GLI⊗MATCH

Project No. 101136670

The malignant Glioma immuno-oncology matchmaker: towards data-driven precision medicine using spatially resolved radio-multiomics

Deliverable 8.2
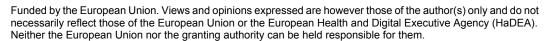# First project data management plan (P-DMP)

WP 8 – Coordination, project and innovation management

Version 1.0

Authors	Frederik De Smet (KUL), Eva Avilla Royo (accelCH)
Lead participant	KUL
Delivery date	30 June 2024
Dissemination level	SEN = Sensitive
Type	DMP = Data Management Plan

Revision history

Author(s)	Description	Date
Frederik De Smet, Chiara Caprioli, Julie Morscio (KUL)	Deliverable draft	25 May 2024
First and last name (partner short name)	Revision 1	dd Month yyyy
First and last name (partner short name)		dd Month yyyy
First and last name (partner short name)		dd Month yyyy
First and last name (partner short name)	Final version	dd Month yyyy

GLI⊛MATCH

Table of Contents

GLI◉MATCH

Partner short names

Short name	Partner
KUL	Katholieke Universiteit Leuven
KUL-LISCO	Dept. of Imaging and Pathology/ Dept. of Human genetics/Leuven Institute for single-cell omics
KUL-UZL	University Hospitals Leuven
EMC	Erasmus Universitair Medisch Centrum Rotterdam
UM	Universiteit Maastricht
OUS	Oslo Universitetssykehus HF
FSJD	Sant Joan de Déu Research Foundation
HSJD	Hospital Sant Joan De Deu
RMC	The Health Corporation - Rambam
UDUS	Heinrich-Heine Universität Düsseldorf
FINCB	Fondazione IRCCS Instituto Neurologico "Carlo Besta'
UU	Uppsala Universitet
AA	Aspect Analytics NV
Timelex	Timelex BV/SRL
accelCH	accelopment Schweiz AG

Abbreviations

Abbreviation	Term
CA	Consortium Agreement
CT	Computerized Tomography
D	Deliverable
EC	European Commission
EtC	Ethical Committee
eCRF	electronic Case Report Form
EU	European Union
GBM	Glioblastoma
HEU	Horizon Europe
IP	Intellectual Property
M	Month
MRI	Magnetic Resonance Imaging
MS	Milestone
NGS	Next Generation Sequencing
phGG	Paediatric High-Grade Glioma
WP	Work Package

Executive summary

Background

The deliverable D8.2: "First project data management plan (P-DMP)" is part of Work Package 8 (WP8): "Co-ordination, project and innovation management", led by KU Leuven.

This deliverable describes a first proposal of the Data Management Plan within the GLIOMATCH project, specifying how research data will be handled throughout the entire lifecycle of the research.

Objectives

D8.2 identifies key tasks and implements strategies to ensure that the generated research data are of high-quality, processed and analysed in the most effective way, stored in a secure environment, accessible and reusable by the relevant subjects for the entire duration of the project and beyond, in agreement with the FAIR principles.

Methodology and implementation

The FAIR principles will be used as a guideline for creation of the data management plan and to design and implement the GLIOMATCH data lake. To comply with the FAIR principles, key information on data collection, storage and sharing will be collected.

A Consortium Agreement will be established among GLIOMATCH partners, to regulate roles and responsibilities of the different aspects of data management in the project and relationship among the parties. This will be also instrumental to regulate long-term usage of the generated data.

Outcomes

The presented DMP will make the research plan operational by enabling collection, storage, maintenance and sharing of clinical and radiological data, experimental data (including raw and processed data and metadata), along with downstream analyses and developed statistical models as foreseen in WPs 2-3-4-5-6.

Impact

Proper data management will be key for carrying out the foreseen research plan in the most effective way and ensuring the reproducibility of results. Moreover, it will enable sharing findings and data throughout the scientific community, which is expected to boost the impact of research results by increasing the number of citations. The presented strategies aim to achieve long-term value and usage of the collected data, preventing digital obsolescence, promoting harmonisation compatibility with future EU initiatives (including the UNCAN data frame) and potentially for commercial exploitation.

