resources*, OmicSHIELD, and dsExposome, with the goal of making DataSHIELD more relevant, adding features to handle a broader range of data types, and ensuring the platform's adaptability and scalability for the future.

The adopted methodology involves a series of software developments, case studies, and real-world applications. Comparative analyses have been used to establish the effectiveness of the new tools and methods created. Additionally, techniques like clustered analysis and differential privacy have been integrated into DataSHIELD's capabilities.

This PhD thesis achieved the objectives by extending DataSHIELD's capabilities to address existing needs in multi-cohort studies. The ShinyDataSHIELD interface encourages a more pleasant and accessible user experience for both novice and experienced researchers. The concept of *resources* is the seminal tool for working with datasets in different formats,

thus expanding DataSHIELD's applicability to multidomain research. OmicSHIELD offers a robust set of tools for analyzing omics data, while dsExposome provides specialized features for exposome data analysis. Both of these additions operate in a federated manner while preserving individuals' privacy. All the objectives achieved in this PhD thesis are attached to an European project (ATHLETE), which required all the developments for the WP3, which is devoted to tools for federated data analysis.

Additionally, the study addresses challenges related to data privacy and collaborations among centers. The platform can effectively manage larger data sets and perform complex analyses without compromising data privacy. This adaptability paves the way for future applications of DataSHIELD in other fields of research, such as neuroimaging and artificial intelligence.

In conclusion, the new tools and features significantly improve DataSHIELD's capacity, scalability, and adaptability. These improvements promise to accelerate the adoption of federated data analysis methods in multi-center studies, thereby advancing research while rigorously maintaining data privacy. Notably, DataSHIELD also contributes to broader goals of reproducibility and transparency in scientific research by allowing results to be easily verified without the need for data sharing, thus overcoming traditional barriers to collaborative research. The work presented in this PhD thesis serves as a critical advancement in federated data analysis, bridging gaps between data privacy, reproducibility, and collaborative research in biomedicine.

Resum

En una era on la privacitat de les dades és crucial, dur a terme anàlisis de dades reproduïbles i segures en un context col·laboratiu entre diversos centres de recerca és una tasca complexa. Amb la creixent rellevància de diversos tipus de dades altament sensibles com clíniques, epidemiològiques o òmiques (genòmiques, transcriptòmiques, exposòmiques, ...), la necessitat de recerca a través de l'anàlisi federada s'ha convertit en una necessitat crítica, especialment degut a la naturalesa sensible de les dades implicades, el que planteja preocupacions significatives sobre la privacitat i l'ètica. Per abordar aquest problema, aquesta tesi doctoral té com a objectiu ampliar les capacitats de DataSHIELD, una plataforma d'anàlisi federada. Aquesta tesi doctoral proporciona mètodes avançats i eines com ShinyDataSHIELD, *resources*, OmicSHIELD i dsExposome, amb l'objectiu de fer que DataSHIELD sigui més rellevant, afegint característiques per gestionar una gamma més ampla de tipus de dades i garantir l'adaptabilitat i escalabilitat de la plataforma per al futur.

La metodologia adoptada implica una sèrie de desenvolupaments de programari, estudis de cas i aplicacions amb dades reals. S'han utilitzat anàlisis comparatius per establir l'eficàcia de les noves eines i mètodes creats. A més, tècniques com l'anàlisi agrupada i la privacitat diferencial s'han integrat en les capacitats de DataSHIELD.

Aquesta tesi doctoral ha aconseguit els objectius ampliant les capacitats de DataSHIELD per abordar les necessitats existents en estudis de múltiples cohorts. La interfície ShinyDataSHIELD fomenta una experiència d'usuari més agradable i accessible tant per a investigadors novells com experimentats. El concepte de *resources* és l'eina seminal per treballar amb conjunts de dades en diferents formats, expandint així l'aplicabilitat de DataSHIELD a la re-

cerca multimodal. OmicSHIELD ofereix un conjunt robust d'eines per analitzar dades òmiques, mentre que dsExposome proporciona característiques especialitzades per a l'anàlisi de dades de l'exposoma. Totes dues addicions operen de manera federada mentre preserven la privacitat dels individus. Tots els objectius assolits en aquesta tesi doctoral estan adjunts a un projecte europeu (ATHLETE), que va requerir tots els desenvolupaments per al WP3, que està dedicat a les eines per a l'anàlisi de dades federades.

A més, l'estudi aborda reptes relacionats amb la privacitat de les dades i les col·laboracions entre centres. La plataforma pot gestionar efectivament conjunts de dades més grans i dur a terme anàlisis complexes sense comprometre la privacitat de les dades. Aquesta adaptabilitat obre el camí per a futures aplicacions de DataSHIELD en altres camps de recerca, com ara la neuroimatge i la intel·ligència artificial.

En conclusió, les noves eines i característiques milloren significativament la capacitat, escalabilitat i adaptabilitat de DataSHIELD. Aquestes millores acceleraran l'adopció de mètodes d'anàlisi de dades federada en estudis multicèntrics, avançant així la recerca mentre es manté rigorosament la privacitat de les dades. Notablement, DataSHIELD també contribueix a objectius més amplis de reproductibilitat i transparència en la recerca científica permetent que els resultats siguin fàcilment verificables sense la necessitat de compartir dades, superant així les barreres tradicionals en la recerca col·laborativa. El treball presentat en aquesta tesi doctoral serveix com un avanç crític en l'anàlisi de dades federada, cobrint els buits entre la privacitat de les dades, la reproductibilitat i la recerca col·laborativa en biomedicina.

Resumen

En una era donde la privacidad de los datos es crucial, llevar a cabo análisis de datos reproducibles y seguros en un contexto colaborativo entre diversos centros de investigación es una tarea compleja. Con la creciente relevancia de varios tipos de datos altamente sensibles como clínicos, epidemiológicos u ómicos (genómicos, transcriptómicos, exposómicos, ...), la necesidad de investigación a través del análisis federado se ha convertido en una necesidad crítica, especialmente debido a la naturaleza sensible de los datos implicados, lo que plantea preocupaciones significativas sobre la privacidad y la ética. Para abordar este problema, esta tesis doctoral tiene como objetivo ampliar las capacidades de DataSHIELD, una plataforma de análisis federado. Esta tesis doctoral proporciona métodos avanzados y herramientas como ShinyDataSHIELD, *resources*, OmicSHIELD y dsExposome, con el objetivo de hacer que DataSHIELD sea más relevante, añadiendo características para gestionar una gama más amplia de tipos de datos y garantizar la adaptabilidad y escalabilidad de la plataforma para el futuro.

La metodología adoptada implica una serie de desarrollos de software, estudios de caso y aplicaciones con datos reales. Se han utilizado análisis comparativos para establecer la eficacia de las nuevas herramientas y métodos creados. Además, técnicas como el análisis agrupado y la privacidad diferencial se han integrado en las capacidades de DataSHIELD.

Esta tesis doctoral ha conseguido los objetivos ampliando las capacidades de DataSHIELD para abordar las necesidades existentes en estudios de múltiples cohortes. La interfaz ShinyDataSHIELD fomenta una experiencia de usuario más agradable y accesible tanto para investigadores novatos como experimentados. El concepto de *resources* es la herramienta seminal para trabajar con conjuntos de datos en diferentes formatos, expandiendo así la aplicabilidad de

DataSHIELD a la investigación multimodal. OmicSHIELD ofrece un conjunto robusto de herramientas para analizar datos ómicos, mientras que dsExposome proporciona características especializadas para el análisis de datos del exposoma. Ambas adiciones operan de manera federada mientras preservan la privacidad de los individuos. Todos los objetivos alcanzados en esta tesis doctoral están adjuntos a un proyecto europeo (ATHLETE), que requirió todos los desarrollos para el WP3, que está dedicado a las herramientas para el análisis de datos federados.

Además, el estudio aborda desafíos relacionados con la privacidad de los datos y las colaboraciones entre centros. La plataforma puede gestionar efectivamente conjuntos de datos más grandes y llevar a cabo análisis complejos sin comprometer la privacidad de los datos. Esta adaptabilidad abre el camino para futuras aplicaciones de DataSHIELD en otros campos de investigación, como la neuroimagen y la inteligencia artificial.

En conclusión, las nuevas herramientas y características mejoran significativamente la capacidad, escalabilidad y adaptabilidad de DataSHIELD. Estas mejoras acelerarán la adopción de métodos de análisis de datos federado en estudios multicéntricos, avanzando así la investigación mientras se mantiene rigurosamente la privacidad de los datos. Notablemente, DataSHIELD también contribuye a objetivos más amplios de reproducibilidad y transparencia en la investigación científica permitiendo que los resultados sean fácilmente verificables sin la necesidad de compartir datos, superando así las barreras tradicionales en la investigación colaborativa. El trabajo presentado en esta tesis doctoral sirve como un avance crítico en el análisis de datos federado, cubriendo los huecos entre la privacidad de los datos, la reproducibilidad y la investigación colaborativa en biomedicina.

Contents

1	Introduction	2
1.1	Background on data sharing	4
1.1.1	Evolution of Data Sharing and Collaboration	4
1.1.2	Challenges in Traditional Data Sharing Approaches	4
1.2	Background on federated data	5
1.2.1	Introduction to Federated Data	5
1.2.2	Federated Learning and Privacy-Preserving Techniques	7
1.2.3	Applications of Federated Data Analysis	10
1.2.4	Future Directions and Challenges	11
1.3	Background on DataSHIELD	13
1.3.1	DataSHIELD introduction	13
1.3.2	DataSHIELD objectives	14
1.3.3	DataSHIELD ethical and legal considerations	15
1.3.4	DataSHIELD Data Privacy and Security	17
1.3.5	Technical Overview of DataSHIELD	18
1.3.6	Data warehouses: Opal and Armadillo	19
1.3.7	Analysis stack. Packages and infrastructure	20
1.3.8	DataSHIELD analysis types	21
1.3.9	DataSHIELD privacy preserving mechanisms	26
1.3.10	DataSHIELD future privacy preserving mechanisms	31
1.3.11	Projects using DataSHIELD	31
1.4	Overview of Exposome Data Analysis	32
1.4.1	The exposome concept	32
1.4.2	Data collection	34
1.4.3	Data integration	39
1.4.4	Statistical analysis methods	40
1.4.5	Challenges and limitations	43
1.5	Overview of Omics Data Analysis	44
1.5.1	Introduction to Omics Data Analysis	45
1.5.2	Types of omic data	45
1.5.3	Data collection	46
1.5.4	Multi-omics data analysis	47
1.5.5	Statistical Analysis in Omics Data	48
2	Hypotheses and objectives	52
2.1	Hypotheses	53
2.2	Objectives	56
3	Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD	59
3.1	Disclaimer	60
3.2	Introduction	60
3.3	Design and implementation	61
3.3.1	The resources architecture	61
3.3.2	The <i>resourcer</i> R package	62
3.4	Results	64
3.4.1	Available resources extensions	67
3.4.2	Real data analyses	68
3.5	Future perspectives	71
3.5.1	DataSHIELD	71
3.5.2	Parallel computing	72
3.5.3	Omic and geographical data	72

3.5.4	Data cataloging	72
3.5.5	Other applications	73
4	OmicSHIELD: Federated privacy-protected meta- and mega-omic data analysis in multi-centre studies	77
4.1	Introduction	78
4.2	Methods	79
4.2.1	Overview of OmicSHIELD	80
4.2.2	Security and privacy	80
4.2.3	Omic analytic capabilities	80
4.2.4	GWAS: Federated population stratification, pooled and IPD meta-analysis and polygenic risk scores	84
4.2.5	Differential gene expression analysis and EWAS	84
4.2.6	Post-omic analyses and visualization	85
4.2.7	Use case 1: Multi-centric GWAS of CINECA data	85
4.2.8	Use case 2: DGE and EWAS analysis of HELIX data	86
4.3	Discussion	87
5	dsExposome: Secure and Privacy-Preserving Exposome Analysis using the DataSHIELD Infrastructure	92
5.1	Introduction	93
5.2	Methods	94
5.2.1	Opal data warehouse	94
5.2.2	Datasets	94
5.2.3	Exposome analysis capabilities	95
5.3	Results	97
5.3.1	Use case 1	97
5.3.2	Use case 2	98
5.3.3	Use case 3	101
5.4	Discussion	102
6	ShinyDataSHIELD - An R Shiny application to perform federated non-disclosive data analysis in multi-cohort studies	107
6.1	Introduction	108
6.2	Implementation	108
6.3	Use	110
6.4	Discussion	111
7	Breakthroughs Beyond Published Work	115
7.1	Introduction	116
7.2	Design and Development	116
7.3	Features and Functionality	117
7.3.1	Singular Value Decomposition (SVD)	117
7.3.2	k-means Clustering	118
7.3.3	k-nearest neighbors (kNN)	121
7.3.4	Factor Analysis of Mixed Data (FAMD)	122
7.4	Current status	126
7.5	Future developments	126
8	Application to real world data	128
8.1	unCoVer project	129
8.1.1	Project description	129
8.1.2	Methodology and Ethical Considerations	129
8.1.3	Tasks developed	130
8.1.4	Development and Improvement of DataSHIELD Packages	132

8.1.5	Academic outcomes and contributions	133
8.2	ATHLETE project	134
8.2.1	Project description	134
8.2.2	DataSHIELD Methods in Work Packages 1, 3 and 4	134
8.2.3	Tasks developed	136
8.2.4	Academic outcomes and contributions	136
9	Discussion	138
9.1	Introduction	139
9.2	Summary of key findings	139
9.3	Interpretation of findings	140
9.4	Strengths and contributions	140
9.5	Limitations	141
9.6	Limitations of the DataSHIELD infrastructure	141
9.7	Future steps	142
10	Conclusions	144
	References	147
	Appendices	162
	Publications derived from this work	163
	PhD Portfolio	163

List of Figures

1	Federated database system	7
2	Example Opal + DataSHIELD infrastructure diagram	21
3	Horizontal and vertical data	22
4	System protection elements flowchart	27
5	Analysis protection elements flowchart	28
6	Governance protection elements flowchart	30
7	Three different domains of the exposome	33
8	Evolution of number of papers related to exposome during last years	34
9	Libelium Air Quality Station installed on a pole	35
10	Personal exposure monitoring devices	36
11	ExWAS visualizations	41
12	Illumina MiniSeq System	46
13	Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer	47
14	Hierarchical clustering dendrogram	49
15	K-means 2D visualization	49
16	Network analysis visualization	50
17	A schematic diagram of a multi-site DataSHIELD infrastructure	64
18	Scheme of DataSHIELD implementation of genomic data analysis.	70
19	Scheme of DataSHIELD implementation of omic-related packages.	82
20	Types of omics data analyses implemented in <i>dsOmicsClient</i>	83
21	Configuration for multi-centric GWAS of CINECA data using OmicSHIELD.	85
22	Locus zoom plots of the top hit for the original data (left) and pooled fast GWAS (right).	86
23	Opal Data infrastructure of HELIX project.	87
24	Architecture of the Exposome Analysis Use Case	97
25	Manhattan plot of the ExWAS analysis	98
26	Architecture of the Geospatial Data Use Case.	99
27	Visualization of Mean Ground-Level Fine Particulate Matter	100
28	Manhattan plot of the ExWAS analysis	101
29	Connections interface. The illustrated configuration is a single server data source configuration with three different tables selected.	110
30	Selecting the tables to use. We can see, following Figure 1, that we have three available tables.	111
31	SVD block method schematic	118
32	K-means flowchart	120
33	K-means scatter plot visualization	121
34	KNN flowchart	122
35	FAMD flowchart	124
36	Life tables flowchart	125
37	unCoVer organization	130
38	unCoVer dashboard. Metadata page	131
39	unCoVer dashboard. Server connection page	131
40	ATHLETE project components	135

List of Tables

1	DataSHIELD Disclosure traps	29
2	Main characteristics of the proposed infrastructure for privacy-protected federated data analyses with big data	65
3	Available resources at the resourcer R package and extensions for genomic data.	67
4	Key aspects of OmicSHIELD.	80
5	Main analysis functions of OmicSHIELD.	81

6	Results presented on the publication by Warembourg et. al (2019) compared to the results obtained on the same analysis using dsExposome.	102
---	--	-----

1 Introduction

1.1 Background on data sharing

All the work conducted on this PhD thesis arises from one challenge, that is data sharing in the scenario of multi-center collaborative studies. So to understand the importance and need of the developments presented, it is key to understand and establish all the background on data sharing.

1.1.1 Evolution of Data Sharing and Collaboration

As the demand for data-driven decision-making has increased over the years, the need for efficient data sharing and collaboration has become more critical. Early data sharing methods were largely manual, involving the physical exchange of data on storage devices like floppy disks or the direct transfer of data between databases through file transfers or data import/export mechanisms [1]. These methods were limited by slow data transfer speeds, storage capacity constraints, and the lack of automation, which often resulted in time-consuming and error-prone processes [2]. Furthermore, these methods offered minimal support for data provenance, data quality, and data integration, making it difficult to trace the origin of data, assess its accuracy, or combine it with other datasets for comprehensive analysis [3].

As the limitations of early data sharing methods became apparent, researchers and practitioners sought ways to overcome these challenges, ultimately leading to the development of distributed computing [4]. The primary objective of distributed computing is to distribute the processing of data and computational tasks across multiple connected computers, often referred to as nodes, which can collaborate to solve complex problems more efficiently than a single machine. The emergence of distributed computing has facilitated more effective data sharing, as it enables data to be stored and processed closer to its source, reducing data transfer latency and improving overall system performance [5]. Additionally, distributed computing allows for enhanced fault tolerance and redundancy, as data and computations can be replicated across multiple nodes, providing greater resilience against hardware failures or data loss [6]. By providing a more robust infrastructure for data sharing, distributed computing has laid the foundation for more advanced collaborative approaches, such as federated data systems, which can address contemporary data privacy, security, and interoperability challenges [7]. Federated data systems are a type of data management architecture that enables the integration, analysis, and sharing of data across multiple, geographically distributed data sources while maintaining the privacy, security, and local control of the original data.

In the domain of life sciences, the need for collaborative data sharing has become increasingly vital, as researchers from various disciplines and institutions work together to address complex challenges such as understanding diseases, discovering new drugs, and advancing personalized medicine [8]. Large-scale, multi-institutional research projects, such as the Human Genome Project [9] and The Cancer Genome Atlas [10], have demonstrated the value of data sharing in driving scientific discoveries and facilitating collaboration among researchers worldwide. However, the sensitive nature of the data involved in these studies, including patient genomic information, clinical records, and biometric data, raises significant privacy and ethical concerns [11]. As a result, the life sciences domain has seen the development of specialized data sharing platforms and federated data systems [12, 13, 14].

In the field of social sciences, the growing availability of large-scale, fine-grained social, economic, and demographic data has created unprecedented opportunities for researchers to study human behavior, social networks, and societal dynamics [15]. However, these data sources, which include government records, social media data, and survey responses, also present unique challenges related to data privacy, security, and data quality [16]. Researchers in social sciences must navigate these challenges while sharing and collaborating with data to ensure the protection of individuals' privacy and comply with regulations like General Data Protection Regulation (GDPR) [17]. Federated data systems and privacy-preserving techniques have therefore gained increasing relevance in the social sciences domain, as they allow researchers to collaboratively analyze data while addressing privacy and security concerns [18, 19].

1.1.2 Challenges in Traditional Data Sharing Approaches

Traditional data sharing approaches, such as direct data transfers or centralized data repositories, have been fundamental to fostering collaborative research and analysis. However, with the advent of the "fourth paradigm" of science [20], characterized by data-intensive scientific discovery and the increasing volume,

variety, and complexity of data, traditional data sharing approaches have encountered several significant challenges that limit their effectiveness and applicability in today’s data-driven environments [21]. These challenges encompass not only privacy and security concerns but also issues related to data integration, data quality, scalability, and the need for real-time analysis. Addressing these challenges associated with traditional data sharing approaches is essential for advancing scientific research and preparing for the future needs of data-driven domains, ensuring that the power of data can be harnessed effectively and responsibly.

To address the privacy concerns, a number of regulations and guidelines have been established worldwide to govern the collection, storage, and sharing of personal data. One of the most prominent regulations is the GDPR [17], which enforces stringent data protection rules for organizations handling the personal data of EU residents. Similarly, in the United States, the Health Insurance Portability and Accountability Act (HIPAA) [22] sets specific requirements for the protection of personal health information. Other countries and regions have also introduced their own data protection laws and guidelines, creating a complex regulatory landscape for organizations and researchers involved in data sharing [23].

These privacy regulations underscore the importance of adopting data sharing techniques that ensure the protection of sensitive information while still allowing researchers and organizations to derive valuable insights from shared data. The development of novel data sharing solutions, such as federated data systems, can play a crucial role in addressing these challenges and promoting responsible data sharing practices in line with privacy regulations.

In addition to data privacy concerns, traditional data sharing approaches also face challenges related to data integration, data quality, scalability, and real-time analysis. Data integration becomes complex as the variety of data sources and formats increases, making it difficult to combine and analyze heterogeneous datasets [24]. Data quality issues, such as inconsistency, incompleteness, and inaccuracy, can hinder the reliability and validity of analyses, leading to potentially misleading conclusions [25]. Scalability is another challenge, as the volume of data continues to grow exponentially, requiring efficient methods to store, manage, and process large-scale datasets [26]. Finally, the need for real-time analysis is becoming increasingly important in many applications, such as monitoring and decision-making in healthcare, finance, and social sciences, necessitating data sharing approaches that enable timely access to and analysis of relevant data [27]. Addressing these challenges is essential for developing more robust data sharing techniques that can meet the demands of modern data-driven environments.

1.2 Background on federated data

Having just explored the complexities and challenges of traditional data sharing methodologies, it becomes evident that the centralization and full data access approach isn’t always achievable. This realization naturally leads us to seek out innovative strategies that address these concerns while still promoting collaboration and knowledge exchange. In the following section federated data will be explored, highlighting its potential in circumventing traditional barriers and opening new avenues for collaborative research.

1.2.1 Introduction to Federated Data

Federated data refers to a distributed data management architecture that addresses the limitations of traditional data sharing methods by enabling the integration, analysis, and sharing of data across multiple, geographically distributed data sources while preserving the privacy, security, and local control of the original data. This approach tackles various challenges related to data management, such as ensuring data confidentiality, accommodating diverse data formats, maintaining data quality and handling ever-growing datasets.

Key characteristics of federated data systems encompass decentralized data storage, privacy preservation, interoperability, scalability and efficiency. In federated data systems, data remains stored locally at its original source, and access is granted to authorized parties without requiring the physical transfer of data. Privacy is maintained through techniques such as differential privacy [28], secure multi-party computation [29], and homomorphic encryption [30], ensuring that sensitive data stays safeguarded during analysis and sharing. Interoperability is achieved by facilitating the integration and analysis of heterogeneous data sources through

common data models, ontologies, and query languages, fostering seamless data sharing and collaboration. Lastly, federated data systems improve scalability and efficiency by distributing computation tasks and processing data locally, which allows for the efficient handling of large-scale datasets and reduction of data transfer latency, ultimately enhancing overall system performance.

1.2.1.1 Federated data vs. centralized data

Federated data systems and centralized data systems exhibit several key differences in terms of data storage, privacy, control, scalability, fault tolerance, and integration. In federated data systems, data remains stored locally at its original source, preserving privacy by allowing data analysis without requiring the physical transfer of data. This approach enables local data sources to maintain control over their data while distributing computation tasks and processing data locally, which enhances scalability and reduces data transfer latency. Furthermore, federated data systems are more fault-tolerant, as the failure of one data source does not necessarily impact the entire system, and they often involve more complex data integration tasks due to the need to accommodate heterogeneous data sources [31].

Alternatively, centralized data systems gather and store data in a central repository, potentially leading to increased privacy risks and necessitating local data sources to cede a degree of control over their data. Centralized systems also face challenges in scaling due to the need to transmit and process all data in a central location and may experience disruptions if the central repository fails [32]. However, centralized data systems can simplify data integration by consolidating data into a unified format and structure within the central repository.

1.2.1.2 Federated database systems

Federated database systems are a way to connect multiple databases, making it easier to access and work with data from different sources. It can be abstractly interpreted as having different books (databases) with information stored in various ways (structures). These database systems act like a librarian (middleware layer) who helps you find the information you need from all the books without having to read each one individually. This is important because it allows people to easily find and use data from different places while keeping the information safe and private in its original location. This idea is illustrated on fig. 1.

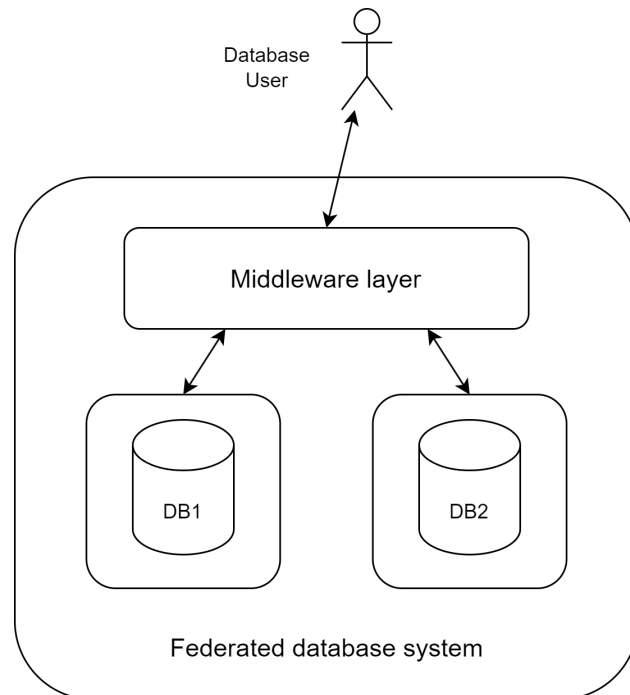


Figure 1: Federated database system: The proposed database system contains two different databases controlled by a middleware layer.

In a federated database system, the middleware layer plays an essential role in controlling data privacy. It is responsible for managing the communication between the participating databases and the end-users or applications, ensuring that only authorized access to the data is granted; it is also responsible of applying the different privacy preserving techniques implemented on the database, whether it is differential privacy, encryption, data anonymization or any other technique.

1.2.2 Federated Learning and Privacy-Preserving Techniques

Federated learning is a decentralized machine learning approach that enables multiple organizations or individuals to collaboratively train a shared model without the need to directly share sensitive or private data [33]. In the era of big data, privacy concerns have grown exponentially due to the vast amount of personal information generated and stored online. Traditional machine learning methods often require the centralization of data, which raises issues related to privacy, security, and data ownership. Federated learning addresses these concerns by allowing participants to maintain control of their data and share only model updates or gradients, thereby reducing the risk of privacy breaches [34].

In a federated learning setup, each participant trains a local model on their own data and communicates the model updates to a central server. The server aggregates the updates, updates the global model, and distributes the updated global model back to the participants. This process is iteratively repeated until the model reaches a desired level of accuracy or convergence. The decentralized nature of federated learning enables better data privacy and security, as well as increased efficiency in certain scenarios where data transmission is a bottleneck.

Several privacy-preserving techniques have been developed to further enhance the security of federated learning. These techniques aim to protect the privacy of individual participants and their data, even when adversaries have access to the shared model updates. As previously stated, these techniques include secure multi-party computation (MPC), homomorphic encryption, and differential privacy. These techniques can be used individually or in combination to ensure the privacy and security of federated learning systems, safeguarding sensitive data from potential breaches or misuse.

1.2.2.1 Secure multi-party computation

Secure Multi-party Computation (MPC) [35] is a cryptographic technique that enables multiple parties to collaboratively compute a function on their private data while ensuring that no information about the data is revealed, apart from the final output of the function. The primary goal of MPC is to maintain the privacy of each party's data throughout the computation process.

Here's a more detailed explanation of how secure multi-party computation works:

1. **Encoding the inputs:** To begin with, each party's data is encoded in such a way that it is split into multiple shares. These shares are distributed among the participating parties using a secret sharing scheme, such as Shamir's secret sharing [36]. Secret sharing ensures that individual shares do not reveal any information about the original data, and a predetermined threshold number of shares is required to reconstruct the original data.
2. **Performing the computation:** After encoding and distributing the data shares, the computation is performed on these shares without directly accessing the original data. The parties collaboratively execute a predefined protocol designed to compute the function on the shares, in a manner that preserves the privacy of the data. The protocol is typically composed of a series of basic operations, such as addition, multiplication, and comparison, which can be performed on the shares while maintaining the privacy of the underlying data.
3. **Reconstructing the output:** Once the computation on the shares is completed, the parties can reconstruct the final output by combining their respective output shares. The output shares are designed in such a way that they do not reveal any information about the original data. The reconstructed output is the same as if the function had been applied directly to the original data, ensuring the correctness of the computation.

Secure MPC can be implemented using various cryptographic primitives and protocols, including secret sharing [36], garbled circuits [37], and oblivious transfer [38]. Each of these techniques has its own advantages and trade-offs in terms of efficiency, security, and communication overhead:

- **Secret sharing:** In this approach, the inputs are divided into shares using a secret sharing scheme, and the computation is performed on the shares using linear operations. This method is efficient for linear operations, but it can be less efficient for non-linear operations, such as multiplication.
- **Garbled circuits:** This technique involves representing the function as a Boolean circuit and encoding the inputs as encrypted labels for the circuit's input wires. The parties exchange encrypted labels during the computation, and the circuit is evaluated in a privacy-preserving manner using Yao's garbled circuit construction. While garbled circuits can be used to evaluate any function, they tend to be less communication-efficient than secret sharing-based methods.
- **Oblivious transfer:** This primitive allows a sender to transmit one of several possible messages to a receiver, without the sender learning which message was selected or the receiver learning anything about the other messages. Oblivious transfer is used in combination with other techniques, such as secret sharing or garbled circuits, to implement secure multi-party computation protocols.

While MPC offers promising privacy-preserving capabilities for life sciences, it faces challenges such as computational and communication overhead, limited scalability, complexity of protocols, restricted support for complex analyses, and trade-offs between privacy and utility. These downsides can hinder the adoption of MPC in life sciences applications, where large-scale data processing, real-time analysis, collaborations involving multiple parties, and high-quality results are crucial. Ongoing research seeks to address these limitations and enhance the applicability of MPC in the life sciences domain.

1.2.2.2 Homomorphic encryption

Homomorphic encryption is a form of encryption that allows computations to be performed directly on encrypted data, without the need for decryption [39]. This powerful cryptographic technique enables privacy-preserving data processing and analysis, as the data remains encrypted throughout the computation process.

In the context of federated learning, homomorphic encryption can be employed to protect the privacy of individual participants and their data when sharing model updates with the central server.

The primary goal of homomorphic encryption is to enable the computation of functions on encrypted data in such a way that the result of the computation, when decrypted, is the same as if the function had been applied to the original plaintext data. This is achieved through the design of encryption schemes that support specific homomorphic operations, such as addition or multiplication, while preserving the encrypted data's structure.

There are various types of homomorphic encryption schemes, including partially homomorphic encryption (PHE), somewhat homomorphic encryption (SHE), and fully homomorphic encryption (FHE) [40, 41, 42]. PHE and SHE support a limited number of homomorphic operations, while FHE allows for an unlimited number of operations on encrypted data. However, FHE schemes typically come with high computational and communication overheads, making them less practical for large-scale federated learning systems.

In recent years, researchers have explored the use of homomorphic encryption in federated learning to enable secure and privacy-preserving model training [43, 44]. For example, Aono et al. (2016) proposed a privacy-preserving linear regression method based on homomorphic encryption, allowing participants to share encrypted model updates with the central server without revealing their private data. Kim et al. (2018) developed a secure federated learning framework using homomorphic encryption and secure MPC, combining the strengths of both techniques to protect participants' privacy while sharing model updates.

Despite its potential, the practical implementation of homomorphic encryption in federated learning remains challenging due to the associated computational and communication overheads. However, ongoing research efforts are focused on developing more efficient homomorphic encryption schemes and tailored protocols that can be effectively integrated into federated learning systems.

1.2.2.3 Differential privacy

Differential privacy is a privacy-preserving technique that provides strong guarantees on the privacy of individual data while allowing statistical analysis of aggregated data [28]. In the context of federated learning, differential privacy can be employed to protect the privacy of individual participants and their data when sharing model updates with the central server.

The main idea behind differential privacy is to introduce carefully controlled noise to the output of a query or a computation, so that the presence or absence of any individual's data in the dataset does not significantly affect the results. This ensures that an adversary cannot infer sensitive information about an individual participant even when they have access to the noisy output.

Differential privacy is formally defined using the concept of ϵ -differential privacy, where ϵ is a parameter that controls the level of privacy. A smaller ϵ value provides stronger privacy guarantees, while a larger ϵ value results in a higher utility or accuracy of the computation. The formal definition of ϵ -differential privacy is as follows:

A randomized mechanism M is ϵ -differentially private if, for any two adjacent datasets D_1 and D_2 that differ in only one individual's data, and for any subset of outputs S , the inequality eq. (1) holds.

$$Pr[A(D_1) \in S] \leq \exp(\epsilon) \cdot Pr[A(D_2) \in S] \quad (1)$$

Where A is a randomizing algorithm that takes a dataset D as input, and ϵ is a positive real number, called the privacy parameter, that defines the level of privacy (closer to 0 indicates more privacy). Formally, S refers to all subsets of the image of A i.e. all possible subsets of the output values of the algorithm A . The inequality refers to datasets D_1 and D_2 that differ on a single element (i.e., the data of one person). If met, it indicates that after application of the randomizing algorithm, the probability that $A(D_1)$ lies in S is less than or equal to the probability that $A(D_2)$ is in S multiplied by the constant $\exp(\epsilon)$ that gets closer to 1.00 as ϵ falls to 0.

In federated learning, differential privacy can be applied by adding noise to the model updates or gradients during the training process [45, 46]. This can be achieved using various techniques, such as the Gaussian mechanism, which adds Gaussian noise to the gradients, or the Laplace mechanism, which adds Laplace noise. The amount of noise added depends on the desired privacy level (ϵ) and the sensitivity of the function being computed, which is the maximum difference in the output when the input data changes by one individual's data.

One of the key challenges in applying differential privacy to federated learning is balancing the trade-off between privacy and utility. As the level of privacy increases, the noise added to the model updates also increases, which can negatively impact the accuracy of the trained model. However, recent research has demonstrated that it is possible to achieve reasonable privacy guarantees while maintaining a high level of utility in federated learning [47].

In conclusion, differential privacy is a powerful technique that can be used to enhance the security of federated learning systems by providing strong privacy guarantees for individual participants. By carefully controlling the amount of noise added to model updates during the learning process, differential privacy ensures that sensitive information about individual participants remains protected while still allowing for accurate model training.

1.2.3 Applications of Federated Data Analysis

In this section, we will explore the applications of federated data analysis in three key areas: healthcare and genomics research, financial services, and fraud protection. We will discuss how federated data systems are transforming these domains by facilitating secure, efficient, and collaborative data analysis, unlocking new opportunities for innovation and improved decision-making.

1.2.3.1 Healthcare and genomics research

Federated data analysis has had a significant impact on the healthcare domain, enabling secure, collaborative research and improved decision-making while addressing privacy concerns and regulatory requirements. It has proven to be transformative in many domains, enabling the analysis of data from multiple institutions and enhance the understanding of complex diseases [48, 49].

A notable example of the application of federated data analysis in healthcare is the i2b2 (Informatics for Integrating Biology and the Bedside) project [50]. This initiative aims to create a scalable informatics framework that allows medical researchers to access patient data from multiple institutions while preserving patient privacy. By utilizing federated data analysis techniques, the i2b2 project has facilitated numerous multi-center studies, leading to significant advancements in the understanding and treatment of various medical conditions.

Another example is the Global Alliance for Genomics and Health (GA4GH), which employs federated data analysis to enable secure, collaborative research on genomic data [51]. By developing standards and tools for sharing and analyzing genomic and clinical data in a federated manner, GA4GH has fostered global collaborations and accelerated the discovery of novel disease biomarkers, therapeutic targets, and diagnostic tools [52].

Several studies have demonstrated the value of federated data analysis in healthcare and genomics research. For instance, Roth et al. [53] employed federated learning to build classification models of breast cancer, using federated data from seven clinical institutions, that approach led to better models than the ones achieved at the institution level. Similarly, Lee et al. [54] applied federated learning to train a model based on pattern mining in order to predict cardiovascular diseases.

These examples demonstrate how federated data analysis has become a critical enabler for healthcare and genomics research, allowing institutions to collaborate and share insights while maintaining data privacy and adhering to regulatory requirements.

1.2.3.2 Financial services and fraud detection

Federated data is not only relevant on the bio-sciences domain, it has also made significant contributions to the domain of financial services, specifically in the area of fraud detection. By allowing financial institutions to securely collaborate and share insights from their data, federated data analysis has improved the accuracy and effectiveness of fraud detection models while preserving the privacy of sensitive customer information [55].

Another application of federated data analysis in the financial services sector is in the area of anti-money laundering (AML). AML efforts often require the collaboration of multiple institutions to identify and report suspicious activities. A tool for that matter has been developed by the AI department of the Chinese bank WeBank [56], they claim to have united several banks to train AML models jointly [57].

Federated data analysis has also been applied in the context of credit scoring and risk assessment. For example, Li et al. [58] proposed a federated learning approach to predict credit risk using data from multiple financial institutions, demonstrating improved performance compared to individual institution models.

These examples illustrate the potential of federated data analysis to transform the financial services sector by enabling secure, collaborative data analysis and improving the efficiency and effectiveness of fraud detection and risk management efforts.

1.2.4 Future Directions and Challenges

Federated learning and federated data are gaining significant attention due to their potential to revolutionize data privacy, distributed computing, and machine learning. In this context, Kairouz et al. [59] noted several key research areas have emerged, such as improving efficiency and effectiveness, preserving the privacy of user data, ensuring fairness and addressing sources of bias, and addressing system challenges. A key property of many of these problems is that they are inherently interdisciplinary, requiring not just machine learning expertise but also techniques from distributed optimization, cryptography, security, differential privacy, fairness, compressed sensing, systems, information theory, statistics, and more [33]. Many of the hardest problems lie at the intersections of these areas, and collaboration will be essential to ongoing progress. In the following sections, we will discuss these critical aspects and their implications for the future of federated learning and federated data systems.

1.2.4.1 Efficiency and effectiveness

Tackling multiple tasks is an aspect of real-world scenarios that federated learning systems should be capable of handling. Smith et al. [60] proposed a multi-task learning framework that learns from multiple tasks. Further research is required to develop efficient multi-task learning approaches that can be applied to federated learning systems, allowing them to handle tasks with diverse objectives and data distributions.

Adapting machine learning workflows to the federated setting poses another challenge. Traditional workflows assume centralized data storage, which is not the case for federated learning systems. Researchers must develop methods to adapt model evaluation, hyperparameter tuning, and other aspects of machine learning workflows for federated settings without compromising data privacy [61].

Communication and compression are vital in federated learning systems, as they impact both efficiency and effectiveness. Strategies like gradient compression [62] can help reduce the communication overhead between devices and central servers. However, further research is needed to develop novel communication and compression techniques that can provide improved trade-offs between accuracy and communication costs.

Lastly, applying federated learning to a broader range of machine learning problems and models is a challenge that must be tackled. Most existing federated learning research focuses on supervised learning and deep learning models. Expanding federated learning to other areas, such as reinforcement learning, unsupervised learning, and different model architectures [63], will help unlock its full potential across various domains.

In conclusion, addressing these challenges related to personalization, multi-task learning, adapting workflows, communication, and expanding the application scope will pave the way for more efficient and effective federated learning systems in the future.

1.2.4.2 Privacy of user data

Addressing privacy concerns in federated learning requires a deeper understanding of both *"what"* functions being computed and the manner in *"how"* computations are executed. This includes considering who can access intermediate results and how to protect against adversarial attacks.

As we have previously discussed, one way to address privacy concerns is through differential privacy, which is a technique that ensures the output of a computation remains indistinguishable when an individual's data is added or removed from the dataset. The challenge lies in adapting this technique to real-world computations, taking into account factors such as the nature of the data (e.g., time-series data), the presence of multiple independent actors, client availability (e.g. are all parties available at all times?); all of that have to be leveraged in order to determine the parameters of differential privacy that provide privacy and usability of the results.

To address the computation execution aspect, several techniques have been proposed, such as multi-party computation (MPC), homomorphic encryption, and trusted execution environments (TEE). MPC allows multiple parties to jointly compute a function while keeping the data private. Although MPC has been deployed at scale, it remains more communication and computation-intensive than its insecure counterparts. For TEE, the challenge is to develop a platform that is free from exploitable vulnerabilities [64].

Another aspect that requires attention is verifiability, which refers to proving that parties have executed their parts of the computation faithfully. The main open problem in this area is protecting federated learning systems against an adversarial server. Designing robust techniques to ensure verifiability while maintaining efficiency and effectiveness is a critical challenge for future research.

In summary, the future challenges in preserving privacy in federated learning involve improving and adapting techniques like differential privacy, multi-party computation, and trusted execution environments, as well as addressing verifiability and protection against adversarial servers.

1.2.4.3 Fairness and addressing sources of bias

Federated learning provides a decentralized approach to machine learning that can offer privacy benefits, but it also faces challenges related to fairness and addressing sources of bias. Machine learning models can sometimes exhibit unintended behaviors, leading to fairness concerns [65].

One such issue is individual fairness, where people with similar characteristics receive different outcomes. Another issue is demographic fairness, where specific groups (e.g., race, gender) receive different outcomes, violating the principle that users should receive the same treatment regardless of their group membership [66].

Bias in training data is a critical factor when considering fairness in federated learning models. Ensuring representative datasets can help improve both the overall quality of downstream models and their fairness.

Having explicit access to demographic information (e.g., race, gender) is essential for many existing fairness criteria, including individual and demographic fairness. However, federated learning contexts often require fairness considerations even when sensitive attributes are unavailable [67]. This situation can arise when developing personalized language models or fair medical image classifiers without knowing additional demographic information about individuals.

One approach to address fairness without access to sensitive attributes is to focus on equal access to effective models. This interpretation of fairness aims to maximize model utility across all individuals, regardless of their (unknown) demographic identities, and regardless of the "goodness" of an individual outcome [68].

Since federated learning is often deployed in privacy and fairness-sensitive contexts, tensions between privacy and fairness objectives can be magnified. Further research is needed to address the potential tension between achieving privacy, fairness, and robustness in both federated and centralized learning [69].

Federated learning presents unique opportunities to improve the diversity of stakeholders and data incorporated into learning, which could enhance both the overall quality of downstream models and their fairness due to more representative datasets. However, federated learning also brings about fairness-related challenges not present in the centralized training regime, necessitating new solutions to address these concerns.

1.2.4.4 System challenges

Frequent and large-scale deployment of updates in federated learning systems poses a challenge, as monitoring and maintaining become increasingly complex in decentralized environments [61]. Research efforts need to focus on developing scalable and reliable methods for managing updates while maintaining the benefits of decentralized learning.

Differences in node availability can introduce various forms of bias, which can affect the overall performance and fairness of federated learning systems [70]. Defining, quantifying, and mitigating these biases remain an essential direction for future research to ensure the robustness and fairness of federated learning.

Tuning system parameters in federated learning is difficult due to the existence of multiple, potentially conflicting objectives. For example, optimizing communication efficiency may conflict with preserving user privacy. Developing strategies that balance these objectives will be crucial to ensure the practicality and adoption of federated learning systems.

Running machine learning workloads on end-user devices is constrained by the lack of a portable, fast, small footprint, and flexible runtime for on-device training [71]. Developing efficient and effective runtimes that can adapt to a wide range of device capabilities will be essential to unlock the full potential of federated learning across diverse applications and settings.

In summary, addressing system challenges in federated learning involves tackling issues related to large-scale updates, monitoring, debugging, device availability biases, system parameter tuning, and developing efficient runtimes for on-device training. Future research efforts should focus on these areas to ensure the successful deployment of federated learning systems in real-world applications.

1.3 Background on DataSHIELD

Having elucidated the nuances of federated data, we appreciate its potential in transforming the data-sharing landscape. However, effectively utilizing federated data requires tools and platforms that can harness its unique structure and challenges. DataSHIELD emerges as one such solution, aiming to streamline federated data analysis.

1.3.1 DataSHIELD introduction

In today's data-driven world, researchers and healthcare professionals often need to analyze sensitive individual-level data, from multiple sources [72]. This data can provide valuable insights into various aspects of human health, behavior, and the environment. However, sharing such sensitive data across institutions can pose significant ethical, legal, and privacy concerns [73]. That's where DataSHIELD [74] comes in.

DataSHIELD is a cutting-edge technology designed to address the challenges of securely accessing and analyzing sensitive data from multiple sources. It enables researchers to work with individual-level data without having to physically share it with others. Although initially developed for use in biomedical and social sciences, DataSHIELD's flexible and versatile nature makes it suitable for any setting where sensitive data analysis is required, but data cannot be physically shared.

Combining diverse data sets enables researchers to access a greater volume of data, which in turn leads to increased statistical power [75]. This enhanced statistical power allows for more accurate assessment of relationships between omics, exposures, genomics, and health conditions. DataSHIELD serves as the enabler

of this innovative approach by providing a secure and privacy-preserving platform for analyzing sensitive data across multiple sources.

At its core, DataSHIELD employs a federated analysis model. This means that researchers can analyze data from multiple sources simultaneously, without actually accessing the raw individual-level data. Instead, DataSHIELD sends analysis requests from a central analysis machine (client) to data-holding machines (servers), which store the harmonized data to be co-analyzed, this approach is summarized as "taking the analysis to the data, not the data to the analysis" [76]. It is important to remark that the data never leaves the analysis servers, the only information leaving the servers is non-disclosive aggregated statistics (e.g. the mean of a dataset).

This innovative approach allows researchers to collaborate and analyze data across institutions while ensuring that sensitive individual-level data remains secure and confidential. Moreover, DataSHIELD's infrastructure is built on free, open-source software, making it accessible and cost-effective [77] for a wide range of users.

One of the key advantages of DataSHIELD being open source is that it fosters a trustworthy relationship between researchers and data owners. By making the source code openly accessible, data owners can independently verify the algorithms and methods being used to analyze their sensitive data. This transparency allows data owners to assess if the provided solutions comply with their security and risk policies, ensuring that their data is handled responsibly and securely.

Moreover, having an open-source platform encourages collaboration and innovation within the research community [78]. Developers, researchers, and other stakeholders can contribute to the improvement and expansion of DataSHIELD's features, making it more adaptable and versatile over time. This collaborative approach accelerates the development of new functionality and ensures that the platform remains up-to-date with the latest advancements in data analysis and privacy protection.

In addition, open-source software like DataSHIELD often benefits from increased security due to the scrutiny of a large community of developers and users. Any potential vulnerabilities can be identified and addressed more quickly, enhancing the platform's overall security and reliability [79].

1.3.2 DataSHIELD objectives

DataSHIELD, a publicly-funded project, has emerged as a vital tool for researchers working with sensitive data by addressing key challenges in privacy, security, and collaboration. With a growing and increasingly diverse user base, the project is establishing itself as a leading platform for secure and privacy-preserving data analysis. As DataSHIELD continues to evolve, its main objectives focus on fostering innovation, ensuring sustainability, and promoting global engagement in order to maintain its commitment to providing a robust, flexible, and reliable solution for the research community. The main objectives of DataSHIELD are:

1. To enable the secure, privacy-preserving analysis of sensitive data from multiple sources, without the need to share or physically transfer individual-level data.

DataSHIELD's primary objective is to facilitate the analysis of sensitive data from multiple sources while maintaining privacy and security. It achieves this by keeping individual-level data at the source and only exchanging aggregated, non-disclosive results across study sites. This approach reduces the risk of data breaches and addresses privacy concerns associated with sharing sensitive data, thereby enabling more collaborative research across different organizations and jurisdictions.

2. To provide a flexible and scalable architecture that supports various data formats and storage systems, as well as a wide range of analytical methods, including individual person data (IPD) and study level meta-analysis (SLMA).

DataSHIELD is designed to support various data formats, storage systems, and analytical methods, making it adaptable to a wide range of research scenarios. Its flexibility allows researchers to work with different types of data, while its scalable architecture ensures that DataSHIELD can accommodate the growing needs of the research community, including complex analyses and increasing data volumes.

3. To continuously develop, update, and extend the functionality of DataSHIELD, including client-side and server-side functions, ensuring quality assurance and comprehensive documentation.

DataSHIELD is committed to the ongoing improvement and expansion of its features, both on the client-side and server-side. The project focuses on ensuring quality assurance, comprehensive documentation, and the integration of new analytical methods, which helps researchers to effectively utilize DataSHIELD and stay up-to-date with the latest advancements in data analysis.

4. To address data governance requirements and facilitate compliance with relevant ethical, legal, and institutional frameworks.

DataSHIELD places a strong emphasis on meeting the ethical, legal, and institutional frameworks that govern sensitive data usage. By incorporating mechanisms to ensure compliance with these requirements, DataSHIELD aims to create a trusted environment for data analysis and collaboration, while also streamlining the process of obtaining necessary permissions and approvals.

5. To transition from a small-scale research software project to a larger, community-driven, meritocratic governance model, overseen by a consortium Steering Committee or Advisory Board.

Recognizing the need to adapt to its growing user base, DataSHIELD aims to transition from a "benevolent dictator" governance model to a more collaborative, meritocratic approach. Establishing a consortium Steering Committee or Advisory Board will help to better engage the global DataSHIELD community and ensure that the project benefits from diverse perspectives and expertise.

6. To explore and implement sustainable funding and resource models, including the provision of training, consultancy, support for implementation, targeted extension of functionality, and the development of specialized add-ons or commercially-oriented product editions.

DataSHIELD is exploring ways to secure long-term funding and resources to support its continued development and growth. This includes offering training, consultancy, implementation support, and targeted functionality extensions through service contracts, as well as developing specialized add-ons or commercially-oriented product editions to generate revenue while meeting the varying needs of its users.

7. To foster collaboration and engagement with the global DataSHIELD community, including researchers, developers, and commercial partners, in order to advance the development and application of the project across various domains, such as large-scale epidemiological studies, health service data, and 'omics research.

DataSHIELD aims to build a thriving community of researchers, developers, and commercial partners who can collectively contribute to the project's advancement. By encouraging collaboration and engagement, DataSHIELD can harness the collective knowledge and experience of its community to address new challenges, explore novel applications, and ultimately improve the privacy and security of data analysis across various domains.

1.3.3 DataSHIELD ethical and legal considerations

As a powerful tool for securely analyzing sensitive data, DataSHIELD brings forth a set of ethical and legal considerations that need to be addressed to ensure the responsible use of the technology. In this section, we will delve into the ethical principles that guide the DataSHIELD project, the legal frameworks and data governance mechanisms it adheres to, and how the project aims to maintain a balance between data protection and enabling valuable research collaborations. We will also discuss how DataSHIELD addresses the challenges of data sharing, consent, and confidentiality, while complying with relevant regulations such as the General Data Protection Regulation (GDPR) in Europe [80]. Furthermore, we will explore the project's initiatives to integrate data governance rights and obligations into data sharing agreements, and its plans to streamline data access through research passporting, enhancing accountability and transparency in the process.

The ethical principles that guide the DataSHIELD project focus on ensuring that the platform is used responsibly while safeguarding the privacy and confidentiality of sensitive data. These principles include:

1. **Respect for autonomy:** DataSHIELD is designed to respect the autonomy of data subjects by enabling data analysis without direct access to individual-level data. This approach ensures that individuals' rights to control their personal information are protected, while still allowing researchers to gain valuable insights from the data.
2. **Non-maleficence and beneficence:** The project aims to minimize any potential harm resulting from the misuse or unauthorized access to sensitive data. By employing secure federated analysis techniques, DataSHIELD ensures that the benefits of research collaborations are maximized without compromising the privacy and well-being of data subjects.
3. **Privacy and confidentiality:** DataSHIELD's technology adheres to the principle of privacy by design, integrating data protection measures at every stage of the data analysis process. This approach not only maintains the confidentiality of sensitive information but also helps build trust among data subjects, researchers, and data custodians.
4. **Transparency and accountability:** DataSHIELD is committed to operating in a transparent and accountable manner. This includes clearly communicating its objectives, methods, and any potential risks or limitations of the platform, as well as maintaining an open dialogue with stakeholders and the wider research community.
5. **Social responsibility:** DataSHIELD recognizes the importance of balancing the need for scientific advancement with the ethical responsibilities that come with handling sensitive data. The project actively promotes responsible data sharing and collaborative research while ensuring compliance with relevant ethical and legal frameworks [81].

These ethical principles guide the development and implementation of DataSHIELD, ensuring that the project remains focused on providing a secure, privacy-preserving platform for data analysis, while upholding the rights and well-being of data subjects and the broader research community.

DataSHIELD adheres to various legal frameworks and data governance mechanisms, depending on the jurisdiction and specific requirements of each data source. Some of the key legal frameworks and governance aspects that DataSHIELD considers are:

1. **General Data Protection Regulation (GDPR):** General Data Protection Regulation (GDPR): The GDPR, in general, is research-friendly and aims to enable the free flow of data. It permits the use of data for research purposes under Article 9(2)j [**<empty citation>**], but leaves the specific regulation and appropriate safeguards to national laws. This creates national differences in the legal basis for processing data for scientific purposes, posing a significant challenge for data sharing across borders. According to GDPR Article 89.1, these safeguards should ensure that technical and organizational measures, such as pseudonymization, are in place to respect the principle of data minimization. When possible, research purposes should be fulfilled using further processing that does not permit or no longer permits the identification of data subjects. DataSHIELD's federated analysis approach aligns with the GDPR's focus on protecting individuals' privacy rights while navigating the complexities arising from differences in national laws.
2. **National and regional data protection laws:** DataSHIELD respects the specific data protection laws and regulations of each country or region in which it operates. This may involve adhering to additional or more stringent privacy requirements, depending on the jurisdiction.
3. **Data Access Committees and ethical approval:** DataSHIELD acknowledges the importance of obtaining permissions from Data Access Committees and securing ethical approval for research projects when required. These governance mechanisms help ensure that data are used responsibly and in accordance with the terms set by data providers and relevant ethics committees.
4. **Legal basis for data processing:** Before any analysis takes place, DataSHIELD users must confirm the legal basis for data processing, which may involve seeking consent from data subjects, complying

with legal obligations, or pursuing legitimate research interests. Ensuring a valid legal basis for data processing is crucial to upholding data subjects' rights and complying with data protection regulations.

5. Data sharing agreements: DataSHIELD aims to incorporate key data governance rights and obligations into data sharing agreements between participating parties. This approach helps streamline data access, while also providing a foundation for enforcing and updating governance structures as needed.
6. Research passporting: DataSHIELD is working towards a system of research passporting, allowing researchers to obtain permission-in-principle to work with specific data sources under predefined conditions. This mechanism simplifies the data access process and provides a means of sanctioning those who violate data governance agreements.

By adhering to these legal frameworks and data governance mechanisms, DataSHIELD ensures that its platform remains compliant with relevant regulations and ethical standards, protecting the rights of data subjects and fostering a responsible research environment.

1.3.4 DataSHIELD Data Privacy and Security

DataSHIELD's platform is designed to prioritize data privacy and security while allowing researchers to perform complex analyses on sensitive datasets. By executing analyses on the server-side, DataSHIELD ensures that individual-level data remains protected and inaccessible to researchers or other parties involved in the research process. This approach minimizes the risk of data leakage, unauthorized access, and inadvertent disclosure of sensitive information.

DataSHIELD employs multiple security mechanisms to protect the sensitive data it processes and ensure the privacy and confidentiality of the individuals whose data are being analyzed. These mechanisms include:

1. Federated Analysis: DataSHIELD's federated analysis approach allows multiple data sources to be analyzed simultaneously without the need to pool raw individual-level data. By sharing only aggregated results, summary statistics, or coefficients, DataSHIELD ensures sensitive information is not disclosed or accessed during the research process.
2. Server-side Processing: All data analyses are executed on the server-side, meaning that individual-level data never leaves the data custodian's server. This prevents unauthorized access or data leakage, as researchers and collaborators only have access to the aggregated results and not the raw data.

DataSHIELD works in conjunction with data warehouses like Opal and Armadillo, which provide essential security features that protect sensitive data throughout the analysis process. These features include:

1. Access Control: Opal and Armadillo incorporate strict access control mechanisms to ensure that only authorized users can access the platform and perform analyses. This may include authentication through usernames and passwords or more advanced methods such as two-factor authentication, depending on the specific implementation.
2. Encrypted Communication: Opal and Armadillo use secure communication protocols, such as HTTPS and SSL/TLS, to encrypt the data transmitted between the client and server during the analysis process. This ensures the confidentiality and integrity of the data while in transit.
3. Audit Logging: Opal and Armadillo maintain audit logs that record all activities performed within the platform, including data access and analysis operations. These logs can be used to monitor user activity, detect potential security breaches, and ensure compliance with relevant data governance policies.

Data privacy and security details are particularly relevant for public health, genetics, and social sciences fields because of the sensitive nature of the data involved. In these fields, data often include personal, medical, genetic, and behavioral information, which, if mishandled or disclosed, could lead to significant consequences for the individuals concerned. These consequences may range from privacy breaches and stigmatization to discrimination and potential misuse of information by unauthorized parties.

Ensuring data privacy and security is crucial for maintaining trust in the research process, ensuring compliance with ethical and legal requirements, and promoting the responsible use of sensitive data for scientific advancement. By effectively protecting individual-level data, researchers in public health, genetics, and social sciences can facilitate collaborations, share data across institutions and countries, and ultimately drive meaningful insights and discoveries to improve human well-being.

