Data Management Plan

CausalPCa

October 16, 2025

PROJECT	
Project number:	[TBD]
Project acronym:	CausalPCa
Project name:	A Causal AI Framework for Longitudinal Modelling of Prostate Cancer

DATA MANAGEMENT PLAN	
Date:	October 16, 2025
Version:	1.0

Data Summary

Will you re-use any existing data and what will you re-use it for? Yes, the project will re-use existing data from public and proprietary sources. Public datasets like The Cancer Genome Atlas (TCGA-PRAD) and ProstateNET will be used to augment our proprietary data, enhance statistical power, and test the generalizability of our models. Proprietary clinical data from our own and collaborating institutions will form the core training dataset.

What types and formats of data will the project generate or re-use? The project will generate and re-use a variety of data types, including:

- **Imaging data:** PET/CT, MRI, SPECT/CT in DICOM format.
- **Clinical data:** Electronic Health Records (EHR), lab values (e.g., PSA), and clinical notes in various formats (e.g., HL7, CSV).
- **Histopathological data:** Whole-slide images (WSI).
- **Omics data:** Genetic, proteomic, and radiomic data.
- **Derived data:** Disentangled latent vectors, synthetic images, and structured reports generated by the AI models.

Raw data will be archived in non-proprietary formats where possible.

What is the purpose of the data generation or re-use and its relation to the objectives of the project? The purpose of the data is to train, validate, and test a novel Causal AI framework for modeling prostate cancer progression. This directly relates to the project's core objectives of creating a "digital twin" for personalized prognosis and treatment selection.

What is the expected size of the data that you intend to generate or re-use? The total amount of generated and re-used data is expected to be in the range of 200-400 TB.

What is the origin/provenance of the data, either generated or re-used? The data originates from our own clinical institution (University Medicine Magdeburg), collaborating institutions (University Medicine Halle, Charité, Rad. Sudenburg), and public archives (TCGA, EUCAIM/ProstateNET).

To whom might your data be useful ('data utility'), outside your project? The data and models will be useful to other researchers in oncology and AI, clinicians, clinical trial designers, and the EUCAIM community.

FAIR data

Making data findable, including provisions for metadata

Will data be identified by a persistent identifier? Yes, all datasets will be assigned a persistent identifier (e.g., DOI) upon deposition in a trusted repository.

Will rich metadata be provided to allow discovery? Yes, rich metadata will be created following the EUCAIM Common Data Model (CDM), based on OMOP-CDM and HL7 FHIR standards. We will map our local schema to the EUCAIM hyper-ontology. Search keywords will be included to optimize discovery.

Will metadata be offered in such a way that it can be harvested and indexed? Yes, metadata will be exposed via standard protocols (e.g., OAI-PMH) to allow harvesting and indexing by external services.

Making data accessible

Repository: Data will be deposited in the secure XNAT platform of the Data Integration Center (DIC) at the University Medicine Magdeburg. Anonymized datasets will be contributed to the EUCAIM platform.

Data: All code will be open-source. Curated, anonymized datasets will be made openly available via the EUCAIM platform. Access to raw clinical data will be restricted due to patient confidentiality, but access can be provided to authorized users under a data sharing agreement.

Metadata: Metadata will be made openly available under a CC0 license. The metadata will contain all necessary information to access the data where possible. Metadata will be preserved even if the data is no longer available.

Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow? We will follow the EUCAIM Common Data Model (CDM), which is based on OMOP-CDM and HL7 FHIR standards. This will ensure interoperability within and across disciplines.

Will your data include qualified references to other data? Yes, the data will include qualified references to other relevant datasets, both within the project and from external sources, to enrich the contextual knowledge.

Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use? Comprehensive documentation, including readme files, codebooks, and information on methodology, will be provided alongside the data.

Will your data be made freely available in the public domain? Anonymized data will be licensed under a permissive license (e.g., CC-BY) to permit the widest re-use possible.

Describe all relevant data quality assurance processes. Data undergoes a rigorous curation and verification process by trained clinicians to ensure it is relevant, representative, and as free of errors as possible.

Other research outputs

Other research outputs, such as software, workflows, and models, will be version-controlled and made available through the project's GitLab repository, provided by the DIC. They will be licensed under an open-source license (e.g., Apache 2.0).

Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project? Costs for data management, storage, and sharing are included in the project budget under the personnel costs for the data scientist and the equipment costs for the storage solution.

Who will be responsible for data management in your project? A dedicated data scientist will be responsible for the day-to-day data management. The project coordinator is ultimately responsible for the data management plan and its execution.

How will long term preservation be ensured? Long-term preservation will be ensured by the DIC at the University Medicine Magdeburg, which has a mandate for sustainable data storage.

Data security

What provisions are or will be in place for data security? All data will be managed within a secure, learning-ready environment based on the XNAT platform. Access to sensitive data will be restricted and personalized, secured by authentication. Data is stored and backed up on the long-term storage cluster of the DIC.

Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? An ethical approval for using the preprocessed data for scientific purposes has

already been obtained. All data processing is covered by approvals from our institutional ethics committee and is fully compliant with the General Data Protection Regulation (GDPR) and the German Federal Data Protection Act.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data? Yes, informed consent procedures will be followed for all newly collected data, ensuring that patients are aware of how their data will be used, shared, and preserved.

Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? Yes, we will adhere to the data handling guidelines of the DFG (code of conduct), Helmholtz Society, and the Leibniz Foundation, as well as the internal RDM policies of the University Medicine Magdeburg.