Next steps

- Set up of the Consortium Agreement
- Completion of ethical and legal assessment for data and sample sharing, along with set up of material and data transfer agreements
- Implementation of operational RedCap and Weave data lake infrastructure
- Generation of project results as foreseen in WP3-5 and storage in data lake
- Provide training and documentation to the relevant users and raise safety awareness
- Design and make operational a long-term data management plan

1 Introduction

GLIOMATCH is a Horizon Europe (HEU) project uniting 14 pan-European partners with the goal of advancing clinical outcomes for adult glioblastoma/paediatric high-grade glioma (GBM/phGG) patients by leveraging state-of-the-art technology. The project aims to pioneer targeted brain cancer treatment through effective patient stratification and personalised pairing of patient and immunotherapy and follow-up of clinical responses to treatment.

To achieve these objectives, consortium partners will integrate spatially resolved, multi-layered tissue maps using integrated single-cell multi-omics with non-invasive MRI images. The GLIOMATCH project is anticipated to generate > 300Tb of data across all stakeholders that will encompass the data lake. The data lake, the largest of its kind ever analysed, will enable the incremental addition of new data, thereby enhancing AI and deep learning models for GBM biomarkers and treatment options.

While the collection, generation and analysis of the data will encompass the 5 years duration of the project, we anticipate the usage of the data will extend beyond this term, as the GLIOMATCH data lake will be maintained and made available for collaborations between academic and industry partners.

As promoter of the GLIOMATCH project, KUL will play a pivotal role in data management and governance. This Project Data Management Plan (P-DMP) aims at the following objectives:

- Outline the **type of data** that the project will generate, collect, and process
- Define and regulate **roles and responsibilities** among all the stakeholders within the GLIOMATCH Consortium
- Allocate the appropriate **resources** for the implementation of the data management plan
- Identify and design appropriate strategies to deal with possible **ethical, legal and safety** issues
- Pursue **high quality standards** of retrospective, prospective and observational (experimental) data collection
- Design and implement **cloud infrastructure** for data collection, storage and analysis
- **Dissemination** of best practices for safe data handling and provide hands-on **training** within the GLIOMATCH Consortium
- Ensure appropriate levels of **security and accessibility** in adherence to FAIR data principles and with Horizon Europe's mandate for Research Data Management.
- Promote **compatibility and harmonisation** with future EU initiatives and commercial exploitation

This report relates to data generated outside of WP5 (Prospective Clinical Study). A complementary Clinical Data Management Report (C-DMP) outlines the processes related to data generated, collected and/or used in the clinical study (D1.2).

This is a working document that is now delivered at a very early stage of the project (M6) and, thus, only presents guiding provisions that will be updated to a final version in M48. Further updates will be made alongside periodic project reporting and evaluation.

2 Data management plan draft

2.1 Definition and regulation of roles and responsibilities

Being the scientific leader of the GLIOMATCH project, KUL leads the overall coordination of the project, including data governance and management. As such, KUL is responsible for developing a data management plan incorporating procedures, technical design and implementation to ensure GLIOMATCH data can be stored safely while being maximally reused. Specifically, KUL will conceptualize a framework for data collection and generation; design a centralized data lake with an appropriate infrastructure (as assisted and upon deployment by AA); define detailed data access conditions, roles and responsibilities, keeping ethical/legal compliance rules (as assisted by Timelex) into account while ensuring security measures (as assisted by AA). Moreover, the KUL Research Office will coordinate and finalise the implementation of the Consortium Agreement (CA), with the purpose of specifying the relationship among the parties with respect to the project, concerning the organization of the work between the parties, the management of the project and the rights and obligations of the Parties concerning inter alia liability, Access Rights and dispute resolution. Such CA will be based on the new DESCA Horizon Europe Consortium Agreement and will be revised by all partners' legal offices before signature.

Roles and responsibilities of each partner in the GLIOMATCH Consortium are outlined in Table 1.