1.3.5 Technical Overview of DataSHIELD

So far, we have provided an overview of DataSHIELD, including its objectives, ethical and legal considerations, as well as the data privacy and security measures it incorporates. We have discussed its significance in research fields like public health, genetics, and social sciences, and how it addresses the challenges of working with sensitive data. In this section, we will delve deeper into the technical aspects of DataSHIELD, exploring its components, structure, and functionalities to better understand how it operates and achieves its goals.

We will delve into the technical aspects of DataSHIELD by examining its underlying structure. At its core, DataSHIELD is a collection of open-source R packages, each designed to perform specific tasks within the framework. By utilizing these packages in combination, DataSHIELD delivers a powerful and versatile solution for the analysis of sensitive data. As an open-source project, DataSHIELD encourages collaboration and contributions from its growing community, ensuring constant improvement and adaptation to the evolving needs of researchers and organizations alike.

R has been chosen as the language to build DataSHIELD for several compelling reasons. Firstly, researchers in health sciences fields typically use R for their analyses, making it a familiar and accessible choice. This allows DataSHIELD to be easily adopted and integrated into existing research workflows.

In addition to this, R is a widely recognized and powerful statistical programming language, offering a vast array of packages and libraries for data analysis, visualization, and manipulation. This enables DataSHIELD to leverage the existing tools and resources within the R ecosystem, providing users with a comprehensive and versatile platform for their research needs.

Moreover, R is an open-source language with a strong community-driven approach, which aligns with DataSHIELD's commitment to openness and collaboration. By using R, DataSHIELD can benefit from the continuous growth and improvement of the language, as well as contribute to the broader R community in return.

The R Bioconductor project has also played a significant role in the development of DataSHIELD. Bioconductor is a collection of packages specifically designed for the analysis of omics data, such as genomics, transcriptomics, and proteomics. These packages offer a wide range of tools and resources tailored to the unique challenges and demands of omics data analysis. The availability of Bioconductor packages has greatly facilitated the development of this thesis, as many of their functions have served as a foundation for the different analysis package developed on this thesis.

DataSHIELD is primarily built on a foundation of three interconnected R packages, namely dsBase, dsBaseClient, and DSI. These packages, which work in tandem, are designed to facilitate the analysis process and ensure seamless communication between the client and server sides. Each package plays a distinct role in the DataSHIELD ecosystem:

dsBase: The dsBase package serves as a fundamental building block within the DataSHIELD framework. It functions as an analysis package, offering capabilities similar to those found in the R base package. Crucial to the software architecture, dsBase is designed to run on the analysis server, ensuring that all data processing and computations take place within a secure environment. This package forms the foundation for DataSHIELD's ability to perform complex analyses while safeguarding sensitive information.

dsBaseClient: The dsBaseClient package acts as a complementary component to the dsBase package within the DataSHIELD framework. Designed to run on the client-side, specifically on the researcher's computer, this package facilitates the communication between the client and the server. Its primary function is to generate the appropriate function calls for the server-side package (dsBase) to execute.

DSI: The DataSHIELD Interface (DSI) package plays a crucial role in the DataSHIELD architecture by managing the communication between the client and the server. It is responsible for transmitting the analysis function calls generated by dsBaseClient to the server, ensuring that the correct instructions are relayed for the server-side package to execute. By doing so, DSI facilitates seamless and secure interactions between the different components within the DataSHIELD framework, further enhancing the privacy and protection of sensitive data during the research process.

1.3.6 Data warehouses: Opal and Armadillo

In the previous section, we provided a technical overview of DataSHIELD, discussing the various R packages that form the foundation of this innovative platform. As we now shift our focus to the data warehouses, Opal and Armadillo, we will explore the technology that enables the deployment of DataSHIELD in real-world applications. These data warehouses play a crucial role in ensuring secure and efficient data management, minimizing the risk of unauthorized access or data breaches. In the following paragraphs, we will explain the unique characteristics and functionalities of Opal and Armadillo, and examine how they contribute to the overall effectiveness of the DataSHIELD infrastructure.

Opal and Armadillo represent two separate data warehouse solutions that function as servers within the DataSHIELD ecosystem. Both have been developed to securely store information and house an R server capable of running DataSHIELD analysis packages like dsBase. In situations involving multiple cohorts, every participating entity, such as a hospital, retains its own on-premises data warehouse, ensuring data remains securely stored at the site without being transferred elsewhere. Opal and Armadillo handle user identification and permission allocation for those seeking to employ the DataSHIELD infrastructure, further reinforcing data security and privacy measures.

1.3.6.1 Opal

Opal, developed by OBiBa, is a comprehensive core data management application that plays a vital role in the DataSHIELD infrastructure. This server application delivers a wide range of tools for importing, transforming, and describing data, as well as managing subject identifiers during data import and export processes. Integrated with R, Opal allows complex statistical analysis and report generation, while also ensuring seamless and secure data import and management through its integration with Onyx and Mica. As a critical component of DataSHIELD, Opal offers robust data management capabilities, enabling efficient and secure data management in real-world applications.

Opal offers a variety of main features and advantages, ensuring efficient and secure data management:

1. **Data Warehouse:** Opal supports various database software backends such as MongoDB, MySQL, MariaDB, and PostgreSQL, and enables the storage of an unlimited number of variables.
2. **Customized variable dictionaries:** Users can create personalized dictionaries for managing and organizing their data.
3. **Multiple data import and export formats:** Opal supports data import from CSV, SPSS, SAS, Stata files, and SQL databases, as well as data export to these formats.
4. **Incremental data importation:** Opal allows for incremental data import, making it easy to update existing data with new information.
5. **Direct connection to multiple data sources:** Opal can directly connect to various data source software, such as SQL databases and LimeSurvey.
6. **Storage of diverse data types:** Users can store data about any type of entity and data of any type, including texts, numbers, geo-localization, images, and videos.
7. **Genotype data storage:** Opal supports the import and storage of genotype data as VCF files (Variant Call format).
8. **Advanced indexing:** Opal uses Elasticsearch for advanced indexing functionality, providing fast and efficient data search capabilities.

9. SQL API: Opal offers an SQL API for selecting, filtering, grouping, and joining table data

1.3.6.2 Armadillo

Armadillo, developed by Molgenis, is a powerful data portal designed for data stewards to share datasets on a server, enabling researchers to perform secure and efficient analysis using DataSHIELD tools. The platform streamlines data sharing and management while maintaining a strong focus on data security, confidentiality, and ease of use.

Armadillo main features can be summarized as:

1. DataSHIELD Integration: Armadillo leverages the DataSHIELD platform, enabling researchers to perform secure, privacy-preserving analysis on shared datasets.
2. Secure Data Sharing: Data stewards can securely share datasets on the Armadillo server, allowing researchers to analyze data while maintaining confidentiality and privacy.
3. Parquet Data Format: Data is stored in the efficient parquet format, which supports fast column selections for analysis.
4. Web User Interface and R Client: Data stewards can manage their data on the Armadillo file server using the web user interface or MolgenisArmadillo R client.
5. Encrypted Data Storage: Armadillo can store data encrypted on its file server for enhanced security.
6. Single Sign-On Authentication: The platform supports OpenID Connect (OIDC) based single sign-on, allowing seamless and secure authentication.
7. Flexible Deployment Options: Armadillo can be installed on various Linux or Unix-based operating systems, and it is also available as a Docker image for quick deployment and testing.

The decision to use one technology or another must be made by each center based on their specific needs and requirements. Furthermore, it is worth mentioning that within a collaborative project, it is entirely possible for different centers to employ distinct data warehouse technologies without hindering the overall goals of the project. Ultimately, the choice of data management and analysis tools should be determined by factors such as data security and compatibility with the center's existing infrastructure and expertise.

1.3.7 Analysis stack. Packages and infrastructure

In the following section, we will bring together all the elements we have discussed so far, illustrating how the various data management and analysis tools, such as Opal, Armadillo, and DataSHIELD, work together. By integrating these tools, we can facilitate the process of managing, sharing, and analyzing data across multiple centers while maintaining data security and privacy. To provide a comprehensive understanding of how these components interact, we will present a simple diagram that demonstrates their synergy and highlights the benefits of employing such a coordinated approach in a real-world scenario. This visual representation will help in grasping the seamless collaboration and interplay between these packages and data warehouse technologies, ultimately showcasing the potential of this integrated framework.

On fig. 2 we can take a look at a diagram showcasing the structure that a two cohort real-world DataSHIELD project would use. As we have previously seen, there are three main R packages within the infrastructure: dsBase, dsBaseClient and DSI. The researcher that is going to use DataSHIELD to perform analysis is located at the center of the image, depicted as "analyst". This analyst is using the dsBaseClient package in RStudio, although any R terminal or software with capabilities to run R code could also be used. The dsBaseClient is then contacting the different analysis servers through the DSI package, which is in charge of sending the DataSHIELD analysis commands. These commands are received by the different Opal servers and sent to the R servers they are running. Inside the R servers, there is the dsBase package which will run the analysis and return (again through Opal and DSI) non-disclosive statistics to the analyst.

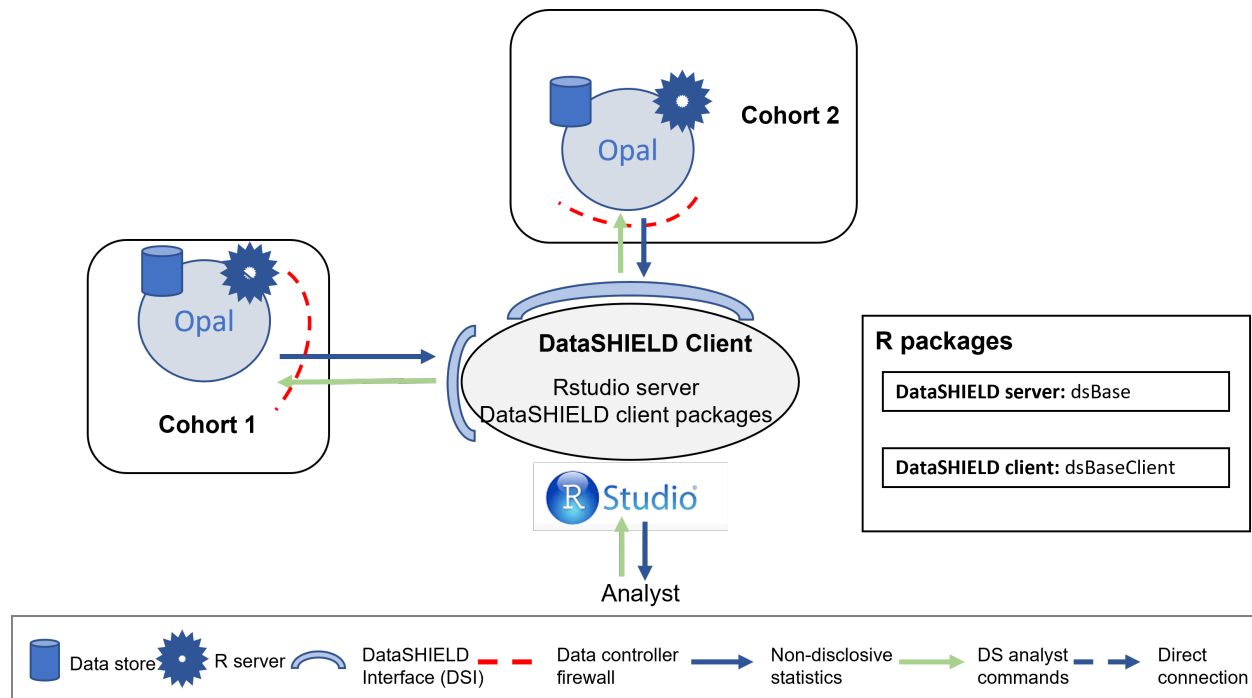


Figure 2: Diagram illustrating the integration of RStudio as a DataSHIELD client with two distinct cohorts using Opal servers. The analyst sends queries through dsBaseClient, and the Opal servers, running the dsBase server package, return non-disclosive statistics to the analyst.

1.3.8 DataSHIELD analysis types

In this section, we will explore the types of analysis that DataSHIELD is capable of performing, highlighting its versatility and adaptability to various research scenarios. DataSHIELD’s ability to work with horizontally partitioned data and conduct both meta-analyses and pooled studies demonstrates its potential to handle complex data structures and meet the needs of diverse research projects. We will develop each of these topics in greater detail, showcasing DataSHIELD’s analytical capabilities.

Horizontally partitioned data and vertically partitioned data are two different ways of organizing and distributing datasets across multiple sources or organizations. Understanding the differences between these two types of data partitioning is essential to appreciate how DataSHIELD operates and adapts to various research contexts.

In horizontally partitioned data, the dataset is divided into multiple parts based on the records or rows. Each partition contains a subset of the complete records, but retains all the variables or columns of the original dataset. This approach is commonly used when different organizations or research centers hold data on distinct individuals or groups, but share the same variables for each record.

On the other hand, vertically partitioned data involves splitting the dataset based on variables or columns. In this case, each partition contains all records but only a subset of the variables. This scenario is typically encountered when various organizations hold different sets of variables or measurements for the same individuals or groups.

This two types of methods of partitioning data are visualized on fig. 3.

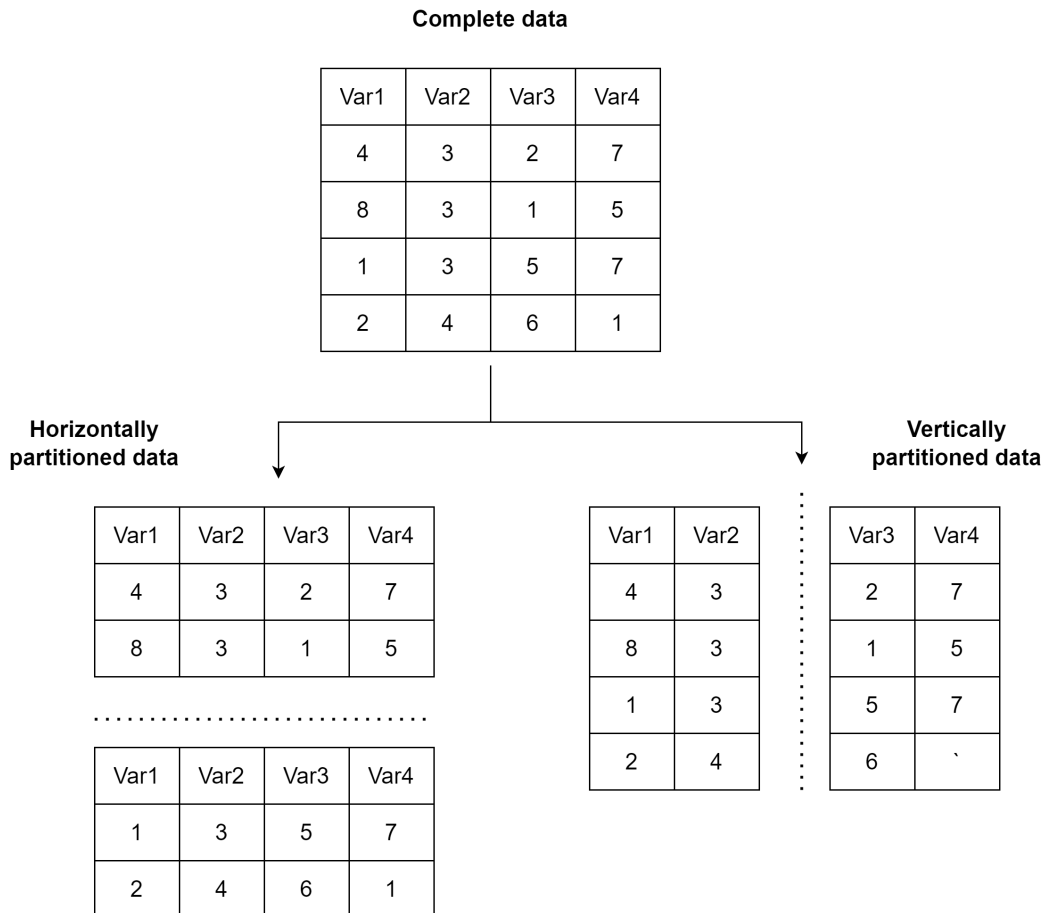


Figure 3: Horizontally and vertically partitioned data example.

DataSHIELD’s focus on horizontally partitioned data is a natural fit for collaborative research projects, as it aligns with the typical configuration of such initiatives [82]. In these projects, multiple centers collect the same type of data from different individuals to address specific research questions, such as health conditions or other relevant factors. By having data from diverse populations, these collaborations aim to achieve greater statistical power and yield more meaningful results. DataSHIELD’s ability to analyze horizontally partitioned data while preserving privacy and data security enables researchers to work collectively, unlocking the potential of shared resources and enhancing the overall quality and impact of their research findings.

DataSHIELD offers two main analysis methods for horizontally partitioned data: meta-analysis and pooled analysis. Both methods enable researchers to combine data from multiple sources while preserving data privacy, but they differ in their approach and capabilities.

Meta-analysis refers to the process of statistically combining results from multiple independent studies to arrive at a single, comprehensive conclusion. In a meta-analysis, summary statistics from each study are used rather than individual-level data. This approach is particularly useful when combining data from different sources or when it is not feasible to share individual-level data due to privacy concerns.

Pooled analysis, on the other hand, involves combining individual-level data from multiple sources into a single dataset for analysis. With DataSHIELD, pooled analysis is performed in a privacy-preserving manner by conducting the analyses on the individual data sources while only sharing aggregated results, the data never leaves the analysis servers. This approach allows researchers to perform more complex statistical analyses, as they have access to a larger, combined dataset without compromising data privacy.

Both meta-analysis and pooled analysis methods offered by DataSHIELD provide researchers with valuable tools to leverage the power of combined data from multiple sources, enabling more robust and meaningful research outcomes while ensuring data privacy and security.

1.3.8.1 Pooled analysis in DataSHIELD: The Generalized linear model (GLM) case

In this section, we will demonstrate an example of a pooled analysis in DataSHIELD, specifically focusing on the implementation of generalized linear models (GLMs) using a modified iterated reweighted least squares (IRLS) method [83, 84]. GLMs are a flexible and powerful technique used to model the relationship between a response variable and one or more explanatory variables. In the following paragraphs, we will describe the step-by-step process of how DataSHIELD performs this privacy-preserving GLM analysis using the modified IRLS method.

First, let's establish the theory behind GLMs. Suppose we have a study with N independent observations. There is a dependent Y variable and a set of covariates q for each observation. The covariates are placed inside a matrix $X^T = (x_1, \dots, x_N)$, where $x_i^T = (x_{i1}, \dots, x_{iq})$ representing the q dimensions for the observation i . We then assume that the relationship between Y and X for the observation i can be expressed by a GLM

$$\eta_i := g(\mu_i) = \beta^T x_i \quad (2)$$

Following the standard notation [84], η_i is the linear predictor, g the function specified by the researcher, μ_i the mean of Y_i having $\mu = (\mu_1, \dots, \mu_N)$ and $\beta^T = (\beta_1, \dots, \beta_q)$. β^T is the parameter vector we wish to estimate.

Following the definition of a GLM, Y must be drawn from a parameterized distribution that belongs to the exponential family (e.g. Gaussian, binomial, Poisson). The probability density f follows:

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - c(\theta_i)}{\phi} + h(h_i, \phi)\right) \quad (3)$$

From (referencia 17 link de dal) we have:

$$\mu_i := E[Y_i] = \frac{dc(\theta_i)}{d\theta_i}; \text{Var}[Y_i] = \phi \frac{d^2c(\theta_i)}{d\theta_i^2} := \phi V_i \quad (4)$$

For a logistic regression ($\mu = p$) we would have eq. (5) and eq. (6).

$$\eta_i = g(p_i) = \log \frac{p_i}{1-p_i}; c(\theta_i) = -\log(1-p_i); \phi = 1 \quad (5)$$

$$\frac{dg(\mu_i)}{d\mu_i} := g'(p_i) = \frac{1}{p_i(1-p_i)}; V_i = p_i(1-p_i) \quad (6)$$

So far, we have provided an overview of the generalized linear model (GLM). In the following lines, we will explain the Iteratively Reweighted Least Squares (IRLS) algorithm, which is designed to estimate the maximum likelihood for the parameter vector β . As an iterative algorithm, IRLS refines its estimates through a series of iterations to reach a convergent solution.

We define β_t as the vector β at the t^{th} iteration. IRLS derives β_{t+1} via eq. (7).

$$\beta_{t+1} = \beta_t + I(\beta_t)^{-1} s(\beta_t) \quad (7)$$

Where I is the expected information matrix and s the score function. We define I and s following eq. (8) and eq. (9).

$$I(\beta_t) = X^T W_t X \quad (8)$$

$$s(\beta_t) = X^T W_t (Y - \mu(t)) g'(\mu(t)) \quad (9)$$

Where W_t is defined following eq. (10) and eq. (11).

$$w_{ii}(t)^{-1} = V_i(t) g'(\mu_i(t))^2 \quad (10)$$

$$g'(\mu_i(t)) = dg(\mu_i(t))/d\mu_i(t) \quad (11)$$

We use $\mu(t)$ since μ depends on β therefore on t . β is updated until a convergence criteria is met.

Now we will see how we can use this method when the data is not pooled into a single large dataframe, but that it can be split in different study servers.

We can write I and s in terms of summations, where N is the number of study servers. We do that following eq. (12) and eq. (13).

$$I(\beta_t) = \sum_{i=1}^N w_{ii}(t) x_i x_i^T \quad (12)$$

$$s(\beta_t) = \sum_{i=1}^N (y_i - \mu_i(t)) g'(\mu_i(t)) w_{ii}(t) x_i \quad (13)$$

With all that information, we can begin the iterative computation process:

1. We estimate $g(\mu_i(t))$ with eq. (2). We know β_t at iteration $(t + 1)$ from the previous iteration.
2. Then, we can estimate $\mu_i(t)$ with the inverse link function $g^{-1}()$. E.g. for logistic regression: $\mu_i(t) := p_i(t) = \exp(\beta_t^T x_i) / (1 + \exp(\beta_t^T x_i))$ which is the expected probability of a positive response at iteration t .
3. We know the general form of function g and the value of $\mu_i(t)$, therefore we can calculate $g'(\mu_i(t))$. We have the general form $g'(\mu_i(t)) := g'(p_i(t))$. Following with the logistic regression example, we can use eq. (5) and eq. (6) giving $g'(p_i(t)) = 1/p_i(t)(1 - p_i(t))$.
4. We know the probability density function for Y , therefore function $c()$ is also known (eq. (3)). We can estimate $V_i(t) = d^2 c(\theta_i) / d\theta_i^2$ for iteration t (eq. (4)). For the logistic regression eq. (5) and eq. (6) yield $V_i(t) = p_i(t)(1 - p_i(t))$
5. Finally, $w_{ii}(t)$ is calculated using $g'(\mu_i(t))$ and $V_i(t)$ with eq. (10).

Now we are in position to compute $I(\beta_t)$ and $s(\beta_t)$ is direct and β_{t+1} can be derived with eq. (7) finishing the iteration process.

For those interested in a more in-depth exploration of the GLM method in DataSHIELD, including the underlying mathematics and practical examples, there is a published paper available that delves into these topics [85]. This resource provides a comprehensive understanding of the method's implementation in DataSHIELD.

In conclusion, the detailed mathematical demonstration provided in this section showcases the intricate and complex mathematics employed in DataSHIELD to perform pooled analysis using the IRLS algorithm for GLM implementation. This example highlights the platform's capability to handle sophisticated statistical techniques while maintaining privacy. However, it is essential to note that for such analyses to be effective,

data across all participating studies must share the same format and scale for both the outcome variable and covariates. Ensuring this uniformity is a fundamental prerequisite for any collaborative research endeavor and helps ensure the accuracy and validity of the results generated.

1.3.8.2 Beyond GLM: Algorithms that can be used to perform pooled analysis

In this section, we will discuss the process of developing or identifying algorithms that can be integrated within the DataSHIELD architecture as pooled methods. By understanding the principles and criteria for developing or identifying suitable algorithms, researchers and developers can contribute to the expansion of DataSHIELD’s capabilities, further promoting secure and collaborative research.

One of the promising approaches to discover new algorithms that could be adapted for use in DataSHIELD’s infrastructure as pooled methods is to explore the field of computer science, specifically focusing on the trend of parallelizing algorithms [86]. Over the past years, parallelization has gained significant attention as it allows for efficient utilization of multi-core CPUs and high-performance GPUs. By leveraging the principles and techniques employed in parallel algorithms, it is possible to identify potential candidates that can be adapted and implemented within DataSHIELD, expanding its capabilities for secure and collaborative analysis across distributed data sources.

When considering the adaptation of parallelized algorithms for use in DataSHIELD, it is essential to establish an equivalence between a computing thread and a study center. This analogy allows us to better understand how these algorithms can be applied in a distributed data setting. However, there are certain characteristics of the algorithms that need to be carefully examined and possibly modified to ensure their suitability for DataSHIELD’s specific requirements. These characteristics may include the communication patterns between threads (or study centers), the nature of the data being processed, and the overall structure of the algorithm. By thoroughly exploring these aspects, we can successfully adapt parallelized algorithms to be effectively employed in DataSHIELD.

When assessing the suitability of an algorithm for use with horizontally partitioned data in DataSHIELD, it is crucial to ensure that the algorithm consistently maintains the same data on each thread (or study center) throughout the computation process, this concept is called non-sharing memory distributed algorithms [87]. This characteristic is essential for preserving the privacy of individual-level data and ensuring that the analysis can be effectively carried out across multiple study centers without compromising data security.

Furthermore, if the algorithm requires any communication between the master (in DataSHIELD, this would be the client) and the threads (in DataSHIELD, these would be the study centers), we must also make certain that this communication only contains aggregated non-disclosive statistics. This additional precaution ensures that the privacy of sensitive data remains protected throughout the entire analysis process.

By carefully selecting algorithms with these features, we can ensure that DataSHIELD continues to provide a secure and robust platform for collaborative, privacy-preserving data analysis.

Once we have identified an algorithm that meets the requirements for handling horizontally partitioned data and ensuring privacy preservation through aggregated non-disclosive statistics, there is still some work to be done to guarantee its security within the DataSHIELD framework. The next step is to add the appropriate disclosure traps to ensure that the results shared with the client consistently do not contain information that would allow reidentification of any individual.

To achieve this, we need to carefully examine the DataSHIELD privacy-preserving mechanisms described in the relevant section and tailor the algorithm to incorporate these protective measures. Depending on the type of data involved, it may be necessary to employ an ad-hoc solution tailored to that specific data.

This step is the most sensitive and critical aspect of adapting an algorithm for use in DataSHIELD. It is always a good practice to consult experts with diverse backgrounds, such as mathematicians, statisticians, biostatisticians, and other professionals, to ensure that the algorithm is secure, effective, and compliant with privacy-preserving principles. Collaboration and thorough review are essential for maintaining the high standards of privacy and security that DataSHIELD strives to achieve.

We will conclude with a simple example of a pooled algorithm by demonstrating the case for the pooled mean. When computing the mean of a variable distributed across different study centers, each center calculates the mean of the variable and reports that value along with the number of observations to the client. The client then aggregates this data using eq. (14) to obtain the combined mean of that variable. On that equation, N is the number of study centers, ρ_i is the mean of the i study center and n_i is the number of observations of the i study center.

$$\frac{\sum_{i=1}^N \rho_i n_i}{\sum_{i=1}^N n_i} \quad (14)$$

To ensure that the information provided by the study centers is non-disclosive, a disclosure trap is placed on each study server. This trap ensures that there is a minimum number of individuals required to compute the mean. If the minimum number of individuals is not met, the server will not return any information to the client. This safeguard helps maintain the privacy and security of the data while still allowing the computation of a pooled mean across multiple study centers.

1.3.9 DataSHIELD privacy preserving mechanisms

To ensure that DataSHIELD maintains a high level of privacy and security, the framework incorporates a multi-layered approach, including system protection elements, analysis protection elements, and governance protection elements. Additionally, the platform is continuously evolving to address potential weaknesses, incorporating new protection elements and future additions as the field of privacy-preserving data analysis advances. The current chapter aims to provide an in-depth understanding of these mechanisms and their implementation in DataSHIELD.

1.3.9.1 System protection elements

DataSHIELD’s privacy-preserving mechanisms encompass various system protection elements that ensure the security and integrity of sensitive data. One of the fundamental aspects of these protection measures is the implementation of robust physical security. This involves secured data centers with access controls and surveillance systems that prevent unauthorized access to the data processing servers and hardware components, mitigating potential breaches and data theft.

Network security is another essential component of DataSHIELD’s system protection. To safeguard the transmission of data between client and server computers, DataSHIELD utilizes encryption techniques, ensuring that the exchanged information remains secure and unintelligible to potential attackers. Moreover, network firewalls are employed to block unauthorized traffic and prevent data from being sent to unapproved locations, further enhancing the platform’s security.

In addition to physical and network security, DataSHIELD employs an R parser to validate client-server commands and their arguments. This process allows only permitted methods to pass from the client to the servers, ensuring that only valid methods are used for data analysis. While the R parser enforces global checks on the commands, it is essential to note its limitations, as it lacks knowledge of the semantics of individual methods. To address this issue, these global checks are augmented by checks within the individual methods themselves.

Furthermore, server-side R in DataSHIELD is only callable through a middleware, such as Opal or Armadillo. This middleware serves as an additional layer of protection, responsible for user authentication and ensuring that only authorized users have access to the data and analytical functions.

Lastly, DataSHIELD maintains logs of all analysis commands executed on the servers. This logging mechanism allows for posterior revision of user activity, providing a means to track and identify potential malicious actors on the platform. By continuously monitoring user activity, DataSHIELD ensures that any

unauthorized or suspicious actions can be promptly investigated and addressed, further strengthening its privacy-preserving mechanisms.

The system structure depicted in this section is illustrated on fig. 4, where there are the two distinct parts differentiated, the client and the server. Communication between the parts is encrypted. Inside the server the communication flow goes from the middleware ("web service"), then passes through the R parser and after that it goes to the R server where data will be analyzed.

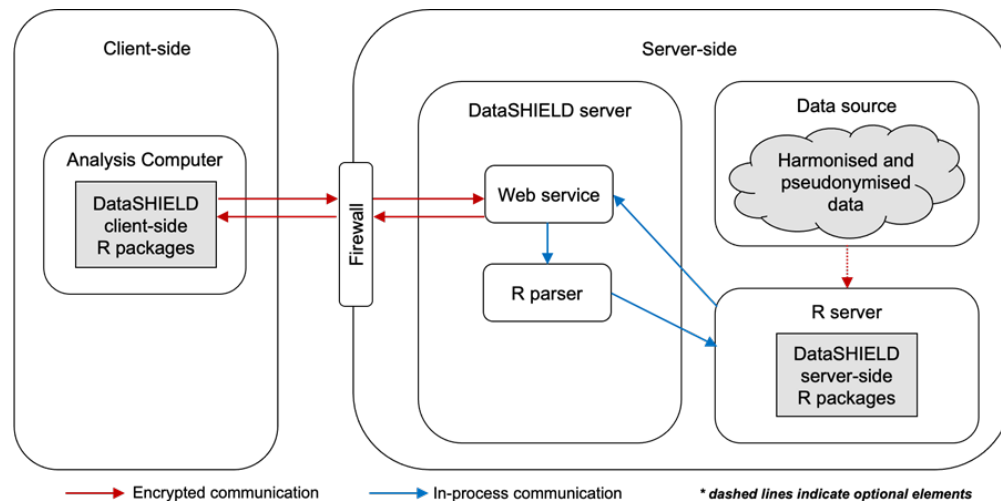


Figure 4: System protection elements flowchart. Figure extracted from 'Mitigating disclosure risk in multi-site federated analysis: the paradigm of DataSHIELD' by Avraam et. al. with authorization from the authors.

1.3.9.2 Analysis protection elements

DataSHIELD's privacy-preserving mechanisms also encompass a variety of analysis protection elements that minimize the risk of data disclosure during statistical analyses. To achieve this, the interactions between the client and server are restricted to "assign" and "aggregate" methods, ensuring that sensitive data remains secure and inaccessible to unauthorized users.

Assign functions play a crucial role in DataSHIELD's analysis protection. These functions generate and save objects on the server-side without returning any information to the client-side, except for study-specific messages indicating the successful creation and expected format of the objects across all studies. Conversely, aggregate functions generate low-dimensional statistical results, which are then returned to the client-side. The design of these aggregate functions inherently limits the potential for disclosive outputs.

While assign functions do not require disclosure traps per se, it is essential to consider whether they could be exploited in a manner that would allow a user to bypass a disclosure trap. To prevent this, assign functions are designed to only interact with server-side objects through aggregate functions, further enhancing DataSHIELD's privacy measures.

Several features have been implemented to limit disclosive outputs in functions. These include removing potentially disclosive outputs (e.g., residuals and predicted values from generalized regression models), sanitizing and validating inputs to ensure proper function behavior, and confirming that error messages do not inadvertently reveal sensitive data.

Disclosure traps play a pivotal role in DataSHIELD's analysis protection, allowing only non-disclosive summary statistics to leave the server. The server-side functions of the dsBase DataSHIELD package version 6.3.0 employ a set of disclosure traps, as listed in table 1. Furthermore, server administrators can add additional disclosure traps when installing new releases of dsBase or other DataSHIELD packages (e.g., dsOmics).

Similar to before, on fig. 5 there is an illustration on how the disclosure traps are integrated on the data flowchart. All the disclosure traps are applied at the R server level given that the function call has been allowed by the R parser, there the outputs are also checked before leaving the server in the form of non-disclosive results, which are passed back to the user.

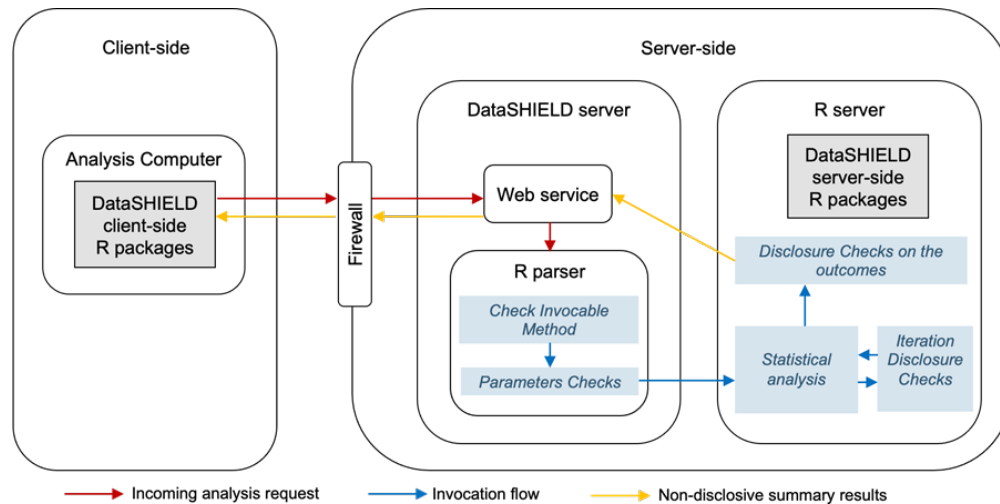


Figure 5: Analysis protection elements flowchart. Figure extracted from 'Mitigating disclosure risk in multi-site federated analysis: the paradigm of DataSHIELD' by Avraam et. al. with authorization from the authors.

The thresholds for these disclosure traps are controlled exclusively by the data custodian, allowing them to maintain full control over the level of protection for their specific data situation. While analysts can view these thresholds, they cannot modify them, ensuring that data custodians can effectively safeguard sensitive information.

1.3.9.3 Governance protection elements

Governance protection elements are crucial in ensuring that DataSHIELD's privacy-preserving mechanisms are robust and effective. These elements establish and enforce appropriate access controls and usage policies to maintain the confidentiality and integrity of sensitive data.

Before researchers can analyze data using DataSHIELD, they must be authorized by the data custodian of the study. This authorization process helps to ensure that only qualified individuals are granted access to sensitive data. Once approved, researchers receive user-specific credentials to log in to data servers, with some organizations even implementing two-factor authentication (2FA) for added security. Certain consortia also require researchers to use a central analysis server, which necessitates an additional set of user-specific credentials.

Data access procedures determine the specific subset of variables that researchers can access, further limiting the potential for unauthorized use or disclosure of sensitive information. To safeguard against potential privacy breaches, data stored on servers is pseudonymized, unlinking it from any personally identifiable information.

In addition to these access controls, DataSHIELD's governance framework also requires researchers to obtain approval from each data owner's ethics committee before publishing any findings. This ensures that the results presented in research papers adhere to the ethical guidelines and data privacy requirements set forth by the data custodians.

Moreover, researchers can be assigned different analysis profiles, each with distinct disclosure trap values. These profiles allow data custodians to customize the level of access and functionality granted to individual

Table 1: DataSHIELD Disclosure traps in dsBase (version 6.3.0) functions

Name	Description
<i>nfilter.tab</i>	The lowest number of non-empty entries that must be present in a table for it to be shared is referred to as the minimum non-zero cell count. This rule applies to tables of different dimensions, including one-, two-, and three-dimensional tables. These tables can be based on counts across one, two, or three factors or can represent the average of a quantitative variable across a factor. The default value for the minimum non-zero cell count is set to 3.
<i>nfilter.subset</i>	The smallest number of non-empty data points (usually representing individuals) that must be present in a specific subset is called the minimum non-zero count. The default value for this requirement is set to 3.
<i>nfilter.glm</i>	The maximum number of factors in a regression model is limited by a certain proportion of the sample size in a study. For example, if a study has 1,000 data points (usually individuals) for a specific analysis and the limit is set to 0.33 (default value), then the model can include up to 330 factors. This restriction helps prevent overly complex models that could potentially disclose sensitive information.
<i>nfilter.string, nfilter.stringShort</i>	The maximum length allowed for a text argument is restricted to ensure its length is tested. Default values for the long and short text limits are 80 and 20 characters, respectively. These restrictions help prevent hackers from embedding harmful code within a valid text argument that could be actively interpreted.
<i>nfilter.levels.density</i>	The highest acceptable ratio of unique categories in a categorical variable to the total number of entries, considered non-disclosive. For instance, if there are 1,000 unique categories from 4,000 entries, this would result in a 0.25 (25%) proportion, which is seen as non-disclosive. The default value is set at 0.33.
<i>nfilter.levels.max</i>	The highest acceptable number of distinct categories in a categorical variable considered non-disclosive. The default value is set at 40.
<i>nfilter.kNN</i>	The smallest allowed value for k in the k-nearest neighbors method, primarily used for certain graphical functions. The default value is set at 3.
<i>nfilter.noise</i>	The smallest amount of noise that can be added to a server-side vector. This value indicates the variance of the added noise. For example, if the minimum noise level is set to 0.25 (the default value), then noise with a zero mean and a variance equal to 25% of the true variance of the vector is added to each individual value in the vector. This "noisy" vector can then be sent back to the client.
<i>datashield.privacyControlLevel</i>	Allow server administrators to operate servers with a limited selection of standard methods. If this option's value is not "permissive", the following server-side methods will be unavailable: dataFrameSubsetDS1, levelsDS, BooleDS, cDS, cbindDS, dataFrameDS, dataFrameSortDS, dataFrameSubsetDS2, dmtC2SDS, rbindDS, recodeLevelsDS, recodeValuesDS, repDS, reShapeDS, seqDS, subsetByClassDS, and subsetDS. The default value is "permissive".

users, further enhancing DataSHIELD’s privacy-preserving mechanisms.

The governance protection is illustrated on fig. 6. On the client side, there are the two different options, one being an analysis computer owned by the researcher and the other one is a central analysis server, the later requires a proprietary authentication protocol to be accessed. Afterwards, both clients connect to the same server, where depending on the specific user, different data and R functions will be available.

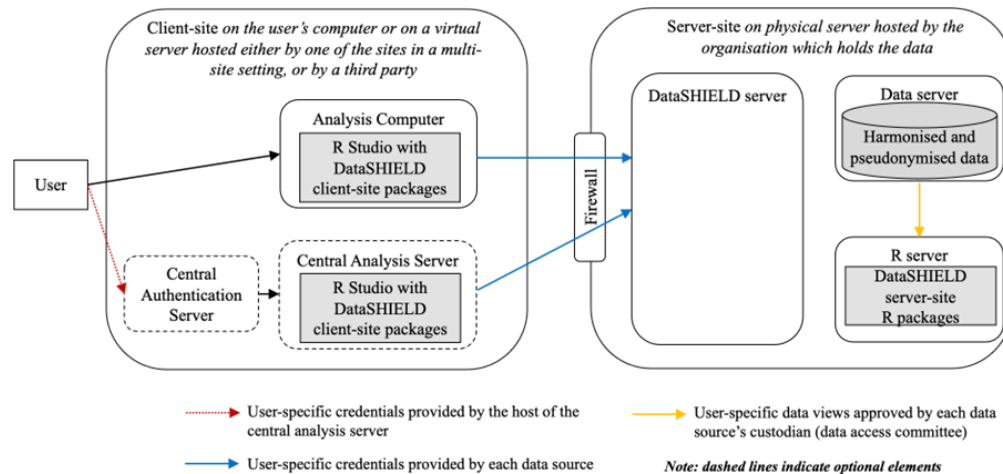


Figure 6: Governance protection elements flowchart. Figure extracted from ‘Mitigating disclosure risk in multi-site federated analysis: the paradigm of DataSHIELD’ by Avraam et. al. with authorization from the authors.

By implementing these governance protection elements, DataSHIELD ensures the responsible use of sensitive data, while maintaining the highest standards of data privacy and security.

1.3.9.4 Future developments

The development of DataSHIELD’s privacy-preserving mechanisms is an ongoing process, with researchers continually exploring new ways to enhance its security and functionality. One promising avenue of development involves leveraging artificial intelligence (AI) to identify potential security risks and attacks. AI can be utilized to monitor analysis commands in real-time, tracking differences and detecting attempts to infer sensitive information. By automating this process, DataSHIELD can further strengthen its protection against inferential disclosure.

The introduction of a Trusted Third Party (TTP) or an intermediate layer offers another potential enhancement to DataSHIELD’s privacy-preserving capabilities. For example, a group of hospitals may be required to publish collective reports on their activities without disclosing individual reports to the public. In such cases, a TTP could be given access to the individual reports, collating and publishing the results in an agreed-upon manner while ensuring data privacy is maintained.

Another area of development involves the use of multiple server domains, which can segregate different types of data, such as injected and processed data, into distinct environments. This approach allows for greater control over data access and reduces the risk of unauthorized disclosure.

Finally, the ability to create scripts that can only be executed if all cohorts agree presents a valuable addition to DataSHIELD’s privacy-preserving mechanisms. This feature enables researchers to collaborate on analyses while adhering to the stringent data privacy requirements of each cohort involved.

1.3.10 DataSHIELD future privacy preserving mechanisms

1.3.11 Projects using DataSHIELD

In this section, we will list various ongoing and pilot projects that utilize DataSHIELD for their research purposes. These projects highlight the growing adoption of DataSHIELD as a powerful and secure data analysis tool within the scientific community. The increasing number of projects and institutions involved in implementing DataSHIELD serves as a strong indicator of the project's success and its potential for further expansion. By looking into these projects, we can better understand how DataSHIELD has been effectively integrated into diverse research initiatives and appreciate the impact it has had on facilitating secure, collaborative research across multiple institutions.

The current projects using DataSHIELD are (extracted from datashield.org on 28/03/2023):

- ATHLETE [88]: Develop advance tools for Human Early Lifecourse Exposome Research and establish a prospective exposome cohort, including a FAIR data infrastructure, by building on Europe's most comprehensive exposome cohorts covering the first 18 years of life.
- BioSHaRE-EU [89] Environmental Core Project for the federated analysis of data from 6 European studies including UK Biobank; Obese Project for the federated analysis of 10 European studies including data from the National Child Development Study.
- ENPADASI [90] (German Institute of Human Nutrition, Max Delbrück Center for Molecular Medicine in the Helmholtz Association): The European Nutritional Phenotype Assessment and Data Sharing Initiative aimed at delivering an open access research infrastructure containing data from a wide variety of nutritional studies, ranging from mechanistic/interventions to epidemiological studies including a multitude of phenotypic outcomes, facilitating combined analyses.
- EUCAN-CONNECT [91]: developing a federated FAIR platform enabling large-scale analysis of high-value cohort data connecting Europe and Canada in personalized health. Collaborating with 173 European population-based cohort studies with 2.5M participants in total. This project aims to coordinate DataSHIELD implementations across LifeCycle, RECAP, InterConnect, Reach, LongITools and Athlete (and future projects that emerge).
- InterConnect [92] (MRC Epidemiology Unit, Cambridge): InterConnect is developing a global collaborative network for diabetes and obesity research, piloting DataSHIELD to facilitate a new approach to data sharing that is secure, scalable and sustainable. This includes data from 43 studies.
- INTIMIC [93] (Max Delbrück Center for Molecular Medicine in the Helmholtz Association): The Intestinal Microbiomics Knowledge Platform (INTIMIC) has the main objective of fostering studies on the microbiota, nutrition and health by assembling available knowledge of the microbiota and the other aspects (e.g. food science and metabolomics) that are relevant in the context of microbiome research in a FAIRyfyed (findable, accessible, interoperable and reusable) fashion to the scientific community, and to share information with the various stakeholders.
- LifeCycle [94]: developing new strategies for optimizing early life that will help to maximize the human developmental potential for current and future European generations. Includes 40 European cohort studies.
- MIRACUM [95] (Medical Informatic in Research and Care in University Medicine): A national German network of 10 University Hospitals to improve healthcare and strengthen Biomedical Informatics in Research and Education.
- RECAP preterm [96]: Research to improve to health, development and quality of life of babies born preterm. Includes 20 population-based cohort studies from Europe.

The projects setting up DataSHIELD pilots are (extracted from datashield.org on 28/03/2023):

- International 100,000+ cohorts consortium [97] : Large cohort studies involving hundreds of thousands of participants have been established or launched in several regions worldwide. Cohorts provide great value for studying diverse populations and key demographic subgroups, rare genotypes and exposures,

and gene-environment interactions. Each cohort is constrained, however, by its size, ancestral origins, and geographical boundaries, which limit the subgroups, exposures, outcomes, and interactions it can examine. Linking data across large cohorts provides a vast digital resource of diverse data to address questions that none of these cohorts can answer alone, enhancing the value of each cohort and leveraging the enormous investments made in them to date.

- LITMUS [98]: Liver Investigation:

Aiming to use DataSHIELD to provide non-disclosive/controlled access to the LITMUS Project’s genetic information, using the dsOmics module developed in collaboration with EUCAN-Connect and ATHLETE. Researchers requiring access to the LITMUS genetic information are able to perform remote analysis operations without needing to directly access highly confidential data, speeding advances in diagnosis and treatment of liver disease.

Testing Marker Utility in Steatohepatitis (LITMUS) funded by the European Innovative Medicines Initiative 2 Joint Undertaking, brings together clinicians and scientists from prominent academic centres across Europe with companies from the European Federation of Pharmaceutical Industries and Associations (EFPIA). Their common goals are developing, validating and qualifying better biomarkers for testing NAFLD.

- LONGITOOLS [99]: a European research project studying the interactions between the environment, lifestyle and health in determining the risks of chronic cardiovascular and metabolic diseases; LongI-Tools is bringing together 25 European cohorts and studies.
- NFDI4Health [100]: the National Research Data Infrastructure for Personal Health Data aims at enabling findability, accessibility, interoperability, and reusability of data generated in clinical trials, epidemiological, and public health studies in Germany to enhance collaboration among research communities while complying with privacy regulations and ethical requirements.

In conclusion, the wide array of projects that currently employ DataSHIELD or are in the process of setting up pilots demonstrates the versatility and effectiveness of this technology in addressing various research challenges. As DataSHIELD continues to gain traction, it is positioned to become an increasingly valuable tool for secure, collaborative data analysis in numerous fields of study.

1.4 Overview of Exposome Data Analysis

After a comprehensive exploration of data sharing, the potential of federated data analysis, and the capabilities and gaps within DataSHIELD, it becomes evident that the manner in which we process and analyze data can be as pivotal as the data itself. Particularly, when we consider complex and multifaceted data types, such as exposome data, the need for sophisticated analytical approaches becomes even more pressing.

1.4.1 The exposome concept

The "exposome" concept, first introduced by Wild in 2005 [101], encompasses the totality of an individual’s environmental exposures and lifestyle factors throughout their lifetime while taking into account the dynamic nature of these exposures as they evolve over time. In 2012, Wild further refined the concept by proposing a three-domain classification system that distinguishes between internal, general external, and specific external factors [102], this distinction is illustrated on fig. 7.

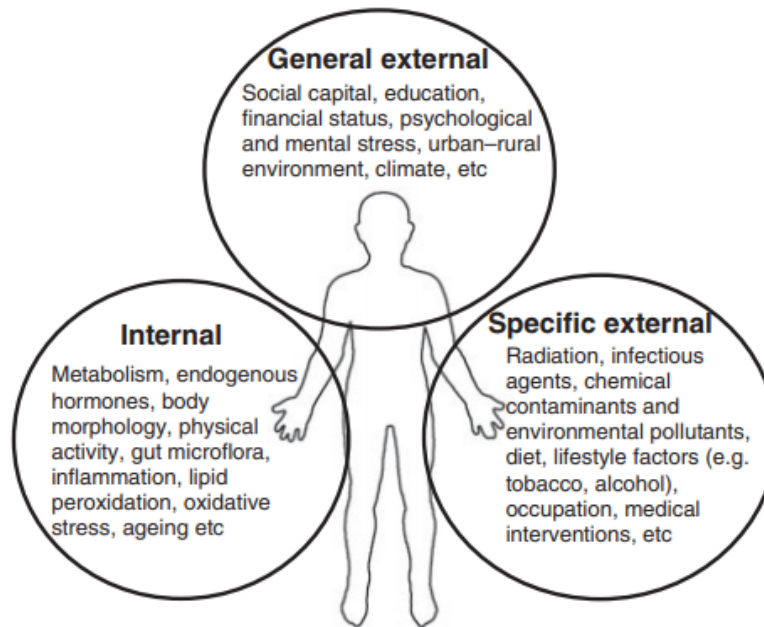


Figure 7: Three different domains of the exposome. Extracted from Wild [102]

The internal domain relates to the biological and toxicological processes that respond to both general and specific external exposures. General external exposures refer to the broader range of factors an individual may encounter, such as mental stress, climate, and living environment. Specific external exposures, on the other hand, cover a wide array of exposures the individual may be subjected to, including lifestyle factors, medical interventions, and environmental pollutants.

Building on the initial understanding of the exposome concept, researchers have since emphasized the importance of integrating multi-omics approaches to comprehensively assess the complex interactions between environmental exposures and genetic factors in relation to human health [103]. These multi-omics integration can provide valuable insights into the molecular mechanisms underlying the exposome and its potential impact on disease development and progression.

A key challenge in exposome research lies in the accurate measurement and assessment of environmental exposures, which are often diverse, dynamic, and multi-dimensional. Advanced technologies such as wearable sensors, personal exposure monitoring devices, and smartphone applications have been employed to collect real-time data on individual exposures, improving the granularity and precision of exposome measurements [104].

Furthermore, the exposome concept emphasizes the importance of considering the temporal dimension of environmental exposures, as both the timing and duration of exposures can significantly influence health outcomes. For instance, critical periods of development, such as prenatal and early postnatal life, may be particularly sensitive to environmental influences, with potential long-lasting effects on an individual's health trajectory [105].

Since its inception in 2005, the exposome concept has undergone significant evolution, driven by advancements in technology and a growing recognition of the importance of considering environmental exposures in the context of human health. As a result, the exposome has emerged as a highly relevant and promising area of research, spanning various fields including environmental health, molecular epidemiology, and precision medicine, this is reflected on the rise of academic papers referring to it, which can be seen on fig. 8.

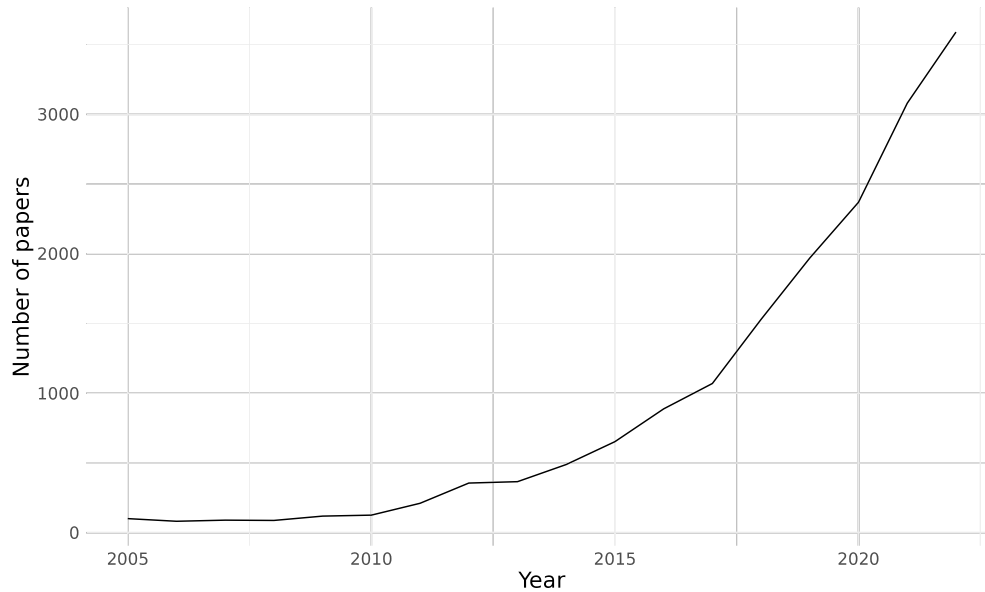


Figure 8: Evolution of number of papers related to exposome during last years. Extracted using Strobel [106]

The exposome concept has also gained increased relevance in the context of precision medicine, an approach that aims to tailor medical treatments to individual patients based on their genetic makeup, environmental exposures, and lifestyle factors. By elucidating the complex interplay between genes and environment, exposome research has the potential to identify novel biomarkers and molecular targets for personalized therapies, as well as inform the development of targeted prevention strategies to reduce disease risk in susceptible individuals [107].

Moreover, the exposome has informed the design of large-scale cohort studies, in which longitudinal data on environmental exposures and health outcomes are collected from thousands of individuals over extended periods. These studies, such as the European Human Early-Life Exposome (HELIX) project and the United States National Institutes of Health’s Environmental Influences on Child Health Outcomes (ECHO) program, are helping to advance our understanding of the long-term health impacts of early-life exposures and identify potential windows of susceptibility for intervention [103, 108].

In summary, the exposome concept has evolved significantly over the years, playing an increasingly important role in shaping research across multiple disciplines. Its emphasis on the comprehensive assessment of environmental exposures, in conjunction with genetic and lifestyle factors, holds great promise for advancing our understanding of human health and disease, as well as informing the development of targeted interventions and prevention strategies.

1.4.2 Data collection

Data collection is a crucial aspect of exposome research, as it forms the foundation for understanding the intricate relationships between environmental exposures, lifestyle factors, and human health. Accurate and comprehensive data on individual exposures is essential for identifying potential links to disease outcomes and uncovering novel insights into disease mechanisms. In this section, we will discuss the various tools and methods employed for capturing exposure data, including static data collection, personal exposure monitoring devices and smartphone applications. Furthermore, we will dig into the use of questionnaires and interviews for gathering information on lifestyle factors, as well as the collection of biological samples, such as blood, urine, and saliva, for analyzing internal exposures and biomarkers of effect. By examining these data collection techniques, we aim to provide a comprehensive overview of the approaches used in exposome research and highlight their significance in advancing our understanding of the complex interplay

between environment, lifestyle, and health.

1.4.2.1 Static data collection

Traditionally, the assessment of exposure to air pollutants has primarily relied on data gathered from fixed-site air quality monitoring networks. While these networks can provide a wealth of data on a broad spectrum of pollutants, they are inherently limited to a single geographic point. This spatial limitation, however, is often circumvented by employing interpolation techniques to generate spatial maps that depict average annual concentrations of air pollutants. These derived pollution surfaces enable researchers to spatially link pollutant concentrations with specific populations or subpopulations, such as asthma patients, children, or pregnant women [109].

This approach is particularly useful for large-scale studies focusing on outdoor air pollution [110], but it is not without its limitations. The reliance on assumptions inherent to this indirect method presents challenges when compared to real exposure scenarios [111]. Specifically, exposure assessments based on averaged measurements may artificially dilute pollution levels and rely heavily on aggregated demographic data [112]. This poses significant issues for personal exposure assessments as they do not accurately reflect an individual's unique exposure. Furthermore, the use of data from fixed-site monitoring stations fails to account for individual mobility patterns, particularly the time spent away from the home environment [113].

Static data collection often employs substantial hardware infrastructure distributed across cities or larger geographic areas. An example of this is the Libelium Air Quality Station, a robust and comprehensive air monitoring system, pictured in fig. 9. This hardware is designed to capture a wide array of pollutant data, including levels of nitrogen dioxide, sulfur dioxide, ozone, and particulate matter [114]. These stations, strategically distributed throughout urban and rural areas, work together to create a network that provides a broad and detailed picture of air quality across a region.