Table 1: Roles and responsibilities

Partner	Responsibilities
KUL	• Overall coordination of the project • Design of centralized data lake • Design of data structure • Resource allocation • CA • IP
AA	• Setup, extend and maintain cloud infrastructure • Train consortium stakeholders • Data harmonisation and connectivity with other European data sources, including the UNCAN platform
Timelex	• Legal/ethical framework of data sharing and access policy
KUL-UZL, EMC, OUS, FSJD, HSJD, RMC, UDUS, FINCB, UU	• Collection of clinical, pathological and available MRI and NGS data

2.2 Allocation of resources

The costs for making data FAIR include the fees for ensuring open access to publications and research data, computational cloud services, storage solutions, and the associated personnel costs. All data will be managed and uploaded to the respective repositories by the data owners, following the provisions set out above and as required by the repository. For the duration of the GLIOMATCH project, all these costs are covered by the project budget as part of the Horizon Europe grant. For the time after the end of the project, currently no decisions on resource allocation have been made. This will be addressed in later stages of the project.

2.3 Ethics, legal and security issues

The data that will be collected, generated and used in the GLIOMATCH project include personal data (e.g., medical history as collected from medical records, radiology scans, genomic data) from patients enrolled in the retrospective and prospective studies, and are confidential and sensitive in their nature. As such, personal data will be collected and processed in accordance with national and European legislation on the protection of individuals, in particular the EU's General Data Protection Regulation 2016/679 (GDPR) and the relevant Belgian laws implementing the GDPR including the Belgian Privacy Act of 30 July 2018 on the protection of privacy in relation to the processing of personal data. Any collection, processing and disclosure of personal data, such as participant health and medical information is subject to compliance with the aforementioned personal data protection laws and will not be processed or shared except in pseudonymized form. In case personal data is transferred outside the European Economic Area, safeguards will be taken by the Sponsor to ensure that appropriate protection travels with the data in accordance with the GDPR. (https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/rules-international-data-transfers_en#documents). Before the start of the research, a study-specific protocol and other related documents will be submitted for review to the KUL EtC for authorization. Moreover, Timelex (i.e., the appointed legal consultant of the GLIOMATCH consortium) will provide the legal/ethical framework of data sharing to ensure legal and ethical compliance of the collection of the applicable health and data protection law. The Study shall not commence until such approvals have been obtained and until other relevant essential study documents, such as duly signed contract agreements, evidence of adequate study financing, informed consent forms etc. are in place. Any personal data shall be treated as confidential at all times including during collection, handling and use or processing, and the personal data (including in any electronic format) shall be stored securely at all times and with all technical and organizational security measures that would be necessary for compliance with EU and national data protection legislation (whichever is more stringent). The Sponsor shall take appropriate measures to ensure the security of all personal data and guard against unauthorized access thereto or disclosure thereof or loss or destruction while in its custody.

2.4 Data collection and generation

2.4.1 Data provenance and purpose

The GLIOMATCH project will generate different types of data: observational (WP2, WP5), experimental (WP3) and statistical/artificial intelligence modelling data (WP4).

Details for each activity and each work package are given in Table 2.

Table 2: Provenance and purpose of data collected in GLIOMATCH

WP	Data provenance	Data purpose
WP2	KUL-UZL, EMC, OUS, FSJD, HSJD, RMC, UDUS, FINCB, UU	Collection of clinical and radiological data from retrospective cohorts
WP3	KUL-LISCO	Generation of spatial multi-omics data and data analysis
WP4	UM	Generation of radio-multiomics predictive model
WP5	KUL-UZL, EMC, OUS, RMC, UDUS, FINCB	Collection of clinical and radiological data from prospective cohorts
WP6????		

2.4.2 Data types, formats, size and quality control

The GLIOMATCH project will include different types of observational, experimental and statistical data:

- Clinical data as collected from clinical records, including but not limited to medical history, pathological reports, treatment details, survival data.
- Radiological data, including MRI and CT scans and radiological reports.
- Experimental data, including genomic data, spatial transcriptomics and spatial proteomics data.
- Public data hosted in public repositories (such as NCI Data Catalog, GEO, etc.), including but not limited to genomic, transcriptomic, proteomic and radiomic data.
- Statistical/artificial intelligence models.

The GLIOMATCH project is anticipated to generate > 300Tb of data across all stakeholders within the 5 years duration of the project. The Consortium aims to use standardized, interchangeable, or open formats as much as possible to ensure the long-term usability of data and compliance with the FAIR principles. Compression methods will be applied to accommodate the substantial volume of data while ensuring effective accessibility, data integrity and affordable long-term storage. Examples of data formats that will be used include standard formats in the field of images (.CZI, .TIFF .OMETIFF, MRI images (DI-COM)), genomic/transcriptomic/epigenomic data (.fastq, .bam, .vcf), data matrices (.txt), electronic case report forms (eCRF) containing clinical data (.txt).

Ensuring high quality at every step of the research lifecycle is a priority within GLIOMATCH and will contribute to generating datasets that can more easily be re-used by others. The responsibility for ensuring the quality of data analysis locally rests with each partners' principal investigator, in many cases supported by specific technical support units acting as a quality assurance. While the ultimate accountability for the data and metadata quality lies with the data owner (the respective partner), at the consortium level there are specific processes in place to ensure that the project's internal progress as well as its outputs in the form of deliverables, publications, or other, are of high quality.

2.4.3   The GLIOMATCH data lake

We will conceptualize a Consortium cloud-based platform infrastructure for management of the data lake and spatial multi-omics analysis within the GLIOMATCH project. This will be deployed, maintained and extended by AA and will serve as the centralized data lake for all consortium members, ensuring the secure storage of all generated clinical data and facilitating seamless data management, accessibility, and sharing among stakeholders. It will also streamline the ingestion process for MRI and NGS data, supporting various data formats for upload. Additionally, the platform will host a managed environment tailored for bioinformatics data processing, analysis, and visualization, also supporting version control. These capabilities will empower researchers to engage in exploratory work and prototyping without the logistical challenges of moving large and sensitive datasets in and out of the platform and among stakeholders, ultimately enabling to achieve the objectives outlined in WP3, WP4 and WP5.

The cloud platform will comprise three core components: a REDCap component for patient and sample information collection via eCRFs, the Weave platform component for data management, visualisation and analysis, and the MRI hub component provided by UM. REDCap is a secure web application which will be hosted by KUL, used for building, and managing online surveys and databases. This will be connected to the Gliomatch data-lake upon appropriate pseudonymization and data protection, to ensure that patient- and sample-level metadata can be linked to the generated experimental data and radiological data. We will outline tailored electronic clinical record forms (eCRFs) that can be used across all 8 clinical centres to which the required pseudonymized clinical data of all the enrolled participants will be entered. For each patient and sample a unique identifier will be defined. The content of eCRFs will be determined in WP2. Clinicians will utilise REDCap to capture sample metadata, which will then be replicated to Weave. The Weave platform, provided by AA, serves as the central platform for spatial multi-omics, accessed by data viewers, spatial biology and MRI data generators, and bioinformaticians via the web API. The MRI hub focuses on in-depth analysis of radiomics models generated by UM. Currently, the pseudonymized MRI raw data upload to Weave is tentative and will require verification by providers at UM. REDCap and MRI Hub platform user instructions and tutorials will be provided by KUL and UM.

2.5   Methodologies - FAIR data

Data collection, handling, processing and transfer for the purpose of this Study will be performed in compliance with applicable regulations, guidelines for clinical studies and internal KUL procedures, under the responsibility of the Investigator.

The European Union's open science policy outlines eight primary objectives, one of which is to encourage the sharing of open data. This concept refers to data that adheres to the principles of being Findable, Accessible, Interoperable, and Reusable (FAIR). As research outcomes are fundamentally based on research data, the availability of open data is critical for verifying and repurposing scientific assertions and knowledge. The GLIOMATCH consortium is committed to managing their data in a way that respects these values throughout the entire research lifecycle, as outlined below.

For data to be of use to others, it first needs to be findable. Following IP screening, the GLIOMATCH datasets will be shared via the GLIOMATCH cloud-based platform (see sections 2.6-8), and also through scientific publications.