Figure 9: Libelium Air Quality Station installed on a pole. Extracted from libelium.com

Despite the inherent spatial limitations of such static data collection methods, they serve as a crucial foundation for understanding regional and city-level trends in air pollution. These trends, in turn, can inform public health initiatives and policy decisions [115].

In the context of exposome research, these traditional methods highlight the need for more sophisticated tools capable of capturing the dynamism and complexity of individual environmental exposures over time

and space.

1.4.2.2 Personal exposure monitoring devices

Personal exposure monitoring (PEM) devices address the limitation of static sensors by offering real-time, individual-level exposure data, thereby providing a more comprehensive picture of the exposome. These devices, often portable and wearable, can monitor a wide range of environmental factors including air pollutants, ultraviolet radiation, noise, temperature, and humidity. By embracing this technology, researchers can further understand the nuances of personal environmental exposure and its impact on health.

Three notable examples of these devices include MicroPEM, Quest Q-300 Noise Dosimeter, and Wristband Passive Samplers. MicroPEM is a portable device that monitors fine particulate matter (PM_{2.5}) in the air, providing researchers with accurate data on exposure to air pollution. The Quest Q-300 Noise Dosimeter is a wearable instrument designed to measure an individual's noise exposure, particularly in occupational settings. Lastly, Wristband Passive Samplers, developed by the Oregon State University Superfund Research Program, are silicone wristbands that absorb and retain various environmental chemicals, allowing for the analysis of personal exposure to a wide range of compounds.

These devices have been utilized in various research projects to assess their effectiveness and gather valuable data on human exposure to environmental factors. For example, the MicroPEM was evaluated in a city with elevated PM_{2.5} levels, demonstrating its utility in monitoring air pollution exposure in urban settings [116]. The Quest Q-300 Noise Dosimeter was employed in a study that investigated the contributions of non-occupational activities to the total noise exposure of construction workers, shedding light on noise exposure levels both inside and outside the workplace [117]. Finally, the Wristband Passive Samplers were featured in a study that highlighted their potential as innovative tools for monitoring personal exposure to various environmental chemicals over time, providing researchers with a unique method for assessing individual exposure to a wide array of compounds [118]. These devices can be seen on fig. 10.



Figure 10: Personal exposure monitoring devices

These devices function by incorporating specialized sensors that respond to specific environmental parameters. For instance, an air pollution monitor may utilize a light scattering sensor to detect airborne particles. The data collected by these devices can then be stored and analyzed, offering detailed insights into the timing, duration, and intensity of exposures.

1.4.2.3 Smartphone applications

In the current digital age, the ubiquity of smartphones has opened up new avenues for collecting personalized environmental exposure data. Most young-adult population carry a smartphone [119], and these devices are equipped with a variety of built-in sensors, making them a convenient and powerful tool for PEM.

A few years ago, the sensing capabilities of smartphones were limited to basic features such as global positioning systems (GPS), cameras, and inertial sensors. However, with rapid technological advancements, smartphones now house a diverse array of sensors with relatively high detection accuracy, transforming them into powerful sensing modules in addition to their primary function as communication hubs. Despite these advancements, the sensors embedded in smartphones still face limitations in monitoring environmental exposures and conditions.

The widespread adoption of smartphones has prompted a shift from stationary environmental sensing methods to more pervasive, personal sensing approaches. Smartphones offer several advantages as tools for individualized environmental sensing. They are affordable, portable, and minimally intrusive, easily blending into daily life without burdening users. However, drawbacks associated with smartphone-based sensing systems include the quality of the embedded sensors and their data, battery life, and user-friendliness [120].

While the data quality of smartphone-based sensors may not be on par with that of larger, more expensive external or static sensors, their low burden enables wide geographic coverage. This broad coverage is facilitated by the portability of smartphones, which necessarily limits the variety of sensors they can carry. Consequently, smartphone-based urban sensing can be classified into opportunistic and participatory sensing [121]. Opportunistic sensing involves the automatic collection of data from embedded smartphone sensors as users go about their daily activities without any explicit input or interaction. On the other hand, participatory sensing requires active user engagement in the data collection process. Users voluntarily contribute data by responding to prompts or actively recording specific environmental conditions, providing a more targeted and context-aware dataset. Both approaches offer unique advantages in exposome research, with opportunistic sensing enabling wider geographic coverage and participatory sensing offering more in-depth and context-specific information about individual exposure.

Different types of exposures and markers can be captured using smartphone applications.

Transportation data

Transportation conditions can be sensed using various features embedded in smartphones, such as accelerometers, GPS, cameras, WiFi, and GSM towers. For example, the Wolverine application primarily uses the accelerometer to estimate road quality, categorizing locations as bumpy, smooth, or experiencing traffic [122]. It assesses traffic based on the number of braking events in forward motion.

Another traffic safety application estimates the time to collision and deceleration rate to avoid a crash (DRAC) using GPS data [123]. These data points help map high-risk zones and segments within the transportation system. To reduce battery drain caused by GPS usage, Panichpapiboon et al. [124] developed an application that uses only the accelerometer to estimate vehicle speed and traffic density, assuming an inverse relationship between density and speed.

General health

Smartphones have become versatile tools for general health monitoring, encompassing physical activity tracking, fall detection, sleep monitoring, general well-being inference, emotion recognition, and even academic performance prediction [125]. These applications typically utilize a variety of sensors embedded in smartphones, including accelerometers, light sensors, GPS, gyroscopes, magnetometers, microphones, and proximity sensors. Custom algorithms are frequently used to process the collected data and ascertain the level of healthy activity.

Sleep monitoring applications have utilized accelerometer and microphone data to estimate sleep duration [126]. However, these applications often fall short in reliability compared to external sensors on the market, such as Jawbone Up, and participatory applications like Sleep-with-phone (SWP).

In the realm of emotional recognition, researchers have used accelerometers, light sensors, and GPS sensors along with machine learning algorithms to measure social activity levels, including phone calls, SMS messages, email activity, and web browsing. The collected data was then used to analyze emotions in terms of pleasure and activeness dimensions [127].

Noise pollution

Noise pollution monitoring is a robust application of environmental sensing that primarily relies on a smartphone's microphone. However, the phone's age and placement during sensing can significantly influence the results.

One challenge with this approach is the variability in microphone quality across different phones. A study that tested noise recording applications on 100 phones [128] (both Android and iOS platforms) found that iOS applications generally provided more accurate results. Android-based applications exhibited higher variability, making them less reliable in scenarios with limited user participation. However, this variability can be mitigated when large numbers of users generate data.

While most applications target the level of sound, some, like DeepEar, aim to identify the type of sound. DeepEar uses deep learning methods to classify sounds into categories like ambient noise, speech, and music [129]. This demonstrates the potential of smartphones to not only measure the intensity of noise pollution but also provide insight into its composition.

1.4.2.4 Questionnaires and interviews

Questionnaires and interviews form an integral part of exposome studies. They serve as primary tools for gathering data on a variety of factors that cannot be measured directly by sensors or biological samples, such as lifestyle habits, dietary intake, occupational exposures, and stress levels. This sort of data has been seen in about 50% of exposome studies used as exposure assessment and in about 25% studies as an outcome according to Haddad et. al. [130].

Typically, these tools are designed and structured to gather comprehensive and accurate data about a participant's exposure history. They can be administered in several ways, including face-to-face interviews, telephone interviews, or self-completed questionnaires, depending on the nature and scope of the study.

The objective of using questionnaires and interviews in exposome studies is twofold. Firstly, they provide insights into the behavioral and social aspects of the exposome. For example, they can capture data on an individual's physical activity levels, dietary habits, and psychological stressors – all of which contribute to health outcomes [131].

Secondly, these tools can help to contextualize and enrich the data obtained from other sources, like personal exposure monitoring devices or biomarker analysis. They can provide information on when and why certain exposures might have occurred, offering a more complete picture of the exposome.

Questionnaires and interviews contribute to exposome studies by enabling researchers to capture a broader spectrum of exposure data. Notably, they allow for the assessment of complex or subjective exposures that are difficult to measure with sensors or biological samples. They also facilitate the capture of longitudinal data, tracking changes in exposure over time.

However, it's important to note that the data obtained through these methods are based on self-reporting, which can be subject to recall bias or reporting errors [132]. Despite these limitations, questionnaires and interviews remain vital for obtaining a holistic understanding of the exposome and its impact on human health.

1.4.2.5 Biological samples

Biological samples, such as blood, urine, and saliva, are crucial components of exposome studies, as they offer valuable insights into internal exposures and biomarkers of effect [133]. By analyzing these samples, researchers can assess the presence and concentration of various chemicals, metabolites, and other substances within the body, which can help to determine the cumulative impact of environmental exposures on human health [134]. Furthermore, biomarkers of effect can provide evidence of early biological responses to these exposures, offering a deeper understanding of the underlying mechanisms linking environmental factors to disease outcomes.

Various techniques are employed for the collection of biological samples in exposome studies, depending on the type of sample and the specific analysis required. Here, we discuss some of the commonly used techniques for collecting blood, urine, and saliva samples.

Blood samples

Blood collection is often performed using venipuncture, where a needle is inserted into a vein to draw the required volume of blood (Lippi et al., 2016). In some cases, capillary blood sampling can be used, which involves collecting blood from a fingertip or heel prick. This method is less invasive and can be more suitable for certain populations, such as infants and children. Blood samples can be processed to obtain plasma, serum, or buffy coat for the analysis of specific biomarkers or chemicals.

Urine samples

Urine collection is a non-invasive method commonly used to assess internal exposures to various environmental chemicals, such as heavy metals or organic pollutants [135]. Urine samples can be collected as spot samples, where a single sample is taken at a specific time, or as 24-hour samples, which provide a more comprehensive representation of the individual's exposure over an entire day.

Saliva samples

Saliva collection is another non-invasive method, often used for measuring hormone levels, oxidative stress markers, or other biomolecules [136]. Saliva can be collected passively by allowing saliva to accumulate in the mouth and then spitting it into a container or by using specialized devices, such as oral swabs or salivettes, which absorb the saliva from the mouth.

Each of these techniques has its advantages and limitations, and the choice of collection method depends on the specific objectives of the exposome study and the nature of the biomarkers or chemicals being analyzed.

Biological samples have been crucial in exposome studies to understand the mechanisms of various health conditions. For example, researchers discovered a connection between prenatal exposure to organophosphate pesticides and altered neurodevelopment in children. They analyzed maternal urine samples and found that higher pesticide exposure was associated with a decrease in cognitive abilities and increased risk of attention problems in the children [137].

Additionally, a study examining the impact of heavy metals on kidney function used blood and urine samples to measure the levels of cadmium and lead in individuals. The findings revealed that exposure to these metals was linked to a higher risk of chronic kidney disease, showcasing the importance of monitoring environmental exposures to protect human health [138].

These examples demonstrate how biological samples can be effectively used in exposome studies to investigate the relationships between environmental exposures, genetic factors, and health outcomes, providing valuable insights into the complex mechanisms underlying various diseases and conditions.

1.4.3 Data integration

This section will discuss the challenges and strategies for integrating diverse and multi-dimensional exposome data. We will delve into the harmonization of data from various sources and the adoption of common data models and ontologies. By addressing these challenges and implementing effective strategies, researchers can gain a more comprehensive understanding of the exposome and its impact on human health, ultimately leading to the development of more effective preventive measures and personalized treatments.

Data harmonization is an essential process when dealing with diverse and multi-dimensional exposome data from different sources, as it helps integrate heterogeneous data types and structures. One of the main challenges is to ensure that variables are measured and represented consistently across all datasets.

The first step towards data harmonization is data standardization [130]. This involves transforming data from multiple sources into a standardized format. For example, researchers may need to convert units, categorize continuous variables, or recode categorical variables using the same coding scheme [139]. This ensures that data can be compared and combined more easily.

Another crucial aspect of data harmonization is data cleaning and quality control [140]. Ensuring data quality is essential for reliable analysis, and this process involves identifying and correcting errors, inconsistencies,

and missing values in the data. Quality control procedures may include range checks, validation of data entry, and cross-checking with other related variables or external sources.

Data privacy and security are crucial aspects of data integration in exposome research, as the collected data often contains sensitive personal information. Various techniques can be employed to safeguard data, such as de-identification, anonymization, and encryption [141]. These methods help minimize the risk of data breaches and protect the privacy of individuals. Additionally, researchers need to comply with data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union. By addressing data privacy and security concerns, researchers can maintain the trust of study participants and ensure the ethical handling of sensitive information.

The use of common data models and ontologies is another important aspect of data integration in exposome research. These standardized frameworks facilitate data integration by providing a consistent way of representing and organizing data from different sources [142]. Adopting common data models and ontologies allows researchers to easily map variables across datasets, ensuring that the integrated data are consistent and meaningful. This approach promotes interoperability among datasets and helps researchers to draw more reliable conclusions from the integrated data. By implementing common data models and ontologies, exposome researchers can effectively tackle the challenges associated with integrating diverse and multi-dimensional data, thus improving the overall quality of their research findings.

By employing these techniques, researchers can effectively harmonize diverse data sources, making it easier to draw meaningful conclusions from integrated exposome data.

1.4.4 Statistical analysis methods

In this section, we present an overview of the key statistical methods used in exposome analysis, highlighting their significance and applications in addressing various research questions.

The exposome encompasses both external and internal factors, including chemical, biological, and lifestyle exposures, as well as genetic, epigenetic, and molecular omics responses. Due to the vast number of exposures and the dynamic nature of biological responses, it is crucial to adopt statistical methods that can account for high-dimensional data and complex correlation structures.

We will discuss various statistical methods employed in exposome research, ranging from dealing with missing data, single-exposure methods, such as Exposome-wide Association Studies (ExWAS), to multi-exposure methods, like variable selection techniques and dimension reduction approaches. Additionally, we will delve into the integration of omics data in exposome research, exploring network-based approaches, cross-omics analyses, and other advanced methods. Lastly, we will touch upon the importance of considering sample size in exposome studies to ensure adequate statistical power.

Understanding and implementing these statistical methods is essential for researchers aiming to explore the exposome and its implications for human health. By leveraging these techniques, it is possible to uncover novel associations, advance our knowledge of the complex interplay between environmental exposures and health, and ultimately contribute to the development of effective prevention and intervention strategies.

1.4.4.1 Missing data

Missing data in an exposome context can be problematic, as the number of complete cases may decrease with the inclusion of more exposures. It is recommended to use imputation techniques, such as multiple imputation, to handle missing data in epidemiological studies [143]. However, applying multiple imputation to large datasets presents additional difficulties [144]. Imputation models should include no more than 15-25 predictors to avoid issues related to convergence due to predictors collinearity [145].

For exposures measured through biochemical assays, some values may be below the limit of detection (LOD). The LOD is the lowest quantity of an exposure that can be detected by a specific method. A common approach is to replace values below the LOD with a fixed value such as the LOD, half the LOD, or $\text{LOD}/\sqrt{2}$ [146]. Single substitution might be acceptable when the proportion of values below the LOD is low (e.g.,

<20%). However, exposures with a high proportion of values below the LOD (e.g., >80%) should either not be used or dichotomized into detected/undetected. Exposures with values below the limit of quantification, which is the lowest quantity of an exposure that can be detected with stated accuracy and precision, should be interpreted with caution.

1.4.4.2 Single-exposure Methods

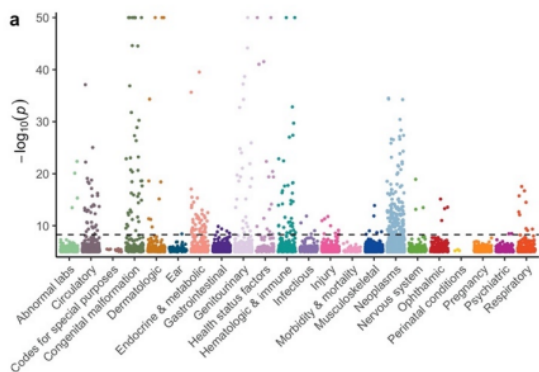
Exposome-wide association studies (ExWAS)

Exposome-wide association studies (ExWAS), pioneered by Juarez et. al. [147], are a comprehensive approach used to investigate the relationship between a wide range of environmental exposures and health outcomes. In these studies, the health outcome of interest is modeled as a function of multiple exposures, with each exposure being treated as a separate covariate in the analysis. By employing linear regression models, researchers can examine the association between individual exposures and the health outcome while adjusting for potential confounding variables.

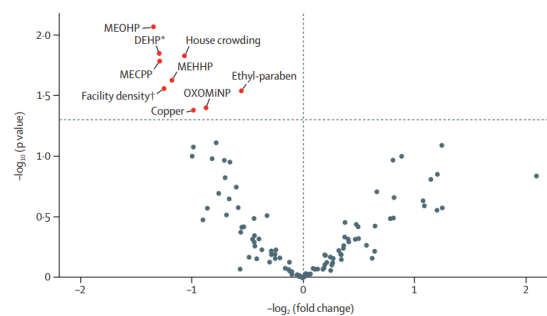
In an ExWAS, a linear regression model is fit for each exposure, with the health outcome as the dependent variable and the exposure as the independent variable, while controlling for other relevant covariates, this is exemplified on eq. (15). This approach allows for the simultaneous evaluation of multiple environmental exposures and their potential impact on health outcomes. The associations found between individual exposures and health outcomes can be used to identify potential environmental risk factors and guide further research into their potential biological mechanisms.

$$\begin{aligned} outcome &\sim exposure_1 + covariate_1 + \dots + covariate_N \\ &\dots \\ outcome &\sim exposure_M + covariate_1 + \dots + covariate_N \end{aligned} \quad (15)$$

However, it is essential to note that ExWAS typically involve a large number of exposures, which can lead to multiple testing issues. To address this, researchers often apply corrections for multiple testing, such as the Bonferroni correction or false discovery rate (FDR) control, to maintain an appropriate level of statistical significance [148]. By using these methods, ExWAS can provide valuable insights into the complex relationships between various environmental exposures and health outcomes, facilitating a more comprehensive understanding of the role of the exposome in human health. ExWAS analysis can effortlessly incorporate multiple imputed datasets. The majority of software applications or tools have the capability to autonomously perform regression analysis on each imputed dataset and merge the outcomes, taking into account the uncertainty arising from imputations. In order to visualize the results produced by this method, Manhattan and Volcano plots are typically used, an example of them can be seen in fig. 11.



(a) Example of Manhattan plot. Extracted from Wang et. al. [149]



(b) Example of Volcano plot. Extracted from Agier et. al. [150]

Figure 11: ExWAS visualizations

1.4.4.3 Multi-exposure methods

Dimension Reduction and Variable Selection are two distinct strategies for addressing the challenges posed by high-dimensional data in exposome research, but they serve different purposes and have different implications.

Dimension Reduction techniques aim to reduce the complexity of the data by transforming a large set of correlated exposures into a smaller set of new variables, often called latent variables or components. These new variables are linear combinations of the original exposures and are designed to capture most of the variance in the original data. This approach simplifies the analysis by reducing the number of variables to a more manageable size, while still retaining the essential information present in the original data. Dimension reduction techniques are especially useful for exploring and visualizing relationships between variables and for studying the combined effects of multiple exposures on health outcomes.

Variable Selection techniques, on the other hand, focus on identifying a subset of the original exposures that are most relevant or informative for predicting a specific health outcome. These techniques search for an optimal set of exposures by assessing the predictive performance of different combinations of exposures in the context of a statistical model. Variable selection techniques can help to identify key exposures that have a strong association with the health outcome of interest, while accounting for multicollinearity and controlling for false positives.

In summary, Dimension Reduction techniques transform the original high-dimensional data into a smaller set of new variables that retain the essential information, while Variable Selection techniques identify a subset of the original exposures that are most relevant for predicting the health outcome of interest. Both strategies are useful for analyzing high-dimensional exposome data, but their applicability and interpretation depend on the specific research question and goals.

Variable Selection Techniques

The Deletion/Substitution/Addition (DSA) algorithm is a model selection technique that constructs and assesses models with different combinations of exposures to identify the optimal set of exposure predictors [151]. The algorithm iteratively removes, adds, or replaces predictors in the model, aiming to minimize the prediction error. It has been applied to exposome research in recent studies of the HELIX project associating early-life exposures and childhood lung function [150].

Another approach, Elastic Net (ENET), is a regularized regression method that combines the L1 and L2 penalties of LASSO and Ridge regression, respectively. This method selects multiple correlated exposure variables, effectively controlling for multicollinearity and allowing the identification of key exposures associated with health outcomes [152]. ENET has been used in exposome research to investigate relationships between multiple environmental contaminants and birth weight [153].

Dimension Reduction Techniques

Principal Component Analysis (PCA) is a widely used dimension reduction technique that transforms a set of correlated exposures into a smaller set of uncorrelated linear combinations known as principal components [154]. These principal components capture most of the variance in the original data, allowing researchers to study the effects of correlated exposures on health outcomes while reducing the complexity of the data. PCA has been applied in exposome research to analyze the daily mortality relationship with air pollution in Beijing [yang2013].

Partial least squares (PLS) regression considers the correlation between outcome and exposure variables by merging principal component analysis (PCA) and multiple regression analysis [155]. PLS regression aims to find a linear breakdown of the exposure matrix that maximizes the covariance between exposure and outcome. The stronger the correlation between an exposure variable and the outcome, the greater the weight assigned to that exposure variable in the linear mix. PLS regression also allows for the inclusion of multiple outcome variables. To determine the ideal number of components, the mean squared error of prediction is used in conjunction with cross-validation [156].

1.4.4.4 Incorporating omics into exposome research

The internal exposome, measured through omics data, which encompass a wide range of high-throughput molecular methodologies such as epigenetics, gene expression, and metabolism, play a crucial role in understanding the complex relationships between environmental exposures and health outcomes. These data can be integrated into exposome research in various ways, such as predictors, mediators, or outcomes. This integration helps to provide deeper insights into the underlying biological mechanisms and pathways linking exposures to health effects.

Network-based approaches are valuable tools for organizing and analyzing high-dimensional omics data in exposome research. These approaches help to visualize and summarize information by identifying hubs of correlated exposures and interpreting systemic biological changes that associate with multiple exposures and health effects [157]. This enables researchers to reveal the grouping of exposures based on their correlation in a population or their chemical or toxicological properties [158].

Dimension reduction techniques, can be employed to analyze omics data in the context of exposome research. As previously stated, these techniques simplify the analysis by reducing the number of variables, while still retaining the essential information present in the original data. They are particularly useful for exploring and visualizing relationships between variables and studying the combined effects of multiple exposures and biomarkers on health outcomes.

Cross-omics analyses investigate how exposure and/or outcome-related signals found at one molecular level correlate with those found at another level, providing insight into the molecular cascades related to specific exposures and/or outcomes [159]. These analyses can be performed in two different ways, (1) relying on pre-existing biological knowledge by connecting the omics layers through a shared gene/pathway identifier and performing pathway enrichment analyses or identifying candidate omics markers after exploratory analysis in a different omics layer, or (2) without prior biological knowledge, employing methods such as multi-block PLS models or canonical regression analyses [160].

Meet-in-the-middle approaches [161] are useful for identifying biomarkers linking exposures and disease outcomes, considering the internal exposome as a mediator of the external exposome-health outcome. This method can help to pinpoint potential targets for intervention and prevention strategies.

1.4.4.5 Sample size in an exposome context

One of the challenges in exposome research is managing the issues related to multiple testing and low-to-moderate effect sizes of individual exposures, this gets aggravated when when a considerable percentage of concentrations fall below the LOD. To partially address these concerns, researchers can increase the sample size of their studies to improve statistical power and reduce the likelihood of false-positive results.

A previous study explored the required sample size for conducting an Exposome-Wide Association Study (ExWAS) approach with 100 exposures in relation to male fertility outcomes [162]. This study found that, in order to achieve 80% power to detect the 95th percentile effect sizes, a sample size of 1,000 to 2,000 subjects would be necessary. This sample size is considerably larger than what would be required when considering a single exposure. It is essential to note that this recommendation may vary depending on the specific exposures, outcomes, and population under investigation.

By increasing the sample size, researchers can better manage the challenges associated with multiple testing and low-to-moderate effect sizes, ultimately contributing to more robust and reliable exposome studies.

1.4.5 Challenges and limitations

Despite the promising potential of exposome research in the field of environmental health, it faces a number of significant challenges and limitations. These range from the complexity and dynamic nature of environmental exposures, to the difficulty in establishing definitive causal relationships between these exposures and health outcomes. In this section we will discuss about these challenges.

1.4.5.1 The Complexity and Dynamic Nature of Environmental Exposures

Environmental exposures represent a vast and varied array of factors, including chemical contaminants, dietary components, physical factors, infectious agents, and more. The exposome encompasses all of these, along with their interactions and cumulative effects over time, adding layers of complexity to its study.

Moreover, these exposures are not static; they are dynamic and evolve throughout an individual's lifespan. This temporal variability adds another dimension of difficulty to accurately measuring and characterizing exposures. For instance, a particular exposure might not only vary in concentration, but also in its nature and impact depending upon the life stage at which it occurs [163].

The sheer volume and variety of potential environmental exposures pose significant challenges for accurate measurement. Traditional methods, such as questionnaires and personal reports, are often subject to recall bias and may not capture the full spectrum of exposures. Biomonitoring, which measures the concentrations of substances in body fluids or tissues, offers a more objective measure of exposure. However, it is limited by technical constraints, including the sensitivity and specificity of the analytic methods, and the temporal relevance of the samples [164].

Furthermore, many environmental factors are interlinked and can interact with each other in complex and often unpredictable ways, further complicating the characterization of the exposome. For example, the impact of a chemical exposure could be modulated by an individual's diet, physical activity levels, or concurrent exposures to other chemicals [165].

In summary, the vast and diverse nature of environmental exposures, their dynamic evolution over time, and the intricate interactions between them pose substantial challenges to the accurate measurement and characterization of the exposome.

1.4.5.2 Statistical Challenges in Exposome Research

Beyond the complexities of measurement and characterization, the exposome presents unique statistical challenges as well. The high-dimensionality of exposome data, resulting from the vast number of potential exposures, poses a significant issue for data analysis. Traditional statistical methods often fall short in managing such high-dimensional data, potentially leading to spurious correlations and inflated false discovery rates [166].

One of the main issues in exposome research is the problem of multiple testing, where a large number of statistical tests are performed simultaneously. As the number of tests increases, so does the likelihood of obtaining significant results purely by chance. This is a particular concern in exposome research given the massive quantity of potential exposures being tested for associations with health outcomes [167].

Moreover, the correlation structure of the exposome data adds another layer of complexity. Environmental exposures are often correlated with each other due to shared sources or common behavioral, socio-demographic, or physiological determinants. This intercorrelation can complicate the interpretation of results and poses challenges for statistical models that assume independence among predictors.

Finally, the temporal dynamics of the exposome introduce additional challenges for statistical analysis. Environmental exposures can vary significantly over time, and different time windows of exposure may have different effects on health outcomes. This necessitates the use of advanced statistical methods capable of handling time-varying exposures and lagged effects, which require sophisticated modeling techniques [168].

In conclusion, the high-dimensionality, multiple testing issue, complex correlation structure, and temporal dynamics of the exposome present significant statistical challenges that must be addressed to unleash the full potential of exposome research.

1.5 Overview of Omics Data Analysis

Building on our understanding of exposome data analysis, it's pertinent to also acknowledge another significant frontier in the realm of complex data: omics data analysis. While the exposome provides a holistic view

of environmental exposures, the omics realm offers an intricate glimpse into the molecular constituents and processes that define living organisms — from genomics and proteomics to metabolomics and beyond.

1.5.1 Introduction to Omics Data Analysis

Omics data analysis is a key component of bioinformatics, a discipline that emerged in response to the need to manage and interpret the massive amount of data generated by genomic (and other omics) research over the past decade. This field represents the convergence of genomics, biotechnology, and information technology, and it involves the analysis and interpretation of data, modeling of biological phenomena, and the development of algorithms and statistics [169].

In the context of omics data analysis, bioinformatics tools are used to analyze and interpret the vast amounts of data generated by various omics technologies, such as genomics, proteomics, and metabolomics, among others. This analysis can provide valuable insights into various biological and disease systems, helping to uncover molecular interactions and potential transduction pathways [170].

Throughout this section, our aim is to give you a comprehensive overview of omics data analysis, highlighting the challenges, methodologies, and potential future directions in this field.

1.5.2 Types of omic data

There are many types of omics data, each type of omics data provides a unique perspective on the biological processes within an organism, and together they offer a comprehensive view of an organism's biology. We will provide definitions of genomics, epigenomics, proteomics, metabolomics, and transcriptomics, given those are some of the most studied ones. By understanding these different types of omics data, we can better appreciate the complexity of biological systems.

1.5.2.1 Genomics

Genomics is the study of the entire genome of an organism. It involves the analysis of the structure, function, and evolution of genes. With the rise of high-throughput sequencing technologies, genomics has become a big data discipline, generating massive amounts of data that need to be processed and interpreted [171]. Genomics research has been key to uncover links between genotype and phenotype [172].

Genomics also involves the investigation of the interplay and influence of genes on each other and the organism's environment. The field has seen the development of robust algorithms, including those based on deep learning [173], to handle the data explosion and provide meaningful interpretations.

1.5.2.2 Epigenomics

Epigenomics is the study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. These modifications, which include DNA methylation and histone modification, among others, can influence gene expression without altering the underlying DNA sequence [174].

Epigenomic processes are increasingly recognized for their fundamental role in diseases such as cancer [175], as they reflect environmental risk factors and can provide a means by which to assess genomic regulatory interactions.

1.5.2.3 Proteomics

Proteomics is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The field of proteomics has seen significant advancements with the advent of mass spectrometry-based proteomics, which has become the tool of choice for identifying and quantifying the proteome of an organism [176].

Proteomics research has resulted in a versatile collection of tools that allow for the uncovering of links between protein structure and function [177].

1.5.2.4 Metabolomics

Metabolomics is the scientific study of chemical processes involving metabolites, the small molecule substrates, intermediates, and products of metabolism. Specifically, is the study of the unique chemical fingerprints that specific cellular processes leave behind.

As with genomics, metabolomics has become a big data discipline, generating massive amounts of data. Metabolomics research has resulted in tools that allow for the understanding of links between metabolic profiles and physiological states [178].

Metabolomics data can be used to infer effects of environmental factors on metabolism [179] and the dynamic changes in metabolic profiles under different conditions [180].

1.5.2.5 Transcriptomics

Transcriptomics is the study of the transcriptome—the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods.

By comparing the transcriptomes of different cells, or the same cells under different conditions, researchers can gain insights into how gene expression changes in response to various factors such as disease states, environmental stressors, or developmental stages. This can be instrumental in understanding complex biological processes and diseases, and in the development of new therapeutic strategies [181].

1.5.3 Data collection

In the era of big data in biology, high-throughput technologies have revolutionized the way we collect, analyze, and interpret biological information [182]. Among these, Next-Generation Sequencing (NGS) of DNA and RNA, and mass spectrometry-based proteomics stand out as particularly impactful in the realm of multi-omics research. They allow for the generation of extensive datasets representing the different layers of biological information, forming the backbone of multi-omics research.

1.5.3.1 Next-Generation Sequencing (NGS)

Next-generation sequencing allows for the sequencing of DNA and RNA much more quickly and cheaply than older sequencing methods. This technology has revolutionized genomics, enabling the sequencing of whole genomes or targeted regions. Different NGS platforms like Illumina, Ion Torrent, and others offer various sequencing capabilities suitable for different research needs. NGS generates vast amounts of data, necessitating the development of sophisticated computational tools and approaches for data processing, variant calling, and interpretation [183].



Figure 12: Illumina MiniSeq System

1.5.3.2 Mass Spectrometry-based Proteomics

Mass spectrometry-based proteomics is a key technique for studying proteins, enabling identification, quantification, and characterization of complex protein mixtures. It involves the use of mass spectrometry, a tool that measures the mass-to-charge ratio of ions to identify and quantify molecules. In proteomics, proteins are typically digested into peptides, which are then ionized and analyzed by the mass spectrometer. This technology has enabled high-throughput protein profiling and biomarker discovery [184].



Figure 13: Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer

1.5.4 Multi-omics data analysis

The integration of multiple omics data types—such as the ones we have just described, provides a more comprehensive view of the biological system being studied. This approach, known as multi-omics data analysis, can help uncover complex molecular interactions and provide better insights into the mechanisms of biological processes.

The use of multi-omics data analysis has led to significant discoveries in various fields. For instance, in cancer research, the use of multi-omics data has been useful in tumor subtyping, prognosis and diagnosis [185]. Similarly, in microbiome research, the combination of metagenomics, metatranscriptomics, and metabolomics data has provided a more holistic view of microbial communities and their functions [186].

However, multi-omics data analysis presents several challenges. Different omics data types can vary greatly in terms of their scale, distribution, and complexity. This makes it difficult to integrate the data and interpret the results. Furthermore, each omics data type has its own set of technical and biological biases, which need to be accounted for in the analysis.

1.5.4.1 Multi-omics data integration

Similar to exposome research, the challenges of multi-omics data are high dimensionality, variability, and complexity of data. Therefore, the methods explained on the exposome section regarding data integration are very similar to the ones used on multi-omics data analysis.

These methods include data pre-processing and normalization to adjust the scale of data, dimensionality reduction techniques like Principal Component Analysis (PCA) to manage data complexity, and various statistical and machine learning techniques for data integration, such as multiple co-inertia analysis (MCIA), canonical correlation analysis (CCA), or partial least squares (PLS).

1.5.5 Statistical Analysis in Omics Data

Once the multi-omics data is integrated, the subsequent step is the analysis of this data to extract meaningful biological insights. Here we will talk about some of the most used methods.

1.5.5.1 Differential Expression Analysis

Differential expression analysis is used to identify genes, proteins, or metabolites whose abundance (like gene expression levels) changes significantly under different experimental conditions. It involves statistical methods that take into account the variability of measurements within and between groups to estimate the likelihood that observed differences occurred by chance.

The most common statistical technique employed is the Student's t-test for comparing two groups, or Analysis of Variance (ANOVA) for comparing more than two groups. These methods rely on the assumption that the data are normally distributed and have similar variance within each group [187].

However, omics data often does not comply with these assumptions. Therefore, different techniques are usually used. These include the limma package [188], which uses an empirical Bayes method to moderate the standard errors of the estimated log-fold changes, providing stable results even for experiments with small numbers of replicates.

For RNA-seq count data, which follows a negative binomial distribution, methods like DESeq2 [189] and edgeR [190] have been developed. These methods use a generalized linear model approach to estimate the variance and handle biological replicates.

All these methods provide a measure of the magnitude of differential expression (fold change), and a p-value indicating the statistical significance of the observed change, which is then adjusted for multiple testing, often using the Benjamini-Hochberg procedure to control the false discovery rate [148].

1.5.5.2 Correlation analysis

Correlation analysis is a statistical method used to evaluate the strength and direction of the relationship between two or more variables. Pearson and Spearman correlation are commonly used, depending on whether the data follow a normal distribution [191].

The Pearson correlation coefficient measures the linear relationship between two variables and assumes that the data are normally distributed. It returns a value between -1 and 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

On the other hand, the Spearman correlation coefficient, a non-parametric measure of correlation, assesses the monotonic relationship between two variables without making any assumptions about the distribution of the data. This means that it only measures the direction of association (positive or negative) but not the exact linear relationship.

In addition, to infer whether a detected correlation is statistically significant, a hypothesis test is often performed. The null hypothesis is that the correlation coefficient in the population from which the sample was drawn is zero. A p-value is then calculated to decide whether to reject the null hypothesis. Afterward, the p-values may need to be adjusted for multiple testing.

It's worth noting that correlation does not imply causation, meaning that even if two variables are correlated, it does not necessarily mean that changes in one variable cause changes in the other.

1.5.5.3 Clustering Analysis

Clustering analysis is a key unsupervised learning technique commonly used in multi-omics data analysis. The goal of clustering is to group or partition the samples (or variables) into clusters so that the samples (or variables) within the same cluster are more similar to each other according to certain criteria, compared to those in other clusters.

Hierarchical Clustering [192] is an unsupervised machine learning method used to group similar objects into clusters. It constructs a hierarchy of clusters, where each node is a cluster comprising the objects and their subclusters. This hierarchy can be visualized as a dendrogram, a tree-like diagram (illustrated on fig. 14). The process continues until all objects are in a single cluster or until a termination condition is met. The resulting structure allows for varying levels of granularity in clustering, offering flexibility based on user requirements.

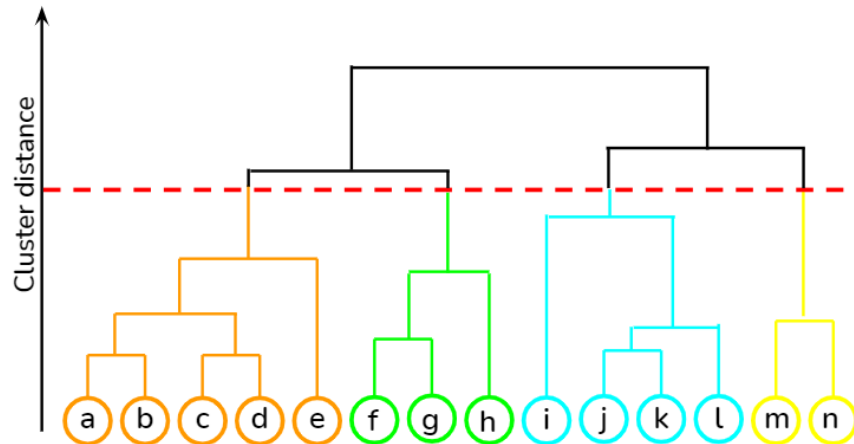


Figure 14: Hierarchical clustering dendrogram

K-means Clustering [193] is another popular method. It partitions the samples into K clusters in which each sample belongs to the cluster with the nearest mean. It is necessary to specify the number of clusters (K) in advance, which can be a disadvantage if the number of clusters is not known a priori. Different from hierarchical clustering, it is common to visualize the results using plain scatter plots (illustrated on fig. 15).

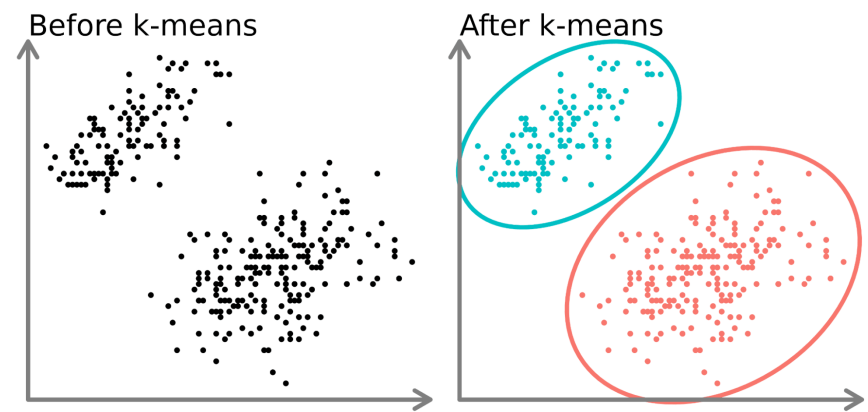


Figure 15: K-means 2D visualization

Determining the number of clusters is a challenge in clustering analysis. Techniques such as the elbow method, silhouette analysis, or gap statistic are used to infer the optimal number of clusters [194].

While clustering techniques can reveal important patterns and groupings in the data, it's important to remember that these are exploratory methods and the resulting clusters need to be validated and interpreted in the context of known biology or followed up with further experiments.

1.5.5.4 Network Analysis

Network analysis is an advanced statistical approach that models and explores the complex interactions and relationships between different biological molecules [195].

A network is a graph that consists of nodes (which represent biological entities such as genes, proteins, or metabolites) and edges (which represent interactions or relationships between these entities). It is typically visualized as on fig. 16.

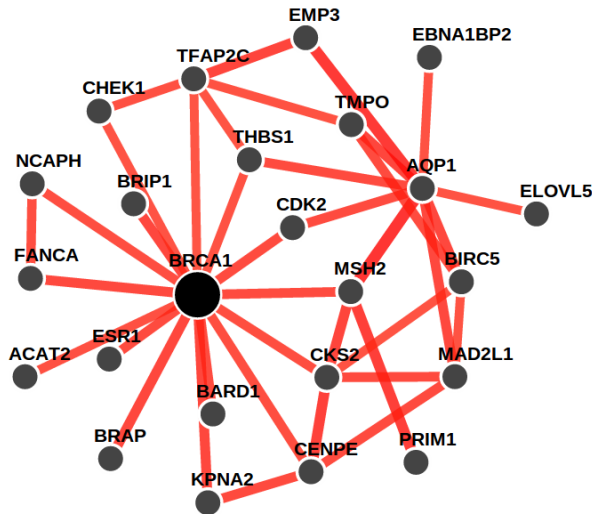


Figure 16: Network analysis visualization: Nodes as black circles, paths as red lines.

Key concepts in network analysis include "degree" (the number of edges connected to a node), "path" (a sequence of nodes and edges connecting two nodes), and "centrality" measures (how important a node is within the network).

Community detection or clustering in networks is another key concept, aiming to identify densely connected groups of nodes, called "modules" or "communities". These modules often correspond to biological pathways.

However, network construction and analysis require careful statistical handling, especially in terms of determining the significance of connections and controlling the false discovery rate. Randomization methods and permutation tests are often used for this purpose.

1.5.5.5 Pathway Analysis

Pathway analysis aims to identify biological pathways significantly enriched in a list of genes, proteins, or metabolites of interest. These pathways often represent a series of actions or changes within a cell that lead to certain cell functions or disease processes [196].

Many online databases and tools are available for pathway analysis, such as KEGG (Kyoto Encyclopedia of Genes and Genomes), Reactome, and MSigDB (Molecular Signatures Database).

Pathway analysis can guide the generation of new hypotheses, the design of subsequent experiments, and the validation of findings. It has proven instrumental in biomedical research, such as identifying the biological roles of potential therapeutic targets for cancer treatment and examining the molecular similarity and dissimilarity between sample groups [197].

It's important to note that while pathway analysis can provide mechanistic insights into the biological processes associated with the list of interest, the results need to be interpreted with caution, especially

considering the incomplete and sometimes inaccurate information in pathway databases.

2 Hypotheses and objectives

2.1 Hypotheses

The DataSHIELD framework stands as a remarkable tool for performing secure and collaborative data analysis, especially accentuated within the ATHLETE project. However, certain limitations have surfaced that require rectification to fully leverage its potential. These limitations encompass: 1) the exigency for more efficient mechanisms to navigate through large and diverse datasets spanning multiple domains within DataSHIELD. 2) The need for privacy-preserving methodologies during omics data analyses; while DataSHIELD offers a secure analysis environment, the privacy preservation in omics data realms demands advanced federated data analysis methodologies to uphold confidentiality without sacrificing analytical accuracy. 3) The imperative for advanced analytical methods to carry out robust, privacy-preserving exposome data analysis. 4) The absence of innovative graphical solutions, which can potentially deter researchers, especially those with limited technical prowess, from adopting DataSHIELD for non-disclosive analysis; a more intuitive and visually guided interface can dramatically lower the entry barrier. 5) The challenge of employing DataSHIELD in a manner that facilitates secure, collaborative multi-center analyses of real-world data. All those limitations are encompassed within the thesis hypothesis.

— **Hypotheses 1: Enhancing DataSHIELD’s data management capabilities could enable efficient, privacy-preserving analysis of larger, more diverse datasets.**

General hypothesis

The implementation of enhanced data management capabilities within the DataSHIELD platform enables federated analysis of larger and more diverse datasets. This extension broadens the utility of the platform into new domains, offering opportunities for deeper insights while preserving privacy and data security.

Specific hypothesis

- Enhanced data source management within DataSHIELD could enable efficient handling and analysis of large and diverse datasets, including geospatial and genomic data.
- Seamless utilization of data structures from existing R packages could significantly improve DataSHIELD’s capability to manage and analyze complex datasets.
- Extending DataSHIELD’s functionality could allow for the analysis of sensitive data from multiple domains without compromising privacy.
- Federated privacy-protecting analysis of genomic and geospatial data may be achievable through new analytical methodologies within DataSHIELD.
- The inclusion of disclosure control in DataSHIELD allows for adjustable levels of privacy protection tailored to the specific data and analysis context, further enhancing its applicability in sensitive data analysis.
- The application of DataSHIELD to larger datasets will be more efficient with functions based on the tidyverse as well as base-R, including the potential integration of the "dplyr" package for operating on tabular datasets.

— **Hypotheses 2: Privacy-preserving omics data analyses can be enabled through advanced federated data analysis methodologies**

General hypothesis

The implementation of a privacy-preserving software solution will facilitate the execution of omics data analyses across multi-centre studies without physically sharing data. This advancement is anticipated to promote enhanced research collaboration, fostering a conducive environment for data-driven research while upholding data privacy standards.

Specific hypothesis

- Implementation of state-of-the-art methods for GWAS, transcriptomic, and epigenomic data analyses in a federated setting may be achieved without compromising privacy.
- A comprehensive approach to quality control and easier integration to available pipelines may surpass existing methodologies like FAHME and sPLINK, particularly in the realm of GWAS.
- Configurable differential privacy and disclosure traps at the data source level may offer a flexible and trustworthy platform for conducting research while protecting sensitive data.
- The ability to perform non-disclosive pooled and IPD meta-analysis may provide researchers with a choice of methods to suit specific data characteristics and study design.
- Integration of advanced omic data analysis methodologies may be successfully achieved in existing large consortia projects that have set up infrastructure for Federated Analysis using DataSHIELD.
- Guided by user needs, the evolution of federated omic data analysis tools may continue to expand and improve, driving new research initiatives and findings.

— **Hypotheses 3: Advanced analytical methods can enable robust, privacy-preserving exposome data analysis.**

General hypothesis

Employing an advanced analytical framework alongside the DataSHIELD platform may facilitate robust, privacy-ensuring analysis of exposome data within a multi-centre study framework. This approach can overcome challenges associated with data sharing, harmonization, and standardization in exposome research.

Specific hypothesis

- An advanced analytical tool can facilitate Exploratory Exposome-Wide Association Studies (ExWAS) using synthetic data, showcasing utility in exploratory analyses of the exposome.
- A new analytical tool may effectively replicate real-world analysis for ExWAS in multi-centre studies, displaying compatibility with real-world data and potential utility for exposome researchers.
- The ability to handle confounding factors in exposome analysis, may confirm the robustness of advanced analytical methods in adjusting for various potential confounders in ExWAS.
- Integration of advanced analytics with the DataSHIELD infrastructure may offer an advantage over traditional meta-analysis methods by enabling pooled analyses in multi-centre studies.

— **Hypotheses 4: Innovative graphical solutions will boost DataSHIELD adoption for non-disclosive analysis**

General hypothesis

A user-friendly and efficient tool will enhance the accessibility and adoption of DataSHIELD infrastructure for federated non-disclosive analysis, catering to both researchers with limited R skills and those experienced in DataSHIELD who seek a platform for quick hypothesis prototyping and analysis.

Specific hypothesis

- An intuitive user interface and step-by-step processes will increase the user base for DataSHIELD by attracting researchers without advanced R skills.
- Experienced DataSHIELD users will benefit from a solution that enables quick hypothesis prototyping and fast analyses without the need to write complex analysis pipelines.

- A solution with comprehensive functionality and modular structure will be easy to upgrade and maintain, ensuring its viability and relevance for future research requirements.
- A user-friendly tool will foster rapid understanding and utilization of DataSHIELD, thus contributing to the broader adoption of non-disclosive analysis methods.

— **Hypotheses 5: Application of DataSHIELD will enable secure, collaborative multi-center analyses of real-world data**

General hypothesis

The application of DataSHIELD to real-world data can effectively address the challenges of data access, sharing, and analysis. Following developments outlined in this thesis, DataSHIELD will be successfully used in real cases, enabling secure multi-center analyses while preserving data privacy and confidentiality.

Specific hypothesis

- The use of DataSHIELD in multi-center studies will enhance collaboration among researchers, enabling more comprehensive and robust analyses without compromising data privacy and confidentiality.
- The application of DataSHIELD to real-world cases will enable the identification of undetected patterns and relationships within data.
- The utilization of DataSHIELD will result in increased reproducibility and transparency in research, contributing to the overall quality and reliability of scientific findings.

2.2 Objectives

Building upon the identified gaps and proposed solutions articulated in the hypothesis, this section delineates the objectives aimed at addressing these challenges within the DataSHIELD framework. These objectives are intricately tied to the ATHLETE project, having been developed specifically to propel the project's federated, non-disclosive analysis capabilities to new heights. The goals outlined here aim to not only ease the user experience for researchers with varying levels of R proficiency but also significantly enhance the efficiency and versatility of handling large and diverse datasets across a multitude of scientific disciplines. The overarching aim is to evolve the DataSHIELD infrastructure into a more robust, flexible, and user-friendly platform that facilitates collaborative research, and can adeptly manage and analyze vast and diverse data types. By achieving these objectives, the path will be paved for groundbreaking discoveries and insights in the ATHLETE project, extending the horizons of what can be accomplished in a secure and collaborative research environment.

— **Objective 1: Showcase the development and expansion of DataSHIELD for efficient, secure, and private federated analysis**

General objective

To elevate the DataSHIELD platform, employing the introduction of resources for adept handling and federated analysis of diverse, large, and sensitive datasets, all while preserving privacy and security.

Specific objective

- Establish enhanced data source management within DataSHIELD, using the concept of "resources" to enable efficient handling and analysis of large and diverse datasets, including exposome and genomic data.
- Elevate DataSHIELD's capability to manage and analyze complex datasets by seamlessly integrating data structures from existing R packages through the "resourcer" package.
- Extend DataSHIELD's functionality to enable analysis of sensitive data from multiple domains without compromising privacy, utilizing the incorporation of resources.
- Achieve federated privacy-protecting analysis of genomic and geospatial data by developing and applying the "dsOmics" and "dsExposome" packages within DataSHIELD.
- Amplify the efficiency of DataSHIELD's application to larger datasets by integrating functions based on the tidyverse as well as base-R, including the potential integration of the "dplyr" package for operating on tabular datasets.

— **Objective 2: Advancing and Validating OmicSHIELD: An Open-Source Initiative for Privacy-Preserved Omics Data Analyses**

General objective

Drive the development and validation of OmicSHIELD, an open-source software, fostering a non-disclosive omics data analyses avenue across multi-centre studies while adhering to stringent privacy standards.

Specific objective

- Implement state-of-the-art methods for GWAS, transcriptomic, and epigenomic data analyses within OmicSHIELD to conduct federated analyses without compromising privacy.
- Establish a robust approach to quality control and population stratification adjustment surpassing existing methodologies like FAHME and sPLINK, particularly within GWAS, utilizing OmicSHIELD's enhanced algorithms.
- Integrate configurable differential privacy and disclosure trap settings at the data source level in OmicSHIELD, offering a flexible and trustworthy platform for privacy-preserving research.

- Enable the execution of non-disclosive pooled and IPD meta-analyses within OmicSHIELD, presenting researchers with methodological choices suited to specific data attributes and study designs.
- Undertake multi-centric GWAS of CINECA data, and DGE and EWAS analysis of HELIX data using OmicSHIELD, aiming to corroborate results with traditional local computation approaches.
- Facilitate the adoption of OmicSHIELD within existing large consortia projects with established infrastructure for Federated Analysis using DataSHIELD, for advanced omic data analysis methodologies.
- Employ user feedback for the continuous evolution of OmicSHIELD, ensuring the tool’s expansion and improvement align with the research community’s needs, propelling new research initiatives and discoveries.

— **Objective 3: Unveiling and Validating the Utility of the dsExposome R Package in Exposome Data Analysis**

General objective

Create and exemplify the effectiveness of ”dsExposome”, a proficient R package devised for analyzing exposome data within a multi-centre study framework utilizing the DataSHIELD infrastructure.

Specific objective

- Facilitate Exploratory Exposome-Wide Association Studies (ExWAS) using exposome data via the ”dsExposome” tool, highlighting its capability for preliminary analyses of the exposome.
- Replicate real-world ExWAS of the HELIX study employing ”dsExposome” in a multi-centre study setting, showcasing the tool’s compatibility and potential utility for exposome researchers.
- Address confounding factors in exposome analysis by reproducing the adjustments utilized in the HELIX study through ”dsExposome”, confirming the robustness of its advanced analytical methods in ExWAS.
- Merge advanced analytics facilitated by ”dsExposome” with the DataSHIELD infrastructure, challenging traditional meta-analysis methods by promoting pooled analyses in multi-centre studies.

— **Objective 4: Bridging the Gap for User-friendly Federated Non-disclosive Analysis**

General objective

Unveil ShinyDataSHIELD, an innovative tool engineered to simplify and broaden the application of DataSHIELD for federated non-disclosive analysis, thus opening the doors of this potent technology to both novice and experienced DataSHIELD users.

Specific objective

- Enhance accessibility to DataSHIELD for researchers lacking advanced R skills by offering an intuitive user interface and step-by-step procedures via ShinyDataSHIELD.
- Expedite hypothesis prototyping and analysis for seasoned DataSHIELD users by streamlining complex analysis pipelines through ShinyDataSHIELD.
- Ensure ease of upgrade and maintenance through a modular and functionally comprehensive structure of ShinyDataSHIELD, keeping pace with evolving research demands.
- Promote a swift grasp and utilization of DataSHIELD through user-friendly navigation and guidance provided by ShinyDataSHIELD, augmenting the acceptance of non-disclosive analysis methods in the research community.

— **Objective 5: Employing DataSHIELD for Secure, Collaborative Multi-center Analyses on Real-world Data**

General objective

Employing DataSHIELD on real-world data opens the avenue for secure, collaborative multi-center analyses, providing a platform to unearth meaningful insights while upholding the integrity of data privacy and confidentiality.

Specific objective

- Foster enhanced collaboration among researchers in multi-center studies by deploying DataSHIELD, ensuring robust and comprehensive analyses whilst safeguarding data privacy and confidentiality.
- Unlock previously undetected patterns and relationships within real-world data by leveraging the capabilities of DataSHIELD.
- Amplify reproducibility and transparency in research through the utilization of DataSHIELD, thereby elevating the overall quality and reliability of scientific findings.

3 Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD

3.1 Disclaimer

From the work presented along the next section, my contributions are centered on the development of exposome and genome data integration, which is linked with the development of the respective analysis packages presented afterwards. Moreover, I also contributed by developing further integrations with GA4GH genome and clinical databases, such as EGA.

3.2 Introduction

Big Data brings new opportunities to biomedicine and challenges to data scientists. These challenges require new computational and statistical paradigms to deal with important principles of data management and data sharing. The new paradigm should consider: ensuring appropriate levels of security and privacy [198]; the rigorous application of the stringent regulations required by governance frameworks such as GDPR in Europe (<https://gdpr-info.eu/>) and similar regulatory mechanisms across North America and elsewhere; and a considered choice between central data warehousing and the distributed (federated) analysis of data that remain with their custodian [199, 200].

Historically there has tended to be a focus on warehousing because of the technical challenges of federation [201], and funder requirements to physically share public data [202]. This requires data custodians to physically transfer data to a central location to make them accessible to analytic users. However, the potential benefits of remote and federated approaches to analysing data are now widely recognised [199, 200]. Most fundamentally, the physical data then remain under the control of their custodian with limited access. This can offer major benefits in terms of: making it easier to meet ethics and governance requirements; enhanced flexibility to refine and rerun analyses quickly without waiting for an analyst at each institution to follow an updated analysis plan; allowing datasets to be updated quickly without needing to be resent to a central location [199, 203].

Anticipating the future growth of federated analysis, the DataSHIELD (www.datashield.ac.uk) and OBiBa (www.obiba.org) projects are now 10 years into the joint development of an open-source analytics platform that enables and simplifies flexible but efficient federated analysis [199, 204, 205]. DataSHIELD is linked to an Opal database designed for data management, harmonization and dissemination [206]. In addition, it actively constrains the risk of information disclosure: i.e. the risk that a data analyst is able—accidentally or deliberately—to infer individual level data [199, 207]. The DataSHIELD platform has a growing user community and a central role in the analytic strategies of several large research consortia focussed primarily on the federated analysis of large cohort studies, particularly in Europe and Canada. These include BioSHaRE [205], EUCAN-connect [208], LifeCycle [209], ATHLETE [210] and InterConnect [211]. This article describes a radical extension to the DataSHIELD/Opal platform—the “Resources” architecture. This allows DataSHIELD to be used in a range of new settings which include the analysis of high-volume data such as omics, geospatial or neuroimaging among others.

Despite the growing confidence users have been placing in DataSHIELD to perform privacy-protected analyses, there have, to date, been several serious limitations: (1) Difficulty in applying DataSHIELD federated analytics across the wide range of data formats, and data sources used in ‘omics-based research. (2) Challenges to the efficient porting of high-volume distributed data into the analytic (R) environments on the remote data servers; single large tranches can overwhelm the handling capacity of the system and regular block-by-block refreshments of the analytic data can be impractically slow. (3) To date DataSHIELD has primarily been applied to research settings (typically large cohort studies) where the emphasis has been on the provision of robust disclosure control for analysing sensitive data. But in big data analyses the emphasis is more typically on fast, efficient analysis and data governance often requires a relatively basic level of disclosure control. For example, many consortium-based ‘omics projects would like data to remain with their usual generators/custodians—i.e. a federated approach avoiding the physical transfer of data to users—and for analysts to be unable to see, copy, capture or otherwise infer, those individual-level data. Crucially, if limitations “1” and “2” could be circumvented it would then be straightforward to ensure that this basic level of disclosure control is embedded into all new functions as they are developed and implemented. This would greatly accelerate the development of new functionality making it realistic to consider rapid implementation of Bioconductor [212], Neuroconductor [213] or R packages designed for big data analyses into DataSHIELD.