Under Horizon Europe, we are required to provide metadata (for datasets but also for deposited publications) that is open under a Creative Commons Public Domain Dedication (CC0) or equivalent, to safeguard legitimate interests or constraints as per the Grant and Consortium Agreements. Each openly shared dataset will be accompanied by appropriate metadata, which will also be dictated by the standards of the specific repository where the datasets will be deposited. Depending on the repository, separate metadata files may need to be submitted alongside the data, whereas in other cases specific fields for additional information will need to be completed upon dataset submission. As a general approach within the consortium, it is agreed that a Readme file written in plain text (.txt) will accompany each research dataset supporting a publication and will detail data and dataset identification, methodologies, equipment used (including brand, model and country of origin), operational conditions, date and time of collection, software needed to open the files, licencing, and other relevant information. In case tabular data is present, the Readme file will also include an explanation for the table headings. When possible, the metadata will follow a standard format.

Key elements to be included in the metadata are, at least:

- Dataset description
- Date of deposit
- Author(s)
- Venue of publication
- Embargo length
- The term 'Horizon Europe'
- Project title
- Acronym
- Grant Agreement number
- Licensing terms
- Persistent identifiers for the dataset, the authors involved in the project and, if possible, their organisations
- Persistent identifiers for related publications and other linked datasets (where applicable)
- For metadata accompanying publications only, not for datasets: persistent identifiers for any research output or any other tools and instruments needed to validate the conclusions of the publication.
- Search keywords in English and/or respecting naming conventions in the field.

More specifically, in the case of search keywords, the data owner (partner(s)) will choose the search keywords manually guided by repository and/or journal keyword guides to confirm appropriateness, using terms that are widely used and accepted in GLIOMATCH fields of research and which describe the dataset, to optimize findability by interested parties with intent on re-use.

The accessibility of data will be supported by the deposition in a trusted repository, as described above. All project partners will ensure open access to the deposited data as soon as possible, ensuring that any embargo periods applied to give time to publish or seek protection of the intellectual property will be in line with the requirements set out in the GA. If open access is not provided to the data needed to validate scientific conclusions, the respective partner will provide the access – digital or physical – needed for validation purposes. Data supporting scientific publications or datasets should be openly published under the most recent Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC0) which facilitates data reuse. A Creative Commons Public Domain Mark (PDM) or equivalent should be applied to raw research data unless the data meet the requirements to be protected by copyright/database rights.

Data interoperability consists in adherence to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins to enable data exchange and re-use between researchers, institutions, organisations, countries.

To facilitate the interoperability of its datasets, GLIOMATCH will make use of open and widely recognised file and data formats, avoiding proprietary ones when possible. Both data and metadata will follow appropriate standards and use a formal, accessible, shared, and broadly applicable language for knowledge representation, to ensure semantic interoperability. Resources such as FAIRsharing will be consulted to identify additional discipline-specific standards and in some cases, the preferred file formats and vocabularies will be dictated by the repositories that will host the data. All files will be in English and the International System of Units or field-specific conventions will be used to further ensure data exchange and re-use.

Data that is findable, accessible, and interoperable as described in the previous sections is generally fit for re-use. The provision of Readme files accompanying the datasets will support the understanding and reuse by third parties and hence also facilitate reproducibility. The licencing conditions for shared datasets will be made transparent through easy-to-use copyright licenses, such as Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0), in line with the Grant Agreement, to promote the widest possible reuse of data. Moreover, the methodology used throughout GLIOMATCH implementation will adhere to standards and standard operating procedures (SOPs) to facilitate the reuse of the GLIOMATCH datasets by others.

2.6 Storage, access and data security

The cloud-based GLIOMATCH platform is designed to be ISO:27001 and GDPR-compliant and will provide a secure environment for the storage, sharing, analysis, and visualisation of large amounts of data by all stakeholders, serving as a centralised repository. It will be accessible via web-browser, eliminating the need

for software installation. Authentication to the platform will be provided through the widely adopted OpenID Connect (OIDC) protocol, with optional identity federation to allow consortium partners to log in using their corporate/university accounts. Single Sign On (SSO) authentication will be implemented, where AA will only receive the email address of the GLIOMATCH end user (*never* the patient), and AA will not be granted admin access to the institution. Following the "principle of least privilege", users will receive minimal necessary access to the platform's relevant tasks. As a fallback, username/password logins can be provided with limited access.