In one stroke, these objectives have all been realised with the development and implementation of the new “Resources” architecture in DataSHIELD/Opal.

This article describes this new facility, illustrates its value with real world examples and considers the exciting implications it has for the future development of the DataSHIELD platform and for the wider adoption of federated approaches to big data analysis. To help researchers use this framework, we present an online book (https://isglobal-brge.github.io/resource_bookdown/) covering installation, sources of help, specialized topics pertaining to specific aspects of privacy-protecting analysis and complete workflows analysing various examples from biomedical, omics and geospatial settings. The packages developed are available through CRAN or Github repositories under open source licenses (GPL3 or MIT).

3.3 Design and implementation

3.3.1 The resources architecture

When analysing data, it is normal to deal with a very wide variety of data formats, data storage systems and programmatic interfaces. The purpose of the work we describe is not to define a new data format or storage system. Instead we have aimed to describe how data can be accessed, in a formal but generic way, to simplify the integration of various data or computation resources in a statistical analysis program. By actively embracing the variety of data formats and computation systems we seek to guarantee that the right tool can always be used for the type and the volume of data that are being considered.

We define a “resource” to be a description of how to access either: (1) data stored and formatted in a particular way or (2) a computation service. Therefore, the descriptors for the resource will contain the following elements: (1) the location of the data or of the computation services, (2) the data format (if this information cannot be inferred from the location property) or the format of the function call to the computation service, (3) the access credentials (if some apply).

Once a resource has been formally defined, it becomes possible to build a programmatic connection object that will make use of the data or computation services described. This resource description is not bound to a specific programmatic language (the URL property is a web standard, other properties are simple strings) and does not enforce the use of a specific software application for building, storing and interpreting a resource object. Section 7.8 in our online book describe some examples of resources available in a demonstration Opal repository.

The data format refers to the intrinsic structure of the data. A very common family of data formats is the tabular format which is made of rows (entities, records, observations etc.) and columns (variables, fields, vectors etc.). Examples of tabular formats are the delimiter-separated values formats (CSV, TSV etc.), the spreadsheet data formats (Microsoft Excel, LibreOffice Calc, Google Sheets etc.), some proprietary statistical software data formats (SPSS, SAS, Stata etc.), the database tables that can be stored in structured database management systems that are row-oriented (MySQL, MariaDB, PostgreSQL, Oracle, SQLite etc.) or column-oriented (Apache Cassandra, Apache Parquet, MariaDB ColumnStore, BigTable etc.), or in semi-structured database management systems such as the document-oriented databases (MongoDB, Redis, CouchDB, Elasticsearch etc.). When the data model is highly structured or particularly complex (data types and objects relationships), a domain-specific data format is sometimes designed to handle the complexity. This then enables statistical analysis and data retrieval to be executed as efficiently as possible. Examples of domain-specific data formats are regularly encountered in the genomic or geospatial fields of research. A data format can also include additional features such as data compression, encoding or encryption. Each data format requires an appropriate reader software library or application to extract the information or perform data aggregation or filtering operations.

Data storage can simply be realised via a file that can be accessed directly from the host’s file system or downloaded from a remote location. More advanced data storage systems can involve software applications that expose an interface to query, extract or analyse the data. These applications can make use of a standard programming interface (e.g. SQL) or expose specific web services (e.g. based on the HTTP communication protocol) or provide a software library (in different programming languages) to access the data.

These different ways of accessing the data are not mutually exclusive. In some cases when the (individual-level) micro-data cannot be extracted, computation services returning aggregated/summary statistics may be provided. The data storage system can also apply security rules, requiring authentication and proper authorisations to access or analyse the data.

The resource location description will make use of the web standard “Uniform Resource Identifier (URI): Generic Syntax”. [214] More specifically, the Uniform Resource Locator (URL) specification is what we need for defining the location of the data or computation resource: the term Uniform allows the resource to be described in the same way, independently of its type, location and usage context; the use of the term Resource does not limit the scope of what might be a “resource”, e.g. a document, a service, a collection of resources, or even abstract concepts (operations, relationships, etc.); the term Locator both identifies the resource and provides a means of locating it by describing its access mechanism (e.g. the network location). The URL syntax is composed of several parts: (1) a scheme, that describes how to access the resource, e.g. the communication protocols “https” (secured HTTP communication), “ssh” (secured shell, for issuing commands on a remote server), or “s3” (for accessing Amazon Web Service S3 file store services), (2) an authority (optional), e.g. a server name address, (3) a path that identifies/locates the resource in a hierarchical way and that can be altered by query parameters.

The resource’s data format might be inferred from the path component of the URL; for example, by using the file name suffix. However, it is not always possible to identify the data format because the path could make sense only for the data storage system, for example when a file store designates a document using an obfuscated string identifier or when a text-based data format is compressed as a zip archive. The format property can provide this information. Although the authority part of the URL can contain user information (such as the username and password), it is discouraged to use this capability for security considerations. The resource’s credentials property will be used instead, and will be composed of an identifier sub-property and a secret sub-property, which can be used for authenticating with a username/password, an access token, a key pair (private and public keys), or any other credentials encoded string. The advantage of separating the credentials property from the resource location property is that a user with limited permissions could have access to the resource’s location information while the credentials are kept secret.

3.3.2 The *resourcer* R package

The *resourcer* package is an R implementation of the data and computation resources description and connection. It reuses many existing R packages for reading various data formats and connecting to external data storage or computation servers. The *resourcer* package’s role is to interpret a resource description object to build the appropriate resource connection object. Because the scope of resources is very wide, the *resourcer* package provides a framework for dynamically extending the interpretation capabilities to new types of resources. Next, we describe the key issues to deal with resources within the R environment. Further details and examples are available in Section 7.8 in our online book.

3.3.2.1 Resources R Implementation

The resource class is a simple R structure that holds the properties of the resource described in the previous section: URL, format, identity and secret. To simplify the designation of a resource, an additional name attribute is defined. This identifier is optional and is not necessarily unique. The ResourceClient is a key R6 class, which wraps a resource object and defines operations that can be performed on it. The *resourcer* package has built-in support for the following use cases:

- Data file resource whose location is defined by the resource’s URL and that can be downloaded in a temporary folder to be read. The file locations that are supported by default are: (1) the local file system (obviously with a no-op download), (2) an HTTP(S) connection, optionally providing basic authentication based on the resource’s credentials, (3) the MongoDB GridFS file store, (4) Opal’s file store and (5) a remote SSH server. For the reading part, the *resourcer* package uses some tidyverse R packages (such as *haven* for the SAS, SPSS and Stata data formats, *readr* for delimited data formats and *readxl* for Excel data formats) or can load an R object based on its class name specified in the resource’s data format property.

- SQL database resource which has a connector based on R's interface for databases (DBI). The resource's URL indicates which database connector is to be used. The SQL databases supported by default are MySQL/MariaDB and PostgreSQL, and some "big data" databases exposing a SQL interface such as PrestoDB and Apache Spark. The resourcer package can be extended to new DBI-compatible databases.
- NoSQL database resource which can be read using connectors from the nodbi package. Only the MongoDB database is supported for now.
- Command-based computation resource. The resourcer package can handle commands to be executed in the local shell and on a remote server through a secure shell (SSH) connection.

3.3.2.2 Interacting with R Resources

As the ResourceClient is simply a connector to a resource, its utility is enhanced by a range of data conversion functions that are defined by default:

- R data.frame, which is the most common representation of tabular data in R. A data frame, as defined in R base, is an object stored in memory that may be not suitable for large to big datasets.
- dplyr tbl, which is another representation of tabular data provided by the dplyr package that nicely integrates with the R interface for databases: filtering, mutation and aggregation operations can be delegated to the underlying SQL database, reducing the R memory and computation footprint. Useful functions are also provided to perform joint operations on relational datasets.

In the case when the resource is a R object, the R data file ResourceClient offers the ability to get the internal raw data object. Then complex data structures, optimized for a specific research domain can be accessed with the most appropriate tools.

When the resource is a computation service provider, the interaction with the resource client will consist of issuing commands/requests with parameters and getting the result from it either as a response object or as a file to be downloaded.

The purpose of the resourcer package is definitely not to substitute itself for the underlying library; rather it is a general approach that facilitates the access to the data and service resources in the most specific way.

3.3.2.3 Extending R Resources

Thanks to its modular and dynamic architecture, the *resourcer* package can easily be extended to:

- new file locations (see for instance the *s3.resourcer* R package which connects Amazon Web Service S3 file stores),
- new file readers (see Section 8 in our online book to see how VCF files are read in the *dsOmics* R package),
- new DBI-compatible databases. For instance, using *bigquery* R package, it would be easy to implement access to a resource stored in a Google's BigQuery database.
- new domain specific applications which would expose data extraction and/or analysis services. The only requirement is that an R connection API exists for the considered resource application.

3.3.2.4 Resources with DataSHIELD/Opal

DataSHIELD infrastructure is a software solution that allows simultaneous co-analysis of multiple data sets stored on different servers without the need to physically pool data or disclose sensitive information. DataSHIELD uses Opal servers to perform such analyses.

At a high level DataSHIELD is set up as a client-server model, each server housing the data for its corresponding study. A request is made from the client to run specific functions on the data held in the remote

servers and that is where the required analyses are actually performed. Non-sensitive and pre-approved summary statistics are returned from each study server to the client where they can then be combined for an overall analysis. An overview of what a multi-site DataSHIELD architecture would look like is illustrated in fig. 17.

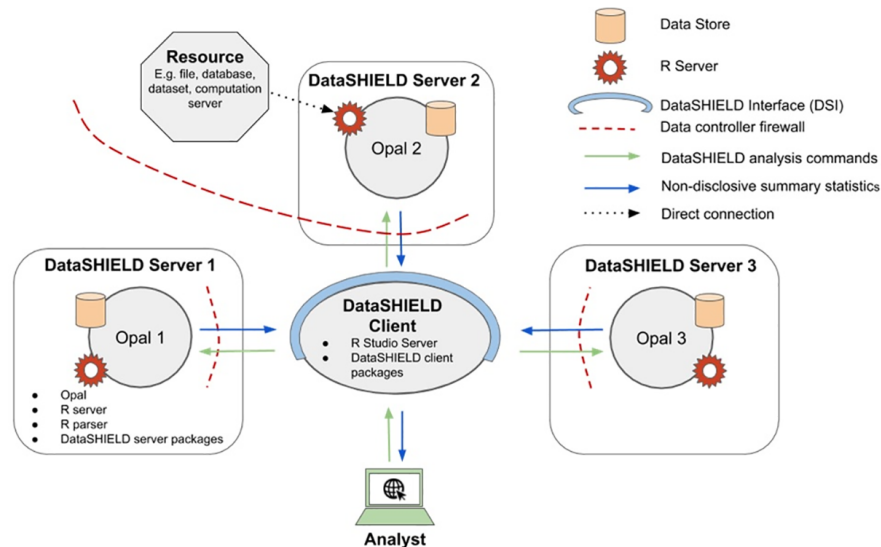


Figure 17: **A schematic diagram of a multi-site DataSHIELD infrastructure.** It includes one central analysis node (the client) and three data nodes (the servers).

The current limitation of this infrastructure is that the data assignment operation consists of extracting the dataset from the data repository (the primary storage can be either a SQL database or a MongoDB database) and pushing it in the R server’s memory as a `data.frame` object. This data assignment process consumes time, physical memory and applies only to datasets in tabular format. Although acceptable for small datasets (usually less than 10M data points), this infrastructure cannot be used for big data systems, complex data structures and existing computation facilities. In order to overcome these limitations, data access can instead be delegated to the R servers using the concept of resources. The DataSHIELD middleware (Opal) will be responsible for performing the assignment of a resource description as a `ResourceClient` object in the R server (as soon as the relevant permissions have been granted). The connection to data or computation resources is then readily usable by the DataSHIELD user. In terms of data access security, the resource credentials are not visible via the DataSHIELD client node, just like the individual-level data.

After converting a resource to a `data.frame` object, existing DataSHIELD analysis R packages can be used in a backward compatible way. New DataSHIELD R packages, such as `dsOmics` and `dsGeo`, described in this paper, can make use of the power of resources to apply DataSHIELD’s data privacy preserving paradigm to large datasets, complex data structures and external computation facilities.

The integration of the resources concept in Opal consists of (1) dynamically discovering the different types of resources that can be handled by the associated R server, (2) providing an appropriate user-friendly graphical interface (GUI) for managing the resource descriptions and access permissions, and (3) assigning resource objects in the R server on user demand. The latest version of Opal (v 3.0) implements these capabilities, making resources accessible to the DataSHIELD framework.

3.4 Results

The table 2 describes the main features of our proposed framework that are mainly driven by DataSHIELD’s capabilities. The table also shows the main advantages or disadvantages that can be found when using the *resources* in federated data analyses. Different features including scalability, disclosure prevention, deployment and future applications to other biomedical areas than genomics are discussed.

Table 2: Main characteristics of the proposed infrastructure for privacy-protected federated data analyses with big data

Feature	Capabilities / Advantages	Limitations / Disadvantages
Resources	Any data source or computation resource that can be accessed from R is made available in DataSHIELD. These include databases of any kind (SQL, NoSQL, distributed Big Data systems), most of the file formats, domain specific applications accessible through web services or remote commands etc. Applies to any scientific domain.	Resource URL design can be complex when resource options need to be specified. R is the required entry point, which complexifies the use of analysis algorithms in Python for instance.
Scalability	Scales with number of studies. The resources can interact with Apache Spark or Hadoop.	Some advanced statistical techniques may not scale well with number of records. The processing power can sometimes suffer from latency. There is a need for investigation share computation over multiple processors.
R programming language	R is an open-source and heterogenous programming language. Interpreters for available for many operating systems. A wide community support its users.	Skills for other functional and statistical programming languages can be transferred to learn R. However not all of the analysts are familiar with other programming languages, which can raise some barriers of adoption.
Disclosure prevention	Several complementary features that provide privacy preservation are implemented in the DataSHIELD architecture. Just one of these is that data custodians/owners (not analysts) have complete control over a series of optional filters that dictate the potential disclosivity of the analytic output. As one example, this includes the minimum count acceptable in a non-empty cell of a contingency table. When rare observations are critical to analysis, data custodians can actively choose to relax the filters to enable meaningful analysis to be undertaken. One of the implications of the combined effect of several of the privacy preservation features is that analysts can never see, copy or abstract the individual-level data on the primary data servers.	Data protection is shared between existing computer systems and DataSHIELD. An existing computer system that has some poorly implemented secured network access and authentication may threaten the quality of disclosure prevention. In other words, as well as the active privacy preservation built into DataSHIELD software, it is also crucial that the hardware on which everything runs must satisfy good practice for conventional privacy protection.
Graphical User Interface (GUI)	The use of R studio and the specialised data warehouse software have a GUI. Analysis can be integrated with GUI packages, such as R Shiny.	An appropriate use of R Scripts, R notebooks and vignettes is helpful. DataSHIELD users then need to learn how to use these approaches before using DataSHIELD. The use of R Studio can make the learning curve less steep.

Feature	Capabilities / Advantages	Limitations / Disadvantages
Applications to other biomedical areas than genomics	Any other specific data infrastructures available in public repositories such as images (OpenNeuro), transcriptomic or epigenomic data (GEO) can be accessed and analyzed in a distributed and privacy preserving way. Applications outside biomedical, health and social science are also entirely possible.	There are no obvious scientific or academic domains to which DataSHIELD could not in principle be applied. Particularly, if the aim is to facilitate the quantitative analysis of individual level data which are sensitive either because of ethico-legal or information-governance restrictions, or because of their intrinsic intellectual or commercial value.
Cost	As everything is based on open-source freeware, the costs are minimized for any organisations who wishes to adopt DataSHIELD.	While DataSHIELD development is substantively funded through grants, there is nevertheless a need to seek support from the user community (particularly large-scale users) to contribute to the core DataSHIELD provision: user training and support; on-line support materials; bug snagging and error correction; continuous testing of evolving code; preparing and smoothly undertaking new releases.
Time	From the user perspective, DataSHIELD is a time-effective tool as does not require iterative in-person communication between the analyst and the data holders. This is most evident in comparing the time commitment required for a standard consortium-based metaanalysis and that required by a centrally controlled study-level meta-analysis via DataSHIELD. The former requires the analysis centre to ask each study to undertake a series of specified analysis and return results (typically several rounds of analysis and return). In contrast the latter is controlled in real time as if one was working directly with the raw data. This can speed things up by orders of magnitude.	Like any specialised open-source software, DataSHIELD analysts, developers and installer need some time to adapt to the concepts of federated systems and disclosure limitations.
Server sharing	DataSHIELD supports multi-tenancy, to permit multiple users to share a DataSHIELD server. It also permits multiple servers to be deployed if needed.	DataSHIELD requires some specialised data warehouse software such as Opal. At the moment there is no white-paper that formally defines standards for further development in this area.

Feature	Capabilities / Advantages	Limitations / Disadvantages
Documentation	DataSHIELD has a wiki that provides beginners training to any new DataSHIELD analysts and developers. Some more advanced tutorials are available on multimedia contents through YouTube. Documentation for the Opal specialised data warehouse software is available online. The online book of this paper will be continuously being updated with more extensions (e.g. transcriptomics, epigenomics, imaging data, longitudinal data analyses, ...).	DataSHIELD documentation needs to be more supportive to the DataSHIELD developers. An online forum is available for community engagement and support. No advanced statistical techniques are explicitly taught, as it is assumed that any analyst planning to use DataSHIELD would already understand the theory and practice underpinning the analysis that is to be undertaken. However, in practice the DataSHIELD team knows that this is not always true and so some analytic theory is increasingly being presented when it is useful.
Deployment	Software solution packages are available for different hosting systems (includes a container-based option) and the resource concept can adapt itself to the existing hosting infrastructure. It is advised that DataSHIELD should only be deployed in a setting in which all hardware and middleware systems satisfy conventional best practice for data management and privacy protection. Similarly, it is assumed that DataSHIELD will not be used if information governance or other restrictions already prescribe the particular analysis proposed.	It is important that there is at least a minimum baseline level of system administration knowledge on the part of the data owner and proper dimensioning of the hardware (especially when targeting multi-user, computation intensive usage). Nobody should be making sensitive data available for analysis via any mechanism—including DataSHIELD—if they do not have a proper understanding of their data systems or of the governance framework under which analysis is to be enacted.

3.4.1 Available resources extensions

We have extended the resources available at the resourcer package into different settings. These extensions as well as the current resources that can be accessed through the Opal servers are described in table 3. So far, we can get data from different locations (Amazon Web Services, HL7 FHIR or Dremio), read other types of files which are specific in genomic studies (BAM, VCF and PLINK) and access data from other infrastructures such as GA4GH, a federated ecosystem for sharing genomic, clinical data [215] and EGA which is a permanent archive that promotes distribution and sharing of genetic and phenotype data consented for specific approved uses [216].

Table 3: Available resources at the resourcer R package and extensions for genomic data.

Type	Resource	Reference	R package	Use
File reader	R data	https://cran.r-project.org/	resourcer	Any
File reader	Tidy data (.csv,.tsv, txt, ...)	https://www.tidyverse.org/	resourcer	Any
File location	S3 compatible file store	https://min.io/ https://aws.amazon.com/	s3.resourcer	Any

Type	Resource	Reference	R package	Use
Database	SQL	https://www.mysql.com/ https://mariadb.org/ https://www.postgresql.org/ https://prestodb.io/	resourcer	Any
Database and Big Data analytics	SQL	https://spark.apache.org/	resourcer	Any
Database	NoSQL	https://www.mongodb.com/	resourcer	Any
Computation service	SSH	https://en.wikipedia.org/wiki/Secure_Shell	resourcer	Any
Domain specific	HL7 FHIR	http://hl7.org/fhir/	fhir.resourcer	Patient's data
Database	SQL	https://www.dremio.com/	odbc.resourcer	Any
File reader	VCF	https://en.wikipedia.org/wiki/Variant_Call_Format	dsOmics	Genomic
File reader	GDS	https://bioconductor.org/packages/release/bioc/html/gdsfmt.html	dsOmics	Genomic
File reader	BAM	http://samtools.github.io/hts-specs/SAMv1.pdf	dsOmics	Genomic
File reader	Bioconductor infrastructures (ExpressionSet, RangedSummarizedExperiment, MultiAssayExperiment, ...)	http://bioconductor.org/	dsOmics	Genomic
Computation service	PLINK	http://zzz.bwh.harvard.edu/plink/	dsOmics	Genomic
Domain specific	GA4GH	https://www.ga4gh.org/	dsOmics	Genomic and clinical
Domain specific	EGA	https://ega-archive.org/	dsOmics	Genomic and clinical

3.4.2 Real data analyses

We illustrate how to perform privacy-protecting big data analyses using our proposed infrastructure. We have set up an Opal demo site (see Chapter 4 in our bookdown) to illustrate how to perform some basic analyses using DataSHIELD as well as how to deal with different resources for genomic and geographical data. These data are publicly available and can be accessed through DataSHIELD or using the URL available in the Opal site. The genomic example describes how to perform genome-wide association (GWAS). The geographic examples describe how to analyse information about journeys undertaken by specified individuals', the environment through which they travel and whether the journeys have an impact on the individuals' health. Chapter IV in our online book provides users with workflows and case studies for downstream analyses and visualizations.

3.4.2.1 Genomic data analysis

Bioconductor provides core data structures and methods that enable genome-scale analysis of high-throughput data in the context of the rich statistical programming environment offered by the R project [212]. We have created two packages to perform privacy-protecting federated genomic data analysis with DataSHIELD and Bioconductor. The *dsOmics* package contains the functions that are used on the server side where the actual analysis is implemented and which specify the privacy-protecting summary statistics that will be send back to the client, while the *dsOmicsClient* has the functions that are used on the client side, to control the commands that are send to the server side and to combine the received outcomes for pooled analysis applications.

Genomic data can be stored in different formats. Variant Call Format (VCF) and PLINK files [217]. are commonly used in genetic epidemiology studies. In order to deal with this type of data, we have extended the resources available at the *resourcer* package to VCF files. We use the Genomic Data Storage (GDS) format which is designed for large-scale data management of genome-wide variants and can efficiently manage VCF files into the R environment. This extension requires the creation of specific *ResourceClient* and *ResourceResolver* classes. This extension is available in the *dsOmics* package (See Section 8 in the supplementary book). Briefly, the client class uses *snpgdsVCF2GDS* function implemented in *SNPrelate* to coerce the VCF file to a GDS object [218]. Then, the GDS object is loaded into R as an object of class *GdsGenotypeReader* from *GWASTools* package [219]. that facilitates downstream analyses such quality control of SNPs and individual data, population stratification and association analyses using *GENESIS* Bioconductor package [220].

The fig. 18 describes how GWAS may be performed using DataSHIELD client-side functions (i.e *dsOmicsClient* package). Basically, data (genomic and phenotypes/covariates) can be stored in different sites (EGA, GA4GH, https, ssh, AWS S3, local, . . .) and are managed with Opal through the *resourcer* package and their extensions implemented in *dsOmics*. The association analyses involving GWAS are based on fitting different generalized linear models (GLMs) for each SNP that can be performed using a base DataSHIELD function (*ds.glm*). The difference between standard data analysis and that done by DataSHIELD is that the analysis is performed at the location of the data (e.g. the data nodes). No data is transferred from that location, only non-disclosive summary statistics. The set of analytical operations which can be requested to be performed at the location of the data, has been careful constructed to prevent any attempt for direct or inferential disclosure of any individual-level information. The R parser also blocks any form of arguments that are not allowed in DataSHIELD. For analytical operations which could potential yield results which are disclosive, for example if a small amount of data is being analysed, the operation will check if the results match the data protection policies of the location's data governance rules, before returning any results back to the client (e.g. the analysis node). If any of the protection rules are violated, the client does not receive any results but gets study-side messages with information about potential disclosure issues [199].

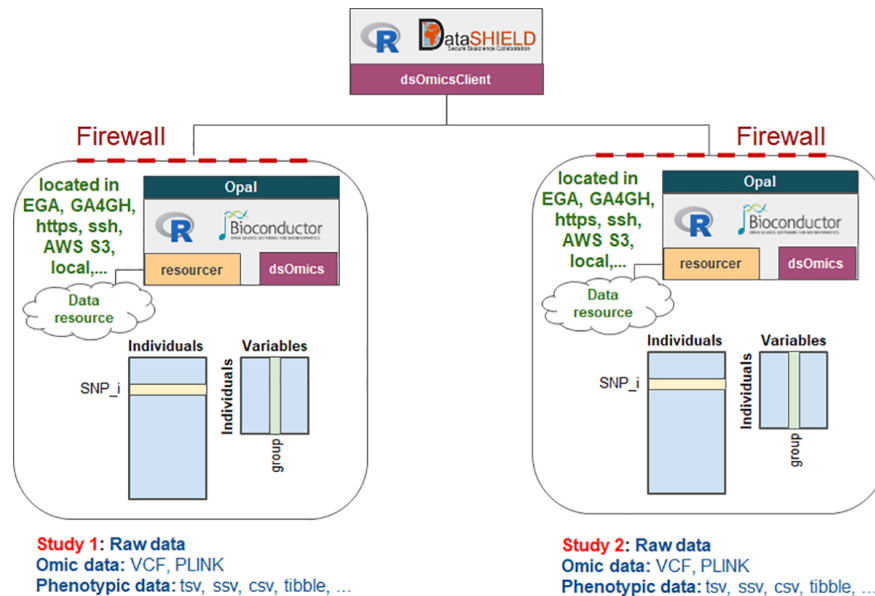


Figure 18: **Scheme of DataSHIELD implementation of genomic data analysis.** The *dsOmics* package contains functions to perform non-disclosive data analyses of resources encoding genomic data that are managed within Opal using the resourcer package. Genomic data normally have two pieces of information, one corresponding to variants (e.g. SNPs) and another for phenotypic data (grouping variable, outcome, covariates, ...). Both can be stored in different resources. BAM/VCF/PLINK for SNPs and text/csv file for phenotypes and covariates. This package should be installed in the Opal server along with their dependences. The package *dsOmicsClient* must be available in the client side and contains functions that allow the interaction between the analysis computer and the servers.

It should be noticed that repeatedly calling `ds.glm` function can be very time consuming when analysing thousands of SNPs which requires multiple iterations of calls over the network between the client and the server. In order to overcome this problem, we also implemented a federated meta-analysis approach that basically runs an independent GWAS at each server and then meta-analyse the results. GDS data at each server are analyzed using GWASTools and GENESIS Bioconductor packages that allows to perform GWAS very quickly. Once the study-specific estimates and standard errors generated by the analyses undertaken on each server have been returned to the client, they can be combined using whatever meta-analysis approaches—and whatever R meta-analysis packages—the user may choose. This methodology has some limitations when data are not properly harmonized (e.g. genotyping in different platforms, different VCF versions, ...). In order to overcome this problem, data format validation can also be performed by the analyst using DataSHIELD functions. In genomics, this can be achieved by first doing imputation and then solving issues concerning genomic strand and file format [221].

GWAS can also be performed using programs that are executed using shell commands. This is the case for PLINK, one of the state-of-the-art programs to run GWAS. Resources also allow the use of secure SSH service to run programs on a remote server accessible through SSH containing data and analysis tools where R is just used for launching the analyses and aggregating results. This feature allows us to create functions to analyze data using specific shell programs. Section 9 in our online book describes how the PLINK program can be used to perform GWAS. In this case, the resource describes that access is given via SSH, the credentials required to connect, and the commands that can be run (of which one is plink).

We would like to emphasize that with DataSHIELD, analysis is performed at the location of the data. The data provider has full control over what information is transferred from their location to the location of the analyst by setting filters for a number of disclosure traps. This means that the results returned to the analyst can be carefully created to be non-disclosive, and match the policies of the data provider's data governance rules.

3.4.2.2 Geographic Information System (GIS) and spatial analysis

The R packages *rgdal*, *rgeos* and *sp* provide core data structures and methods that enable analysis of geospatial data in the context of the rich statistical programming environment offered by the R project. We have created two packages to perform privacy-protecting federated GIS data analysis with DataSHIELD and these packages. The *dsGeo* package contains the functions used on the server side to assure privacy-protecting analyses, while *dsGeoClient* has functions that command the data analyses from the client side and enable integration across studies.

The *resourcer* package allows large geospatial datasets to be handled. These include data derived from standard storage systems, such as relational databases, making use of existing *sp* data structures such as *SpatialPoints* and *SpatialLinesDataFrame* among others. These types of data are the core of geospatial analysis in R, allowing users to work with geometries and their descriptive attributes. For example, we might want to know that a GPS trace corresponds to someone who is 45 years old, or that a region defined by a polygon has a particular air pollution level. As described in the methods section, resources can be extended to any type of data that can be managed within R. Here we describe how to extend the resources to the case of analysing Geographic Positioning System (GPS) traces and other geolocation data, combined with phenotypic data.

Section 13 in our online book illustrates how to perform a realistic analysis of GPS traces and geolocation data. In this example, building on the work of Burgoine et al., [222] we consider daily commutes captured as GPS traces by 810 individuals in the eastern suburbs of London. We also have data on the location of 6100 fast food or takeaway outlets in the same area. Further data are available, relating to each individual's Body Mass Index (BMI), age, sex, total household income, highest educational qualification and smoking status. These data therefore allow us to test the association between exposure to takeaway food on a commute and on individual's BMI. We illustrate how the tools available in the *dsGeo* package allow this question to be addressed.

We created three resources in Opal that contain the GPS journey data (*SpatialLinesDataFrame*), the locations of the food outlets (*SpatialPointsDataFrame*) and the phenotypic data (*data.frame*). These data can be manipulated to give a measure of exposure to the food outlets for each individual. Our package allows us to transform the point data denoting food outlets into buffer regions surrounding each point. The idea is that if an individual's GPS trace falls within this buffer then we can say that the individual is 'exposed' to that food outlet. Thus, we find the intersection of each individual's trace with each buffered region to obtain a vector of the intersecting buffers for each individual. This is further processed into a count of 'exposures' per individual. Finally we can run GLMs using the DataSHIELD base function `ds.glm()` to test the association between BMI and food outlet exposure, correcting for potential confounding factors such as income.

3.5 Future perspectives

3.5.1 DataSHIELD

The development of the new "resources" component of the OBiBa middle-ware that underpins the DataSHIELD platform has profound implications for the future of the overall DataSHIELD project. It has greatly relaxed constraints on the source, format and volume of the data that can be ingested into a DataSHIELD session while continuing to provide full flexibility in terms of the capacity to enact active disclosure control at a level appropriately tailored to the context of the particular analytic problem at hand. This has already allowed us to embark on exploring extensions of functionality to encompass some of the most widely used classes of contemporary research data, for example, Omic data and images. In tandem with recent advances we have made in learning how to simplify the extension of functionality in any field provided that a relevant R-package already exists, DataSHIELD is now on the cusp of being able to provide a generic easily usable and simply extendable approach to truly federated analysis across many science and technology domains. This will have a myriad of applications in academic research, commercial settings and health & social care systems. The ability to finely tune disclosure controls in a manner that can only be modified by the data custodian will make this approach particularly attractive for anybody wishing to work with data that are sensitive. This not only includes human (or other) data subject to the appropriately stringent requirements

of contemporary law, ethics and broader frameworks for data governance, but also encompasses data that are sensitive for reasons of intellectual property investment or commercial value. However, if disclosure controls are set to be very permissive or off, the convenience, flexibility and potential extendibility of the system could ultimately make this an approach of choice for any federated analysis even if the data are not especially sensitive.

In light of the development of resources, it is envisaged that in the future there will be at least three flavours of DataSHIELD that will vary in their convenience of use, flexibility and ease of extension of functionality: (1) **All disclosure checks set to off**. This will permit maximum flexibility for analysing non-sensitive federated data and easily extendable by adding new functionality; (2) **Disclosure checks on but minimal**—e.g. preventing users from seeing or copying the individual level data but avoiding restrictions on the analyses themselves. This flavour is likely to be most useful when data from a large research platform cannot physically be shared but the analysis required is based on data objects that are fundamentally privacy-protecting such as Omics data or images based on internal scanning; (3) **Full disclosure checks**; this will be equivalent to the default situation that applies now.

Now that it is possible for DataSHIELD to work much more readily with large/big datasets, we anticipate that DataSHIELD and the *resourcer* R package will offer functions based on the tidyverse as well as base-R. For example, this will include the option of using *dplyr* R package for operating on tabular datasets (see **Interacting with R Resources** section) and on tibbles as well as standard R data-frame objects. A resources integration improvement might therefore be to use the *dplyr* API for delegating as much as possible data filtering and mutating to the underlying data storage system (e.g. databases exposing a SQL query interface). This extension is already being explored.

3.5.2 Parallel computing

Working effectively with large data may also require programming practices that match the available computer hardware infrastructure, both processing and storage. R is efficient when operating on vectors or arrays, so a pattern used by high-performing and scalable algorithms is to split the data into manageable chunks and to iterate over them. Chunks can be evaluated in parallel to gain speed. There are several R packages that can be used to this end (parallel, foreach, ...) as well as *BiocParallel* Bioconductor package that facilitates parallel evaluation across different computing environments while allowing users from having to configure the technicalities. DataSHIELD analysis is parallelized by design (i.e. each server is working independently of the others). Therefore, for a server instance, the best approach is to use data structures and analysis tools that perform computations efficiently. This is the strongest point of the resources as we have done with the integration of *dplyr* and Bioconductor packages as well as those that can be implemented using, for instance, *sparklyr* [223].

3.5.3 Omic and geographical data

We have provided some of the functionalities offered by Bioconductor and R packages in the DataSHIELD context that allow to analyse genomic and geographical data. These packages are extensive and more work is needed to repackage a more complete range of operations available in a privacy-protecting way. For instance, *ds.Omics* can be easily be extended to other omic data analyses such as differential gene expression analysis of methylation data analyses using the same strategy as the one used for genomic data. Visualisation can add great value and will be covered when being implemented in a privacy-protecting way. The *dsBaseClient* and *dsBase* packages already contain functions for privacy-protecting plots such as heatmaps and scatter plots. The privacy of data is protected in these cases by effectively blurring the data or by removing outlying points. These techniques could be adapted, for instance, to allow geospatial data to be visualised in a privacy-protecting way.

3.5.4 Data cataloging

The next step of the resources integration in Opal is to make their meta-data findable to a researcher: exposing the data dictionaries, annotated with taxonomy or ontological terms, would benefit the research community when looking for datasets for a research question. OBiBa software application suite provides

both Opal, the data repository (or data integration system, using the resources) and Mica [206], the data web portal application. Mica operates by extracting from Opal the dataset dictionaries to build a searchable data catalogue, with basic summary statistics and by allowing the submission of data access requests. Resources registered in Opal should be made visible from Mica as well.

3.5.5 Other applications

We would like to highlight that there are dozens of disciplines other than genomics and geospatial that could also benefit from our infrastructure. For instance, extending the resources to other settings such as neuroimaging by using libraries from Neuroconductor, a similar project to Bioconductor for computational imaging, would be an important advance in that field since data confidentiality may also be an issue. Also, it is worth noting that one of the main advantages of using the resources is that we do not need to move data from their original repositories which can present a serious problem when dealing with neuroimaging data [224]. Another area that can readily benefit from our new framework is artificial intelligence. Big data and machine learning have applied innovatively many advanced statistical methodologies such as deep learning which is driving the creation of new and innovative clinical diagnostic applications among others [225]. The current trend is to include machine learning algorithms within Cloud capacities in different biomedical problems [226, 227, 228]. Our framework can interface with “Apache Spark”, a fast and general engine for big data processing [229], through the sparklyr R package that will allow the use of different machine learning algorithms for big data.

References

- [198] Luca Bonomi, Yingxiang Huang, and Lucila Ohno-Machado. “Privacy challenges and research opportunities for genomic data sharing”. In: *Nature genetics* 52.7 (2020), pp. 646–654.
- [199] Amadou Gaye et al. “DataSHIELD: taking the analysis to the data, not the data to the analysis”. In: *International journal of epidemiology* 43.6 (2014), pp. 1929–1944.
- [200] Google AI Blog. “Federated analytics: Collaborative data science without data collection”. In: <https://ai.googleblog.com/2020/03/federated-analytics-collaborative-data.html> (2020).
- [201] aridhia. “Is Federated Analysis the Way Forward for Genomics? — Trusted Digital Research Environment”. In: <https://www.aridhia.com/blog/is-federated-analysis-the-way-forward-for-genomics/> (2015).
- [202] Mark Walport and Paul Brest. “Sharing research data to improve public health”. In: *The Lancet* 377.9765 (2011), pp. 537–539.
- [203] Paul R Burton et al. “Data Safe Havens in health research and healthcare”. In: *Bioinformatics* 31.20 (2015), pp. 3241–3248.
- [204] Michael Wolfson et al. “DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data”. In: *International journal of epidemiology* 39.5 (2010), pp. 1372–1382.
- [205] Dany Doiron et al. “Data harmonization and federated analysis of population-based studies: the BioSHaRE project”. In: *Emerging themes in epidemiology* 10.1 (2013), pp. 1–8.
- [206] Dany Doiron et al. “Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination”. In: *International journal of epidemiology* 46.5 (2017), pp. 1372–1378.
- [207] Hampapuram K Ramapriyan. “NASA EOSDIS Data Identifiers: Approach and System”. In: (2017).
- [208] EUCAN Connect. In: <https://www.eucanconnect.eu/> (2020).
- [209] Home—LifeCycle. In: <https://lifecycle-project.eu/> (2020).
- [210] Martine Vrijheid et al. “Advancing tools for human early lifecourse exposome research and translation (ATHLETE): Project overview”. In: *Environmental Epidemiology* 5.5 (2021).
- [211] Interconnect Project. In: <https://interconnectproject.eu/> (2020).
- [212] Wolfgang Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. In: *Nature methods* 12.2 (2015), pp. 115–121.
- [213] John Muschelli et al. “Neuroconductor: an R platform for medical imaging analysis”. In: *Biostatistics* 20.2 (2019), pp. 218–239.
- [214] Larry Masinter, Tim Berners-Lee, and Roy T Fielding. “Uniform resource identifier (URI): Generic syntax”. In: *Network Working Group: Fremont, CA, USA* (2005).
- [215] Global Alliance for Genomics and Health*. “A federated ecosystem for sharing genomic, clinical data”. In: *Science* 352.6291 (2016), pp. 1278–1280.
- [216] Ilkka Lappalainen et al. “The European Genome-phenome Archive of human data consented for biomedical research”. In: *Nature genetics* 47.7 (2015), pp. 692–695.
- [217] S Purcell et al. *PLINK: Whole genome data analysis toolset*. 2013.
- [218] Xiuwen Zheng et al. “A high-performance computing toolset for relatedness and principal component analysis of SNP data”. In: *Bioinformatics* 28.24 (2012), pp. 3326–3328.
- [219] Stephanie M Gogarten et al. “GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies”. In: *Bioinformatics* 28.24 (2012), pp. 3329–3331.
- [220] Stephanie M Gogarten et al. “Genetic association testing using the GENESIS R/Bioconductor package”. In: *Bioinformatics* 35.24 (2019), pp. 5346–5348.
- [221] Patrick Deelen et al. “Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration”. In: *BMC research notes* 7 (2014), pp. 1–4.

-
- [222] Thomas Burgoine et al. “Associations between exposure to takeaway food outlets, takeaway food consumption, and body weight in Cambridgeshire, UK: population based, cross sectional study”. In: *Bmj* 348 (2014).
- [223] sparklyr. In: <https://spark.rstudio.com/> (2023).
- [224] Stephen M Smith and Thomas E Nichols. “Statistical challenges in “big data” human neuroimaging”. In: *Neuron* 97.2 (2018), pp. 263–268.
- [225] Jie Xu, Kanmin Xue, and Kang Zhang. “Current status and future trends of clinical diagnoses via image-based deep learning”. In: *Theranostics* 9.25 (2019), p. 7556.
- [226] Mani Abedini et al. “A cloud-based infrastructure for feedback-driven training and image recognition”. In: *MEDINFO 2015: eHealth-enabled Health*. IOS Press, 2015, pp. 691–695.
- [227] Peipei Ping et al. “Biomedical informatics on the cloud: a treasure hunt for advancing cardiovascular medicine”. In: *Circulation research* 122.9 (2018), pp. 1290–1301.
- [228] Celio de Sousa et al. “Cloud-computing and machine learning in support of country-level land cover and ecosystem extent mapping in Liberia and Gabon”. In: *PLoS One* 15.1 (2020), e0227438.
- [229] Runxin Guo et al. “Bioinformatics applications on apache spark”. In: *GigaScience* 7.8 (2018), giy098.

**4 OmicSHIELD: Federated
privacy-protected meta- and
mega-omic data analysis in
multi-centre studies with a fully open
source analytic platform**

4.1 Introduction

Contemporary data analytics in health and biological sciences include a central focus on the analysis and interpretation of high volume ‘omics data’ (genomic, epigenomic, or metabolomic data). An important requirement for fully exploiting the potential of such data is to make large amounts of clinical, epidemiological and omic information accessible and interoperable to researchers. This can be achieved through data sharing. Historically, data-sharing has been based on central warehousing; this requires data generators to physically transfer data or summary statistics to make them accessible to analytic users. This approach has been adopted, for instance, by the vast majority of consortia devoted to the analysis of genomic data. Under this setting, each data provider runs their own genome-wide association studies (GWAS) independently and shared summary statistics are meta-analyzed by one or two data analysts [230]. Alternatively, a recent and increasingly used analytic trend known as federated analysis (FA) permits analysis of multiple decentralized datasets without accessing disclosive or individual-level information (i.e., migrating the analysis to the data) [231]. Motivated by the delicate nature of genetic and health data and the ethical and legal issues behind sharing this kind of information, the potential benefits of FA are being increasingly recognized widely [232, 233, 234].

Meta-analysis is widely adopted for the combination of GWAS [235] and is also being used in differential gene expression and epigenome wide association studies (EWAS) by combining results from different populations [236]. Individual participant data (IPD) meta-analysis is an increasingly popular tool used as an alternative to traditional aggregate data meta-analysis, especially as it avoids reliance on published results and provides an opportunity to investigate individual-level interactions, stratified models or adjusting for other covariates. Being capable of performing both IPD meta-analyses and mega-analyses (i.e. pooled analyses) in a federated framework would be a cutting-edge advance for the biomedical field, allowing, among other possibilities, to choose the best and most convenient approach to be applied depending on data characteristics and designs. Unfortunately, most currently available FA systems for omics data are only intended to perform pooled analyses, arguing that this approach substantially increases statistical power [237]. However, pooled analysis in a multi-cohort setting is not recommended when data are heterogeneous among cohorts or when data are not properly harmonized (e.g. gene expression normalized using different methods, or GWAS data maintained on different platforms) as substantive heterogeneity in the nature of the data can lead to biased results. This includes “confounding by study” which can be very severe when an outcome and an explanatory covariate vary in tandem (or in a reciprocal manner) across study populations [238].

Omic FA has a key constraint since it includes sensitive information. It requires ensuring appropriate levels of security and privacy and the judicious application of the stringent regulations implicit to contemporary governance frameworks such as the General Data Protection Regulation (GDPR) in Europe. In order to address this important issue, different privacy-protecting techniques such as federated learning (FL), differential privacy (DP), homomorphic encryption (HE), and secure multi-party computation (SMPC) have been developed, some or all of which may be adopted [239]. Furthermore, algorithms developed by researchers could potentially be used alongside genotype-phenotype associations from genetic association studies by an attacker to predict genotypes and phenotypes of target individuals based on genome information shared by individuals or their relatives [240]. A secure FA platform should thus have solutions to minimise the risk of potential attacks.

The burden of data sharing on multi-center studies that will deal with omic data has enormously increased in the last few years. These include large projects such as ORCHESTRA, MIRACUM, unCoVer, LifeCycle, HELIX and ATHLETE [241, 242, 243, 244, 245, 246] among many others. Having software solutions to FA in omic studies using privacy-protected techniques is therefore an urgent need. In the omic setting, infrastructures for federated networks, FA of GWAS (FAHME [237] and sPLINK [247]) and transcriptomic data (Flimma [248]) have been proposed. These existing tools have important limitations including that they may require their own data infrastructures and different programming languages – some of which are not open source, hence making the implementation of new features difficult. Another important limitation of existing solutions is that downstream analyses (e.g. data visualization, post-omic data analyses) are poorly integrated into analytical pipelines.

In this paper, we introduce OmicSHIELD, a software analytic platform for omic multi-center studies that

overcomes these limitations. Our platform is based on DataSHIELD which is a software created to allow analysis of data at individual-level using disclosure-preventing methods that address ethical-legal restrictions surrounding confidentiality [238, 249]. Key aspects of DataSHIELD include: a) client-server architecture (“taking the analysis to the data”), b) analytical methods for FA including both pooled and meta-analyses, and (c) tailored multi-layer disclosure controls (the bottom line being that the analyst cannot see, copy or extract individual level data held by individual studies) [250]. DataSHIELD has been used in different multi-consortia projects, details on how they leveraged DataSHIELD to guarantee confidentiality of their data can be found elsewhere [251, 252, 253]. OmicSHIELD has also implemented protection layers to prevent malicious attacks that have the aim of recovering individual-level data through omics data analyses. These include different privacy-preserving methods, filters to avoid getting disclosive information (e.g. related to low allele frequencies).

OmicSHIELD covers analytical techniques for transcriptomics, epigenomics and genomics omics data. It allows both pooled analyses and IPD meta-analyses. The analysis are available for horizontally partitioned data, this approach is commonly used when different organizations or research centers hold data on distinct individuals or groups, but share the same variables for each record. A key feature is that DataSHIELD is open-source, written in R and licensed under the GPL, thus facilitating downstream analyses within a single pipeline by interacting with other programming languages (e.g. Python) and with other R or Bioconductor packages. Data warehousing is based on Opal which is integrated within DataSHIELD, thus offering a complete software solution, more information on Supplementary data 2. Another huge advantage of using this approach is that DataSHIELD has implemented state-of-the-art methods to perform the standard statistical analyses applied in different disciplines, including biomedicine, epidemiology and the social sciences, in a non-disclosive manner. This facilitates, for instance, performing federated descriptive analyses before commencing omic studies using the same platform. OmicSHIELD is based upon our recent development, the “resources” architecture which is a new DataSHIELD infrastructure that allows: 1) the use of large data in their original repositories; 2) working with original data formats (e.g. PLINK, VCF, ExpressionSet, Range-SummarizedExperiments); 3) interactions with other programming languages (including shell commands) and softwares (R, Neuroconductor, Bioconductor, Python); and 4) interfacing with “Apache Spark”, a fast and general purpose analytical engine for big data and deep learning [254]. Furthermore, OmicSHIELD incorporates both disclosure controls and differential privacy approaches to assure privacy-preserving data analyses. Therefore, our approach has the potential to fulfil the stringent requirements made by data controllers (e.g. hospitals) often hinder multicentric medical studies, since differentially private learning has been positioned as one of the preferred methods for GDPR-compliant recommender systems [255].

To help users leverage these frameworks, we present an online book available at <https://isglobal-brge.github.io/OmicSHIELD/>. It covers installation, sources of help, and complete workflows illustrating examples of omics data analyses using freely available datasets. It also includes material describing different use cases corresponding to real world data applications from different existing projects. Data used in the examples are fully available at two Opal servers which readers can access as DataSHIELD users. Therefore, results can be fully reproduced and users can perform additional analyses using other covariates or conditions. Developed packages are available through CRAN and GitHub repositories under open source licenses (GPLv3 or MIT).

4.2 Methods

In this section, we start by giving a general overview of OmicSHIELD functionalities and solutions, highlighting their virtues and indications for dealing with some of the challenges mentioned in this manuscript. Then, we demonstrate the applicability of OmicSHIELD to two real life cases of omic data analyses using state-of-the-art methods. We illustrate how to perform a FA of genomic data in the first example and how to analyse transcriptomic and epigenomic data in the second example. To this end, we provide one Opal server with all the required data to reproduce the use cases (see Section 1.4 in the online book).

4.2.1 Overview of OmicSHIELD

Herein, we provide a global overview of main features implemented in OmicSHIELD. The table 4 describes the key aspects that make OmicSHIELD the right platform to perform FA of omics data. Minimum hardware requirements to run OmicSHIELD are very different depending on the type of omics data; for genomics, storage capability is most important as genomic data sets tend to be very large, whereas analyses are usually highly optimized and do not load all the available data, thus meaning that computational power and RAM requirements are considerably lower; for other omics (e.g. transcriptomics, epigenomics), RAM requirements are more demanding and may therefore be higher. Indicative requirements are available in Section 19 of the online book.

Table 4: Key aspects of OmicSHIELD.

Key issue for omic FA	OmicSHIELD solution
Different types of federated analyses	Pooled analysis and IPD meta-analysis
Non-disclosive analyses	Those by DataSHIELD
Privacy-protected / attacks	Differential privacy, filters, audit activity
Different omic data	Functions for genomics, transcriptomics and epigenomic with easy extension to metagenomics
Open source	GPL3 and MIT license
Interaction with other tools	Post omic analyses with R/Bioconductor, other FA for clinical or epidemiological analyses with DataSHIELD

4.2.2 Security and privacy

Aside from the inherent disclosure control techniques of DataSHIELD (<https://data2knowledge.atlassian.net/wiki/spaces/DSDEV/pages/714768398/Disclosure+control>), there are additional techniques for OmicSHIELD code to ensure that server-side functions do not return disclosive information. These can be summarized as disclosure traps embedded in the functional analytic code that runs on the data processing servers, allowing only non-disclosive low-dimensional summary statistics to leave the server and therefore filtering the information that the client receives. The behaviour of these techniques can be configured by each study's data custodians in the Opal server, thus allowing individual studies to adopt a set of disclosure controls that complies with their own local required regulations (see online book Section 20), the final choice of the parameters in k-anonymity and k-nearest neighbours depends on a rigorous assessment of the real risks of disclosure and the magnitude of the information loss generated by applying the anonymization process. In general, this association is highly context dependent, and therefore the selection of the parameters must be specified based on each specific data situation. Differential privacy methods (Supplementary data 4) are also implemented to enable an additional layer of protection to results that are returned by study servers to the central analysis node. Differential privacy has been defined as the inability of an attacker to distinguish whether a single individual is present in a dataset [256]. Adding stochastic noise to function outputs is a way of achieving this: different types of noise can be used [257], with fine-tuned Laplace noise being the chosen method for our implementation (see online book Section 16). This approach has been adopted as a countermeasure to inference attacks using complex queries [258]: as such attacks can make use of GWAS results and allele frequencies, the differential privacy mechanisms we implement are intended to cause additional difficulty for such attacks. There are other strategies that can be performed using minor allele frequencies (MAF) [259]. In order to prevent these attacks, we offer an extra layer of protection by having a filter that blocks the output for SNPs with a MAF lower than a pre-specified threshold. This threshold is configurable by data owners and does not need to be the same across all the study servers.

4.2.3 Omic analytic capabilities

OmicSHIELD contains functionalities to perform three types of omic data analysis: GWAS, DGE and EWAS. The table 5 describes the main functions available in OmicSHIELD. The fig. 19 demonstrates how omic association analyses can be performed using DataSHIELD client-side functions (i.e. using the dsOmicsClient

package). Data (omics and phenotypes/covariates) are stored in their native formats at different sites (for example, remote servers accessible via https or ssh, Amazon S3, locally, etc.) managed from Opal through the resourcer R package. Analysis follows the DataSHIELD client-server architecture, implemented through a pair of libraries with dsOmics implemented server-side and dsOmicsClient client-side. Most association analyses involving omics data are based on fitting different generalized linear models (GLMs) for each feature (e.g. SNP, CpG, gene, transcript, etc.) and this forms the basis of the methods we have implemented, which includes two different types of analysis: pooled and IPD meta-analyses.

Table 5: **Main analysis functions of OmicSHIELD.** For the complete list of functions and the complete details refer to the available online guide.

	Function	Description
Genomics	ds.fastGWAS	Performs a pooled fast GWAS using the algorithm described in the “Methods” section
	ds.metaGWAS	Performs a IPD meta-analysis GWAS using the GENESIS BioConductor library
	ds.alleleFrequency	Calculates the allele frequencies. Can be used pooled or as a IPD meta-analysis
	ds.exactHWE	Calculates the exact HWE test using Fisher’s method. There is the option of only using the controls to calculate this test
	ds.PCA	Performs a pooled PCA using only the SNPs that have been linked to differentiate ethnic groups
	ds.PRS	Calculates the polygenic risk scores of the individuals sourcing the risk SNPs and weights on the PGSCatalog
	ds.PLINK	Creates a remote connection to a machine with PLINK to remotely run analysis commands using traditional PLINK syntaxis
	ds.snptest	Creates a remote connection to a machine with SNPTEST to remotely run analysis commands using traditional SNPTEST syntaxis
	manhattan	Plots a Manhattan plot using the results from ds.fastGWAS and ds.metaGWAS
	LocusZoom	Plots a LocusZoom plot using the results from ds.fastGWAS and ds.metaGWAS. It can retrieve the genes present on the region of interest using BioMaRT and TxDb.Hsapiens.UCSC.hgXX.knownGene (XX can be 37 or 38)
plotPCA	Plots the results of ds.PCA. The plot can be color coded using categorical variables of the genomic data	
Other Omics	ds.addPhenoData2eSet	Auxiliary function to add phenotype data to the ExpressionSets that contain the omic data
	ds.limma	Fits a limma + voom model
	ds.edgeR	Fits an edgeR model
	ds.DESeq2	Fits a DESeq2 model

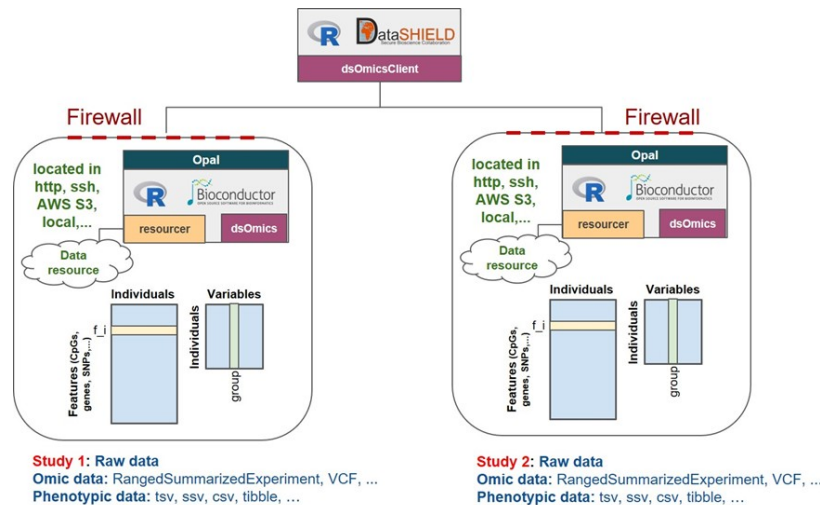


Figure 19: **Scheme of DataSHIELD implementation of omic-related packages.** The dsOmics package contains functions to perform non-disclosive data analyses of resources encoding omic data that are managed within Opal using the *resourcer* package. Omic data normally have two pieces of information, one corresponding to features (CpGs, SNPs, genes, ...) and another for phenotypic data (grouping variable, outcome, covariates, ...) that can be stored in different resources (e.g. PLINK and table in genomics) or in a specific resource designed for that purpose in R/Bioconductor (e.g. *ExpressionSet* or *RangedSummarizedExperiment*). This package should be installed in the Opal server along with their dependences. The package *dsOmicsClient* must be available in the client side and contains functions that allow the interaction between the analysis computer and the servers.

The “pooled approach” (fig. 20(A)) is recommended when the user wants to analyse omics data from different sources and obtain results as if the data were physically located in a single database. This can be very time consuming when using base DataSHIELD functions (such as *ds.glm*) which require multiple iterations of calls across the network between the client and the servers. We circumvented this problem by implementing a fast algorithm for massive generalized linear models (see Supplementary data 6). This approach is not recommended for data not fully harmonized – that is, it should not be performed when gene expressions are normalized using different methods, or when GWAS data use different platforms, as substantive heterogeneity in the nature of the data can lead to biased results [260]. The “IPD meta-analysis approach” (fig. 20(B)) overcomes limitations raised when performing pooled analyses. In particular, computation issues are addressed by using scalable and fast methods to perform data analyses at the whole-genome level at each server. Best practice methods to perform meta GWAS should be adopted [261]. We have implemented these methodologies in OmicSHIELD (see online book Section 14). Adopting a meta-analytic approach is also recommended when there is large heterogeneity in the trait of interest between cohorts.

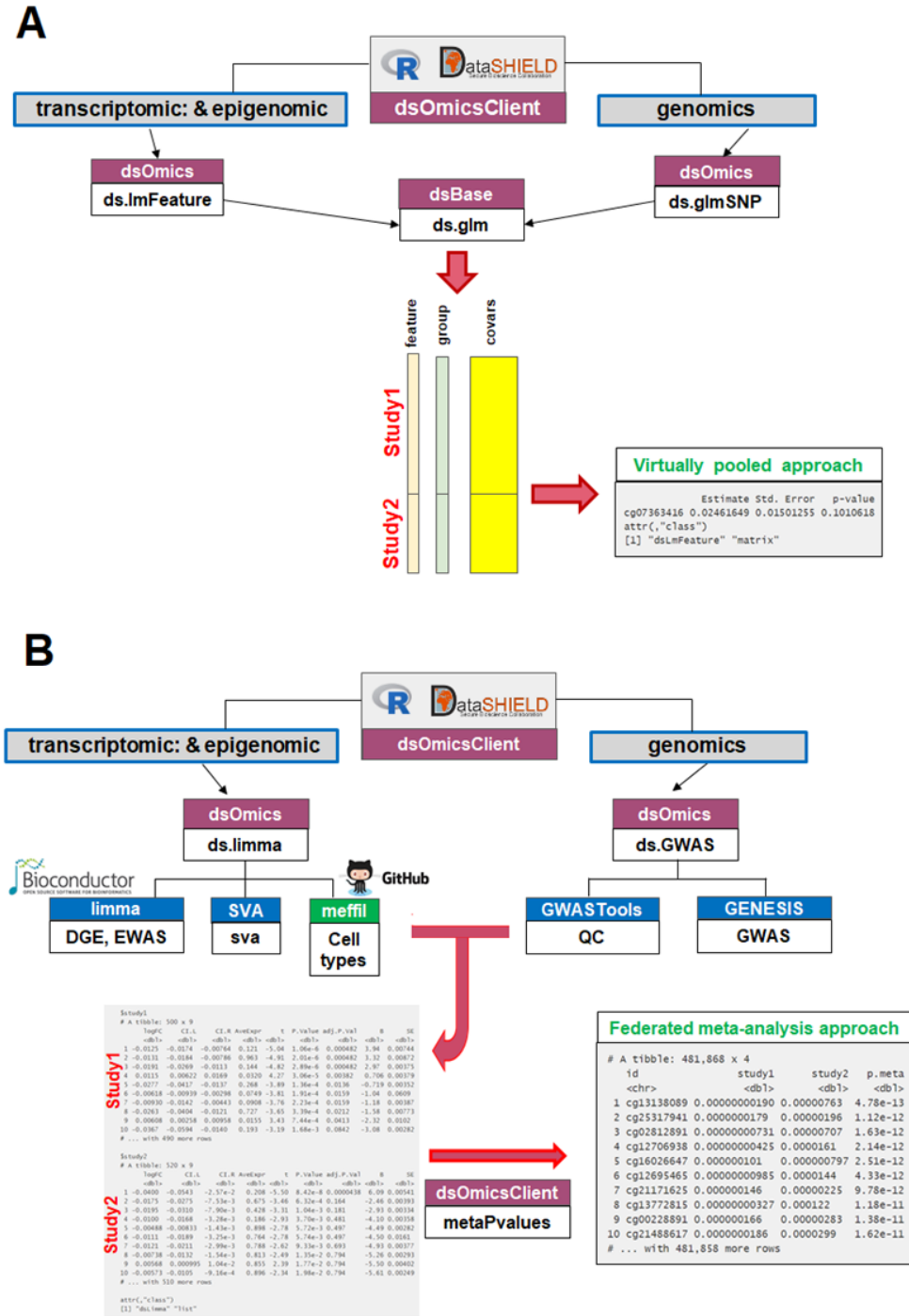


Figure 20: **Types of omics data analyses implemented in *dsOmicsClient*.** We have implemented two different types of analyses: the virtually pooled and the federated meta-analysis. **Panel A** shows the “virtually pooled approach” that is recommended when the user wants to analyse omic data from different sources (e.g studies) and obtain results as if the data were located in a single data warehouse. **Panel B** depicts the “federated meta-analysis approach” that overcomes the limitations raised when performing pooled analyses: computing time and data harmonization. The computation issue is addressed by using scalable and fast methods to perform data analysis at whole-genome level at each server. The data harmonization issue is addressed by combining results using p-values that are independent of how omic data in features have been recorded.

4.2.4 GWAS: Federated population stratification, pooled and IPD meta-analysis and polygenic risk scores

Genomic data are analysed using the GWASTools and GENESIS Bioconductor packages that allow quality control (QC) and GWAS using the Genomic Data Storage (GDS) infrastructure to be performed [262]. Section 5.1 in the online book describes how to perform analysis using a single centre containing data from the CINECA study (see Supplementary data 1), while in Section 5.2 an example of multi-centre data analyses is described using CINECA data split into three cohorts. As DataSHIELD can deal with computational resources [254], we have also implemented methods to perform meta-analyses using PLINK and SNPTEST which are standardly used to perform GWAS using genotyped and imputed SNPs, respectively.

We use the PGS Catalog to calculate polygenic risk scores (PRS, Supplementary data 8) from curated literature [263]. As this information can be disclosive, OmicSHIELD calculates PRS for each individual at each server without any interaction between the cohorts (see Section 6 in the online book). The PRS are stored on the servers and are considered thereafter as any other covariable. That is, PRS can be used as part of an association model, but extractions can be obtained only through summary statistics subject to usual disclosure controls provided by DataSHIELD.

Computing principal components analyses (PCA) is the standard methodology to address populations in GWAS, but computing federated PCA is missed in other federated GWAS solutions (e.g. FAHME [237] or sPLINK [247]). Existing approaches use principal components (PCs) estimated at each cohort and these covariates are then used for the adjustment of association models. However, PCs should be computed using the entire population to capture genetic differences among individuals [264]. OmicSHIELD is able to circumvent this issue by adopting the block approach (Supplementary data 5), thus providing a better solution than any other available elsewhere.

4.2.5 Differential gene expression analysis and EWAS

The DGE and EWAS meta-analyses provided by OmicSHIELD make use of the widely used limma package [265] that uses ExpressionSet, RangedSummarizedExperiment or GenomicRatioSet Bioconductor infrastructures to deal with omic and phenotypic (e.g. covariates) information. Section 9 in our online book describes how to perform DGE from data available in a public repository such as the The Cancer Genome Atlas (TCGA) project. This corresponds to RNA-seq data which are analysed using limma+voom [266]. We have implemented functions for using other methods such as DESeq2 and edgeR [267] as well as methods to analyse microarray data using limma. In this example, we use TCGA data available through the recount project. There is no need to store either the data or a copy in a new location, even temporarily. Specifically, it does not need to be loaded into R or uploaded to an Opal server; we simply create a resource in the Opal server called tcga.liver such that the URL is the one available from recount. Then, analysis can be directed from the local computer, with data access being managed through the Opal server in such a manner that the risk of disclosure is appropriately controlled: i.e. with inferences based exclusively on the manipulation of low dimensional summary statistics.

Section 10 in the online book illustrates how to perform EWAS. In this example, we describe how to carry out analyses using data from two different sources (e.g. two different cohorts or studies). DNA methylation profiling (Illumina 450K array) of 190 individuals (100 in study 1 and 90 in study 2) is undertaken in the superior temporal gyrus and prefrontal cortical brain regions of patients with Alzheimer’s (GEO accession number GSE66351). We are interested in determining differentially methylated probes (DMPs) between the two regions of the brain. Two resources are created in the Opal server that contain ExpressionSets with the CpG sites from each study. In this situation, a range of different analyses might be performed. One may be interested in assessing whether a given CpG is associated with a particular trait or covariate via an analysis equivalent to one applied if one has access to the whole data set (i.e. 190 individuals) in a single machine. This would represent a “pooled approach” in a single-site DataSHIELD structure. The analyses are normally performed using GLMs, as in the case of gene expression data. We can run GLMs using the ds.glm() DataSHIELD base function, using an approach that is mathematically equivalent to placing individual-level data from all sources in one central warehouse and analysing those data using the conventional glm() function in R. Our package permits the analysis of several (or even all) CpGs using

this approach. In order to speed up the process, we can run `limma` at each server using our `ds.limma()` function and, once the results from each study have been returned to the client, they can be combined using study-level meta-analysis techniques.

4.2.6 Post-omic analyses and visualization

The pooled approach returns the size effects, standard errors and some annotations of the specific features. The IPD meta-analysis approach returns study-specific estimates and standard errors generated by the analyses undertaken on each server. These results can then be combined using different meta-analytic techniques. GWAS use effect size and standard errors, while DGE and EWAS use p-values [236]. Both approaches are implemented in OmicSHIELD (see Section 5.2.8 in the online book). Once analyses are performed, different visualizations can be obtained. The results obtained from this analysis are non-disclosive since they include only the names of the features (genes, CpGs, SNPs, etc.), the annotation and the corresponding size effect, p-values and adjusted p-values. For all types of analysis, state-of-the-art plots such as qq-plots and Manhattan plots can be obtained. For GWAS, a locus zoom plot can also be created.

4.2.7 Use case 1: Multi-centric GWAS of CINECA data

To evaluate our new approach to GWAS analysis, we used a public dataset of synthetic genotype data from CINECA, information about the data can be found on Supplementary data 1. This dataset has been split into three different virtual study servers to act as individual study centres. The virtual configuration is illustrated in fig. 21. The resources with this data are available at the demo Opal server hosted by Obiba (<https://opal-demo.obiba.org/>, user: dsuser; password: P@ssw0rd) on the project called OMICS.

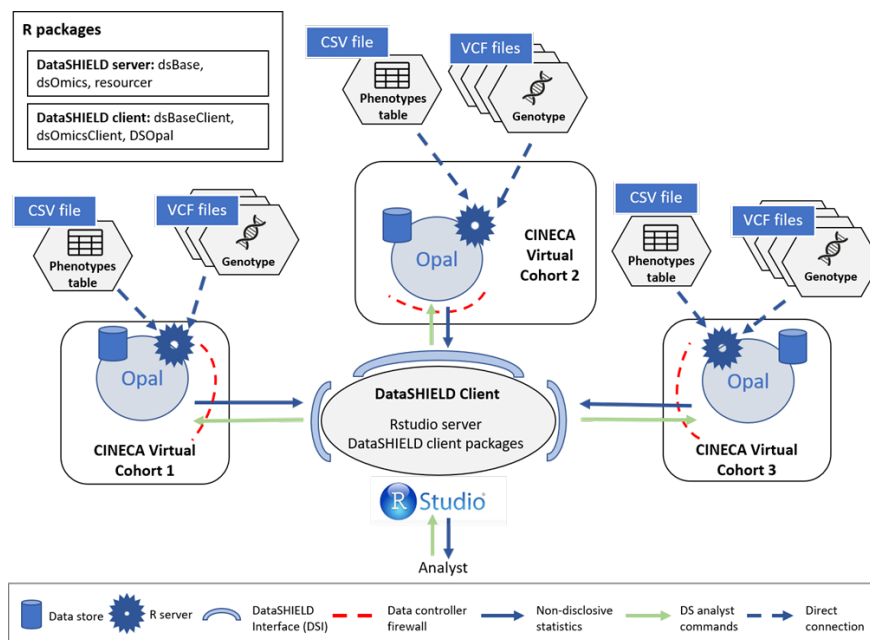


Figure 21: **Configuration for multi-centric GWAS of CINECA data using OmicSHIELD.** To achieve this configuration, the data from CINECA has been partitioned into three different virtual cohorts, containing 817, 1073 and 614 individuals respectively. Each virtual cohort contains the genotype information of the individuals, the amount of variants present for each individual is 865 thousands.

In order to offer new developments to researchers as well as provide traditional methodologies, we describe how two different approaches can be performed. In the IPD meta-analysis, results are computed at each study separately and then combined through meta-analysis, while the pooled approach results are computed using a technique that optimises the model across all servers simultaneously and therefore allows virtually

pooled results to be obtained without sharing data between the servers (hence reducing both computational and networking loads). Here, we present in detail the fast pooled GWAS using differential privacy. A complete example describing traditional meta-analysis is available in Section 5 of the online book.

We compare the results obtained using OmicSHIELD with those obtained by pulling the three datasets into a single dataset and being analysed with a single computer. We are interested in assessing associations between SNPs and diabetes, information that is obtained from a variable called ‘diabetes_diagnosed_doctor’. We adjust for other covariates including sex, age and high-density lipoprotein (HDL) cholesterol. The effect sizes (i.e. beta values) of the top 20 SNPs obtained with OmicSHIELD are compared with the effects obtained using a single dataset. This comparison yielded a mean square error of 5.3×10^{-4} and a bias of -2.6×10^{-3} which is almost negligible in practical terms (bias in the risk of a given SNP is to the order of 1 in 10000). The Manhattan plot depicted in fig. 22 shows that the top hits among the p-values and the general trend of significance levels are accurately replicated using OmicSHIELD. We can see that the noise added by the differential privacy method (-privacy: 3) allows trends to be replicated while ensuring that the top hits remain significant.

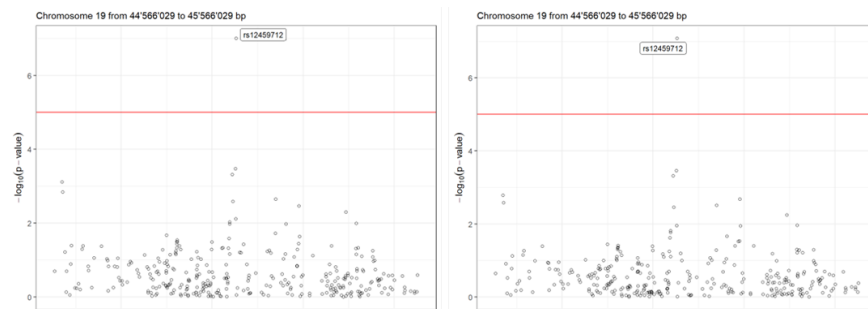


Figure 22: **Locus zoom plots of the top hit for the original data (left) and pooled fast GWAS (right)**. The red line corresponds to a threshold of significance of $-\log_{10}(P) \geq 5 \times 10^{-5}$. Trends and top hits are reproduced on the OmicSHIELD analysis, the major differences being found on the SNPs clearly not relevant.

4.2.8 Use case 2: DGE and EWAS analysis of HELIX data

Here we illustrate how to perform DGE and EWAS of HELIX data, information about the data can be found on Supplementary data 1. The data infrastructure for this project is depicted in fig. 23. Transcriptome data are stored in an Opal server as ExpressionSet objects while Epigenome data are stored as GenomicRatioSet, which are two of the standard Bioconductor infrastructures to deal with this type of omic data [268]. Note that both types of Bioconductor objects contain phenotypic data (i.e. metadata) encapsulated jointly with the omic data. All of the datasets are available in the same Opal server using different ‘resources’. In both examples, we are interested in comparing gene expression and methylation between males and females focusing only on the autosomes.

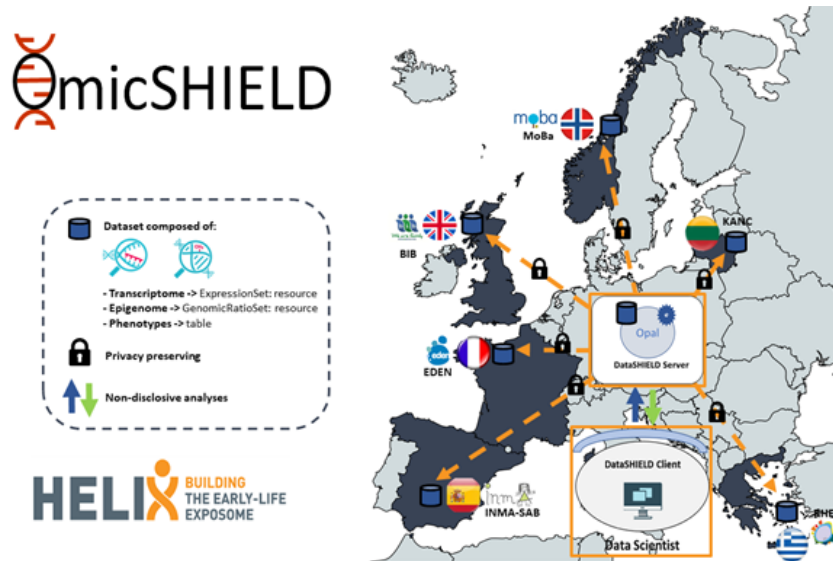


Figure 23: Opal Data infrastructure of HELIX project.

For the DGE analysis, we analyse microarray data derived using the "Human Transcriptome Array 2.0" of Affymetrix. Among the different analyses, we show how to apply IPD meta-analysis with the functions `ds.limma()` and `metaPvalues()`. For each microarray probe, this analysis implements multiple generalized linear models separately (one per study) and combine the results using study-specific derived p-values. As a result, we identified a list of 325 probes (mapping 287 genes) differentially expressed between boys and girls of the HELIX project, and which further passed multiple-testing correction filters (P-value threshold=1.74e-06). As proof of the ability of OmicSHIELD to be integrated with other R functionalities and Bioconductor packages, we continued the pipeline presenting results of a functional enrichment analysis (FEA). This analysis shows that significant differentially expressed genes participate in processes with evident and previously-described sexual dimorphism such is the case of "Longevity regulating pathways" (KEGG Pathway: hsa04211) (See Section 11.5 in the online book).

For the EWAS data we illustrate how to perform DNA methylation differential analysis using microarray data obtained with the "Infinium HumanMethylation450k" platform of Illumina. Section 12 in the online book describes how to compare the DNA methylation levels between boys and girls in the HELIX cohorts. We illustrate how to perform an epigenome-wide meta-analysis with and without adjusting for surrogate variables. We also adjusted our models by confounders including age and ethnicity. Consequently, from the initial list of almost 300k CpGs, we identify a total of 10,417 differential methylated probes between boys and girls from which only 3 passed the strict Bonferroni multiple-testing correction. Interestingly, two of these 3 probes, cg12052203 and cg25650246 (mapping the B3GNT1 and RFTN1 respectively), have been previously associated with sex methylation differences (<http://www.ewascatalog.org>). The FEA showed that significant CpGs map genes participating in processes with strong sex differences such is the case of bone formation ("Endocrine and other factor-regulated calcium reabsorption", KEGG Pathway: hsa04961).

Aside from the use cases, we have also performed a validation of our software using data from the ATHLETE project, on the validation we made sure that the results are consistent with the typical local-computation approach. The results yielded by OmicSHIELD using that data will be included on a methylation research manuscript. All the details of the validation can be found on Supplementary data 3.

4.3 Discussion

We have presented a software to perform omic analyses using multi-centre studies (i.e. federated data) with active disclosure protection during analysis and for outputs. Such a tool, based on the paradigm of DataSHIELD and Opal, provides a great opportunity for researchers to enhance multi centre collaboration by

establishing a trustworthy platform that brings the analysis to the data, hence avoiding onerous data sharing procedures. The solution we present is fully open source, enabling researchers not only to contribute their own developments, but to personally assess and control the disclosure-prevention and differential-privacy mechanisms. This guarantees the data contributors (e.g. research participants) and custodians (e.g. data controllers) complete transparency on how data is utilised, something that is central to the philosophy of the GDPR but is not achievable with commercially licensed software.

OmicSHIELD has implemented state-of-the-art methods in GWAS, transcriptomic and epigenomic data analysis including methodologies that are missing from existing approaches. For example, for GWAS, we have implemented quality control (Supplementary data 7) of individual studies before performing pooled- or meta-analyses [261], and non-disclosive pooled principal component analysis (PCA) to adjust for population stratification in pooled GWAS (something not addressed in FAHME [237] or sPLINK [247]). For transcriptomic and epigenomic studies, we have implemented outlier removal and surrogate variable analysis, and differential expression analyses using not only limma and voom but also other existing approaches, including edgeR and DESeq, that could be required, for metagenomics data analyses [267].

OmicSHIELD provides the results one would expect when having all the data combined on a single machine, however, the data are neither shared between study servers nor stored on an intermediate server. Simply stated, data never leave the study centers where they are hosted and thus remain fully controlled by their data owners. Pooled analysis can be beneficial to improve statistical power. However, this approach is not useful when huge imbalances exist between different datasets. Our software solution includes approaches for researchers to select the best methods taking into consideration how data have been collected or harmonized, different types of study designs, and data heterogeneity. Both implemented methods are privacy-protected using disclosure traps and differential privacy. The disclosure traps and differential privacy are configurable at the data source level; the decision of whether to apply differential privacy (and it's -differential privacy) is dictated by the data manager of each site, and it is possible to have different sites using different configurations, thus offering flexibility to multi-center studies where there are different views and regulations on data protection and disclosure risks.

Currently, there are several European projects that have set up an infrastructure using DataSHIELD to perform FA. These include UnCoVer [242], ATHLETE [244], LifeCycle [243], InterConnect [269], in addition to national consortia established in Germany (e.g. INTIMIC [270] and MIRACUM [271]) and Sweden [272]. They started by analysing data addressing clinical and epidemiological scientific questions, but are now also moving towards including omic data analyses. In this regard, OmicSHIELD will provide a great solution for performing FA in these large consortia and allow the examination, for instance, of the impact of the exposome on the epigenome, discover new genetic risk factor for persistent COVID or knowing how the exposome impacts on human health, among others.

The presented iteration of OmicSHIELD has the potential to reproduce many published papers as well as be the main driver of new investigative projects; nevertheless, there are many ways this software could be expanded in the future. Discussion to determine future directions to be taken will be held with researchers that use our tool for their work; in this way, we can guarantee that future versions of OmicSHIELD will contain functionalities required by real life projects, ensuring its longevity and quality. Besides having access to new developments of OmicSHIELD, researchers that choose to use the DataSHIELD ecosystem will also benefit from the growing array of available open source libraries (<https://www.datashield.org/help/community-packages>), thus enabling them to use a wide variety of tools to perform non-disclosive statistical analyses on their data.

References

- [230] Mary K Wojczynski, Michael A Province, et al. “A meta-analysis of genome-wide association studies identifies multiple longevity genes”. In: (2019).
- [231] Amadou Gaye et al. “DataSHIELD: taking the analysis to the data, not the data to the analysis”. In: *International journal of epidemiology* 43.6 (2014), pp. 1929–1944.
- [232] Karthik V Sarma et al. “Federated learning improves site performance in multicenter deep learning without data sharing”. In: *Journal of the American Medical Informatics Association* 28.6 (2021), pp. 1259–1264.
- [233] Micah J Sheller et al. “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data”. In: *Scientific reports* 10.1 (2020), p. 12598.
- [234] Ittai Dayan et al. “Federated learning for predicting clinical outcomes in patients with COVID-19”. In: *Nature medicine* 27.10 (2021), pp. 1735–1743.
- [235] Evangelos Evangelou and John PA Ioannidis. “Meta-analysis methods for genome-wide association studies and beyond”. In: *Nature Reviews Genetics* 14.6 (2013), pp. 379–389.
- [236] Daniel Toro-Domínguez et al. “A survey of gene expression meta-analysis: methods and applications”. In: *Briefings in Bioinformatics* 22.2 (2021), pp. 1694–1705.
- [237] David Froelicher et al. “Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption”. In: *Nature communications* 12.1 (2021), p. 5910.
- [238] Dingyi Xiang, Wei Cai, et al. “Privacy protection and secondary use of health data: Strategies and methods”. In: *BioMed Research International* 2021 (2021).
- [239] Zaobo He et al. “Inference attacks and controls on genotypes and phenotypes for individual genomic data”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.3 (2018), pp. 930–937.
- [240] Dana Mateş et al. “ORCHESTRA project in Romania—a prospective occupational cohort to study the impact of COVID-19 pandemic on healthcare workers”. In: *Romanian Journal of Occupational Medicine* 72.1 (2021), pp. 54–58.
- [241] Christopher Hampf et al. “A survey on the current status and future perspective of informed consent management in the MIRACUM consortium of the German Medical Informatics Initiative”. In: *Translational Medicine Communications* 6 (2021), pp. 1–11.
- [242] José L Peñalvo et al. “Unravelling data for rapid evidence-based response to COVID-19: a summary of the unCoVer protocol”. In: *BMJ open* 11.11 (2021), e055630.
- [243] Vincent WV Jaddoe et al. “The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents”. In: *European journal of epidemiology* 35 (2020), pp. 709–724.
- [244] Martine Vrijheid et al. “Advancing tools for human early lifecourse exposome research and translation (ATHLETE): Project overview”. In: *Environmental Epidemiology* 5.5 (2021).
- [245] Martine Vrijheid et al. “The human early-life exposome (HELIX): project rationale and design”. In: *Environmental health perspectives* 122.6 (2014), pp. 535–544.
- [246] L Jonathan Dursi et al. “CanDIG: Federated network across Canada for multi-omic and health data discovery and analysis”. In: *Cell Genomics* 1.2 (2021).
- [247] Reza Nasirigerdeh et al. “sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies”. In: *Genome Biology* 23.1 (2022), pp. 1–24.
- [248] Olga Zolotareva et al. “Flimma: a federated and privacy-aware tool for differential gene expression analysis”. In: *Genome biology* 22.1 (2021), pp. 1–26.
- [249] Isabelle Budin Ljøsne et al. “DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis”. In: (2014).
- [250] Rebecca C Wilson et al. “DataSHIELD—new directions and dimensions”. In: *Data Science Journal* (2017).

- [251] 2020. URL: <https://www.mrc-epid.cam.ac.uk/interconnect>.
- [252] Dany Doiron et al. “Data harmonization and federated analysis of population-based studies: the BioSHaRE project”. In: *Emerging themes in epidemiology* 10.1 (2013), pp. 1–8.
- [253] Stefan Johansson et al. “Risk of high blood pressure among young men increases with the degree of immaturity at birth”. In: *Circulation* 112.22 (2005), pp. 3430–3436.
- [254] Yannick Marcon et al. “Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD”. In: *PLoS computational biology* 17.3 (2021), e1008880.
- [255] Rachel Cummings and Deven Desai. “The role of differential privacy in gdpr compliance”. In: *FAT’18: Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2018, p. 20.
- [256] Ilya Mironov et al. “Computational differential privacy”. In: *Annual International Cryptology Conference*. Springer. 2009, pp. 126–142.
- [257] Quan Geng and Pramod Viswanath. “The optimal noise-adding mechanism in differential privacy”. In: *IEEE Transactions on Information Theory* 62.2 (2015), pp. 925–951.
- [258] Nour Almadhoun, Erman Ayday, and Özgür Ulusoy. “Inference attacks against differentially private query results from genomic datasets including dependent tuples”. In: *Bioinformatics* 36.Supplement_1 (2020), pp. i136–i145.
- [259] Kerem Ayoç et al. “The effect of kinship in re-identification attacks against genomic data sharing beacons”. In: *Bioinformatics* 36.Supplement_2 (2020), pp. i903–i910.
- [260] Maria Blettner et al. “Traditional reviews, meta-analyses and pooled analyses in epidemiology.” In: *International journal of epidemiology* 28.1 (1999), pp. 1–9.
- [261] Thomas W Winkler et al. “Quality control and conduct of genome-wide association meta-analyses”. In: *Nature protocols* 9.5 (2014), pp. 1192–1212.
- [262] Stephanie M Gogarten et al. “Genetic association testing using the GENESIS R/Bioconductor package”. In: *Bioinformatics* 35.24 (2019), pp. 5346–5348.
- [263] SA Lambert et al. “The Polygenic Score (PGS) Catalog: an open database to enable reproducibility and systematic evaluation”. In: *EUROPEAN JOURNAL OF HUMAN GENETICS*. Vol. 28. SUPPL 1. SPRINGER NATURE CAMPUS, 4 CRINAN ST, LONDON, N1 9XW, ENGLAND. 2020, pp. 135–135.
- [264] Emil Uffelmann, Qin Qin Huang, and Nchangwi Syntia Munung. “Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. “Genome-Wide Association Studies.””. In: *Nature Reviews Methods Primers* 1.1 (), pp. 1–21.
- [265] Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47–e47.
- [266] Charity W Law et al. “voom: Precision weights unlock linear model analysis tools for RNA-seq read counts”. In: *Genome biology* 15.2 (2014), pp. 1–17.
- [267] Christopher A Miller et al. “Visualizing tumor evolution with the fishplot package for R”. In: *BMC genomics* 17.1 (2016), pp. 1–3.
- [268] Wolfgang Huber et al. “Orchestrating high-throughput genomic analysis with Bioconductor”. In: *Nature methods* 12.2 (2015), pp. 115–121.
- [269] José M Terras et al. “Fostering the relation and the connectivity between smart homes and grids—InterConnect project”. In: *CIREN 2020 Berlin Workshop (CIREN 2020)*. Vol. 2020. IET. 2020, pp. 761–764.
- [270] Valeria Agamennone et al. “HDHL-INTIMIC: a European knowledge platform on food, diet, intestinal microbiomics, and human health”. In: *Nutrients* 14.9 (2022), p. 1881.
- [271] Hans-Ulrich Prokosch et al. “MIRACUM: medical informatics in research and care in university medicine”. In: *Methods of information in medicine* 57.S 01 (2018), e82–e91.
- [272] Johan Sundström et al. “Rationale for a Swedish cohort consortium”. In: *Uppsala Journal of Medical Sciences* 124.1 (2019), pp. 21–28.

**5 dsExposome: Secure and
Privacy-Preserving Exposome
Analysis using the DataSHIELD
Infrastructure**

5.1 Introduction

The exposome [273] is a term used to describe the cumulative impact of environmental exposures on human health throughout an individual’s lifetime. It encompasses a wide range of exposures, including physical, chemical, and biological factors, such as air pollution, diet, alcohol consumption, and stress. The exposome represents a holistic view of the environment and its impact on human health [274], beyond the traditional focus on specific exposures and diseases. The study of the exposome has the potential to inform the development of new prevention and treatment strategies for a wide range of health conditions [275]. For example, research has shown that exposure to air pollution can increase the risk of cardiovascular disease [275], and that exposure to toxic chemicals can increase the risk of reproductive problems [276]. By identifying the environmental exposures that contribute to these and other health conditions, the exposome provides a framework for understanding the causes and mechanisms of disease, and for developing targeted and effective prevention strategies. In addition, the exposome has the potential to improve our understanding of individual susceptibility to disease. For example, genetic and epigenetic variations may interact with environmental exposures to influence an individual’s risk of disease [277], and some individuals may be more susceptible to certain exposures due to their genetic profile. By considering the entire exposome, rather than just individual exposures, researchers can gain a more comprehensive understanding of the factors that contribute to health and disease.

Despite the potential of the exposome to advance our understanding of human health, the study of the exposome is challenging, particularly when working with large and complex datasets that may contain sensitive personal information [278]. This poses significant privacy and security risks, as well as ethical and regulatory challenges. For example, the collection and analysis of data on individuals’ lifestyles, behaviors, and health status may be subject to privacy laws and regulations, which may raise concerns about the use of personal information for research purposes. To address these challenges, it is important to adopt secure and privacy-preserving methods for collecting and analyzing exposome data. This involves using appropriate methods to protect sensitive information and to ensure that only non-disclosive aggregated results are made available to researchers. In addition, it is important to consider the ethical implications of exposome research, such as the potential for harm or discrimination, and to ensure that all participants are fully informed about the purposes and methods of the study.

To achieve the full potential of exposome research while also protecting the privacy and security of participants and their data, the use of privacy-preserving methods is crucial. In this context, DataSHIELD [279] offers a valuable solution. DataSHIELD is a decentralized infrastructure designed to allow for secure and privacy-preserving analysis of sensitive data. By enabling researchers to perform complex analyses on datasets without having to access them directly, DataSHIELD provides a secure and ethical framework for the study of the exposome. DataSHIELD works by allowing researchers to perform complex analyses on safely stored data without directly accessing it. This is achieved through data warehouses, using Opal [280] or Armadillo software [281]. Researchers can then perform their analyses on the remote data through a client-server architecture, with the client (researcher) contacting the server, which returns the results to the researcher in an aggregated, non-disclosive format. This infrastructure ensures that the underlying data remains secure and confidential, while still allowing for meaningful and robust analysis to be performed.

In this context, we present dsExposome, a DataSHIELD package for performing Exposome data analysis. The package has been designed to include the required functionalities on a typical Exposome data analysis pipeline, that is data preprocessing and normalization, identify differentially exposed features, and model exposome-outcome associations. The application fields of dsExposome include but are not limited to environmental health, epidemiology, toxicology, and precision medicine. The package can be used to analyze data from various sources, including population-based studies, cohort studies, and case-control studies. It can be used to identify potential environmental risk factors for various diseases, including cancer, respiratory and cardiovascular diseases. dsExposome can also be used for the identification of biomarkers for exposure to specific environmental agents, which can be used for monitoring environmental health and for the development of targeted interventions to reduce exposure. The development of the package has been possible thanks to the recent advances of DataSHIELD allowing all types of objects to be used (through what is called resources [282]), which catalyzed the development of our work, as we are able to use a type of objects called ‘ExposomeSet’, which are defined on the exposome analysis package ‘rexposome’ [283].

5.2 Methods

5.2.1 Opal data warehouse

Opal is a data warehousing system that was created for hosting, documenting and processing data from epidemiological studies, but has since evolved into a data hub application that controls access to various external resources and their usage in R analysis environments. As the middleware application for each data node within the DataSHIELD infrastructure, Opal performs user authentication and authorization, makes data available in private R server-side sessions, controls which operations can be executed, and tracks user activity. It acts as a broker between researchers and institutions, with institutions having control over user access to data and operations.

Opal controls input requests while the algorithms return non-disclosive information through real-time programmatic interactions via web services. The DataSHIELD client-side API is currently available in R but can be extended, while the server-side API is written in R and can be used as an entry point to other computational resources. All user interactions with the R environment are recorded by Opal for auditing purposes.

5.2.2 Datasets

5.2.2.1 Use case 1

The INMA-Sabadell Cohort's exposome data analysis focuses on investigating the effects of environmental factors on the health of children. The data, belonging to the INMA-Sabadell birth cohort [284], contains information on 88 environmental exposures, 4 health outcomes (rhinitis, whistling chest, flu, and wheezing), and 4 covariates (sex, age, BMI, and blood pressure) of 109 individuals. The individuals are children born in Sabadell (Spain) over the years 2004 and 2005, there is almost equality on the sex of the individuals.

5.2.2.2 Use case 2

The synthetic clinical data has been generated based on a study linking long-term fine particulate matter (PM_{2.5}) exposure and IHR risk [285]. The generated data contains information on 3888 patients, with the following phenotypes: IHR condition, gender, age, body mass index (BMI), and smoking status, as well as patient location information in terms of longitude and latitude. The exposure data includes PM_{2.5}, sulfate (SO₄), nitrate (NO₃), ammonium (NH₄), organic matter (OM), black carbon (BC), mineral dust (SOIL), and sea salt (SS), which are the annual mean estimates for 2020 obtained from various data sources, including NASA MODIS [286] and MISR [287], and estimated by the Atmospheric Composition Analysis Group at Washington University of St. Louis [288].

5.2.2.3 Use case 3

The HELIX project is aimed at understanding the impact of environmental risk factors on the health of mothers and children, with a special focus on molecular health profiles (omics data). The project is based on data from six European birth cohorts, including the BiB (Born in Bradford, United Kingdom), EDEN (Étude des Déterminants pré et postnatals du développement et de la santé de l'ENfant, France), INMA (Infancia y Medio Ambiente, Spain), KANC (Kaunus Cohort, Lithuania), MoBa (Norwegian Mother and Child Cohort Study, Norway), and Rhea (Mother-Child Cohort in Crete, Greece) cohorts, and includes a total of 31,472 mother-child pairs. For the purpose of this study, a sample of 1,301 children between the ages of 6 and 11, with omics data available, complete environmental history, and no serious health problems, was selected. The pre-processed transcriptomic and epigenomic data from HELIX are publicly accessible and have been added to the ISGlobal Opal server as part of the "HELIX" project, providing a platform for federated exposome, transcriptomic and epigenomic data analysis. This data set is used on the third use case.

5.2.3 Exposome analysis capabilities

Along this section we highlight two distinct analysis capabilities: server-wise and pooled. In server-wise analysis, the results obtained by the researcher are independent of the server where the data is stored. For instance, when conducting an association analysis, we obtain the results from each server, which can be further analyzed using meta-analysis techniques to obtain a global result. Pooled techniques refer to algorithms that can combine data without physically placing it together. This technique can be advantageous because, with meta-analysis techniques, statistical power can be lost.

5.2.3.1 Exposome Wide Association Studies

dsExposome main capability is to perform Exposome Wide Association Studies (ExWAS). An ExWAS is a kind of analysis where multiple generalized linear models (e.g. linear or logistic regressions) are fitted. The linear models follow the formula:

$$\text{phenotype_of_interest} \sim \text{exposure} + \text{covariates}$$

The same model is fitted for all the available exposures (or the ones we manually select). The results will then show how each exposure is related to a phenotype, which is represented by a regression coefficient and a p-value corrected for multiple testing.

Based on inputs from exposome researchers, we also included a variation of the ExWAS, which we call inverse ExWAS. In this model, we invert the position of the phenotype and exposure variables in the model, i.e.:

$$\text{exposure} \sim \text{phenotype_of_interest} + \text{covariates}$$

ExWAS can be performed server-wise as a individual patient data IPD meta-analysis approach and it can also be performed using a pooled approach. To perform the pooled analysis, the function ‘ds.glm’ from the dsBase package (core DataSHIELD package) is used. The pooled approach can bring very valuable results, as it performs the analysis as if all the data was contained in the same data set, although data are not put together, at any stage of the algorithm.

5.2.3.2 Principal component analysis

Principal component analysis (PCA) is a dimensionality reduction technique. In the study of exposome such tools are very valuable due to the high dimensionality exposome sets can have [289]. We have developed our PCA functionality both as server-wise and pooled analysis. In order to do a pooled PCA, we have used a block method [290].

The PCA method has the drawback of only working with numerical data. Most exposure data is numeric, although we may need to analyze datasets that contain categorical variables. In order to still be able to perform dimensionality reduction on them, we have implemented a factor analysis of mixed data (FAMD). This method can be interpreted as a PCA for the numerical variables and a multiple correspondence analysis (MCA) for the categorical ones. This method is only implemented server-wise, not pooled.

After performing a PCA/FAMD analysis on the exposome data, the typical following step is to apply a clustering algorithm in order to understand how the individuals can be grouped. In order to do that we implemented a hierarchical clustering of principal components (HCPC) [291].

5.2.3.3 Exposome data exploration

Aside from the core analytical functionalities, we also included methods aimed at performing exposome data exploration. We have included two different methods to check for exposure normality, the methods included are the Shapiro test and the Anderson-Darling test.

There is also the option to retrieve the summary of a given variable, whether it is numeric or categorical the summary information will be distinct:

- Numerical

- Number of observations
- Quantiles
- Mean
- Categorical
 - Number of observations
 - Category labels
 - Counts for each category

Finally, there is a method to retrieve the amount of missing exposure information. The researcher can obtain a table of percentage of missing and count of missing per exposure.

5.2.3.4 Exposure transformations

If a certain exposure does not follow a gaussian distribution, we can try to apply a conversion to it. For that reason, we included the option of applying the most typical transformation to selected exposures. The available transformations methods are 1) logarithm, 2) exponential and 3) square root.

5.2.3.5 Visualization

In order to provide a complete analysis experience, the only missing piece is good visualization of the results. We provide visualizations for all the different methods of our package.

For the ExWAS we provide a Manhattan plot. On a Manhattan plot the features (exposures) are on the x axis and on the y axis there are the p-values of the association of the features to the phenotype of interest. We also included the option of plotting the beta values instead of the p-values. The same plot can also be used for the inverse ExWAS.

There is a function to visualize the results of the PCA/FAMD analysis. This plotting functionality has been designed to replicate the type of plots that can be achieved using the traditional on-premises analysis package ‘rexposome’. It has capabilities of displaying the PCA of the exposures as well as the phenotypes.

For the descriptive analysis, there are a various type of visualizations available:

- Visualization of missing values: Simple bar plot to visually inspect the amount (or percentage) of missing values present on the exposome data.
- Visualization of exposures: We can visualize using boxplots the numerical exposures. We can filter by families if we are not interested in all the exposures.
- Visualization of single exposures: We can visualize selected exposures using binned histograms. There is the option of visualizing the current histogram plus the histograms of the transformed exposure (using logarithmic, exponential and square root transformations), normality test scores are given in all cases to assess whether the exposure needs to be transformed.

5.2.3.6 Miscellaneous

Bundled into dsExposome we have included many miscellaneous functionalities that while they do not necessarily contribute to the analysis, they are of extreme importance when using DataSHIELD. That is because since we do not have the actual data loaded into our computer, we can’t check for the family names of the exposures, the number of individuals, exposure names, etc. For that reason, we have distinct functions to obtain this information that we later need when calling certain functions.

A good example of why those functions are necessary would be visualizing a histogram of a certain exposure, we need to be able to retrieve the exposure names in order to pass that information to the plotting function. Also, if we want to perform an ExWAS, it is important to know which covariables are available and their exact naming. In theory most of that information should be available to the researchers via codebooks established

by the different cohorts; nevertheless, it is always better to have actual access to that information just in case there is a mismatch between the dataset and the codebook.

5.3 Results

5.3.1 Use case 1

In the first use case, we will demonstrate the capabilities of the dsExposome package and DataSHIELD infrastructure for performing a pooled Exposome-Wide Association Study (ExWAS) on exposome data from the INMA-Sabadell cohort. The data consists of 88 exposures and 109 individuals, divided into two cohorts of 51 and 58 individuals respectively, stored on two different Opal servers. The fig. 24 illustrates the infrastructure of this use case, showcasing the decentralization of the data and the secure nature of the analysis performed through DataSHIELD. By utilizing dsExposome and DataSHIELD, we aim to demonstrate the potential of this approach for large-scale exposome studies, advancing our understanding of the relationship between environmental exposures and human health.

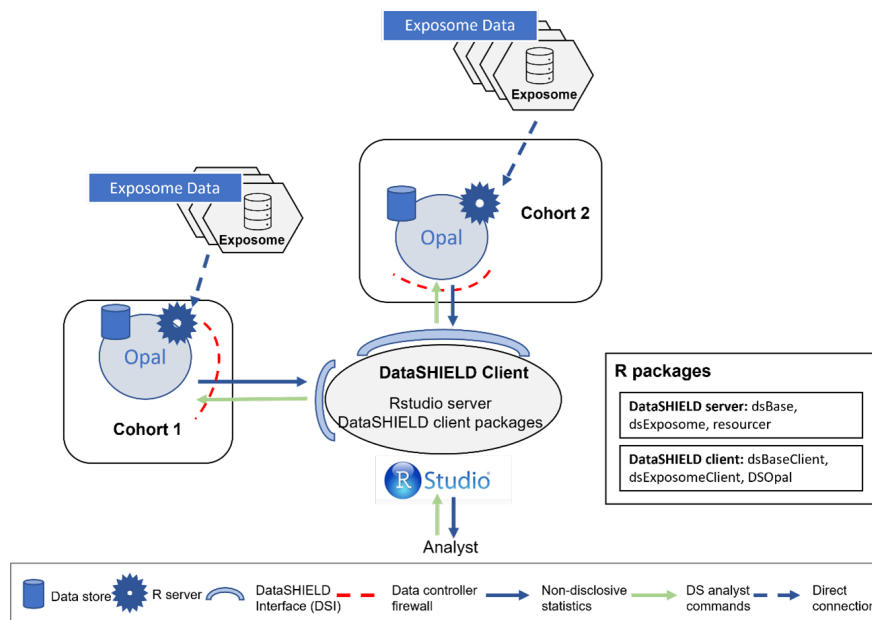


Figure 24: Architecture of the Exposome Analysis Use Case, which consists of two different Opal servers storing the INMA-Sabadell exposome data of 88 exposures and 109 individuals, divided into two cohorts of 51 and 58 individuals, respectively. The data is securely stored on each server and is available to the researcher through non-disclosive aggregated results.

Performing a pooled analysis with the dsExposome package on the DataSHIELD infrastructure means that the results we receive are the same as if we had all the data together in the same dataset. This provides a significant advantage over traditional meta-analysis studies, where results are often limited by the heterogeneity of the data or the inability to pool data from multiple sources. This offers the ability to analyze the combined data from two different opal servers as if it were one dataset, providing a robust and secure approach to exposome analysis.

The results of the analysis of the relationship between the air pollutants and metal exposures with the flu health condition can be found in fig. 25, this figure has been generated with functions already available on dsExposome, there is no need of sourcing third party libraries for visualizing results. The figure presents the results in the form of a Manhattan plot, which graphically depicts the impact of different exposures on the incidence of the flu health condition. The Manhattan plot provides valuable insights into the potential relationship between the exposures and the health condition, allowing for easy interpretation and analysis

Overall, this use case showcases the versatility and power of DataSHIELD in the analysis of complex, multi-modal sensitive data.

The fig. 26 provides an overview of the architecture for this use case, showcasing how the synthetic clinical data and NetCDF sources are all stored on the Opal server, linked together, and made available for analysis.

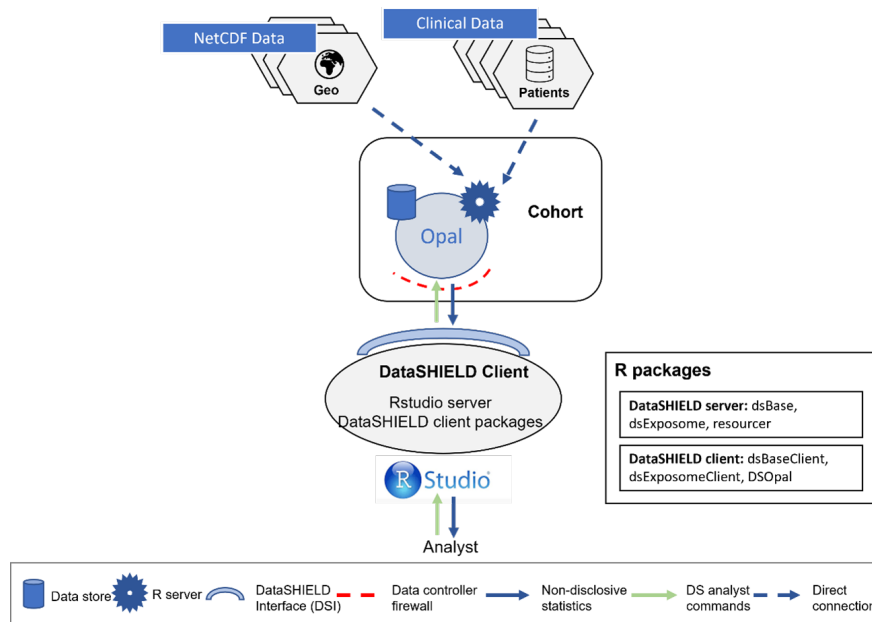


Figure 26: Architecture of the Geospatial Data Use Case. The diagram illustrates the setup of the study, with the clinical data and geospatial data securely stored on a single Opal server.

In fig. 22, we showcase the geospatial representation of the mean ground-level PM2.5 data of 2020 in the USA. It is important to note that, despite the ability to use this information as a covariable in our analysis, the client does not have access to the raw data and is unable to generate this map themselves. The map is shown purely to provide an insight into the appearance of the data.

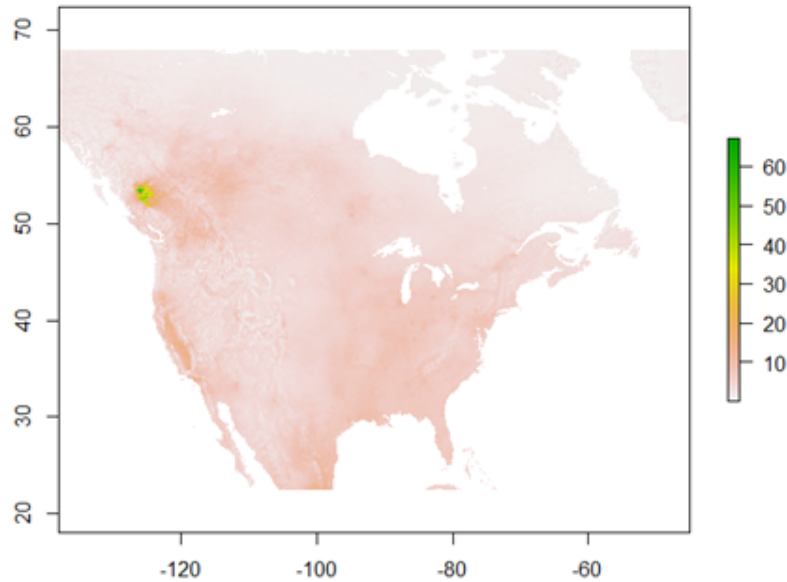


Figure 27: Visualization of Mean Ground-Level Fine Particulate Matter (PM2.5) Data for the USA in 2020. This map serves only for illustrative purposes and the client is not able to generate it. The data has been extracted from a NetCDF data source that is securely stored on the Opal server.

With all this data at our disposal, we can use DataSHIELD and dsExposome to perform the proposed ExWAS analysis. More precisely, we will assess the relationship of the different geo-located exposures to the IHR condition adjusted by gender and smoke condition (smoker/non-smoker). The results of this analysis can be seen on the fig. 23 in the form of a Manhattan plot. This plot displays the association between each exposure and the IHR health condition. The plot illustrates the impact of including this additional information on the analysis, and provides insight into the potential relationship between the different exposures and the IHR health condition. As expected, we obtain a significant association between the health condition and PM2.5 as we generated our synthetic data to have this association. Nevertheless, it is an indicator that 1) we have successfully related patient location data to geo-location exposure data, and 2) the ExWAS results of the dsExposome package are reliable.

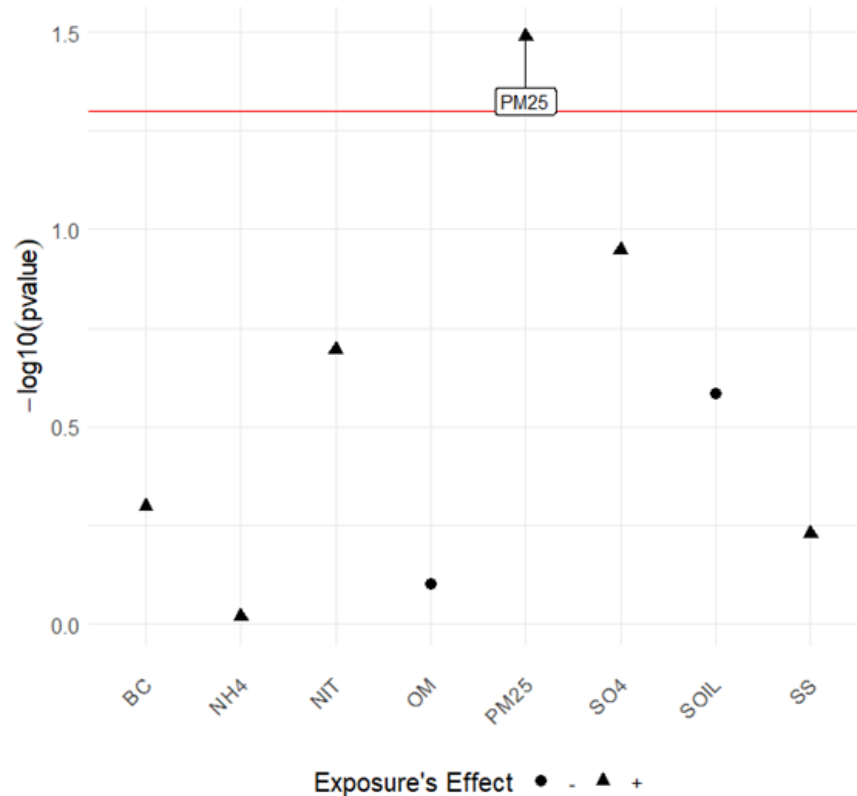


Figure 28: Manhattan plot of the ExWAS analysis results showing the relationship between different geo-coded air exposures and IHR health condition. The horizontal red line represents the significance threshold of p-value = 0.05.

5.3.3 Use case 3

This third use case aims to demonstrate the capabilities of the dsExposome package and DataSHIELD infrastructure for performing an exposome analysis on a multi-centre study framework, this time with real data from the HELIX study [294], a European multi-cohort exposome collaborative effort. The focus is on applying an ExWAS to obtain a real relationship between a health condition and different exposures. To highlight the potential of DataSHIELD in this context, the use case replicates published analyses investigating the association between a range of prenatal and postnatal exposures and blood pressure in children, which was conducted by Warembourg et al. [295] following traditional techniques: data transfers between collaborating centers and study of it on-premises.

The original study found that a variety of internal and external exposures impact both diastolic and systolic blood pressure readings. For the purpose of simplicity, we will only replicate the results reported for postnatal exposures and systolic blood pressure. These results use as confounding factors the child age, child height, child sex, child cohort, mother’s age at the moment of birth and mother’s BMI at the moment of birth.

As seen on the first use case, when performing an ExWAS with dsExposome, we can do it as a pooled analysis, which provides a great advantage over traditional meta-analysis methods. However, when dealing with real world data instead of synthetic datasets generated for demonstration purposes, we do not always have the opportunity to use this advance, as poorly harmonized datasets could play against our purposes. Fortunately, the HELIX data used for this use case was well harmonized, so we performed a pooled analysis.

The comparison of our analysis results with the published ones is presented in table 6, we present the significant exposures found by the publication and their p-values and compare it to the p-values we found. We observe that the most significant findings obtained using dsExposome are like those in the paper we

aimed to replicate. However, it’s worth noting that our pipeline doesn’t exactly reproduce the methodology used in the publication and thus, the results may not be exactly the same. We can clearly see how overall the p-value magnitude order as well as beta sign is consistent across the board; the beta values present a greater value variance, but never a sign difference.

Table 6: Results presented on the publication by Warembourg et. al (2019) compared to the results obtained on the same analysis using dsExposome. The table contains information on the 11 top hits from the publication.

Exposure	Publication p-val	dsExposome p-val	Publication beta	dsExposome beta
DDE	0.00000	0.00000	-2.10	-1.47
HCB	0.00000	0.00000	-2.05	-3.97
PCB 153	0.00009	0.00012	-1.90	-2.24
PCB 170	0.00068	0.00036	-1.73	-1.14
PBDE 153	0.00093	0.00298	-1.66	-0.49
PCB 138	0.00099	0.00061	-1.54	-1.89
PCB 180	0.00142	0.00037	-1.77	-1.23
PCB 118	0.00337	0.00603	-1.07	-1.80
MBzP	0.01069	0.00741	-0.87	-0.89
MEHP	0.02466	0.04348	-0.80	-0.73
Indoor benzene	0.04105	0.04424	0.78	2.47

In summary, with this use case we highlight the ability of dsExposome to provide valuable results in real world multi-centre studies, such as the HELIX study. By replicating the findings of Warembourg et al. and demonstrating the compatibility with real world data, we further establish dsExposome as a valuable tool for exposome researchers.

If the reader is interested in replicating the use cases, we have made available a selection of supplementary material at: <https://isglobal-brge.github.io/Supplementary-Material/>. The user guides contain step-by-step commented code of how to reproduce the presented use cases using dsExposome and DataSHIELD. Also, we have uploaded the raw data for use cases 1 and 2, so that the reader can reproduce the results both using DataSHIELD and traditional techniques (on-premises analysis) in order to properly assess the veracity of the results provided by our solution. For the use case 3 the raw data and DataSHIELD access will be provided upon reasonable request.

5.4 Discussion

The study of the exposome has the potential to revolutionize our understanding of human health and disease, by providing a holistic view of the impact of environmental exposures on health. However, the study of the exposome is challenging, particularly when working with large and complex datasets that may contain sensitive personal information, which poses significant privacy and security risks, as well as ethical and regulatory challenges. To address these challenges, it is important to adopt secure and privacy-preserving methods for collecting and analyzing exposome data.

The DataSHIELD infrastructure provides a solution to these challenges by enabling secure and privacy-preserving analysis of sensitive data. By allowing researchers to perform complex analyses on datasets without having to access them directly, DataSHIELD provides a secure and ethical framework for the study of the exposome. This is achieved using data warehouses, such as Opal or Armadillo, and the client-server architecture of DataSHIELD, which ensures that the underlying data remains secure and confidential, while still allowing for meaningful and robust analysis to be performed.

Here we present dsExposome, an open-source package that gathers the main tools and methods employed in the study of the exposome. These tools include data preprocessing, machine learning methods such as PCA, clustering and association analyses. Our solution is written using R and is embedded inside the DataSHIELD ecosystem of packages, which further increases the value of our proposition, given that the researchers can

integrate different techniques on their analyses. For example, they can also have access to non-disclosive Lasso regression methods to be used with the exposome data. Such integration is a key for the quality of the exposome research of the future, which as we demonstrated can rely on non-disclosive techniques.

This study showcases the effectiveness of the dsExposome package and DataSHIELD infrastructure in conducting meaningful exposome analysis across various scenarios. Especially it is important to remark the third use case, which indicates the potential and viability of this approach in conducting exposome research while ensuring the protection of individuals' privacy and data security.

The results we presented offer various perspectives on what our software is capable of. Initially, we emphasized on two separate use cases using both synthetic and public data, allowing interested readers to conduct the same analysis locally and verify the results our solution provides for transparency. In the third use case, we demonstrated the effectiveness of our product by obtaining meaningful results in a real-world application. Our findings were not an exact match to the previously published study, but this was not the goal of our package. Instead, we aimed to create a flexible and adaptable foundation for non-disclosive exposome data analysis, and we achieved this by making the software open-source and receptive to new functionalities from developers. This approach allows researchers to tailor the package to their specific analysis needs on future projects, ensuring the most useful and relevant results.

Overall, the DataSHIELD infrastructure and dsExposome package provide a valuable solution for the study of the exposome, by enabling secure and privacy-preserving analysis of sensitive data, and by allowing researchers to perform complex analyses without compromising the privacy of the individuals represented in the data. As such, this approach represents a significant step forward in the development of a secure and ethical framework for the study of the exposome.