Given the sensitive nature of project data, robust data security measures will be integral to the cloud platform. Operational measures, including awareness training sessions on security threats and best practices will be conducted for end users. AA's ISO:27001 certification (https://www.iso.org/standard/27001) for information security management and business continuity ensures that the core development team is well-versed in best practices and operational procedures. As part of stakeholder training, AA will provide guidelines on security best practices and processes based on its own ISO:27001-certified ISMS, to strike a pragmatic balance within the consortium between security and practical usage. AA will provide the consortium members with an end user licence agreement (EULA), which must be agreed upon before accessing the platform. The EULA will include information on security best practices, GDPR controls, and other legal requirements. Any updates to the agreement during the project will require the end user's agreement. Platform access will be revoked in case of noncompliance and/or suspicious behaviour. From an authorization standpoint, a flexible role-based access control (RBAC) system will be implemented. This system will enable stakeholders to determine data sharing settings, ranging from private access within individual teams of a consortium partner to consortium-wide data sharing to facilitate high-end integrated data analysis. Computational stakeholders will have programmatic access through AA provided SDKs.

2.7 Copyright and intellectual property

The foreground copyright and IP as related to the GLIOMATCH project, including information on who owns the data arising from research and the intellectual property rights relating to them, innovation management regarding the non-commercial and commercial exploitation of results, publication ruling, and restrictions on the reuse of third-party data will be managed according to the rules set out in the Consortium Agreement and necessary licenses will be negotiated between consortium members in collaboration with Task 8.7 and WP8.

2.8 Long-term management

The GLIOMATCH data lake storing the spatial multiomics map of tumor-host interactions, along with the linked radiomics model and detailed clinical annotation, will be the first repository of this kind in the field of GBM immunotherapy. As such, it is anticipated that the GLIOMATCH data will provide an invaluable resource for researchers beyond the 5 years duration of the project. Some foreseeable research purposes for data reuse include consultation, usage to train and validate newer prediction models and for atlas studies; exploration of the potential exploitable results and their evaluation for future non-commercial and commercial exploitation will be based on SWOT and PESTEL analyses and other relevant factors.

To valorise long-term use of the collected data and prevent digital obsolescence, as well as enable collaborations between academic and industry partners, we will host and maintain the data lake using its pre-existing cloud solution platform for data management and spatial multi-omics, promoting data harmonisation for compatibility with future EU initiatives (including the UNCAN data frame).

Although no long-term preservation plan for GLIOMATCH research data is currently in place, we anticipate that a second P-DMP (M48) will extensively cover the main aspects of this task, including decisions on which data and for which uses long-term preservation is needed, how long data needs to be preserved, specifications about repository and file formats, how preservation will be ensured, who and with which role (custodian and host) will manage the long-term data, what metadata will be made available, additional costs that come with using the repository or data archive and how these will be covered.

3 Future work

This Data Management Plan will be maintained and updated by the GLIOMATCH Consortium, following the EC Guidelines on Data Management and with contributions from all partners. The presented guiding provisions will be updated to a final version in M48 (second P-DMP, D8.6), and further updates will be made alongside periodic project reporting and evaluation (D8.3, D8.4).

Future work will address the following items:

- Set up of the Consortium Agreement ()
- Completion of ethical and legal assessment for data and sample sharing, along with set up of material and data transfer agreements (D1.8, D1.12, D2.3, D1.3, D1.6)
- Implementation of operational RedCap (D1.4, D2.1) and Weave data lake infrastructure (D1.1, D1.7, D1.11, D2.2)
- Generation of project results as foreseen in WP3-5 and storage in data lake
- Provide training and documentation to the relevant users and raise safety awareness (D1.5)
- Design and make operational a long-term data management plan (D1.3)