References

- [273] Christopher Paul Wild. “The exposome: from concept to utility”. In: *International journal of epidemiology* 41.1 (2012), pp. 24–32.
- [274] Germaine M Buck Louis, Melissa M Smarr, and Chirag J Patel. “The exposome research paradigm: an opportunity to understand the environmental basis for human health and disease”. In: *Current environmental health reports* 4 (2017), pp. 89–98.
- [275] Andreas Daiber et al. “The “exposome” concept—how environmental risk factors influence cardiovascular health”. In: *Acta Biochimica Polonica* 66.3 (2019), pp. 269–283.
- [276] John D Meeker et al. “Semen quality and sperm DNA damage in relation to urinary bisphenol A among men from an infertility clinic”. In: *Reproductive toxicology* 30.4 (2010), pp. 532–539.
- [277] Henrik Christian Bidstrup Leffers et al. “The study of interactions between genome and exposome in the development of systemic lupus erythematosus”. In: *Autoimmunity reviews* 18.4 (2019), pp. 382–392.
- [278] Fernando Martin Sanchez et al. “Exposome informatics: considerations for the design of future biomedical research information systems”. In: *Journal of the American Medical Informatics Association* 21.3 (2014), pp. 386–390.
- [279] Amadou Gaye et al. “DataSHIELD: taking the analysis to the data, not the data to the analysis”. In: *International journal of epidemiology* 43.6 (2014), pp. 1929–1944.
- [280] Dany Doiron et al. “Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination”. In: *International journal of epidemiology* 46.5 (2017), pp. 1372–1378.
- [281] Molgenis. *Molgenis Software*. <https://www.molgenis.org/>. Accessed: Feb. 02, 2023. 2023.
- [282] Yannick Marcon et al. “Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD”. In: *PLoS computational biology* 17.3 (2021), e1008880.
- [283] Carles Hernandez-Ferrer et al. “Comprehensive study of the exposome and omic data using rexposome Bioconductor Packages”. In: *Bioinformatics* 35.24 (2019), pp. 5344–5345.
- [284] Núria Ribas-Fitó et al. “Child health and the environment: the INMA Spanish Study”. In: *Paediatric and perinatal epidemiology* 20.5 (2006), pp. 403–410.
- [285] Stacey E Alexeeff et al. “Long-term PM_{2.5} exposure and risks of ischemic heart disease and stroke events: review and meta-analysis”. In: *Journal of the American Heart Association* 10.1 (2021), e016890.
- [286] C. O. Justice et al. *An overview of MODIS Land data processing and product status*. <http://www.edc.usgs.gov/programs/sddm/modisdist/index.shtml>. Accessed: Feb. 07, 2023. 2023.
- [287] David J Diner et al. “Multi-angle Imaging SpectroRadiometer (MISR) instrument description and experiment overview”. In: *IEEE Transactions on Geoscience and Remote Sensing* 36.4 (1998), pp. 1072–1087.
- [288] Aaron Van Donkelaar et al. “Monthly global estimates of fine particulate matter and their uncertainty”. In: *Environmental Science & Technology* 55.22 (2021), pp. 15287–15300.
- [289] Vrinda Kalia et al. “Unsupervised dimensionality reduction for exposome research”. In: *Current opinion in environmental science & health* 15 (2020), pp. 32–38.
- [290] Michael W Berry et al. “Parallel algorithms for the singular value decomposition”. In: *Statistics Textbooks and Monographs* 184.117 (2006), p. 31.
- [291] F. Husson et al. *Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data?* <http://www.agrocampus-ouest.fr/math/>. Accessed: Feb. 10, 2023. 2023.
- [292] Courtney D Kozul et al. “Low-dose arsenic compromises the immune response to influenza A infection in vivo”. In: *Environmental health perspectives* 117.9 (2009), pp. 1441–1447.
- [293] Russ Rew and Glenn Davis. “NetCDF: an interface for scientific data access”. In: *IEEE computer graphics and applications* 10.4 (1990), pp. 76–82.

-
- [294] Léa Maitre et al. “Human Early Life Exposome (HELIX) study: a European population-based exposome cohort”. In: *BMJ open* 8.9 (2018), e021311.
- [295] Charline Warembourg et al. “Early-life environmental exposures and blood pressure in children”. In: *Journal of the American College of Cardiology* 74.10 (2019), pp. 1317–1328.

6 ShinyDataSHIELD - An R Shiny application to perform federated non-disclosive data analysis in multi-cohort studies

6.1 Introduction

Data privacy continues to be a central concern in contemporary research [296]. There are many ethico-legal considerations, including requirements under General Data Protection Regulation (GDPR), that must be taken into account when planning and configuring an analysis involving sensitive data; in particular mitigating the risk of individual identification. Such considerations have a huge impact on the feasibility and time to carry out multi-cohort studies and genomic studies, which rely on obtaining permissions to access and share sensitive data [297]. DataSHIELD is an open-source software infrastructure aimed at facilitating an effective and appropriate response to such challenges [298, 299]. To achieve this, DataSHIELD represents an infrastructure where the researchers only ever receive sufficient statistics (low dimensional data transformations/aggregations containing all of the information needed to drive whatever analysis is required) from each of the different data servers, while the servers themselves manage their data using Obiba's Opal technology [300]. The data owners/custodians manage these servers, and have sole control of the disclosure filters that are applied to the outputs, as well as the set of DataSHIELD functions that can be used on their data. This enables researchers to perform analyses on federated data without the need to possess physical copies of the data from each source. On the DataSHIELD website (<https://www.datashield.org/help>) the reader can find information on how to:

1. Conduct basic statistical analyses
2. Administrate the servers
3. Deploy DataSHIELD functions and packages

Recently, DataSHIELD has seen a major upgrade focused on expanding the scope of which types of data can be analysed [301], which results in the ability to analyse high volume, potentially non-tabular, data such as genomics data structures among many others. Many science fields can now make use of this extension, and therefore it is important to make DataSHIELD easier to use for non-technical users. The DataSHIELD infrastructure includes a series of R packages that enables the remote and non-disclosive analysis of sensitive data (<https://www.datashield.org/help/community-packages>). The software described in this article uses a subset of functionalities from the dsBase [302] package for data shaping, analysis and presentation methods, the dsSurvival [303] package for survival analysis and the dsOmics [304] package for omics analysis. We present ShinyDataSHIELD, an R Shiny [305] application that enables interaction with the DataSHIELD analysis infrastructure via a web application, providing capabilities to perform federated non-disclosive analysis for non-technical users. We have designed the application to provide a user-friendly experience that frees the researcher from writing any analysis code.

6.2 Implementation

The following list describes all the functionalities of the software presented, which can be used in two configurations: (1) single data sources; (2) multiple data sources in a federated configuration. The single-source configuration invokes all of the privacy protection features of DataSHIELD but, by definition, has no need to activate the routines and algorithms permitting federated co-analysis across multiple sites. For further information we have made available an online user-guide with different use cases and technical information (https://isglobal-brge.github.io/ShinyDataSHIELD_bookdown/). All the functionalities use DataSHIELD disclosure controls.

Tabular data functionalities:

1. Data column types: When dealing with tabular data a researcher may be interested in: 1) Assessing the column class; and 2) Transforming the class of a column. Both functionalities are available.
2. Descriptive statistics: There are a number of functions that can display descriptive statistics in two different ways:
 - (a) Summary statistics: Available for numeric and categorical variables. It displays a table with summary statistics. For categorical variables it displays the number of counts in each category, and for numerical variables, the quartiles and the mean values.

- (b) Graphical representations: 1) Scatter plot, to visualize the relationship between two numerical variables; 2) Histogram, to visualize the distribution of a numerical variable; 3) Heatmap, to visualize the density of counts in grids formed by two numerical variables; 4) Boxplot, to visualize the locality and spread of one or more numerical variables, with the option of performing two groupings using categorical variables. All graphical representations preserve data privacy as they are generated through anonymisation techniques or disclosure controls [306].
3. Statistical modelling: There are three classes of statistical models that can be fit.
- (a) Generalized linear models (GLM): GLM models can be fitted using pooled techniques or meta-analysis techniques (study-level meta-analysis fitting will also yield a forest plot of the results as well as the regression results table). For both approaches the user can specify the error distribution to be either Gaussian, Poisson or binomial for linear, poisson and logistic regressions respectively.
 - (b) Generalized linear mixed effects models (GLMer): GLMer models are fitted using meta-analysis techniques (a forest plot is displayed alongside the results table). The user can specify the error distribution to be either Poisson or binomial.
 - (c) Survival analysis: Survival analysis can be performed via a study-level meta-analysis of Cox regression models. The models can be fitted for different types of censoring: left, right, counting and intervals. The regression results are displayed in a table and a forest plot, while privacy-preserving survival curves are also displayed.

Resources functionalities:

1. Genomics: Genome-wide association study (GWAS) can be performed using two types of resources: VCF (Variant Call Format) files and PLINK containers. The VCF files are analysed using BioConductor libraries (GWASTools [307] mainly), while PLINK containers are analysed using the PLINK software [308]. The GWAS results can be visualized as a table or as a Manhattan plot [309].
2. Omics: Perform association analysis using Limma [310]. The accepted resources for this analysis are ExpressionSet and RangedSummarizedExperiment containers [311].

Miscellaneous functionalities:

The miscellaneous functionalities are not part of DataSHIELD, they are part of the software that improves the user experience.

1. Plot editor: There is a built-in plot editor that allows some simple customization for the generated plots. This editor has been built using the ggplot2 and the ggthemr [312] R packages. The options that the plot editor offers are: 1) Change text size; 2) Change X-axis text angle; 3) Add title, subtitle and caption; 4) Custom labels for X and Y axes; 5) Custom legend label (if there is a legend to be customized); 6) Colour themes.

ShinyDataSHIELD has been implemented using a modular approach where all the modules have the same design, so there is no confusion when using different functionalities. Each one of the different modules performs a single task: a module for statistical modelling, a module for descriptive analysis, etc. This guarantees that the application will be easy to upgrade and the source code will be easier to read for future maintainers.

The language for interacting with the modules follows a similar pattern, which in our application is the following: 1) When entering a module, the researcher must select the tables or resources to be used. Internal checks are always performed on the selected items to ensure the functions of the module will not crash; 2) The buttons are disabled until the researcher performs an operation that requires them, e.g. the visualization buttons for the survival models are not available until a survival model is fitted. This shared language ensures a consistent experience.

Operating in the background, R Shiny provides executive control of the functions in the different DataSHIELD packages, thus providing a seamless experience for the researcher, who simply receives the aggregated results as tables and figures interacting with a web-application.

6.3 Use

In this section we will explain how to use the software. As previously mentioned, there is a common structure across all the functionalities described that provides for a friendly user experience. With this provision, the typical experience of a ShinyDataSHIELD user should be both easy and intuitive. A key component of our software is that there are multiple checks that display human readable error messages, helping the users understand the reason something is not working. If the reader wishes to reproduce the displayed screenshots using our software, he/she can refer to the online user guide.

The first step is to define which Opal servers will be used, the credentials to be applied to them and which tables or resources have to be loaded into the remote R sessions hosted at the Opal servers. It is in this step where we define whether we want to use a single data source or multiple. To define multiple data sources, we just have to add more servers. This allows us to perform pooled analysis – combining inferences across all the specified servers. Once we have performed the connections, we only have access to the particular set of datasets we have selected. If we wish to use different ones, we have to disconnect and reconnect again specifying the new set of data sets we now require. This step is illustrated in fig. 29.

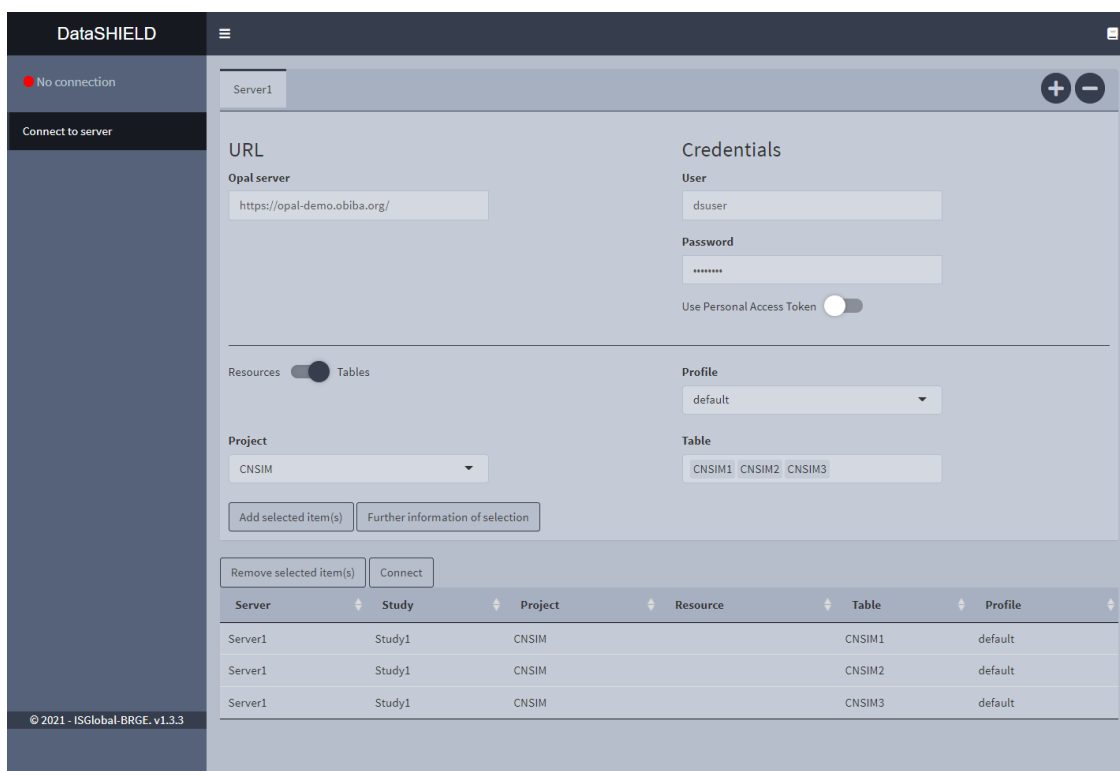


Figure 29: Connections interface. The illustrated configuration is a single server data source configuration with three different tables selected.

The next step is to select which of the array of data available on each server should be used. This can sound counterintuitive given we just selected the data to load in the previous step. However, it adds flexibility, for example we can load multiple datasets and study them separately without having to disconnect. Moreover, it is at this point where the data is checked for integrity for doing pooled analysis. This ensures that the tables contain equivalent variables. This step is illustrated in fig. 30.

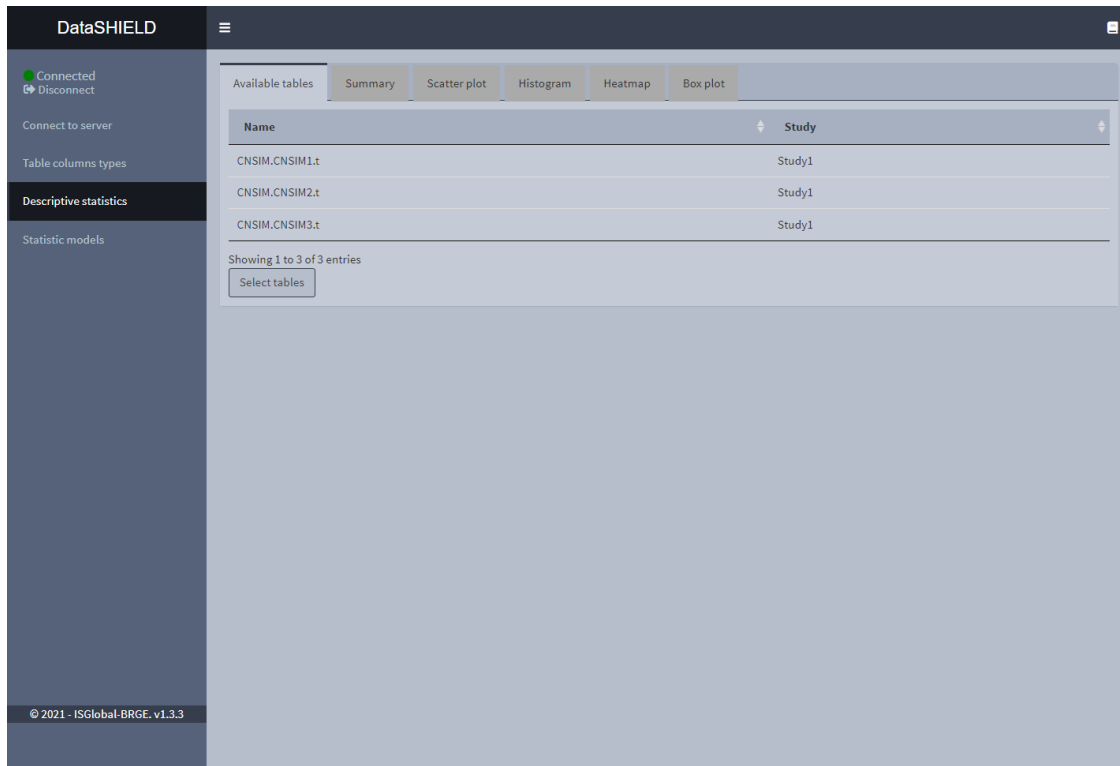


Figure 30: Selecting the tables to use. We can see, following Figure 1, that we have three available tables.

With this last step completed, we are inside the module that we have selected. Now we can finally use our data to undertake the desired statistical data analyses. All the modules have multiple functionalities, for example the descriptive analysis module has different visualization options (see Section 3.3 in the bookdown), all the functionalities available are displayed along the navigation tabs of each module (see fig. 30). Linear, generalized linear and survival models can also be fitted using our shiny app as described in section 3.4 in the bookdown.

There are other modules that have functionalities that can only be used when a certain action has been performed. This is very easy and intuitive for the users, for example, the genomics module can perform a GWAS and it can also plot the results (e.g. Manhattan plot). But, the plot can only be created once a GWAS analysis has been performed (see Section 3.5 in the bookdown).

Having followed all the steps described above, the user is now in a position to perform any of the analyses described in the Implementation section.

6.4 Discussion

ShinyDataSHIELD is a novel R Shiny application that enables federated non-disclosive analyses for non-technical users. The goal of our software is to make the DataSHIELD infrastructure more accessible to researchers without R skills, as well as providing a platform for researchers experienced in DataSHIELD to perform quick hypothesis prototypes and quick analyses without the burden of writing a new analysis pipeline within R. For reference, even a basic pipeline aimed at fitting a linear model, while necessarily incorporating appropriate check code, may typically require anywhere from 25 to 50 lines of code. However, despite benefitting from the simplicity provided by ShinyDataSHIELD the researcher is not freed from the need to ensure that he/she is correctly interpreting the statistical results obtained from the application and that all models are built upon correct assumptions.

Our software is designed around core DataSHIELD functionalities and will be expanded as new functionalities

and packages are available. With the current version of the Shiny app the plot editor can only be used for customizing Box plots, but we aim to make this functionality available for all other types of plots in a future release. Also, new functionalities will be added if researchers request them. Moreover, other research groups could take the source code and modify it to suit their particular needs. Since our software wraps DataSHIELD functions, periodical revisions from the maintainers will be required when new versions of DataSHIELD are released, to ensure that no wrappers are broken.

In conclusion, helping researchers to adopt non-disclosive methods for potentially federated analyses is inevitably a challenging task. When using DataSHIELD via its traditional integrated development environment this involves learning DataSHIELD syntax. By providing a user-friendly Shiny R based tool that can simplify the procedures and shorten the learning curve, we believe that this article and the technical work program underpinning it can contribute towards an accelerated adoption of such methods as well as demonstrating the capabilities of DataSHIELD. We hope that this will encourage researchers interested in the technology to explore its capabilities and test them out for themselves. Given ongoing developments in ShinyDataSHIELD as well as rapid evolution of the DataSHIELD infrastructure itself, ShinyDataSHIELD will be actively maintained and upgraded in the years to come, so that all the novel functionalities introduced by the DataSHIELD community can be used and be utilized in on our application.

References

- [296] Maria Petrescu and Anjala S Krishen. “Analyzing the analytics: data privacy concerns”. In: *Journal of Marketing Analytics* 6 (). DOI: 10.1057/s41270-018-0034-x. URL: <https://doi.org/10.1057/s41270-018-0034-x>.
- [297] Karim Abouelmehdi et al. “Big data security and privacy in healthcare: A Review”. In: vol. 113. Elsevier B.V., Jan. 2017, pp. 73–80. DOI: 10.1016/j.procs.2017.08.292.
- [298] Amadou Gaye et al. “DataSHIELD: taking the analysis to the data, not the data to the analysis”. In: *International Journal of Epidemiology* 43 (6 Dec. 2014), pp. 1929–1944. ISSN: 0300-5771. DOI: 10.1093/IJE/DYU188. URL: <https://academic.oup.com/ije/article/43/6/1929/707730>.
- [299] Rebecca C. Wilson et al. “DataSHIELD - New directions and dimensions”. In: *Data Science Journal* 16 (Apr. 2017). DOI: 10.5334/DSJ-2017-021.
- [300] Dany Doiron et al. “Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination”. In: (). DOI: 10.1093/ije/dyx180. URL: www.obiba.org.
- [301] Yannick Marcon et al. “Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD”. In: *PLOS Computational Biology* 17 (3 Mar. 2021), e1008880. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1008880. URL: <https://doi.org/10.1371/journal.pcbi.1008880>.
- [302] DataSHIELD Core Development Team. *dsBase: v6.1.1*. 2020. URL: <https://github.com/datashield/dsBase>.
- [303] Soumya Banerjee and Tom R P Bishop. “neelsoumya/dsSurvival: v1.0.0 Survival models in DataSHIELD”. In: (June 2021). DOI: 10.5281/ZENODO.4917552. URL: <https://zenodo.org/record/4917552>.
- [304] Xavier; Marcon Yannick González Juan R.; Escriba-Montagut. *dsOmics: v1.0.7*. 2020. URL: <https://github.com/isglobal-brge/dsOmics>.
- [305] JJ Allaire Yihui Xie Jonathan McPherson Winston Chang Joe Cheng. *shiny: Web Application Framework for R. R package version 1.5.0*. 2020. URL: <https://cran.r-project.org/package=shiny>.
- [306] Demetris Avraam et al. “Privacy preserving data visualizations”. In: *EPJ Data Sci.* (2021). DOI: 10.1140/epjds/s13688-020-00257-4. URL: <https://doi.org/10.1140/epjds/s13688-020-00257-4>.
- [307] Stephanie M. Gogarten et al. “GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies”. In: *Bioinformatics* 28 (24 Dec. 2012), pp. 3329–3331. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTS610. URL: <https://academic.oup.com/bioinformatics/article/28/24/3329/246030>.
- [308] Shaun Purcell et al. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses”. In: *The American Journal of Human Genetics* 81 (3 Sept. 2007), pp. 559–575. ISSN: 0002-9297. DOI: 10.1086/519795.
- [309] Stephen D Turner. “Annotated Manhattan plots and QQ plots for GWAS using R, Revisited”. In: (2011). DOI: 10.1038/npre.2011.6070.1. URL: <http://www.stephenturner.us/>.
- [310] Gordon K Smyth. *Limma, Linear Models for Microarray Data: v3.48.3*. 2021. URL: <https://bioconductor.org/packages/release/bioc/html/limma.html>.
- [311] *SummarizedExperiment for Coordinating Experimental Assays, Samples, and Regions of Interest*. URL: <https://bioconductor.org/packages/release/bioc/vignettes/SummarizedExperiment/inst/doc/SummarizedExperiment.html>.
- [312] Mikata-Project. *ggthemr: v1.1.0*. 2020. URL: <https://github.com/Mikata-Project/ggthemr>.

7 Breakthroughs Beyond Published Work

7.1 Introduction

Alongside developing the contents presented on this manuscript in the form of published papers, some other work was developed but was never intended to be published. This work is a DataSHIELD analysis package, named `dsMLClient`. This package was designed with the intention of integrating various machine learning methodologies within the DataSHIELD infrastructure. The `dsMLClient` tried to offer a new way of doing machine learning tasks. It did this by stepping away from the normal server-based approach used in the current DataSHIELD machine learning tools.

The primary goal was to adapt traditional machine learning methods and transform them into a comprehensive package that could operate as a pooled algorithm across multiple servers. The intent was to construct an analysis package that could extend the reach of machine learning methods into the DataSHIELD ecosystem.

Even though it was envisioned as a proof-of-concept, there was still a great emphasis on disclosure prevention, so all the methodologies were developed and implemented with that in mind. The other emphasis of this work was to implement all the methodologies so they could operate in a pooled configuration, making them very useful for multi-cohort projects.

In addition to `dsMLClient`, a significant contribution was made to extend the existing `dsSurvival` package. Originating from a specific need in the unCoVer study for pooled survival analysis, a new method based on survival tables was added. This extension not only allows for meta-analysis but also provides the capability for pooled survival analysis, thereby significantly enhancing the package's versatility and the scope of research that can be carried out using DataSHIELD.

On the following sections we will discuss the work what was done, pointing out the technical details of the different methods implemented.

7.2 Design and Development

Machine learning

The beginning of the `dsMLClient` package originated in the need for a more sophisticated range of methods that could enhance the DataSHIELD ecosystem's capacity. We recognized a growing demand among current and potential users for more advanced, yet accessible, methodologies within their projects. It became clear that there was a gap in the capabilities of DataSHIELD.

One of the available packages at the time was `dsSwissKnife`, developed by the Swiss Bioinformatics Institute. Despite its potential, it fell short in addressing some of the fundamental requirements of the DataSHIELD community. Specifically, it lacked support for pooled analysis methods, a very powerful feature for facilitating analysis across multiple servers. Furthermore, it failed to implement non-disclosure mechanisms, essential for preserving data privacy within the DataSHIELD infrastructure.

Motivated by the limitations of `dsSwissKnife` and driven by the need for a more comprehensive solution, we embarked on developing the `dsMLClient` package. Our strategy was to initially focus on implementing basic methods as a proof of concept. This would allow us to demonstrate the viability of the concept before committing to more complex functionality.

Even though this was a proof of concept phase, we still committed our efforts to make sure that there were non-disclosure checks and mechanisms. Making sure that the prototype is well suited to comply with the DataSHIELD standards. As a result, the final product is suitable to be considered to be used with real world data.

Survival analysis

The enhancement of the `dsSurvival` package was born out of both necessity and opportunity. While the original package had made strides in providing survival analysis capabilities within the DataSHIELD ecosystem, it was not without its limitations. Specifically, the package was initially tailored to meta-study survival

analyses, which although valuable, did not fully capitalize on DataSHIELD’s inherent strength—pooling data across multiple centers to achieve higher statistical power.

Enter the unCoVer study—a pivotal catalyst that accelerated the timeline of this development. While the unCoVer study initially reached out to us for aid in implementing pooled survival analysis, the need for such a functionality had been looming in our minds. The essence of DataSHIELD lies in its capability to facilitate secure multi-center analyses, a promise that could only be fully realized through pooled analyses.

The journey of incorporating this new feature was relatively streamlined in terms of time but presented its unique set of challenges. The core hurdle was ensuring the harmonious merging of non-disclosive data received from different study servers. Another layer of complexity was added as we entered debates about how the returned data to the client should be appropriately filtered to remain non-disclosive.

The result is a significantly augmented `dsSurvival` package that is capable of elevating any survival analysis that involves multi-study, harmonized data. This extension is not just a reactionary development to meet the immediate needs of a specific project; it represents a thoughtful expansion that underscores DataSHIELD’s commitment to both data privacy and rigorous, multi-faceted statistical analyses.

7.3 Features and Functionality

Machine learning

The `dsMLClient` package was designed with a blend of diverse methodologies to offer users a useful range of analytical tools within the DataSHIELD ecosystem. Four key methods form the core of this package: 1) Singular Value Decomposition (SVD), 2) k-means clustering, 3) k-nearest neighbors (kNN), and 4) Factor Analysis of Mixed Data (FAMD). These were selected based on their wide-ranging applications and potential to improve the quality and depth of analyses within the DataSHIELD ecosystem.

7.3.1 Singular Value Decomposition (SVD)

SVD is a matrix factorization method commonly used for dimensionality reduction, noise reduction, and the identification of underlying latent variables. The implementation of SVD offers users a powerful tool for handling high-dimensional data and uncovering patterns and relationships that might otherwise be missed.

7.3.1.1 Statistical method implemented

Implementing Singular Value Decomposition (SVD) in a distributed environment presents unique challenges, primarily due to the need to work across separate datasets while still extracting meaningful, unified patterns. To address this, a block method was utilized in `dsMLClient`[313].

The block method is a procedure that hinges on the division of the entire dataset into multiple sub-datasets or ‘blocks’. These blocks are processed individually, and a SVD is computed for each, yielding respective U (left singular vectors) and Σ (singular values). This approach allows for the partitioning of computational load across servers, thereby optimizing performance.

After the individual SVD computations, the results are merged into a composite matrix, essentially creating a condensed representation of the original data from across all servers. This merged matrix then undergoes a final SVD. The end product is a set of singular vectors and singular values that reflect the structure of the entire dataset, despite it being processed in blocks.

The steps can be summarized as follows:

1. **Divide the Dataset:** Partition the entire dataset into manageable ‘blocks’ or sub-datasets.
2. **Compute Individual SVDs:** Perform SVD on each sub-dataset, resulting in sets of U and Σ .
3. **Merge Results:** Combine all U and Σ into a single matrix that encapsulates the information from all sub-datasets.

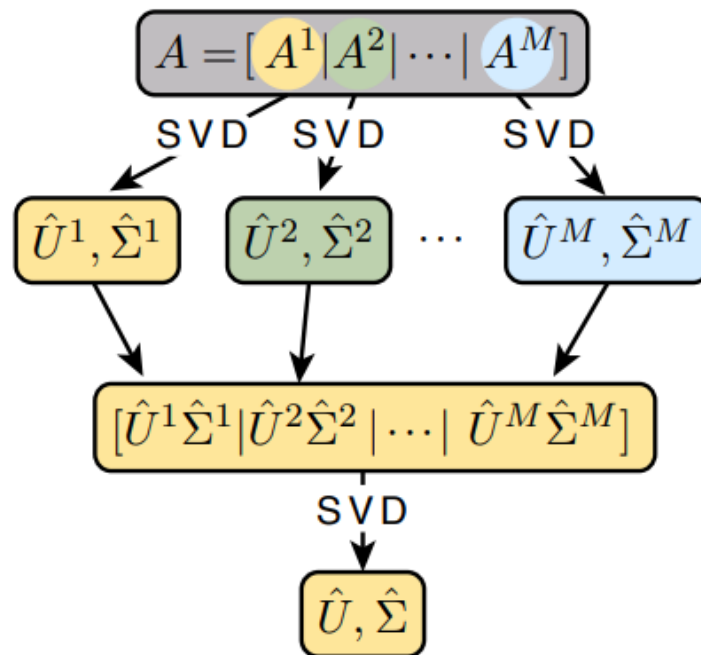


Figure 31: SVD block method schematic. Extracted from Iwen et al.[313]

4. **Perform Final SVD:** Execute an additional SVD on the merged matrix, resulting in a final set of U (left singular vectors), Σ (singular values), and V (right singular vectors).

This steps are illustrated on fig. 31.

For this SVD implementation, only the numerical columns of the datasets are considered, while non-numerical columns (e.g., factors, characters) are omitted from the analysis. This decision ensures the mathematical validity of the SVD, as the method is primarily designed to work with numerical data.

To accommodate datasets of varying sizes, the client-side SVD offers two versions: a full SVD and a truncated SVD. The full SVD is generally used for smaller datasets where computational resources are not a limiting factor. For larger datasets, where computing the full SVD might be resource-intensive, a truncated SVD is employed. The truncated SVD provides an approximation of the full SVD and significantly reduces computational load, thereby enabling analysis of larger datasets.

In sum, the block method implementation of SVD in the `dsMLClient` package presents a robust, privacy-preserving, and scalable solution for distributed data analysis within the DataSHIELD ecosystem.

7.3.2 k-means Clustering

k-means is an unsupervised learning algorithm often used to divide datasets into meaningful groups or clusters. The inclusion of k-means in the `dsMLClient` package equips users with the ability to explore and interpret the natural structure of their datasets, which can be crucial for a range of tasks including market segmentation, anomaly detection, and data pre-processing.

7.3.2.1 Statistical method implemented

For the implementation of k-means clustering within the `dsMLClient` package, a parallel k-means algorithm was used[314]. This approach builds on the traditional k-means algorithm, allowing it to efficiently work across distributed datasets, which is a key characteristic of the DataSHIELD ecosystem.

In a parallel k-means algorithm, the k-means clustering process is independently computed for each block or partition of the dataset (horizontally partitioned). This computation generates a set of 'local' centroids for each block. These local centroids are then transmitted to the client, where they are merged and averaged, taking into account the number of data points associated with each centroid.

The parallel k-means algorithm follows an iterative process. After the initial computation of local centroids and their subsequent merging, the new set of 'global' centroids is sent back to the servers. Each server then recomputes the k-means clustering based on these new centroids, and the process repeats. This iterative procedure continues until one of two stopping conditions is met:

1. The change in centroid positions (known as the learning rate) falls below a pre-specified threshold. This condition suggests that the centroids have largely stabilized and further iterations are unlikely to yield significant changes.
2. A maximum number of iterations is reached. This condition serves as a fail-safe to prevent the algorithm from running indefinitely in cases where the centroids continue to shift.

Here is a high-level description of the process (a detailed flowchart is available on fig. 32):

1. **Initial Centroid Selection:** Initial centroids are either randomly selected or specified by the user.
2. **Local k-means Calculation:** Each server independently applies the k-means algorithm to its local dataset, using the current global centroids.
3. **Centroid Aggregation:** Each server calculates the local centroids and the count of data points assigned to each centroid. This information is sent to the client.
4. **Global Centroid Calculation:** At the client side, the received local centroids and counts are used to compute the new global centroids. The global centroids are calculated as the weighted average of local centroids, with the weights being the number of data points assigned to each centroid.
5. **Centroid Update and Iteration:** The new global centroids are sent back to the servers and used for the next round of local k-means calculation. The process repeats until the stopping conditions are met.

From a data privacy perspective, the only information sent back to the client are the local centroids, the count of data points for each centroid, and the assignment labels. The local centroids are aggregated results that can't be traced back to the original data, preserving privacy. The counts are also non-disclosive as they merely provide a count of data points per cluster. Finally, the assignment labels, which indicate which cluster each data point belongs to, contain no intrinsic information about the data points themselves, further protecting privacy.

7.3.2.2 Results Visualization for k-means Clustering

In the `dsMLClient` package, k-means clustering results are visualized using a scatter plot enhanced with cluster ellipses. This graphical representation is a straightforward and effective way to display the results of the k-means algorithm.

Remember, the key to interpreting such a scatter plot is to consider not just the individual points, but also the larger patterns formed by the clusters and their ellipses. These patterns can provide key insights into the structure of your data and the effectiveness of your clustering.

An example of the visualization created by `dsMLClient` can be seen in fig. 33.

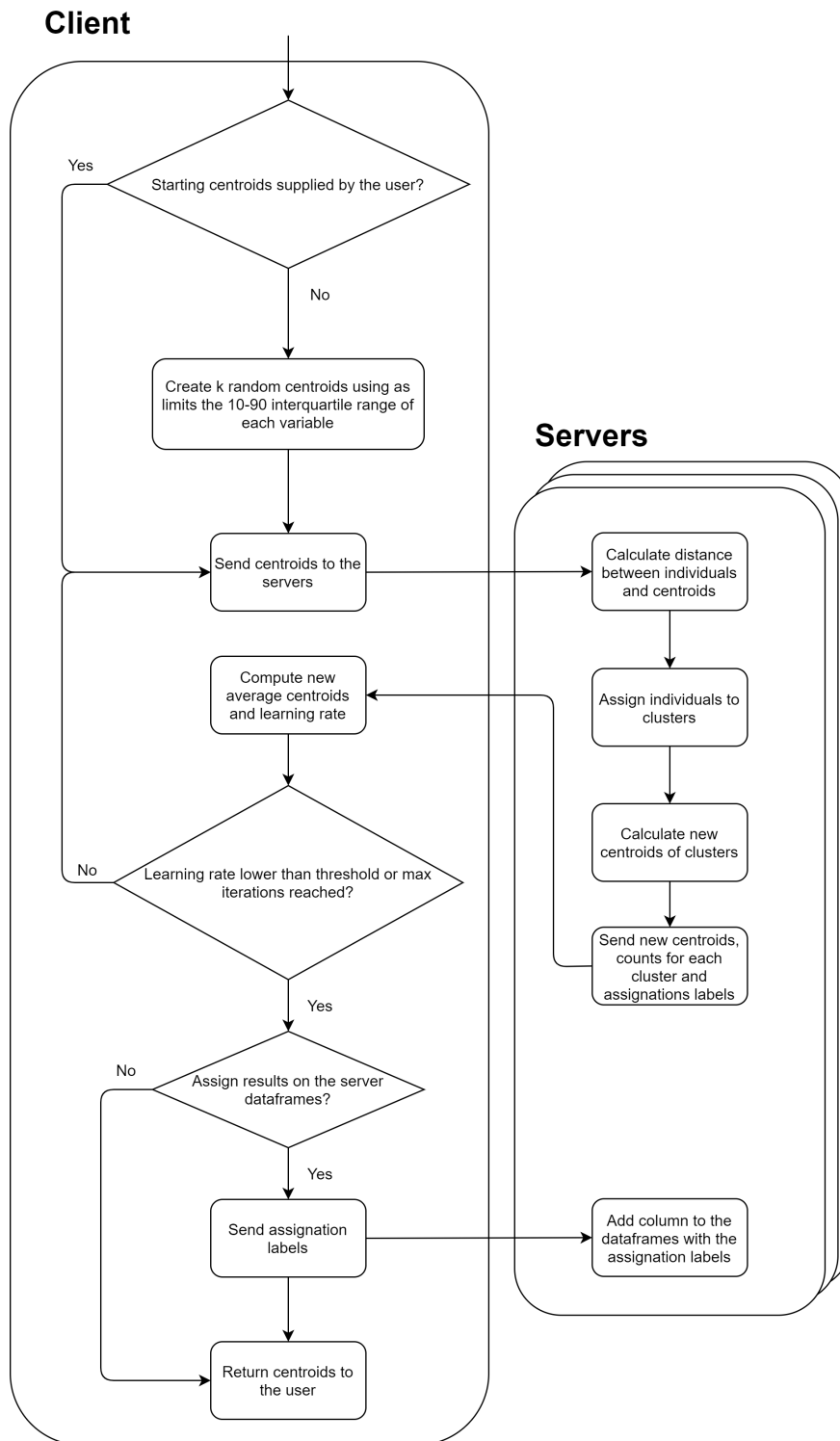


Figure 32: K-means flowchart

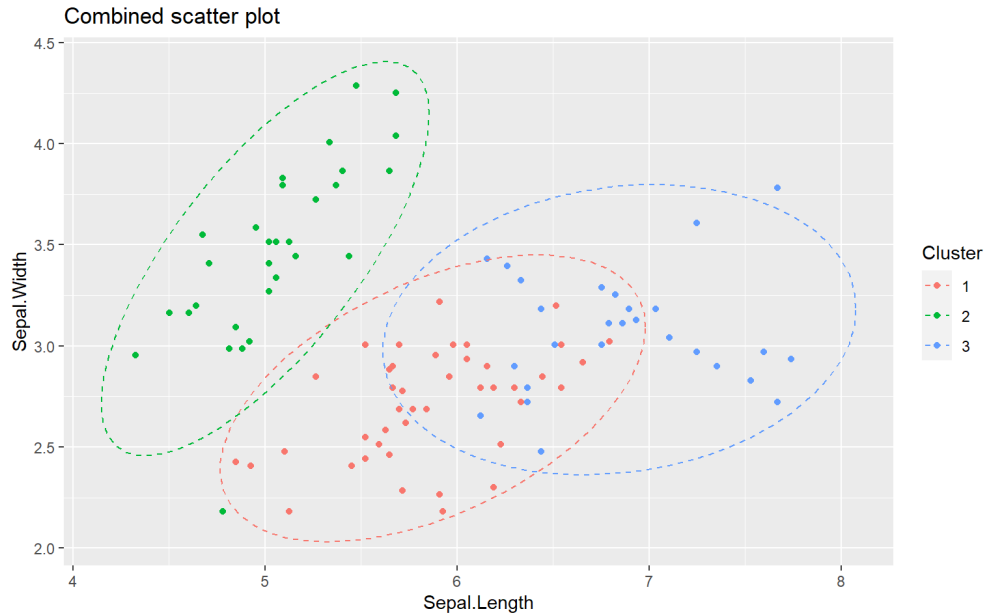


Figure 33: K-means scatter plot visualization

7.3.3 k-nearest neighbors (kNN)

kNN is a type of instance-based learning method that can be used for both classification and regression tasks. With kNN, users of `dsMLClient` can draw on their data's inherent structure to predict outcomes for new instances, based on the characteristics of 'neighboring' data points.

7.3.3.1 Statistical method implemented

In the `dsMLClient` package, a parallel version of the k-Nearest Neighbors (kNN) algorithm has been implemented, inspired by the work of Liang et al [315]. This version is adapted to work with distributed datasets in a manner similar to the implementation of the k-means algorithm.

The core idea behind the kNN algorithm is to classify a data point based on the labels of its 'k' nearest neighbors in the feature space. In the context of parallel computation across distributed datasets, this process involves several key steps (illustrated on fig. 34):

1. **Data Point Transmission:** A data point that needs to be classified is sent to all servers.
2. **Local kNN Calculation:** Each server finds the 'k' closest data points to the query data point within its local dataset, along with their classification labels.
3. **Results Aggregation:** The sets of 'k' closest neighbors and their labels from all servers are sent back to the client.
4. **Final Decision:** The client aggregates all the received neighbors and their labels. The final classification of the query data point is determined based on majority voting among the aggregated labels.

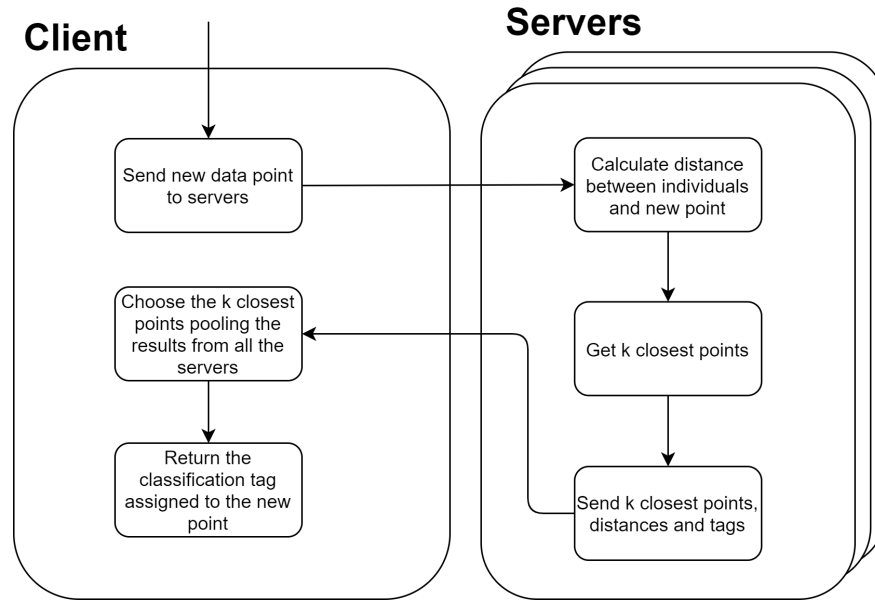


Figure 34: KNN flowchart

Unlike the k-means algorithm, there is no iterative process involved in the kNN algorithm. Once the final decision is made based on the aggregated closest neighbors, the classification for the query data point is finalized.

The implementation of the parallel kNN algorithm in `dsMLClient` employs a privacy-preserving method to mitigate potential data disclosure risks. This approach addresses a specific risk in the kNN algorithm: if a user queries a data point that exactly matches a point in the dataset, the returned distance from the server will be zero, revealing that an exact match exists in the dataset.

To circumvent this disclosure risk, the implementation utilizes a similar anonymization process as used in the `scatterPlotDS`¹ function. The principle idea is to add noise to the dataset when computing the distance. This additional step ensures that the client can never definitively confirm whether a queried point exactly matches a point in the dataset.

However, the introduction of noise brings a trade-off: while it reduces the risk of data disclosure, it introduces a degree of uncertainty into the classification process. The altered distances due to noise addition can potentially affect the identification of 'k' nearest neighbors and, consequently, the final classification decision.

Despite this trade-off, the privacy-preserving implementation of kNN within the `dsMLClient` package offers a balance between maintaining data privacy and enabling distributed classification tasks.

7.3.4 Factor Analysis of Mixed Data (FAMD)

FAMD is a multivariate data analysis method suitable for datasets with a mixture of continuous and categorical variables. FAMD enables users to explore relationships between different types of variables and to reduce dimensionality without losing critical information inherent in the data.

7.3.4.1 Statistical method implemented

The implementation of Factor Analysis of Mixed Data (FAMD) in the `dsMLClient` package is based on the method proposed by Jérôme Pagès[316]. FAMD is a versatile technique that allows the simultaneous analysis of both quantitative (continuous) and qualitative (categorical) variables.

¹<https://github.com/datashield/dsBase/blob/v6.2/R/scatterPlotDS.R>

The steps involved in the method are as follows (illustrated on fig. 35):

1. **Dummy Encoding:** Categorical variables are converted into a set of binary (dummy) variables, each representing a unique category within the original variable. This process is known as one-hot encoding or dummy variable encoding.
2. **Standardization of Continuous Variables:** Each continuous variable is standardized by dividing it by its standard deviation. This ensures that all continuous variables have the same scale and that none dominates the analysis due to its numerical scale.
3. **Normalization of Dummy Variables:** Each dummy variable is divided by the square root of its proportion (p_j), where p_j is the proportion of individuals that take the category j . This normalization is done to make the scale of the dummy variables comparable to that of the continuous variables.
4. **Principal Component Analysis (PCA):** PCA is performed on the resulting matrix, which includes both the standardized continuous variables and the normalized dummy variables. The PCA will yield principal components that are linear combinations of all the variables (both continuous and categorical).

Survival analysis

The new extension employs actuarial life tables to facilitate pooled survival analyses. This methodology is particularly well-suited for examining survival probabilities at fixed time intervals, making it extremely practical for a wide range of survival analysis applications, such as monitoring the survival rates of cancer patients over time.

In essence, the methodology involves the construction of a life table based on two key inputs: time intervals and events. Time intervals define the periods within which events can occur, while events (e.g., patient death) represent the outcomes of interest in the survival analysis. Importantly, this life table can be independently constructed at each server, which is essential for achieving pooled analysis capabilities across different study centers.

The steps involved in the method are as follow (illustrated on fig. 36):

1. **Survival object creation:** By using the already laid out `dsSurvival` package, a survival object has to be created. On the creation of this object, the objective event and stratification variables can be specified.
2. **Extract data of interest:** From the survival object created, only some data of interest has to be extracted, that being the times, individuals at risk (still at the study for a given time) and individuals with an event.
3. **Merge data on the client:** Join the data by time points by adding the values.
4. **Compute survival probability:** Once the data has been merged, it is trivial to obtain the probability of survival (given eq. (16) and eq. (17)).

$$\text{Survival rate} = 1 - \frac{\text{Ids with event}}{\text{Ids at risk}} \quad (16)$$

$$\text{Probability of survival} = \text{cumprod}(\text{Survival rate}) \quad (17)$$

The primary challenge encountered on the development of this method was a discussion whether returning the number of individuals with an event on a time frame is disclosive; the discussions concluded that it is not disclosive, although from the DataSHIELD community would be of great interest on this particular matter.

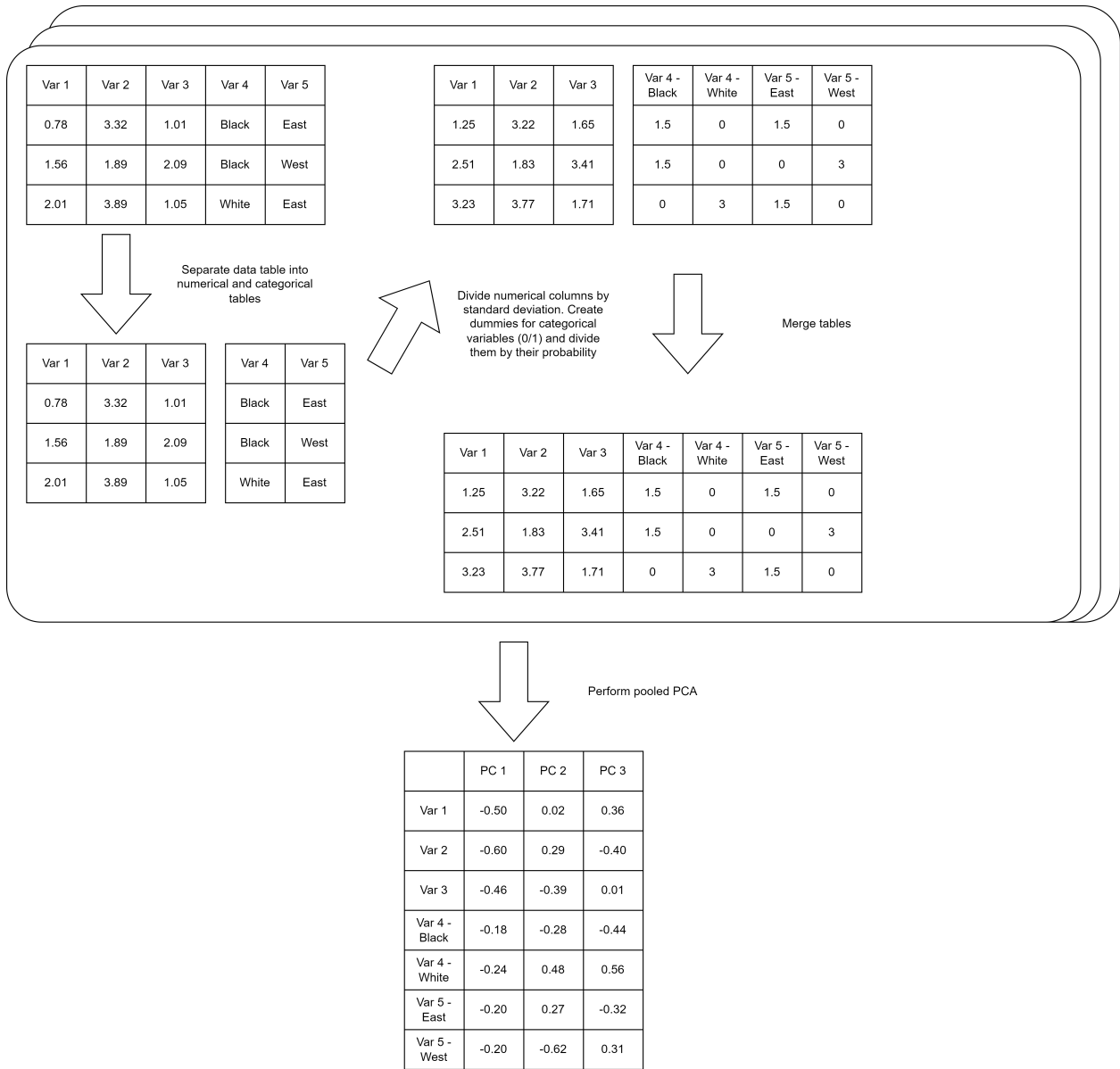


Figure 35: FAMD flowchart

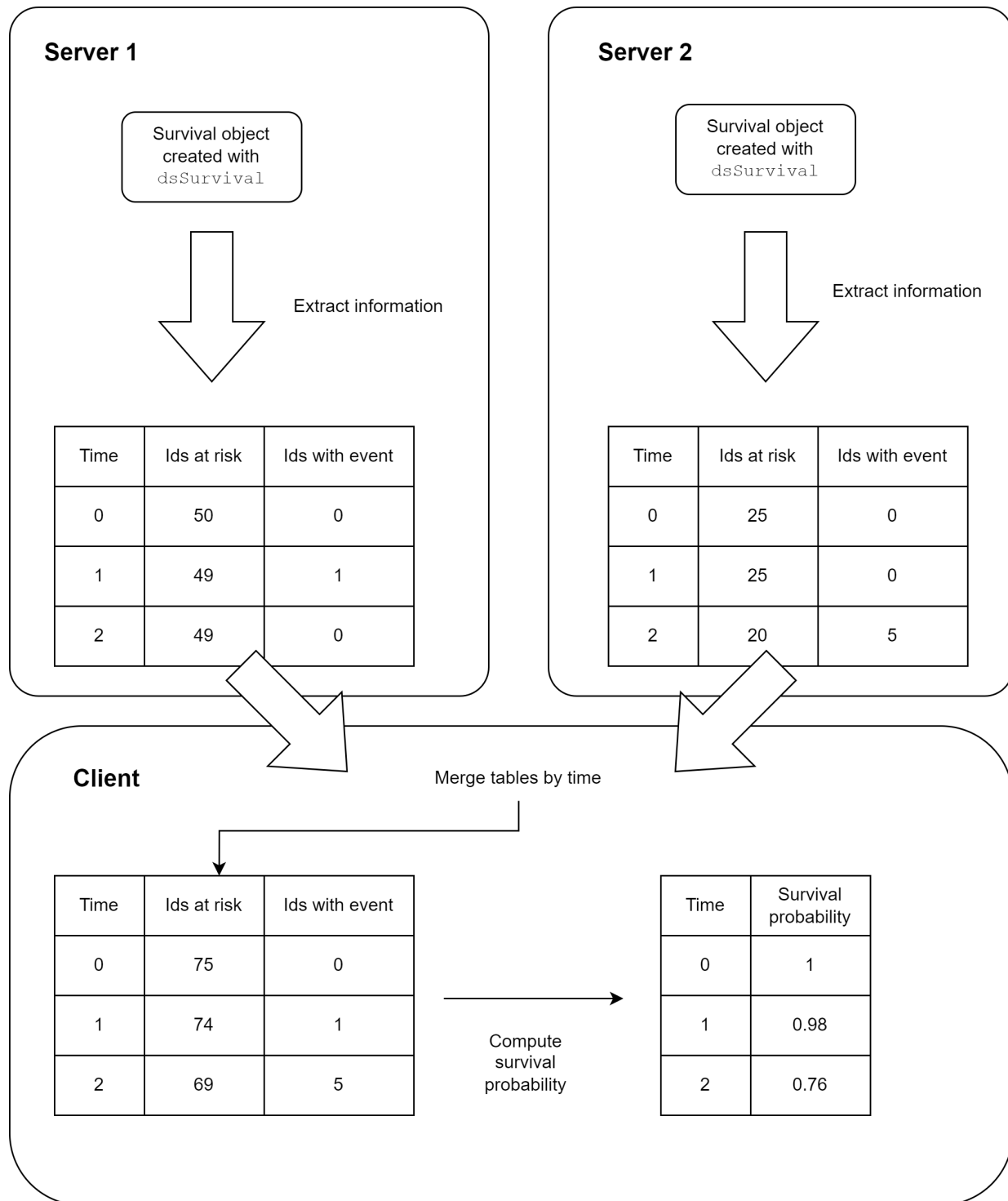


Figure 36: Life tables flowchart. Exemplified with two different servers performing a pooled survival analysis.

7.4 Current status

Machine learning

As of today, `dsMLClient` stands as a successfully developed proof of concept, with a firm focus on data privacy and functionality. Its development has demonstrated how various machine learning methodologies can be adapted and implemented in the DataSHIELD infrastructure for pooled analysis across different servers.

`dsMLClient`'s development and testing phase has yielded accurate results, giving confidence in the correctness of its implemented methods. The package incorporates a diverse range of methodologies, which have been carefully chosen to accommodate a wide range of analysis providing comprehensive functionality.

At its current stage, `dsMLClient` could be deployed in real-world projects. Its capabilities are such that it can serve users who need to conduct analyses across distributed datasets; as long as they are comfortable with the implemented methods for mitigating data disclosure risks.

Survival analysis

At the time of writing the manuscript of this thesis, the developments have been submitted as a pull request to the original repository containing the 'dsSurvival' package. An open communication is being maintained with the original creator of the package, Dr. Soumya Banerjee, which whom I have collaborated before.

Moreover, on the upcoming DataSHIELD 2023 conference I have been approved for a talk about this pooled survival method. On the conference I expect to withstand further discussions with DataSHIELD members about the method, its security and whether it has any hidden flaws I might have missed.

7.5 Future developments

Machine learning

As we move forward, the future of `dsMLClient` could be very promising, with opportunity of growth and development. There is already interest from various researchers and projects keen on utilizing its current functionalities. These preliminary expressions of interest indicate that the package, in its existing form, can begin to serve a wider user base, fulfilling real-world analysis needs.

However, the journey of `dsMLClient` is far from complete. Future work on the package will be guided by two main priorities.

First, there is a need for more thorough assessment of the data disclosure risk mitigation mechanisms currently in place. Although the prototype was primarily developed with a focus on non-disclosure mechanisms, as the package moves into wider use, the robustness of these mechanisms will be a critical aspect to ensure the privacy of the data used. The plan is to undertake rigorous testing and validation of these mechanisms, refining and enhancing them as necessary to ensure they effectively maintain data privacy while facilitating meaningful analysis.

Second, there is a need to broaden the scope of machine learning methodologies available in the package. While the package already supports a variety of methods, there is room to incorporate more advanced techniques, making the tool even more versatile and valuable to researchers. The selection of these new methods will be informed by the needs and feedback of the users, as well as the latest developments in the field of machine learning.

In sum, the future of `dsMLClient` will be one of evolution and refinement. There is a firm commitment to making the package a great tool for DataSHIELD users.

Survival analysis

Upon discussing over the package and its security on the DataSHIELD 2023 conference, it is to be expected the new method to be merged on the original 'dsSurvival' repository. Following that no further developments

are foreseen for this package.

8 Application to real world data

8.1 unCoVer project

8.1.1 Project description

The unCoVer project, an initiative named for its mission to "Unravel Data for Rapid Evidence-Based Response to COVID-19," emerged as a collaborative scientific project in the middle of the COVID-19 pandemic. The project was initiated in November 2020 and was set to span two years until November 2022. At its core, the unCoVer project represents a dynamic and responsive alliance, coordinated by the Prins Leopold Instituut voor Tropische Geneeskunde in Belgium, consisting of 29 dedicated partners across 18 countries worldwide.

The aims of unCoVer are centered around harnessing the power of real-world data derived from the health systems' response and patient care during the COVID-19 crisis across Europe and beyond. UnCoVer's efforts involve synchronizing research on a global scale to effectively combat the ongoing COVID-19 pandemic. The project seeks to streamline access to and usage of COVID-19 related real-world data, capitalizing on the potential of data being generated routinely and reflecting common medical practices.

Moreover, unCoVer aims to identify potential data gaps and underrepresented populations, gathering information with existing and planned COVID-19 related clinical databases. It provides an innovative platform for the aggregation of dissimilar data sources, anticipating the needs for data harmonization and addressing ethical and legal considerations.

Leveraging expertise in advanced computational, epidemiological and biostatistical methods, unCoVer is designed to handle the complexity of heterogeneous and multi-layered information. This allows for rapid queries and the generation of robust findings related to various facets of COVID-19, including SARS-CoV-2 infection, prognosis determinants, treatment safety and effectiveness, as well as the disease's impact on health system resources.

unCoVer aspires to extend the use and results of the platform, inviting new partners with existing similar networks on both European and international levels. The ultimate goal is to maximize the project's impact in saving lives and optimizing resources in the fight against COVID-19. It is through this objective that unCoVer's network works towards a future where the world is better equipped to tackle public health emergencies like the COVID-19 pandemic.

8.1.2 Methodology and Ethical Considerations

The unCoVer project, funded by Horizon 2020, employs methodologies and ethical guidelines in the objective to uncover comprehensive insights into the COVID-19 pandemic. This network, consisting of 29 partners from 18 countries, collects and uses real-world data (RWD) derived from the care and response of health systems to COVID-19 patients across Europe and internationally. unCoVer's aim is to utilize the full potential of this information to rapidly address the pressing clinical and epidemiological research questions that the pandemic continually presents.

From the start of the COVID-19 pandemic, the partners have been gathering RWD from electronic health records, which now includes information from over 22,000 hospitalised COVID-19 patients. Additionally, the project has also information from national surveillance and screening data, and registries with over 1.9 million COVID-19 cases across Europe. This data, subject to continuous updates, represent a very powerful resource for analysis.

The diverse datasets are meticulously catalogued, harmonised, and integrated into a multi-user data repository operated through Opal-DataSHIELD, an interoperable open-source server application. Federated data analyses are conducted without sharing or disclosing any individual-level data. The primary objective of these analyses is to reveal patient baseline characteristics, biomarkers, determinants of COVID-19 prognosis, the safety and effectiveness of treatments, potential strategies against COVID-19, and epidemiological patterns.

These analyses serve to supplement evidence from clinical trials, which often exclude more complex, heterogeneous populations and those most at risk of severe COVID-19. This ensures that unCoVer's insights are inclusive and representative of the diverse global population affected by the pandemic.

Ethics are central to unCoVer’s planification. Databases are available through a federated data analysis platform that processes available COVID-19 RWD without disclosing identification information to analysts, and it limits output to data aggregates. The dissemination of unCoVer’s activities, which include the access and use of diverse RWD and the results generated by pooled analyses, take place through training and educational activities, scientific publications, and conference communications, thereby maximizing the impact of the project’s findings.

The overall project organization is summarized on fig. 37.

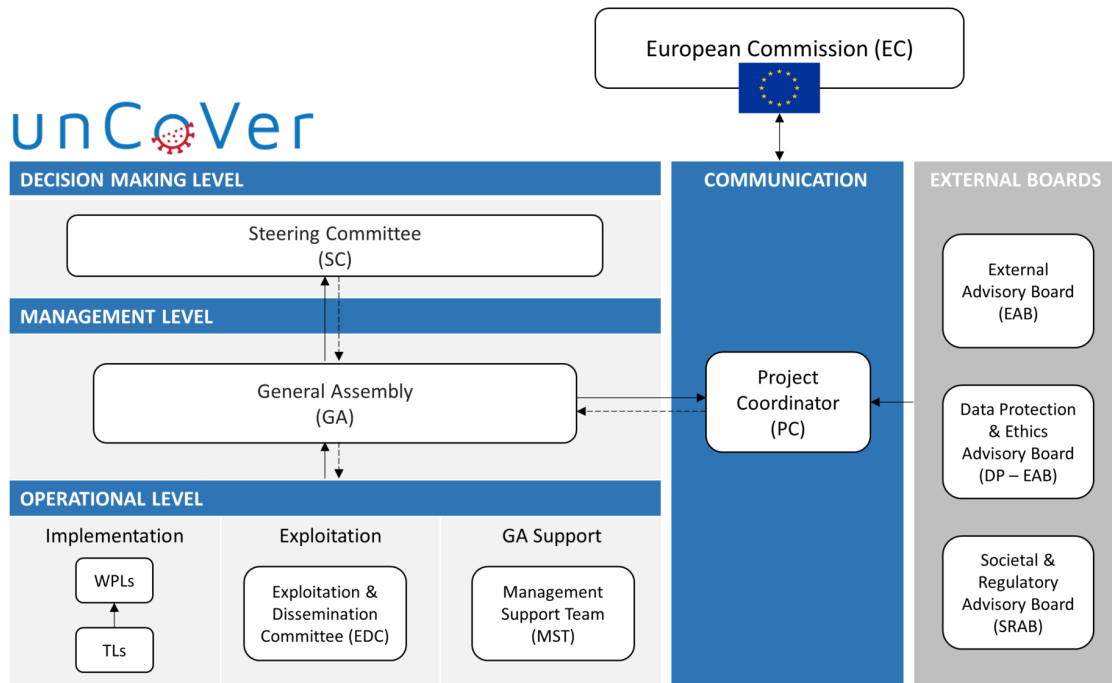


Figure 37: unCoVer organization. Taken from <https://uncover-eu.net/mission>

8.1.3 Tasks developed

As an associate member of the unCoVer project, I developed different tasks to contribute to the analysis of the data and the overall development of the project.

8.1.3.1 Application Development and Utilization

One of my significant contributions to the unCoVer project is the design and development of an R Shiny application, a data handling solution crafted with the objective of streamlining the data analysis process. This application is designed as a user-friendly interface for exploiting the data gathered during the project. Its origins are based on the research I conducted and compiled into the ShinyDataSHIELD paper, a piece of work previously presented on this thesis manuscript.

This application’s practicality lies in its ability to heighten the efficiency of data usage within the project. By breaking down barriers between valuable data and the team members who need access to it, we managed to optimize our approach to comprehensive data analysis and interpretation. This tool facilitated a smoother data navigation experience, making data interpretation less complicated and more accessible to all members of the project, regardless of their computational expertise. A couple of screenshots of the application are on fig. 38 and fig. 39.

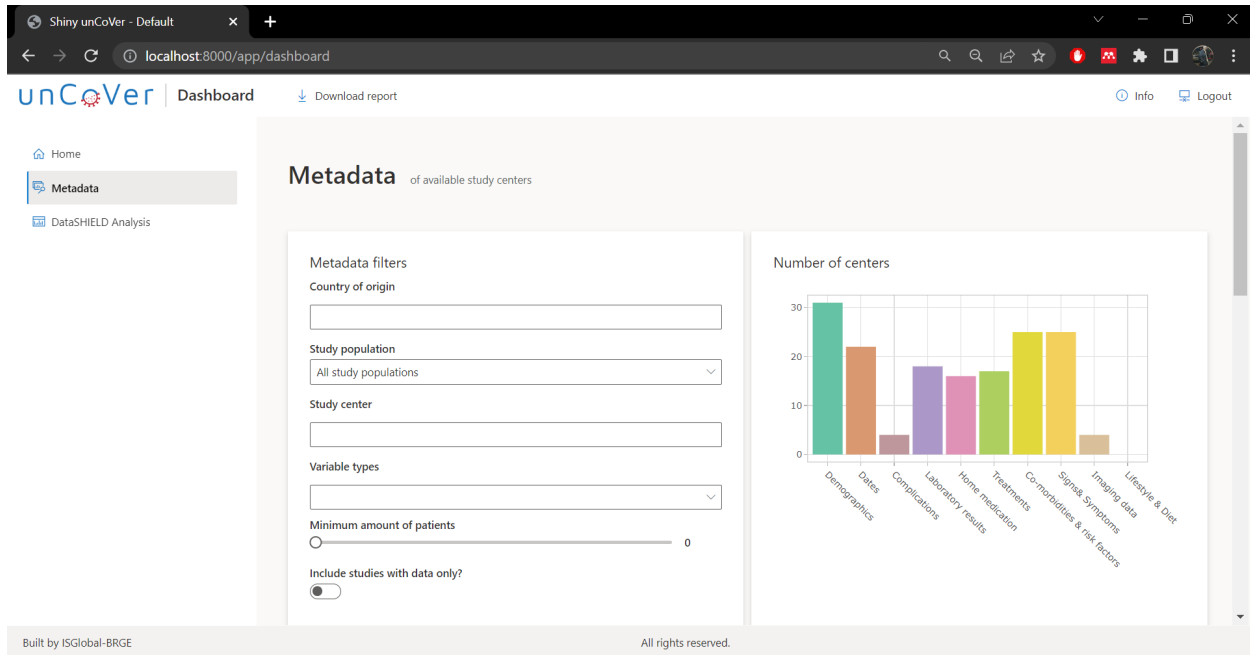


Figure 38: unCoVer dashboard. Metadata page

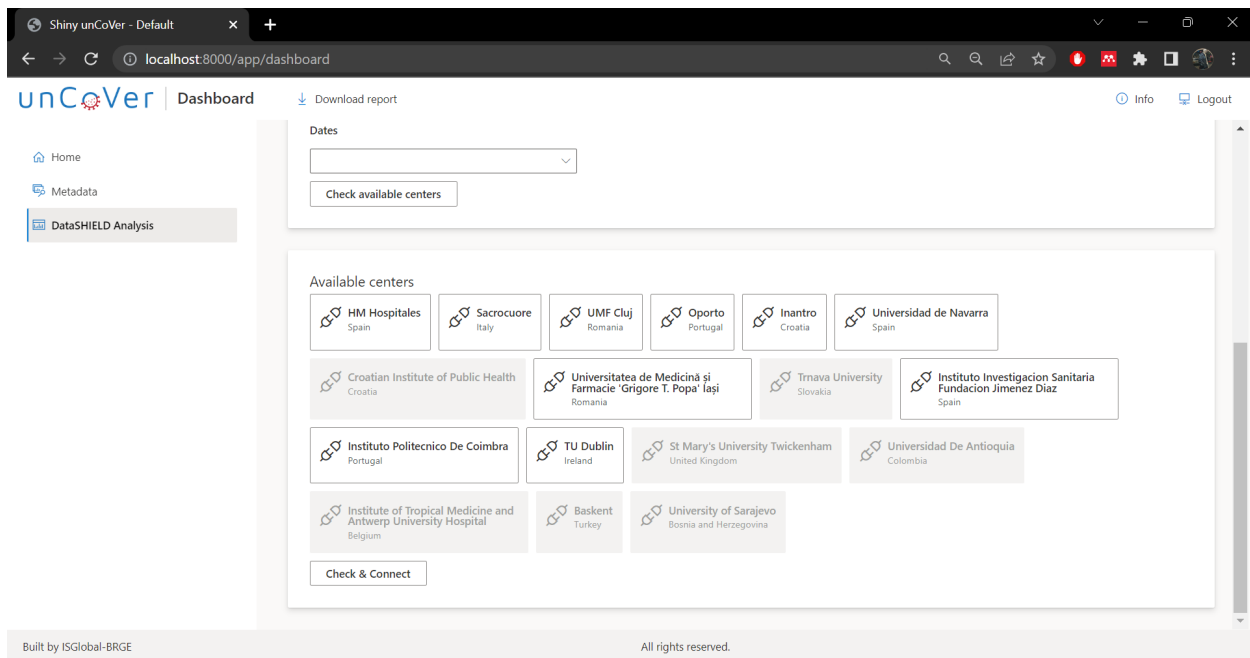


Figure 39: unCoVer dashboard. Server connection page

8.1.3.2 Analytical Workshops and Skills Transfer

Beyond the development of the R Shiny application, my contributions to the unCoVer project also extended to skill transfer and education. Recognizing the importance of empowering the project's analysts with the knowledge and skills necessary to effectively utilize the DataSHIELD infrastructure, I developed and delivered a series of analysis workshops. The material has been made openly available to be used as a reference by

other projects, it can be found at https://isglobal-brge.github.io/UnCoVer_workshop/.

These workshops were conducted in-person, creating an environment for real-time interactions and hands-on resolution of any arising problems or questions. The Universidad Politécnica de Madrid (UPM) served as the venue for these workshops on two separate occasions. The workshops were delivered in collaboration with Juan R. González, director of this thesis.

8.1.3.3 Impactful Data Analysis

The final significant aspect of my involvement in the unCoVer project was conducting data analyses. Utilizing DataSHIELD, I studied the effects of comorbidities on the mortality rate of COVID-19. This exploration involved several analytical strategies, including exploratory data analysis, linear regression analysis, and survival analysis.

These methodologies provided a detailed understanding of the subject, shedding light on the relation between comorbidities and the subsequent impact on patient COVID outcome.

8.1.4 Development and Improvement of DataSHIELD Packages

To achieve the objectives of the unCoVer project, and at the same time improve the DataSHIELD infrastructure, I developed and improved a couple of DataSHIELD packages. These packages were key on handling the project's data and enabled more sophisticated analyses.

8.1.4.1 Development of dsDates Package

The first package, named `dsDates`, was developed to manage and operate on date class variables. This functionality was crucial for the unCoVer project, as it has data comprised a multitude of dates related to patient care, such as admission and discharge dates from hospitals. These variables required proper manipulation, particularly for computations like determining the length of a patient's hospital stay - a critical factor in survival analysis. The `dsDates` package facilitated these operations, providing a reliable tool for working with date-related data.

8.1.4.2 Enhancements to the dsSurvival Package

Additionally, the project called for improvements to an existing package - `dsSurvival`. This package, designed for conducting survival analyses within the DataSHIELD infrastructure, needed enhancements to meet the project's specific needs.

The first enhancement involved improving the visual representation of survival curves. The base R package initially plotted these, but for publication purposes, a more polished and comprehensive visualization was needed. To fulfill this requirement, I implemented improvements using the `ggplot2` package, elevating the quality and clarity of survival curve visualizations.

The second enhancement was the use of stratified variables in survival analysis (for instance, analyzing survival rates based on gender). While this feature was technically available in the package, it was compromised by a bug. I addressed this issue, resolving the bug and enabling reliable stratified survival analysis.

Finally, the third enhancement involved the implementation of a method for pooled survival analysis. Prior to this, the package was only capable of conducting server-wise (meta) survival analysis. Recognizing the potential of pooled survival analysis, particularly given the unCoVer project's vast data distributed across multiple servers, I developed a method using tables of survival. This new functionality enables pooled survival analysis, allowing for data from various servers to be analyzed as if contained within a single table. This has significantly increased statistical power, yielding more robust results and advancing our understanding of COVID-19 survival rates.

8.1.5 Academic outcomes and contributions

The unCoVer project has led to the creation of substantial academic output, a testament to its significant research. One notable outcome is the formulation of a peer-reviewed paper, provisionally titled "Cardiometabolic comorbidities and COVID-19 outcomes: a case-example of federated learning of real-world data from hospitals across Europe".

In this paper, the results of my extensive DataSHIELD analyses, as detailed earlier, form the analysis foundation. These analyses investigate the impact of cardiometabolic comorbidities on the outcomes of COVID-19, showcasing the power of federated learning to gain insights from real-world data collected from various hospitals across Europe.

As of now, the manuscript is in its developmental stage, with rigorous efforts being made to refine its content and structure. Upon completion, it will be submitted to a scientific journal for consideration of publication, extending the reach of the unCoVer project's findings. The specific journal for submission is yet to be decided, with the selection process aiming to ensure maximum visibility and impact within the scientific community.

The culmination of these results into a peer-reviewed paper marks a significant academic contribution, enhancing the existing body of knowledge on COVID-19 and providing valuable insights that could influence future research and clinical practices.

8.2 ATHLETE project

8.2.1 Project description

The Advancing Tools for Human Early Lifecourse Exposome Research and Translation (ATHLETE) project is a comprehensive initiative focusing on the systematic analysis of the exposome - the totality of environmental exposures encountered from conception onwards - during the early life stages from early pregnancy through adolescence. The primary objective of ATHLETE is to establish a toolbox of exposome tools and an Europe-wide exposome cohort that will evaluate the impact of a wide array of environmental risk factors on mental, cardiometabolic, and respiratory health outcomes, along with the associated biological pathways.

This project utilizes the data and resources of 16 existing longitudinal population-based birth cohort studies across 11 European countries, with approximately 80,000 mother-child pairs. This range of groups effectively illustrates the variety within European communities, while also guaranteeing a comprehensive coverage of existing exposome data.

The project also incorporates newly established birth cohorts that offer improved sampling strategies for exposure assessment and advanced outcome assessments. These new cohorts enable the evaluation of new chemicals that have been produced in high volumes more recently.

The project consists of three interlinked components focusing on data and tools, evidence, and translation. Specific efforts are directed towards creating a findable, accessible, interoperable, reusable (FAIR) data infrastructure, developing advanced statistical and toxicological strategies for analyzing complex multidimensional exposome data, and implementing intervention strategies to improve early life urban and chemical exposomes. Moreover, the project aims to translate the resulting evidence into policy recommendations and prevention strategies.

Ultimately, ATHLETE aspires to generate a substantial body of knowledge and tools that will be instrumental in better understanding and preventing health damage from environmental exposures. All data, tools, and results will be assembled in an openly accessible toolbox, providing a valuable resource for researchers, policymakers, and other stakeholders well beyond the duration of the project.

The work package structure of the ATHLETE project is illustrated on fig. 40.

8.2.2 DataSHIELD Methods in Work Packages 1, 3 and 4

Work Package 1 (WP1) focuses on the construction of a FAIR (Findable, Accessible, Interoperable, Reusable) data infrastructure for the ATHLETE Exposome cohort. This involves gathering exposome data from disparate sources into an open-access platform that is easy to navigate for researchers. One significant addition to this infrastructure is the integration of the HELIX subcohort data. To ensure interoperability of the exposome data, harmonization protocols have been implemented, which standardize the exposome variables across all cohorts.

Data access is organized into a federated and a centralized system. For the federated access protocol, DataSHIELD is employed. This tool allows for remote data analysis without the need for data sharing or release, thereby circumventing potential governance restrictions or data access delays. DataSHIELD enables access from the R statistical environment using MOLGENIS or Opal software.

Work Package 3 (WP3) is primarily concerned with the development of exposome data analysis tools. This involves tackling analytical challenges in exposome research, including estimating combined effects of exposures, integrating exposome and cross-omics data, and incorporating a priori knowledge on causal structures and mediators to enhance causal inference.

Importantly, WP3 aims to expand DataSHIELD tools for remote and non-disclosive data analysis to better accommodate exposome data visualization and analysis. New functionalities have been developed to handle complex big data, within DataSHIELD through the Opal (and MOLGENIS) data warehouse.

Work Package 4 (WP4) centers on finding the biological pathways from the exposome to health. Leveraging omics technologies, WP4 aims to understand early, preclinical perturbations of biological pathways in

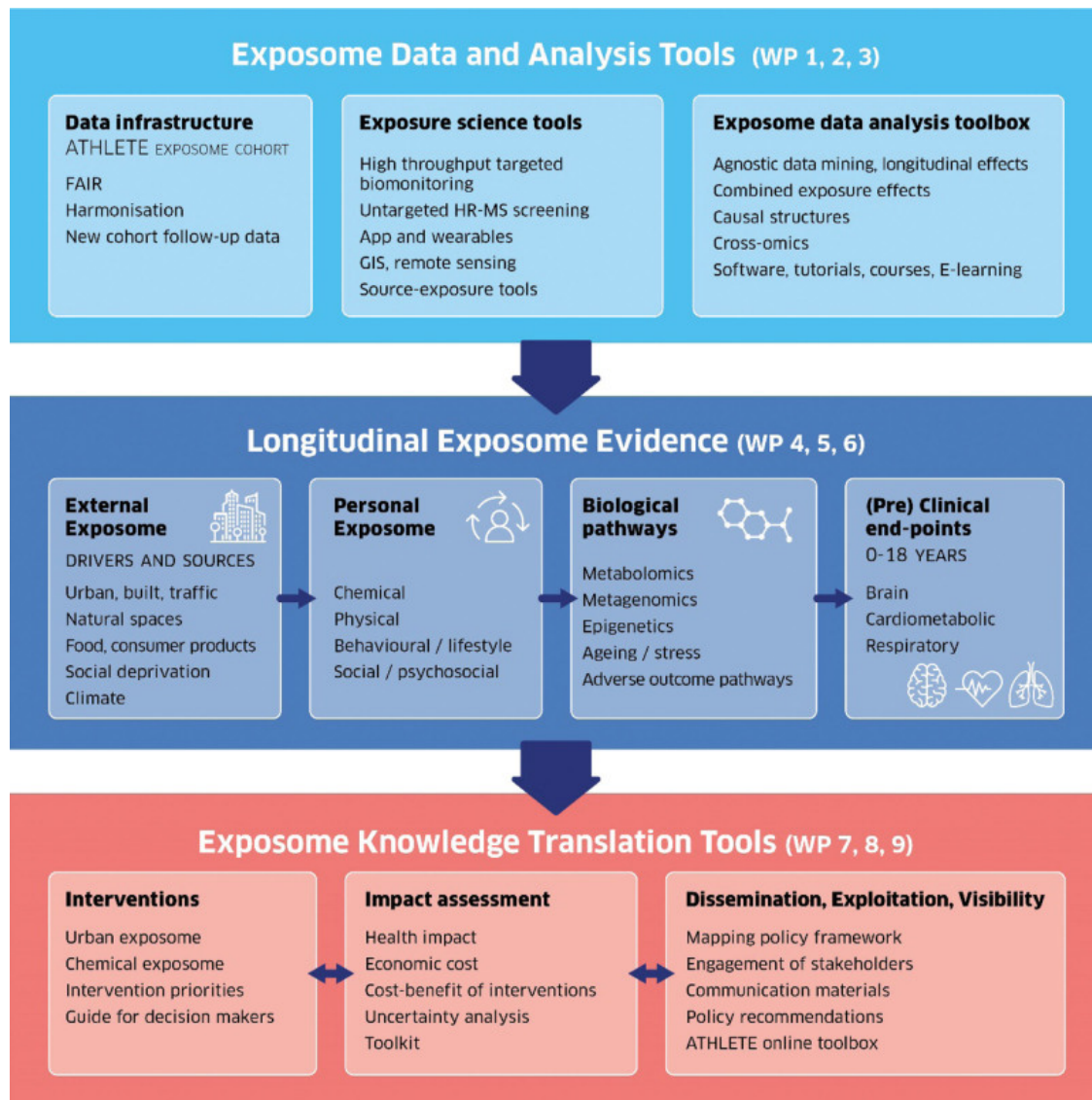


Figure 40: ATHLETE project components. Extracted from [88]

response to environmental exposures.

DataSHIELD will be extended to handle omics data. In addition, open-source software, front-end applications, tutorials, and e-learning materials will be developed to make these analytical tools more accessible to a broad audience of researchers.

8.2.3 Tasks developed

The ATHLETE project has been a comprehensive and collaborative effort involving multiple parties. My role in this project has been primarily concentrated in the areas of software development, technical support, infrastructure maintenance and infrastructure deployment.

8.2.3.1 Deployment and Maintenance of the Opal Node at ISGlobal

One of the critical tasks I executed was the deployment and maintenance of the Opal node at ISGlobal. This data infrastructure plays a pivotal role, as it hosts the INMA data of the ATHLETE project. Overseeing the setup and continued functioning of the Opal node required an in depth understanding of data hosting needs and procedures of the project. This task also demanded ongoing troubleshooting and problem-solving to ensure that data access remained smooth and consistent.

8.2.3.2 Assistance to Researchers and Enhancement of dsExposome

Working closely with researchers has been a key part of my role. Specifically, I've provided technical assistance to the researchers making use of the `dsExposome` package to analyze exposome data through DataSHIELD. In response to their specific analytical needs, I helped to expand the functionalities of this tool, streamlining their workflow and improving the efficiency of their data analysis processes. The modifications implemented in `dsExposome` were meant to cover the unique needs of exposome analyses in the ATHLETE project.

8.2.3.3 Collaboration and Support in dsOmics Utilization

Furthermore, my work extended to the use of the `dsOmics` package in the project. This task involved providing support to researchers, similar to my role with `dsExposome`. I also implemented additional functionalities in response to researchers' specific needs for their analyses. A notable part of this task was my collaboration with Sofía Aguilar Lacasaña, who utilized the package for her paper's analysis. Her efforts effectively served to validate the `dsOmics` package results in a multi-cohort scenario, contributing significantly to the software's overall success.

8.2.3.4 Development of an Experimental Version of the limma Package

Lastly, I undertook an innovative task to enhance the efficiency of omics analysis within the project. Given the constraints of the DataSHIELD infrastructure, where analyses extending over 30 minutes encountered issues, there was a need to optimize computational speed. To address this, I developed an experimental version of the `limma` package, a popular tool for the analysis of gene expression data in the R environment. Through parallelization, this experimental modification allowed a 30x speedup of computations, drastically improving the efficiency and feasibility of extensive omics analyses in the ATHLETE project.

8.2.4 Academic outcomes and contributions

The ATHLETE project has significantly contributed to the scientific community through a series of high-quality research outputs and advancements in data analysis tools. This body of work has broadened our understanding of the exposome and its associations with health outcomes, thereby influencing both public health and epidemiology fields.

8.2.4.1 Published Papers

A substantial part of the project's academic outcomes includes a series of papers published in various scientific journals. These papers encompass a wide range of topics related to the exposome, demonstrating the depth

of the research conducted within the ATHLETE project.

For instance, several papers focused on the development and validation of analytical tools for exposome data. This includes papers on the `dsExposome` and `dsOmics` packages, which discuss how these tools can enhance the analysis of complex exposome and omics data. Furthermore, details on the optimization of the `limma` package for efficient omics analysis within DataSHIELD infrastructure represents an important contribution to computational biology and bioinformatics.

Moreover, numerous papers have presented findings on the relationships between environmental exposures and health outcomes, shedding light on important public health issues. They have highlighted the role of early life exposures on later life health and the role of the exposome in disease risk prediction. Additionally, the inclusion of the gut microbiome as a new omics layer in exposome research is a significant advancement in the field.

8.2.4.2 Thesis Contributions

The different manuscripts presented along this thesis are all part of the work and efforts that have been placed on the ATHLETE project. They provide the software foundations that have been used along the project to implement the DataSHIELD infrastructure and analysis.

In summary, the academic outcomes of the ATHLETE project have greatly enriched scientific literature on the exposome, as well as enriching the DataSHIELD analysis capabilities. The advancements made in this project hold great potential to influence future research directions in environmental health and epidemiology.

8.2.4.3 Past Contributions

In addition to the above-mentioned contributions, the ATHLETE project also encompasses significant work conducted during my Master's thesis, which was also undertaken at ISGlobal. This research culminated in the publication of a paper also part of the ATHLETE project's academic output.

The paper represents a contribution to the study of the exposome. It illustrates my early engagement with the field and forms the foundation upon which my subsequent research activities in the ATHLETE project were built. It was through this initial study that I gained valuable insights into the complexities of exposome research and the analytical challenges it presents.

9 Discussion

9.1 Introduction

The primary focus of this research is distributed across three different fronts. First, it aims to catalyze the adoption of a specialized non-disclosive, open-source technology known as DataSHIELD. Second, it seeks to extend the DataSHIELD ecosystem by implementing new methods and packages. These enhancements enable researchers to work with novel types of data, such as omics and exposome data, while ensuring a non-disclosive analysis approach. Lastly, the study is designed to create user-friendly tools that make interaction with DataSHIELD more accessible to both new and experienced developers. This layered approach addresses key barriers to the wider application of DataSHIELD technology.

The need for such advancements is underlined by the challenges that researchers currently face in the field of scientific data analysis, particularly when dealing with sensitive genomics data. Traditional methods often require cumbersome data sharing agreements, which can slow down the pace of discovery and limit collaborative efforts. The non-disclosive features of DataSHIELD can alleviate these issues, enabling a higher quality and more democratic research landscape.

In addition to offering a more streamlined and secure way to handle sensitive data, DataSHIELD also brings an added layer of credibility to scientific research. Its non-disclosive nature assures that the data of individual participants remains confidential. Moreover, it allows for greater transparency in research findings. Reviewers and readers alike can more easily reproduce results and validate that the published findings indeed address the research question comprehensively.

Another significant advantage is the speed and efficiency that DataSHIELD technology can bring to research projects. Traditional multi-center studies often require laborious coordination between different data managers, delaying the testing of new hypotheses. DataSHIELD enables researchers to conduct multi-center analyses right from their own computers, significantly speeding up the research process. This acceleration could lead to faster insights and more robust scientific outcomes.

In summary, the research addresses critical gaps in the existing DataSHIELD ecosystem. Through its contributions, it paves the way for more secure, efficient, and democratic data analysis in scientific research.

9.2 Summary of key findings

The cornerstone achievements of this research are encapsulated in four pivotal papers that form the backbone of this thesis. Firstly, the resourcer system was developed to extend DataSHIELD's capabilities to support a diverse range of data formats. While DataSHIELD was initially designed to work with tabular data, resourcer now enables compatibility with a plethora of file types, including but not limited to R data files, plain text data, s3 compatible databases, SQL databases, and specialized formats such as VCF and BAM. This allows researchers greater flexibility in the data they can use, opening the door for more comprehensive and varied analyses.

Secondly and thirdly, specialized packages were developed to expand DataSHIELD's utility in handling omics and exposome data. Consultation with experts in both fields ensured the development of robust and relevant functionalities. Among the standouts are pooled ExWAS and pooled PCA for exposome analysis, and super-fast pooled GWAS and PRS for omic data. These functionalities not only serve theoretical needs but have proven effective in real-world applications.

Fourthly, a user-friendly graphical platform was introduced to simplify the DataSHIELD experience. This web-based application removes the need for researchers to familiarize themselves with DataSHIELD's intricacies or to install R and related dependencies. It offers a range of essential functionalities for hypothesis testing, such as descriptive analysis, GLM, survival models, GWAS, and ExWAS, with the added convenience of a built-in plot editor for generating publication-ready figures. Future updates, based on user feedback, are also in the pipeline to make the platform even more comprehensive.

Beyond these key deliverables, it is noteworthy that the developed tools are already being actively used in research projects. In particular, the ATHLETE and HELIX projects stand out as primary beneficiaries of these advancements, further underlining the real-world applicability and impact of this research.

In summary, the findings of this research have achieved the overarching goals of enhancing DataSHIELD’s capabilities, extending its applications in scientific research, and making it more accessible to the broader scientific community. These developments directly address the initial problems outlined in this research, offering a more secure, versatile, and user-friendly platform for non-disclosive data analysis.

9.3 Interpretation of findings

While the core of this research lies in tool development rather than traditional scientific inquiry, its relevance to the existing body of work on DataSHIELD and non-disclosive analysis cannot be understated. Prior to this research, DataSHIELD had a more limited scope, being predominantly used for multi-center studies that focused on tabular epidemiological data. The addition of new data types—particularly omics and exposome data—represents a significant departure from prior studies and has the potential to enrich existing research by enabling new scientific hypotheses and investigations.

Although the findings of this thesis may not be described as scientific knowledge advances in the traditional sense, their utility and potential for broad impact are remarkable. The extensions and enhancements introduced here stand to make non-disclosive analysis more robust, versatile, and applicable to a wider range of scientific questions. This opens the door for new projects concerned with data privacy to conduct rigorous and proper research without compromising on data integrity or security.

By focusing on tool development, this research has filled a unique and necessary gap in the existing landscape of non-disclosive data analysis. The tools have practical applications and are already being adopted in significant research endeavors, notably the ATHLETE and HELIX projects. The ability to seamlessly integrate omics or exposome data with traditional epidemiological data is particularly promising, as it allows for a more comprehensive understanding of various phenomena and can lead to new, impactful scientific results.

In sum, the work presented in this thesis complements and extends the existing capabilities of DataSHIELD, thereby broadening its applicability and making it a more valuable resource for secure, efficient, and comprehensive data analysis in scientific research.

9.4 Strengths and contributions

One of the foremost strengths of this research lies in its capacity to significantly improve the DataSHIELD infrastructure. This research not only elevates the platform’s capabilities but also fosters its broader adoption within the scientific community. The focus on tailoring tools to meet the specific needs of actual researchers sets this work apart, offering practical solutions that are immediately applicable and beneficial for ongoing research. The provision of documentation and the promise of continued support for new tool integration reflect a deep commitment to creating a sustainable and user-friendly environment for secure, efficient data analysis.

In terms of capabilities, this research fills a void in the existing DataSHIELD ecosystem by introducing the ability to analyze new types of data—specifically, omics and exposome data. Moreover, it does so without compromising the essential feature of non-disclosiveness. This enhancement opens new avenues for researchers interested in a wide variety of scientific questions that go beyond traditional epidemiological data. The research also furnishes all the required functionalities for comprehensive studies in these new domains, thus offering an all-inclusive platform for non-disclosive data analysis.

The novel contributions of this work have a broader implication for the intersection of bioinformatics and non-disclosive data analysis. For instance, the implementation of fast pooled GWAS and differential privacy mechanisms represent significant advances in the field. Similarly, the introduction of pooled PCA and survival analysis methods, designed specifically for DataSHIELD, fill specific gaps in the current analytical landscape. Although these contributions have been implemented within the DataSHIELD platform, their underlying principles and methodologies could well be adapted for other platforms, expanding their potential impact.

This thesis, therefore, serves as a landmark in the field, marking the integration of multiple facets of bioinformatics and non-disclosive data analysis. Through its various improvements and novel contributions, the research has the potential to catalyze significant advancements in secure, efficient, and comprehensive data analysis in scientific research.

9.5 Limitations

One significant limitation of this research revolves around the assurance of non-disclosure in the analysis packages developed. While the responsibility of ensuring non-disclosure typically falls on the developer, the lack of standardized protocols or a dedicated team within the DataSHIELD community to vet new packages for this quality adds an element of uncertainty. Although my work has benefited from the input and expertise of multiple individuals, including my thesis director and other members of the DataSHIELD community, the absence of an established verification process leaves room for potential lapses in the non-disclosive guarantee.

Another perceived limitation concerns the comprehensiveness of the analysis packages for omics and exposure data. While the open-source nature of these tools allows for future expansion, the current versions may not cater to all the diverse methodologies that different researchers might prefer. It's worth noting, however, that this limitation is inherent to any tool aiming to be comprehensive in a rapidly evolving field. The current set of functionalities is, nevertheless, robust and designed to meet the most pressing needs of researchers in these domains.

Lastly, the lack of a dedicated team or centralized review mechanism within the DataSHIELD community to ensure the non-disclosive nature of newly developed packages is both a limitation and a constraint. It poses a challenge to the expansion and acceptance of new tools within the DataSHIELD ecosystem, given the imperative of maintaining data privacy in non-disclosive analysis.

9.6 Limitations of the DataSHIELD infrastructure

DataSHIELD holds significant promise for advancing non-disclosive data analysis in bioinformatics, but the adoption and successful implementation of this infrastructure come with challenges. A primary technical obstacle hindering its wider adoption is the difficulty of use, which is intricately tied to the setup of the infrastructure itself. Implementing DataSHIELD for multi-center studies presents an even greater challenge, as the necessity to deploy new technology across different research centers can create friction between the desired goals of a study and the practical limitations of the infrastructure. This friction often dissuades new projects from adopting the technology, damaging its reputation before it even has a chance to demonstrate its capabilities.

Adding to these challenges are the financial constraints. Projects rarely budget for the new machines required to host the DataSHIELD infrastructure, often leading to the use of older, less reliable machines. This has a cascading effect, as researchers sometimes find themselves unable to conduct their analyses due to server downtime at one or more centers. While latency and scalability have not proven to be significant issues, the overall impression of DataSHIELD can suffer when researchers encounter these sorts of roadblocks.

The ecosystem's usability also limits its adoption. Currently, the connection packages for DataSHIELD are exclusively available in R, alienating researchers who may be more comfortable with Python or other programming languages. Although there are ongoing efforts to make DataSHIELD accessible through Python, this limitation can deter potential users.

Another critical issue affecting DataSHIELD's reputation and adoption is the prevailing misconception about its inherent security features. While DataSHIELD is designed to be a secure method for data analysis, the actual security lies in the analysis packages themselves, not the infrastructure. A poorly designed package can compromise the entire system, thereby defeating its core purpose of providing a non-disclosive platform. This calls into question the rigorosity of the security measures in place, and exposes the lack of a dedicated team or standardized procedures for vetting new packages for non-disclosive compliance.

Though DataSHIELD promotes itself as a secure data analysis environment, it has not undergone comprehensive penetration testing. This leaves a significant gap in its security posture, creating reservations among IT professionals who may advocate for other, more rigorously tested solutions. While VPN access for the analysis servers has been cited as a countermeasure, the evolving nature of security threats demands a more robust and proactive approach to security.

The DataSHIELD community is active and supportive, especially through public forums, but there are areas that require improvement. Documentation for developers is notably lacking, leading to a learning curve that could discourage new contributors. While the core functionalities of DataSHIELD are actively developed, the project does not have a sufficient mechanism for scrutinizing new packages to ensure their security and non-disclosive properties. This is not a small oversight, but a critical missing component that could significantly affect its long-term viability and trustworthiness.

The project's funding model also poses limitations. Relying mainly on public grants has not allowed for the kind of rapid development and feature expansion possible with private funding. This financial limitation inhibits the ecosystem from keeping pace with privately funded alternatives and even affects its ability to conduct essential security tests like penetration testing.

Given these challenges, a restructuring of priorities could significantly benefit DataSHIELD. The focus should shift from adding new features to the core system, which is already stable and functional, to ensuring the non-disclosive properties of new packages. This will not only make the system more secure but will also build trust, encouraging its adoption across new projects. While the community can contribute to some extent, organizational restructuring and additional funding avenues need to be explored to address these limitations effectively.

9.7 Future steps

The landscape of secure, non-disclosive data analysis is continually evolving, and DataSHIELD is poised to adapt and grow in response to emerging needs and technologies. One of the most immediate areas for expansion is the development of new statistical methodologies. As the scientific community's requirements become increasingly complex, the need for a richer repertoire of analytical tools within DataSHIELD becomes crucial. By incorporating a wider range of statistical methods, DataSHIELD can cater to a broader audience of researchers with diverse analytical needs.

Excitingly, one of the upcoming features that has been highly requested is Python support. DataSHIELD has primarily operated within the R programming ecosystem, which, while robust, limited its accessibility to a subset of the scientific community. The extension into Python will undoubtedly broaden its user base and facilitate integration with a multitude of data analysis pipelines, thereby boosting its applicability and utility.

In the short term, a key focus is on the extensive testing of new functions and packages. Given the project's open-source nature, the community plays a pivotal role in identifying bugs and limitations, thereby ensuring that DataSHIELD remains a reliable tool for secure data analysis. These efforts align well with the long-term strategic objectives, which include establishing DataSHIELD as a highly secure and rigorously tested platform. While there are no known plans for collaborations or partnerships to accelerate these goals, the role of community contributions cannot be overstated. As the community grows and becomes increasingly knowledgeable about the project's nuances, its members will likely make more meaningful contributions to overcoming existing limitations.

The potential applications of DataSHIELD are not limited to its current primary users in multi-center bioscience research projects. There is considerable scope for expansion into healthcare, where the tool could serve to connect hospitals and inform treatment decisions. Its non-disclosive, secure nature makes it an excellent fit for sensitive medical data. Moreover, as the toolset expands, it may find applications in social sciences, particularly if capabilities for Geographical Information Systems (GIS) analyses are developed. However, certain industries like banking may not align well with the project's goals and features.

Ensuring robust security measures is vital to maintaining the core value proposition of DataSHIELD. Given its mission to provide a non-disclosive analysis platform, any strides made in enhancing security protocols

and passing penetration tests would contribute significantly to its credibility. Simultaneously, efforts are underway to improve documentation and user-friendliness, which will undoubtedly make the platform more accessible to newcomers and contribute to its ongoing success.

In summary, the future of DataSHIELD appears promising, with planned expansions in both methodological and technological dimensions. Its success hinges on its adaptability, the contributions of an engaged community, and its ability to maintain a strong focus on its core principles of secure, non-disclosive data analysis.

10 Conclusions

Expanding the Capabilities of DataSHIELD with resources

The introduction of **resources** in the DataSHIELD platform represents a pivotal advancement, acting as a catalyzer that greatly extends the platform capabilities to use virtually any type of data as well as computing resources. This enables federated privacy-protecting analyses across multiple domains, offering a powerful prospect for academic, commercial, and healthcare sectors.

Incorporating the notion of **resources** has exponentially enhanced DataSHIELD's capabilities. It can now handle and analyze larger, more intricate datasets while maintaining privacy and security. This robust solution plays a pivotal role in managing and parsing vast datasets, ensuring seamless interoperability with other R packages.

Further extending its reach, the birth of the **dsOmics** and **dsExposome** packages has unlocked newer horizons for DataSHIELD. These allow genomic and exposome data analyses in a federated and privacy-conscious manner. Those two packages can't exist without the advances introduced by the **resources**, as the data format in which genomic and exposome data is stored, does not comply with what DataSHIELD accepted before as inputs, that being plain tables. The advances done developing this work is what unlocked the potential to have omics and exposome data analysis in DataSHIELD, and on the future that could even be expanded to image analysis and geospatial data analysis.

OmicSHIELD: A Beacon for Privacy-Protected Omics Analysis

In the realm of omics research, OmicSHIELD emerges as a revolutionary, open-source solution. It carves a niche by facilitating privacy-protected, non-disclosive omics data analyses across multi-center studies. In doing so, it paves the way for greater collaboration, helping omics research to advance leaps and bounds without having to be slowed by the always time-consuming data sharing agreements and handling of sensitive data.

Answering the call for a robust solution in omics data analysis, OmicSHIELD seamlessly integrates with multi-center studies. Its capabilities encompass a comprehensive suite of tools adept at analyzing genomic, transcriptomic, and epigenomic data. Among its capabilities, it is to be mentioned the state-of-the-art methods for pooled GWAS, filling the gaps left by predecessors like FAHME and sPLINK. Furthermore, the software harnesses differential privacy and employs disclosure trap mechanisms, ensuring a robust shield of privacy.

Beyond privacy, OmicSHIELD showcases its prowess in pooled and meta-analyses, ensuring the privacy of individual-level data. When compared with traditional methods, it stands tall, delivering consistent pooled results without the need of physically pulling the data on the same server, guaranteeing that sensitive data never leave the study servers. With prospects of integration into large consortia projects and an open-source foundation that thrives on continuous user feedback, OmicSHIELD promises to remain a standard in the ever-evolving domain of omic data analyses.

The dsExposome Package: Bridging Exposome Analyses with Data Privacy

Within the intricate framework of exposome analyses, the **dsExposome** package emerges as a new solution. It's not just any tool, but one that has been crafted for the DataSHIELD infrastructure, promising a robust solution for multi-center studies. Its power lies in balancing the scales between detailed exposome analysis and aiding data privacy concerns.

The **dsExposome** tool has showcased its merits by efficiently conducting an Exposome-Wide Association Study (ExWAS) using synthetic data sets as well as with real-world data. Its seamless compatibility with the DataSHIELD framework ensures data privacy, proving invaluable for multi-center studies where data sharing and harmonization are formidable challenges. Furthermore, its application in replicating real-world exposome analysis, such as the HELIX study, stands testament to its capability.

In addition to its analytical strengths, **dsExposome** has the ability to handle various confounding factors in exposome analyses. This ensures that the outcomes are not just accurate, but also scientifically insightful. A significant advantage it offers is its ability to perform a pooled analysis, which stands out as a more efficient

solution than traditional meta-analysis methods, especially for complex multi-source studies. And while it might chart a different course than traditional on-premises analyses, its results are comparable, proving its utility. The provision of comprehensive user guides and access to test data accentuates its commitment to transparency and collaborative research.

ShinyDataSHIELD: Streamlining Federated Non-Disclosive Analysis

At the forefront of federated non-disclosive analysis, ShinyDataSHIELD stands as a user-friendly utility. As an R Shiny application, it has been meticulously crafted to augment the usability of the DataSHIELD infrastructure. It achieves this by striking a balance, making DataSHIELD accessible to both new and seasoned researchers.

With its vast repertoire, ShinyDataSHIELD aims towards a wide array of research needs. From basic data column transformations to intricate statistical modeling, its scope spans wide. Its intuitive design ensures smooth user interactions, enabling fast hypothesis testing and efficient analysis, all without demanding programming or scripting know-how.

The platform also ensures data integrity with its rigorous checks and understandable error messages, empowering users to identify and rectify issues on the fly. As it continues to evolve, the promise of extended plotting functionalities and the incorporation of new DataSHIELD features ensures it remains at the cutting edge. Given its potential, it is poised to play a fundamental role in promoting the adoption of non-disclosive analysis methods, underscoring the immense promise of such methods in advancing research while upholding data privacy.

DataSHIELD in Action: Real-World Applications for Collaborative Multi-Center Analyses

DataSHIELD's application on tangible, real-world data has substantiated its potential as a robust tool for secure, collaborative multi-center analyses. Its powers goes beyond theoretical promises, offering researchers a platform that seamlessly integrates meaningful insights with data privacy and confidentiality.

In direct application scenarios, improved DataSHIELD methodologies exhibited commendable efficiency and accuracy. When juxtaposed against traditional statistical methods, the disparities were non-significant, demonstrating the methodological success and adaptability of DataSHIELD. But beyond just methodology, DataSHIELD proved to be an instrumental bridge for researchers across multi-center studies. By allowing researchers to directly obtain results without sending analysis scripts between centers, DataSHIELD cross-center collaboration has become more streamlined and efficient for researchers.

One of DataSHIELD's standout contributions has been its indomitable impact on the reproducibility and transparency facets of research. Given that there is no need to share data to reproduce the results of a certain peer reviewed paper, research developed using DataSHIELD can easily share the scripts and DataSHIELD access so that the results can be reproduced and tested by readers and reviewers, providing an exceptionally valuable resource to guaranteeing good practices and credibility on analyses. Such guarantees are almost non-existent on traditional research given the data sharing agreements impediments and the overall reluctance to share data just to be used to reproduce some results.

With such practical applications and outcomes, DataSHIELD has firmly established itself as more than just a tool—it's a catalyst for reshaping the landscape of collaborative research while upholding the data privacy.

References

- [1] Christine L. Borgman. “The conundrum of sharing research data”. In: *Journal of the American Society for Information Science and Technology* 63 (6 June 2012), pp. 1059–1078. ISSN: 1532-2890. DOI: 10.1002/ASI.22634. URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/asi.22634><https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22634><https://onlinelibrary.wiley.com/doi/10.1002/asi.22634>.
- [2] Jim Gray and Alex Szalay. “Where the Rubber Meets the Sky: Bridging the Gap between Databases and Science”. In: (2004).
- [3] Ann S. Zimmerman. “New Knowledge from Old Data”. In: <http://dx.doi.org/10.1177/0162243907306704> 33 (5 Sept. 2008), pp. 631–652. ISSN: 01622439. DOI: 10.1177/0162243907306704. URL: <https://journals.sagepub.com/doi/abs/10.1177/0162243907306704?journalCode=sthd>.
- [4] George F Coulouris, Jean Dollimore, and Tim Kindberg. *Distributed systems: concepts and design*. pearson education, 2005.
- [5] Kasame Tritrakan and Veera Muangsin. “Using peer-to-peer communication to improve the performance of distributed computing on the Internet”. In: *Proceedings - International Conference on Advanced Information Networking and Applications, AINA 2 (2005)*, pp. 295–298. ISSN: 1550445X. DOI: 10.1109/AINA.2005.337.
- [6] Arif Sari et al. “Fault Tolerance Mechanisms in Distributed Systems”. In: *International Journal of Communications, Network and System Sciences* 8 (12 Dec. 2015), pp. 471–482. ISSN: 1913-3715. DOI: 10.4236/IJCSNS.2015.812042. URL: http://www.scirp.org/Html/1-9702032_61986.htm<http://www.scirp.org/Journal/Paperabs.aspx?paperid=61986>.
- [7] Susanne Busse et al. “Federated Information Systems: Concepts, Terminology and Architectures”. In: (2007).
- [8] Jonathan A Sagotsky et al. “Life Sciences and the web: a new era for collaboration”. In: *Molecular Systems Biology* (2008). DOI: 10.1038/msb.2008.39. URL: <http://openwetware.org>.
- [9] James D. Watson. “The Human Genome Project: Past, Present, and Future”. In: *Science* 248 (4951 1990), pp. 44–49. ISSN: 00368075. DOI: 10.1126/SCIENCE.2181665. URL: <https://www.science.org/doi/10.1126/science.2181665>.
- [10] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”. In: *Contemporary Oncology/Współczesna Onkologia* 2015 (1 2015), pp. 68–77. ISSN: 1428-2526. DOI: 10.5114/WO.2014.47136. URL: <https://tcga-data.nci.nih.gov/datareports/codeTablesReport..>
- [11] Jane Kaye. “The Tension Between Data Sharing and the Protection of Privacy in Genomics Research”. In: (2012). DOI: 10.1146/annurev-genom-082410-101454. URL: <http://www.epigenome.org>.
- [12] Jennifer Harrow et al. “ELIXIR: providing a sustainable infrastructure for life science data at European scale”. In: *Bioinformatics* 37 (16 Aug. 2021), pp. 2506–2511. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTAB481. URL: <https://academic.oup.com/bioinformatics/article/37/16/2506/6310171>.
- [13] Nirav Merchant et al. “The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences”. In: *PLOS Biology* 14 (1 Jan. 2016), e1002342. ISSN: 1545-7885. DOI: 10.1371/JOURNAL.PBIO.1002342. URL: <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002342>.
- [14] Ali Hasnain et al. “BioFed: Federated query processing over life sciences linked open data”. In: *Journal of Biomedical Semantics* 8 (1 Mar. 2017), pp. 1–19. ISSN: 20411480. DOI: 10.1186/S13326-017-0118-0/FIGURES/4. URL: <https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0118-0>.
- [15] David Lazer et al. “Life in the network: the coming age of computational social science”. In: *Science* 323 (5915 2009), pp. 721–723. DOI: 10.1126/science.1167742.

- [16] Linnet Taylor and Ralph Schroeder. “Is bigger better? The emergence of big data as a tool for international development policy”. In: *GeoJournal* 80 (4 Aug. 2015), pp. 503–518. ISSN: 03432521. DOI: 10.1007/S10708-014-9603-5/METRICS. URL: <https://link.springer.com/article/10.1007/s10708-014-9603-5>.
- [17] General Data Protection Regulation. “General data protection regulation (GDPR)”. In: *Intersoft Consulting, Accessed in October 24* (1 2018).
- [18] Alexandre Passant et al. “Federating distributed social data to build an interlinked online information society”. In: *IEEE Intelligent Systems* 24 (6 2009), pp. 44–48. ISSN: 15411672. DOI: 10.1109/MIS.2009.113.
- [19] Michael Boniface et al. “The Social Data Foundation model: Facilitating health and social care transformation through datatrust services”. In: *Data Policy* 4 (2022), e6. ISSN: 2632-3249. DOI: 10.1017/DAP.2022.1. URL: <https://www.cambridge.org/core/journals/data-and-policy/article/social-data-foundation-model-facilitating-health-and-social-care-transformation-through-datatrust-services/CD882977DA412B4020945C3FFE8725A0>.
- [20] Tony Hey et al. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.
- [21] Lawrence Mbuagbaw et al. “Challenges to complete and useful data sharing”. In: *Trials* 18 (1 Feb. 2017), pp. 1–3. ISSN: 17456215. DOI: 10.1186/S13063-017-1816-8/PEER-REVIEW. URL: <https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-017-1816-8>.
- [22] Accountability Act. “Health insurance portability and accountability act of 1996”. In: *Public law* 104 (1996), p. 191.
- [23] Graham Greenleaf. “Global data privacy laws 2017: 120 national data privacy laws, including Indonesia and Turkey”. In: *Including Indonesia and Turkey (January 30, 2017)* 145 (2017), pp. 10–13.
- [24] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of data integration*. Elsevier, 2012.
- [25] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. “Data quality assessment”. In: *Communications of the ACM* 45 (4 Apr. 2002), pp. 211–218. ISSN: 15577317. DOI: 10.1145/505248.506010. URL: <https://dl.acm.org/doi/10.1145/505248.506010>.
- [26] Jeffrey Dean and Sanjay Ghemawat. “MapReduce: simplified data processing on large clusters”. In: *Communications of the ACM* 51 (1 2008), pp. 107–113.
- [27] Gianpaolo Cugola and Alessandro Margara. “Processing flows of information”. In: *ACM Computing Surveys (CSUR)* 44 (3 June 2012). ISSN: 03600300. DOI: 10.1145/2187671.2187677. URL: <https://dl.acm.org/doi/10.1145/2187671.2187677>.
- [28] Cynthia Dwork. “Differential privacy”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4052 LNCS (2006), pp. 1–12. ISSN: 16113349. DOI: 10.1007/11787006_1/COVER. URL: https://link.springer.com/chapter/10.1007/11787006_1.
- [29] Peter Bogetoft et al. “Secure multiparty computation goes live”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5628 LNCS (2009), pp. 325–343. ISSN: 03029743. DOI: 10.1007/978-3-642-03549-4_20/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-642-03549-4_20.
- [30] Xun Yi, Russell Paulet, and Elisa Bertino. “Homomorphic encryption”. In: *SpringerBriefs in Computer Science* 0 (9783319122281 2014), pp. 27–46. ISSN: 21915776. DOI: 10.1007/978-3-319-12229-8_2/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-319-12229-8_2.
- [31] Amit P. Sheth and James A. Larson. “Federated database systems for managing distributed, heterogeneous, and autonomous databases”. In: *ACM Computing Surveys (CSUR)* 22 (3 Sept. 1990), pp. 183–236. ISSN: 15577341. DOI: 10.1145/96602.96604. URL: <https://dl.acm.org/doi/10.1145/96602.96604>.
- [32] Gio Wiederhold. “Mediators in the Architecture of Future Information Systems”. In: *Computer* 25 (3 1992), pp. 38–49. ISSN: 00189162. DOI: 10.1109/2.121508.

- [33] Tian Li et al. “Federated Learning: Challenges, Methods, and Future Directions”. In: *IEEE Signal Processing Magazine* 37 (3 May 2020), pp. 50–60. ISSN: 15580792. DOI: 10.1109/MSP.2020.2975749.
- [34] Qiang Yang et al. “Federated Machine Learning”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10 (2 Jan. 2019). ISSN: 21576912. DOI: 10.1145/3298981. URL: <https://dl.acm.org/doi/10.1145/3298981>.
- [35] Oded Goldreich. “Secure multi-party computation”. In: *Manuscript. Preliminary version* 78 (110 1998).
- [36] Adi Shamir. “How to share a secret”. In: *Communications of the ACM* 22 (11 Nov. 1979), pp. 612–613. ISSN: 15577317. DOI: 10.1145/359168.359176. URL: <https://dl.acm.org/doi/10.1145/359168.359176>.
- [37] Andrew Chi Chih Yao. “HOW TO GENERATE AND EXCHANGE SECRETS.” In: *Annual Symposium on Foundations of Computer Science (Proceedings)* (1986), pp. 162–167. ISSN: 02725428. DOI: 10.1109/SFCS.1986.25.
- [38] Michael O. Rabin. “How To Exchange Secrets with Oblivious Transfer”. In: *Cryptology ePrint Archive* (2005).
- [39] Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. “On data banks and privacy homomorphisms”. In: *Foundations of secure computation* 4 (11 1978), pp. 169–180.
- [40] Craig Gentry. “Fully Homomorphic Encryption Using Ideal Lattices”. In: *Proceedings of the Annual ACM Symposium on Theory of Computing* (2009), pp. 169–178. ISSN: 07378017. DOI: 10.1145/1536414.1536440. URL: <https://dl.acm.org/doi/10.1145/1536414.1536440>.
- [41] Zvika Brakerski and Vinod Vaikuntanathan. “Fully homomorphic encryption from ring-LWE and security for key dependent messages”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6841 LNCS (2011), pp. 505–524. ISSN: 16113349. DOI: 10.1007/978-3-642-22792-9_29/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-642-22792-9_29.
- [42] Marten Van Dijk et al. “Fully homomorphic encryption over the integers”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6110 LNCS (2010), pp. 24–43. ISSN: 03029743. DOI: 10.1007/978-3-642-13190-5_2/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-642-13190-5_2.
- [43] Yoshinori Aono et al. “Scalable and secure logistic regression via homomorphic encryption”. In: *CO-DASPY 2016 - Proceedings of the 6th ACM Conference on Data and Application Security and Privacy* (Mar. 2016), pp. 142–144. DOI: 10.1145/2857705.2857731. URL: <https://dl.acm.org/doi/10.1145/2857705.2857731>.
- [44] Miran Kim et al. “Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation”. In: *JMIR Med Inform 2018;6(2):e19* <https://medinform.jmir.org/2018/2/e19> 6 (2 Apr. 2018), e8805. ISSN: 22919694. DOI: 10.2196/MEDINFORM.8805. URL: <https://medinform.jmir.org/2018/2/e19>.
- [45] Martín Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the ACM Conference on Computer and Communications Security* 24-28-October-2016 (Oct. 2016), pp. 308–318. ISSN: 15437221. DOI: 10.1145/2976749.2978318. URL: <https://dl.acm.org/doi/10.1145/2976749.2978318>.
- [46] H. Brendan McMahan et al. “Learning Differentially Private Recurrent Language Models”. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (Oct. 2017). URL: <https://arxiv.org/abs/1710.06963v3>.
- [47] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. “Differential Privacy Has Disparate Impact on Model Accuracy”. In: *Advances in Neural Information Processing Systems* 32 (2019).

- [48] Alexander Chowdhury et al. “A Review of Medical Federated Learning: Applications in Oncology and Cancer Research”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12962 LNCS (2022), pp. 3–24. ISSN: 16113349. DOI: 10.1007/978-3-031-08999-2_1/FIGURES/2. URL: https://link.springer.com/chapter/10.1007/978-3-031-08999-2_1.
- [49] Xiaoxiao Li et al. “Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results”. In: *Medical Image Analysis* 65 (Oct. 2020), p. 101765. ISSN: 1361-8415. DOI: 10.1016/J.MEDIA.2020.101765.
- [50] Shawn N. Murphy et al. “Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)”. In: *Journal of the American Medical Informatics Association* 17 (2 Mar. 2010), pp. 124–130. ISSN: 1067-5027. DOI: 10.1136/JAMIA.2009.000893. URL: <https://academic.oup.com/jamia/article/17/2/124/2909101>.
- [51] Sharon F. Terry. “The global alliance for genomics health”. In: *Genetic testing and molecular biomarkers* 18 (6 June 2014), pp. 375–376. ISSN: 1945-0257. DOI: 10.1089/GTMB.2014.1555. URL: <https://pubmed.ncbi.nlm.nih.gov/24896853/>.
- [52] Lena Dolman et al. “ClinGen advancing genomic data-sharing standards as a GA4GH driver project”. In: *Human Mutation* 39 (11 Nov. 2018), pp. 1686–1689. ISSN: 1098-1004. DOI: 10.1002/HUMU.23625. URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/humu.23625><https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.23625><https://onlinelibrary.wiley.com/doi/10.1002/humu.23625>.
- [53] Holger R. Roth et al. “Federated Learning for Breast Density Classification: A Real-World Implementation”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12444 LNCS (2020), pp. 181–191. ISSN: 16113349. DOI: 10.1007/978-3-030-60548-3_18/COVER. URL: https://link.springer.com/chapter/10.1007/978-3-030-60548-3_18.
- [54] Eric W. Lee et al. “Privacy-preserving Sequential Pattern Mining in distributed EHRs for Predicting Cardiovascular Disease”. In: *AMIA Summits on Translational Science Proceedings 2021* (2021), p. 384. ISSN: 1942597X. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378625/>[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378625/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378625/?report=abstract).
- [55] Wenbo Zheng et al. “Federated meta-learning for fraudulent credit card detection”. In: 2021, pp. 4654–4660.
- [56] Yang Liu et al. “FATE: An Industrial Grade Platform for Collaborative Learning With Data Protection”. In: *Journal of Machine Learning Research* 22 (2021), pp. 1–6. DOI: 10.5555/3546258. URL: <https://www.fedai.org>.
- [57] FedAI-WeBank. *Utilization of FATE in Anti Money Laundering Through Multiple Banks*. URL: <https://www.fedai.org/cases/utilization-of-fate-in-anti-money-laundering-through-multiple-banks/>.
- [58] Yan Li and Guihua Wen. “Research and Practice of Financial Credit Risk Management Based on Federated Learning.” In: *Engineering Letters* 31 (1 2023).
- [59] Peter Kairouz et al. *Advances and Open Problems in Federated Learning*. Vol. 14. Now Publishers, Inc., June 2021, pp. 1–210. ISBN: 9781680837704. DOI: 10.1561/22000000083. URL: <http://dx.doi.org/10.1561/22000000083>.
- [60] Virginia Smith et al. “Federated Multi-Task Learning”. In: *Advances in Neural Information Processing Systems* 30 (2017).
- [61] Keith Bonawitz et al. “Towards Federated Learning at Scale: System Design”. In: *Proceedings of Machine Learning and Systems* 1 (Apr. 2019), pp. 374–388.
- [62] Dan Alistarh et al. “QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding”. In: *Advances in Neural Information Processing Systems* 30 (2017).
- [63] Shaoxiong Ji et al. “Emerging Trends in Federated Learning: From Model Fusion to Federated X Learning”. In: (Feb. 2021). URL: <http://arxiv.org/abs/2102.12920>.

- [64] Patrick Jauernig, Ahmad Reza Sadeghi, and Emmanuel Stempf. “Trusted execution environments: Properties, applications, and challenges”. In: *IEEE Security and Privacy* 18 (2 Mar. 2020), pp. 56–60. ISSN: 15584046. DOI: 10.1109/MSEC.2019.2947124.
- [65] Solon Barocas, Moritz Hardt, and Arvind Narayanan. “Fairness in machine learning”. In: *Nips tutorial* 1 (2017), p. 2017.
- [66] Cynthia Dwork et al. “Fairness through awareness”. In: *ITCS 2012 - Innovations in Theoretical Computer Science Conference (2012)*, pp. 214–226. DOI: 10.1145/2090236.2090255. URL: <https://dl.acm.org/doi/10.1145/2090236.2090255>.
- [67] Shen Yan, Hsien Te Kao, and Emilio Ferrara. “Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes”. In: *International Conference on Information and Knowledge Management, Proceedings (Oct. 2020)*, pp. 1715–1724. DOI: 10.1145/3340531.3411980. URL: <https://dl.acm.org/doi/10.1145/3340531.3411980>.
- [68] Moritz Hardt et al. “Equality of Opportunity in Supervised Learning”. In: *Advances in Neural Information Processing Systems* 29 (2016).
- [69] Lingjuan Lyu et al. “Towards Fair and Privacy-Preserving Federated Deep Models”. In: *IEEE Transactions on Parallel and Distributed Systems* 31 (11 Nov. 2020), pp. 2524–2541. ISSN: 15582183. DOI: 10.1109/TPDS.2020.2996273.
- [70] Ahmed M. Abdelmoniem et al. “On the Impact of Device and Behavioral Heterogeneity in Federated Learning”. In: (Feb. 2021). URL: <https://arxiv.org/abs/2102.07500v1>.
- [71] Seyed Ali Osia et al. “A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics”. In: *IEEE Internet of Things Journal* 7 (5 May 2020), pp. 4505–4518. ISSN: 23274662. DOI: 10.1109/JIOT.2020.2967734.
- [72] Kenney Ng et al. “Curating and Integrating Data from Multiple Sources to Support Healthcare Analytics”. In: *Studies in Health Technology and Informatics* 216 (2015), pp. 1056–1056. ISSN: 18798365. DOI: 10.3233/978-1-61499-564-7-1056. URL: <https://ebooks.iospress.nl/doi/10.3233/978-1-61499-564-7-1056>.
- [73] Bradley Malin, David Karp, and Richard H Scheuermann. “Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research NIH Public Access”. In: *J Investig Med* 58 (1 2010), pp. 11–18. DOI: 10.231/JIM.0b013e3181c9b2ea.
- [74] Isabelle Budin-Ljøsne et al. “DataSHIELD: An Ethically Robust Solution to Multiple-Site Individual-Level Data Analysis”. In: *Public Health Genomics* 18 (2015), pp. 87–96. DOI: 10.1159/000368959. URL: www.karger.com/OA-license.
- [75] H J Lamberink et al. “Statistical power of clinical trials increased while effect size remained stable Citation for published version (APA)”. In: *Journal of Clinical Epidemiology* 102 (2014), pp. 123–128. DOI: 10.1016/j.jclinepi.2018.06.014. URL: <https://doi.org/10.1016/j.jclinepi.2018.06.014>.
- [76] Amadou Gaye et al. “DataSHIELD: taking the analysis to the data, not the data to the analysis”. In: *International Journal of Epidemiology* 43 (6 Dec. 2014), pp. 1929–1944. ISSN: 0300-5771. DOI: 10.1093/IJE/DYU188. URL: <https://academic.oup.com/ije/article/43/6/1929/707730>.
- [77] Emmanouil Prokakis. “Free and Open-Source Software: Freedom, Transparency and Efficiency in the Digitalization Era”. In: *Journal of Politics and Ethics in New Technologies and AI* 1 (1 Aug. 2022), e31230–e31230. ISSN: 2944-9243. DOI: 10.12681/JPENTAI.31230. URL: <https://ejournals.epublishing.ekt.gr/index.php/jpentai/article/view/31230>.
- [78] Yuriy Tymchuk, Andrea Mocchi, and Michele Lanza. “Collaboration in Open-Source Projects: Myth or Reality?” In: (2014). URL: <http://smalltalkhub.com>.
- [79] Dale Murray. “Open source and security: why transparency now equals strength”. In: [https://doi.org/10.1016/S1353-4858\(20\)30082-9](https://doi.org/10.1016/S1353-4858(20)30082-9) 2020 (7 Nov. 2021), pp. 17–19. ISSN: 13534858. DOI: 10.1016/S1353-4858(20)30082-9. URL: <https://www.magonlinelibrary.com/doi/10.1016/S1353-4858%2820%2930082-9>.

- [80] Victoria Chico. “The impact of the General Data Protection Regulation on health research”. In: *British Medical Bulletin* 128 (1 Dec. 2018), pp. 109–118. ISSN: 0007-1420. DOI: 10.1093/BMB/LDY038. URL: <https://academic.oup.com/bmb/article/128/1/109/5184942>.
- [81] Shona Kalkman et al. “Responsible data sharing in international health research: A systematic review of principles and norms”. In: *BMC Medical Ethics* 20 (1 Mar. 2019), pp. 1–13. ISSN: 14726939. DOI: 10.1186/S12910-019-0359-9/TABLES/8. URL: <https://link.springer.com/articles/10.1186/s12910-019-0359-9><https://link.springer.com/article/10.1186/s12910-019-0359-9>.
- [82] Xing Zhao et al. “Cohort Profile: the China Multi-Ethnic Cohort (CMEC) study on behalf of the China Multi-Ethnic Cohort (CMEC) collaborative group”. In: (). DOI: 10.1093/ije/dyaa185. URL: <https://academic.oup.com/ije/article/50/3/721/6000341>.
- [83] Peter McCullagh. *Generalized linear models*. Routledge, 2019.
- [84] Murray A Aitkin et al. *Statistical modelling in GLIM 4*. Vol. 32. Oxford University Press, USA, 2005.
- [85] E. M. Jones et al. “DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective”. In: *Norsk Epidemiologi* 21 (2 Apr. 2012), pp. 231–239. ISSN: 0803-2491. DOI: 10.5324/NJE.V21I2.1499. URL: <https://www.ntnu.no/ojs/index.php/norepid/article/view/1499>.
- [86] Sandeep Sen and Amit Kumar. “Parallel Algorithms”. In: *Design and Analysis of Algorithms* (May 2019), pp. 277–307. DOI: 10.1017/9781108654937.015. URL: <https://www.cambridge.org/core/books/design-and-analysis-of-algorithms/parallel-algorithms/6078AD64E93EA5E10A690E427E82950D>.
- [87] Richard J. Anderson and Lawrence Snyder. “A Comparison of Shared and Nonshared Memory Models of Parallel Computation”. In: *Proceedings of the IEEE* 79 (4 1991), pp. 480–487. ISSN: 15582256. DOI: 10.1109/5.92042.
- [88] Martine Vrijheid et al. “Advancing tools for human early lifecourse exposome research and translation (ATHLETE): Project overview”. In: *Environmental Epidemiology* 5 (5 Oct. 2021), E166. ISSN: 24747882. DOI: 10.1097/EE9.000000000000166. URL: </pmc/articles/PMC8683140//pmc/articles/PMC8683140/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8683140/>.
- [89] Dany Doiron et al. “Data harmonization and federated analysis of population-based studies: The BioSHaRE project”. In: *Emerging Themes in Epidemiology* 10 (1 Nov. 2013), pp. 1–8. ISSN: 17427622. DOI: 10.1186/1742-7622-10-12/FIGURES/2. URL: <https://ete-online.biomedcentral.com/articles/10.1186/1742-7622-10-12>.
- [90] Mariona Pinart et al. “Joint Data Analysis in Nutritional Epidemiology: Identification of Observational Studies and Minimal Requirements”. In: *The Journal of Nutrition* 148 (2 Feb. 2018), pp. 285–297. ISSN: 0022-3166. DOI: 10.1093/JN/NXX037. URL: <https://academic.oup.com/jn/article/148/2/285/4913031>.
- [91] *EUCAN-CONNECT*. URL: <https://eucanconnect.com/project-description/>.
- [92] *InterConnect - global data for diabetes and obesity research - MRC Epidemiology Unit*. URL: <https://www.mrc-epid.cam.ac.uk/interconnect/>.
- [93] Valeria Agamennone et al. “HDHL-INTIMIC: A European Knowledge Platform on Food, Diet, Intestinal Microbiomics, and Human Health”. In: *Nutrients* 14 (9 May 2022), p. 1881. ISSN: 20726643. DOI: 10.3390/NU14091881/S1. URL: <https://www.mdpi.com/2072-6643/14/9/1881/htmhttps://www.mdpi.com/2072-6643/14/9/1881>.
- [94] Vincent W.V. Jaddoe et al. “The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents”. In: *European Journal of Epidemiology* 35 (7 July 2020), pp. 709–724. ISSN: 15737284. DOI: 10.1007/S10654-020-00662-Z/TABLES/3. URL: <https://link.springer.com/article/10.1007/s10654-020-00662-z>.

- [95] Hans Ulrich Prokosch et al. “MIRACUM: Medical Informatics in Research and Care in University Medicine”. In: *Methods of information in medicine* 57 (S 01 July 2018), e82–e91. ISSN: 2511705X. DOI: 10.3414/ME17-02-0025/ID/JR0025-22. URL: <http://www.thieme-connect.com/products/ejournals/html/10.3414/ME17-02-0025><http://www.thieme-connect.de/DOI/DOI?10.3414/ME17-02-0025>.
- [96] Deborah Bamber et al. “Development of a data classification system for preterm birth cohort studies: the RECAP Preterm project”. In: *BMC Medical Research Methodology* 22 (1 Dec. 2022), pp. 1–9. ISSN: 14712288. DOI: 10.1186/S12874-021-01494-5/FIGURES/4. URL: <https://bmcmredresmethodol.biomedcentral.com/articles/10.1186/s12874-021-01494-5>.
- [97] Teri A. Manolio, Peter Goodhand, and Geoffrey Ginsburg. “The International Hundred Thousand Plus Cohort Consortium: integrating large-scale cohorts to address global scientific challenges”. In: *The Lancet Digital Health* 2 (11 Nov. 2020), e567–e568. ISSN: 25897500. DOI: 10.1016/S2589-7500(20)30242-9. URL: <http://www.thelancet.com/article/S2589750020302429/fulltext><http://www.thelancet.com/article/S2589750020302429/abstract>[https://www.thelancet.com/journals/landig/article/PIIS2589-7500\(20\)30242-9/abstract](https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30242-9/abstract).
- [98] Yasaman Vali et al. “Biomarkers for staging fibrosis and non-alcoholic steatohepatitis in non-alcoholic fatty liver disease (the LITMUS project): a comparative diagnostic accuracy study”. In: *The Lancet Gastroenterology Hepatology* 0 (0 Mar. 2023). ISSN: 2468-1253. DOI: 10.1016/S2468-1253(23)00017-1. URL: <http://www.thelancet.com/article/S2468125323000171/fulltext><http://www.thelancet.com/article/S2468125323000171/abstract>[https://www.thelancet.com/journals/langas/article/PIIS2468-1253\(23\)00017-1/abstract](https://www.thelancet.com/journals/langas/article/PIIS2468-1253(23)00017-1/abstract).
- [99] Justiina Ronkainen et al. “LongITools: Dynamic longitudinal exposome trajectories in cardiovascular and metabolic noncommunicable diseases”. In: *Environmental Epidemiology* 6 (1 Feb. 2022). ISSN: 24747882. DOI: 10.1097/EE9.000000000000184. URL: <https://pubmed.ncbi.nlm.nih.gov/35657/><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8835657/><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8835657/?report=abstract><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8835657/>.
- [100] J. Fluck et al. “NFDI4Health - Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten”. In: (July 2021). DOI: 10.17192/BFDM.2021.2.8331. URL: <https://edoc.mdc-berlin.de/20895/>.
- [101] Christopher Paul Wild. “Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology”. In: *Cancer Epidemiology, Biomarkers Prevention* 14 (8 Aug. 2005), pp. 1847–1850. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-05-0456. URL: <https://aacrjournals.org/cebpa/article/14/8/1847/258124/Complementing-the-Genome-with-an-Exposome-The>.
- [102] Christopher Paul Wild. “The exposome: from concept to utility”. In: *International journal of epidemiology* 41 (1 Feb. 2012), pp. 24–32. ISSN: 1464-3685. DOI: 10.1093/IJE/DYR236. URL: <https://pubmed.ncbi.nlm.nih.gov/22296988/>.
- [103] Martine Vrijheid et al. “The human early-life exposome (HELIX): Project rationale and design”. In: *Environmental Health Perspectives* 122 (6 2014), pp. 535–544. ISSN: 15529924. DOI: 10.1289/EHP.1307204. URL: <http://dx.doi.org/10.1289/ehp.1307204>.
- [104] Mark J. Nieuwenhuijsen et al. “Using Personal Sensors to Assess the Exposome and Acute Health Effects”. In: *International Journal of Environmental Research and Public Health* 2014, Vol. 11, Pages 7805-7819 11 (8 Aug. 2014), pp. 7805–7819. ISSN: 1660-4601. DOI: 10.3390/IJERPH110807805. URL: <https://www.mdpi.com/1660-4601/11/8/7805/htm><https://www.mdpi.com/1660-4601/11/8/7805>.
- [105] R. Barouki et al. “Epigenetics as a mechanism linking developmental exposures to long-term toxicity”. In: *Environment International* 114 (May 2018), pp. 77–86. ISSN: 0160-4120. DOI: 10.1016/J.ENVINT.2018.02.014.
- [106] Volker Strobel. “Pold87/academic-keyword-occurrence: First release”. In: (Apr. 2018). DOI: 10.5281/ZENODO.1218409. URL: <https://doi.org/10.5281/zenodo.1218409#ZFPKmPuGqMA.mendeley>.

- [107] Paolo Vineis and Marc Chadeau-Hyam. “Integrating biomarkers into molecular epidemiological studies”. In: *Current Opinion in Oncology* 23 (1 Jan. 2011), pp. 100–105. ISSN: 10408746. DOI: 10.1097/CCO.0B013E3283412DE0. URL: https://journals.lww.com/co-oncology/Fulltext/2011/01000/Integrating_biomarkers_into_molecular.17.aspx.
- [108] T. Michael O’Shea et al. “Environmental influences on child health outcomes: cohorts of individuals born very preterm”. In: *Pediatric Research* 2022 93:5 93 (5 Aug. 2022), pp. 1161–1176. ISSN: 1530-0447. DOI: 10.1038/s41390-022-02230-5. URL: <https://www.nature.com/articles/s41390-022-02230-5>.
- [109] E. Nethery et al. “From measures to models: an evaluation of air pollution exposure assessment for epidemiological studies of pregnant women”. In: *Occupational and Environmental Medicine* 65 (9 Sept. 2008), pp. 579–586. ISSN: 1351-0711. DOI: 10.1136/OEM.2007.035337. URL: <https://oem.bmj.com/content/65/9/579https://oem.bmj.com/content/65/9/579.abstract>.
- [110] Judith C. Chow et al. “Designing monitoring networks to represent outdoor human exposure”. In: *Chemosphere* 49 (9 Dec. 2002), pp. 961–978. ISSN: 0045-6535. DOI: 10.1016/S0045-6535(02)00239-4.
- [111] Andrea Cattaneo et al. “Comparison between Personal and Individual Exposure to Urban Air Pollutants”. In: <http://dx.doi.org/10.1080/02786821003662934> 44 (5 May 2010), pp. 370–379. ISSN: 02786826. DOI: 10.1080/02786821003662934. URL: <https://www.tandfonline.com/doi/abs/10.1080/02786821003662934>.
- [112] Charles E. Rodes, Richard M. Kamens, and Russell W. Wiener. “The Significance and Characteristics of the Personal Activity Cloud on Exposure Assessment Measurements for Indoor Contaminants”. In: *Indoor Air* 1 (2 July 1991), pp. 123–145. ISSN: 1600-0668. DOI: 10.1111/J.1600-0668.1991.03-12.X. URL: [https://onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0668.1991.03-12.X. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0668.1991.03-12.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0668.1991.03-12.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1600-0668.1991.03-12.x](https://onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0668.1991.03-12.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0668.1991.03-12.xhttps://onlinelibrary.wiley.com/doi/10.1111/j.1600-0668.1991.03-12.x).
- [113] Eleanor Setton et al. “The impact of daily mobility on exposure to traffic-related air pollution and health effect estimates”. In: *Journal of Exposure Science Environmental Epidemiology* 2011 21:1 21 (1 June 2010), pp. 42–48. ISSN: 1559-064X. DOI: 10.1038/jes.2010.14. URL: <https://www.nature.com/articles/jes201014>.
- [114] Libelium. *Air Quality Station - Outdoor quality monitor in cities industries - Libelium*. URL: <https://www.libelium.com/iot-products/air-quality-station/>.
- [115] Vivek Shandas et al. “Integrating High-Resolution Datasets to Target Mitigation Efforts for Improving Air Quality and Public Health in Urban Neighborhoods”. In: *International Journal of Environmental Research and Public Health* 2016, Vol. 13, Page 790 13 (8 Aug. 2016), p. 790. ISSN: 1660-4601. DOI: 10.3390/IJERPH13080790. URL: <https://www.mdpi.com/1660-4601/13/8/790/htmlhttps://www.mdpi.com/1660-4601/13/8/790>.
- [116] Yanjun Du et al. “Assessment of PM_{2.5} monitoring using MicroPEM: A validation study in a city with elevated PM_{2.5} levels”. In: *Ecotoxicology and Environmental Safety* 171 (Apr. 2019), pp. 518–522. ISSN: 0147-6513. DOI: 10.1016/J.ECOENV.2019.01.002.
- [117] Richard Neitzel et al. “Contributions of Non-occupational Activities to Total Noise Exposure of Construction Workers”. In: *The Annals of Occupational Hygiene* 48 (5 July 2004), pp. 463–473. ISSN: 0003-4878. DOI: 10.1093/ANNHYG/MEH041. URL: <https://academic.oup.com/annweh/article/48/5/463/229579>.
- [118] Steven G. O’Connell, Laurel D. Kincl, and Kim A. Anderson. “Silicone wristbands as personal passive samplers”. In: *Environmental Science and Technology* 48 (6 Mar. 2014), pp. 3327–3335. ISSN: 15205851. DOI: 10.1021/ES405022F/SUPPL_FILE/ES405022F_SI_001.PDF. URL: <https://pubs.acs.org/doi/full/10.1021/es405022f>.
- [119] Anabela Berenguer et al. “Are Smartphones Ubiquitous?: An in-depth survey of smartphone adoption by seniors”. In: *IEEE Consumer Electronics Magazine* 6 (1 Jan. 2017), pp. 104–110. ISSN: 21622256. DOI: 10.1109/MCE.2016.2614524.

- [120] Tiago C. De Araújo, Lígia T. Silva, and Adriano J.C. Moreira. “Data quality issues in environmental sensing with smartphones”. In: *SENSORNETS 2017 - Proceedings of the 6th International Conference on Sensor Networks 2017-January* (2017), pp. 59–68. DOI: 10.5220/0006201600590068.
- [121] Ebrahim Nemati, Christina Batteate, and Michael Jerrett. “Opportunistic Environmental Sensing with Smartphones: a Critical Review of Current Literature and Applications”. In: *Current environmental health reports* 4 (3 Sept. 2017), pp. 306–318. ISSN: 21965412. DOI: 10.1007/s40572-017-0158-8/METRICS. URL: <https://link.springer.com/article/10.1007/s40572-017-0158-8>.
- [122] Ravi Bhoraskar et al. “Wolverine: Traffic and road condition estimation using smartphone sensors”. In: *2012 4th International Conference on Communication Systems and Networks, COMSNETS 2012* (2012). DOI: 10.1109/COMSNETS.2012.6151382.
- [123] Giuseppe Guido et al. “Estimation of Safety Performance Measures from Smartphone Sensors”. In: *Procedia - Social and Behavioral Sciences* 54 (Oct. 2012), pp. 1095–1103. ISSN: 1877-0428. DOI: 10.1016/J.SBSPRO.2012.09.824.
- [124] Sooksan Panichpapiboon and Puttipong Leakkaw. “Traffic Sensing Through Accelerometers”. In: *IEEE Transactions on Vehicular Technology* 65 (5 May 2016), pp. 3559–3567. ISSN: 00189545. DOI: 10.1109/TVT.2015.2448237.
- [125] Rui Wang et al. “Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones”. In: *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014), pp. 3–14. DOI: 10.1145/2632048.2632054. URL: <https://dl.acm.org/doi/10.1145/2632048.2632054>.
- [126] Zhenyu Chen et al. “Unobtrusive sleep monitoring using smartphones”. In: 2013, pp. 145–152.
- [127] Dianxi Shi et al. “User emotion recognition based on multi-class sensors of smartphone”. In: 2015, pp. 478–485.
- [128] Enda Murphy and Eoin A. King. “Testing the accuracy of smartphones and sound level meter applications for measuring environmental noise”. In: *Applied Acoustics* 106 (May 2016), pp. 16–22. ISSN: 0003-682X. DOI: 10.1016/J.APACOUST.2015.12.012.
- [129] Nicholas D. Lane, Petko Georgiev, and Lorena Qendro. “DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning”. In: *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Sept. 2015), pp. 283–294. DOI: 10.1145/2750858.2804262. URL: <https://dl.acm.org/doi/10.1145/2750858.2804262>.
- [130] Nadine Haddad, Xanthi D. Andrianou, and Konstantinos C. Makris. “A Scoping Review on the Characteristics of Human Exposome Studies”. In: *Current Pollution Reports* 5 (4 Dec. 2019), pp. 378–393. ISSN: 21986592. DOI: 10.1007/s40726-019-00130-7/FIGURES/3. URL: <https://link.springer.com/article/10.1007/s40726-019-00130-7>.
- [131] Liudmila Liutsko et al. “Type of Physical Activity, Diet, BMI and Tobacco/Alcohol Consumption Relationship: How They are Associated with Our Health?” In: (Mar. 2021). DOI: 10.20944/PREPRINTS202103.0766.V1. URL: <https://www.preprints.org/manuscript/202103.0766/v1>.
- [132] Chong ho Yu. “Reliability of self-report data”. In: *Retrieved August 13* (2010), p. 2011.
- [133] Gary W. Miller and Dean P. Jones. “The Nature of Nurture: Refining the Definition of the Exposome”. In: *Toxicological Sciences* 137 (1 Jan. 2014), pp. 1–2. ISSN: 1096-6080. DOI: 10.1093/TOXSCI/KFT251. URL: <https://academic.oup.com/toxsci/article/137/1/1/1647257>.
- [134] Stephen M. Rappaport. “Genetic Factors Are Not the Major Causes of Chronic Diseases”. In: *PLOS ONE* 11 (4 Apr. 2016), e0154387. ISSN: 1932-6203. DOI: 10.1371/JOURNAL.PONE.0154387. URL: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154387>.
- [135] Katie M. O’Brien et al. “Environmental chemicals in urine and blood: Improving methods for creatinine and lipid adjustment”. In: *Environmental Health Perspectives* 124 (2 Feb. 2016), pp. 220–227. ISSN: 15529924. DOI: 10.1289/EHP.1509693. URL: <http://dx.doi.org/10.1289/ehp.1509693>.

- [136] Peter Gallagher et al. “Assessing cortisol and dehydroepiandrosterone (DHEA) in saliva: effects of collection method”. In: <http://dx.doi.org/10.1177/0269881106060585> 20 (5 Sept. 2006), pp. 643–649. ISSN: 02698811. DOI: 10.1177/0269881106060585. URL: <https://journals.sagepub.com/doi/abs/10.1177/0269881106060585?journalCode=jopa>.
- [137] Stephanie M. Engel et al. “Prenatal exposure to organophosphates, paraoxonase 1, and cognitive development in childhood”. In: *Environmental Health Perspectives* 119 (8 Aug. 2011), pp. 1182–1188. ISSN: 00916765. DOI: 10.1289/EHP.1003183.
- [138] Nam Hee Kim et al. “Environmental Heavy Metal Exposure and Chronic Kidney Disease in the General Population”. In: *Journal of Korean Medical Science* 30 (3 Feb. 2015), pp. 272–277. ISSN: 15986357. DOI: 10.3346/JKMS.2015.30.3.272. URL: <https://synapse.koreamed.org/articles/1022847>.
- [139] Isabel Fortier et al. “Maelstrom Research guidelines for rigorous retrospective data harmonization”. In: *International Journal of Epidemiology* 46 (1 Feb. 2017), pp. 103–105. ISSN: 0300-5771. DOI: 10.1093/IJE/DYW075. URL: <https://academic.oup.com/ije/article/46/1/103/2617181>.
- [140] Isabel Fortier et al. “Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies”. In: *International Journal of Epidemiology* 39 (5 Oct. 2010), pp. 1383–1393. ISSN: 0300-5771. DOI: 10.1093/IJE/DYQ139. URL: <https://academic.oup.com/ije/article/39/5/1383/805481>.
- [141] Clete A. Kushida et al. “Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies”. In: *Medical care* 50 (Suppl July 2012), S82. ISSN: 00257079. DOI: 10.1097/MLR.0B013E3182585355. URL: <https://pubmed.ncbi.nlm.nih.gov/24652465/>.
- [142] O. Bodenreider. “Biomedical ontologies in action: role in knowledge management, data integration and decision support.” In: *Yearbook of medical informatics* 17 (01 2008), pp. 67–79. ISSN: 09434747. DOI: 10.1055/S-0038-1638585/ID/JR1638585-54. URL: <http://www.thieme-connect.de/products/ejournals/html/10.1055/s-0038-1638585><http://www.thieme-connect.de/DOI/DOI?10.1055/s-0038-1638585>.
- [143] A. Rogier T. Donders et al. “Review: A gentle introduction to imputation of missing values”. In: *Journal of Clinical Epidemiology* 59 (10 Oct. 2006), pp. 1087–1091. ISSN: 0895-4356. DOI: 10.1016/J.JCLINEPI.2006.01.014.
- [144] Elizabeth A. Stuart et al. “Multiple Imputation With Large Data Sets: A Case Study of the Children’s Mental Health Initiative”. In: *American Journal of Epidemiology* 169 (9 May 2009), pp. 1133–1139. ISSN: 0002-9262. DOI: 10.1093/AJE/KWP026. URL: <https://academic.oup.com/aje/article/169/9/1133/125871>.
- [145] Stef Van Buuren and Catharina G M Oudshoorn. *Multivariate imputation by chained equations*. 2000.
- [146] Richard W. Hornung and Laurence D. Reed. “Estimation of Average Concentration in the Presence of Nondetectable Values”. In: <https://doi.org/10.1080/1047322X.1990.10389587> 5 (1 Jan. 2011), pp. 46–51. ISSN: 1047322X. DOI: 10.1080/1047322X.1990.10389587. URL: <https://www.tandfonline.com/doi/abs/10.1080/1047322X.1990.10389587>.
- [147] Paul D. Juarez et al. “The Public Health Exposome: A Population-Based, Exposure Science Approach to Health Disparities Research”. In: *International Journal of Environmental Research and Public Health* 2014, Vol. 11, Pages 12866-12895 11 (12 Dec. 2014), pp. 12866–12895. ISSN: 1660-4601. DOI: 10.3390/IJERPH111212866. URL: <https://www.mdpi.com/1660-4601/11/12/12866><https://www.mdpi.com/1660-4601/11/12/12866>.
- [148] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1 Jan. 1995), pp. 289–300. ISSN: 2517-6161. DOI: 10.1111/J.2517-6161.1995.TB02031.X. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1995.tb02031.x><https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x><https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1995.tb02031.x>.

- [149] Quanli Wang et al. “Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 UK Biobank participants”. In: *bioRxiv* (11 Dec. 2020), p. 2020.12.13.422582. ISSN: 26928205. DOI: 10.1101/2020.12.13.422582. URL: <https://www.biorxiv.org/content/10.1101/2020.12.13.422582v1><https://www.biorxiv.org/content/10.1101/2020.12.13.422582v1/abstract>.
- [150] Lydiane Agier et al. “Early-life exposome and lung function in children in Europe: an analysis of data from the longitudinal, population-based HELIX cohort”. In: *The Lancet Planetary Health* 3 (2 Feb. 2019), e81–e92. ISSN: 25425196. DOI: 10.1016/S2542-5196(19)30010-5. URL: <http://www.thelancet.com/article/S2542519619300105/fulltext><http://www.thelancet.com/article/S2542519619300105/abstract>[https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196\(19\)30010-5/abstract](https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(19)30010-5/abstract).
- [151] Sandra E. Sinisi and Mark J. Van Der Laan. “Deletion/substitution/addition algorithm in learning with applications in genomics”. In: *Statistical Applications in Genetics and Molecular Biology* 3 (1 Aug. 2004). ISSN: 15446115. DOI: 10.2202/1544-6115.1069/MACHINEREADABLECITATION/RIS. URL: <https://www.degruyter.com/document/doi/10.2202/1544-6115.1069/html>.
- [152] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2 Apr. 2005), pp. 301–320. ISSN: 1467-9868. DOI: 10.1111/J.1467-9868.2005.00503.X. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2005.00503.x><https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x><https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x>.
- [153] Virissa Lenters et al. “Prenatal phthalate, perfluoroalkyl acid, and organochlorine exposures and term birth weight in three birth cohorts: Multi-pollutant models based on elastic net regression”. In: *Environmental Health Perspectives* 124 (3 Mar. 2016), pp. 365–372. ISSN: 15529924. DOI: 10.1289/EHP.1408933.
- [154] Theodore Wilbur Anderson. “An introduction to multivariate statistical analysis”. In: (1962).
- [155] Herman Wold. “Estimation of principal components and related models by iterative least squares”. In: *Multivariate analysis* (1966), pp. 391–420.
- [156] Ron Wehrens and B-H Mevik. “The pls package: principal component and partial least squares regression in R”. In: (2007).
- [157] Vrinda Kalia, Dean P. Jones, and Gary W. Miller. “Networks at the nexus of systems biology and the exposome”. In: *Current Opinion in Toxicology* 16 (Aug. 2019), pp. 25–31. ISSN: 2468-2020. DOI: 10.1016/J.COTOX.2019.03.008.
- [158] Olivier Taboureau and Karine Audouze. “Human environmental disease network: A computational model to assess toxicology of contaminants”. In: *ALTEX - Alternatives to animal experimentation* 34 (2 May 2017), pp. 289–300. ISSN: 1868-8551. DOI: 10.14573/ALTEX.1607201. URL: <https://altex.org/index.php/altex/article/view/58>.
- [159] Jingwen Yan et al. “Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data”. In: *Briefings in Bioinformatics* 19 (6 Nov. 2018), pp. 1370–1381. ISSN: 1467-5463. DOI: 10.1093/BIB/BBX066. URL: <https://academic.oup.com/bib/article/19/6/1370/3904446>.
- [160] Sijia Huang, Kumardeep Chaudhary, and Lana X. Garmire. “More is better: Recent progress in multi-omics data integration methods”. In: *Frontiers in Genetics* 8 (JUN June 2017), p. 84. ISSN: 16648021. DOI: 10.3389/FGENE.2017.00084/BIBTEX.
- [161] Marc Chadeau-Hyam et al. “Meeting-in-the-middle using metabolic profiling – a strategy for the identification of intermediate biomarkers in cohort studies”. In: <http://dx.doi.org/10.3109/1354750X.2010.533285> 16 (1 Feb. 2011), pp. 83–88. ISSN: 1354750X. DOI: 10.3109/1354750X.2010.533285. URL: <https://www.tandfonline.com/doi/abs/10.3109/1354750X.2010.533285>.
- [162] Ming Kei Chung et al. “Exposome-wide association study of semen quality: Systematic discovery of endocrine disrupting chemical biomarkers in fertility require large sample sizes”. In: *Environment International* 125 (Apr. 2019), pp. 505–514. ISSN: 0160-4120. DOI: 10.1016/J.ENVINT.2018.11.037.

- [163] Bruna Galobardes, John W. Lynch, and George Davey Smith. “Childhood Socioeconomic Circumstances and Cause-specific Mortality in Adulthood: Systematic Review and Interpretation”. In: *Epidemiologic Reviews* 26 (1 July 2004), pp. 7–21. ISSN: 0193-936X. DOI: 10.1093/EPIREV/MXH008. URL: <https://academic.oup.com/epirev/article/26/1/7/384224>.
- [164] Ken Sexton, Larry L. Needham, and James L. Pirkle. “Human Biomonitoring of Environmental Chemicals: Measuring chemicals in human tissues is the” gold standard” for assessing people’s exposure to pollution”. In: *American Scientist* 92 (1 2004), pp. 38–45.
- [165] Amelia K. Wesselink, Elizabeth E. Hatch, and Lauren A. Wise. “Invited Commentary: Interaction Between Diet and Chemical Exposures”. In: *American Journal of Epidemiology* 188 (9 Sept. 2019), pp. 1605–1607. ISSN: 0002-9262. DOI: 10.1093/AJE/KWZ152. URL: <https://academic.oup.com/aje/article/188/9/1605/5523271>.
- [166] John P.A. Ioannidis. “Exposure-wide epidemiology: revisiting Bradford Hill”. In: *Statistics in Medicine* 35 (11 May 2016), pp. 1749–1762. ISSN: 1097-0258. DOI: 10.1002/SIM.6825. URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.6825><https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6825><https://onlinelibrary.wiley.com/doi/10.1002/sim.6825>.
- [167] Chirag J. Patel and John P.A. Ioannidis. “Placing epidemiological results in the context of multiplicity and typical correlations of exposures”. In: *J Epidemiol Community Health* 68 (11 Nov. 2014), pp. 1096–1100. ISSN: 0143-005X. DOI: 10.1136/JECH-2014-204195. URL: <https://jech.bmj.com/content/68/11/1096><https://jech.bmj.com/content/68/11/1096.abstract>.
- [168] Tomohiro Shinozaki and Etsuji Suzuki. “Understanding Marginal Structural Models for Time-Varying Exposures: Pitfalls and Tips”. In: *Journal of Epidemiology* 30 (9 Sept. 2020), pp. 377–389. ISSN: 0917-5040. DOI: 10.2188/JEA.JE20200226. URL: <https://doi.org/10.2188/jea.JE20200226>.
- [169] Biswapriya B. Misra et al. “Integrated Omics: Tools, Advances, and Future Approaches”. In: *Journal of molecular endocrinology* 62 (1 Jan. 2018), R21–R45. ISSN: 1479-6813. DOI: 10.1530/JME-18-0055. URL: <https://pubmed.ncbi.nlm.nih.gov/30006342/>.
- [170] Ryo Yamada et al. “Interpretation of omics data analyses”. In: *Journal of Human Genetics* 2020 66:1 66 (1 May 2020), pp. 93–102. ISSN: 1435-232X. DOI: 10.1038/s10038-020-0763-5. URL: <https://www.nature.com/articles/s10038-020-0763-5>.
- [171] Uwe Röhm and José A Blakeley. “Data Management for High-Throughput Genomics”. In: (Sept. 2009). URL: <https://arxiv.org/abs/0909.1764v1>.
- [172] Peter M. Visscher et al. “Five Years of GWAS Discovery”. In: *The American Journal of Human Genetics* 90 (1 Jan. 2012), pp. 7–24. ISSN: 0002-9297. DOI: 10.1016/J.AJHG.2011.11.029.
- [173] Jianxiao Liu et al. “Application of deep learning in genomics”. In: *Science China Life Sciences* 63 (12 Dec. 2020), pp. 1860–1878. ISSN: 18691889. DOI: 10.1007/S11427-020-1804-5/METRICS. URL: <https://link.springer.com/article/10.1007/s11427-020-1804-5>.
- [174] François Fuks. “DNA methylation and histone modifications: teaming up to silence genes”. In: *Current Opinion in Genetics Development* 15 (5 Oct. 2005), pp. 490–495. ISSN: 0959-437X. DOI: 10.1016/J.GDE.2005.08.002.
- [175] Martin Widschwendter et al. “Epigenome-based cancer risk prediction: rationale, opportunities and challenges”. In: *Nature Reviews Clinical Oncology* 2018 15:5 15 (5 Feb. 2018), pp. 292–309. ISSN: 1759-4782. DOI: 10.1038/nrclinonc.2018.30. URL: <https://www.nature.com/articles/nrclinonc.2018.30>.
- [176] Benjamin F. Cravatt, Gabriel M. Simon, and John R. Yates. “The biological impact of mass-spectrometry-based proteomics”. In: *Nature* 2007 450:7172 450 (7172 Dec. 2007), pp. 991–1000. ISSN: 1476-4687. DOI: 10.1038/nature06525. URL: <https://www.nature.com/articles/nature06525>.
- [177] Hedi Hegyi and Mark Gerstein. “The relationship between protein structure and function: a comprehensive survey with application to the yeast genome”. In: *Journal of Molecular Biology* 288 (1 Apr. 1999), pp. 147–164. ISSN: 0022-2836. DOI: 10.1006/JMBI.1999.2661.

- [178] Wenbin Zhou et al. “The Effect of Exhaustive Exercise on Plasma Metabolic Profiles of Male and Female Rats”. In: *Journal of Sports Science Medicine* 18 (2 2019), p. 253. ISSN: 13032968. URL: [/pmc/articles/PMC6543993//pmc/articles/PMC6543993/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6543993/](https://pubmed.ncbi.nlm.nih.gov/pmc/articles/PMC6543993/).
- [179] John L. Harwood. “Environmental factors which can alter lipid metabolism”. In: *Progress in Lipid Research* 33 (1-2 Jan. 1994), pp. 193–202. ISSN: 0163-7827. DOI: 10.1016/0163-7827(94)90022-1.
- [180] Hanna Johansson Jänkänpää et al. “Metabolic profiling reveals metabolic shifts in Arabidopsis plants grown under different light conditions”. In: *Plant, Cell Environment* 35 (10 Oct. 2012), pp. 1824–1836. ISSN: 1365-3040. DOI: 10.1111/J.1365-3040.2012.02519.X. URL: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1365-3040.2012.02519.x><https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-3040.2012.02519.x>
- [181] Amelia Casamassimi et al. “Transcriptome Profiling in Human Diseases: New Advances and Perspectives”. In: *International Journal of Molecular Sciences 2017, Vol. 18, Page 1652* 18 (8 July 2017), p. 1652. ISSN: 1422-0067. DOI: 10.3390/IJMS18081652. URL: [https://www.mdpi.com/1422-0067/18/8/1652](https://www.mdpi.com/1422-0067/18/8/1652/htmhttps://www.mdpi.com/1422-0067/18/8/1652).
- [182] Jason A. Reuter, Damek V. Spacek, and Michael P. Snyder. “High-Throughput Sequencing Technologies”. In: *Molecular Cell* 58 (4 May 2015), pp. 586–597. ISSN: 1097-2765. DOI: 10.1016/J.MOLCEL.2015.05.004.
- [183] Sophia Yohe and Bharat Thyagarajan. “Review of Clinical Next-Generation Sequencing”. In: *Archives of Pathology Laboratory Medicine* 141 (11 Nov. 2017), pp. 1544–1557. ISSN: 0003-9985. DOI: 10.5858/ARPA.2016-0501-RA. URL: <https://dx.doi.org/10.5858/arpa.2016-0501-RA>.
- [184] Paola Indovina et al. “Mass spectrometry-based proteomics: The road to lung cancer biomarker discovery”. In: *Mass Spectrometry Reviews* 32 (2 Mar. 2013), pp. 129–142. ISSN: 1098-2787. DOI: 10.1002/MAS.21355. URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/mas.21355https://onlinelibrary.wiley.com/doi/abs/10.1002/mas.21355https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/mas.21355>.
- [185] Otilia Menyhart and Balázs Györfy. “Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis”. In: *Computational and Structural Biotechnology Journal* 19 (Jan. 2021), pp. 949–960. ISSN: 2001-0370. DOI: 10.1016/J.CSBJ.2021.01.009.
- [186] Vanessa Aguiar-Pulido et al. “Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis”. In: *Evolutionary Bioinformatics* 12 (May 2016), pp. 5–16. ISSN: 11769343. DOI: 10.4137/EBO.S36436/ASSET/IMAGES/LARGE/10.4137_EBO.S36436-FIG3.JPEG. URL: <https://journals.sagepub.com/doi/10.4137/EBO.S36436>.
- [187] Xiangqin Cui and Gary A. Churchill. “Statistical tests for differential expression in cDNA microarray experiments”. In: *Genome Biology* 4 (4 Apr. 2003), pp. 1–10. ISSN: 14656906. DOI: 10.1186/GB-2003-4-4-210/FIGURES/1. URL: <https://link.springer.com/articles/10.1186/gb-2003-4-4-210https://link.springer.com/article/10.1186/gb-2003-4-4-210>.
- [188] Matthew E. Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic Acids Research* 43 (7 Apr. 2015), e47–e47. ISSN: 0305-1048. DOI: 10.1093/NAR/GKV007. URL: <https://dx.doi.org/10.1093/nar/gkv007>.
- [189] Michael I. Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biology* 15 (12 Dec. 2014), pp. 1–21. ISSN: 1474760X. DOI: 10.1186/S13059-014-0550-8/FIGURES/9. URL: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>.
- [190] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26 (1 Jan. 2010), pp. 139–140. ISSN: 1367-4803. DOI: 10.1093/BIOINFORMATICS/BTP616. URL: <https://dx.doi.org/10.1093/bioinformatics/btp616>.

- [191] R Artusi, P Verderio, and EJTIjobm Marubini. “Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval”. In: *The International journal of biological markers* 17 (2 2002), pp. 148–151.
- [192] Frank Nielsen. “Hierarchical Clustering”. In: (2016), pp. 195–211. DOI: 10.1007/978-3-319-21903-5_8. URL: https://link.springer.com/chapter/10.1007/978-3-319-21903-5_8.
- [193] J. A. Hartigan and M. A. Wong. “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Applied Statistics* 28 (1 1979), p. 100. ISSN: 00359254. DOI: 10.2307/2346830.
- [194] Noviyanti T M Sagala, Alexander Agung, and Santoso Gunawan. “Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods”. In: *ComTech: Computer, Mathematics and Engineering Applications* 13 (1 Feb. 2022), pp. 1–10. ISSN: 2476-907X. DOI: 10.21512/COMTECH.V13I1.7270. URL: <https://journal.binus.ac.id/index.php/comtech/article/view/7270>.
- [195] Steve Horvath. *Weighted network analysis: applications in genomics and systems biology*. Springer Science Business Media, 2011.
- [196] Miguel A. García-Campos, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. “Pathway analysis: State of the art”. In: *Frontiers in Physiology* 6 (DEC Dec. 2015), p. 170515. ISSN: 1664042X. DOI: 10.3389/FPHYS.2015.00383/BIBTEX.
- [197] Ori Folger et al. “Predicting selective drug targets in cancer through metabolic networks”. In: *Molecular systems biology* 7 (2011). ISSN: 1744-4292. DOI: 10.1038/MSB.2011.35. URL: <https://pubmed.ncbi.nlm.nih.gov/21694718/>.
- [313] M. A. Iwen and B. W. Ong. “A Distributed and Incremental SVD Algorithm for Agglomerative Data Analysis on Large Networks”. In: <https://doi.org/10.1137/16M1058467> 37 (4 Nov. 2016), pp. 1699–1718. ISSN: 10957162. DOI: 10.1137/16M1058467. URL: <https://epubs.siam.org/doi/10.1137/16M1058467>.
- [314] Barbara Hohlt. “Pthread parallel k-means”. In: *UC Berkeley* (2001), pp. 1–9.
- [315] Shenshen Liang et al. “Design and evaluation of a parallel K-nearest neighbor algorithm on CUDA-enabled GPU”. In: *Proceedings - 2010 IEEE 2nd Symposium on Web Society, SWS 2010* (2010), pp. 53–60. DOI: 10.1109/SWS.2010.5607480.
- [316] Jérôme Pagès. *Multiple factor analysis by example using R*. CRC Press, 2014.

Appendices

Publications derived from this work

Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD

Yannick Marcon^{1*}, Tom Bishop², Demetris Avraam³, **Xavier Escriba-Montagut**^{4,5}, Patricia Ryser-Welch³, Stuart Wheeler⁶, Paul Burton³, Juan R. González^{4,5,7,8*}

Published on: PLOS Computational Biology. IF 4.779 (2021). Position Q1

¹Epigeny, St Ouen, France, ²MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, ³Population Health Sciences Institute, Newcastle University, Newcastle, United Kingdom, ⁴Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, ⁵Universitat Pompeu Fabra (UPF), Barcelona, Spain, ⁶Arjuna Technologies, Newcastle, United Kingdom, ⁷Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain, ⁸Dept. of Mathematics, Universitat Autònoma de Barcelona (UAB), Bellaterra (Barcelona), Spain. *Email: yannick.marcon@obiba.org
juanr.gonzalez@isglobal.org

OmicSHIELD: Federated privacy-protected meta- and mega-omic data analysis in multi-centre studies with a fully open source analytic platform

Xavier Escriba-Montagut^{1,2}, Yannick Marcon³, Augusto Anguita-Ruiz^{1,2}, Sofia Aguilar^{1,2}, Demetris Avraam^{6,7}, Jose Urquiza^{1,2}, Andrei S. Morgan^{4,5}, Rebecca C. Wilson^{6,7}, Paul Burton⁶ and Juan R. Gonzalez^{1,2,8*}

Submitted

¹Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, ²Universitat Pompeu Fabra (UPF), Barcelona, Spain, ³Epigeny, St Ouen, France, ⁴Université Paris Cité, Centre of Research in Epidemiology and Statistics (CRESS), Obstetrical Perinatal and Pediatric Epidemiology Research Team (EPOPé), INSERM, INRAE, F-75006, Paris, France, ⁵Elizabeth Garrett Anderson Institute for Women's Health London, University College London, London, UK, ⁶Population Health Sciences Institute, Newcastle University, Newcastle, United Kingdom, ⁷Department of Public Health, Policy and Systems, University of Liverpool, Liverpool, United Kingdom, ⁸Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain. *Email: juanr.gonzalez@isglobal.org

dsExposome: Secure and Privacy-Preserving Exposome Analysis using the DataSHIELD Infrastructure

Xavier Escriba-Montagut^{1,2}, Augusto Anguita-Ruiz^{1,2,5}, Yannick Marcon⁴, Xavier Basagaña^{1,2} and Juan R. Gonzalez^{1,2,3*}

Submitted

¹Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, ²Universitat Pompeu Fabra (UPF), Barcelona, Spain, ³Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain, ⁴Epigeny, St Ouen, France, ⁵CIBEROBN (Physiopathology of Obesity and Nutrition Network CB12/03/30038), Institute of Health Carlos III (ISCIII), 28029 Madrid, Spain. *Email: juanr.gonzalez@isglobal.org

Software application profile: ShinyDataSHIELD - An R Shiny application to perform federated non-disclosive data analysis in multi-cohort studies

Xavier Escriba-Montagut^{1,2}, Yannick Marcon³, Demetris Avraam⁴, Soumya Banerjee⁵, Tom R. P. Bishop⁵, Paul Burton⁴ and Juan R. Gonzalez^{1,2,6*}

Published on: International Journal of Epidemiology. IF 9.685 (2021). Position Q1

¹Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain, ²Universitat Pompeu Fabra (UPF), Barcelona, Spain, ³Epigeny, St Ouen, France, ⁴Population Health Sciences Institute, Newcastle University, Newcastle, United Kingdom, ⁵MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK, ⁶Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain. *Email: juanr.gonzalez@isglobal.org

PhD Portfolio

Other tasks developed

- Maintainer of R BioConductor packages of the research group: `MEAL`, `MultiDataSet`, `rexposome`, `omicRexposome` and `CTDquerier`.
- Supervisor of technical computer scientist intern from June 2021 to July 2021.
- Opal node of ISGlobal. Tasks developed:
 - Migation of user and database to new Opal node
 - Deplyment of node using a Docker swarm stack
 - Regular maintainment of node
 - Update of node to comply with LifeCycle analysis requirements
- Opal node for HELIX. Tasks developed:
 - Deployment of the node in parallel to the one for ISGlobal
 - Upload of HELIX data
- UnCoVer project. Tasks developed:
 - Creation of analysis scripts in DataSHIELD. Descriptive tables, linear modeling and survival analysis
 - Development of R Shiny application to interact with the project DataSHIELD infrastructure and perform analysis
 - Development of package to work with temporal (dates) data in DataSHIELD (`dsDates`)
 - Development of custom features on the `dsSurvival` package to fulfill their analysis needs:
 - * Improvement of survival curves visualization using `ggplot2` package
 - * Stratification of survival curves
 - * Survival analysis using life tables to enable pooled survival analysis with multi-center data
- CADSET project. Tasks developed:
 - Setup of prototype Opal nodes
 - Technical support for Opal deployment to the different cohorts involved
 - Online workshops for the data managers. Data upload, user management and permission management
- COVICAT project. Tasks developed:
 - Setup of Opal node for the project at the BSC-CNS infrastructure
- Technical support for group students bachelor's and master's thesis:
 - Carla Casanova. RNA-seq analysis using radiomic features calculated from lungs of patients with COPD. Master in Bioinformatics, Universitat Autònoma de Barcelona (UAB). July 2022.
 - Carlos Lopez. RNA-seq analysis using radiomic features calculated from lungs of patients with COPD. Master in Bioinformatics, Universitat Autònoma de Barcelona (UAB). July 2022.
 - Elisabet-Vera Matamala. Implementation of a 3D CNN for COPD classification. Bahcelor's degree in Biomedical Engineering, Universitat Autònoma de Barcelona (UAB). June 2023.

Presentations in congress

- ShinyDataSHIELD: An R Shiny application to perform federated non-disclosive data analysis in multi-cohort studies. Talk presented at the DataSHIELD conference 2021 held online (10-11 of November 2021)
- Non-disclosive federated exposome data analysis with DataSHIELD and Bioconductor in multicohort consortia. Talk presented at the EHEN Scientific Meeting held in Barcelona, Spain (24-25 of May 2022)
- OmicSHIELD: privacy-protected federated omic data analysis in multi-center studies with Bioconductor through DataSHIELD. Talk presented at the EuroBioConductor Conference held in Heidelberg, Germany (14-16 of September 2022)
- Differential privacy: a new disclosure control method to the DataSHIELD ecosystem. Talk presented at the DataSHIELD conference 2022 held in Barcelona, Spain (19-21 of October 2022)
- dsOmics in ATHLETE: Problems and current status "EWAS Green Spaces blood". Talk presented in collaboration with Sofía Aguilar at the ATHLETE Consortia meeting held in Barcelona, Spain (24 of January 2023)
- Enhancing Survival Analysis with survival tables. Talk presented at the DataSHIELD conference 2023 held in Groningen, Netherlands (11-13 of October 2023)
- Survival analysis in DataSHIELD and future directions with OMOP CMD. Talk presented at the ATHLETE Consortia meeting held in Grenoble, France (22-25 of January 2024)

Workshops delivered

- Utilization of the unCoVer toolbox for COVID-19 data analysis, Workshop delivered in collaboration with Juan R. Gonzalez at the UnCoVer project meeting held at Universidad Politecnica de Madrid (6 of May 2022)
- Técnicas ómicas en el diagnóstico de enfermedades raras, Workshop delivered in collaboration with Laura Balagué and Natàlia Carreras at the "La Marató" iGenCO workshops at Parc Científic de Barcelona (16-17 of November 2022)
- Joint development of federated analyses within unCoVer's Opal/Datashield Infrastructure, Workshop delivered in collaboration with Juan R. Gonzalez at the UnCoVer project meeting held at Universidad Politecnica de Madrid (12 of April 2023)